Numerička linearna algebra

Vježbe

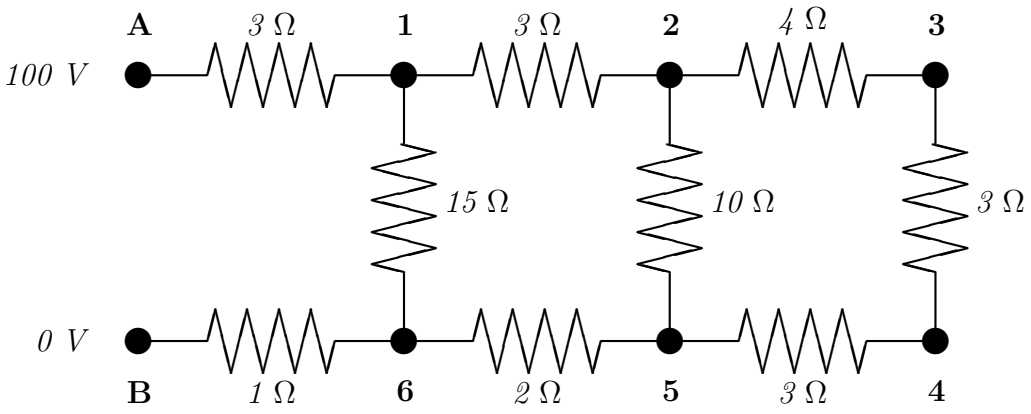
1 Sustavi linearnih jednadžbi

1.1 Gaussove eliminacije i LU faktorizacija

1.1.1 Opis problema

Započet ćemo sa primjerom iz prakse, u kojem se pojavljuje problem rješavanja sustava linearnih jednadžbi. To će nam biti motivacija za proučavanje danog problema, naime sustavi linearnih jednadžbi pojavljuje se kao posljedica rješavanja mnogih problema u fizici, kemiji, biologiji, strojarstvu, gradevini . . .

Primjer 1.1 *Problem od kojeg polazimo je računanje potencijala u električnoj mreži prikazanoj na Slici 1.*



Slika 1: Električna mreža

Otpori u otpornicima su dani na slici, a razlika potencijala izmedu točaka **A** *i* **B** *je 100 V. Iz* Ohmovog zakona *slijedi da je jakost struje I_{pq} koja struji od točke p do točke q u grani mreže pq, dana sa*

$$I_{pq} = \frac{v_p - v_q}{R_{pq}},$$

gdje su v_p i v_q potencijali u točkama p i q, a R_{pq} je otpor grane pq. Prema Kirchoffovom zakonu*, suma jakosti struja koje završavaju u jednom čvoru mora biti jednaka nuli, i to vrijedi za svaki čvor mreže. To je zapravo zakon*

sačuvanja naboja, i on ukazuje na to da se struja ne može akumulirati niti generirati u bilo kojem čvoru mreže. Primjena tih dvaju zakona na čvor **1** *vodi do sljedeće jedadžbe*

$$I_{A1} + I_{21} + I_{61} = \frac{100 - v_1}{3} + \frac{v_2 - v_1}{3} + \frac{v_6 - v_1}{15} = 0$$

ili u sređenom obliku

$$11v_1 - 5v_2 - v_6 = 500.$$

Na sličan način mogu se napisati jednažbe za svaki preostali čvor u mreži, čime dobivamo sustav od 6 jedadžbi, sa 6 nepoznanica. Nepoznanice su v_i*,* $i = 1, \ldots, 6$*, potencijali u čvorovima.*

$$\begin{array}{rcrcrcrcrcr}
11v_1 & - & 5v_2 & & & & & - & v_6 & = & 500 \\
-20v_1 & + & 41v_2 & - & 15v_3 & & - & 6v_5 & & = & 0 \\
 & - & 3v_2 & + & 7v_3 & - & 4v_4 & & & = & 0 \\
 & & & - & v_3 & + & 2v_4 & - & v_5 & = & 0 \\
 & - & 3v_2 & & & - & 10v_4 & + & 28v_5 & - & 15v_6 & = & 0 \\
-2v_1 & & & & & & & - & 15v_5 & + & 47v_6 & = & 0
\end{array}$$

Dakle problem smo sveli na rješavanje sustava, kojeg matrično možemo zapisati kao $Av = b$*, pri čemu je* $A \in \mathbb{R}^{6\times 6}$ *matrica, a* $v, b \in \mathbb{R}^6$ *vektori, i oni su oblika*

$$\underbrace{\begin{bmatrix} 11 & -5 & 0 & 0 & 0 & -1 \\ -20 & 41 & -15 & 0 & -6 & 0 \\ 0 & -3 & 7 & -4 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & -3 & 0 & -10 & 28 & -15 \\ -2 & 0 & 0 & 0 & -15 & 47 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{bmatrix}}_{v} = \underbrace{\begin{bmatrix} 500 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{b}.$$

Primjer 1.2 *Zadani su parovi točaka* (x_i, y_i)*,* $i = 0, \ldots, n$*, gdje su* $y_i = f(x_i)$ *izmjerene (ili na neki drugi način dobivene) vrijednosti funkcije* $f(x)$ *koju želimo aproksimirati polinomom* $p(x)$ *stupnja* n*. Pretpostavljamo i da je* $x_i \neq x_j$ *za* $i \neq j$*. Kriterij za odabir polinoma* $p(x)$ *je da u točkama* x_i *ima vrijednost* $p(x_i) = y_i = f(x_i)$*. (Govorimo o* **interpolacijskom** *polinomu.) Ako* $p(x)$ *prikažemo u kanonskom obliku*

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n = \sum_{j=0}^{n} a_j x^j$$

onda treba odrediti koeficijente $a_0, \ldots, a_n$ tako da vrijede uvjeti interpolacije $p(x_i) = y_i$, $i = 0, \ldots, n$, tj.

$$\begin{aligned}
a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_{n-1} x_0^{n-1} + a_n x_0^n &= y_0 \\
a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_{n-1} x_1^{n-1} + a_n x_1^n &= y_1 \\
\vdots \quad \vdots \quad \vdots \quad \cdots \quad \vdots \quad \vdots \quad \vdots & \\
a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_{n-1} x_i^{n-1} + a_n x_i^n &= y_i \\
a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_{n-1} x_n^{n-1} + a_n x_n^n &= y_n.
\end{aligned}$$

Vidimo da svaki uvjet interpolacije daje jednu jednadžbu u kojoj se nepoznati koeficijenti pojavljuju linearno. Sve jednadžbe čine sustav linearnih jednadžbi kojeg možemo matrično zapisati u obliku $Va = y$, tj.

$$\underbrace{\begin{bmatrix}
1 & x_0 & x_0^2 & x_0^3 & \cdots & x_0^{n-1} & x_0^n \\
1 & x_1 & x_1^2 & x_1^3 & \cdots & x_1^{n-1} & x_1^n \\
 & & \cdots & & \cdots & & \\
1 & x_i & x_i^2 & x_i^3 & \cdots & x_i^{n-1} & x_i^n \\
 & & \cdots & & \cdots & & \\
1 & x_n & x_n^2 & x_n^3 & \cdots & x_n^{n-1} & x_n^n
\end{bmatrix}}_{V}
\underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix}}_{a}
=
\underbrace{\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix}}_{y}. \tag{1}$$

Matrica V zove se Vandermondeova matrica. Zahvaljujući njenom specijalnom obliku, moguće je efikasno odrediti $a = V^{-1}y$.

Kao što vidimo u primjerima, a ima ih još puno takvih, sustavi linearnih jednadžbi se pojavljuju vrlo često, zato je jedan od osnovnih problema numeričke matematike rješavanje sustava linearnih jednadžbi. U ovom poglavlju istraživat ćemo metode za rješavanje kvadratnih $n \times n$ sustava, tj. sustava s n jednadžbi i n nepoznanica,

$$\begin{aligned}
a_{11} x_1 + a_{12} x_2 + \cdots + a_{1j} x_j + \cdots + a_{1n} x_n &= b_1 \\
a_{21} x_1 + a_{22} x_2 + \cdots + a_{2j} x_j + \cdots + a_{2n} x_n &= b_2 \\
\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots & \\
a_{i1} x_1 + a_{i2} x_2 + \cdots + a_{ij} x_j + \cdots + a_{in} x_n &= b_i \\
\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots & \\
a_{n1} x_1 + a_{n2} x_2 + \cdots + a_{nj} x_j + \cdots + a_{nn} x_n &= b_n
\end{aligned}$$

Matrica $A = [a_{ij}]_{i,j=1}^n \in \mathbb{R}^{n \times n}$ je **matrica sustava**, a njeni elementi su **koeficijenti sustava**. Vektor $b = [b_i]_{i=1}^n \in \mathbb{R}^n$ je **vektor desne strane**

sustava. Treba odrediti **vektor nepoznanica** $x = [x_i]_{i=1}^n \in \mathbb{R}^n$ tako da vrijedi $Ax = b$.

Kada je matrica A kvadratna i regularna, onda znamo da je problem teorijski vrlo lagano riješiti: $x = A^{-1}b$, međutim, ako to želimo riješiti na računalu tada će se pojaviti neki problemi.

- Usprkos brzini današnjih računala, kod računanja sustava velikih dimenzija, vrijeme izvršavanja može postati neprimjereno dugo. Naročito kada koristimo Gaussovu metodu eliminacija, koja dolazi do rješenja u $O(n^3)$ elementarnih operacija $(+, -, *, /)$.
- Rezultati koje računalo izračuna mogu biti netočni, a neki čak i s neprihvatljivo velikom greškom.

Zato su razvijene razne metode za rješavanje sustava linearnih jednadžbi,kako bi se minimizirali gore navedeni problemi.

1.1.2 Gaussove eliminacije i LU faktorizacija

Primjer 1.3 *Riješimo sljedeći sustav jednadžbi:*

$$\begin{array}{rcl} 5x_1 + x_2 + 4x_3 = & 19 & \\ 10x_1 + 4x_2 + 7x_3 = & 39 & \equiv \\ -15x_1 + 5x_2 - 9x_3 = & -32 & \end{array} \quad \underbrace{\begin{bmatrix} 5 & 1 & 4 \\ 10 & 4 & 7 \\ -15 & 5 & -9 \end{bmatrix}}_{A=[a_{ij}]_{i,j=1}^3} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 19 \\ 39 \\ -32 \end{bmatrix}}_{b=[b_i]_{i=1}^3}. \tag{2}$$

Koristimo metodu supstitucija, odnosno eliminacija. Prvo iz prve jednadžbe izrazimo x_1 pomoću x_2 i x_3, te to uvrstimo u zadnje dvije jednadžbe, koje postaju dvije jednadžbe s dvije nepoznanice (x_2 i x_3). Dobivamo

$$x_1 = \frac{1}{5}(19 - x_2 - 4x_3),$$

pa druga jednadžba sada glasi

$$\frac{10}{5}(19 - x_2 - 4x_3) + 4x_2 + 7x_3 = 39,$$

tj.

$$-\frac{10}{5}(x_2 + 4x_3) + 4x_2 + 7x_3 = 39 + \left(-\frac{10}{5}19\right).$$

Dakle, efekt ove transformacije je ekvivalentno prikazan kao rezultat množenja prve jednadžbe s

$$-\frac{a_{21}}{a_{11}} = -\frac{10}{5} = -2$$

i zatim njenim dodavanjem (pribrajanjem) drugoj jednadžbi. Druga jednadžba sada glasi

$$2x_2 - x_3 = 1.$$

Ako ovu transformaciju sustava zapišemo matrično, imamo

$$\underbrace{\begin{bmatrix} 5 & 1 & 4 \\ 10 & 4 & 7 \\ -15 & 5 & -9 \end{bmatrix}}_{A} \quad \mapsto \quad \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{L^{(2,1)}} \underbrace{\begin{bmatrix} 5 & 1 & 4 \\ 10 & 4 & 7 \\ -15 & 5 & -9 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ -15 & 5 & -9 \end{bmatrix}}_{A^{(1)} = \left[a_{ij}^{(1)}\right]_{i,j=1}^{3}}.$$

Nepoznanicu x_1 eliminiramo iz zadnje jednadžbe ako prvu pomnožimo s

$$-\frac{a_{31}^{(1)}}{a_{11}^{(1)}} = -\frac{-15}{5} = 3$$

i onda je pribrojimo zadnjoj. To znači sljedeću promjenu matrice koeficijenata:

$$\underbrace{\begin{bmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ -15 & 5 & -9 \end{bmatrix}}_{A^{(1)}} \quad \mapsto \quad \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}}_{L^{(3,1)}} \underbrace{\begin{bmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ -15 & 5 & -9 \end{bmatrix}}_{A^{(1)}} = \underbrace{\begin{bmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ 0 & 8 & 3 \end{bmatrix}}_{A^{(2)} = \left[a_{ij}^{(2)}\right]_{i,j=1}^{3}}.$$

Vektor desne strane je u ove dvije transformacije promijenjen u

$$\underbrace{\begin{bmatrix} 19 \\ 39 \\ -32 \end{bmatrix}}_{b} \quad \mapsto \quad \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{L^{(2,1)}} \begin{bmatrix} 19 \\ 39 \\ -32 \end{bmatrix} = \underbrace{\begin{bmatrix} 19 \\ 1 \\ -32 \end{bmatrix}}_{b^{(1)}}$$

$$\mapsto \quad \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}}_{L^{(3,1)}} \begin{bmatrix} 19 \\ 1 \\ -32 \end{bmatrix} = \underbrace{\begin{bmatrix} 19 \\ 1 \\ 25 \end{bmatrix}}_{b^{(2)}}.$$

Novi, ekvivalentni, sustav je $A^{(2)}x = b^{(2)}$, tj.

$$\begin{aligned} 5x_1 + \ x_2 + 4x_3 &= 19 \\ 2x_2 - \ x_3 &= 1 \\ 8x_2 - 3x_3 &= 25, \end{aligned} \tag{3}$$

u kojem su druga i treća jednadžba sustav od dvije jednadžbe s dvije nepoznanice. Očito je da rješenje $x = (x_1, x_2, x_3)^\tau$ sustava (2) zadovoljava i sustav (3). Obratno, ako trojka x_1, x_2, x_3 zadovoljava (3), onda množenjem prve jednadžbe u (3) s 2 i zatim pribrajanjem drugoj jednadžbi, dobijemo drugu jednadžbu sustava (2). Na sličan način iz prve i treće jednadžbe sustava (3) rekonstruiramo treću jednadžbu polaznog sustava (2). U tom smislu kažemo da su sustavi (2) i (3) ekvivalentni: imaju isto rješenje.

Nadalje, primijetimo da smo proces eliminacija (tj. izražavanja nepoznanice x_1 pomoću x_2 i x_3 i eliminiranjem x_1 iz zadnje dvije jednadžbe) jednostavno opisali matričnim operacijama. Eliminaciju nepoznanice x_1 smo prikazali kao rezultat množenja matrice koeficijenata i vektora desne strane s lijeva jednostavnim matricama $L^{(2,1)}$ i $L^{(3,1)}$.

Jasno je da je sustav (3) jednostavniji od polaznog. Zato sada nastavljamo s primjenom iste strategije: iz treće jednadžbe eliminiramo x_2 tako što drugu jednadžbu pomnožimo s

$$-\frac{a_{32}^{(2)}}{a_{22}^{(2)}} = -4$$

i pribrojimo je trećoj. Tako treća jednadžba postaje $7x_3 = 21$, a cijeli sustav ima oblik

$$\begin{aligned} 5x_1 + \ x_2 + 4x_3 &= 19 \\ 2x_2 - \ x_3 &= 1 \\ 7x_3 &= 21. \end{aligned}$$

Transformaciju eliminacije x_2 iz treće jednadžbe možemo matrično zapisati kao transformaciju matrice koeficijenata

$$\underbrace{\begin{bmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ 0 & 8 & 3 \end{bmatrix}}_{A^{(2)}} \mapsto \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -4 & 1 \end{bmatrix}}_{L^{(3,2)}} \underbrace{\begin{bmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ 0 & 8 & 3 \end{bmatrix}}_{A^{(2)}} = \underbrace{\begin{bmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ 0 & 0 & 7 \end{bmatrix}}_{A^{(3)}} \tag{4}$$

i transformaciju vektora desne strane

$$\underbrace{\begin{bmatrix} 19 \\ 1 \\ 25 \end{bmatrix}}_{b^{(2)}} \mapsto \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -4 & 1 \end{bmatrix}}_{L^{(3,2)}} \underbrace{\begin{bmatrix} 19 \\ 1 \\ 25 \end{bmatrix}}_{b^{(2)}} = \underbrace{\begin{bmatrix} 19 \\ 1 \\ 21 \end{bmatrix}}_{b^{(3)} = \left[b_i^{(3)}\right]_{i=1}^{3}}. \tag{5}$$

Sustav (4), koji je ekvivalentan polaznom, lako riješimo.

1. *Iz treće jednadžbe je* $x_3 = \dfrac{21}{7} = 3.$

2. *Iz druge jednadžbe je* $x_2 = \dfrac{1}{2}\,(1 + x_3) = 2.$

3. *Iz prve jednadžbe je* $x_1 = \dfrac{1}{5}\,(19 - x_2 - 4x_3) = 1.$

Jednostavna provjera potvrđuje da su x_1, x_2, x_3 *rješenja polaznog sustava (2).*

Analizirajmo postupak rješavanja u prethodnom primjeru. Pažljivo pogledajmo oblik matrica u realaciji

$$A^{(3)} = L^{(3,2)} L^{(3,1)} L^{(2,1)} A.$$

Matrica $A^{(3)}$ je gornjetrokutasta, a produkt $L^{(3,2)} L^{(3,1)} L^{(2,1)}$ je donjetrokutasta matrica. Dakle, polaznu matricu A smo množenjem slijeva donjetrokutastom matricom načinili gornjetrokutastom. To možemo pročitati i ovako:

$$A = LA^{(3)}, \quad L = (L^{(2,1)})^{-1} (L^{(3,1)})^{-1} (L^{(3,2)})^{-1},$$

gdje je L donjetrokutasta matrica. Lako se provjerava da je

$$L = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{(L^{(2,1)})^{-1}} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix}}_{(L^{(3,1)})^{-1}} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 4 & 1 \end{bmatrix}}_{(L^{(3,2)})^{-1}} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -3 & 4 & 1 \end{bmatrix}.$$

Dakle, matricu A smo napisali kao produkt donjetrokutaste i gornjetrokutaste matrice, $A = LA^{(3)}$. Gornjetrokutastu matricu u ovom kontekstu obično označavamo s $U = A^{(3)}$, pa je A rastavljena na produkt $A = LU$. Govorimo o **LU faktorizaciji** matrice A (neki tu faktorizaciju zovu i LR faktorizacija matrice). Uočimo da je računanje produkta koji definira matricu L jednostavno. Inverze matrica $L^{(2,1)}$, $L^{(3,1)}$ i $L^{(3,2)}$ dobijemo samo promjenom predznaka netrivijalnih elemenata u donjem trokutu, a cijeli produkt jednostavno je stavljanje tih elemenata na odgovarajuće pozicije u matrici L. Sada još primijetimo da je relacija (4) zapravo linearni sustav $Ux = b^{(3)}$, gdje je $b^{(3)} = L^{(3,2)} L^{(3,1)} L^{(2,1)} b = L^{-1} b$. Jasno,

$$x = A^{-1} b = (LU)^{-1} b = U^{-1} L^{-1} b.$$

Dakle, u terminima matrice A i vektora b, linearni sustav u primjeru 1.3 riješen je metodom koja se sastoji od tri glavna koraka.

1. Matricu sustava A treba faktorizirati u obliku $A = LU$, gdje je L donjetrokutasta, a U gornjetrokutasta matrica.

2. Rješavanjem donjetrokutastog sustava $Ly = b$ treba odrediti vektor $y = L^{-1}b$.

3. Rješavanjem gornjetrokutastog sustava $Ux = y$ treba odrediti vektor $x = U^{-1}y = U^{-1}(L^{-1}b)$.

1.1.3 Trokutasti sustavi: rješavanje supstitucijama unaprijed i unazad

Trokutasti sustavi jednadžbi lako se rješavaju. Pogledajmo, na primjer, donjetrokutasti sustav $Lx = b$ dimenzije $n = 4$:

$$\begin{bmatrix} \ell_{11} & 0 & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} & 0 \\ \ell_{41} & \ell_{42} & \ell_{43} & \ell_{44} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}.$$

Neka je matrica L regularna. To znači da su $\ell_{ii} \neq 0$ za $i = 1, 2, 3, 4$. Očito je

$$\begin{aligned} x_1 &= \frac{b_1}{\ell_{11}} \\ x_2 &= \frac{1}{\ell_{22}}(b_2 - \ell_{21}x_1) \\ x_3 &= \frac{1}{\ell_{33}}(b_3 - \ell_{31}x_1 - \ell_{32}x_2) \\ x_4 &= \frac{1}{\ell_{44}}(b_4 - \ell_{41}x_1 - \ell_{42}x_2 - \ell_{43}x_3). \end{aligned}$$

Vidimo da x_1 možemo odmah izračunati, a za $i > 1$ formula za x_i je funkcija od b_i, i-tog retka matrice L i nepoznanica $x_1, \ldots, x_{i-1}$ koje su prethodno već izračunate. Dakle, prvo izračunamo x_1, pa tu vrijednost uvrstimo u izraz koji daje x_2; zatim x_1 i x_2 uvrstimo u izraz za x_3, itd. Ovakav postupak zovemo **supstitucija unaprijed**.

Algoritam 1.1 *Rješavanje linearnog sustava jednadžbi $Lx = b$ s regularnom donjetrokutastom matricom $L \in \mathbb{R}^{n\times n}$.*

/* Supstitucija unaprijed za $Lx = b$ */

$x_1 = \dfrac{b_1}{\ell_{11}}$;

za $i = 2, \ldots, n$ {

$$x_i = \left(b_i - \sum_{j=1}^{i-1} \ell_{ij} x_j \right) \Big/ \ell_{ii}; \}$$

Prebrojimo operacije u gornjem algoritmu:

- dijeljenja: n ;
- množenja: $1 + 2 + \cdots + (n-1) = \dfrac{1}{2}\, n(n-1)$;
- zbrajanja i oduzimanja: $1 + 2 + \cdots + (n-1) = \dfrac{1}{2}\, n(n-1)$.

Dakle, ukupna složenost je $O(n^2)$.

Gornjetrokutaste sustave rješavamo na sličan način. Ako je sustav $Ux = b$ oblika

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}, \quad \prod_{i=1}^{4} u_{ii} \neq 0,$$

onda, polazeći od zadnje jednadžbe unazad, imamo

$$\begin{aligned} x_4 &= \frac{b_4}{u_{44}} \\ x_3 &= \frac{1}{u_{33}} (b_3 - u_{34} x_4) \\ x_2 &= \frac{1}{u_{22}} (b_2 - u_{23} x_3 - u_{24} x_4) \\ x_1 &= \frac{1}{u_{11}} (b_1 - u_{12} x_2 - u_{13} x_3 - u_{14} x_4). \end{aligned}$$

Ovakav postupak zovemo **supstitucija unazad**.

Algoritam 1.2 *Rješavanje linearnog sustava jednadžbi $Ux = b$ s regularnom gornjetrokutastom matricom $U \in \mathbb{R}^{n \times n}$.*

/∗ Supstitucija unazad za $Ux = b$ ∗/

$x_n = \dfrac{b_n}{u_{nn}}$;

za $i = n-1, \ldots, 1$ {

$$x_i = \left(b_i - \sum_{j=i+1}^{n} u_{ij} x_j \right) \Big/ u_{ii}; \}$$

Kao i kod supstitucija naprijed, složenost ovog algoritma je $O(n^2)$.

1.1.4 LU faktorizacija sa i bez pivotiranja

Sada kad smo uočili da se rješavanje linearnog sustava $Ax = b$ faktoriziranjem matrice A svodi na trokutaste sustave, ostaje nam posebno proučiti faktorizaciju matrice $A \in \mathbb{R}^{n\times n}$ na produkt donje i gornjetrokutaste matrice. Zanima nas proizvoljna dimenzija n, ali ćemo zbog jednostavnosti razmatranja na početku sve ideje ilustrirati na primjeru $n = 5$. Neka je

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix}.$$

Sjetimo se, eliminacija prve nepoznanice manifestira se poništavanjem koeficijenata na pozicijama $(2,1),(3,1),\ldots,(n,1)$. To možemo napraviti u jednom potezu[1]. Ako definiramo matricu

$$L^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -\dfrac{a_{21}}{a_{11}} & 1 & 0 & 0 & 0 \\ -\dfrac{a_{31}}{a_{11}} & 0 & 1 & 0 & 0 \\ -\dfrac{a_{41}}{a_{11}} & 0 & 0 & 1 & 0 \\ -\dfrac{a_{51}}{a_{11}} & 0 & 0 & 0 & 1 \end{bmatrix},$$

onda je x_1 eliminiran iz svih jednadžbi osim prve, tj.

$$A^{(1)} \equiv L^{(1)}A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & a_{34}^{(1)} & a_{35}^{(1)} \\ 0 & a_{42}^{(1)} & a_{43}^{(1)} & a_{44}^{(1)} & a_{45}^{(1)} \\ 0 & a_{52}^{(1)} & a_{53}^{(1)} & a_{54}^{(1)} & a_{55}^{(1)} \end{bmatrix}.$$

Objasnimo oznake koje koristimo za elemente matrice $A^{(1)}$. Općenito, elementi $A^{(1)}$ označeni su s $a_{ij}^{(1)}$, $1 \le i, j \le n$. Međutim, elementi prvog retka u

[1] U primjeru 1.3 smo zbog jednostavnosti poništavali koeficijente jedan po jedan.

$A^{(1)}$ jednaki su prvom retku u A, $a^{(1)}_{1j} = a_{1j}$, $1 \leq j \leq n$, pa smo to eksplicitno naznačili u zapisu matrice $A^{(1)}$.

Primijetimo da je transformaciju $A \mapsto A^{(1)}$ moguće izvesti samo ako je

$$a_{11} \neq 0. \tag{6}$$

Također, lako se uvjerimo da je

$$(L^{(1)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \dfrac{a_{21}}{a_{11}} & 1 & 0 & 0 & 0 \\ \dfrac{a_{31}}{a_{11}} & 0 & 1 & 0 & 0 \\ \dfrac{a_{41}}{a_{11}} & 0 & 0 & 1 & 0 \\ \dfrac{a_{51}}{a_{11}} & 0 & 0 & 0 & 1 \end{bmatrix},$$

te da iz $A = (L^{(1)})^{-1}A^{(1)}$ slijedi

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \dfrac{a_{21}}{a_{11}} & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ 0 & a^{(1)}_{22} \end{bmatrix}.$$

Jednostavno, dobili smo faktorizaciju vodeće 2×2 podmatrice od A. Uvjet za izvod ove faktorizacije bio je (6). Stavimo

$$\alpha_2 \equiv \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a^{(1)}_{22}.$$

Ako je $\alpha_2 \neq 0$, onda je i $a^{(1)}_{22} \neq 0$ pa je dobro definirana matrica

$$L^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -\dfrac{a^{(1)}_{32}}{a^{(1)}_{22}} & 1 & 0 & 0 \\ 0 & -\dfrac{a^{(1)}_{42}}{a^{(1)}_{22}} & 0 & 1 & 0 \\ 0 & -\dfrac{a^{(1)}_{52}}{a^{(1)}_{22}} & 0 & 0 & 1 \end{bmatrix} \quad \text{i njen inverz} \quad (L^{(2)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & \dfrac{a^{(1)}_{32}}{a^{(1)}_{22}} & 1 & 0 & 0 \\ 0 & \dfrac{a^{(1)}_{42}}{a^{(1)}_{22}} & 0 & 1 & 0 \\ 0 & \dfrac{a^{(1)}_{52}}{a^{(1)}_{22}} & 0 & 0 & 1 \end{bmatrix}.$$

Vrijedi

$$A^{(2)} \equiv L^{(2)}A^{(1)} = L^{(2)}L^{(1)}A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} & a_{45}^{(2)} \\ 0 & 0 & a_{53}^{(2)} & a_{54}^{(2)} & a_{55}^{(2)} \end{bmatrix}. \qquad (7)$$

Uočimo da oznake u relaciji (7) naglašavaju da je u matrici $A^{(2)} = \left[a_{ij}^{(2)}\right]_{i,j=1}^{n}$ prvi redak jednak prvom retku matrice A, a drugi redak jednak drugom retku matrice $A^{(1)}$. Ako sada u relaciji $A = (L^{(1)})^{-1}(L^{(2)})^{-1}A^{(2)}$ izračunamo produkt $(L^{(1)})^{-1}(L^{(2)})^{-1}$ dobivamo

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{a_{21}}{a_{11}} & 1 & 0 & 0 & 0 \\ \frac{a_{31}}{a_{11}} & \frac{a_{32}^{(1)}}{a_{22}^{(1)}} & 1 & 0 & 0 \\ \frac{a_{41}}{a_{11}} & \frac{a_{42}^{(1)}}{a_{22}^{(1)}} & 0 & 1 & 0 \\ \frac{a_{51}}{a_{11}} & \frac{a_{52}^{(1)}}{a_{22}^{(1)}} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} & a_{45}^{(2)} \\ 0 & 0 & a_{53}^{(2)} & a_{54}^{(2)} & a_{55}^{(2)} \end{bmatrix}, \qquad (8)$$

odakle zaključujemo da vrijedi

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{a_{21}}{a_{11}} & 1 & 0 \\ \frac{a_{31}}{a_{11}} & \frac{a_{32}^{(1)}}{a_{22}^{(1)}} & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} \\ 0 & 0 & a_{33}^{(2)} \end{bmatrix}.$$

Dakle, ako je $a_{11} \neq 0$ i $\alpha_2 \neq 0$, onda smo dobili trokutastu faktorizaciju vodeće 3×3 podmatrice od A. Stavimo

$$\alpha_3 \equiv \det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = a_{11}a_{22}^{(1)}a_{33}^{(2)}.$$

Ako je $\alpha_3 \neq 0$ onda je i $a_{33}^{(2)} \neq 0$ pa su dobro definirane matrice

$$L^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{a_{43}^{(2)}}{a_{33}^{(2)}} & 1 & 0 \\ 0 & 0 & -\frac{a_{53}^{(2)}}{a_{33}^{(2)}} & 0 & 1 \end{bmatrix}, \quad (L^{(3)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{a_{43}^{(2)}}{a_{33}^{(2)}} & 1 & 0 \\ 0 & 0 & \frac{a_{53}^{(2)}}{a_{33}^{(2)}} & 0 & 1 \end{bmatrix},$$

i vrijedi

$$A^{(3)} \equiv L^{(3)}A^{(2)} = L^{(3)}L^{(2)}L^{(1)}A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} & a_{45}^{(3)} \\ 0 & 0 & 0 & a_{54}^{(3)} & a_{55}^{(3)} \end{bmatrix}.$$

Ako izračunamo produkt $(L^{(1)})^{-1}(L^{(2)})^{-1}(L^{(3)})^{-1}$, onda vidimo da vrijedi

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{a_{21}}{a_{11}} & 1 & 0 & 0 & 0 \\ \frac{a_{31}}{a_{11}} & \frac{a_{32}^{(1)}}{a_{22}^{(1)}} & 1 & 0 & 0 \\ \frac{a_{41}}{a_{11}} & \frac{a_{42}^{(1)}}{a_{22}^{(1)}} & \frac{a_{43}^{(2)}}{a_{33}^{(2)}} & 1 & 0 \\ \frac{a_{51}}{a_{11}} & \frac{a_{52}^{(1)}}{a_{22}^{(1)}} & \frac{a_{53}^{(2)}}{a_{33}^{(2)}} & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} & a_{45}^{(3)} \\ 0 & 0 & 0 & a_{54}^{(3)} & a_{55}^{(3)} \end{bmatrix},$$

te da je

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{a_{21}}{a_{11}} & 1 & 0 & 0 \\ \frac{a_{31}}{a_{11}} & \frac{a_{32}^{(1)}}{a_{22}^{(1)}} & 1 & 0 \\ \frac{a_{41}}{a_{11}} & \frac{a_{42}^{(1)}}{a_{22}^{(1)}} & \frac{a_{43}^{(2)}}{a_{33}^{(2)}} & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} \end{bmatrix}.$$

Ponovo zaključujemo na isti način: definiramo

$$\alpha_4 \equiv \det \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = a_{11} a_{22}^{(1)} a_{33}^{(2)} a_{44}^{(3)}.$$

Ako je $\alpha_4 \neq 0$, onda je i $a_{44}^{(3)} \neq 0$, pa su dobro definirane matrice

$$L^{(4)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -\frac{a_{54}^{(3)}}{a_{44}^{(3)}} & 1 \end{bmatrix}, \quad (L^{(4)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{a_{54}^{(3)}}{a_{44}^{(3)}} & 1 \end{bmatrix}. \quad (9)$$

Lako provjerimo da vrijedi

$$A^{(4)} \equiv L^{(4)} A^{(3)} = L^{(4)} L^{(3)} L^{(2)} L^{(1)} A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} & a_{45}^{(3)} \\ 0 & 0 & 0 & 0 & a_{55}^{(4)} \end{bmatrix}.$$

te da je, nakon računanja produkta $(L^{(1)})^{-1}(L^{(2)})^{-1}(L^{(3)})^{-1}(L^{(4)})^{-1}$,

$$A = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{a_{21}}{a_{11}} & 1 & 0 & 0 & 0 \\ \frac{a_{31}}{a_{11}} & \frac{a_{32}^{(1)}}{a_{22}^{(1)}} & 1 & 0 & 0 \\ \frac{a_{41}}{a_{11}} & \frac{a_{42}^{(1)}}{a_{22}^{(1)}} & \frac{a_{43}^{(2)}}{a_{33}^{(2)}} & 1 & 0 \\ \frac{a_{51}}{a_{11}} & \frac{a_{52}^{(1)}}{a_{22}^{(1)}} & \frac{a_{53}^{(2)}}{a_{33}^{(2)}} & \frac{a_{54}^{(3)}}{a_{44}^{(3)}} & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} & a_{45}^{(3)} \\ 0 & 0 & 0 & 0 & a_{55}^{(4)} \end{bmatrix}}_{U}. \qquad (10)$$

Vidimo da je izvedivost operacija koje su dovele do faktorizacije $A = LU$ ovisila o uvjetima

$$a_{11} \neq 0, \quad a_{22}^{(1)} \neq 0, \quad a_{33}^{(2)} \neq 0, \quad a_{44}^{(3)} \neq 0.$$

Također, uočili smo da su ti uvjeti osigurani ako su u matrici A determinante glavnih podmatrica dimenzija $1, 2, \ldots, n-1$ različite od nule. To je u našem primjeru značilo uvjete

$$\alpha_1 \equiv a_{11} \neq 0, \quad \alpha_2 \neq 0, \quad \alpha_3 \neq 0, \quad \alpha_4 \neq 0.$$

Brojeve a_{11}, $a_{22}^{(1)}$, $a_{33}^{(2)}$, $a_{44}^{(3)}$ zovemo **pivotni elementi** ili kratko **pivoti**. Brojevi α_1, α_2, α_3, α_4 su **glavne minore** matrice A.

Dakle, možemo zaključiti sljedeće:

- *Ako je prvih $n-1$ minora matrice A različito od nule, onda su i svi pivotni elementi različiti od nule i Gaussove eliminacije daju LU faktorizaciju matrice A.*

U tom slučaju sljedeći algoritam računa faktorizaciju $A = LU$.

Algoritam 1.3 *Računanje LU faktorizacije matrice A.*

$L = I$;
za $k = 1, \ldots, n-1$ {
  za $j = k+1, \ldots, n$ {

$$\ell_{jk} = \frac{a_{jk}^{(k-1)}}{a_{kk}^{(k-1)}};$$

$a_{jk}^{(k)} = 0;\ \}$
za $j = k+1, \ldots, n$ $\{$
za $i = k+1, \ldots, n$ $\{$
$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \ell_{ik} a_{kj}^{(k-1)};\ \}\ \}\ \}$
$U = A^{(n-1)} = \left[a_{ij}^{(n-1)} \right].$

Jedan, očit problem, s LU faktorizacijom koju smo opisali u prethodnom odjeljku je da za njeno računanje, prema opisanom algoritmu, matrica A mora imati specijalnu strukturu: sve njene glavne podmatrice do uključivo reda $n-1$ moraju biti regularne. Sljedeći primjer ilustrira taj problem.

Primjer 1.4 *Neka je matrica sustava $Ax = b$ dana s*

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

Ova matrica je regularna, $\det A = -1$*, pa sustav uvijek ima rješenje, ali A očito nema LU faktorizaciju, jer*

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \ell_{21} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}$$

povlači da je

$$\begin{aligned} 1 \cdot u_{11} &= 0 \\ 1 \cdot u_{12} &= 1 \\ \ell_{21} \cdot u_{11} &= 1 \\ \ell_{21} \cdot u_{12} + u_{22} &= 1. \end{aligned}$$

Iz prve jednadžbe odmah vidimo da mora biti $u_{11} = 0$, a iz treće slijedi da $\ell_{21} \cdot 0 = 1$, što je nemoguće.

S druge strane, matrica A reprezentira linearni sustav

$$\begin{aligned} 0x_1 + x_2 &= b_1 \\ x_1 + x_2 &= b_2 \end{aligned}$$

koji uvijek ima rješenje $x_1 = b_2 - b_1$, $x_2 = b_1$, i kojeg možemo ekvivalentno zapisati kao[2]

$$x_1 + x_2 = b_2$$

[2]Zamjena redoslijeda jednadžbi ne mijenja rješenje sustava.

$$0x_1 + x_2 \;=\; b_1.$$

Matrica ovog sustava je

$$A' = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

i očito ima jednostavnu LU faktorizaciju s $L = I$, $U = A'$. Vezu između A i A' zapisujemo matrično:

$$A' = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}}_{A}.$$

Matricu P zovemo **matrica permutacije** *ili jednostavno permutacija. Njeno djelovanje na matricu A je jednostavno permutiranje redaka.*

Da bismo ilustrirali kako zamjenama redaka uvijek možemo dobiti LU faktorizaciju, vratimo se našem 5×5 primjeru i pogledajmo npr. relacije (7), (8):

$$A^{(2)} \equiv L^{(2)}A^{(1)} = L^{(2)}L^{(1)}A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} & a_{45}^{(2)} \\ 0 & 0 & a_{53}^{(2)} & a_{54}^{(2)} & a_{55}^{(2)} \end{bmatrix},$$

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \dfrac{a_{21}}{a_{11}} & 1 & 0 & 0 & 0 \\ \dfrac{a_{31}}{a_{11}} & \dfrac{a_{32}^{(1)}}{a_{22}^{(1)}} & 1 & 0 & 0 \\ \dfrac{a_{41}}{a_{11}} & \dfrac{a_{42}^{(1)}}{a_{22}^{(1)}} & 0 & 1 & 0 \\ \dfrac{a_{51}}{a_{11}} & \dfrac{a_{52}^{(1)}}{a_{22}^{(1)}} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} & a_{45}^{(2)} \\ 0 & 0 & a_{53}^{(2)} & a_{54}^{(2)} & a_{55}^{(2)} \end{bmatrix}.$$

Neka je $a_{33}^{(2)} = 0$. Dakle, više ne možemo kao ranije definirati $L^{(3)}$. Pogledajmo elemente $a_{43}^{(2)}$ i $a_{53}^{(2)}$. Ako su oba jednaka nuli, onda možemo staviti $L^{(3)} = I$ i nastaviti dalje, jer je cilj transformacije $L^{(3)}$ poništiti $a_{43}^{(2)}$ i $a_{53}^{(2)}$. Ako su oni već jednaki nuli onda u ovom koraku ne treba ništa raditi pa je transformacija jednaka jediničnoj matrici. Neka je sada npr. $a_{53}^{(2)} \neq 0$. Ako definiramo matricu

$$P^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \quad \text{onda je} \quad P^{(3)}A^{(2)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{53}^{(2)} & a_{54}^{(2)} & a_{55}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} & a_{45}^{(2)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \end{bmatrix}.$$

Sada možemo definirati matrice

$$L^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\dfrac{a_{43}^{(2)}}{a_{53}^{(2)}} & 1 & 0 \\ 0 & 0 & -\dfrac{a_{33}^{(2)}}{a_{53}^{(2)}} & 0 & 1 \end{bmatrix}, \quad (L^{(3)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \dfrac{a_{43}^{(2)}}{a_{53}^{(2)}} & 1 & 0 \\ 0 & 0 & \dfrac{a_{33}^{(2)}}{a_{53}^{(2)}} & 0 & 1 \end{bmatrix}$$

i postići

$$A^{(3)} \equiv L^{(3)}P^{(3)}A^{(2)} = L^{(3)}P^{(3)}L^{(2)}L^{(1)}A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{53}^{(2)} & a_{54}^{(2)} & a_{55}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} & a_{45}^{(3)} \\ 0 & 0 & 0 & a_{54}^{(3)} & a_{55}^{(3)} \end{bmatrix}.$$

Primijetimo da je treći redak matrice $A^{(3)}$ jednak petom retku matrice $A^{(2)}$. Za sljedeći korak eliminacija provjeravamo vrijednost $a_{44}^{(3)}$. Ako je $a_{44}^{(3)} \neq 0$, postupamo kao i ranije, tj. definiramo matricu $L^{(4)}$ kao u relaciji (9). Ako je $a_{44}^{(3)} = a_{54}^{(3)} = 0$, onda možemo staviti $L^{(4)} = I$, jer je u tom slučaju $A^{(3)}$ već

gornjetrokutasta. Neka je $a_{44}^{(3)} = 0$, ali $a_{54}^{(3)} \neq 0$, tako da $L^{(4)}$ nije definirana. Lako provjerimo da permutacijska matrica

$$P^{(4)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad \text{daje} \quad P^{(4)}A^{(3)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{53}^{(2)} & a_{54}^{(2)} & a_{55}^{(2)} \\ 0 & 0 & 0 & a_{54}^{(3)} & a_{55}^{(3)} \\ 0 & 0 & 0 & a_{44}^{(3)} & a_{45}^{(3)} \end{bmatrix}.$$

Kako je po pretpostavci $a_{44}^{(3)} = 0$, možemo staviti $L^{(4)} = I$ i matrica $U = L^{(4)}P^{(4)}A^{(3)}$ je gornjetrokutasta. Sve zajedno, vrijedi relacija

$$U = L^{(4)}P^{(4)}L^{(3)}P^{(3)}L^{(2)}L^{(1)}A.$$

Vidjeli smo ranije da je množenje inverza trokutastih matrica $L^{(k)}$ jednostavno. Međutim, mi sada imamo permutacijske matrice između, pa ostaje istražiti kako one djeluju na strukturu produkta. Primijetimo,

$$P^{(4)}L^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{a_{43}^{(2)}}{a_{53}^{(2)}} & 1 & 0 \\ 0 & 0 & -\frac{a_{33}^{(2)}}{a_{53}^{(2)}} & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{a_{33}^{(2)}}{a_{53}^{(2)}} & 0 & 1 \\ 0 & 0 & -\frac{a_{43}^{(2)}}{a_{53}^{(2)}} & 1 & 0 \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{a_{33}^{(2)}}{a_{53}^{(2)}} & 1 & 0 \\ 0 & 0 & -\frac{a_{43}^{(2)}}{a_{53}^{(2)}} & 0 & 1 \end{bmatrix}}_{\widetilde{L}^{(3)}} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} = \widetilde{L}^{(3)}P^{(4)}.$$

Dakle, $P^{(4)}$ možemo **prebaciti s lijeve na desnu stranu od** $L^{(3)}$, ako u $L^{(3)}$ permutiramo elemente ispod dijagonale u trećem stupcu. Tako dobivena

matrica $\widetilde{L}^{(3)}$ ima istu strukturu kao i $L^{(3)}$. Na isti način je $P^{(3)}L^{(2)}L^{(1)} = \widetilde{L}^{(2)}\widetilde{L}^{(1)}P^{(3)}$ i $P^{(4)}\widetilde{L}^{(2)}\widetilde{L}^{(1)} = \widetilde{\widetilde{L}}^{(2)}\widetilde{\widetilde{L}}^{(1)}P^{(4)}$ pa je

$$U = L^{(4)}P^{(4)}L^{(3)}P^{(3)}L^{(2)}L^{(1)}A = L^{(4)}\widetilde{L}^{(3)}\widetilde{\widetilde{L}}^{(2)}\widetilde{\widetilde{L}}^{(1)}P^{(4)}P^{(3)}A,$$

tj.

$$\underbrace{P^{(4)}P^{(3)}}_{P}A = \underbrace{(L^{(4)})^{-1}(\widetilde{L}^{(3)})^{-1}(\widetilde{\widetilde{L}}^{(2)})^{-1}(\widetilde{\widetilde{L}}^{(1)})^{-1}}_{L}U.$$

Produkt koji definira matricu L je iste strukture kao i ranije, dakle, imamo jednostavno slaganje odgovarajućih elemenata. Nadalje matrica

$$P = P^{(4)}P^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

je opet matrica permutacije.

Jasno je kako bi ovaj postupak izgledao općenito. Na kraju eliminacija bi vrijedilo

$$U = A^{(n-1)} = L^{(n-1)}P^{(n-1)}(\cdots(L^{(3)}P^{(3)}\underbrace{(L^{(2)}P^{(2)}\underbrace{\underbrace{(L^{(1)}P^{(1)}A)}_{A^{(1)}}}_{A^{(2)}})}_{A^{(3)}})\cdots), \quad (11)$$

i $P = P^{(n-1)}P^{(n-2)}\cdots P^{(2)}P^{(1)}$, gdje neke od permutacija $P^{(k)}$ mogu biti jednake identitetama (jediničnim matricama). Ova metoda naziva se **Gaussove eliminacije sa parcijalnim pivotiranjem**.

Ilustrirajmo opisanu proceduru jednim numeričkim primjerom.

Primjer 1.5 *Neka je*

$$A = \begin{bmatrix} 1 & 1 & 4 & 1 \\ 2 & 1 & 1 & 6 \\ 5 & 1 & 1 & 0 \\ 1 & 4 & 1 & 3 \end{bmatrix}.$$

Najveći element u prvom stupcu od A je na poziciji $(3,1)$ – to znači da prvi pivot maksimiziramo ako zamijenimo prvi i treći redak od A. Tu zamjenu realizira permutacija $P^{(1)}$, gdje je

$$P^{(1)} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad P^{(1)}A = \begin{bmatrix} 5 & 1 & 1 & 0 \\ 2 & 1 & 1 & 6 \\ 1 & 1 & 4 & 1 \\ 1 & 4 & 1 & 3 \end{bmatrix}.$$

Sada definiramo

$$L^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{2}{5} & 1 & 0 & 0 \\ -\frac{1}{5} & 0 & 1 & 0 \\ -\frac{1}{5} & 0 & 0 & 1 \end{bmatrix}, \quad \textit{pa je} \quad A^{(1)} = L^{(1)}P^{(1)}A = \begin{bmatrix} 5 & 1 & 1 & 0 \\ 0 & \frac{3}{5} & \frac{3}{5} & 6 \\ 0 & \frac{4}{5} & \frac{19}{5} & 1 \\ 0 & \frac{19}{5} & \frac{4}{5} & 3 \end{bmatrix}.$$

Sljedeći pivot je maksimiziran permutacijom $P^{(2)}$, gdje je

$$P^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad P^{(2)}A^{(1)} = \begin{bmatrix} 5 & 1 & 1 & 0 \\ 0 & \frac{19}{5} & \frac{4}{5} & 3 \\ 0 & \frac{4}{5} & \frac{19}{5} & 1 \\ 0 & \frac{3}{5} & \frac{3}{5} & 6 \end{bmatrix}.$$

Sljedeći korak eliminacija glasi

$$L^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -\frac{4}{19} & 1 & 0 \\ 0 & -\frac{3}{19} & 0 & 1 \end{bmatrix}, \quad A^{(2)} = L^{(2)}P^{(2)}A^{(1)} = \begin{bmatrix} 5 & 1 & 1 & 0 \\ 0 & \frac{19}{5} & \frac{4}{5} & 3 \\ 0 & 0 & \frac{69}{19} & \frac{7}{19} \\ 0 & 0 & \frac{9}{19} & \frac{105}{19} \end{bmatrix}.$$

Sljedeća permutacija je identiteta, $P^{(3)} = I$, pa u zadnjem koraku imamo

$$L^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{9}{69} & 1 \end{bmatrix}, \quad A^{(3)} = L^{(3)}P^{(3)}A^{(2)} = \begin{bmatrix} 5 & 1 & 1 & 0 \\ 0 & \frac{19}{5} & \frac{4}{5} & 3 \\ 0 & 0 & \frac{69}{19} & \frac{7}{19} \\ 0 & 0 & 0 & \frac{7182}{1311} \end{bmatrix}.$$

Sada primijetimo da je $A^{(3)} = L^{(3)} I L^{(2)} P^{(2)} L^{(1)} P^{(1)} A$, *gdje je*

$$P^{(2)} L^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{5} & 0 & 0 & 1 \\ -\frac{1}{5} & 0 & 1 & 0 \\ -\frac{2}{5} & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{5} & 1 & 0 & 0 \\ -\frac{1}{5} & 0 & 1 & 0 \\ -\frac{2}{5} & 0 & 0 & 1 \end{bmatrix} P^{(2)} = \widetilde{L}^{(1)} P^{(2)}.$$

Dakle, $U \equiv A^{(3)} = L^{(3)} L^{(2)} \widetilde{L}^{(1)} P^{(2)} P^{(1)} A$. *Ako stavimo* $P = P^{(2)} P^{(1)}$, *onda vrijedi*

$$\begin{aligned} PA &= \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 4 & 1 \\ 2 & 1 & 1 & 6 \\ 5 & 1 & 1 & 0 \\ 1 & 4 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 5 & 1 & 1 & 0 \\ 1 & 4 & 1 & 3 \\ 1 & 1 & 4 & 1 \\ 2 & 1 & 1 & 6 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{5} & 1 & 0 & 0 \\ \frac{1}{5} & 0 & 1 & 0 \\ \frac{2}{5} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{4}{19} & 1 & 0 \\ 0 & \frac{3}{19} & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{9}{69} & 1 \end{bmatrix} U \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{5} & 1 & 0 & 0 \\ \frac{1}{5} & \frac{4}{19} & 1 & 0 \\ \frac{2}{5} & \frac{3}{19} & \frac{9}{69} & 1 \end{bmatrix} \begin{bmatrix} 5 & 1 & 1 & 0 \\ 0 & \frac{19}{5} & \frac{4}{5} & 3 \\ 0 & 0 & \frac{69}{19} & \frac{7}{19} \\ 0 & 0 & 0 & \frac{7182}{1311} \end{bmatrix}. \end{aligned}$$

Dakle, možemo zaključiti sljedeće:

- *Za proizvoljnu* $n \times n$ *matricu* A *postoji permutacija* P *tako da Gaussove eliminacije daju* LU *faktorizaciju od* PA, *tj.* $PA = LU$, *gdje je* L *donjetrokutasta matrica s jedinicama na dijagonali, a* U *je gornjetrokutasta matrica. Permutaciju* P *možemo odabrati tako da su svi elementi matrice* L *po apsolutnoj vrijednosti najviše jednaki jedinici.*

Pogledajmo kako algoritam možemo implementirati na računalu s minimalnim korištenjem dodatnog memorijskog prostora. Prisjetimo se našeg 5×5 primjera i relacije (10):

$$A = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{a_{21}}{a_{11}} & 1 & 0 & 0 & 0 \\ \frac{a_{31}}{a_{11}} & \frac{a_{32}^{(1)}}{a_{22}^{(1)}} & 1 & 0 & 0 \\ \frac{a_{41}}{a_{11}} & \frac{a_{42}^{(1)}}{a_{22}^{(1)}} & \frac{a_{43}^{(2)}}{a_{33}^{(2)}} & 1 & 0 \\ \frac{a_{51}}{a_{11}} & \frac{a_{52}^{(1)}}{a_{22}^{(1)}} & \frac{a_{53}^{(2)}}{a_{33}^{(2)}} & \frac{a_{54}^{(3)}}{a_{44}^{(3)}} & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} & a_{45}^{(3)} \\ 0 & 0 & 0 & 0 & a_{55}^{(4)} \end{bmatrix}}_{U}.$$

Vidimo da je za spremanje svih elemenata matrica L i U dovoljno n^2 varijabli (lokacija u memoriji), dakle onoliko koliko zauzima originalna matrica A. Ako pažljivo pogledamo proces računanja LU faktorizacije, uočavamo da ga možemo izvesti tako da matrica U ostane zapisana u gornjem trokutu matrice A, a strogo donji trokut matrice L bude napisan na mjestu elemenata strogo donjeg trokuta polazne matrice A. Kako matrica L po definiciji ima jedinice na dijagonali, te elemente ne treba nigdje posebno zapisivati. Na taj način se elementi polazne matrice gube, a računanje možemo shvatiti kao promjenu sadržaja polja A koje sadrži matricu A:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix} \quad \mapsto \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ \frac{a_{21}}{a_{11}} & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ \frac{a_{31}}{a_{11}} & \frac{a_{32}^{(1)}}{a_{22}^{(1)}} & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ \frac{a_{41}}{a_{11}} & \frac{a_{42}^{(1)}}{a_{22}^{(1)}} & \frac{a_{43}^{(2)}}{a_{33}^{(2)}} & a_{44}^{(3)} & a_{45}^{(3)} \\ \frac{a_{51}}{a_{11}} & \frac{a_{52}^{(1)}}{a_{22}^{(1)}} & \frac{a_{53}^{(2)}}{a_{33}^{(2)}} & \frac{a_{54}^{(3)}}{a_{44}^{(3)}} & a_{55}^{(4)} \end{bmatrix}.$$

Sve matrice $A^{(k)}$, $k = 1, 2, \ldots, n-1$ su pohranjene u istom $n \times n$ polju koje na početku sadrži matricu $A \equiv A^{(0)}$. Na ovaj način zapis algoritma 1.3 postaje još jednostavniji i elegantniji.

Algoritam 1.4 *Računanje LU faktorizacije matrice A bez dodatne memorije.*

za $k = 1, \ldots, n-1$ {
  za $j = k+1, \ldots, n$ {

$$A(j,k) = \frac{A(j,k)}{A(k,k)};\ \}$$

  za $j = k+1, \ldots, n$ {
    za $i = k+1, \ldots, n$ {
      $A(i,j) = A(i,j) - A(i,k)A(k,j);\ \}\ \}\ \}$

Primijetimo da smo koristili oznake uobičajene u programskim jezicima – element matrice (dvodimenzionalnog polja) označili smo s $A(i,j)$. Isto tako, vidimo da konkretna realizacija algoritma na računalu uključuje dodatne trikove i modifikacije kako bi se što racionalnije koristili resursi računala (npr. memorija).

Na kraju ovog odjeljka, pokažimo kako cijeli algoritam na računalu možemo implementirati bez dodatne memorije. Kako smo prije vidjeli, LU faktorizaciju možemo napraviti tako da L i U smjestimo u matricu A. Sada još primijetimo da sustave $Ly = b$ i $Ux = y$ možemo riješiti tako da y i x u memoriju zapisujemo na mjesto vektora b. Tako dobijemo sljedeću implementaciju Gaussovih eliminacija:

Algoritam 1.5 *Rješavanje sustava jednadžbi $Ax = b$ Gaussovim eliminacijama bez dodatne memorije.*

/* LU faktorizacija, $A = LU$ */
za $k = 1, \ldots, n-1$ {
  za $j = k+1, \ldots, n$ {

$$A(j,k) = \frac{A(j,k)}{A(k,k)};\ \}$$

  za $j = k+1, \ldots, n$ {
    za $i = k+1, \ldots, n$ {
      $A(i,j) = A(i,j) - A(i,k)A(k,j);\ \}\ \}\ \}$
/* Rješavanje sustava $Ly = b$, y napisan na mjesto b. */
za $i = 2, \ldots, n$ {
  za $j = 1, \ldots, i-1$ {
    $b(i) = b(i) - A(i,j)b(j);\ \}\ \}$
/* Rješavanje sustava $Ux = y$, x napisan na mjesto b. */

$$b(n) = \frac{b(n)}{A(n,n)};$$

za $i = n-1, \ldots, 1$ {
  za $j = i+1, \ldots, n$ {
    $b(i) = b(i) - A(i,j)b(j)$; }
  $b(i) = b(i)/A(i,i)$; }

Zadatak 1.1 *Implementirajte Gaussove eliminacije bez pivotiranja i algoritam za rješavanje linearnih sustava pomoću Gaussovih elimininacija u* Octave*-i. Dokumentaciju o* Octave*-i možete naći na adresi http://www.gnu.org/software/octave/docs.html.*

1.1.5 Numerička svojstva Gaussovih eliminacija

Računalo je ograničen, konačan stroj. Imamo ograničenu količinu memorijskog prostora u kojem možemo držati polazne podatke, međurezultate i rezultate računanja[3]. Umjesto skupa realnih brojeva $\mathbb{R}$ imamo njegovu aproksimaciju pomoću konačno mnogo prikazivih brojeva (realni brojevi koje računalo koristi su zapravo konačan skup razlomaka) što znači da računske operacije ne možemo izvršavati niti točno niti rezultat možemo po volji dobro aproksimirati. Koristimo svojstvo operacija u aritmetici konačne preciznosti:

$$x \odot y = \mathrm{fl}(x \circ y) = \mathrm{round}(x \circ y) = (x \circ y)(1+\delta), \quad |\delta| \leq \varepsilon,$$

gdje su $\circ \in \{+, -, \times, /\}$, $\odot \in \{\oplus, \ominus, \otimes, \oslash\}$, a x i y su reprezentabilni brojevi.

Primjer 1.6 *Neka je α mali parametar, $|\alpha| \ll 1$, i neka je matrica A definirana s*

$$A = \begin{bmatrix} \alpha & 1 \\ 1 & 1 \end{bmatrix}.$$

U egzaktnom računanju imamo

$$L^{(2,1)} = \begin{bmatrix} 1 & 0 \\ -\dfrac{1}{\alpha} & 1 \end{bmatrix}, \quad L^{(2,1)}A = \begin{bmatrix} \alpha & 1 \\ 0 & 1 - \dfrac{1}{\alpha} \end{bmatrix},$$

pa je LU faktorizacija matrice A dana s

$$\underbrace{\begin{bmatrix} \alpha & 1 \\ 1 & 1 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 1 & 0 \\ \dfrac{1}{\alpha} & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} \alpha & 1 \\ 0 & 1 - \dfrac{1}{\alpha} \end{bmatrix}}_{U}.$$

[3]Svaka operacija zahtijeva izvjesno vrijeme izvršavanja pa je ukupno trajanje algoritma također važan faktor. U ovom odjeljku prvenstveno ćemo analizirati problem točnosti.

Pretpostavimo sada da ovaj račun provodimo na računalu u aritmetici s 8 decimalnih znamenki, tj. točnosti $\varepsilon \approx 10^{-8}$. Neka je $|\alpha| < \varepsilon$, npr. neka je $\alpha = 10^{-10}$. Kako je problem jednostavan, vrijedi

$$\begin{aligned}
\tilde{\ell}_{21} &= \ell_{21}(1+\epsilon_1), \quad |\epsilon_1| \le \varepsilon, \\
\tilde{u}_{11} &= u_{11}, \\
\tilde{u}_{12} &= u_{12}, \\
\tilde{u}_{22} &= 1 \ominus 1 \oslash \alpha = -1 \oslash \alpha = -\frac{1}{\alpha}(1+\epsilon_1).
\end{aligned}$$

Primijetimo da je

$$\left| \frac{\tilde{u}_{22} - u_{22}}{u_{22}} \right| \le \frac{2\varepsilon}{1-\varepsilon}.$$

Dakle svi elementi matrica $\widetilde{L}$ i $\widetilde{U}$ izračunati su s malom relativnom pogreškom. Sjetimo se da je ovaj primjer najavljen kao primjer numeričke nestabilnosti procesa eliminacija, odnosno LU faktorizacije. Gdje je tu nestabilnost ako su svi izračunati elementi matrica $\widetilde{L}$ i $\widetilde{U}$ gotovo jednaki točnim vrijednostima? Odstupanje (relativna greška) je najviše reda veličine dvije greške zaokruživanja – gdje je onda problem?

Izračunajmo (egzaktno) $\widetilde{L}\widetilde{U}$:

$$\widetilde{L}\widetilde{U} = \begin{bmatrix} 1 & 0 \\ 1 \oslash \alpha & 1 \end{bmatrix} \begin{bmatrix} \alpha & 1 \\ 0 & -1 \oslash \alpha \end{bmatrix} = \begin{bmatrix} \alpha & 1 \\ 1 & 0 \end{bmatrix} = \underbrace{\begin{bmatrix} \alpha & 1 \\ 1 & 1 \end{bmatrix}}_{A} + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}}_{\delta A}.$$

Primijetimo da δA ne možemo smatrati malom perturbacijom polazne matrice A – jedan od najvećih elemenata u matrici A, $a_{22} = 1$, je promijenjen u nulu. Ako bismo koristeći $\widetilde{L}$ i $\widetilde{U}$ pokušali riješiti linearni sustav $Ax = b$, zapravo bismo radili na sustavu $(A+\delta A)x = b$. Tek da dobijemo osjećaj kako katastrofalno loš rezultat možemo dobiti, pogledajmo linearne sustave

$$\begin{bmatrix} \alpha & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \begin{bmatrix} \alpha & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Njihova rješenja su

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \dfrac{-1}{\alpha - 1} \\ \dfrac{2\alpha - 1}{\alpha - 1} \end{bmatrix} \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 - 2\alpha \end{bmatrix}.$$

Vidimo da se x_1 i $\tilde{x}_1$ potpuno razlikuju. Zaključujemo da Gaussove eliminacije mogu biti numerički nestabilne – dovoljna je jedna greška zaokruživanja

"u krivo vrijeme na krivom mjestu" pa da dobiveni rezultat bude potpuno netočan.

Napomena 1.1 *I ovaj primjer zaslužuje komentar. Vidimo da katastrofalno velika greška nije uzrokovana akumuliranjem velikog broja grešaka zaokruživanja. Cijeli problem je u samo jednoj aritmetičkoj operaciji (pri računanju $\tilde{u}_{22}$) koja je zapravo izvedena jako točno, s malom greškom zaokruživanja.*

Cilj numeričke analize algoritma je da otkrije moguće uzroke nestabilnosti, objasni fenomene vezane za numeričku nestabilnost i ponudi rješenja za njihovo uklanjanje.

Nestabilnost ilustrirana primjerom u skladu je s teoremom s predavanja, koji tvrdi sljedeće.

- *Neka je algoritam 1.4 primijenjen na matricu $A \in \mathbb{R}^{n\times n}$ i neka su uspješno izvršene sve njegove operacije. Ako su $\widetilde{L}$ i $\widetilde{U}$ izračunati trokutasti faktori, onda je*

$$\widetilde{L}\widetilde{U} = A + \delta A, \quad |\delta A| \leq \frac{n\varepsilon}{1-n\varepsilon}(|A| + |\widetilde{L}|\,|\widetilde{U}|) \leq \frac{2n\varepsilon}{1-2n\varepsilon}\,|\widetilde{L}|\,|\widetilde{U}|,$$

  gdje prva nejednakost vrijedi za $n\varepsilon < 1$, a druga za $2n\varepsilon < 1$.

Naime, ako izračunamo $|\widetilde{L}|\,|\widetilde{U}|$ dobijemo

$$|\widetilde{L}|\,|\widetilde{U}| = \begin{bmatrix} |\alpha| & 1 \\ 1+\epsilon & 2|1 \oslash \alpha| \end{bmatrix},$$

gdje je $1+\epsilon = \alpha(1 \oslash \alpha)$, $|\epsilon| \leq \varepsilon$. Kako je na poziciji $(2,2)$ u matrici $|\widetilde{L}|\,|\widetilde{U}|$ element koji je reda veličine $1/|\alpha| > 1/\varepsilon$, vidimo da nam teorem ne može garantirati mali δA.

Jasno nam je da je, zbog nenegativnosti matrica $|\widetilde{L}|$ i $|\widetilde{U}|$, mali produkt $|\widetilde{L}|\,|\widetilde{U}|$ moguć samo ako su elementi od $\widetilde{L}$ i $\widetilde{U}$ mali po apsolutnoj vrijednosti. Pogledajmo nastavak primjera 1.6.

Primjer 1.7 *Neka je A matrica iz primjera 1.6. Zamijenimo joj poredak redaka,*

$$A' = PA = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} \alpha & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \alpha & 1 \end{bmatrix}.$$

LU faktorizacija matrice $A' = LU$ je

$$\begin{bmatrix} 1 & 1 \\ \alpha & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \alpha & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 1-\alpha \end{bmatrix}.$$

Ako je $|\alpha| < \varepsilon$, onda su izračunate matrice

$$\widetilde{L} = \begin{bmatrix} 1 & 0 \\ \alpha & 1 \end{bmatrix}, \quad \widetilde{U} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

i vrijedi

$$\widetilde{L}\widetilde{U} = \begin{bmatrix} 1 & 1 \\ \alpha & 1+\alpha \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1 \\ \alpha & 1 \end{bmatrix}}_{A'} + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & \alpha \end{bmatrix}}_{\delta A'}, \quad |\delta A'| \leq \varepsilon |A|.$$

Primijetimo i da je produkt

$$|\widetilde{L}|\,|\widetilde{U}| = \begin{bmatrix} 1 & 1 \\ |\alpha| & 1+\alpha \end{bmatrix}$$

po elementima istog reda veličine kao i $|A'|$. Dakle, u ovom primjeru je bilo dovoljno zamijeniti poredak redaka u A (redoslijed jednadžbi) pa da imamo garantirano dobru faktorizaciju u smislu da je $\widetilde{L}\widetilde{U} = A' + \delta A'$ s malom perturbacijom $\delta A'$.

Zadatak 1.2 *U* Octave*-i izračunajte LU faktorizaciju za matrice iz primjera 1.6 i 1.7, te riješite sa njima sustave za $b = [1,2]^T$, odnosno $b = [2,1]^T$. Uzmite da je $\alpha = eps/4$.*

Matrične norme

Definicija 1.1 *Preslikavanje $\nu : \mathbb{C}^{m\times n} \longrightarrow \mathbb{R}$ je* **matrična norma** *na $\mathbb{C}^{m\times n}$ ako zadovoljava sljedeće uvjete:*

1. *$\nu(A) \geq 0$, za svako $A \in \mathbb{C}^{m\times n}$*
2. *$\nu(A) = 0$ ako i samo ako je $A = 0$*
3. *$\nu(\alpha A) = |\alpha|\nu(A)$, za $\alpha \in \mathbb{C}$, $A \in \mathbb{C}^{m\times n}$*
4. *$\nu(A+B) \leq \nu(A) + \nu(B)$, za sve $A, b \in \mathbb{C}^{m\times n}$*

(1.–2.) $\rightarrow$ pozitivna definitnost, (3.) $\rightarrow$ homogenost, (4.) $\rightarrow$ nejednakost trokuta.

Definicija 1.2 *Neka su μ, ν i ρ matrične norme na $\mathbb{C}^{m\times n}$, $\mathbb{C}^{n\times k}$ i $\mathbb{C}^{m\times k}$ redom. One su* **konzistentne** *ako je*

$$\rho(AB) \leq \mu(A)\nu(B),$$

za svaki izbor $A \in \mathbb{C}^{m\times n}$ *i* $B \in \mathbb{C}^{n\times k}$.

Specijalno, matrična norma ν *na* $\mathbb{C}^{n\times n}$ *je* **konzistentna** *ako je*

$$\nu(AB) \leq \nu(A)\nu(B),$$

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

Napomena 1.2

- *Gornja definicija obuhvaća i konzistentnost matrične i vektorske norme, jer prirodno identificiramo* $\mathbb{C}^{n\times 1}$ *i* $\mathbb{C}^n$.
- *Ako je* ν *konzistentna matrična norma na* $\mathbb{C}^{n\times n}$ *i* A_1, A_2,...,$A_m \in \mathbb{R}^{n\times n}$ *proizvoljne matrice, indukcijom se odmah vidi da je*

  $$\nu(A_1A_2\cdots A_m) \leq \nu(A_1)\nu(A_2)\cdots\nu(A_m).$$

  Specijalno, za svako $A \in \mathbb{C}^{n\times n}$ *i* $m \in \mathbb{N}$ *je*

  $$\nu(A^m) \leq \nu(A)^m.$$

Standardna Euklidska vektorska norma ima jedno povoljno svojstvo, a to je:

$$\|Ux\|_2 = \|x\|_2, \quad x \in \mathbb{C}^n,\ U \in \mathbb{C}^{n\times n}\ U^*U = UU^* = I,$$

Budući da je U **unitarna matrica** (u ravnini to su rotacije i refleksije) ovo svojstvo se zove unitarna invarijantnost vektorske norme, pri čemu djelovanje matrice U čuva udaljenosti. Takvo svojstvo se može definirati i za matrične norme.

Definicija 1.3 *Norma* ν *na* $\mathbb{C}^{m\times n}$ *je* **unitarno invarijantna** *ako je:*

$$\nu(U^*AV) = \nu(A),$$

za sve unitarne matrice $U \in \mathbb{C}^{m\times m}$, $V \in \mathbb{C}^{n\times n}$ *i sve* $A \in \mathbb{C}^{m\times n}$.

Teorem 1.1 *Ako je* ν *konzistentna matrična norma na* $\mathbb{C}^{n\times n}$, *onda postoji vektorska norma* $\|\ \|$ *na* $\mathbb{C}^n$ *koja je konzistentna sa* ν.

Definicija 1.4 *Neka je* $A \in \mathbb{C}^{n\times n}$, *tada je sa*

$$spr(A) = \rho(A) = \max\{|\lambda| \ :\ \lambda \in \sigma(A)\}$$

definiran **spektralni radijus** *matrice* A.

Teorem 1.2 *Neka je ν konzistentna matrična norma na $\mathbb{C}^{n\times n}$. Tada za svaku matricu $A \in \mathbb{C}^{n\times n}$ vrijedi*

$$\rho(A) \leq \nu(A).$$

Teorem 1.3 *Neka je $\| \|$ proizvoljna norma na $\mathbb{C}^n$. Preslikavanje $\nu : \mathbb{C}^{n\times n} \longrightarrow \mathbb{R}$,*

$$\nu(A) = \max_{\|x\|=1} \|Ax\|,$$

za $A \in \mathbb{C}^{n\times n}$, je konzistentna matrična norma na $\mathbb{C}^{n\times n}$, konzistentna sa $\| \|$, i zove se **operatorska norma na $\mathbb{C}^{n\times n}$, inducirana vektorskom normom** $\| \|$.

Primjer 1.8 *Neka je $A = [a_{ij}] \in \mathbb{C}^{n\times n}$. Sljedeća preslikavanja definiraju konzistentne matrične norme na $\mathbb{C}^{n\times n}$.*

- $\| \|_F : \mathbb{C}^{n\times n} \longrightarrow \mathbb{R}$,

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{tr(A^*A)},$$

  zove se **Frobeniusova** *ili* **Euklidska norma**. *(Na $\mathbb{C}^{n\times 1} \cong \mathbb{C}^n$ je $\| \|_F = \| \|_2$.)*

  - *Frobeniusova matrična norma $\| \|_F$ i euklidska vektorska norma $\| \|_2$ su konzistentne jer je za $x \in \mathbb{C}^n$*

$$\|Ax\|_F \leq \|A\|_F \|x\|_F = \|A\|_F \|x\|_2.$$

  - *Frobeniusova norma $\| \|_F$ je unitarno invarijantna.*

- $\| \|_2 : \mathbb{C}^{n\times n} \longrightarrow \mathbb{R}$,

$$\|A\|_2 = \sqrt{\rho(A^*A)},$$

  zove se **spektralna norma**.

  - *Spektralna matrična norma $\| \|_2$ je operatorska norma na $\mathbb{C}^{n\times n}$ inducirana vektorskom normom $\| \|_2$*

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$

  - *Vrijedi konzistentnost s vektorskom normom, za $x \in \mathbb{C}^n$*

$$\|Ax\|_2 \leq \|A\|_2 \|x\|_2$$

– *Spektralna norma* $\| \ \|_2$ *je unitarno invarijantna.*

- $\| \ \|_1 : \mathbb{C}^{n\times n} \longrightarrow \mathbb{R}$,

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|.$$

– *Matrična norma* $\| \ \|_1$ *je operatorska norma na* $\mathbb{C}^{n\times n}$ *inducirana vektorskom normom* $\| \ \|_1$

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1$$

– *Vrijedi konzistentnost s vektorskom normom, za* $x \in \mathbb{C}^n$

$$\|Ax\|_1 \leq \|A\|_1 \|x\|_1$$

- $\| \ \|_\infty : \mathbb{C}^{n\times n} \longrightarrow \mathbb{R}$,

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

– *Matrična norma* $\| \ \|_\infty$ *je operatorska norma na* $\mathbb{C}^{n\times n}$ *inducirana vektorskom normom* $\| \ \|_\infty$

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty$$

– *Vrijedi konzistentnost s vektorskom normom, za* $x \in \mathbb{C}^n$

$$\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty$$

Sve prikazane norme mogu se definirati i na $\mathbb{C}^{m\times n}$.

Obzirom da je $\mathbb{C}^{m\times n} \cong \mathbb{C}^{mn}$ a na $\mathbb{C}^{mn}$ su sve vektorske norme ekvivalentne, to su i sve matrične norme ekvivalentne.

Napomena 1.3 *Neka su* $\| \ \|_p$ *i* $\| \ \|_q$ *matrične norme na* $\mathbb{C}^{m\times n}$, *tada je za svaku matricu* $A \in \mathbb{C}^{m\times n}$

$$\|A\|_p \leq \alpha_{pq} \|A\|_q,$$

pri čemu se jednakost dostiže, a konstante α_{pq} *su tabelirane u sljedećoj tablici.*

$\| \ \|_p$ \ $\| \ \|_q$	$\| \ \|_1$	$\| \ \|_2$	$\| \ \|_\infty$	$\| \ \|_F$
$\| \ \|_1$	1	$\sqrt{m}$	m	$\sqrt{m}$
$\| \ \|_2$	$\sqrt{n}$	1	$\sqrt{m}$	1
$\| \ \|_\infty$	n	$\sqrt{n}$	1	$\sqrt{n}$
$\| \ \|_F$	$\sqrt{n}$	$\sqrt{rang(A)}$	$\sqrt{m}$	1

Zadatak 1.3 *Zadane su dvije matrice* $A = [a_{ij}], B = [b_{ij}] \in \mathbb{R}^{5\times 5}$, *pri čemu je*

$$a_{ij} = \frac{1}{i+j-1}, \qquad A \text{ je Hilbertova matrica}$$

$$b_{ij} = diag(1,2,3,4,5), \qquad B \text{ je dijagonalna matrica}$$

Izračunajte norme $\| \ \|_1$, $\| \ \|_2$, $\| \ \|_\infty$ *i* $\| \ \|_F$, *za te matrice, i provjerite da su te norme konzistentne. Također provjerite da su* $\| \ \|_2$ *i* $\| \ \|_F$ *unitarno invarijantne, da* $\| \ \|_F$ *nije operatorska norma, a da ostale norme to jesu.*

Napomena 1.4 *Dobivamo*

$$\begin{aligned} \|A\|_1 &= 2.2833 \\ \|A\|_2 &= 1.5671 \\ \|A\|_\infty &= 2.2833 \\ \|A\|_F &= 1.5809 \\ \|B\|_1 &= 5 \\ \|B\|_2 &= 5 \\ \|B\|_\infty &= 5 \\ \|B\|_F &= 7.4162 \end{aligned}$$

i

$$\begin{aligned} 3.7282 = \|A\cdot B\|_1 &\leq \|A\|_1 \cdot \|B\|_1 = 2.2833 \cdot 5 = 11.4167 \\ 3.3455 = \|A\cdot B\|_2 &\leq \|A\|_2 \cdot \|B\|_2 = 1.5671 \cdot 5 = 7.8353 \\ 5 = \|A\cdot B\|_\infty &\leq \|A\|_\infty \cdot \|B\|_\infty = 2.2833 \cdot 5 = 11.4167 \\ 3.3690 = \|A\cdot B\|_F &\leq \|A\|_F \cdot \|B\|_F = 1.5809 \cdot 7.4162 = 11.7243 \end{aligned}$$

Da bi provjerili da $\| \ \|_F$ *nije operatorska norma, primjetimo sljedeće svojstvo operatorskih normi:*

Neka je $\| \ \|$ *operatorska norma na* $\mathbb{C}^{n\times n}$ *inducirana vektorskom normom* $\| \ \|$. *Tada je*

$$\|I\| = \max_{\|x\|=1} \|Ix\| = \max_{\|x\|=1} \|x\| = 1.$$

Nužni uvjet *da bi matrična norma bila* **operatorska** *je* $\|I\| = 1$.

Provjerimo taj uvjet za $\| \ \|_F$ *na* $\mathbb{R}^{5\times 5}$.

$$\|I\|_F = \sqrt{5} \neq 1,$$

dakle $\| \ \|_F$ *zaista nije operatorska norma.*

Za provjeru ostalih normi, uzet ćemo proizvoljni vektor norme 1 i pokazati konzistentnost sa vektorskom normom, a zatim ćemo definirati vektor za koji se postiže jednakost u definiciji operatorske norme.

$$\begin{aligned} 1.0192 = \|Ax_1\|_1 &\leq \|A\|_1 = 2.2833 \\ 0.9046 = \|Ax_2\|_2 &\leq \|A\|_2 = 1.5671 \\ 0.8832 = \|Ax_\infty\|_\infty &\leq \|A\|_\infty = 2.2833 \end{aligned}$$

- $\| \ \|_1$
  - *Biramo vektor* y_1 *takav da je* $\|Ay_1\|_1 = \|A\|_1$
  - *Definira se kao* $y_1 = e_k$*, pri čemu je* $k \in \{1, \ldots, 5\}$ *takav da je* $\|A(:,k)\|_1 = \max_{1 \leq j \leq 5} \|A(:,j)\|_1 = \|A\|_1$
  - *U našem slučaju to je*

$$y_1 = e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

  -

$$\|Ay_1\|_1 = \|A\|_1 = 2.2833$$

- $\| \ \|_2$
  - *Biramo vektor* y_2 *takav da je* $\|Ay_2\|_2 = \|A\|_2$
  - *Definira se kao svojstveni vektor norme 1, najveće svojstvene vrijednosti* $\lambda_{max}(A^*A)$ *matrice* A^*A
    $\|y_2\|_2 = 1$, $A^*Ay_2 = \lambda_{max}(A^*A)y_2$
  - *U našem slučaju to je*

$$y_2 = \begin{bmatrix} 0.7679 \\ 0.4458 \\ 0.3216 \\ 0.2534 \\ 0.2098 \end{bmatrix}$$

  -

$$\|Ay_2\|_2 = \|A\|_2 = 1.5671$$

- $\| \ \|_\infty$
  - *Biramo vektor* y_∞ *takav da je* $\|Ay_\infty\|_\infty = \|A\|_\infty$
  - *Definira se tako da za* $k \in \{1, \ldots, 5\}$ *gdje je* $\|A(k,:)\|_1 = \max_{1 \le i \le 5} \|A(i,:)\|_1 = \|A\|_\infty$

$$y_\infty(j) = \left\{ \begin{array}{ll} \frac{a_{kj}}{|a_{kj}|}, & a_{kj} \neq 0 \\ 1, & a_{kj} = 0 \end{array} \right\}, j = 1, \ldots, 5$$

  - *U našem slučaju to je za* $k = 1$

$$y_\infty = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

  -

$$\|Ay_\infty\|_\infty = \|A\|_\infty = 2.2833$$

Primjena matričnih normi

Iz teorema sa predavanja možemo zaključiti sljedeće.

- Neka je $\tilde{x}$ izračunato rješenje regularnog $n \times n$ sustava jednadžbi $Ax = b$, dobiveno Gaussovim eliminacijama bez obzira na pivotiranje. Tada postoji perturbacija ΔA za koju vrijedi

$$(A + \Delta A)\tilde{x} = b, \quad \|\Delta A\|_F \le O(n^3)\varepsilon\rho\|A\|_F,$$

  gdje je ρ **faktor rasta elemenata u LU faktorizaciji** i definiran je sa

$$\rho = \frac{\max\limits_{i,j,k} |\tilde{a}_{ij}^{(k)}|}{\max\limits_{i,j} |a_{ij}|}.$$

  Ovime je zadana ocjena za **povratnu grešku** ili **grešku unazad**.

- Za grešku $\tilde{x} - x$, uz uvjet da je $\|A^{-1}\Delta A\|_F < 1$, tada vrijedi

$$\begin{aligned} \tilde{x} - x &= (A + \Delta A)^{-1}b - x = (I + A^{-1}\Delta A)^{-1}A^{-1}b - x = \\ &= [(I + A^{-1}\Delta A)^{-1} - I]x = \\ &= [(I - A^{-1}\Delta A + (A^{-1}\Delta A)^2 - (A^{-1}\Delta A)^3 + \cdots) - I]x = \\ &= -A^{-1}\Delta A(I - A^{-1}\Delta A + (A^{-1}\Delta A)^2 - \cdots)x = \\ &= -A^{-1}\Delta A(I + A^{-1}\Delta A)^{-1}x \end{aligned}$$

Relativna greška po normi tada iznosi

$$\frac{\|\tilde{x}-x\|_2}{\|x\|_2} \leq \frac{\|A^{-1}\|_F\|\Delta A\|_F}{1-\|A^{-1}\|_F\|\Delta A\|_F} \leq \frac{O(n^3)\varepsilon\rho\kappa_F}{1-O(n^3)\varepsilon\rho\kappa_F(A)},$$

uz uvjet $O(n^3)\varepsilon\rho\kappa_F(A) < 1$, pri čemu je $\kappa_F(A) = \|A^{-1}\|_F\|A\|_F$ **broj uvjetovanosti matrice A**. Ovime je dana ocjena za **grešku unaprijed**.

Zadatak 1.4 *Za sustave iz zadatka 1.2 u* Octave*-i izračunajte* ρ, $\kappa_F(A)$ *i rezidual* $r = b - A\tilde{x}$ *te* $\|r\|_2$. *Zaključite zašto je došlo do velike greške u naprijed kod rješenja prvog sustava, i zašto je došlo do male greške kod rješenja drugog sustava.*

Zadatak 1.5 *Egzaktno i u* Octave*-i riješite sustav* $Ax = b$ *pomoću Gaussovih eliminacija bez pivotiranja, pri čemu su*

$$\begin{bmatrix} 1 & 2\alpha \\ 1 & \alpha \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1+2\alpha \\ 1+\alpha \end{bmatrix}$$

za $\alpha = eps/2$. *Izračunajte relativnu grešku unaprijed* $\|\tilde{x}-x\|_2/\|x\|_2$, ρ, $\kappa_F(A)$ *i* $\|r\|_2$, *te zaključite zašto je došlo do tako velike relativne greške.*

Napomena 1.5 *U egzaktnoj aritmetici imamo*

$$\left[\begin{array}{cc|c} 1 & 2\alpha & 1+2\alpha \\ 1 & \alpha & 1+\alpha \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 2\alpha & 1+2\alpha \\ 0 & -\alpha & -\alpha \end{array}\right] \sim$$
$$\sim \left[\begin{array}{cc|c} 1 & 2\alpha & 1+2\alpha \\ 0 & 1 & 1 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 1 \end{array}\right],$$

tj. točno rješenje je

$$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

U aritmetici konačne preciznosti je $fl(1+\alpha) = 1$, *pa imamo*

$$\left[\begin{array}{cc|c} 1 & 2\alpha & fl(1+2\alpha) \\ 1 & \alpha & fl(1+\alpha) \end{array}\right] = \left[\begin{array}{cc|c} 1 & 2\alpha & 1+2\alpha \\ 1 & \alpha & 1 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 2\alpha & 1+2\alpha \\ 0 & -\alpha & -2\alpha \end{array}\right] \sim$$
$$\sim \left[\begin{array}{cc|c} 1 & 2\alpha & 1+2\alpha \\ 0 & 1 & 2 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 0 & 1-2\alpha \\ 0 & 1 & 2 \end{array}\right],$$

tj. izračunato rješenje je

$$\hat{x} = \begin{bmatrix} 1-2\alpha \\ 2 \end{bmatrix} \neq \begin{bmatrix} 1 \\ 1 \end{bmatrix} = x,$$

što je vrlo netočno.

Provjerimo još na kraju uvjetovanost matrice i njen pivotni rast. Dobivamo da je

$$\kappa_F(A) \approx 1.8014 \cdot 10^{16},$$

što je jako veliko, a pivotni rast je

$$\rho = 1,$$

što je minimalno. Dakle možemo zaključiti da u ovom primjeru glavni uzrok velike relativne greške je loša uvjetovanost matrice sustava.

Zadatak 1.6 *Neka su zadani* $A \in \mathbb{R}^{4\times 4}$ *i* $b \in \mathbb{R}^4$

$$A = \begin{bmatrix} 10^{-10} & 2 & -3 & 300 \\ 2 & -2 & 100 & 10^5 \\ -111 & 1 & 0 & -1 \\ 2222 & 4 & -1 & -1 \end{bmatrix}, \qquad b = \begin{bmatrix} 299 \\ 100100 \\ -111 \\ 2224 \end{bmatrix}.$$

Egzaktno rješenje ovog sustava je

$$x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Riješite dani sustav metodom Gaussovih eliminacija bez i sa pivotiranjem, i provjerite valjanost ocjena greški.

Napomena 1.6 *Dobivena izračunata rješenja su* $\tilde{x}_1$ *za Gussove eliminacije bez pivotiranja, i* $\tilde{x}_2$ *za Gussove eliminacije sa pivotiranjem, pri čemu su*

$$\tilde{x}_1 = \begin{bmatrix} 0.999875737761613 \\ 1.000333614228450 \\ 1.000202791544997 \\ 0.999999803820594 \end{bmatrix}, \qquad \tilde{x}_2 = \begin{bmatrix} 1.000000000000010 \\ 0.999999999994659 \\ 0.999999999999958 \\ 1.000000000000000 \end{bmatrix}.$$

Relativne greške iznose

$$\frac{\|x - \tilde{x}_1\|_2}{\|x\|_2} \approx 2.04856 \cdot 10^{-4}, \quad \frac{\|x - \tilde{x}_2\|_2}{\|x\|_2} \approx 2.67056 \cdot 10^{-12},$$

Pogledajmo sada ocjene greške. Kao aproksimaciju pivotnog rasta ćemo uzeti veličinu

$$\rho \gtrsim \frac{\max\limits_{i,j} |\hat{u}_{ij}|}{\max\limits_{i,j} |a_{ij}|} = \rho_U,$$

i ona u našem primjeru iznosi

$$\rho_{U,1} = 5.9999999 \cdot 10^7, \quad \rho_{U,2} = 1.001203064.$$

Ono što nam je potrebno za ocjenu povratne greške i greške u naprijed su norma i broj uvjetovanosti matrice A.

$$\|A\|_F = 1.00025 \cdot 10^5, \qquad \kappa_F(A) = 1.04886 \cdot 10^5.$$

Dakle, ocjenu za povratnu grešku označit ćemo sa

$$oc_p \approx 64u\rho_U\|A\|_F,$$

a za grešku unaprijed sa

$$oc_u \approx \frac{64u\rho_U\kappa_F}{1 - 64u\rho_U\kappa_F}.$$

Za oba dvije metode Gaussovih eliminacija imamo

$$oc_{p,1} = 4.2643 \cdot 10^{-2}, \qquad oc_{u,1} = 4.6809 \cdot 10^{-2},$$

$$oc_{p,2} = 7.1158 \cdot 10^{-10}, \qquad oc_{u,2} = 7.4616 \cdot 10^{-10}.$$

Napomena 1.7 *Kod rješavanja sustava $Ax = b$, postupak Gaussovih eliminacija bez ili sa pivotiranjem je stabilniji ukoliko su elementi u matrici A istog reda veličine.*

Neka su D_1 i D_2 regularne dijagonalne matrice. Sustav $Ax = b$ ekvivalentan je sustavu $D_1Ax = D_1b$, a uz $D_2D_2^{-1} = I$ imamo:

$$\underbrace{(D_1AD_2)}_{\bar{A}}\underbrace{D_2^{-1}x}_{y} = \underbrace{D_1b}_{\bar{b}},$$

pri čemu je $y = D_2^{-1}x$ rješenje sustava $\bar{A}y = \bar{b}$. To je **skaliranje sustava** *$Ax = b$. Naime množenje dijagonalnom matricom slijeva matrice A znači množenje svakog retka od A odgovarajućim dijagonalnim elementom od D_1, a množenje s D_2 zdesna odgovara množenju stupca od A dijagonalnim elementom od D_2. Odgovarajućim izborom elemenata matrica D_1 i D_2 se element koji je puno manji ili veći od ostalih elemenata u matrici A može svesti na isti red veličine, tj. rješavanje skaliranog sustava je stabilnije. Povratna veza je tada vrlo jednostavna:*

$$x = D_2y.$$

Općenito kriterij za izbor matrica D_1 i D_2 je takav, da skalirana matrica $\bar{A} = D_1AD_2$ ima manji broj uvjetovanosti, i time predstavlja sustav čije rješavanje je stabilnije.

Primjer 1.9 *Neka je dan sustav $Ax = b$, takav da je*

$$A = \begin{bmatrix} 1 & 2 \cdot 10^2 & -1 \cdot 10^6 \\ 3 \cdot 10^5 & 2 \cdot 10^7 & 0 \\ -4 \cdot 10^{10} & 5 \cdot 10^{12} & 10^{16} \end{bmatrix}, \qquad b = \begin{bmatrix} -1.000000000001 \cdot 10^{12} \\ -3 \cdot 10^5 \\ 1.000000000004 \cdot 10^{22} \end{bmatrix}.$$

Egzaktno rješenje ovog sustava je

$$x = \begin{bmatrix} -1 \\ 0 \\ 10^6 \end{bmatrix}.$$

Riješimo li ovaj sustav u Octave*-i pomoću Gaussovih eleiminacija sa pivotiranjem redaka, u dvostrukoj preciznosti dobit ćemo rezultat*

$$\tilde{x} = \begin{bmatrix} -0.9999745024 \\ 0 \\ 10^6 \end{bmatrix}, \qquad \frac{\|x - \tilde{x}\|_2}{\|x\|_2} = 2.54976 \cdot 10^{-11}.$$

S druge strane vidimo da su elementi matrice A vrlo različitih redova veličine. Definirajmo zato

$$D_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 10^{-5} & 0 \\ 0 & 0 & 10^{-10} \end{bmatrix}, \qquad D_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 10^{-2} & 0 \\ 0 & 0 & 10^{-6} \end{bmatrix}.$$

Skalirani sustav $\bar{A}y = \bar{b}$ sada izgleda ovako

$$\bar{A} = D_1 A D_2 = \begin{bmatrix} 1 & 2 & -1 \\ 3 & 2 & 0 \\ -4 & 5 & 1 \end{bmatrix}, \qquad \bar{b} = D_1 b = \begin{bmatrix} -1.000000000001 \cdot 10^{12} \\ -3 \\ 1.000000000004 \cdot 10^{12} \end{bmatrix}.$$

Egzaktno rješenje ovog sustava je

$$y = \begin{bmatrix} -1 \\ 0 \\ 10^{12} \end{bmatrix}.$$

Riješimo li i ovaj sustav u Octave*-i na isti način dobit ćemo*

$$\tilde{y} = \begin{bmatrix} -9.999960194463315 \cdot 10^{-1} \\ -2.122961956521739 \cdot 10^{-5} \\ 1.000000000000000 \cdot 10^{12} \end{bmatrix}, \quad \tilde{x}_y = D_2 \tilde{y} = \begin{bmatrix} -9.999960194463315 \cdot 10^{-1} \\ -2.122961956521739 \cdot 10^{-7} \\ 1.000000000000000 \cdot 10^{6} \end{bmatrix},$$

$$\frac{\|x - \tilde{x}_y\|_2}{\|x\|_2} = 3.98621 \cdot 10^{-12}.$$

Dakle, možemo zaključiti da je računanje skaliranog sustava stabilnije od početnog sustava, čija matrica ima vrlo različite elemente po veličini. Također kada provjerimo brojeve uvjetovanosti matrica A i $\bar{A}$:

$$\kappa_F(A) \approx 7.4 \cdot 10^{14}, \qquad \kappa_F(\bar{A}) \approx 8.2$$

vidimo da je skalirana matrica puno bolje uvjetovana od početne, pa je rezultat u skladu sa teorijom perturbacija kod rješavanja linearnih sustava.

1.1.6 Pozitivno definitni sustavi i faktorizacija Choleskog

Kažemo da je simetrična $n \times n$ matrica A **pozitivno definitna** ako za sve $x \in \mathbb{R}^n$, $x \neq 0$, vrijedi

$$x^\tau A x > 0.$$

Primjer 1.10 *Na primjeru sa predavanja bilo je pokazano kako se problem aproksimativnog rješavanja rubnog problema može svesti na rješavanje linearnog sustava. Rubni problem*

$$\begin{aligned} -\frac{d^2}{dx^2}u(x) &= f(x), \quad 0 < x < 1, \\ u(0) &= u(1) = 0. \end{aligned}$$

je diskretizacijom na segmentu $[0,1]$, sa mrežom od $n+2$ točaka

$$h = \frac{1}{n+1}, \quad x_i = ih, \quad i = 0, \ldots, n+1. \tag{12}$$

sveden na linearni sustav

$$\underbrace{\begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{bmatrix}}_{T_n} \underbrace{\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{n-2} \\ u_{n-1} \\ u_n \end{bmatrix}}_{u} = h^2 \underbrace{\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{n-2} \\ f_{n-1} \\ f_n \end{bmatrix}}_{f}. \tag{13}$$

Tvrdimo da je matrica sustava T_n simetrična pozitivno definitna matrica. Da je matrica simetrična to se odmah vidi, ali da li je i pozitivno definitna, to moramo provjeriti.

Jedan od kriterija pozitivne definitnosti simetrične matrice A je sljedeći:

- *Simetrična matrica A je pozitivno definitna ako i samo ako su sve njene vodeće minore strogo pozitivne.*

Provjerimo da li je to istina za matricu T_n. Dokaz se provodi matematičkom indukcijom. Označimo sa m_k vodeću minoru reda k matrice T_n. Tada imamo:

$$m_1 = T_n(1,1) = 2 > 0,$$

$$m_2 = \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} = 3 > 0.$$

Pretpostavimo da je $m_i = i + 1 > 0$ za $i = 1, 2, \ldots, k < n$ i provjerimo pozitivnost minore m_{k+1}. Razvijanjem po prvom retku dobivamo sljedeću jednakost

$$m_{k+1} = \underbrace{\begin{vmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{vmatrix}}_{(k+1)\times(k+1)} =$$

$$= 2 \cdot \underbrace{\begin{vmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & \cdots & \cdots & \cdots & & & \\ & & -1 & 2 & -1 & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & \end{vmatrix}}_{k\times k} + 1 \cdot \underbrace{\begin{vmatrix} -1 & -1 & & & & & \\ & 2 & -1 & & & & \\ & & \cdots & \cdots & & & \\ & & -1 & 2 & -1 & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & \end{vmatrix}}_{k\times k}$$

$$= 2m_k - m_{k-1} = 2(k+1) - k = k + 2 > 0.$$

Dakle matrica T_n je zaista pozitivno definitna.

Iz analize slične onoj za LU faktorizaciju, možemo zaključiti sljedeće.

- Neka je $A \in \mathbb{R}^{n\times n}$ pozitivno definitna matrica. Tada postoji jedinstvena gornjetrokutasta matrica $R \in \mathbb{R}^{n\times n}$ s pozitivnom dijagonalom, takva da je $A = R^T R$.

- Gore opisana trokutasta faktorizacija (faktorizacija Choleskog) je provediva ako i samo ako je matrica A simetrična i pozitivno definitna, pa je ona kriterij za pozitivnu definitnost simetrične realne matrice.

Faktorizaciju $A = R^\tau R$ zovemo **faktorizacija Choleskog** ili **trokutasta faktorizacija** simetrične pozitivno definitne matrice, i možemo je izračunati sljedećim algoritmom.

Algoritam 1.6 *Računanje faktorizacije Choleskog simetrične pozitivno definitne matrice* $A \in \mathbb{R}^{n\times n}$.

za $i = 1, \ldots, n$ {

$r_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2}$;

/* za $i = 1$, $r_{ii} = \sqrt{a_{ii}}$ */

za $j = i+1, \ldots, n$ {

$r_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}\right) \Big/ r_{ii}$; } }

Primjer 1.11 *Pomoću faktorizacije Choleskog provjerit ćemo sada da je matrica T_n iz primjera 1.10 zaista pozitivno definitna matrica.*

Prema algoritmu 1.17, lako se vidi da je $T_n = R_n^T R_n$, pri čemu je

$$R = \begin{bmatrix} \sqrt{2} & -\frac{1}{\sqrt{2}} & & & & & \\ & \frac{\sqrt{3}}{\sqrt{2}} & -\frac{\sqrt{2}}{\sqrt{3}} & & & & \\ & & \frac{2}{\sqrt{3}} & -\frac{\sqrt{3}}{2} & & & \\ & & & \frac{\sqrt{5}}{2} & -\frac{2}{\sqrt{5}} & & \\ & & & & \cdots & \cdots & \\ & & & & & \frac{\sqrt{n}}{\sqrt{n-1}} & -\frac{\sqrt{n-1}}{\sqrt{n}} \\ & & & & & & \frac{\sqrt{n+1}}{\sqrt{n}} \end{bmatrix},$$

dakle, faktorizacija Choleskog matrice T_n je provediva, pa je ona zaista pozitivno definitna matrica.

Prođimo sada korak po korak algoritma 1.17 za računanje faktorizacije Choleskog matrice T_5

$$T_5 = \begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

U prvom koraku izračunamo korijen od elementa na $(1,1)$ *poziciji, i njime podijelimo sve elemente prvog retka, dakle*

$$r_{1,1} = \sqrt{2}, \quad r_{12} = -\frac{1}{\sqrt{2}}, \quad r_{13} = r_{14} = r_{15} = 0$$

i

$$T_5^{(1)} = \begin{bmatrix} \sqrt{2} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

U drugom koraku računamo

$$r_{22} = \sqrt{2 - \left(-\frac{1}{\sqrt{2}}\right)^2} = \sqrt{2 - \frac{1}{2}} = \frac{\sqrt{3}}{\sqrt{2}},$$

a budući da se na pozicijama $(1,3)$*,* $(1,4)$ *i* $(1,5)$ *u matrici* $T_5^{(1)}$ *nalaze nule, imamo da je*

$$r_{23} = \frac{-1}{\frac{\sqrt{3}}{\sqrt{2}}} = -\frac{\sqrt{2}}{\sqrt{3}}, \qquad r_{24} = r_{25} = 0,$$

pa je

$$T_5^{(2)} = \begin{bmatrix} \sqrt{2} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ 0 & \frac{\sqrt{3}}{\sqrt{2}} & -\frac{\sqrt{2}}{\sqrt{3}} & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

U trećem koraku je

$$r_{33} = \sqrt{2 - \left(-\frac{\sqrt{2}}{\sqrt{3}}\right)^2} = \sqrt{2 - \frac{2}{3}} = \frac{2}{\sqrt{3}},$$

a budući da se na pozicijama $(1,4)$*,* $(2,4)$*,* $(1,5)$ *i* $(2,5)$ *u matrici* $T_5^{(2)}$ *nalaze nule, imamo da je*

$$r_{34} = \frac{-1}{\frac{2}{\sqrt{3}}} = -\frac{\sqrt{3}}{2}, \qquad r_{35} = 0,$$

odakle je

$$T_5^{(3)} = \begin{bmatrix} \sqrt{2} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ 0 & \frac{\sqrt{3}}{\sqrt{2}} & -\frac{\sqrt{2}}{\sqrt{3}} & 0 & 0 \\ 0 & 0 & \frac{2}{\sqrt{3}} & -\frac{\sqrt{3}}{2} & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

U četvrtom koraku je

$$r_{44} = \sqrt{2 - \left(-\frac{\sqrt{3}}{2}\right)^2} = \sqrt{2 - \frac{3}{4}} = \frac{\sqrt{5}}{2},$$

a budući da se na pozicijama $(1,5)$, $(2,5)$ *i* $(3,5)$ *u matrici* $T_5^{(3)}$ *nalaze nule, imamo da je*

$$r_{4,5} = \frac{-1}{\frac{\sqrt{5}}{2}} = -\frac{2}{\sqrt{5}},$$

pa je

$$T_5^{(4)} = \begin{bmatrix} \sqrt{2} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ 0 & \frac{\sqrt{3}}{\sqrt{2}} & -\frac{\sqrt{2}}{\sqrt{3}} & 0 & 0 \\ 0 & 0 & \frac{2}{\sqrt{3}} & -\frac{\sqrt{3}}{2} & 0 \\ 0 & 0 & 0 & \frac{\sqrt{5}}{2} & -\frac{2}{\sqrt{5}} \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

I napokon u zadnjem, petom koraku, moramo izračunati samo dijagonalni element

$$r_{55} = \sqrt{2 - \left(-\frac{2}{\sqrt{5}}\right)^2} = \sqrt{2 - \frac{4}{5}} = \frac{\sqrt{6}}{\sqrt{5}}.$$

Konači oblik faktora je

$$R_5 = T_5^{(5)} = \begin{bmatrix} \sqrt{2} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ 0 & \frac{\sqrt{3}}{\sqrt{2}} & -\frac{\sqrt{2}}{\sqrt{3}} & 0 & 0 \\ 0 & 0 & \frac{2}{\sqrt{3}} & -\frac{\sqrt{3}}{2} & 0 \\ 0 & 0 & 0 & \frac{\sqrt{5}}{2} & -\frac{2}{\sqrt{5}} \\ 0 & 0 & 0 & 0 & \frac{\sqrt{6}}{\sqrt{5}} \end{bmatrix}.$$

Zadatak 1.7 *U* Octave*-i izračunajte gornjetrokutasti faktor Choleskog matrice* T_5.

Primjer 1.12 *Vidjeli smo da je provedivost faktorizacije Choleskog nužan i dovoljan uvjet da je simetrična matrica pozitivno definitna, u* **egzaktnoj aritmetici**. *Da li isto vrijedi i u aritmetici konače preciznosti? Odgovor je: uglavnom DA. To znači da u velikoj većini slučajeva, ako je faktorizacija Choleskog provediva u aritmetici konačne preciznosti, matrica je zaista pozitivno definitna, a ako nije provediva, matrica nije pozitivno definitna. Međutim, u rijetkim slučajevima se može dogoditi da je matrica pozitivno definitna, ali vrlo loše uvjetovana, a da zbog grešaka zaokruživanja računanje faktora Choleskog bude prekinuto prije kraja. To se događa kod računanja dijagonalnog elementa gornjetrokutastog faktora, kada se broj koji je pod korijenom zaokruži na nulu ili na negativni broj. Takav primjer je sljedeća* 2×2 *matrica: (izvedite na* **student***-u)*

$$A = \begin{bmatrix} 25 & 1 \\ 1 & 0.04000000000000001 \end{bmatrix}.$$

U egzaktnoj aritmetici dobivamo sljedeće

$$\begin{aligned} r_{11} &= \sqrt{25} = 5 \\ r_{12} &= \frac{1}{r_{11}} = \frac{1}{5} = 0.2 \\ r_{22} &= \sqrt{(0.04 + 10^{-17}) - 0.2^2} = \sqrt{(0.04 + 10^{-17}) - 0.04} = \sqrt{10^{-17}} > 0, \end{aligned}$$

što znači da je u egzaktnoj aritmetici faktorizacija Choleskog matrice A provediva, i ona glasi

$$A = G^T G, \qquad G = \begin{bmatrix} 5 & 0.2 \\ 0 & \sqrt{10^{-17}} \end{bmatrix}.$$

S druge strane ako ovaj primjer želimo izračunati u Octave*-i, dobit ćemo sljedeću poruku:*
??? Error using ==> chol
Matrix must be positive definite.
Što se tu zapravo dogodilo? Kod računanja izraza

$$fl\left(0.04000000000000001 - \left(\frac{1}{5}\right)^2\right) = 0$$

broj 10^{-17} *je zaokružen na 0.*

Promotrimo sada linearni sustav jednadžbi u kojem je $A \in \mathbb{R}^{n\times n}$ simetrična, pozitivno definitna matrica. Ako je $A = R^\tau R$ trokutasta faktorizacija, onda rješenje

$$x = A^{-1}b = R^{-1}R^{-\tau}b$$

možemo dobiti tako da prvo nađemo rješenje y sustava $R^\tau y = b$, a zatim riješimo sustav $Rx = y$. Kako je R gornjetrokutasta matrica, cijeli postupak je vrlo jednostavan i možemo ga zapisati na sljedeći način:

Algoritam 1.7 *Rješavanje linearnog sustava jednadžbi $Ax = b$ s pozitivno definitnom matricom $A \in \mathbb{R}^{n\times n}$.*

/* Trokutasta faktorizacija $A = R^\tau R$ */
za $i = 1, \ldots, n$ {
$$r_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2};$$
/* za $i = 1$, $r_{ii} = \sqrt{a_{ii}}$ */
za $j = i+1, \ldots, n$ {
$$r_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj} \right) \Big/ r_{ii}; \} \}$$
/* Supstitucije naprijed za $R^\tau y = b$ */
$$y_1 = \frac{b_1}{r_{11}};$$
za $i = 2, \ldots, n$ {
$$y_i = \left(b_i - \sum_{j=1}^{i-1} r_{ji} y_j \right) \Big/ r_{ii}; \}$$
/* Supstitucije unazad za $Rx = y$ */
$$x_n = \frac{y_n}{r_{nn}};$$
za $i = n-1, \ldots, 1$ {
$$x_i = \left(y_i - \sum_{j=i+1}^{n} r_{ij} x_j \right) \Big/ r_{ii}; \}$$

1.1.7 Numerička svojstva faktorizacije Choleskog

Iz teorema sa predavanja, kao i kod LU faktorizacije možemo zaključiti

- Neka je za zadanu $n\times n$ pozitivno definitnu matricu A algoritam 1.17 uspješno izvršio sve operacije u aritmetici konačne preciznosti s greškom zaokruživanja ε. Ako je $\widetilde{R}$ izračunata aproksimacija matrice R, onda je
$$\widetilde{R}^\tau \widetilde{R} = A + \delta A,$$

gdje je $\delta A = [\delta a_{ij}]$ simetrična matrica i za sve $1 \leq i, j \leq n$ vrijedi

$$|\delta a_{ij}| \leq O(n)\varepsilon\sqrt{a_{ii}a_{jj}}.$$

- Neka je $n \times n$ pozitivno definitna matrica A i neka je $\tilde{x}$ izračunata aproksimacija točnog rješenja $x = A^{-1}b$ u aritmetici konačne preciznosti, dobivena pomoću faktorizacije Choleskog. Tada postoji simetrična perturbacija ΔA takva da je

  $$(A + \Delta A)\tilde{x} = b.$$

  Pri tome je

  $$|\delta a_{ij}| \leq O(n^2)\varepsilon\sqrt{a_{ii}a_{jj}}$$

  i

  $$\|\Delta A\|_F \leq O(n^{\frac{5}{2}})\varepsilon\|A\|_F.$$

- Relativna greška unaprijed je dana sa

  $$\frac{\|x - \tilde{x}\|_2}{\|x\|_2} \leq \frac{O(n^{5/2})\varepsilon\kappa_F(A)}{1 - O(n^{5/2})\varepsilon\kappa_F(A)},$$

  uz uvjet $O(n^{5/2})\varepsilon\kappa_F(A) < 1$.

Zadatak 1.8 *Neka su zadani simetrična pozitivno definitna matrica* $A \in \mathbb{R}^{4\times4}$ *i vektor* $b \in \mathbb{R}^4$

$$A = \begin{bmatrix} 10^8 & 0 & 2\cdot10^4 & -3\cdot10^4 \\ 0 & 484 & -11 & 22 \\ 2\cdot10^4 & -11 & 4.2501 & -6.44 \\ -3\cdot10^4 & 22 & -6.44 & 47 \end{bmatrix},$$

$$b = \begin{bmatrix} 9.999\cdot10^7 \\ 495 \\ 19986.8101 \\ -29937.44 \end{bmatrix}.$$

Egzaktno rješenje ovog sustava je

$$x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

U Octave-*i riješite dani sustav pomoću faktorizacije Choleskog, i provjerite valjanost ocjene greški.*

Napomena 1.8 *Dobiveno rješenje je*

$$\tilde{x} = \begin{bmatrix} 1.000000000090355 \\ 0.9999999897241033 \\ 0.9999995493222721 \\ 1.000000000730864 \end{bmatrix}.$$

Relativna greška iznosi

$$\frac{\|x - \tilde{x}\|_2}{\|x\|_2} \approx 2.253977322520550 \cdot 10^{-7}.$$

Pogledajmo sada ocjene greške. Ono što nam je potrebno za ocjenu greške u naprijed je broj uvjetovanosti matrice A.

$$\kappa_F(A) \approx 3.701934 \cdot 10^{13}$$

Ocjena za grešku unaprijed aproksimativno iznosi

$$oc_u \approx \frac{32u\kappa_F(A)}{1 - 32u\kappa_F(A)} = 1.514358177218385 \cdot 10^{-1}.$$

Za grešku unazad dovoljno je provjeriti da za izračunati faktor Choleskog $\widetilde{R}$ matrice A

$$\|A - \widetilde{R}^T\widetilde{R}\|_F \lesssim 4^{\frac{3}{2}}\varepsilon\|A\|_F.$$

U ovom slučaju potrebne vrijednosti iznose

$$\|A - \widetilde{R}^T\widetilde{R}\|_F = 0$$

$$\|A\|_F = 1.0000001 \cdot 10^8$$

$$oc_p \approx 8\varepsilon\|A\|_F = 8.881785351738714 \cdot 10^{-8}.$$

1.2 Iterativne metode

1.2.1 Opis problema

Problemi koji se javljaju kod rješavanja linearnih sustava Gaussovim eliminacijama su:

- Elemente matrice sustava velikih dimenzija je problematično spremiti u memoriju, zbog ograničemosti radne memorije.
- Vrijeme izvršavanja Gaussovih eliminacija nad matricama velikih dimenzija je neprihvatljivo dugo.

- U primjeni se često pojavljuju matrice velikih dimenzija, ali koje su strukturirane i rijetko popunjene (puno nula). U tim slučajevim nepotrebno je spremati elemente koji su jednaki nula, pa se matrica sprema u posebnom formatu. Problem je što Gaussove eliminacije mogu upropastiti tu specijalnu strukturu, i rezultat se ne može spremiti u istom formatu.

Prethodna diskusija nas motivira da potražimo i drugačije pristupe za rješavanje linearnog sustava $Ax = b$. Primijetimo da ne moramo nužno težiti pronalaženju egzaktnog rješenja – umjesto toga želimo **dovoljno dobru** aproksimaciju $\tilde{x}$. Zato ima smisla pokušati konstruirati niz $x^{(0)}, x^{(1)}, \ldots, x^{(k)}, \ldots$ vektora iz $\mathbb{R}^n$ sa sljedećim svojstvima:

(i) za svaki k formula za računanje $x^{(k)}$ je jednostavna;

(ii) $x^{(k)}$ teži prema $x = A^{-1}b$ i za neki k (obično $k \ll n$) je $x^{(k)}$ prihvatljiva aproksimacija za x.

Rješavamo, dakle, sustav $Ax = b$, $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$. Iterativnu metodu pokušavamo naći u obliku

$$x^{(k+1)} = Tx^{(k)} + c, \qquad k = 0, 1, 2, \ldots, \quad x^{(0)} \text{ zadan},$$

gdje je $T \in \mathbb{R}^{n\times n}$ **iterativna matrica** i $c \in \mathbb{R}^n$.

Jedan način odabira iterativne matrice T je taj da matricu sustava A rastavimo na

$$A = M - N, \qquad M \text{ regularna}.$$

Tada se polazni linearni sustav transformira u

$$\begin{aligned} (M - N)x &= b \\ Mx &= Nx + b \\ x &= M^{-1}Nx + M^{-1}b \\ x &= Tx + c \end{aligned}$$

gdje je

$$T = M^{-1}N, \qquad c = M^{-1}b,$$

pri čemu je onda rješenje sustava fiksna točka iteracija

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b, \qquad k = 0, 1, 2, \ldots.$$

O konvergenciji iteracija možemo reći sljedeće.

- Neka je $b \in \mathbb{R}^n$ i $A = M - N \in \mathbb{R}^{n\times n}$ regularna matrica. Ako je M regularna matrica i $\rho(M^{-1}N) < 1$, tada niz iteracija $\{x^{(k)}, k \geq 0\}$ definiran sa $x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b$, $k = 0, 1, 2, \ldots$ konvergira prema $x = A^{-1}b$ za proizvoljnu početnu iteraciju $x^{(0)}$. Tvrdnja teorema vrijedi i ako je $\|M^{-1}N\| < 1$ za bilo koju konzistentnu matričnu normu $\| \|$.

Napomena 1.9 *Dokaz gornje tvrdnje je vrlo jednostavan. Definirajmo grešku u svakom koraku kao*

$$e^{(k)} = x^{(k)} - x, \qquad k = 0, 1, 2, \ldots$$

Tada za $x = A^{-1}b$ vrijedi

$$\begin{aligned} x^{(k+1)} &= M^{-1}Nx^{(k)} + M^{-1}b \\ x &= M^{-1}Nx + M^{-1}b \\ x^{(k+1)} - x &= M^{-1}N(x^{(k)} - x) \\ e^{(k+1)} &= M^{-1}Ne^{(k)} = (M^{-1}N)^2e^{(k-1)} = \cdots = (M^{-1}N)^{k+1}e^{(0)}. \end{aligned}$$

Prema jednom teoremu o matričnim normama, za svako $\varepsilon > 0$ (mi ćemo uzeti $\varepsilon < 1 - \rho(M^{-1}N)$) postoji konzistentna matrična norma $\|\cdot\|$ na $\mathbb{R}^{n\times n}$ takva da vrijedi:

$$\|M^{-1}N\| \leq \rho(M^{-1}N) + \varepsilon < 1,$$

odakle je

$$\|e^{(k+1)}\| \leq \|M^{-1}N\|^{k+1}\|e^{(0)}\| \longrightarrow 0, \ \textit{kad } k \to \infty.$$

Znači:

$$\begin{aligned} \lim_{k\to\infty} e^{(k)} &= 0 \\ \lim_{k\to\infty} x^{(k)} &= x. \end{aligned}$$

Zadatak 1.9 *Neka vrijede gornje pretpostavke za $T = M^{-1}N$ i pretpostavimo da tražimo aproksimaciju rješenja takvu da vrijedi*

$$\|x^{(k)} - x\| < \varepsilon, \tag{14}$$

gdje je $\|\cdot\|$ neka od normi $\|\cdot\|_1$, $\|\cdot\|_2$ ili $\|\cdot\|_\infty$. Nađite kriterij zaustavljanja, tj. broj iteracije k za koju će vrijediti (14).

Napomena 1.10 *Za $k, p \in \mathbb{N}_0$ promatrimo sljedeće:*

$$\begin{aligned}
&\|x^{(k+p)} - x^{(k)}\| = \|x^{(k+p)} - x^{(k+p-1)} + x^{(k+p-1)} - x^{(k+p-2)} + \cdots + x^{(k+1)} - x^{(k)}\| \leq \\
&\leq \|x^{(k+p)} - x^{(k+p-1)}\| + \|x^{(k+p-1)} - x^{(k+p-2)}\| + \cdots + \|x^{(k+1)} - x^{(k)}\| = \\
&= \|T^{p-1}(x^{(k+1)} - x^{(k)})\| + \|T^{p-2}(x^{(k+1)} - x^{(k)})\| + \cdots + \|(x^{(k+1)} - x^{(k)})\| \leq \\
&\leq \left(\|T\|^{p-1} + \|T\|^{p-2} + \cdots + 1\right) \|x^{(k+1)} - x^{(k)}\| \leq \\
&\leq \left(\|T\|^{p-1} + \|T\|^{p-2} + \cdots + 1\right) \|T\|^k \|x^{(1)} - x^{(0)}\| \leq \\
&\leq \frac{\|T\|^k}{1 - \|T\|} \|x^{(1)} - x^{(0)}\|.
\end{aligned}$$

Kad pustimo da $p \to \infty$ dobivamo

$$\|x - x^{(k)}\| \leq \frac{\|T\|^k}{1 - \|T\|} \|x^{(1)} - x^{(0)}\|,$$

pa je dovoljno tražiti da je

$$\frac{\|T\|^k}{1 - \|T\|} \|x^{(1)} - x^{(0)}\| < \varepsilon,$$

odnosno

$$k > \frac{\ln\left(\frac{\varepsilon(1-\|T\|)}{\|x^{(1)} - x^{(0)}\|}\right)}{\ln(\|T\|)}.$$

Zadatak 1.10 *Zadan je sustav, gdje je*

$$A = \begin{bmatrix} 5 & 0 & 0 & 0.1 & 0.1 \\ 5 & 5 & 0 & 0 & 0.1 \\ 0.1 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0.1 \\ 0.1 & 0.1 & 5 & 0 & 5 \end{bmatrix}, \qquad b = \begin{bmatrix} -9.7 \\ -14.8 \\ -0.2 \\ 5.2 \\ 9.7 \end{bmatrix}.$$

Odredite rastav matrice A kao $A = M - N$ i odgovarajućom iterativnom metodom riješite sustav, tako da greška u svakoj nepoznanici bude manja od 10^{-3}.

Napomena 1.11 *Rastavimo matricu $A = M - N$ na sljedeći način:*

$$M = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 \\ 5 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 5 & 0 & 5 \end{bmatrix}, \qquad N = \begin{bmatrix} 0 & 0 & 0 & -0.1 & -0.1 \\ 0 & 0 & 0 & 0 & -0.1 \\ -0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.1 \\ -0.1 & -0.1 & 0 & 0 & 0 \end{bmatrix}.$$

Lako se može provjeriti da je

$$M^{-1} = \begin{bmatrix} 0.2 & 0 & 0 & 0 & 0 \\ -0.2 & 0.2 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & -0.2 & 0 & 0.2 \end{bmatrix}$$

i

$$T = M^{-1}N = \begin{bmatrix} 0 & 0 & 0 & 0.02 & 0.02 \\ 0 & 0 & 0 & -0.02 & 0 \\ 0.02 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.02 \\ 0 & 0.02 & 0 & 0 & 0 \end{bmatrix}, \quad c = M^{-1}b = \begin{bmatrix} -1.94 \\ -1.02 \\ -0.04 \\ 1.04 \\ 1.98 \end{bmatrix}.$$

Za iteracije

$$x^{(k+1)} = Tx^{(k)} + c, \qquad k = 0, 1, 2, \ldots$$

najprije trebamo provjeriti da li konvergiraju.

$$\rho(T) \leq \|T\|_\infty = 0.04 < 1,$$

dakle iteracije će konvergirati. Prema prethodnom zadatku pronaći ćemo broj iteracija koji je potreban za postizanje aproksimacije rješenja, čija greška ima $\|\cdot\|_\infty$ *normu manju od* $\varepsilon = 10^{-3}$*. Uzet ćemo da je* $x^{(0)} = 0$ *i onda je* $x^{(1)} = c$*.*

$$\begin{aligned} \frac{\|T\|_\infty^k}{1 - \|T\|_\infty}\|c\|_\infty &< 10^{-3} \\ \frac{0.04^k}{0.96} \cdot 1.98 &< 10^{-3} \\ 0.04^k &< 4.85 \cdot 10^{-4} \\ -3.2189k &< -7.6317 \\ k &> 2.3709, \end{aligned}$$

dakle mora biti

$$k = 3.$$

Kad se izračuna dobivamo:

$$x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \; x^{(1)} = \begin{bmatrix} -1.94 \\ -1.02 \\ -0.04 \\ 1.04 \\ 1.98 \end{bmatrix}, x^{(2)} = \begin{bmatrix} -2.0004 \\ -0.9992 \\ -0.0012 \\ 1.0004 \\ 2.0004 \end{bmatrix}, x^{(3)} = \begin{bmatrix} -2.000016 \\ -0.999992 \\ 0.000008 \\ 0.999992 \\ 1.999984 \end{bmatrix},$$

a egzaktno rješenje je

$$x = \begin{bmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{bmatrix},$$

pa vidimo da je naša ocjena točna, jer je greška

$$e^{(3)} = x - x^{(3)} \approx \begin{bmatrix} 1.6 \cdot 10^{-5} \\ -6 \cdot 10^{-6} \\ -6 \cdot 10^{-6} \\ 6 \cdot 10^{-6} \\ 1.6 \cdot 10^{-5} \end{bmatrix}.$$

1.2.2 Jacobijeva metoda

Matricu $A \in \mathbb{R}^{n \times n}$ rastavimo kao

$$A = L + D + R, \tag{15}$$

tako da su

$$\begin{array}{rcl} L & = & \text{donji trokut od } A \\ D & = & \text{dijagonala od } A \\ R & = & \text{gornji trokut od } A \end{array}$$

uz pretpostavku da A nema nula na dijagonali.

Jacobijeva metoda je iterativna metoda oblika

$$x^{(k+1)} = M_J^{-1} N_J x^{(k)} + M_J^{-1} b, \quad k = 0, 1, 2, \ldots$$

pri čemu su

$$M_J = D, \qquad N_J = -(L + R),$$

odnosno, radi se o iteracijama oblika

$$x^{(k+1)} = T_J x^{(k)} + c_J, \quad k = 0, 1, 2, \ldots$$

za koje su

$$T_J = -D^{-1}(L + R), \qquad c_J = D^{-1} b.$$

Algoritam 1.8 *Rješavanje linearnog sustava pomoću Jacobijeve metode.*

x_0 *fiksiran;*
for k=0,1,2,...
begin
for $i = 1, \ldots, n$
begin

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right);$$

end*;*
end

Uvjete za koje Jacobijeva metoda konvergira, daje sljedeća tvrdnja.

- Ako je matrica sustava $A \in \mathbb{R}^{n \times n}$ strogo dijagonalno dominantna matrica, tj. ako vrijedi

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| < |a_{ii}|, \quad i = 1, \ldots, n,$$

tada Jacobijeva metoda konvergira za svaku početnu iteraciju.

Primjer 1.13 *Neka je*

$$A = \begin{bmatrix} 2 & 0.1 \\ -0.1 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 19.9 \\ -3 \end{bmatrix}, \quad x = A^{-1}b = \begin{bmatrix} 10 \\ -1 \end{bmatrix}.$$

Matricu A napišimo kao u relaciji (15). Za početnu iteraciju uzmimo vektor

$$x^{(0)} = D^{-1}b = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 19.9 \\ -3 \end{bmatrix} = \begin{bmatrix} 9.949999999999999 \\ -1.5 \end{bmatrix}.$$

Primijetimo da u početnoj iteraciji pokušavamo "pogoditi" rješenje. Ponekad je dovoljno uzeti slučajno odabran vektor. Ipak, poželjno je da je polazna iteracija što je moguće bliže cilju. Naš izbor je bio rezultat jednostavne ideje: matricu A aproksimiramo s D (jer su dijagonalni elementi veći od izvandijagonalnih), pa $A^{-1}b$ aproksimiramo s $D^{-1}b$. Naravno da je ovo gruba aproksimacija, ali ipak ima smisla. Sada iteriramo i dobijemo

$$\begin{aligned} x^{(1)} &= \begin{bmatrix} 1.002500000000000e{+}001 \\ -1.002500000000000e{+}000 \end{bmatrix}, \\ x^{(2)} &= \begin{bmatrix} 1.000012500000000e{+}001 \\ -9.987500000000000e{-}001 \end{bmatrix}, \end{aligned}$$

$$
\begin{aligned}
x^{(3)} &= \begin{bmatrix} 9.999937500000000e{+}000 \\ -9.999937500000000e{-}001 \end{bmatrix}, \\
x^{(4)} &= \begin{bmatrix} 9.999999687499999e{+}000 \\ -1.000003125000000e{+}000 \end{bmatrix}, \\
x^{(5)} &= \begin{bmatrix} 1.000000015625000e{+}001 \\ -1.000000015625000e{+}000 \end{bmatrix}.
\end{aligned}
$$

Ako izračunamo relativne greške $\epsilon_k = \|x - x^{(k)}\|_\infty / \|x\|_\infty$, *onda je*

$$
\begin{aligned}
\epsilon_0 &= 5.000000000000000e{-}001 \\
\epsilon_1 &= 2.499999999999858e{-}002 \\
\epsilon_2 &= 1.249999999999973e{-}003 \\
\epsilon_3 &= 6.250000000029843e{-}005 \\
\epsilon_4 &= 3.125000000103739e{-}006 \\
\epsilon_5 &= 1.562499996055067e{-}007.
\end{aligned}
$$

Zadatak 1.11 *Jacobijevom metodom riješite sustav* $Ax = b$ *u* Octave-*i, ako je*

$$
A = \begin{bmatrix} 10 & 1 & 0 & 1 \\ 1 & 10 & 1 & 0 \\ 0 & 1 & 10 & 1 \\ 1 & 0 & 1 & 10 \end{bmatrix}, \qquad b = \begin{bmatrix} -8 \\ 0 \\ 12 \\ 20 \end{bmatrix},
$$

tako da greška u svakoj koordinati ne prelazi 10^{-3}.

Napomena 1.12 *Vidimo da je*

$$
D = 10I, \quad \Longrightarrow D^{-1} = \frac{1}{10} I,
$$

odakle slijedi

$$
T_J = -D^{-1}(L + R) = -\frac{1}{10} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix},
$$

$$
c_J = D^{-1} b = \begin{bmatrix} -0.8 \\ 0 \\ 1.2 \\ 2 \end{bmatrix}.
$$

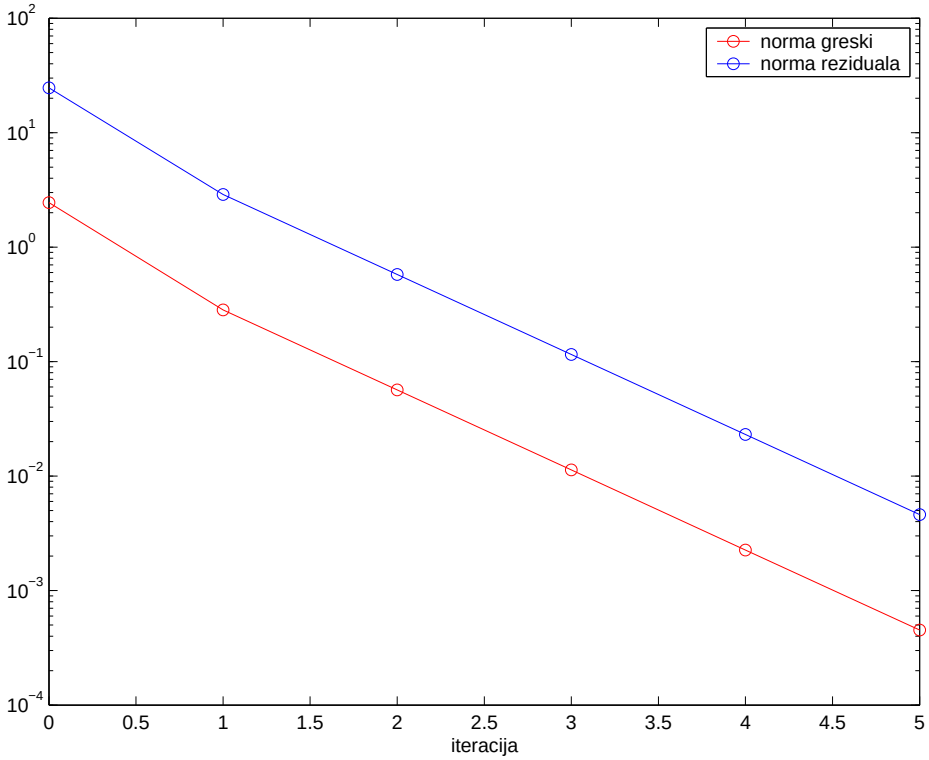
Prvo provjeravamo da li iterativna metoda uopće konvergira:

$$
\rho(T_J) \leq \|T_J\|_\infty = \frac{2}{10} = 0.2 < 1,
$$

dakle, metoda konvergira. Dalje, određujemo koliko koraka moramo izvršiti da bi postigli željenu točnost, za $x_0 = 0$.

$$\begin{aligned} \|x^{(k)} - x\|_\infty &\leq \frac{\|T_J\|_\infty^k}{1 - \|T_J\|_\infty} \|c_J\|_\infty = \frac{0.2^k}{0.8} \cdot 2 < 10^{-3}, \Rightarrow \\ 0.2^k &< 0.0004 \Rightarrow \\ k &> 4.8614. \end{aligned}$$

Moramo izvršiti $k = 5$ *koraka. Dobit ćemo aproksimacije rješenja:*



Slika 2: Norme greški i reziduala u svakoj iteraciji Jacobijeve metode, za matricu A iz zadatka 1.11.

$$x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad x^{(1)} = \begin{bmatrix} -0.8 \\ 0 \\ 1.2 \\ 2 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} -1 \\ -0.04 \\ 1 \\ 1.96 \end{bmatrix}, \quad x^{(3)} = \begin{bmatrix} -0.992 \\ 0 \\ 1.008 \\ 2 \end{bmatrix},$$

$$x^{(4)} = \begin{bmatrix} -1 \\ -0.0016 \\ 1 \\ 1.9984 \end{bmatrix}, \quad x^{(5)} = \begin{bmatrix} -0.99968 \\ 0 \\ 1.00032 \\ 2 \end{bmatrix},$$

dok je egzaktno rješenje jednako

$$x = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 2 \end{bmatrix}.$$

1.2.3 Gauss–Seidelova metoda

Matricu $A \in \mathbb{R}^{n\times n}$ rastavimo kao u (15).

Gauss–Seidelova metoda je iterativna metoda oblika

$$x^{(k+1)} = M_{GS}^{-1} N_{GS} x^{(k)} + M_{GS}^{-1} b, \quad k = 0, 1, 2, \ldots$$

pri čemu su

$$M_{GS} = D + L, \qquad N_{GS} = -R,$$

odnosno, radi se o iteracijama oblika

$$x^{(k+1)} = T_{GS} x^{(k)} + c_{GS}, \quad k = 0, 1, 2, \ldots$$

za koje su

$$T_{GS} = -(D + L)^{-1} R, \qquad c_{GS} = (D + L)^{-1} b.$$

Algoritam 1.9 *Rješavanje linearnog sustava pomoću Gauss–Seidelove metode.*

x_0 *fiksiran;*
for k=0,1,2,. . .
  begin
  for $i = 1, \ldots, n$
    begin
    $x_i^{(k+1)} = \dfrac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right)$*;*
    end*;*
  end

Uvjete za koje Gauss–Seidelova metoda konvergira, daje sljedeća tvrdnja.

- Ako je matrica sustava $A \in \mathbb{R}^{n\times n}$ simetrična pozitivno definitna matrica, tada Gauss–Seidelova metoda konvergira za svaku početnu iteraciju.

Zadatak 1.12 *Neka je zadan sustav kao u Zadatku 1.11, ali ovaj puta ga rješavamo Gauss-Seidelovom metodom.*

Napomena 1.13 *U ovom slučaju vrijedi:*

$$D+L=\begin{bmatrix} 10 & 0 & 0 & 0 \\ 1 & 10 & 0 & 0 \\ 0 & 1 & 10 & 0 \\ 1 & 0 & 1 & 10 \end{bmatrix}, \quad R=\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Lako se vidi da je

$$(D+L)^{-1}=\begin{bmatrix} 0.1 & 0 & 0 & 0 \\ -0.01 & 0.1 & 0 & 0 \\ 0.001 & -0.1 & 0.1 & 0 \\ -0.0101 & 0.001 & -0.01 & 0.1 \end{bmatrix}.$$

Iz toga slijedi da je

$$T_{GS}=-(D+L)^{-1}R=\begin{bmatrix} 0 & -0.1 & 0 & -0.1 \\ 0 & 0.01 & -0.1 & 0.01 \\ 0 & -0.001 & 0.01 & -0.101 \\ 0 & 0.0101 & -0.001 & 0.0201 \end{bmatrix},$$

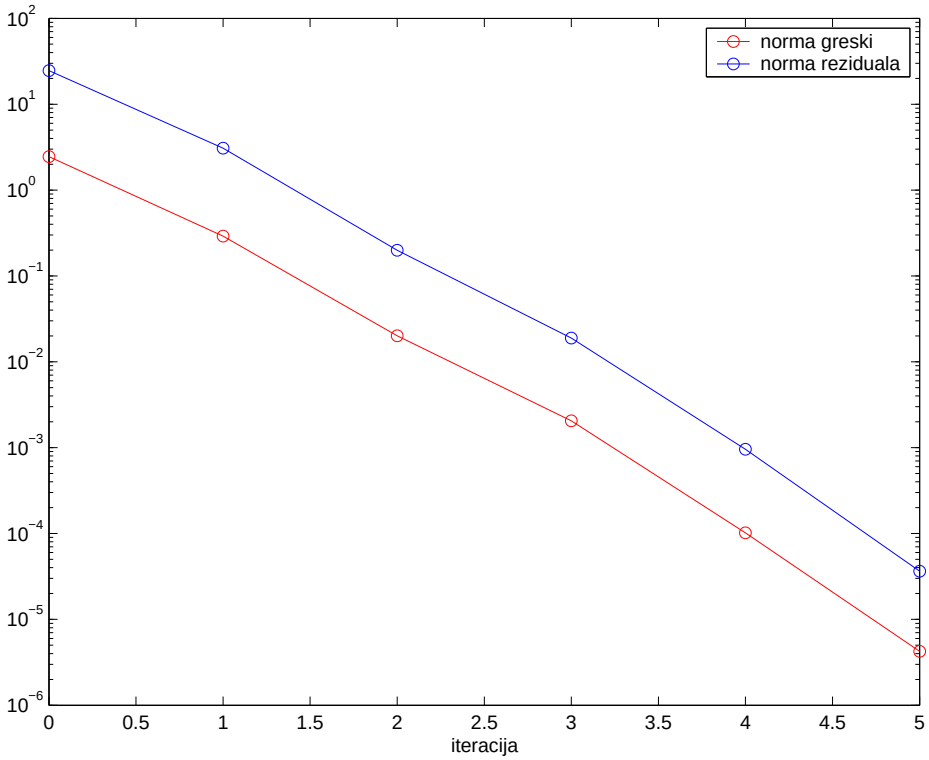
$$c_{GS}=(D+L)^{-1}b=\begin{bmatrix} -0.8 \\ 0.08 \\ 1.192 \\ 1.9608 \end{bmatrix}.$$

Opet provjeravamo da li iterativna metoda uopće konvergira:

$$\rho(T_{GS}) \le \|T_{GS}\|_\infty = 0.2 < 1,$$

dakle, metoda konvergira. Dalje, određujemo koliko koraka moramo izvršiti da bi postigli istu točnost, za $x_0 = 0$.

$$\begin{aligned} \|x^{(k)}-x\|_\infty &\le \frac{\|T_{GS}\|_\infty^k}{1-\|T_{GS}\|_\infty}\|c_{GS}\|_\infty = \frac{0.2^k}{0.8}\cdot 1.9608 < 10^{-3}, \Rightarrow \\ 0.2^k &< 0.000408 \Rightarrow \\ k &> 4.8491. \end{aligned}$$



Slika 3: Norme greški i reziduala u svakoj iteraciji Gauss–Seidelove metode, za matricu A iz zadatka 1.12.

Moramo izvršiti opet $k = 5$ koraka. Dobit ćemo aproksimacije rješenja:

$$x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad x^{(1)} = \begin{bmatrix} -0.8 \\ 0.08 \\ 1.192 \\ 1.9608 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} -1.004 \\ -0.01879 \\ 1.05799 \\ 1.99983 \end{bmatrix},$$

$$x^{(3)} = \begin{bmatrix} -0.99810 \\ -0.00077 \\ 1.00009 \\ 1.99980 \end{bmatrix}, \quad x^{(4)} = \begin{bmatrix} -0.99990 \\ -0.000019 \\ 1.00002 \\ 1.99999 \end{bmatrix}, \quad x^{(5)} = \begin{bmatrix} -0.999997 \\ -0.000002 \\ 1.000001 \\ 1.999999 \end{bmatrix},$$

dok je egzaktno rješenje opet jednako

$$x = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 2 \end{bmatrix}.$$

U ovom slučaju vidimo da, iako nam je ocjena predvidjela da će nam trebati 5 koraka za postizanje točnosti od 10^{-3}, to smo već postigli u 4. koraku.

1.2.4 JOR metoda

Primijećeno je da, kada je spektralni radius iterativne matrice blizu jedan, iteracije vrlo sporo konvergiraju. Zato se u iteracije uvodi **parametar relaksacije** koji nastoji smanjiti spektralni radijus iterativne matrice i ubrzati konvergenciju. To se radi pomoću sljedećeg rastava:

$$A = \frac{1}{\omega}D + \frac{\omega - 1}{\omega}D + L + R.$$

JOR metoda (*Jacobi over relaxation*) je tada iterativna metoda oblika

$$x^{(k+1)} = M_{JOR,\omega}^{-1} N_{JOR,\omega} x^{(k)} + M_{JOR,\omega}^{-1} b, \quad k = 0, 1, 2, \ldots$$

pri čemu su

$$M_{JOR,\omega} = \frac{1}{\omega}D, \qquad N_{JOR,\omega} = \frac{1-\omega}{\omega}D - (L + R),$$

odnosno, radi se o iteracijama oblika

$$x^{(k+1)} = T_{JOR,\omega} x^{(k)} + c_{JOR,\omega}, \quad k = 0, 1, 2, \ldots$$

za koje su

$$T_{JOR,\omega} = (1-\omega)I - \omega D^{-1}(L + R) = (1-\omega)I + \omega T_J, \qquad c_{JOR,\omega} = \omega D^{-1} b.$$

Za $\omega = 1$ JOR iteracije se svode na Jacobijeve iteracija.

Algoritam 1.10 *Rješavanje linearnog sustava pomoću JOR metode.*

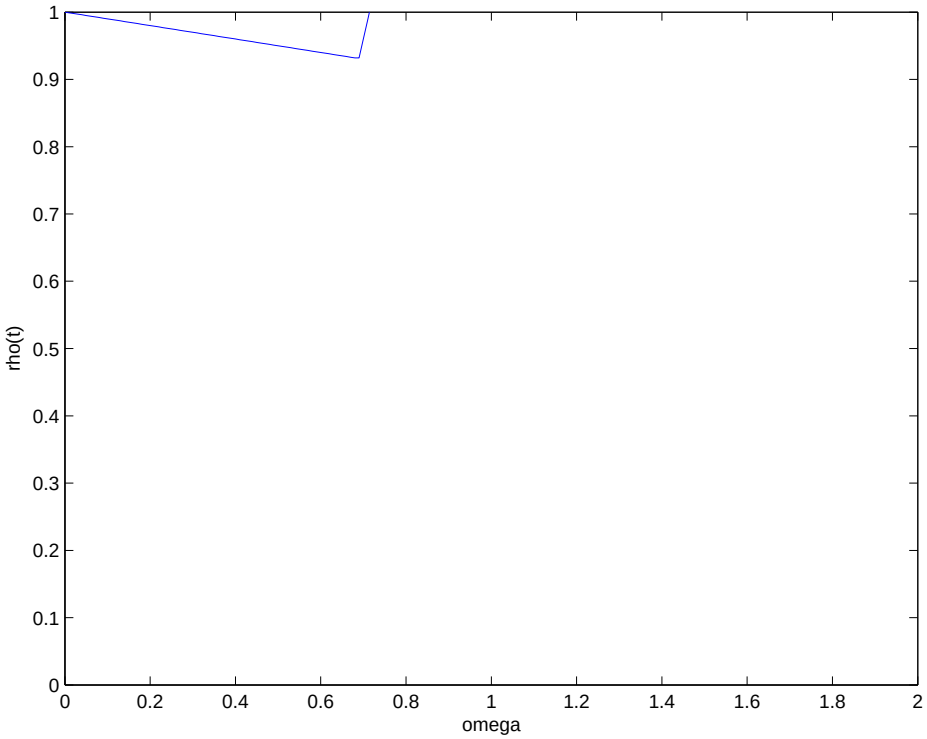
x_0 *fiksiran;*
for k=0,1,2,. . .
  begin
  for $i = 1, \ldots, n$
    begin
    $$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right);$$
    end;
  end

O konvergenciji JOR metode, govore sljedeće tvrdnje.

- Ako Jacobijeva metoda konvergira, onda konvergira i JOR metoda za $\omega \in \langle 0, 1]$ i za svaku početnu iteraciju.
- Ako je matrica sustava $A \in \mathbb{R}^{n\times n}$ simetrična pozitivno definitna matrica, i ako je $|\lambda| < 1$ za sve $\lambda \in \sigma(T_J)$, i ako je $\lambda_{min} = \min\{\lambda : \lambda \in \sigma(T_J)\}$ ($\lambda_{min} \leq 0$), onda
  - JOR metoda konvergira za
$$0 < \omega < \frac{2}{1 - \lambda_{min}} \leq 2$$
    za svaku početnu iteraciju.
  - JOR metoda ne konvergira za
$$\omega < 0 \quad \& \quad \omega \geq 2.$$
- JOR metoda ne konvergira za $\omega < 0$ i $\omega \geq 2$.

Zadatak 1.13 *Odredite ω za koje JOR metoda konvergira ako je matrica sustava $Ax = b$*
$$A = \begin{bmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{bmatrix}.$$

Napomena 1.14 *JOR metoda će konvergirati ako i samo ako je $\rho(T_{JOR,\omega}) < 1$. Zato gledamo*
$$\begin{aligned} T_{JOR,\omega} &= (1-\omega)I + \omega T_J \Rightarrow \\ \sigma(T_{JOR,\omega}) &= (1-\omega) + \omega\sigma(T_J), \end{aligned}$$
odakle je
$$\sigma(T_{JOR,\omega}) \in \{1 - \omega + \omega\lambda : \lambda \in \sigma(T_J)\}.$$
Imamo
$$T_J = \begin{bmatrix} 0 & -0.9 & -0.9 \\ -0.9 & 0 & -0.9 \\ -0.9 & -0.9 & 0 \end{bmatrix},$$
Odakle slijedi da je
$$\sigma(T_J) = \{0.9, -1.8\}.$$



Slika 4: Spektralni radijus JOR iterativne matrice za matricu A iz zadatka 1.13. Minimum se postiže za $\omega \approx 0.7$.

Dakle

$$\sigma(T_{JOR,\omega}) = \{1 - \omega + 0.9\omega, 1 - \omega - 1.8\omega\} = \{1 - 0.1\omega, 1 - 2.8\omega\},$$

pa za spektralni radijus mora vrijediti

$$\rho(T_{JOR,\omega}) = \max\{|1 - 0.1\omega|, |1 - 2.8\omega|\} < 1.$$

To će vrijediti ako i samo ako

$$|1 - 0.1\omega| < 1, \quad \& \quad |1 - 2.8\omega| < 1.$$

Dalje, računamo

$$-1 < 1 - 0.1\omega < 1 \;\Leftrightarrow\; -2 < -0.1\omega < 0 \;\Leftrightarrow\; 0 < \omega < 20$$

$$-1 < 1 - 2.8\omega < 1 \;\Leftrightarrow\; -2 < -2.8\omega < 0 \;\Leftrightarrow\; 0 < \omega < 0.7143$$

Budući da obje nejednakosti moraju biti zadovoljene, konačno rješenje je

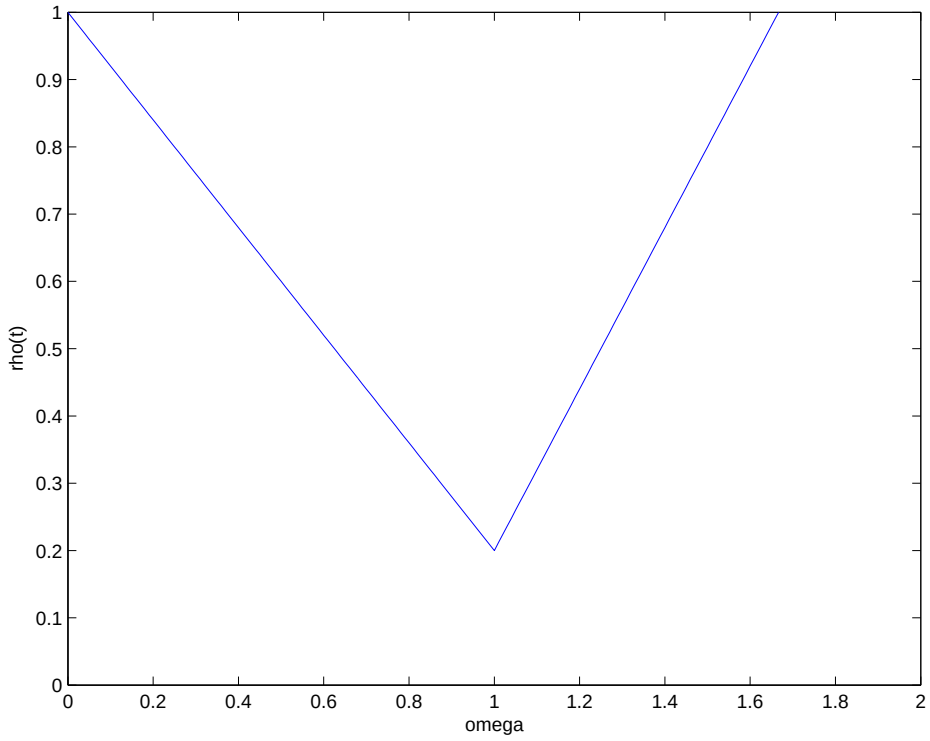
$$\omega \in \langle 0, 0.7143 \rangle.$$

Ako želimo naći ω za kojeg je konvergencija najbrža, tada zahtijevamo

$$\max\{|1 - 0.1\omega|, |1 - 2.8\omega|\} \longrightarrow \min .$$

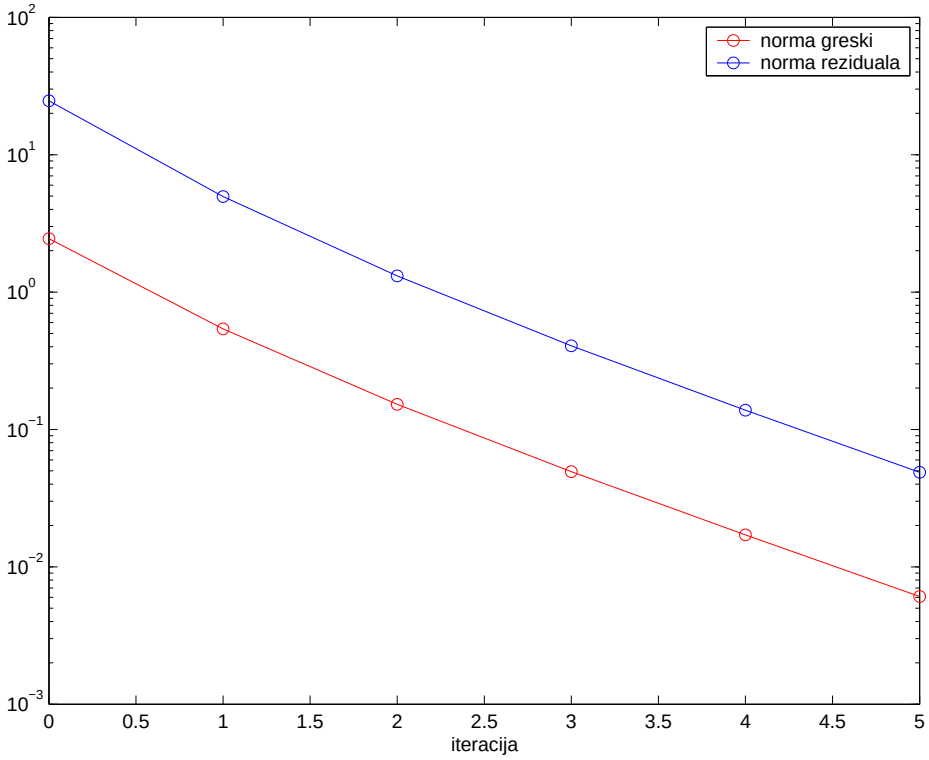
Zadatak 1.14 *Neka je zadan sustav kao u Zadatku 1.11, ali ovaj puta ga rješavamo JOR metodom.*

Napomena 1.15 *Prvo pogledajmo kada će metoda konvergirati.*

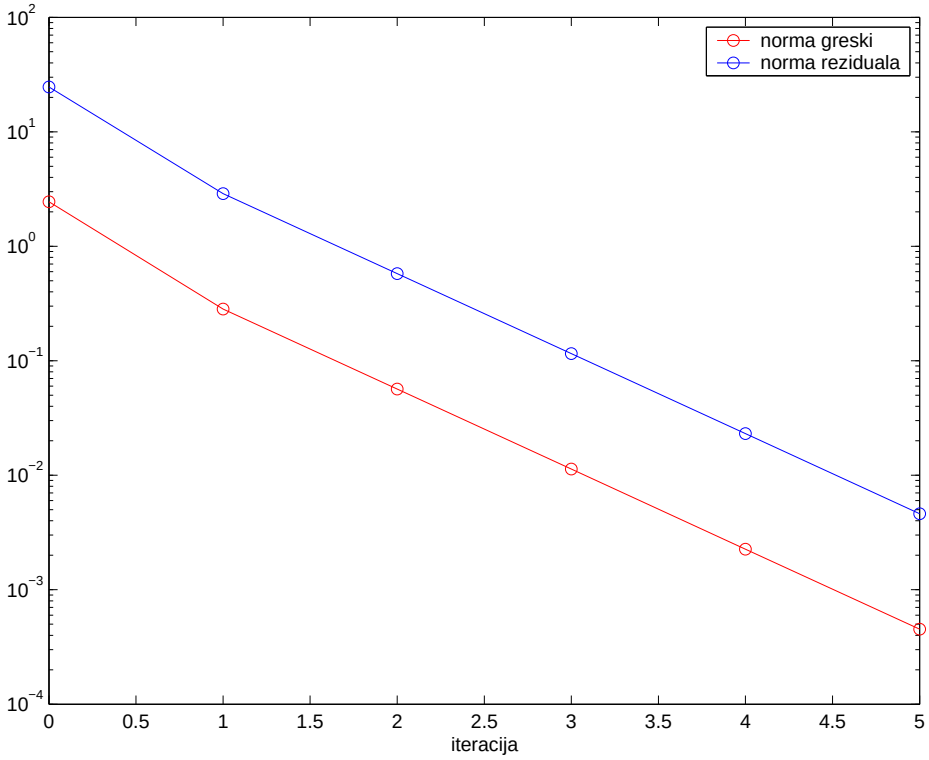


Slika 5: Spektralni radijus JOR iterativne matrice za matricu A iz zadatka 1.14.

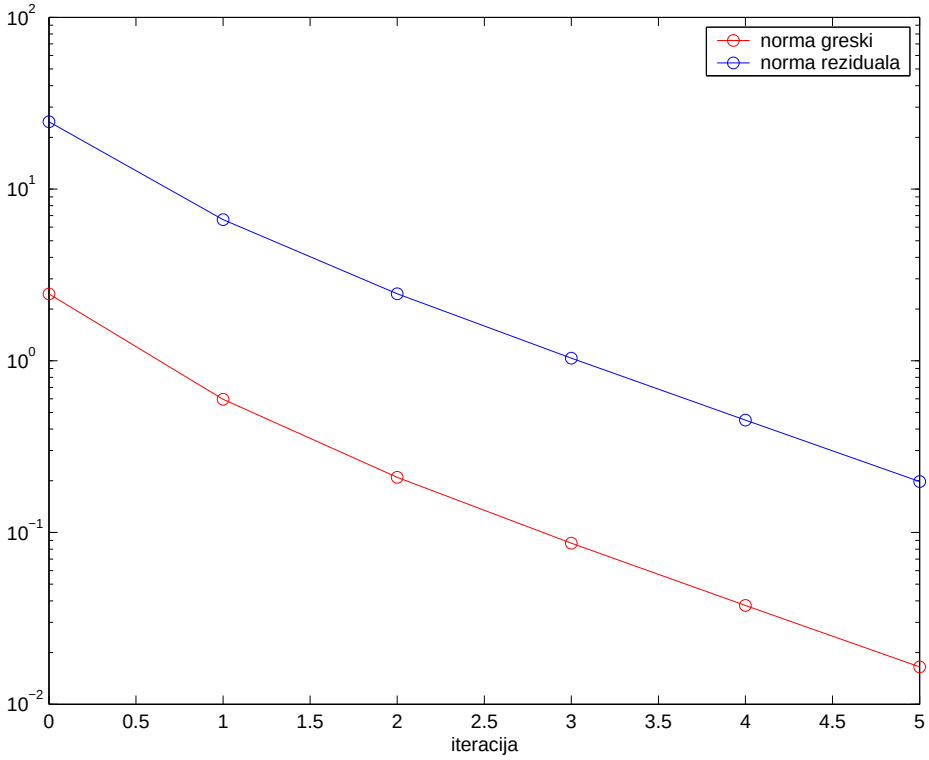
Iz grafa vidimo da za ovu matricu, JOR metoda konvergira za otprilike $\omega \in \langle 0, 1.65 \rangle$, a da najbržu konvergenciju dostiže za $\omega = 1$, tj. upravo za Jacobijevu iterativnu metodu. Riješite sustav za $\omega = 0.8, 1, 1.2$. Brzina konvergencije u tim numeričkim primjerima je u skladu sa slikom 5.



Slika 6: Norme greški i reziduala u svakoj iteraciji JOR metode, za matricu A iz zadatka 1.14 i parametar relaksacije $\omega = 0.8$.



Slika 7: Norme greški i reziduala u svakoj iteraciji JOR metode, za matricu A iz zadatka 1.14 i parametar relaksacije $\omega = 1$.



Slika 8: Norme greški i reziduala u svakoj iteraciji JOR metode, za matricu A iz zadatka 1.14 i parametar relaksacije $\omega = 1.2$.

1.2.5 SOR metoda

Opet imamo rastav:

$$A = \frac{1}{\omega}D + L + \frac{\omega - 1}{\omega}D + R.$$

SOR metoda(*Succesive over relaxation*) je tada iterativna metoda oblika

$$x^{(k+1)} = M_{SOR,\omega}^{-1} N_{SOR,\omega} x^{(k)} + M_{SOR,\omega}^{-1} b, \quad k = 0, 1, 2, \ldots$$

pri čemu su

$$M_{SOR,\omega} = \frac{1}{\omega}D + L, \qquad N_{SOR,\omega} = \frac{1-\omega}{\omega}D - R,$$

odnosno, radi se o iteracijama oblika

$$x^{(k+1)} = T_{SOR,\omega} x^{(k)} + c_{SOR,\omega}, \quad k = 0, 1, 2, \ldots$$

za koje su

$$T_{SOR,\omega} = (1-\omega)(D+\omega L)^{-1}D - \omega(D+\omega L)^{-1}R, \qquad c_{SOR,\omega} = \omega(D+\omega L)^{-1}b.$$

Ako se iteracije napišu na drugačiji način, onda se dobiva:

$$x^{(k+1)} = (1-\omega)x^{(k)} + \omega D^{-1}(b - Lx^{(k+1)} - Rx^{(k)}) = (1-\omega)x^{(k)} + \omega x_{GS}^{(k+1)}.$$

Za $\omega = 1$ SOR iteracije se svode na Gauss–Seidelove iteracije.

Algoritam 1.11 *Rješavanje linearnog sustava pomoću SOR metode.*

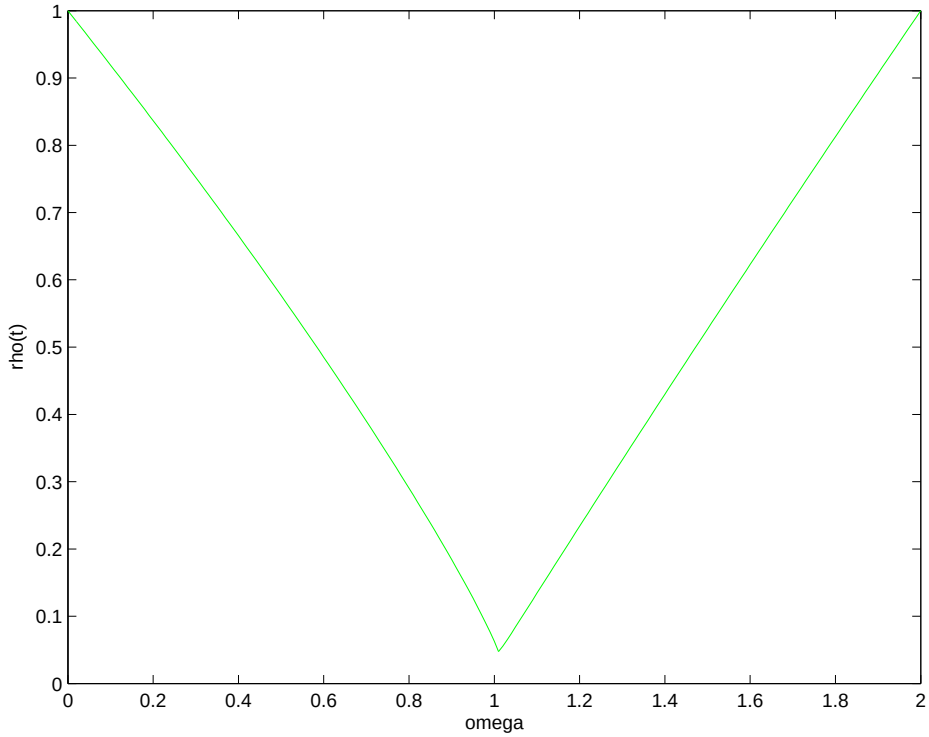
x_0 *fiksiran;*
for k=0,1,2,. . .
  begin
  for $i = 1, \ldots, n$
    begin
    $$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right);$$
    end*;*
  end

O konvergenciji SOR iterativnog postupka, govore sljedeće tvrdnje.

- Ako je matrica sustava $A \in \mathbb{R}^{n\times n}$ simetrična pozitivno definitna matrica, tada SOR metoda konvergira za $\omega \in \langle 0, 2\rangle$ i za svaku početnu iteraciju.
- SOR metoda ne konvergira za $\omega < 0$ i $\omega \geq 2$.

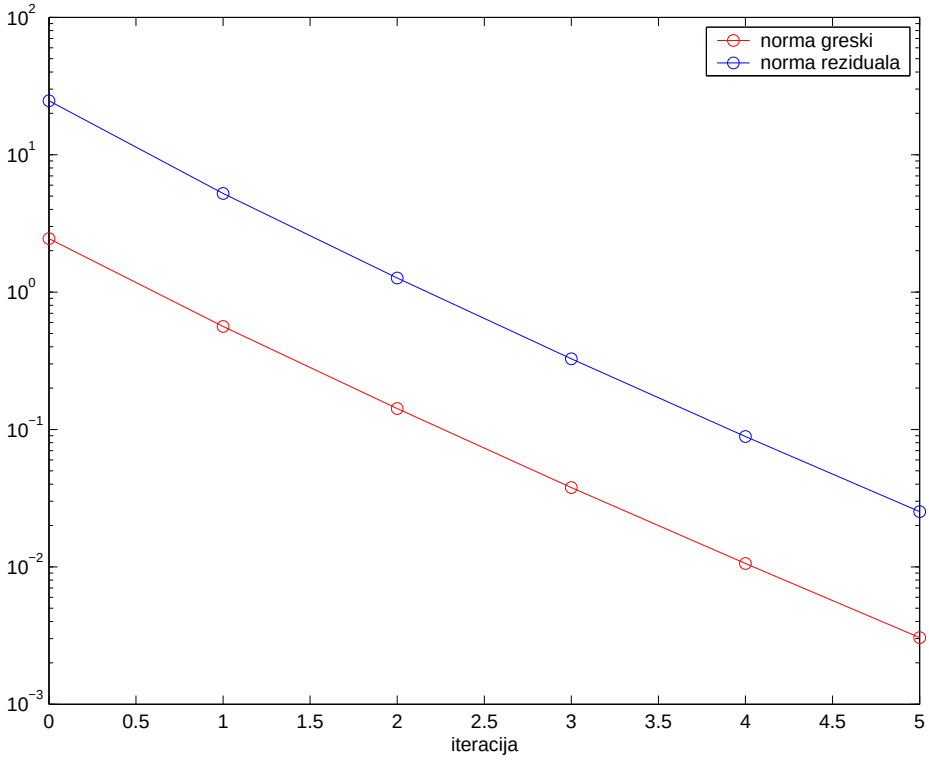
Zadatak 1.15 *Neka je zadan sustav kao u Zadatku 1.11, ali ovaj puta ga rješavamo SOR metodom.*

Napomena 1.16 *Prvo pogledajmo kada će metoda konvergirati.*

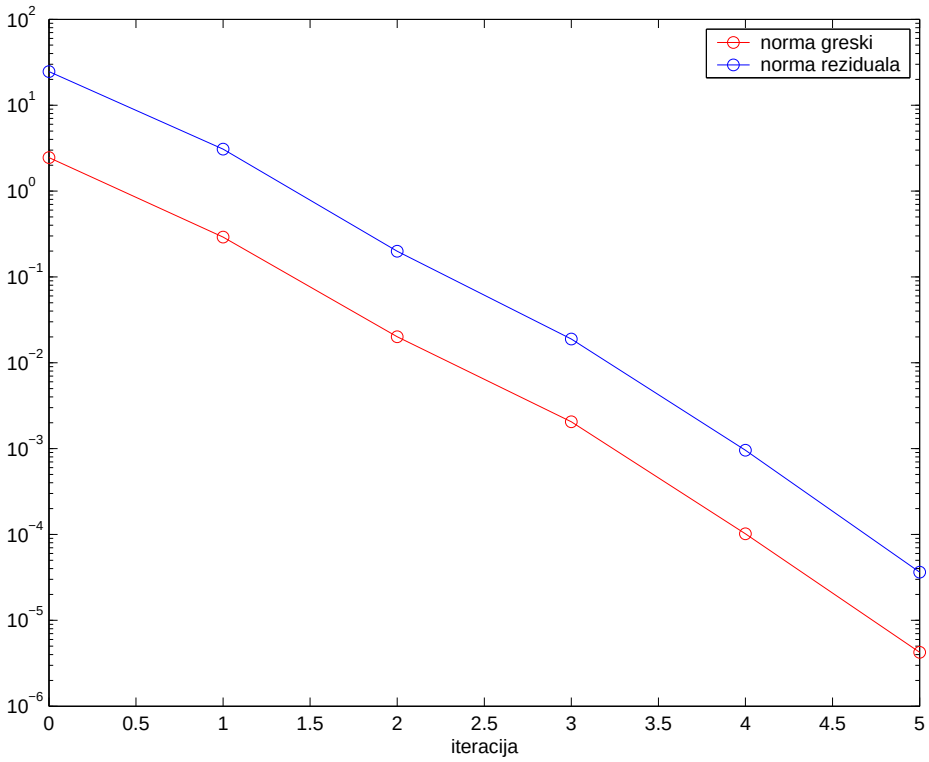


Slika 9: Spektralni radijus SOR iterativne matrice za matricu A iz zadatka 1.15.

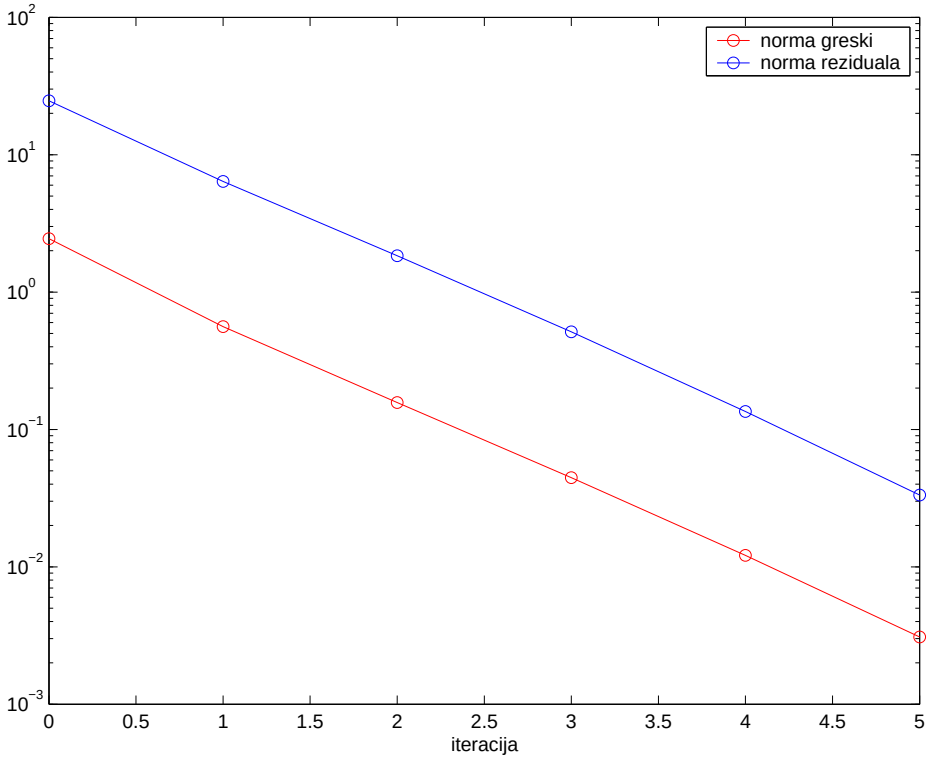
Iz grafa vidimo da za ovu matricu, SOR metoda konvergira za $\omega \in \langle 0, 2\rangle$, a da najbržu konvergenciju dostiže opet za $\omega = 1$, tj. upravo za Gauss–Seidelovu iterativnu metodu. Riješite sustav za $\omega = 0.8, 1, 1.2$. Brzina konvergencije u tim numeričkim primjerima je u skladu sa slikom 9.



Slika 10: Norme greški i reziduala u svakoj iteraciji SOR metode, za matricu A iz zadatka 1.15 i parametar relaksacije $\omega = 0.8$.



Slika 11: Norme greški i reziduala u svakoj iteraciji SOR metode, za matricu A iz zadatka 1.15 i parametar relaksacije $\omega = 1$.



Slika 12: Norme greški i reziduala u svakoj iteraciji SOR metode, za matricu A iz zadatka 1.15 i parametar relaksacije $\omega = 1.2$.

1.2.6 Iteracije iz Krylovljevih potprostora. Konjugirani gradijenti

Prisjetimo se najprije rezultata iz linearne algebre, koji tvrdi da svaka matrica poništava svoj karakteristični i minimalni polinom. Za $A \in \mathbb{C}^{n\times n}$ i $b \in \mathbb{C}^n$, to možemo zapisati na sljedeći način

$$\kappa_A(A) = a_0 I + a_1 A + \cdots + a_{n-1}A^{n-1} + a_n A^n = 0,$$

gdje je $\kappa_A(\lambda) = \det(A - \lambda I) = \sum_{i=0}^{n} a_i \lambda^i$ karakteristični polinom matrice A.

U slučaju kada je matrica regularna, nula ne može biti korijen karakterističnog polinoma, pa je $a_0 \neq 0$. Odavde jednostavnim računom možemo dobiti da vrijedi

$$\begin{aligned} &-\frac{1}{a_0}(a_1 I + \cdots + a_{n-1}A^{n-2} + a_n A^{n-1}) \cdot A = \\ &= \quad A \cdot \left(-\frac{1}{a_0}\right)(a_1 I + \cdots + a_{n-1}A^{n-2} + a_n A^{n-1}) = I, \end{aligned}$$

to jest, da je

$$A^{-1} = -\frac{1}{a_0}(a_1 I + \cdots + a_{n-1}A^{n-2} + a_n A^{n-1}). \tag{16}$$

Budući da rješenje sustava $Ax = b$ možemo zapisati kao $x = A^{-1}b$, uz uvažavanje prethodnog zapisa za A^{-1} možemo zaključiti da je

$$x = -\frac{a_1}{a_0}b - \cdots - \frac{a_{n-1}}{a_0}A^{n-2}b - \frac{a_n}{a_0}A^{n-1}b,$$

odnosno

$$x \in \ \text{span}\{b, Ab, \ldots, A^{n-1}b\} = \mathcal{K}_n(A, b). \tag{17}$$

Prostor koji se pojavljuje na desnoj strani u (17) zovemo **Krylovljevim prostorom** matrice A i inicijalnog vektora b. Upravo iz (17) dobivamo ideju za iterativne metode rješavanja sustava linearnih jednadžbi koje bi se temeljile na aproksimacijama iz Krylovljevih potprostora.

Jedan način definicije iteracije je

$$x_{k+1} = x_k + \alpha_k(b - Ax_k) = x_k + \alpha_k r_k \tag{18}$$

gdje je α_k dinamički parametar koji se određuje iz nekih optimizirajućih uvjeta. Najprije definirajmo osnovne pojmove:

$$\begin{aligned} x_0 &= \text{početna aproksimacija} \\ e_k &= x - x_k = \text{greška}, \qquad k = 0, 1, \ldots \quad (x = A^{-1}b) \\ r_k &= b - Ax_k = Ae_k = \text{rezidual}, \qquad k = 0, 1, \ldots \end{aligned}$$

Pogledajmo sada u kojim potprostorima se nalaze iteracije (18).

$$\begin{aligned}
x_0 &= x - e_0 \quad \in x + \operatorname{span}\{e_0\} \\
x_1 &= x_0 + \alpha_0 A e_0 = x - e_0 + \alpha_0 A e_0 \quad \in x + \operatorname{span}\{e_0, Ae_0\} \\
x_2 &= x_1 + \alpha_1 A e_1 = x - e_0 + \alpha_0 A e_0 + \alpha_1 A(e_0 - \alpha_0 A e_0) = \\
&= x - e_0 + (\alpha_0 + \alpha_1) A e_0 - \alpha_0 \alpha_1 A^2 e_0 \quad \in x + \operatorname{span}\{e_0, Ae_0, A^2 e_0\} \\
&\vdots \\
x_k &= x_{k-1} + \alpha_{k-1} A e_{k-1} \quad \in x + \operatorname{span}\{e_0, Ae_0, \ldots, A^k e_0\}.
\end{aligned}$$

Dakle, općenito vrijedi za x_k definiran u (18)

$$x_k \in x + \mathcal{K}_{k+1}(A, e_0), \qquad k = 0, 1, \ldots$$

Postavlja se pitanje na koji ćemo način birati parametar α_k.

U ovom odjeljku opisat ćemo dobivanje iterativne metode iz Krylovljevih potprostora, za rješavanje sustava $Ax = b$, $A \in \mathbb{R}^{n\times n}$, $b, x \in \mathbb{R}^n$, pri čemu je matrica sustava A *simetrična pozitivno definitna*. Ideja odabira parametra α_k u k-toj iteraciji metode je ta da se minimizira neka norma greške $e_{k+1} = e_k - \alpha_k r_k$. Problem je što nam je greška jednako tako nepoznata kao i samo rješenje, pa norma $\|\cdot\|_2$ ne dolazi u obzir. Ono što mi možemo izračunati je rezidual. $r_{k+1} = Ae_{k+1} = r_k - \alpha_k A r_k$. Zato za simetričnu pozitivno definitnu matricu A ima smisla definirati *A-normu* $\|\cdot\|_A$

$$\|v\|_A = \sqrt{\langle v, v\rangle_A} = \sqrt{v^T A v}.$$

Za domaću zadaću možete provjeriti da je $\|\cdot\|_A$ zaista norma na $\mathbb{R}^n$.

Definirajmo sada funkciju $f : \mathbb{R} \longrightarrow \mathbb{R}$ kao

$$\begin{aligned}
f(\alpha_k) &= e_{k+1}^T A e_{k+1} = \\
&= \alpha_k^2 r_k^T A r_k - 2\alpha_k r_k^T A e_k + e_k^T A e_k = \\
&= \alpha_k^2 r_k^T A r_k - 2\alpha_k r_k^T r_k + e_k^T A e_k.
\end{aligned}$$

Traženje minimuma funkcije $f(\alpha_k)$ je ekvivalentno traženju minimuma $\|e_{k+1}\|_A$. Funkcija $f(\alpha_k)$ je kvadratna funkcija po varijabli α_k, i parametar uz α_k^2 je $r_k^T A r_k \geq 0$, što znači da funkcija poprima minimum u tjemenu, koje je jedina nultočka derivacije funkcije f'.

$$0 = f'(\alpha_k) = 2\alpha_k r_k^T A r_k - 2 r_k^T r_k,$$

odakle slijedi da se minimalna A-norma greške e_{k+1} postiže za

$$\alpha_k = \frac{r_k^T r_k}{r_k^T A r_k}.$$

Zbog ovakvog odabira parametra α_k vrijedi da je r_{k+1} okomit na r_k, tj.

$$r_k^T r_{k+1} = r_k^T r_k - \alpha_k r_k^T A r_k = r_k^T r_k - \frac{r_k^T r_k}{r_k^T A r_k} r_k^T A r_k = 0.$$

Važno je još primijetiti, da tako dugo dok nismo našli egzaktno rješenje, to jest dok je $r_k \neq 0$, α_k je strogo veći od nule. Zbog okomitosti r_{k+1} i r_k slijedi

$$\|e_k\|_A^2 = \|e_{k+1}\|_A^2 + \alpha_k^2 \|r_k\|_A^2 > \|e_{k+1}\|_A^2,$$

odakle se vidi da se A-norma greške smanjuje u svakom koraku. Ova metoda, zbog načina odabira parametra, zove se **metoda najbržeg silaska**.

Algoritam 1.12 *Rješavanje linearnog sustava pomoću metode najbržeg silaska.*

x_0 *fiksiran;*
$r_0 = b - Ax_0$*;*
for k=0,1,2,. . .
  $\alpha_k = \frac{r_k^T r_k}{r_k^T A r_k}$*;*
  $x_{k+1} = x_k + \alpha_k r_k$*;*
  $r_{k+1} = r_k - \alpha_k A r_k$*;*
end

Primijećeno je, međutim, da ova metoda može dosta sporo konvergirati,jer se često događa da ona radi korake u smjeru kojim je neki raniji korak već prošao. Da bi se to izbjeglo, unaprijed odabiremo skup A-ortogonalnih vektora, odnosno smjerove traganja d_0, d_1,. . . , d_{n-1}. Dva vektora d_i i d_j su A-ortogonalna ili *konjugirana* ako vrijedi da je

$$\langle d_i, d_j \rangle_A = d_j^T A d_i = 0.$$

Lagano se može provjeriti da su A-ortogonalni vektori linearno nezavisni. Znači u svakom koraku biramo točku

$$x_{k+1} = x_k + \alpha_k d_k \tag{19}$$

s minimalnom A-normom greške.

Dakle, u svakom smjeru d_k napravit ćemo točno jedan korak, i taj korak će biti takve dužine da ćemo poništiti komponentu vektora greške e_k u smjeru Ad_k. Nakon n koraka bit ćemo gotovi. U $(k+1)$-om koraku onda biramo e_{k+1} takav da bude jednak početnoj grešci, kojoj su odstranjene sve komponente u smjerovima Ad_0,. . . ,Ad_k, odnosno on je A-ortogonalan na d_0,. . . ,d_k. A-ortogonalnost između e_{k+1} i d_k je ekvivalentna nalaženju točke minimuma duž smjera traganja d_k, kao i u metodi najbržeg silaska. Da bi to vidjeli,

ponovo ćemo derivirati po α_k funkciju $g(\alpha_k) = e_{k+1}^T A e_{k+1}$ i izjednačiti je s nulom, samo što je u tom slučaju r_{k+1} okomit na d_k. Ako opet uvrstimo da je $r_{k+1} = r_k - \alpha_k A d_k$ i $e_{k+1} = e_k - \alpha_k d_k$, dobit ćemo izraz za α_k

$$\alpha_k = \frac{d_k^T r_k}{d_k^T A d_k}. \tag{20}$$

Ovako dobivena metoda naziva se **metoda konjugiranih smjerova**.

Algoritam 1.13 *Rješavanje linearnog sustava pomoću metode konjugiranih smjerova.*

x_0 *fiksiran;*
A-ortogonalni vektori $d_0, d_1, \ldots, d_{n-1}$ *fiksirani;*
$r_0 = b - Ax_0$*;*
for k=0,1,2,. . .
  $\alpha_k = \frac{d_k^T r_k}{d_k^T A d_k}$*;*
  $x_{k+1} = x_k + \alpha_k d_k$*;*
  $r_{k+1} = r_k - \alpha_k A d_k$*;*
end

Svojstva metode konjugiranih smjerova su sljedeća.

- Za metodu konjugiranih smjerova vrijede sljedeća svojstva:

$$d_j^T A d_i = 0 \quad (i \neq j) \tag{21}$$

$$d_j^T r_i = d_j^T A e_i = 0 \quad (j < i) \tag{22}$$

$$d_i^T r_0 = d_i^T r_1 = \cdots = d_i^T r_i. \tag{23}$$

  Skalar α_i može se zato napisati kao

$$\alpha_k = \frac{d_k^T r_0}{d_k^T A d_k}. \tag{24}$$

- Metoda konjugiranih smjerova je m-koračna metoda ($m \leq n$), u smislu da je u m-tom koraku aproksimacija x_m jednaka rješenju $x = A^{-1}b$.

Ostaje još pronalaženje odgovarajućih vektora $d_0, d_1, \ldots, d_{n-1}$. Skup A-ortogonalnih smjerova $\{d_i\}$ možemo dobiti uz pomoć Gramm–Schmidtove metode A-ortogonalizacije na niz linearno nezavisnih vektora $u_0, \ldots, u_{n-1}$ sa skalarnim produktom $\langle \cdot, \cdot \rangle_A$. Dakle, A-ortogonalne vektore možemo dobiti kao

$$d_k = u_k + \sum_{i=0}^{k-1} \beta_{ki} d_i, \tag{25}$$

pri čemu su koeficijenti oblika

$$\beta_{ki} = -\frac{d_i^T A u_k}{d_i^T A d_i}. \tag{26}$$

Konkretan odabir vektora $u_0,\ldots,u_{n-1}$ vodi nas do **metode konjugiranih gradijenata**.

Metoda konjugiranih gradijenata (CG) je, zapravo, metoda konjugiranih smjerova kod koje se smjerovi traganja konstruiraju primjenom Gram–Schmidtove metode A-ortogonalnosti na reziduale, tj. uzima se da je $u_i = r_i$. Činjenica da su vektori r_i dobiveni metodom konjugiranih smjerova linearno nezavisni, može se provjeriti uz pomoć (22) i (23). Ponovo vrijedi

$$\text{span}\{d_0, d_1, \ldots, d_{k-1}\} = \text{span}\{r_0, r_1, \ldots, r_{k-1}\},$$

i budući da je r_k ortogonalan na prethodne smjerove traganja zbog (22), on je onda zbog prethodne tvrdnje, ortogonalan i na prethodne reziduale, odnosno vrijedi

$$r_i^T r_j = 0, \qquad i \neq j. \tag{27}$$

Promatramo sljedeći skalarni produkt

$$r_k^T r_{i+1} = r_k^T r_i - \alpha_i r_k^T A d_i,$$

pa odavde vrijedi

$$d_i^T A r_k = \frac{1}{\alpha_i}(r_k^T r_i - r_k^T r_{i+1}). \tag{28}$$

Za $i < k-1$ lijeva strana u (28) je jednaka 0, pa su $\beta_{ki} = 0$ za $i = 0, 1, \ldots, k-2$, a za $\beta_k = \beta_{k,k-1}$ vrijedi

$$\begin{aligned} \beta_k &= \frac{r_k^T r_k}{d_{k-1}^T r_{k-1}} = \qquad \text{zbog (20)}, \\ &= \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}}, \qquad \text{zbog (22) i (25)}. \end{aligned} \tag{29}$$

Zbog (20), (22) i (25) možemo i α_k napisati u ljepšem obliku

$$\alpha_k = \frac{r_k^T r_k}{d_k^T A d_k}, \tag{30}$$

odakle se vidi, da ukoliko nismo našli egzaktno rješenje u k-tom koraku, α_k je pozitivan.

Sada smo u potpunosti definirali metodu konjugiranih gradijenata.

Algoritam 1.14 *Rješavanje linearnog sustava pomoću metode konjugiranih gradijenata.*

x_0 *fiksiran;*
$d_0 = r_0 = b - Ax_0$*;*
for k=0,1,2,. . .
$\alpha_k = \frac{r_k^T r_k}{d_k^T A d_k}$*;*
$x_{k+1} = x_k + \alpha_k d_k$*;*
$r_{k+1} = r_k - \alpha_k A d_k$*;*
$\beta_{k+1} = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$*;*
$d_{k+1} = r_{k+1} + \beta_{k+1} d_k$*;*
end

Navedimo još nekoliko svojstava, koja su vezana uz metodu konjugiranih gradijenata.

- Greška e_k dobivena u k-tom koraku metode konjugiranih gradijenata ima najmanju A-normu na prostoru

$$e_0 + \text{span}\{Ae_0, A^2e_0, \ldots, A^k e_0\}. \tag{31}$$

- U svakom koraku CG algoritma, duljina vektora greške $e_k = x - x_k$ se reducira, pri čemu je $A^{-1}b = x = x_m$, za neki $m \leq n$.

O analizi greške i konvergenciji metode konjugiranih gradijenata govore sljedeće opservacije.

- Zbog relacije (31), greška u k-tom koraku metode ima oblik

$$e_k = e_0 + \sum_{i=1}^{k} \psi_i A^i e_0 = \left(I + \sum_{i=1}^{k} \psi_i A^i \right) e_0.$$

  Koeficijenti ψ_i su u linearnoj vezi sa koeficijentima α_i i β_i, a metoda konjugiranih gradijenata bira ψ_j takve da oni minimiziraju $\|e_k\|_A$. Tada, izraz za grešku možemo izraziti kao

$$e_k = p_k(A)e_0, \tag{32}$$

  gdje je p_k polinom k-tog stupnja kod kojeg zahtijevamo da je $p_k(0) = 1$.

  Kako je matrica A simetrična i pozitivno definitna, tada matricu možemo zapisati kao produkt matrica $A = U\Lambda U^T$, pri čemu su za $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, $\lambda_1, \ldots, \lambda_n$ svojstvene vrijednosti od A i $U^T U = UU^T = I$. Još imamo

da je $p_k(A) = Up_k(\Lambda)U^T$. $A^{1/2}$ je hermitski drugi korijen od A i vrijedi $A^{1/2} = U\Lambda^{1/2}U^T$, pa komutira sa A i sa $p_k(A)$. Zbog toga slijedi

$$\begin{aligned}
\|e_k\|_A &= \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|p_k(A)e_0\|_A = \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \sqrt{e_0^T p_k(A)Ap_k(A)e_0} = \\
&= \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|A^{1/2}p_k(A)e_0\|_2 = \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|p_k(A)A^{1/2}e_0\|_2 \leq \\
&\leq \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|p_k(A)\|_2\|A^{1/2}e_0\|_2 = \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|p_k(\Lambda)\|_2\|e_0\|_A,
\end{aligned}$$

dakle konačno možemo napisati

$$\|e_k\|_A \leq \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \max_{i=1,\ldots,n} |p_k(\lambda_i)|\|e_0\|_A. \tag{33}$$

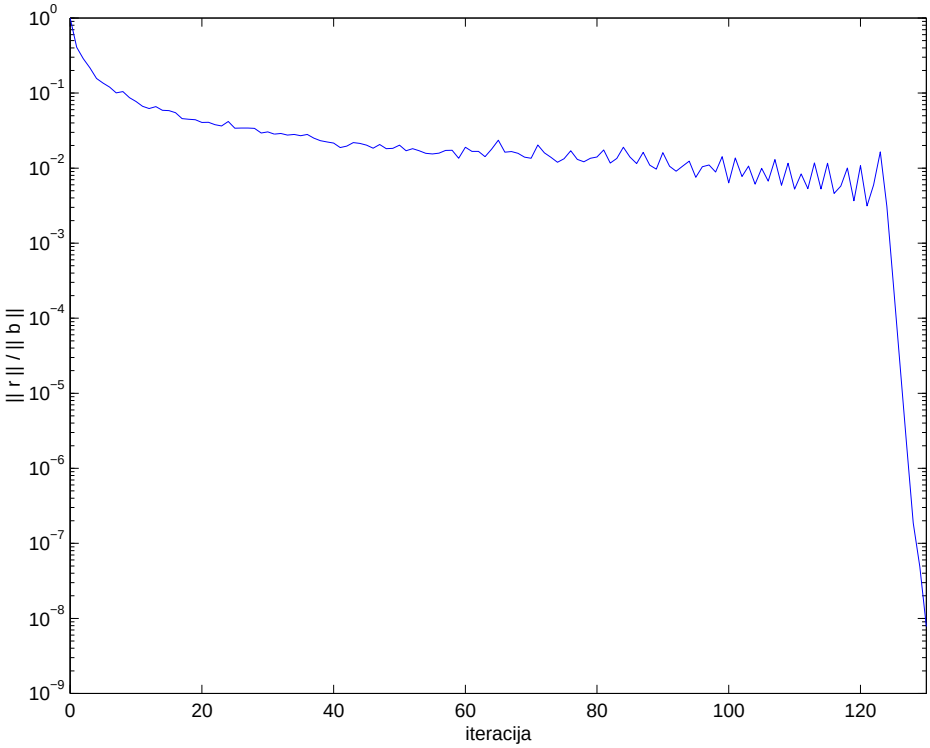
- Primijenjiva ocjena greške u k-tom koraku metode konjugiranih gradijenata je dana sa

$$\|e_k\|_A \leq 2\left(\frac{\sqrt{\kappa_2(A)}-1}{\sqrt{\kappa_2(A)}+1}\right)^k \|e_0\|_A. \tag{34}$$

pri čemu je $\kappa(A)_2 = \|A\|_2 \cdot \|A^{-1}\|_2 = \lambda_{max}/\lambda_{min}$ broj uvjetovanosti matrice A.

Zadatak 1.16 *Matrica sustava u ovom zadatku je simetrična pozitivno definitna* 100×100 *matrica, sa svojstvenim vrijednostima* $\lambda(A) \in \{1, 4, 9, \ldots, 10000\}$, *a dobivena je kao produkt* $A = Q\Lambda Q^T$, *pri čemu je* Λ *dijagonalna matrica svojstvenih vrijednosti, a* Q *slučajna ortogonalna matrica. Uvjetovanost joj je jednaka* $\kappa(A) = 10^4$. *Za početnu iteraciju uzet ćemo* $x_0 = [0\ 0\ \ldots\ 0]^T$, *a za desnu stranu sustava,* b *je određena tako da rješenje sustava bude jednako* $x = [1\ 1\ \ldots\ 1]^T$, *odnosno da je* $b = A \cdot x$. *Riješite ovaj sustav u* Octave-*i, pomoću metode konjugiranih gradijenata i u svakom koraku* k *kontrolirajte relativnu normu reziduala* $\|r_k\|_2/\|b\|_2$. *Iteriranje se treba zaustaviti kada je ona manja od* $tol = 10^{-8}$.

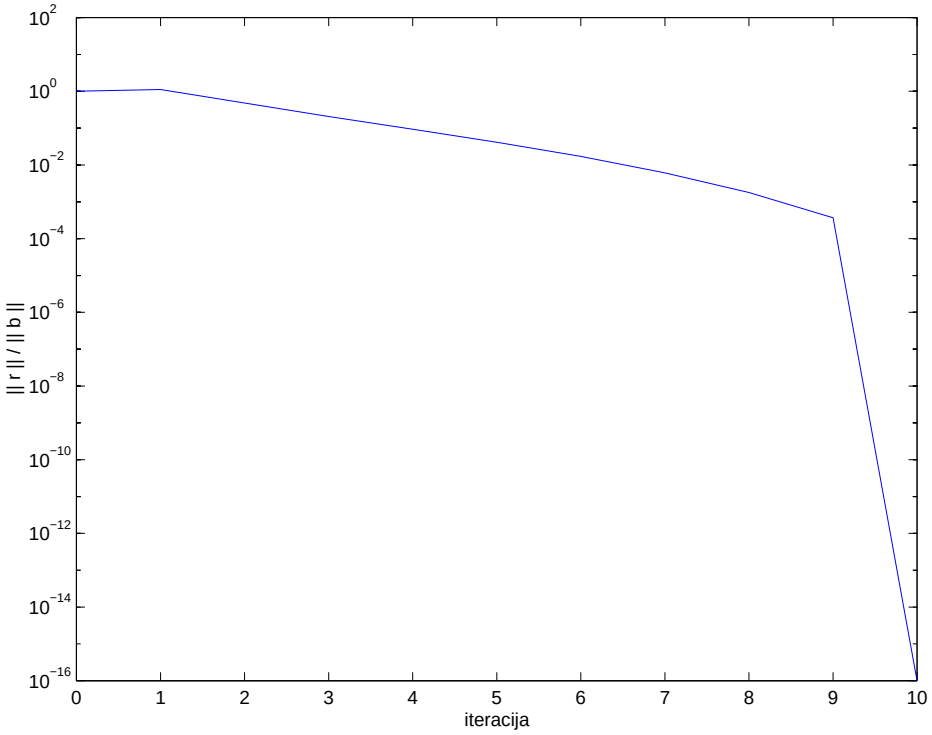
Napomena 1.17 *Metoda konjugiranih gradijenata bi u egzaktnoj aritmetici trebala konvergirati u najviše 100 iteracija, međutim budući da je matrica* A *loše uvjetovana, i budući da broj iteracija do postizanja željene točnosti ovisi o* $\mathcal{O}(\sqrt{\kappa})$, *broj iteracija do postizanja tolearncije tol u aritmetici konačne preciznosti je veći od 100.*



Slika 13: Relativne norme reziduala u svakoj iteraciji metode konjugiranih gradijenata, za matricu A iz zadatka 1.16.

Zadatak 1.17 *Situacija u ovom zadatku je slična prethodnom, samo što pozitivno definitna matrica A ima deset različitih svojstvenih vrijednosti, svaka od njih kratnosti 10. Dakle, $A = Q\Lambda Q^T$, gdje je Λ dijagonalna matrica svojstvenih vrijednosti $\lambda(A) \in \{1, 2, \ldots, 10\}$, a Q slučajna ortogonalna matrica. Uvjetovanost matrice A iznosi $\kappa(A) = 10$. b i x_0 se određuju kao u prethodnom zadatku. Riješite ovaj sustav u* Octave*-i, pomoću metode konjugiranih gradijenata i u svakom koraku k kontrolirajte relativnu normu reziduala $\|r_k\|_2/\|b\|_2$. Iteriranje se treba zaustaviti kada je ona manja od $tol = 10^{-8}$.*

Napomena 1.18 *Budući da, konvergencija CG metoda ovisi o stupnju polinoma koji minimizira vrijednost polinoma u različitim svojstvenim vrijednostima matrice A, takav polinom može imati minimalno stupanj 10, pa je za konvergenciju dovoljno 10 iteracija metode. Vidi ocjenu (33).*



Slika 14: Relativne norme reziduala u svakoj iteraciji metode konjugiranih gradijenata, za matricu A iz zadatka 1.17.

1.2.7 GMRES metoda

Jedna od metoda koja se može primijeniti na sve vrste matrica je **GMRES**. *GMRES metoda* (Generalized minimal residual algorithm) ili generalizirani algoritam minimalnog reziduala koristi modificirani Gram–Schmidtov postupak kako bi konstruirao ortonormiranu bazu za niz Krylovljevih potprostora $\text{span}\{r_0, Ar_0, \ldots, A^k r_0\}$. Kada se modificirani Gram–Schmidtov postupak primijeni na ovakav prostor dobiva se *Arnoldijev algoritam.*

Algoritam 1.15 *Arnoldijev algoritam.*
Dan je vektor q_1 *sa* $\|q_1\|_2 = 1$*;*
for $j = 1, 2, \ldots, n-1$
  $\tilde{q}_{j+1} = Aq_j$*;*
  for $i = 1, \ldots, j$
    $h_{i,j} = q_i^T \tilde{q}_{j+1}$*;*
    $\tilde{q}_{j+1} := \tilde{q}_{j+1} - h_{i,j} q_i$*;*
  end

$h_{j+1,j} = \|\tilde{q}_{j+1}\|_2$;
$q_{j+1} = \dfrac{\tilde{q}_{j+1}}{h_{j+1,j}}$;
end

U svakom koraku se minimizira 2–norma reziduala, pa je konačan izgled GMRES algoritma sljedeći.

Algoritam 1.16 *Rješavanje linearnog sustava pomoću GMRES metode.*
x_0 *fiksiran;*
$r_0 = b - Ax_0$;
$\beta = \|r_0\|_2$;
$q_1 = \dfrac{r_0}{\beta}$;
$l = [1, 0, \ldots, 0]^T$;
for $k = 1, 2, \ldots$
Izračunaj q_{k+1} *i* $h_{i,k} = H(i,k)$ *za* $i = 1, \ldots, k+1$, *koristeći Arnoldijev algoritam.*
Primijeni $F_1, \ldots, F_{k-1}$ *na zadnji stupac od* H, *odnosno:*
for $i = 1, \ldots, k-1$
$\begin{bmatrix} H(i,k) \\ H(i+1,k) \end{bmatrix} := \begin{bmatrix} c_i & s_i \\ -\bar{s}_i & c_i \end{bmatrix} \begin{bmatrix} H(i,k) \\ H(i+1,k) \end{bmatrix}$;
Izračunaj k*-tu Givensovu rotaciju* F_k *kako bi se poništio* $(k+1,k)$ *element od* H*:*
$c_k = \dfrac{|H(k,k)|}{\sqrt{|H(k,k)|^2 + |H(k+1,k)|^2}}$;
if $c_k \neq 0$
$s_k = c_k \dfrac{\overline{H(k+1,k)}}{\overline{H(k,k)}}$;
else
$s_k = 1$;
end
Primijeni k*-tu rotaciju na* l *i na zadnji stupac od* H*:*
$\begin{bmatrix} l(k) \\ l(k+1) \end{bmatrix} := \begin{bmatrix} c_k & s_k \\ -\bar{s}_k & c_k \end{bmatrix} \begin{bmatrix} l(k) \\ 0 \end{bmatrix}$;
$H(k,k) := c_k H(k,k) + s_k H(k+1,k)$;
$H(k+1,k) = 0$;
end
if *ocjena norme reziduala* $\beta|l(k+1)|$ *dovoljno mala*
Riješi gornje trokutasti sustav $H_{k\times k} y_k = \beta l_{k\times 1}$.
$x_k = x_0 + Q_k y_k$;
end

Neka svojstva GMRES metode:

- Rezidual u k-tom koraku zadovoljava

$$r_k = b - Ax_k = Q_{k+1}(\beta\xi_1 - H_{k+1,k}y_k) = Q_{k+1}F^{(k)*}((g^{(k)})_{k+1}\xi_{k+1}), \quad (35)$$

pri čemu su ξ_1 prvi jedinični vektor $\xi_1 = [1, 0, \ldots, 0]^T$, $F^{(k)} = F_k F_{k-1} \cdots F_1$, a $g^{(k)} = \beta F^{(k)}\xi_1$. Kao rezultat dobivamo

$$\|r_k\|_2 = \|b - Ax_k\|_2 = |(g^{(k)})_{k+1}|. \quad (36)$$

- Neka je A regularna matrica. Tada se GMRES algoritam prekida u k-tom koraku ($h_{k+1,k} = 0$) ako i samo ako je aproksimacija x_k jednaka egzaktnom rješenju.

O konvergenciji GMRES metode govore sljedeće tvrdnje.

- Neka je x_k aproksimacija rješenja ostvarena u k-tom koraku GMRES algoritma, i neka je $r_k = b - Ax_k$. Tada postoji $q_{k-1} \in \mathbb{P}_{k-1}$ takav da je x_k oblika

$$x_k = x_0 + q_{k-1}(A)r_0$$

i

$$\|r_k\|_2 = \min_{q_{k-1}\in\mathbb{P}_{k-1}} \|(I - Aq_{k-1}(A))r_0\|_2.$$

- Neka je dan nerastući niz $f(0) \geq f(1) \geq \cdots \geq f(n-1) > 0$ pozitivnih brojeva i skup kompleksnih brojeva različitih od nule $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, tada postoji matrica A sa svojstvenim vrijednostima $\lambda_1, \lambda_2, \ldots, \lambda_n$ i desna strana sustava b sa $\|b\|_2 = f(0)$ takvi da reziduali r_k u svakom koraku GMRES algoritma primijenjenog na sustav $Ax = b$ sa $x_0 = 0$, zadovoljavaju $\|r_k\|_2 = f(k)$, $k = 1, 2, \ldots, n-1$.

Primjer 1.14 *U ovom primjeru promatramo samo GMRES metodu, i pokazat ćemo da se zaista može konstruirati matrica sa u naprijed određenom krivuljom konvergencije, za bilo koji skup svojstvenih vrijednosti. Definirat ćemo najprije skup svojstvenih vrijednosti* $\lambda(A) \in \{-50, -49, \ldots, -1, 2, 4, \ldots, 100\}$, *i niz vrijednosti* $f(0) = 100$, $f(1) = 99$, $f(2) = 98, \ldots, f(99) = 1$, $f(100) = 0$. *Nadalje, definirajmo*

$$g(k) = \sqrt{(f(k-1))^2 - (f(k))^2} = \sqrt{(100-k+1)^2 - (100-k)^2}, \quad k = 1, \ldots, 100.$$

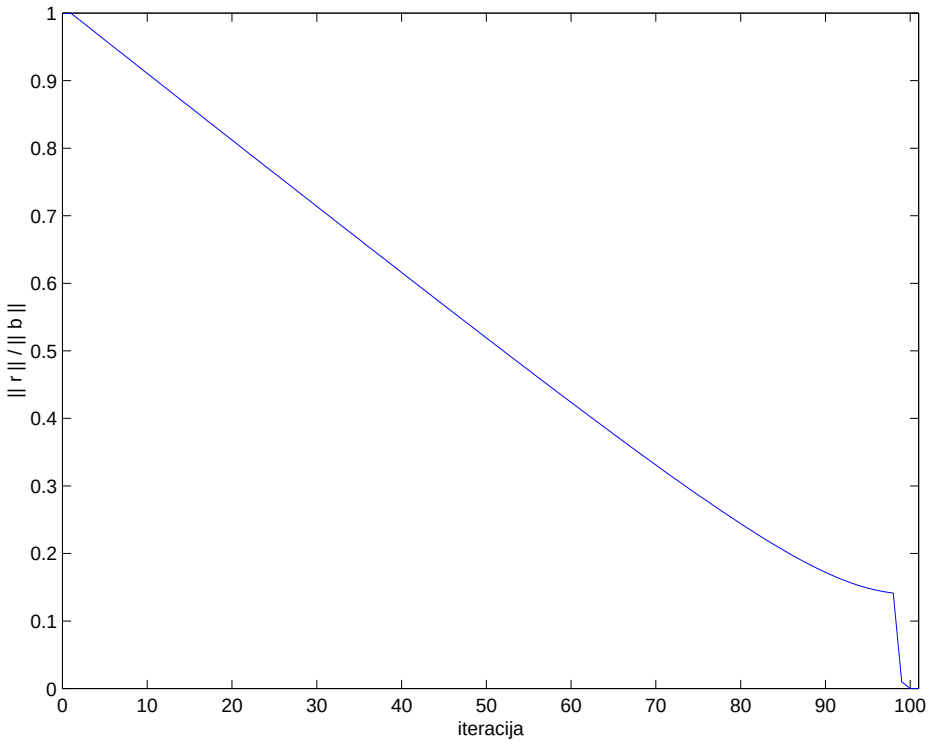
Matrica $V = [v_1\ v_2\ \cdots\ v_{100}]$ *je slučajna ortogonalna matrica, i ako definiramo* $b = \sum_{i=1}^{100} g(i)v_i$, *tada je* $\|b\|_2 = f(0)$. *U nastavku, definirajmo još matricu* $B = [b\ v_1\ \cdots\ v_{99}]$ *i izračunajmo koeficijente polinoma*

$$a(z) = z^{100} - \sum_{i=0}^{99} \alpha_i z^i = (z - \lambda_1(A)) \cdots (z - \lambda_{100}(A)),$$

i na kraju konstruirajmo matricu A kao

$$A = B \cdot \begin{bmatrix} 0 & \cdots & 0 & \alpha_0 \\ 1 & \cdots & 0 & \alpha_1 \\ & \ddots & \vdots & \vdots \\ & & 1 & \alpha_{99} \end{bmatrix} \cdot B^{-1}.$$

Početna iteracija je $x_0 = 0$, i $tol = 10^{-5}$. Prema teoretskim rezultatima, ovako definirani sustav $Ax = b$ ima matricu sa gore definiranim svojstvenim vrijednostima i sa rezidualima, takvim da nakon svake iteracije GMRES metode vrijedi $\|r_k\|_2 = f(k)$. Dobiveni rezultati u Slici 15 to potvrđuju, do na numeričke greške.



Slika 15: Relativne norme reziduala u svakoj iteraciji GMRES metode, za matricu A iz primjera 1.14.

1.2.8 Ubrzanje konvergencije. Pojam prekondicioniranja

Najjednostavnije, prekondicioniranje možemo opisati kao bilo kakvo modificiranje originalnog linearnog sustava, koje na neki način olakšava rješavanje

danog sustava. Konvergencija iterativnih metoda ovisi o nekim svojstvima matrice sustava, pri čemu se najčešće radi o svojstvima spektra ili singularnih vrijednosti. Zato je cilj takve modifikacije transformirati linearni sustav u ekvivalentan sustav koji ima isto rješenje, ali koji ima bolja svojstva, npr. bolja spektralna svojstva matrice sustava. Matrica prekondicioniranja je matrica koja utječe na transformaciju sustava. Na primijer, ako matrica M aproksimira matricu sustava A na neki način, transformirani sustav

$$M^{-1}Ax = M^{-1}b \tag{37}$$

ima isto rješenje kao i originalni sustav $Ax = b$, ali svojstva matrice $M^{-1}A$ mogu biti bolja.

U slučaju da je matrica A simetrična, tada se od matrice M može zahtijevati da bude simetrična i pozitivno definitna. Matrica M se može onda faktorizirati npr. metodom Choleskog na $M = LL^T$, jer onda možemo definirati i *dvostrano* prekondicioniranje pomoću faktora, pri čemu se tada rješava sustav

$$L^{-1}AL^{-T}y = L^{-1}b, \qquad x = L^{-T}y. \tag{38}$$

pa matrica prekondicioniranog sustava $L^{-1}AL^{-T}$ ostaje simetrična. U ovom slučaju treba napomenuti da, kod onih iterativnih metoda kod kojih konvergencija ovisi o spektru matrice sustava, matrice $M^{-1}A$ i $L^{-1}AL^{-T}$ imaju iste svojstvene vrijednosti, jer su slične.

Uvjeti odabira matrice prekondicioniranja M su sljedeći:

- Rješavanje prekondicioniranog sustava iterativnom metodom je brže, u smislu da će metoda zahtijevati manje iteracija do postizanja konvergencije od rješavanja originalnog sustava ($M^{-1}A \approx I$).

- Za metodu konjugiranih gradijenata primijenjenu na simetričnu pozitivno definitni problem, težnja nam je imati simetričnu prekondicioniranu matricu $L^{-1}AL^{-T}$ sa brojem uvjetovanosti blizu jedan. S druge strane, znamo da za konvergenciju metode konjugiranih gradijenata značajnu ulogu igra i distribucija svojstvenih vrijednosti. U tom smislu mi možemo zahtijevati da prekondicionirana matrica ima, na primjer, samo nekoliko velikih svojstvenih vrijednosti, a ostatak gusto nakupljnih oko jedne točke, ili da prekondicionirana matrica ima samo nekoliko različitih svojstvenih vrijednosti. U svakom slučaju svojstvene vrijednosti bi trebale biti distribuirane tako da polinom malog stupnja, koji je jednak jedinici u ishodištu, bude mali u svim svojstvenim vrijednostima.

Još je jedan važan rezultat dao Van der Sluis, a tiče se **prekondicioniranja dijagonalnom matricom**.

- Ako simetrična pozitivno definitna matrica A ima sve dijagonalne elemente jednake 1, tada

$$\kappa(A) \leq m \cdot \min_{D \in \mathcal{D}} \kappa(DAD), \tag{39}$$

gdje je $\mathcal{D} = \{\text{pozitivno definitne dijagonalne matrice}\}$, a m je maksimalan broj elemenata, različitih od nula, u bilo kojem retku od A.

Prekondicionirani sustav je tada oblika

$$DADy = Db, \qquad x = Dy.$$

Napomena 1.19 *Van der Sluisov teorem tvrdi da je za bilo koju simetričnu pozitivno definitnu $n \times n$ matricu $A = [a_{ij}]$,* **dijagonalna matrica prekondicioniranja** *dana sa $D = diag(\sqrt{a_{11}^{-1}}, \ldots, \sqrt{a_{nn}^{-1}})$ optimalna u određenom smislu.*

Drugi način odabira matrice prekondicioniranja su **nekompletne faktorizacije**. Nekompletnima smatramo one faktorizacije, kojima se tokom samog procesa faktorizacije određeni elementi zanemaruju. To na primjer mogu biti netrivijalni elementi u faktorizaciji na pozicijama u kojima originalna matrica sustava ima nulu. Takve se matrice prekondicioniranja tada uvijek ostavljaju u faktoriziranom obliku. Njihova efikasnost onda ovisi o tome kako dobro njihov inverz aproksimira A^{-1}.

U slučaju simetrične pozitivno definitne matrice sustava A, za prekondicioniranje se može upotrijebiti *nekompletna faktorizacije Choleskog (IC).* Njegov algoritam je vrlo sličan originalnoj faktorizaciji Choleskog.

Algoritam 1.17 *Računanje nekompletne faktorizacije Choleskog (IC) simetrične pozitivno definitne matrice $A \in \mathbb{R}^{n \times n}$.*

za $i = 1, \ldots, n$ {

$r_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2}$;

/* za $i = 1$, $r_{ii} = \sqrt{a_{ii}}$ */

za $j = i+1, \ldots, n$ {

ako $a_{ij} \neq 0$ {

$r_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj} \right) \Big/ r_{ii}$; }

inače {

$r_{ij} = 0$ } } }

Ako definiramo skup $\mathcal{P}$ kao podskup indeksa $\{(i,j) : \ j \neq i, \ i,j = 1,\ldots,n\}$, takvih da je $a_{ij} = 0$, tada matrica R dobivena iz nekompletne faktorizacije Choleskog ima istu strukturu nula kao i početna matrica A u gornjem trokutu, tj. $r_{ij} = 0$ za $(i,j) \in \mathcal{P}$. Za određenu klasu matrica može se pokazati da je $R^T R \approx A$. Prekondicionirani sustav tada glasi

$$R^{-T} A R^{-1} y = R^{-T} b, \qquad x = R^{-1} y.$$

Pogledajmo sada kako bi rješavali prekondicionirani pozitivno definitni sustav (38) pomoću metode konjugiranih gradijenata. Jedna mogućnost je jednostavno na njega primijeniti algoritam 1.14, ali tada ćemo trebati eksplicitno izračunati faktor L, i rješavati sustave sa matricama L i L^T. S druge strane, algoritam možemo prepraviti tako da se rješavaju sustavi samo sa $M = LL^T$. To možemo učiniti na sljedeći način. Prvo označimo sve veličine vezane uz prekondicionirani sustav (38) sa $\hat{}$, a veličine vezane uz početni sustav $Ax = b$ sa standardnim oznakama, tada uz pomoć relacija

$$x_k = L^{-T} \hat{x}_k, \qquad r_k = L \hat{r}_k,$$

$$d_k = L^{-T} \hat{d}_k, \qquad M = LL^T,$$

CG metodu primijenjenu na sustav (38) možemo transformirati u algoritam, koji ne ovisi o faktorizaciji matrice prekondicioniranja M. Imamo:

$$\begin{aligned}
\alpha_{k-1} &= \frac{\hat{r}_{k-1}^T \hat{r}_{k-1}}{\hat{d}_{k-1}^T L^{-1} A L^{-T} \hat{d}_{k-1}} = \frac{r_{k-1}^T L^{-T} L^{-1} r_{k-1}}{d_{k-1}^T L L^{-1} A L^{-T} L^T d_{k-1}} = \frac{r_{k-1}^T M^{-1} r_{k-1}}{d_{k-1}^T A d_{k-1}} \\
x_k &= L^{-T} \hat{x}_k = L^{-T} \hat{x}_{k-1} + \alpha_{k-1} L^{-T} \hat{d}_{k-1} = x_{k-1} + \alpha_{k-1} d_{k-1} \\
r_k &= L \hat{r}_k = L \hat{r}_{k-1} - \alpha_{k-1} L L^{-1} A L^{-T} \hat{d}_{k-1} = r_{k-1} - \alpha_{k-1} A d_{k-1} \\
\beta_k &= \frac{\hat{r}_k^T \hat{r}_k}{\hat{r}_{k-1}^T \hat{r}_{k-1}} = \frac{r_k^T L^{-T} L^{-1} r_k}{r_{k-1}^T L^{-T} L^{-1} r_{k-1}} = \frac{r_k^T M^{-1} r_k}{r_{k-1}^T M^{-1} r_{k-1}} \\
d_k &= L^{-T} \hat{d}_k = L^{-T} \hat{r}_k + \beta_k L^{-T} \hat{d}_{k-1} = L^{-T} L^{-1} r_k + \beta_k d_{k-1} = M^{-1} r_k + \beta_k d_{k-1}
\end{aligned}$$

Zato konačno dobivamo algoritam za prekondicionirane konjugirane gradijente.

Algoritam 1.18 *Rješavanje linearnog sustava pomoću prekondicionirane metode konjugiranih gradijenata.*

x_0 *fiksiran;*
$r_0 = b - Ax_0$*;*
riješi $Mp_0 = r_0$*;*
$d_0 = p_0$*;*

for k=0,1,2,...

$\alpha_k = \frac{r_k^T p_k}{d_k^T A d_k}$;

$x_{k+1} = x_k + \alpha_k d_k$;

$r_{k+1} = r_k - \alpha_k A d_k$;

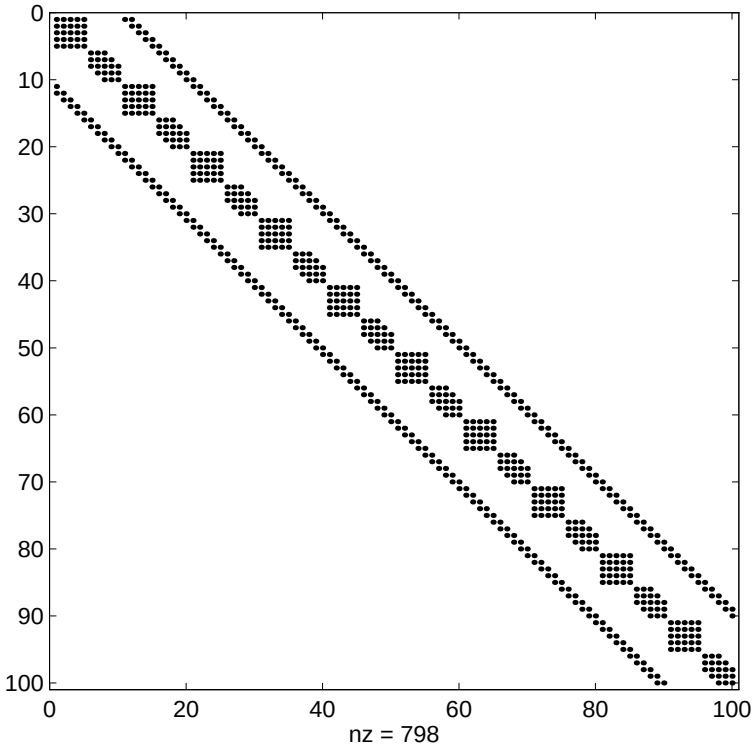
riješi $M p_{k+1} = r_{k+1}$;

$\beta_{k+1} = \frac{r_{k+1}^T p_{k+1}}{r_k^T p_k}$;

$d_{k+1} = p_{k+1} + \beta_{k+1} d_k$;

end

Primjer 1.15 *Matrica sustava ovog primjera je rijetko popunjena Stieltjesova matrica, čiji raspored netrivijalnih elemenata je dan u Slici 16. Svojstvene*



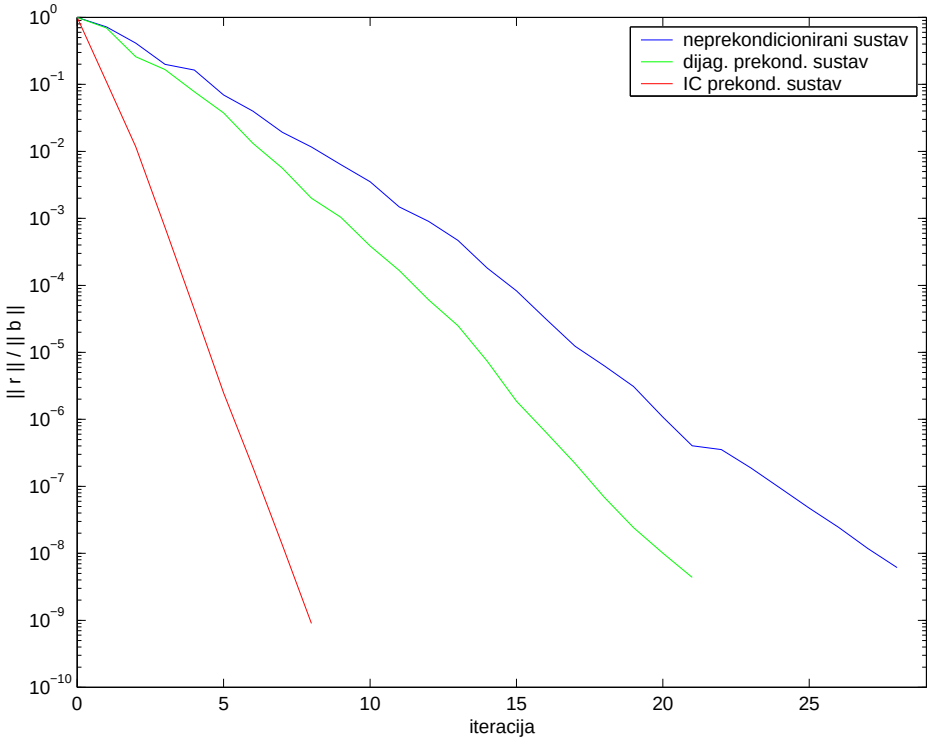
Slika 16: Raspored netrivijalnih elemenata matrice sustava A iz primjera 1.15.

vrijednosti ove matrice nalaze se u intervalu $\lambda(A) \in \langle 3.23, 47.07 \rangle$*, i mnoge su vrlo blizu jedne drugima, a uvjetovanost iznosi* $\kappa(A) = 14.5627$.

Ova matrica je pogodna za prekondicioniranje sa nekompletnom faktorizacijom Choleskog. Takvo prekondicioniranje ćemo usporediti sa dijago-

nalnim prekondicioniranjem, i sa originalnim sustavom bez prekondicioniranja, kada se rješavaju pomoću metode konjugiranih gradijenata. Vektor desne strane b je izračunat tako da je egzaktno rješenje $x = [1, 1, \ldots, 1]^T$*, a* $x_0 = [0, 0 \ldots, 0]$*. Odnosi između uvjetovanosti i broja iteracija potrebnih za postizanje iste točnosti od* $tol = 10^{-8}$ *tih triju sustava, dani su u sljedećoj tablici.*

	neprekondicionirani sustav	**dijagonalno prekondicionirani sustav**	**IC prekondicionirani sustav**
$\kappa(A)$	14.5627	7.8162	1.5025
k	28	21	8



Slika 17: Relativne norme reziduala u svakoj iteraciji prekondicionirane metode konjugiranih gradijenata, za matricu A iz primjera 1.15.

1.3 Primjer numeričkog rješavanja obične diferencijalne jednadžbe

Na kraju poglavlja o rješavanju sustava linearnih jednadžbi, numerički ćemo riješiti jednu konkretnu običnu diferencijalnu jednadžbu (rubni problem), i usporediti dobiveno rješenje sa egzaktnim rješenjem jednadžbe. Dakle, imamo sljedeći problem:

$$\begin{aligned} -\frac{d^2}{dx^2}y(x) - y(x) &= 2\sin(x), \quad x = \langle 0, 1\rangle, \\ y(0) &= 0 \\ y(1) &= \cos(1). \end{aligned} \tag{40}$$

Lako se može provjeriti da je egzaktno rješenje dano sa

$$y(x) = x\cos(x).$$

Njega ćemo diskretizirati konačnim diferencijama na segmenu $[0, 1]$, sa mrežom od 101 točaka, pri čemu je

$$h = 0.01, \quad x_i = ih, \quad y_i = y(x_i), \quad f_i = 2\sin x_i, \quad i = 0, 1, \ldots, 100.$$

Tada imamo:

$$\begin{aligned} y_0 &= 0 \\ \frac{-y_{i-1} + 2y_i - y_{i+1}}{h^2} - y_i &= f_i, \quad i = 1, \ldots, 99 \\ y_{100} &= \cos(1). \end{aligned}$$

Kad to sredimo dobit ćemo

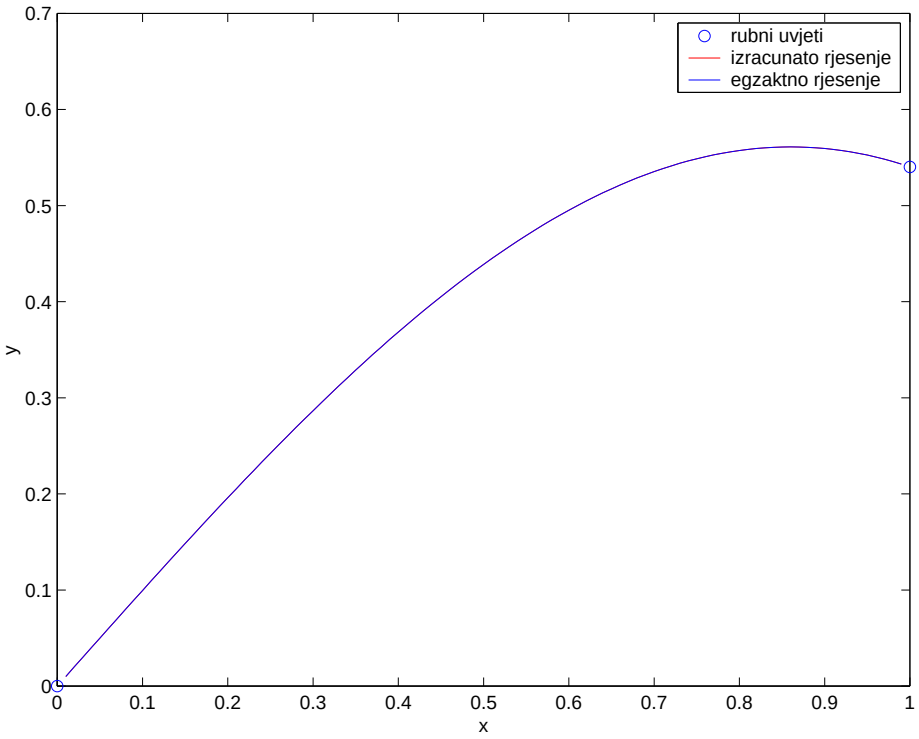
$$\begin{aligned} (2 - h^2)y_1 - y_2 &= h^2 f_1 \\ -y_{i-1} + (2 - h^2)y_i - y_{i+1} &= h^2 f_i, \quad i = 2, \ldots, 98 \\ -y_{98} + (2 - h^2)y_{99} &= h^2 f_{99} + y_{100}, \end{aligned}$$

odnosno dobit ćemo sustav $Ay = b$, pri čemu su

$$A = \begin{bmatrix} 1.9999 & -1 & & & & \\ -1 & 1.9999 & -1 & & & \\ & & \cdots & \cdots & \cdots & \\ & & & -1 & 1.9999 & -1 \\ & & & & -1 & 1.9999 \end{bmatrix}, A \in \mathbb{R}^{99\times 99}$$

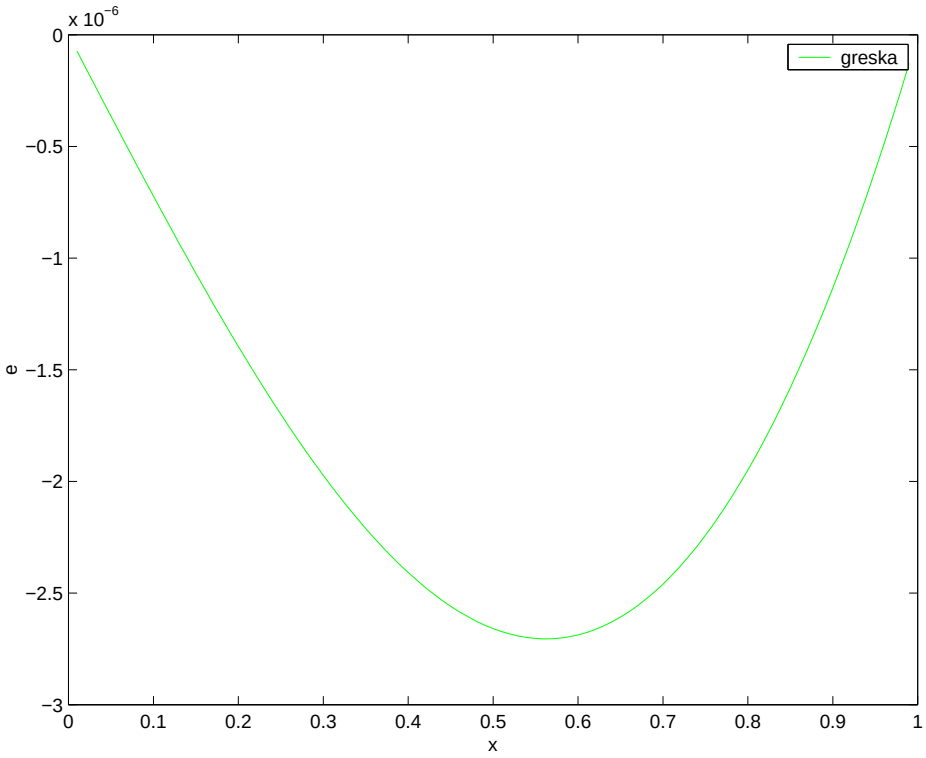
$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{98} \\ y_{99} \end{bmatrix}, \quad b = \begin{bmatrix} 0.0002\sin(0.01) \\ 0.0002\sin(0.02) \\ \vdots \\ 0.0002\sin(0.98) \\ 0.0002\sin(0.99) + \cos(1) \end{bmatrix} \in \mathbb{R}^{99}.$$

Lako se može provjeriti da je matrica pozitivno definitna (npr. u *Octave*-i izračunati njezinu faktorizaciju Choleskog). Njezina uvjetovanost je $\kappa_2(A) = 4.5090 \cdot 10^3$, što nije tako loše, pa sustav možemo riješiti pomoću faktorizacije Choleskog i metodom konjugiranih gradijenata. Izračunata rješenja se razlikuju najviše za $3.3307 \cdot 10^{-15}$ u svakoj komponenti, tako da je skoro sve jedno koju od tih dviju metoda koristimo. Uspoređivat ćemo egzaktno rješenje rubnog problema y sa rješenjem koje je dala metoda konjugiranih gradijenata kod rješavanja sustava dobivenog diskretizacijom $\tilde{y}$.



Slika 18: Numeričko i egzaktno rješenje rubnog problema (40).

Kao što vidimo iz slike 18, izračunato rješenje je prilično dobra aproksimacija egzaktnog rješenja. Na slici 19 je prikazana greška, u kojoj prevladava greška diskretizacije. Greška nastala zbog primjene metode konjugiranih gradijenata u aritmetici konačne preciznosti u ovom slučaju je zanemariva.



Slika 19: Greška $(y - \tilde{y})$ između egzaktnog i numeričkog rješenje rubnog problema (40) .

2 Problem najmanjih kvadrata

2.1 Opis problema

Pretpostavimo da imamo skup mjerenih podataka (t_k, y_k), $k = 1, \ldots, m$, i želimo taj model aproksimirati funkcijom oblika $\varphi(t)$. Ako je $\varphi(t)$ linearna, tj. ako je

$$\varphi(t) = x_1\varphi_1(t) + \cdots + x_n\varphi_n(t),$$

onda bismo željeli pronaći parametre x_j tako da mjereni podaci (t_k, y_k) zadovoljavaju

$$y_k = \sum_{j=1}^{n} x_j\varphi_j(t_k), \quad k = 1, \ldots, m.$$

Ako označimo

$$a_{kj} = \varphi_j(t_k), \quad b_k = y_k, \tag{41}$$

onda prethodne jednadžbe možemo u matričnom obliku pisati kao

$$Ax = b,$$

pri čemu je $A = [a_{ij}] \in \mathbb{R}^{m\times n}$ i $b = [b_i] \in \mathbb{R}^m$. Ako je mjerenih podataka više nego parametara, tj. ako je $m > n$, onda ovaj sustav jednadžbi ima više jednadžbi nego nepoznanica, pa je *preodređen*.

Postoji mnogo načina da se odredi "najbolje" rješenje,

- zbog statističkih razloga to je često metoda najmanjih kvadrata.
- Funkcija φ određuje se iz uvjeta da euklidska norma (norma 2) vektora pogrešaka u čvorovima aproksimacije bude najmanja moguća, tj. tako da minimiziramo S,

$$S = \sum_{k=0}^{m}(y_k - \varphi(t_k))^2 \to \min.$$

- tj. određujemo x tako da minimizira rezidual $r = Ax - b$

$$\min_x \|r\|_2 = \min_x \|Ax - b\|_2, \quad A \in \mathbb{R}^{m\times n}, \quad b \in \mathbb{R}^m, \quad x \in \mathbb{R}^n. \tag{42}$$

- Ako je $\text{rang}(A) < n$, onda rješenje x ovog problema očito **nije** jedinstveno, jer mu možemo dodati bilo koji vektor iz nul-potprostora od A, a da se rezidual ne promijeni.

- Među svim rješenjima x problema najmanjih kvadrata uvijek postoji jedinstveno rješenje x najmanje norme, tj. koje još minimizira i $\|x\|_2$.

Iz geometrijske interpretacije problema najmanjih kvadrata odmah vidimo da je za rješenje x, Ax ortogonalna projekcija vektora b na $\mathcal{R}(A)$. To se lako može provjeriti ako definiramo diferencijabilnu funkciju

$$\phi(x) = \frac{1}{2}\|Ax - b\|_2^2,$$

i izjednačimo $\nabla\phi(x) = 0$ (ekvivalentno traženju minimuma $\min_x \|Ax - b\|_2$). Tada možemo raspisati $\phi(x)$ kao

$$\begin{aligned}
\phi(x) &= \frac{1}{2}(Ax-b)^T(Ax-b) = \frac{1}{2}x^TA^TAx - x^TA^Tb + \frac{1}{2}b^Tb = \\
&= \frac{1}{2}\sum_{i,j=1}^{n} x_i(A^TA)_{ij}x_j - \sum_{i=1}^{n} x_i(A^Tb)_i + \frac{1}{2}b^Tb = \\
&= \frac{1}{2}\sum_{i}^{n}(A^TA)_{ii}x_i^2 + \frac{1}{2}\sum_{i\neq j}(A^TA)_{ij}x_ix_j - \sum_{i=1}^{n}(A^Tb)_ix_i + \frac{1}{2}b^Tb.
\end{aligned}$$

Izračunajmo sada k-tu parcijalnu derivaciju od $\phi(x)$ i izjednačimo ju sa nulom.

$$\begin{aligned}
\frac{\partial}{\partial x_k}\phi(x) &= (A^TA)_{kk}x_k + \frac{1}{2}\sum_{j\neq k}(A^TA)_{kj}x_j + \frac{1}{2}\sum_{i\neq k}(A^TA)_{ik}x_i - (A^Tb)_k = \\
&= \overset{\{(A^TA)_{ji}=(A^TA)_{ij}\}}{\hookrightarrow} \sum_{i=1}^{n}(A^TA)_{ki}x_i - (A^Tb)_k = (A^TAx - A^Tb)_k.
\end{aligned}$$

Dakle,

$$\nabla\phi(x) = A^TAx - A^Tb,$$

a iz $\nabla\phi(x) = 0$ slijedi

$$A^T(Ax - b) = A^Tr = 0. \tag{43}$$

Da se zaista radi o minimumu, provjerimo Hessian.

$$\frac{\partial^2}{\partial x_l \partial x_k}\phi(x) = (A^TA)_{kl},$$

što znači da je

$$H\phi = A^TA$$

i on je

- pozitivno definitan u slučaju da je matrica A punog stupčanog ranga, pa tada postoji jedinstveni minimum, i on je rješenje sustava $A^TAx = A^Tb$,

- pozitivno semidefinitan u slučaju da matrica A nema puni stupčani rang, pa se tada minimum postiže na čitavom afinom potprostu. To možemo provjeriti na sljedeći način. Neka je x rješenje problema najmanjih kvadrata, i neka je i $x+z$ također rješenje istog problema. Tada x i $x+z$ moraju zadovoljavati (43), pa imamo

$$0 = A^T[A(x+z) - b] = A^T(Ax - b) + A^TAz = A^TAz.$$

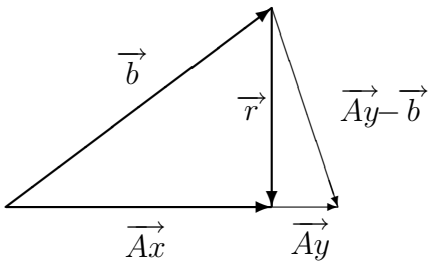
  Ako gornju jednakost skalarno pomnožimo sa z, dobit ćemo da je $\|Az\|_2 = 0$, odakle slijedi da je $Az = 0$ odnosno $z \in \mathcal{N}(A)$ (z je u jezgri od A). Dakle skup rješenja u ovom slučaju čini skup

$$x + \mathcal{N}(A).$$

Na kraju, ako sa $\overrightarrow{b}$, $\overrightarrow{Ax}$ i $\overrightarrow{r}$ označimo vektore u vektorskom prostoru $\mathbb{R}^m$, pri čemu je x je rješenje problema najmanjih kvadrata, tada imamo da je $\overrightarrow{b} = \overrightarrow{Ax} - \overrightarrow{r}$, a zbog (43) je $(Ay)^Tr = 0$ za svaki $y \in \mathbb{R}^n$, odnosno

$$\overrightarrow{r} \perp \mathcal{R}(A).$$

Na kraju možemo zaključiti da je $\overrightarrow{Ax}$ dobiven iz $\overrightarrow{b}$, tako što mu se oduzela komponenta okomita na $\mathcal{R}(A)$, pa je $\overrightarrow{Ax}$ zaista ortogonalna projekcija od $\overrightarrow{b}$ na $\mathcal{R}(A)$.



Slika 20: Okomitost reziduala rješenja x problema $\|Ax - b\|_2 \to \min$ na $\mathcal{R}(A)$.

Matrični problem najmanjih kvadrata može se riješiti na više načina, koji uključuju neke istaknute faktorizacije matrice A.

2.2 QR faktorizacija

Definiciju i egzistenciju QR faktorizacije daje sljedeći teorem.

Teorem 2.1 (QR dekompozicija)

- *Neka je $A \in \mathbb{C}^{m\times n}$, uz $m \geq n$. Tada postoji unitarna matrica $Q \in \mathbb{C}^{m\times m}$ takva da je*

$$Q^* A = R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix},$$

  gdje je $R \in \mathbb{C}^{m\times n}$, a $R_1 \in \mathbb{C}^{n\times n}$ gornjetrokutasta matrica s nenegativnim dijagonalnim elementima.

- *Neka je $A \in \mathbb{R}^{m\times n}$, uz $m \geq n$. Tada postoji ortogonalna matrica $Q \in \mathbb{R}^{m\times m}$ takva da je*

$$Q^T A = R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix},$$

  gdje je $R \in \mathbb{R}^{m\times n}$, a $R_1 \in \mathbb{R}^{n\times n}$ gornjetrokutasta matrica s nenegativnim dijagonalnim elementima.

U oba slučaja je $A = QR$.

Napomena 2.1 *Ako izvršimo particiju matrica*

$$Q = \underset{n}{[\,Q_1} \quad \underset{m-n}{Q_2\,]}$$

onda iz Teorema 2.1 slijedi

$$A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1.$$

Dakle, QR dekompoziciju možemo napisati u skraćenom obliku

$$A = Q_1 R_1,$$

pri čemu je $Q_1 \in \mathbb{C}^{m\times n}$ ortonormirana matrica, a $R_1 \in \mathbb{C}^{n\times n}$ gornjetrokutasta matrica s nenegativnim dijagonalnim elementima.

Neka svojstva QR faktorizacije dana su u sljedećim tvrdnjama.

- Ako je $A \in \mathbb{C}^{n\times n}$ regularna matrica, tada je matrica Q jedinstvena.

- Neka je $A \in \mathbb{C}^{n\times n}$ matrica sa stupcima a_i, $i = 1, \ldots, n$. Tada je

$$|\det A| \leq \prod_{i=1}^{n} \|a_i\|_2,$$

pri čemu se jednakost postiže ako i samo ako su stupci od A međusobno ortogonalni. (**Hadamardova nejednakost**)

QR faktorizaciju možemo izračunati na više načina. Mi ćemo obraditi dva načina računanja, kod kojih se ortogonalna matrica Q dobiva uzastopnim množenjem elementarnih ortogonalnih matrica, kao što su: reflektori ili rotacije.

2.2.1 QR faktorizacija pomoću Householderovih reflektora

Za zadani vektor $a \in \mathbb{R}^m$, $a \neq 0$, tražimo ortogonalnu matricu $H \in \mathbb{R}^{m\times m}$ takvu da je

$$Ha = -\alpha e, \text{ gdje je } e \in \mathbb{R}^m, \ \|e\|_2 = 1 \text{ zadani vektor.}$$

(Za $a = 0$ je $H = I$ i nužno je $\alpha = 0$.) Za H zahtijevamo da je oblika

$$H = I - \frac{1}{\gamma} vv^T, \qquad \text{gdje je } \gamma > 0, \ v \neq 0.$$

Matrica H je **Householderov reflektor**.

- Lako se vidi da je H simetrična matrica, tj. $H^T = H$.
- Za $\gamma = \frac{\|v\|_2^2}{2}$ je

$$\begin{aligned} H^T H &= H^2 = \left(I - \frac{2}{\|v\|_2^2} vv^T\right)\left(I - \frac{2}{\|v\|_2^2} vv^T\right) = \\ &= I - \frac{2}{\|v\|_2^2} vv^T - \frac{2}{\|v\|_2^2} vv^T + \frac{4}{\|v\|_2^4} vv^T vv^T = \\ &= I - \frac{4}{\|v\|_2^2} vv^T + \frac{4}{\|v\|_2^2} vv^T = I \end{aligned}$$

i H je ortogonalna matrica.

Zbog ortogonalnosti od H mora biti

$$\|a\|_2 = \|Ha\|_2 = \| - \alpha e\|_2 = |-\alpha| \|e\|_2 = |\alpha|.$$

Ako definiramo

$$\alpha = \begin{cases} \|a\|_2, & e^T a \geq 0 \\ -\|a\|_2, & e^T a < 0 \end{cases}$$

(Predznak se bira zbog stabilnosti metode, da izbjegnemo fatalno kraćenje.)

$$\begin{aligned} v &= a + \alpha e \\ \gamma &= \frac{\|a\|_2^2 + 2\alpha e^T a + \alpha^2}{2} = \frac{2\|a\|_2^2 + 2\|a\|_2 |e^T a|}{2} = \\ &= \|a\|_2(\|a\|_2 + |e^T a|) \end{aligned}$$

Tada vrijedi

$$\begin{aligned} Ha &= \left(I - \frac{1}{\gamma} v v^T\right) a = a - \frac{(a+\alpha e)^T a}{\|a\|_2(\|a\|_2 + |e^T a|)}(a + \alpha e) = \\ &= a - \frac{a^T a + \alpha e^T a}{\|a\|_2(\|a\|_2 + |e^T a|)}(a + \alpha e) = \\ &= a - \frac{\|a\|_2(\|a\|_2 + |e^T a|)}{\|a\|_2(\|a\|_2 + |e^T a|)}(a + \alpha e) = a - a - \alpha e = \\ &= -\alpha e, \end{aligned}$$

što smo i tražili.

Napomena 2.2 *Da bi računali sa Householderovim reflektorom H uopće ga ne trebamo posebno računati kako bi dobili njegov matrični oblik.*

(a) Za $x \in \mathbb{R}^m$ je:

$$Hx = \left(I - \frac{1}{\gamma} v v^T\right) x = x - \frac{v^T x}{\gamma} v.$$

Dakle, potrebno je izračunati samo skalarni produkt $v^T x$ i $\mu = \frac{v^T x}{\gamma} \in \mathbb{R}$, odakle je

$$Hx = x - \mu v,$$

što je manje operacija nego tražiti matricu H i množiti je vektorom.

(b) Za $A = [\ a_1 \ \cdots \ a_n\ ] \in \mathbb{R}^{m\times n}$ je:

$$HA = H[\ a_1 \ \cdots \ a_n\ ] = [\ Ha_1 \ \cdots \ Ha_n\ ],$$

pri čemu Ha_i računamo kao pod (a). To je opet manje operacija nego što bi bilo da izvedemo puno matrično množenje $H \cdot A$.

(c) Za $A \in \mathbb{R}^{m\times m}$, $A^T = A$ je:

$$\begin{aligned} H^TAH &= HAH = \left(I - \frac{1}{\gamma}vv^T\right) A \left(I - \frac{1}{\gamma}vv^T\right) = \\ &= A - \frac{1}{\gamma}vv^TA - \frac{1}{\gamma}Avv^T + \frac{1}{\gamma^2}vv^TAvv^T. \end{aligned}$$

Ako stavimo da je

$$u = \frac{1}{\gamma}Av \in \mathbb{R}^m, \quad p = \frac{1}{2\gamma}v^Tu \in \mathbb{R}, \quad w = u - pv \in \mathbb{R}^m,$$

tada imamo

$$\begin{aligned} H^TAH &= A - vu^T - uv^T + \frac{1}{\gamma}v\left(\frac{1}{\gamma}Av\right)^T vv^T = \\ &= A - vu^T - uv^T + 2pvv^T = A - v(u^T - pv^T) - (u - pv)v^T = \\ &= A - vw^T - wv^T. \end{aligned}$$

Ovo je manji broj operacija nego direktno množenje.

Householderove reflektore možemo primijeniti na traženje QR faktorizacije matrice A. Radimo direktno nad stupcima matrice, i to od dijagonale na dolje.

Neka je $A = [\ a_1^{(1)} \ \cdots \ a_n^{(1)}\ ] \in \mathbb{R}^{m\times n}$ za $m \geq n$. Ako je $a_1^{(1)} \neq 0$, stavimo li

$$e^{(1)} = e_1 \in \mathbb{R}^m,$$

znamo naći Householderov reflektor H_1 takav da je

$$H_1a_1^{(1)} = -\alpha_1 e^{(1)}.$$

Tada je

$$\begin{aligned} A^{(2)} &= H_1A^{(1)} = [\ H_1a_1^{(1)} \ \cdots \ H_1a_n^{(1)}\ ] = \\ &= \begin{bmatrix} -\alpha_1 & * & * & \cdots & * \\ 0 & & & & \\ \vdots & a_2^{(2)} & a_3^{(2)} & \cdots & a_n^{(2)} \\ 0 & & & & \end{bmatrix}. \end{aligned}$$

Ako je $a_1^{(1)} = 0$, stavimo $H_1 = I$.

Ako je $a_2^{(2)} \neq 0 \in \mathbb{R}^{m-1}$, postoji Householderova matrica $\bar{H}_2 \in \mathbb{R}^{(m-1)\times(m-1)}$ takva da je

$$\bar{H}_2a_2^{(2)} = -\alpha_2 e_1,$$

uz $e_1 \in \mathbb{R}^{m-1}$. Za

$$H_2 = \begin{bmatrix} 1 & 0 \\ 0 & \bar{H}_2 \end{bmatrix}$$

je

$$H_2 A^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & \bar{H}_2 \end{bmatrix} \begin{bmatrix} -\alpha_1 & * & * & \cdots & * \\ 0 & & & & \\ \vdots & a_2^{(2)} & a_3^{(2)} & \cdots & a_n^{(2)} \\ 0 & & & & \end{bmatrix} =$$

$$= \begin{bmatrix} -\alpha_1 & * & * & \cdots & * \\ 0 & -\alpha_2 & & & \\ \vdots & \vdots & \bar{H}_2 a_3^{(2)} & \cdots & \bar{H}_2 a_n^{(2)} \\ 0 & 0 & & & \end{bmatrix} = \begin{bmatrix} -\alpha_1 & * & * & \cdots & * \\ 0 & -\alpha_2 & * & \cdots & * \\ 0 & 0 & & & \\ \vdots & \vdots & a_3^{(3)} & \cdots & a_n^{(3)} \\ 0 & 0 & & & \end{bmatrix}$$

Nastavljamo tako dalje, svaki puta smanjujući dimenziju problema i radeći sa

$$A^{(k)}(k:m, k:n) \quad \text{i} \quad \bar{H}_k \in \mathbb{R}^{(m-k+1)\times(m-k+1)},$$

a $H_k \in \mathbb{R}^{m\times m}$ definiramo sa

$$H_k = \begin{bmatrix} I_{k-1} & 0 \\ 0 & \bar{H}_k \end{bmatrix}.$$

Na kraju imamo

$$H_n H_{n-1} \cdots H_1 A = \begin{bmatrix} -\alpha_1 & * & * & \cdots & * \\ 0 & -\alpha_2 & * & \cdots & * \\ & & \ddots & & \\ & & & \ddots & \\ 0 & 0 & & & -\alpha_n \\ 0 & 0 & & & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & & & 0 \end{bmatrix} = R, \qquad (44)$$

tj. $A = QR$, gdje je $Q = H_1 \cdots H_n$. Želim li u R nenegativnu dijagonalu, u (44) slijeva još pomnožimo matricom

$$H_{n+1} = \operatorname{diag}(-\operatorname{sgn}(\alpha_1), \ldots, -\operatorname{sgn}(\alpha_n), 1, \ldots, 1),$$

koja je ortogonalna.

Na računalu štedimo na memoriji tako da vektore kojima realiziramo Householderove reflektore $\bar{H}_k$ ($v_k \in \mathbb{R}^{m-k+1}$, $k = 1, \ldots, n$) stavljamo u donji trokut od A, dok je u gornjem R, čija dijagonala je u posebnom polju D.

Algoritam 2.1 *Za zadanu matricu $A \in \mathbb{R}^{m\times n}$, algoritam izračunava QR faktorizaciju pomoću Householderovih reflektora. Ortogonalna matrica Q se dobiva kao produkt $Q = H_1 \cdots H_n$, pri čemu je $H_i = I - \frac{1}{\gamma_i} v^{(i)} v^{(i)^T}$, a R je gornjetrokutasta.*

for $j = 1, \ldots, n$
begin
$na = \sqrt{\sum_{i=j}^{m} a_{ij}^2}$;
if $a_{jj} > 0$ **then**
begin
$\alpha = na$;
end
else
begin
$\alpha = -na$;
end
$\gamma_j = na(na + |a_{jj}|)$;
for $i = 1, \ldots, j-1$
begin
$v_i^{(j)} = 0$;
end
$v_j^{(j)} = a_{jj} + \alpha$;
$a_{jj} = -\alpha$;
for $i = j+1, \ldots, m$
begin
$v_i^{(j)} = a_{ij}$;
$a_{ij} = 0$;
end;
for $k = j+1, \ldots, n$
begin
$\mu = \frac{1}{\gamma} \sum_{i=j}^{m} v_i^{(j)} a_{ik}$;
for $i = j, \ldots, m$
begin
$a_{ik} = a_{ik} - \mu v_i^{(j)}$;
end
end;
end;
$R = A$;

Napomena 2.3 *I kod QR faktorizacije se može* **pivotirati** *i to tako da se stupac najveće norme (od dijagonale na dolje $a_k^{(k)}, \ldots, a_n^{(k)}$) dovede na pivotno*

mjesto i njega se poništi ispod dijagonale. To se koristi kad želimo naći rang matrice, jer su dijagonalni elementi matrice R sortirani padajuće po apsolutnim vrijednostima. Imamo

$$H_n(\cdots H_2((H_1(AI_{1,j_1}))I_{2,j_2})\cdots I_{n,j_n}) = R,$$

tj.

$$Q^T AP = R, \quad \Longrightarrow \quad AP = QR.$$

Primjer 2.1 *Pogledajmo sada na konkretnom primjeru računanje QR faktorizacije pomoću Householderovih reflektora. Za primjer ćemo uzeti matricu* $A \in \mathbb{R}^{4\times 3}$

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 4 & 2 \\ -2 & 0 & 3 \\ 6 & -1 & 2 \end{bmatrix}.$$

Za dobivanje matrice Q trebat će nam tri Householderova reflektora, koji će poništiti elemente ispod dijagonale za svaki stupac posebno. Prvi reflektor $H_1 \in \mathbb{R}^{4\times 4}$ *treba transformirati 1. stupac* $A(1:4,1)$*, tako da*

$$H_1 \begin{bmatrix} 1 \\ 2 \\ -2 \\ 6 \end{bmatrix} = \begin{bmatrix} -\alpha_1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Prvo provjerimo

$$e_1^T A(1:4,1) = A(1,1) = 1 > 0$$

odakle slijedi

$$\begin{aligned} \alpha_1 &= \|A(1:4,1)\|_2 = \sqrt{1^2+2^2+(-2)^2+6^2} = \sqrt{45} = 6.7082, \\ \gamma_1 &= 6.7082(6.7082+1) = 51.7082, \end{aligned}$$

i vektor $v^{(1)}$ *iznosi*

$$v^{(1)} = A(1:4,1) + \alpha e_1 = \begin{bmatrix} 1 \\ 2 \\ -2 \\ 6 \end{bmatrix} + \begin{bmatrix} 6.7082 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 7.7082 \\ 2 \\ -2 \\ 6 \end{bmatrix},$$

a reflektor H_1 *je oblika*

$$H_1 = I - \frac{1}{\gamma_1} v^{(1)} v^{(1)T} = \begin{bmatrix} -0.1491 & -0.2981 & 0.2981 & -0.8944 \\ -0.2981 & 0.9226 & 0.0774 & -0.2321 \\ 0.2981 & 0.0774 & 0.9226 & 0.2321 \\ -0.8944 & -0.2321 & 0.2321 & 0.3038 \end{bmatrix}.$$

Dalje, slijedi

$$H_1 A(1:4,1) = \begin{bmatrix} -6.7082 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

$$H_1 A(1:4,2) = \begin{bmatrix} -0.5963 \\ 3.3264 \\ 0.6736 \\ -3.0209 \end{bmatrix},$$

$$H_1 A(1:4,3) = \begin{bmatrix} -1.3416 \\ 1.9114 \\ 3.0886 \\ 1.7341 \end{bmatrix},$$

odnosno

$$A^{(2)} = H_1 A = \begin{bmatrix} -6.7082 & -0.5963 & -1.3416 \\ 0 & 3.3264 & 1.9114 \\ 0 & 0.6736 & 3.0886 \\ 0 & -3.0209 & 1.7341 \end{bmatrix}.$$

Sljedeći korak je određivanje reflektora H_2, oblika

$$H_2 = \begin{bmatrix} 1 & 0 \\ 0 & \bar{H}_2 \end{bmatrix}, \qquad \bar{H}_2 \in \mathbb{R}^{3\times 3}, \quad \bar{H}_2 = I - \frac{1}{\gamma_2} \bar{v}^{(2)} \bar{v}^{(2)T}.$$

On mora transformirati odgovarajući dio 2. stupca $A^{(2)}(2:4,2)$, tako da

$$\bar{H}_2 \begin{bmatrix} 3.3264 \\ 0.6736 \\ -3.0209 \end{bmatrix} = \begin{bmatrix} -\alpha_2 \\ 0 \\ 0 \end{bmatrix}.$$

Dobit ćemo sljedeće vrijednosti

$$\begin{aligned} \alpha_2 &= 4.5436 \\ \gamma_2 &= 35.7581 \end{aligned}$$

$$\bar{v}^{(2)} = \begin{bmatrix} 7.8700 \\ 0.6736 \\ -3.0209 \end{bmatrix},$$

$$H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.7321 & -0.1483 & 0.6649 \\ 0 & -0.1483 & 0.9873 & 0.0569 \\ 0 & 0.6649 & 0.0569 & 0.7448 \end{bmatrix}.$$

H_2 ne mijenja 1. redak, a sva akcija se događa na recima od 2. do 4. djelovanjem matrice $\bar{H}_2$. Još k tome, $\bar{H}_2$ ne mijenja niti odgovarajući dio 1. stupca $A^{(2)}(2:4,1)$ jer je on jednak nul-vektoru.

$$\bar{H}_2 A^{(2)}(2:4,2) = \begin{bmatrix} -4.5436 \\ 0 \\ 0 \end{bmatrix},$$

$$\bar{H}_2 A^{(2)}(2:4,3) = \begin{bmatrix} -0.7043 \\ 2.8648 \\ 2.7381 \end{bmatrix},$$

odnosno

$$A^{(3)} = H_2 A^{(2)} = \begin{bmatrix} -6.7082 & -0.5963 & -1.3416 \\ 0 & -4.5436 & -0.7043 \\ 0 & 0 & 2.8648 \\ 0 & 0 & 2.7381 \end{bmatrix}.$$

Posljednji korak treba odrediti reflektor H_3, oblika

$$H_3 = \begin{bmatrix} I_2 & 0 \\ 0 & \bar{H}_3 \end{bmatrix}, \qquad \bar{H}_3 \in \mathbb{R}^{2\times 2}, \quad \bar{H}_3 = I - \frac{1}{\gamma_3}\bar{v}^{(3)}\bar{v}^{(3)T}.$$

On mora transformirati odgovarajući dio 3. stupca $A^{(3)}(3:4,4)$, tako da

$$\bar{H}_3 \begin{bmatrix} 2.8648 \\ 2.7381 \end{bmatrix} = \begin{bmatrix} -\alpha_3 \\ 0 \end{bmatrix}.$$

Dobit ćemo sljedeće vrijednosti

$$\begin{aligned} \alpha_3 &= 3.9628 \\ \gamma_3 &= 27.0565 \end{aligned}$$

$$\bar{v}^{(3)} = \begin{bmatrix} 6.8276 \\ 2.7381 \end{bmatrix},$$

$$H_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -0.7229 & -0.6909 \\ 0 & 0 & -0.6909 & 0.7229 \end{bmatrix}.$$

H_3 ne mijenja 1. i 2. redak, a sva akcija se događa na recima od 3. do 4. djelovanjem matrice $\bar{H}_3$. Još k tome, $\bar{H}_3$ ne mijenja niti odgovarajuće djelove 1. i 2. stupca $A^{(3)}(3:4,1:2)$ jer su oni jednaki nul-vektorima.

$$\bar{H}_3 A^{(3)}(3:4,3) = \begin{bmatrix} -3.9628 \\ 0 \end{bmatrix}.$$

Konačno dobivamo

$$R_H = \begin{bmatrix} -6.7082 & -0.5963 & -1.3416 \\ 0 & -4.5436 & -0.7043 \\ 0 & 0 & -3.9628 \\ 0 & 0 & 0 \end{bmatrix},$$

$$Q_H = H_1 H_2 H_3 = \begin{bmatrix} -0.1491 & -0.4206 & 0.3776 & -0.8114 \\ -0.2981 & -0.8412 & -0.2542 & 0.3726 \\ 0.2981 & -0.0391 & -0.8510 & -0.4305 \\ -0.8944 & 0.3375 & -0.2619 & -0.1325 \end{bmatrix}.$$

2.2.2 QR faktorizacija pomoću Givensovih rotacija

Mnogo točnije je raditi QR faktorizaciju pomoću Givensovih rotacija, iako taj postupak zahtijeva više operacija.

Givensove rotacije su kvadratne matrice koje su dobivene ulaganjem dvodimenzionalnih rotacija u veću jediničnu matricu.

$$R(p,q;\phi) = \begin{bmatrix} 1 & & & \vdots & & & & \vdots & & & & \\ & \ddots & & \vdots & & & & \vdots & & 0 & & \\ & & 1 & \vdots & & & & \vdots & & & & \\ \cdots & \cdots & \cdots & c & \cdots & \cdots & \cdots & -s & \cdots & \cdots & \cdots & \\ & & & \vdots & 1 & & & \vdots & & & & \\ & & & \vdots & & \ddots & \vdots & & & & & \\ & & & \vdots & & & 1 & \vdots & & & & \\ \cdots & \cdots & \cdots & s & \cdots & \cdots & \cdots & c & \cdots & \cdots & \cdots & \\ & & & \vdots & & & & \vdots & 1 & & & \\ & 0 & & \vdots & & & & \vdots & & \ddots & & \\ & & & \vdots & & & & \vdots & & & 1 & \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ p \\ \\ \\ \\ q \\ \\ \\ \end{matrix}$$

$$\begin{matrix} & & & p & & & & q & & & & \end{matrix}$$

gdje su

$$\begin{aligned} R(p,q) = R(p,q;\phi) &\in \mathbb{R}^{n\times n} \\ c &= \cos\phi \\ s &= \sin\phi \\ \phi &\in [0, 2\pi\rangle, \end{aligned}$$

a p i q su pivotni indeksi i smatramo da je $p < q$.

Matrica $R(p, q; \phi)$ je očito ortogonalna i vrijedi

$$R(p, q; \phi)^{-1} = R(p, q; \phi)^T = R(p, q; -\phi).$$

Pomnožimo li matricu $A \in \mathbb{R}^{m \times n}$ slijeva sa $R(p, q; \phi)^T$, u A se promijeni samo p-ti i q-ti redak, a sve ostalo ostaje isto. Zato umjesto velike matrice možemo gledati pripadnu ravninsku rotaciju

$$\bar{R} = \bar{R}(p, q; \phi) = \begin{bmatrix} c & -s \\ s & c \end{bmatrix}$$

i samo p-ti i q-ti redak od A.

Neka su

$$[a_1 \ a_2 \ \cdots \ a_n] \quad i \quad [b_1 \ b_2 \ \cdots \ b_n]$$

p-ti i q-ti redak od A i neka je $\bar{A} = \bar{R}^T A$. Zapravo mijenjamo samo ovo:

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} a_1 & a_2 & \cdots & a_n \\ b_1 & b_2 & \cdots & b_n \end{bmatrix} = \begin{bmatrix} \bar{a}_1 & \bar{a}_2 & \cdots & \bar{a}_n \\ \bar{b}_1 & \bar{b}_2 & \cdots & \bar{b}_n \end{bmatrix}.$$

ϕ ćemo odabrati tako da se u A poništi element na mjestu (q, r), tj. tako da je $\bar{b}_r = 0$. Imamo:

$$\begin{aligned} ca_i + sb_i &= \bar{a}_i \\ -sa_i + cb_i &= \bar{b}_i, \qquad i = 1, \ldots, n \end{aligned}$$

Iz uvjeta $\bar{b}_r = 0$, je

$$cb_r = sa_r,$$

$$\bar{R}^T \begin{bmatrix} a_r \\ b_r \end{bmatrix} = \begin{bmatrix} \bar{a}_r \\ 0 \end{bmatrix}.$$

Budući da je $\bar{R}$ ortogonalna vrijedi

$$|\bar{a}_r| = \left\| \begin{bmatrix} \bar{a}_r \\ 0 \end{bmatrix} \right\|_2 = \left\| \bar{R}^T \begin{bmatrix} a_r \\ b_r \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} a_r \\ b_r \end{bmatrix} \right\|_2 = \sqrt{a_r^2 + b_r^2}.$$

$\bar{a}_r$ biramo tako da bude pozitivan:

$$\bar{a}_r = \sqrt{a_r^2 + b_r^2} > 0.$$

Ako je $a_r = b_r = 0$ tada $\bar{R} = I$. Napokon, dobivamo

$$c = \frac{a_r}{\bar{a}_r}, \qquad s = \frac{b_r}{\bar{a}_r}.$$

Napomena 2.4 *Zbog točnijeg računanja u aritmetici konačne preciznosti, c i s se često računaju kao*

- $|b_r| > |a_r|$

$$\tau = \frac{a_r}{b_r}, \quad s = \frac{sign(b_r)}{\sqrt{1+\tau^2}}, \quad c = s\tau,$$

- $|b_r| \leq |a_r|$

$$\tau = \frac{b_r}{a_r}, \quad c = \frac{sign(a_r)}{\sqrt{1+\tau^2}}, \quad s = c\tau.$$

Givensove rotacije poništavaju element po element matrice A. Za dobivanje QR faktorizacije potrebno je poništiti sve elemente donjeg trokuta matrice A, i to tako da se jednom poništeni element (jednak nuli) više ne mijenja. Način na koji biramo kojim redom ćemo ih poništavati se zove **pivotna strategija**.

Najčešća pivotna strategija je poništavanje po stupcima:

$$\begin{bmatrix} * & * & * & * & * \\ 5 & * & * & * & * \\ 4 & 9 & * & * & * \\ 3 & 8 & 12 & * & * \\ 2 & 7 & 11 & 14 & * \\ 1 & 6 & 10 & 13 & 15 \end{bmatrix},$$

i to tako da se na poziciji (i, j) element poništi Givensovom rotacijom $R_j(i-1, i)$. Na kraju, za $A \in \mathbb{R}^{m\times n}$, $m \geq n$ dobivamo da je

$$R_n(n, n+1)^T \cdots R_n(m-2, m-1)^T R_n(m-1, m)^T \cdots R_2(2, 3)^T \cdots R_2(m-2, m-1)^T \cdot \\ \cdot R_2(m-1, m)^T \cdot R_1(1, 2)^T \cdots R_1(m-2, m-1)^T R_1(m-1, m)^T A = R,$$

tj. $A = QR$, gdje se matrica Q tada dobiva kao produkt odgovarajućih Givensovih rotacija

$$Q = R_1(m-1, m) \cdots R_1(1, 2) R_2(m-1, m) \cdots R_2(2, 3) \cdots R_n(m-1, m) \cdots R_n(n, n+1).$$

Algoritam 2.2 *Za zadanu matricu $A \in \mathbb{R}^{m\times n}$, algoritam izračunava QR faktorizaciju pomoću Givensovih rotacija. Ortogonalna matrica Q se dobiva kao produkt $Q = R_1(m-1, m) \cdots R_1(1, 2) R_2(m-1, m) \cdots R_2(2, 3) \cdots R_n(m-1, m) \cdots R_n(n, n+1)$, pri čemu je $\bar{R}_j(i-1, i) = \begin{bmatrix} c_j^{(i-1,i)} & -s_j^{(i-1,i)} \\ s_j^{(i-1,i)} & c_j^{(i-1,i)} \end{bmatrix}$ a R je gornjetrokutasta.*

for $j = 1, \ldots, n$
  begin
    for $i = m, m-1, \ldots j+1$
      begin
        if $|a_{ij}| > |a_{i-1,j}|$
          begin
            $\tau = \frac{a_{i-1,j}}{a_{ij}}$;
            $s_j^{(i-1,i)} = \frac{sign(a_{ij})}{\sqrt{1+\tau^2}}$;
            $c_j^{(i-1,i)} = s_j^{(i-1,i)}\tau$;
          end
        else
          begin
            $\tau = \frac{a_{ij}}{a_{i-1,j}}$;
            $c_j^{(i-1,i)} = \frac{sign(a_{i-1,j})}{\sqrt{1+\tau^2}}$;
            $s_j^{(i-1,i)} = c_j^{(i-1,i)}\tau$;
          end
        $a_{i-1,j} = \sqrt{a_{i-1,j}^2 + a_{ij}^2}$;
        $a_{ij} = 0$;
        for $k = j+1 : n$
          begin
            $pom = a_{i-1,k}$;
            $a_{i-1,k} = c_j^{(i-1,i)} \cdot a_{i-1,k} + s_j^{(i-1,i)} \cdot a_{ik}$;
            $a_{ik} = -s_j^{(i-1,i)} \cdot pom + c_j^{(i-1,i)} \cdot a_{ik}$;
          end
      end
  end
$R = A$;

Primjer 2.2 *Rješavamo isti problem kao u primjeru 2.1. Dakle, tražimo QR faktorizaciju matrice*

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 4 & 2 \\ -2 & 0 & 3 \\ 6 & -1 & 2 \end{bmatrix},$$

ali ovaj puta pomoću Givensovih rotacija. Poništavanje elemenata pod dijagonalom izvršavat ćemo po redu, kako je opisano u algoritmu 2.2.

Dakle, prvo želimo poništiti element u matrici A na poziciji $(4,1)$, pomoću elementa na poziciji $(3,1)$. Zato tražimo Givensovu rotaciju $\bar{R}_1(3,4)$, oblika

$$\bar{R}_1(3,4) = \begin{bmatrix} c_1^{(3,4)} & -s_1^{(3,4)} \\ s_1^{(3,4)} & c_1^{(3,4)} \end{bmatrix},$$

takvu da $\bar{R}_1(3,4)^T [a_{3,1}\ a_{4,1}]^T = [a_{3,1}^{(2)}\ 0]^T$, gdje je $a_{3,1}^{(2)} = \sqrt{a_{3,1}^2 + a_{4,1}^2}$, odnosno

$$\begin{bmatrix} c_1^{(3,4)} & s_1^{(3,4)} \\ -s_1^{(3,4)} & c_1^{(3,4)} \end{bmatrix} \begin{bmatrix} -2 \\ 6 \end{bmatrix} = \begin{bmatrix} 6.3246 \\ 0 \end{bmatrix},$$

gdje je $6.3246 = \sqrt{(-2)^2 + 6^2} = \sqrt{40}$. Budući da je $|a_{4,1}| > |a_{3,1}|$, dalje računamo

$$\begin{aligned} \tau_1^{(3,4)} &= \frac{a_{3,1}}{a_{4,1}} = \frac{-2}{6} = -\frac{1}{3} = -0.3333 \\ s_1^{(3,4)} &= \frac{1}{\sqrt{1 + (-0.3333)^2}} = 0.9487 \\ c_1^{(3,4)} &= 0.9487 \cdot (-0.3333) = -0.3162 \end{aligned}$$

Konačni izgled Givensove rotacije u prvom koraku je

$$R_1^{(3,4)} = \begin{bmatrix} I_2 & 0 \\ 0 & \bar{R}_1^{(3,4)} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -0.3162 & -0.9487 \\ 0 & 0 & 0.9487 & -0.3162 \end{bmatrix}.$$

Kada sa $R_1(3,4)^T$ pomnožimo s desna matricu A, dobit ćemo matricu $A^{(2)}$, kojoj će 1. i 2. redak biti jednak istim recima u matrici A, a sva akcija se događa samo u 3. i 4. retku djelovanjem matrice $\bar{R}_1(3,4)$. Imamo

$$\bar{R}_1(3,4)A(3:4,1) = \begin{bmatrix} 6.3246 \\ 0 \end{bmatrix},$$

$$\bar{R}_1(3,4)A(3:4,2) = \begin{bmatrix} -0.9487 \\ 0.3162 \end{bmatrix},$$

$$\bar{R}_1(3,4)A(3:4,3) = \begin{bmatrix} 0.9487 \\ -3.4785 \end{bmatrix},$$

odnosno

$$A^{(2)} = R_1(3,4)^T A = \begin{bmatrix} 1.0000 & 2.0000 & -1.0000 \\ 2.0000 & 4.0000 & 2.0000 \\ 6.3246 & -0.9487 & 0.9487 \\ 0 & 0.3162 & -3.4785 \end{bmatrix}.$$

U sljedećem koraku želimo poništiti element u matrici $A^{(2)}$ na poziciji $(3,1)$, pomoću elementa na poziciji $(2,1)$. To radimo sa Givensovom rotacijom $R_1(2,3)$ oblika

$$R_1(2,3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \bar{R}_1(2,3) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \bar{R}_1(2,3) = \begin{bmatrix} c_1^{(2,3)} & -s_1^{(2,3)} \\ s_1^{(2,3)} & c_1^{(2,3)} \end{bmatrix},$$

takvom da je

$$\bar{R}_1(2,3)^T \begin{bmatrix} 2 \\ 6.3246 \end{bmatrix} = \begin{bmatrix} \sqrt{2^2 + 6.3246^2} \\ 0 \end{bmatrix}.$$

Dobit ćemo sljedeće vrijednosti

$$\begin{aligned} \tau_1^{(2,3)} &= 0.3162 \\ s_1^{(2,3)} &= 0.9535 \\ c_1^{(2,3)} &= 0.3015 \end{aligned}$$

Kada sa $R_1(2,3)^T$ pomnožimo s desna matricu $A^{(2)}$, dobit ćemo matricu $A^{(3)}$, kojoj će 1. i 4. redak biti jednak istim recima u matrici $A^{(2)}$, a sva akcija se događa samo u 2. i 3. retku djelovanjem matrice $\bar{R}_1(2,3)$. Imamo

$$\bar{R}_1(2,3)A^{(2)}(2:3,1) = \begin{bmatrix} 6.6332 \\ 0 \end{bmatrix},$$

$$\bar{R}_1(2,3)A^{(2)}(2:3,2) = \begin{bmatrix} 0.3015 \\ -4.0999 \end{bmatrix},$$

$$\bar{R}_1(2,3)A^{(2)}(2:3,3) = \begin{bmatrix} 1.5076 \\ -1.6209 \end{bmatrix},$$

odnosno

$$A^{(3)} = R_1(2,3)^T A^{(2)} = \begin{bmatrix} 1.0000 & 2.0000 & -1.0000 \\ 6.6332 & 0.3015 & 1.5076 \\ 0 & -4.0999 & -1.6209 \\ 0 & 0.3162 & -3.4785 \end{bmatrix}.$$

Dalje, želimo poništiti element u matrici $A^{(3)}$ na poziciji $(2,1)$, pomoću elementa na poziciji $(1,1)$. To radimo sa Givensovom rotacijom $R_1(1,2)$ oblika

$$R_1(1,2) = \begin{bmatrix} \bar{R}_1(1,2) & 0 \\ 0 & I_2 \end{bmatrix}, \quad \bar{R}_1(1,2) = \begin{bmatrix} c_1^{(1,2)} & -s_1^{(1,2)} \\ s_1^{(1,2)} & c_1^{(1,2)} \end{bmatrix},$$

takvom da je

$$\bar{R}_1(1,2)^T \begin{bmatrix} 1 \\ 6.6332 \end{bmatrix} = \begin{bmatrix} \sqrt{1^2 + 6.6332^2} \\ 0 \end{bmatrix}.$$

Dobit ćemo sljedeće vrijednosti

$$\begin{aligned} \tau_1^{(1,2)} &= 0.1508 \\ s_1^{(1,2)} &= 0.9888 \\ c_1^{(1,2)} &= 0.1491 \end{aligned}$$

Kada sa $R_1(1,2)^T$ *pomnožimo s desna matricu* $A^{(3)}$*, dobit ćemo matricu* $A^{(4)}$*, kojoj će 3. i 4. redak biti jednak istim recima u matrici* $A^{(3)}$*, a sva akcija se događa samo u 1. i 2. retku djelovanjem matrice* $\bar{R}_1(1,2)$*. Imamo*

$$\bar{R}_1(1,2)A^{(3)}(1:2,1) = \begin{bmatrix} 6.7082 \\ 0 \end{bmatrix},$$

$$\bar{R}_1(1,2)A^{(3)}(1:2,2) = \begin{bmatrix} 0.5963 \\ -1.9327 \end{bmatrix},$$

$$\bar{R}_1(1,2)A^{(3)}(1:2,3) = \begin{bmatrix} 1.3416 \\ 1.2136 \end{bmatrix},$$

odnosno

$$A^{(4)} = R_1(1,2)^T A^{(3)} = \begin{bmatrix} 6.7082 & 0.5963 & 1.3416 \\ 0 & -1.9327 & 1.2136 \\ 0 & -4.0999 & -1.6209 \\ 0 & 0.3162 & -3.4785 \end{bmatrix}.$$

Sada prelazimo u 2. stupac, u kojem želimo poništiti element u matrici $A^{(4)}$ *na poziciji* $(4,2)$*, pomoću elementa na poziciji* $(3,2)$*. To radimo sa Givensovom rotacijom* $R_2(3,4)$ *oblika*

$$R_2(3,4) = \begin{bmatrix} I_2 & 0 \\ 0 & \bar{R}_2(3,4) \end{bmatrix}, \quad \bar{R}_2(3,4) = \begin{bmatrix} c_2^{(3,4)} & -s_2^{(3,4)} \\ s_2^{(3,4)} & c_2^{(3,4)} \end{bmatrix},$$

takvom da je

$$\bar{R}_2(3,4)^T \begin{bmatrix} -4.0999 \\ 0.3162 \end{bmatrix} = \begin{bmatrix} \sqrt{(-4.0999)^2 + 0.3162^2} \\ 0 \end{bmatrix}.$$

Dobit ćemo sljedeće vrijednosti

$$\begin{aligned} \tau_2^{(3,4)} &= -0.0771 \\ c_2^{(3,4)} &= 0.0769 \\ s_2^{(3,4)} &= -0.9970 \end{aligned}$$

Kada sa $R_2(3,4)^T$ *pomnožimo s desna matricu* $A^{(4)}$*, dobit ćemo matricu* $A^{(5)}$*, kojoj će 1. i 2. redak biti jednak istim recima u matrici* $A^{(4)}$*, a sva akcija se događa samo u 3. i 4. retku djelovanjem matrice* $\bar{R}_2(3,4)$*. Još k tome,* $\bar{R}_2(3,4)$ *ne mijenja niti odgovarajući dio 1. stupca* $A^{(4)}(3:4,1)$ *jer je on jednak nul-vektoru. Imamo*

$$\bar{R}_2(3,4)A^{(4)}(3:4,2) = \begin{bmatrix} 4.1121 \\ 0 \end{bmatrix},$$

$$\bar{R}_2(3,4)A^{(4)}(3:4,3) = \begin{bmatrix} 1.3486 \\ 3.5929 \end{bmatrix},$$

odnosno

$$A^{(5)} = R_2(3,4)^T A^{(4)} = \begin{bmatrix} 6.7082 & 0.5963 & 1.3416 \\ 0 & -1.9327 & 1.2136 \\ 0 & 4.1121 & 1.3486 \\ 0 & 0 & 3.5929 \end{bmatrix}.$$

Dalje, želimo poništiti element u matrici $A^{(5)}$ *na poziciji* $(3,2)$*, pomoću elementa na poziciji* $(2,2)$*. To radimo sa Givensovom rotacijom* $R_2(2,3)$ *oblika*

$$R_2(2,3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \bar{R}_2(2,3) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \bar{R}_2(2,3) = \begin{bmatrix} c_2^{(2,3)} & -s_2^{(2,3)} \\ s_2^{(2,3)} & c_2^{(2,3)} \end{bmatrix},$$

takvom da je

$$\bar{R}_2(2,3)^T \begin{bmatrix} -1.9327 \\ 4.1121 \end{bmatrix} = \begin{bmatrix} \sqrt{(-1.9327)^2 + 4.1121^2} \\ 0 \end{bmatrix}.$$

Dobit ćemo sljedeće vrijednosti

$$\begin{aligned} \tau_2^{(2,3)} &= -0.4700 \\ s_2^{(2,3)} &= 0.9050 \\ c_2^{(2,3)} &= -0.4254 \end{aligned}$$

Kada sa $R_2(2,3)^T$ *pomnožimo s desna matricu* $A^{(5)}$*, dobit ćemo matricu* $A^{(6)}$*, kojoj će 1. i 4. redak biti jednak istim recima u matrici* $A^{(5)}$*, a sva akcija se događa samo u 2. i 3. retku djelovanjem matrice* $\bar{R}_2(2,3)$*. Još k tome,* $\bar{R}_2(2,3)$ *ne mijenja niti odgovarajući dio 1. stupca* $A^{(5)}(2:3,1)$ *jer je on jednak nul-vektoru. Imamo*

$$\bar{R}_2(2,3)A^{(5)}(2:3,2) = \begin{bmatrix} 4.5436 \\ 0 \end{bmatrix},$$

$$\bar{R}_2(2,3)A^{(5)}(2:3,3) = \begin{bmatrix} 0.7043 \\ -1.6719 \end{bmatrix},$$

odnosno

$$A^{(6)} = R_2(2,3)^T A^{(5)} = \begin{bmatrix} 6.7082 & 0.5963 & 1.3416 \\ 0 & 4.5436 & 0.7043 \\ 0 & 0 & -1.6719 \\ 0 & 0 & 3.5929 \end{bmatrix}.$$

I konačno u zadnjem koraku prelazimo u 3. stupac, u kojem želimo poništiti element u matrici $A^{(6)}$ *na poziciji* $(4,3)$*, pomoću elementa na poziciji* $(3,3)$*. To radimo sa Givensovom rotacijom* $R_3(3,4)$ *oblika*

$$R_3(3,4) = \begin{bmatrix} I_2 & 0 \\ 0 & \bar{R}_3(3,4) \end{bmatrix}, \quad \bar{R}_3(3,4) = \begin{bmatrix} c_3^{(3,4)} & -s_3^{(3,4)} \\ s_3^{(3,4)} & c_3^{(3,4)} \end{bmatrix},$$

takvom da je

$$\bar{R}_3(3,4)^T \begin{bmatrix} -1.6719 \\ 3.59291 \end{bmatrix} = \begin{bmatrix} \sqrt{(-1.6719)^2 + 3.5929^2} \\ 0 \end{bmatrix}.$$

Dobit ćemo sljedeće vrijednosti

$$\begin{aligned} \tau_3^{(3,4)} &= -0.4654 \\ s_3^{(3,4)} &= 0.9066 \\ c_3^{(3,4)} &= -0.4219 \end{aligned}$$

Kada sa $R_3(3,4)^T$ *pomnožimo s desna matricu* $A^{(6)}$*, dobit ćemo matricu* $A^{(7)}$*, kojoj će 1. i 2. redak biti jednak istim recima u matrici* $A^{(6)}$*, a sva akcija se događa samo u 3. i 4. retku djelovanjem matrice* $\bar{R}_3(3,4)$*. Još k tome,* $\bar{R}_3(3,4)$ *ne mijenja niti odgovarajuće djelove 1. i 2. stupca* $A^{(6)}(3:4,1:2)$ *jer su oni jednaki nul-vektorima. Imamo*

$$\bar{R}_2(2,3)A^{(5)}(2:3,2) = \begin{bmatrix} 3.9628 \\ 0 \end{bmatrix},$$

odnosno, konačno dobivamo

$$R_G = \begin{bmatrix} 6.7082 & 0.5963 & 1.3416 \\ 0 & 4.5436 & 0.7043 \\ 0 & 0 & 3.9628 \\ 0 & 0 & 0 \end{bmatrix},$$

$$Q_G = R_1(3,4)R_1(2,3)R_1(1,2)R_2(3,4)R_2(2,3)R_3(3,4) = \\ = \begin{bmatrix} 0.1491 & 0.4206 & -0.3776 & -0.8114 \\ 0.2981 & 0.8412 & 0.2542 & 0.3726 \\ -0.2981 & 0.0391 & 0.8510 & -0.4305 \\ 0.8944 & -0.3375 & 0.2619 & -0.1325 \end{bmatrix}.$$

Ako sada usporedimo R_H i Q_H, QR faktore dobivene pomoću Householderovih reflektora, sa R_G i Q_G, QR faktorima dobivenih pomoću Givensovih rotacija, tada možemo primijetiti da nisu posve jednaki. Radi se samo o različitim predznacima. Naime, R_H nema pozitivnu dijagonalu, dok R_G ju ima. Ako definiramo

$$D = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

tada imamo

$$\begin{aligned} Q_H D &= Q_G \\ D R_H &= R_G \end{aligned}$$

2.2.3 Rješavanje problema najmanjih kvadrata pomoću QR faktorizacije

Postoje dvije različite situacije kod rješavanja problema najmanjih kvadrata $\|Ax - b\|_2 \to \min$.

- **Matrica A je punog stupčanog ranga**
  Kao što smo vidjeli u opisu problema u tome slučaju je rješenje problema jednako
  $$x = (A^T A)^{-1} A^T b.$$
  Sada napišemo QR faktorizaciju matrice A
  $$A = QR = Q_1 R_1,$$

gdje je Q_1 ortonormalna matrica tipa $m \times n$, a R_1 trokutasta tipa $n \times n$ i uvrstimo u rješenje. Dobivamo

$$\begin{aligned} x &= (A^T A)^{-1} A^T b = (R_1^T Q_1^T Q_1 R_1)^{-1} R_1^T Q_1^T b \\ &= (R_1^T R_1)^{-1} R_1^T Q_1^T b = R_1^{-1} R_1^{-T} R_1^T Q_1^T b = R_1^{-1} Q_1^T b, \end{aligned}$$

tj. x se dobiva primjenom "invertirane" skraćene QR faktorizacije od A na b (po analogiji s rješavanjem linearnih sustava, samo što A ne mora imati inverz).

Preciznije, da bismo našli x, rješavamo trokutasti linearni sustav

$$R_1 x = Q_1^T b.$$

Na ovakav se način najčešće rješavaju problemi najmanjih kvadrata. Nije teško pokazati da je cijena računanja $2nm^2 - \frac{2}{3}m^3$.

- **Matrica A nema puni stupčani rang**
  U ovom slučaju prvo trebamo odrediti rang matrice A i tada se koristi **QR faktorizacija sa stupčanim pivotiranjem** (na prvo mjesto dovodi se stupac čiji "radni dio" ima najveću normu).

  Ako matrica A ima rang $r < n$, onda njena QR faktorizacija ima oblik

$$AP = QR = Q \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} r \\ n-r \\ m-n \end{matrix}, \qquad (45)$$
$$\begin{matrix} r & n-r \end{matrix}$$

  gdje je R_{11} regularna reda r, R_{12} neka $r \times (n-r)$ matrica, a matrica P je $n \times n$ matrica permutcija. Kad se nakon pivotiranja u tekućem koraku, i poništavanja ispoddijagonalnih elemenata u tekućem stupcu, na dijagonali nađe 0, tada znamo da je donji desni $(n-r) \times (n-r)$ blok matrice R_1 jednak nulmatrici. Kod rješavanja problema najmanjih kvadrata tada imamo

$$\begin{aligned} \|b - Ax\|_2^2 &= \|Q^T b - (Q^T A P)(P^T x)\|_2^2 \\ &= \|(c - R_{12} z) - R_{11} y\|_2^2 + \|d\|_2^2, \end{aligned}$$

  gdje je

$$P^T x = \begin{bmatrix} y \\ z \end{bmatrix} \begin{matrix} r \\ n-r \end{matrix}, \quad i \quad Q^T b = \begin{bmatrix} c \\ d \end{bmatrix} \begin{matrix} r \\ m-r \end{matrix}.$$

Prema tome, ako tražimo x koji minimizira normu reziduala, tada on mora zadovoljavati

$$x = P \begin{bmatrix} R_{11}^{-1}(c - R_{12}z) \\ z \end{bmatrix}.$$

Ako stavimo da je $z = 0$ tada dobivamo **osnovno rješenje**

$$x = P \begin{bmatrix} R_{11}^{-1}c \\ 0 \end{bmatrix},$$

koje samo ne mora imati minimalnu normu u skupu svih rješenja, ali ga je jednostavno izračunati i ima najviše r elemenata različitih od nule.

Do rješenja problema najmanjih kvadrata sa minimalnom normom možemo, s druge strane, doći pomoću **potpune ortogonalne dekompozicije**. U jednakosti (45) možemo izvesti još jednu QR faktorizaciju, i to na sljedeći način. Trebamo izračunati $n \times n$ ortogonalnu matricu Z takvu da je

$$Z \begin{bmatrix} R_{11}^T \\ R_{12}^T \end{bmatrix} = \begin{bmatrix} L_{11}^T \\ 0 \end{bmatrix} \begin{matrix} r \\ n-r \end{matrix}$$

gdje je L_{11}^T $r \times r$ gornjetrokutasta matrica. Tada slijedi

$$Q^T A S = L = \begin{bmatrix} L_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} r \\ m-r \end{matrix},$$
$$\begin{matrix} r & n-r \end{matrix}$$

gdje je $S = PZ^T$. Primijetimo da je $\mathcal{N}(A) = \mathcal{R}(S(1:n, r+1:n))$. Kod rješavanja problema najmanjih kvadrata tada imamo

$$\|Ax - b\|_2^2 = \|(Q^T A S)S^T x - Q^T b\|_2 = \|L_{11}w - c\|_2^2 + \|d\|_2^2,$$

gdje je

$$S^T x = \begin{bmatrix} w \\ \omega \end{bmatrix} \begin{matrix} r \\ n-r \end{matrix} \qquad Q^T b = \begin{bmatrix} c \\ d \end{bmatrix} \begin{matrix} r \\ m-r \end{matrix}.$$

Jasno je, da ako x treba minimizirati normu reziduala, tada moramo imati $w = L_{11}^{-1}c$, a da bi x imao minimalnu normu tada mora biti $\omega = 0$. Dakle u ovom slučaju rješenje problema najmanjih kvadrata sa minimalnom normom glasi

$$x = S \begin{bmatrix} L_{11}^{-1}c \\ 0 \end{bmatrix}.$$

Napomena 2.5 *Zbog grešaka zaokruživanja, umjesto pravog R, izračunamo*

$$R' = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \\ 0 & 0 \end{bmatrix}.$$

Naravno, željeli bismo da je $\|R_{22}\|_2$ vrlo mala, reda veličine $\varepsilon\|A\|_2$, pa da je možemo "zaboraviti", tj. staviti $R_{22} = 0$ i tako odrediti rang od A. Nažalost, to nije uvijek tako. Na primjer, bidijagonalna matrica

$$A = \begin{bmatrix} \frac{1}{2} & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \frac{1}{2} \end{bmatrix}$$

je skoro singularna ($\det(A) = 2^{-n}$), njena QR faktorizacija je $Q = I$, $R = A$, i nema niti jednog R_{22} koji bi bio po normi malen.

Zbog toga koristimo pivotiranje, koje R_{11} pokušava držati što bolje uvjetovanim, a R_{22} po normi što manjim.

Zadatak 2.1 *Zadane su točke u ravnini*

$$(1, 3.5),\ (2, 4.9),\ (3, 6.8)\ (4, 9.3),\ (5, 10.9),\ (6, 13.4),\ (7, 15.1),\ (8, 16.7),$$

$$(9, 19)\ (10, 21.2)$$

koje treba aproksimirati pravcem

$$f(x) = a_0 + a_1 x,$$

koristeći metodu najmanjih kvadrata. Iz ovih podataka možemo zaključiti da se radi o malo perturbiranim točkama sa pravca $p(x) = 2x + 1$.

Napomena 2.6 *Želimo minimizirati sljedeći funkcional*

$$S(a_0, a_1) = \sum_{k=0}^{n} (y_k - a_0 - a_1 x_k)^2 \to \min,$$

pri čemu je $\phi_1(x) = 1$, a $\phi_2(x) = x$. Da bismo kreirali matricu $A = [a_{ij}]$, trebamo naći njene elemente. Prema opisu problema sa početka poglavlja (41), za njih vrijedi

$$a_{k,1} = 1, \qquad a_{k,2} = x_k,$$

dakle

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \\ 1 & 9 \\ 1 & 10 \end{bmatrix}, \qquad b = \begin{bmatrix} 3.5 \\ 4.9 \\ 6.8 \\ 9.3 \\ 10.9 \\ 13.4 \\ 15.1 \\ 16.7 \\ 19 \\ 21.2 \end{bmatrix}.$$

Prvo računamo QR faktorizaciju matrice A, recimo pomoću Householderovih reflektora. Dobit ćemo

$$R_1 = \begin{bmatrix} -3.1623 & -17.3925 \\ 0 & 9.0830 \end{bmatrix}, \quad Q_1 = \begin{bmatrix} -0.3162 & -0.4954 \\ -0.3162 & -0.3853 \\ -0.3162 & -0.2752 \\ -0.3162 & -0.1651 \\ -0.3162 & -0.0550 \\ -0.3162 & 0.0550 \\ -0.3162 & 0.1651 \\ -0.3162 & 0.2752 \\ -0.3162 & 0.3853 \\ -0.3162 & 0.4954 \end{bmatrix},$$

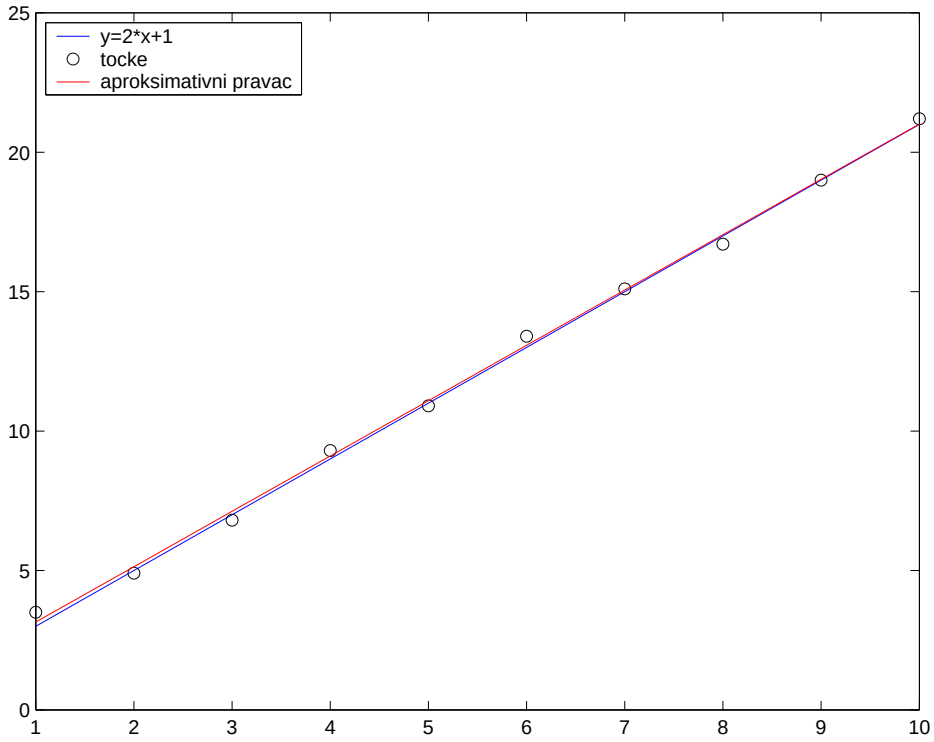
a rješenje problema najmanjih kvadrata je dano sa

$$x = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = R_1^{-1} Q_1^T b = \begin{bmatrix} 1.1667 \\ 1.9842 \end{bmatrix},$$

odnosno, ovime smo izračunali tražene koeficijente pravca

$$a_0 = 1.1667 \qquad a_1 = 1.9842.$$

Dakle aproksimativni pravac je $\hat{p}(x) = 1.1667 + 1.9842x$*, i prikazan je na slici 21.*



Slika 21: Aproksimativni pravac za točke iz zadatka 2.1, dobiven kao rezultat problema najmanjih kvadrata, riješenih pomoću QR faktorizacije.

2.3 Dekompozicija singularnih vrijednosti (SVD)

Množenjem unitarnim matricama možemo proizvoljnu pravokutnu matricu svesti na dijagonalni oblik.

Teorem 2.2 (Dekompozicija singularnih vrijednosti (SVD)) *Neka je $A \in \mathbb{C}^{m\times n}$ matrica ranga r. Tada postoje unitarne matrice $U \in \mathbb{C}^{m\times m}$ i $V \in \mathbb{C}^{n\times n}$ takve da je*

$$U^*AV = \begin{bmatrix} \Sigma_+ & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} r \\ m-r \end{matrix}$$
$$\begin{matrix} r & n-r \end{matrix}$$

gdje je $\Sigma_+ = diag(\sigma_1, \ldots, \sigma_r)$, uz $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.
(Tada je $A = U\Sigma V^*$.)

Definicija 2.1 *Pozitivni skalari $\sigma_1, \ldots, \sigma_r$ zovu se* **singularne vrijednosti**, *a stupci matrica U i V zovu se* **lijevi i desni singularni vektori** *matrice A.*

Napomena 2.7 *Ako izvršimo particiju matrica*

$$U = \underset{r}{[U_1} \quad \underset{m-r}{U_2]} \qquad V = \underset{r}{[V_1} \quad \underset{n-r}{V_2]}$$

onda iz Teorema 2.2 slijedi

$$\begin{bmatrix} \Sigma_+ & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} U_1^* \\ U_2^* \end{bmatrix} A[\ V_1 \quad V_2\ ],$$

odnosno

$$\Sigma_+ = U_1^* A V_1.$$

Dakle, dekompoziciju singularnih vrijednosti možemo napisati u skraćenom obliku

$$A = U_1 \Sigma_+ V_1^*,$$

pri čemu su $U_1 \in \mathbb{C}^{m\times r}$ *i* $V_1 \in \mathbb{C}^{n\times r}$ *ortonormalne matrice.*

Napomena 2.8 *Za matricu* $A \in \mathbb{C}^{m\times n}$ *ranga* $r \leq \min(m,n)$*, matrice* $A^*A \in \mathbb{C}^{n\times n}$ *i* $AA^* \in \mathbb{C}^{m\times m}$ *su simetrične i pozitivno semidefinitne. Vrijedi:*

- $$V^*A^*AV = diag(\sigma_1^2, \ldots, \sigma_r^2, \underbrace{0, \ldots, 0}_{n-r}), \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

  tj, kvadrati singularnih vrijednosti matrice A *su svojstvene vrijednosti matrice* A^*A*, samo što se među njima nalazi* $n-r$ *nula, a stupci matrice* V *su njeni svojstveni vektori.*

- $$U^*AA^*U = diag(\sigma_1^2, \ldots, \sigma_r^2, \underbrace{0, \ldots, 0}_{m-r}), \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

  tj, kvadrati singularnih vrijednosti matrice A *su svojstvene vrijednosti matrice* AA^**, samo što se među njima nalazi* $m-r$ *nula, a stupci matrice* U *su njeni svojstveni vektori.*

SVD ima sljedeće korisno svojstvo.

- Neka je SVD matrice $A \in \mathbb{C}^{m\times n}$ dana kao u Teoremu 2.2. Ako je $k < r = \text{rang}(A)$ i

  $$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^*$$

  tada

  $$\min_{\text{rang}(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}.$$

2.3.1 Rješavanje problema najmanjih kvadrata pomoću SVD dekompozicije

I u ovom slučaju postoje dvije različite situacije kod rješavanja problema najmanjih kvadrata $\|Ax - b\|_2 \to \min$.

- Ako A ima puni rang, onda je rješenje problema najmanjih kvadrata

$$\min_x \|Ax - b\|_2$$

jednako

$$x = V_1 \Sigma_+^{-1} U_1^T b,$$

tj. dobiva se primjenom "invertiranog" skraćenog SVD-a od A na b, pri čemu je $V_1 = V$.

To lako možemo provjeriti na sljedeći način. Prvo primijetimo da vrijedi

$$\|Ax - b\|_2^2 = \|U_1 \Sigma_+ V^T x - b\|_2^2.$$

Budući da je A punog ranga, to je i Σ. Zbog unitarne ekvivalencije 2-norme, vrijedi

$$\begin{aligned} \|U_1 \Sigma_+ V^T x - b\|_2^2 &= \|U^T (U_1 \Sigma_+ V^T x - b)\|_2^2 = \left\| \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} (U_1 \Sigma_+ V^T x - b) \right\|_2^2 \\ &= \left\| \begin{bmatrix} \Sigma_+ V^T x - U_1^T b \\ -U_2^T b \end{bmatrix} \right\|_2^2 = \|\Sigma_+ V^T x - U_1^T b\|_2^2 + \|U_2^T b\|_2^2. \end{aligned}$$

Prethodni izraz se minimizira ako je prvi član jednak 0, tj. ako je

$$x = V \Sigma_+^{-1} U_1^T b.$$

Usput dobivamo i vrijednost minimuma $\min_x \|Ax - b\|_2 = \|U_2^T b\|_2$.

- SVD se primjenjuje u metodi najmanjih kvadrata i kad matrica A nema puni stupčani rang. Rješenja su istog oblika (sjetite se, više ih je), samo što moramo znati izračunati "inverz" matrice Σ kad ona nije punog ranga, tj. kad ima neke nule na dijagonali. Takav inverz zove se generalizirani inverz i označava sa Σ^+ ili $\Sigma^\dagger$.

U slučaju da je

$$\Sigma = \begin{bmatrix} \Sigma_+ & 0 \\ 0 & 0 \end{bmatrix},$$

pri čemu je Σ_+ regularna, onda je

$$\Sigma^\dagger = \begin{bmatrix} \Sigma_+^{-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Neka matrica A ima rang $r < n$. Rješenje x koje minimizira $\|Ax - b\|_2$ može se karakterizirati na sljedeći način. Neka je $A = U\Sigma V^T$ SVD od A i neka je

$$A = U\Sigma V^T = [U_1, U_2, U_3] \begin{bmatrix} \Sigma_+ & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} [V_1, V_2]^T = U_1 \Sigma_+ V_1^T,$$

gdje je Σ_+ nesingularna, reda r, matrice U_1 i V_1 imaju r stupaca, matrice U_2 i V_2 imaju $n - r$ stupaca, a matrica U_3 ima $m - n$ stupaca. Neka je

$$\sigma := \sigma_{\min}(\Sigma_+),$$

najmanja ne-nula singularna vrijednost od A. Tada se sva rješenja problema najmanjih kvadrata mogu napisati u formi

$$x = V_1 \Sigma_+^{-1} U_1^T b + V_2 z,$$

gdje je z proizvoljni vektor. Rješenje x koje ima minimalnu 2-normu je ono za koje je $z = 0$, tj.

$$x = V_1 \Sigma_+^{-1} U_1^T b,$$

i vrijedi ocjena

$$\|x\|_2 \leq \frac{\|b\|_2}{\sigma}.$$

Gornje tvrdnje ćemo sada provjeriti. Korištenjem unitarne invarijantnosti 2-norme, dobivamo

$$\begin{aligned} \|Ax - b\|_2^2 &= \|U^T(Ax - b)\|_2^2 = \left\| \begin{bmatrix} U_1^T \\ U_2^T \\ U_3^T \end{bmatrix} (U_1 \Sigma_+ V_1^T x - b) \right\|_2^2 \\ &= \left\| \begin{bmatrix} \Sigma_+ V_1^T x - U_1^T b \\ -U_2^T b \\ -U_3^T b \end{bmatrix} \right\|_2^2 = \|\Sigma_+ V_1^T x - U_1^T b\|_2^2 + \|U_2^T b\|_2^2 + \|U_3^T b\|_2^2. \end{aligned}$$

Očito, izraz je minimiziran kad je prva od tri norme u posljednjem redu jednaka 0, tj. ako je

$$\Sigma_+ V_1^T x = U_1^T b,$$

ili

$$x = V_1 \Sigma_+^{-1} U_1^T b.$$

Stupci matrica V_1 i V_2 su međusobno ortogonalni, pa je $V_1^T V_2 z = 0$ za sve vektore z. Odavde vidimo da x ostaje rješenje problema najmanjih kvadrata i kad mu dodamo $V_2 z$, za bilo koji z, tj. ako je

$$x = V_1 \Sigma_+^{-1} U_1^T b + V_2 z.$$

To su ujedno i sva rješenja, jer stupci matrice V_2 razapinju nul-potprostor $\mathcal{N}(A)$. Osim toga, zbog spomenute ortogonalnosti vrijedi i

$$\|x\|_2^2 = \|V_1 \Sigma_+^{-1} U_1^T b\|_2^2 + \|V_2 z\|_2^2,$$

a to je minimalno za $z = 0$. Na kraju, za to minimalno rješenje vrijedi ocjena

$$\|x\|_2 = \|V_1 \Sigma_+^{-1} U_1^T b\|_2 = \|\Sigma_+^{-1} U_1^T b\|_2 \leq \frac{\|U_1^T b\|_2}{\sigma} = \frac{\|b\|_2}{\sigma}.$$

Primjerom se lako pokazuje da je ova ocjena dostižna.

Zadatak 2.2 *Rješavamo isti problem kao u zadatku 2.1, samo što ćemo problem najmanjih kvadrata ovaj puta rješavati pomoću SVD dekompozicije.*

Napomena 2.9 *Podsjetimo se, matrica i desna strana problema su oblika*

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \\ 1 & 9 \\ 1 & 10 \end{bmatrix}, \qquad b = \begin{bmatrix} 3.5 \\ 4.9 \\ 6.8 \\ 9.3 \\ 10.9 \\ 13.4 \\ 15.1 \\ 16.7 \\ 19 \\ 21.2 \end{bmatrix}.$$

Sada računamo SVD dekompoziciju matrice A. Dobit ćemo

$$U_1 = \begin{bmatrix} 0.0571 & -0.5850 \\ 0.1070 & -0.4869 \\ 0.1570 & -0.3887 \\ 0.2069 & -0.2906 \\ 0.2569 & -0.1925 \\ 0.3068 & -0.0944 \\ 0.3567 & 0.0037 \\ 0.4067 & 0.1019 \\ 0.4566 & 0.2000 \\ 0.5065 & 0.2981 \end{bmatrix}, \quad \Sigma_+ = \begin{bmatrix} 19.8217 & 0 \\ 0 & 1.4491 \end{bmatrix}, \quad V = \begin{bmatrix} 0.1422 & -0.9898 \\ 0.9898 & 0.1422 \end{bmatrix},$$

a rješenje problema najmanjih kvadrata je dano sa

$$x = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = V\Sigma_+^{-1}U_1^T b = \begin{bmatrix} 1.1667 \\ 1.9842 \end{bmatrix},$$

odnosno, ovime smo izračunali tražene koeficijente pravca

$$a_0 = 1.1667 \qquad a_1 = 1.9842.$$

Dakle aproksimativni pravac je $\hat{p}(x) = 1.1667 + 1.9842x$*, i on jednak je onome iz zadatka 2.1.*

2.3.2 Generalizirani inverz

Ako želimo proširiti pojam inverza (X) i na matrice koje nisu regularne ili čak nisu kvadratne, onda zahtijevamo da on mora zadovoljavati malo oslabljene uvjete nego standardni inverz:

$$AX = XA = I.$$

Najpoznatiji generalizirani inverz je tzv. *Moore–Penroseov inverz*, koji je određen sa sljedeća četiri uvjeta.
Moore–Penroseovi uvjeti:

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

Za generalizirani inverz vrijede sljedeća svojstva

- Neka je $A \in \mathbb{C}^{m\times n}$. Tada postoji jedinstvena matrica $X \in \mathbb{C}^{n\times m}$, koja zadovoljava Penroseove uvjete. Ta matrica ima oblik

  $$A^\dagger = V \begin{bmatrix} \Sigma_+^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^*, \quad \text{pri čemu je } A = U \begin{bmatrix} \Sigma_+ & 0 \\ 0 & 0 \end{bmatrix} V^*$$

  singularna dekompozicija matrice A.

- Za proizvoljnu matricu $A \in \mathbb{C}^{m\times n}$ vrijedi:

  1. $\left(A^\dagger\right)^\dagger = A$

2. $\left(\bar{A}\right)^{\dagger} = \overline{(A^{\dagger})}$
3. $\left(A^{T}\right)^{\dagger} = \left(A^{\dagger}\right)^{T}$
4. $\text{rang}(A) = \text{rang}(A^{\dagger}) = \text{rang}(AA^{\dagger}) = \text{rang}(A^{\dagger}A)$
5. Ako matrica $A \in \mathbb{C}^{m\times n}$ ima rang n, tada je

$$A^{\dagger} = (A^{*}A)^{-1}A^{*} \quad \text{i} \quad A^{\dagger}A = I_n.$$

6. Ako matrica $A \in \mathbb{C}^{m\times n}$ ima rang m, tada je

$$A^{\dagger} = A^{*}(AA^{*})^{-1} \quad \text{i} \quad AA^{\dagger} = I_m.$$

7. Ako je $A = FG$ i $\text{rang}(A) = \text{rang}(F) = \text{rang}(G)$, tada je

$$A^{\dagger} = G^{\dagger}F^{\dagger}.$$

8. Ako su U i V unitarne matrice, tada je

$$(UAV)^{\dagger} = V^{*}A^{\dagger}U^{*}.$$

3 Problem svojstvenih vrijednosti

3.1 Opis problema

I ovaj puta ćemo započeti promatranje danog problema primjerima iz prakse. Rješavanje problema svojstvenih vrijednosti pojavljuje se kao posljedica rješavanja raznih problema u ekonomiji, fizici ...

Primjer 3.1 *Pretpostavimo da korporacije mogu biti u jednoj od n mogućih kreditnih kategorija ("credit rating"), i da one mogu preći iz jedne kategorije u bilo koju drugu u diskretnim jedinicam vremena, recimo svake godine. Neka je a_{ij} vjerojatnost da korporacija prijeđe u kategoriju i sljedeće godine, ako se trenutno nalazi u kategoriji j. Pretpostavimo da je ovaj sustav zapravo* **Markovljev lanac**, *tj. da vjerojatnosti prelaska ovise samo o trenutnoj kategoriji, a ne o prošlim kategorijama. To je samo aproksimacija realnih sustava.*

Svojstva matrice $A = [a_{ij}]$:

- $0 \leq a_{ij} \leq 1$, *jer se radi o vjerojatnostima.*
- $\sum_i a_{ij} = 1$, *za svako j, budući da sustav uvijek mora preći u neku novu kategoriju.*
- *Kvadratna matrica $A = [a_{ij}]$ ima nenegativne elemente, i suma elemenata svakog stupca je 1.*

Pretpostavimo da imamo velik skup korporacija, i neka u_j predstavlja udio u tom skupu onih korporacija, koje su u kategoriji j u početnom trenutku, uz svojstva $0 \leq u_j \leq 1$ i $\sum_j u_j = 1$. Ako je skup dovoljno velik, i ako se prelazak iz kategorije u kategoriju svake korporacije odvija neovisno o drugima, tada se udio korporacija u skupu svih korporacija koje će se nakon jedne godine nalaziti u kategoriji i, označen sa v_i, dobiva kao

$$v_i = \sum_j a_{ij} u_j, \qquad ili \qquad v = Au.$$

Primijetimo da

$$\sum_i v_i = \sum_i \sum_j a_{ij} u_j = \sum_j \left(\sum_i a_{ij} \right) u_j = \sum_j u_j = 1.$$

Općenito, ako sa $u^{(k)}$ označimo vektor gustoće nakon k koraka, tada

$$u^{(k)} = Au^{(k-1)} = A^k u^{(0)}.$$

Prema tome dugoročno ponašanje gustoće ovisi o svojstvima visokih potencija matrice A.

Prema gornjim pretpostavkama, moguće je procijeniti vjerojatnosti prelaska na osnovu povijesnih podataka. U tablici 1 nalaze se vjerojatnosti prelaska izraženi u postocima, za jednu godinu, objavljeni u Credit Metrics *za 2001. godinu.*

Konačna	Početna kategorija							
kategorija	AAA	AA	A	BBB	BB	B	CCC	D
AAA	90.81	0.70	0.09	0.02	0.03	0	0.22	0
AA	8.33	90.65	2.27	0.33	0.14	0.11	0	0
A	0.68	7.79	91.05	5.95	0.67	0.24	0.22	0
BBB	0.06	0.64	5.52	86.93	7.73	0.43	1.30	0
BB	0.12	0.06	0.74	5.30	80.53	6.48	2.38	0
B	0	0.14	0.26	1.17	8.84	83.46	11.24	0
CCC	0	0.02	0.01	0.12	1.00	4.07	64.86	0
D	0	0	0.06	0.18	1.06	5.20	19.79	100

Tablica 1: Tablica vjerojatnost prelaska izražena u %

Sada se postavlja pitanje što se događa kad $k \to \infty$? Da li se sustav smiruje u ravnotežnom stanju? Ako postoji ravnotežno stanje $u^{(\infty)} = \bar{u}$, tada mora vrijediti

$$A\bar{u} = \bar{u},$$

tako da se ono ne mijenja u nadolazećim godinama. Dakle, $\bar{u}$ mora biti svojstveni vektor matrice A koji pripada svojstvenoj vrijednosti jednakoj 1. Ako pogledamo tablicu, također je jasno da je jedan takav svojstveni vektor jednak $[0, \ldots, 0, 1]^T$, tj. ako su svi u kategoriji D tada svi i ostaju u toj kategoriji. To nužno ne mora značiti, da svi teže ka kategoriji D. To ćemo sada provjeriti.

Pretpostavimo da A ima n linearno nezavisnih svojstvenih vektora $v_1, \ldots, v_n$ i n svojstvenih vrijednosti $\lambda_1, \ldots, \lambda_n$, i pretpostavimo da je v_1 svojstveni vektor koji pripada svojstvenoj vrijednosti $\lambda_1 = 1$. Tada $u^{(k)}$ možemo raspisati po komponentama u smjerovima $v_1, \ldots, v_n$ kao

$$u^{(k)} = \nu_1^{(k)} v_1 + \cdots + \nu_n^{(k)} v_n.$$

Tada imamo

$$u^{(k+1)} = Au^{(k)} = \nu_1^{(k)} Av_1 + \cdots + \nu_n^{(k)} Av_n = \lambda_1 \nu_1^{(k)} v_1 + \cdots + \lambda_n \nu_n^{(k)} v_n.$$

Prema tome dobiva se da je $\nu_j^{(k+1)} = \lambda_j \nu_j^{(k)}$, *odnosno*

$$\nu_j^{(k)} = \lambda_j^k \nu_j^{(0)}.$$

Komponenta vektora u smjeru j*-tog svojstvenog vektora ili raste ili trne eksponencijalno kad* $k \to \infty$, *ovisno o tome da li je odgovarajuća svojstvena vrijednost veća ili manja od 1 po apsolutnoj vrijednosti.*

Jasno je da niti jedna svojstvena vrijednost od A *ne može biti veća od 1 po apsolutnoj vrijednosti, jer da to nije tako, apsolutna vrijednost od* $u^{(k)}$ *bi rasla eksponencijalno, što je u suprotnosti sa činjenicom da je suma svih komponenti od* $u^{(k)}$ *jednaka 1. Mi znamo da postoji najmanje jedna svojstvena vrijednost jednaka 1. Prema tome, ako su sve ostale svojstvene vrijednosti po apsolutnoj vrijednosti manje od 1, tada će njihove komponente utrnuti, i dugoročno gledano katergorija kojoj će svi težiti je kategorija* D.

Provjerim to sad u Octave-*i. Matricu* A *dobit ćemo iz tablice 1 tako da svaki njen element podijelimo sa 100. Dakle imamo*

$$A = \begin{bmatrix} 0.9081 & 0.0070 & 0.0009 & 0.0002 & 0.0003 & 0 & 0.0022 & 0 \\ 0.0833 & 0.9065 & 0.0227 & 0.0033 & 0.0014 & 0.0011 & 0 & 0 \\ 0.0068 & 0.0779 & 0.9105 & 0.0595 & 0.0067 & 0.0024 & 0.0022 & 0 \\ 0.0006 & 0.0064 & 0.0552 & 0.8693 & 0.0773 & 0.0043 & 0.0130 & 0 \\ 0.0012 & 0.0006 & 0.0074 & 0.0530 & 0.8053 & 0.0648 & 0.0238 & 0 \\ 0 & 0.0014 & 0.0026 & 0.0117 & 0.0884 & 0.8346 & 0.1124 & 0 \\ 0 & 0.0002 & 0.0001 & 0.0012 & 0.0100 & 0.0407 & 0.6486 & 0 \\ 0 & 0 & 0.0006 & 0.0018 & 0.0106 & 0.0520 & 0.1979 & 1.0000 \end{bmatrix},$$

njene svojstvene vrijednosti su dane sa

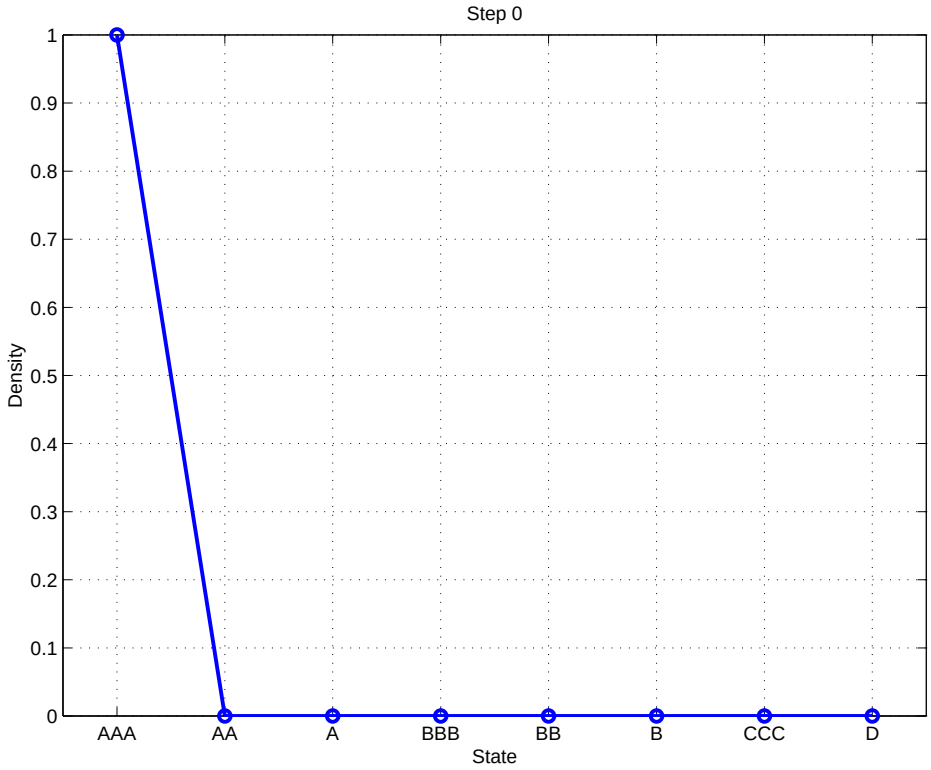
$$\sigma(A) = \{\ 0.6260 \quad 0.7318 \quad 0.8259 \quad 0.8725 \quad 0.9058 \quad 0.9326 \quad 0.9882 \quad 1.0000\ \}$$

dok su odgovarajući svojstveni vektori dani sa

$$\begin{bmatrix} -0.0064 \\ 0.0034 \\ 0.0079 \\ -0.0616 \\ 0.0785 \\ -0.4626 \\ 0.8028 \\ -0.3625 \end{bmatrix}, \quad \begin{bmatrix} 0.0007 \\ -0.0124 \\ 0.1325 \\ -0.4423 \\ 0.7400 \\ -0.4429 \\ -0.1339 \\ 0.1581 \end{bmatrix}, \quad \begin{bmatrix} -0.0058 \\ 0.1476 \\ -0.5818 \\ 0.6458 \\ 0.0775 \\ -0.4072 \\ -0.0849 \\ 0.2087 \end{bmatrix}, \quad \begin{bmatrix} 0.0908 \\ -0.4831 \\ 0.3625 \\ 0.4326 \\ -0.1563 \\ -0.5181 \\ -0.0991 \\ 0.3703 \end{bmatrix},$$

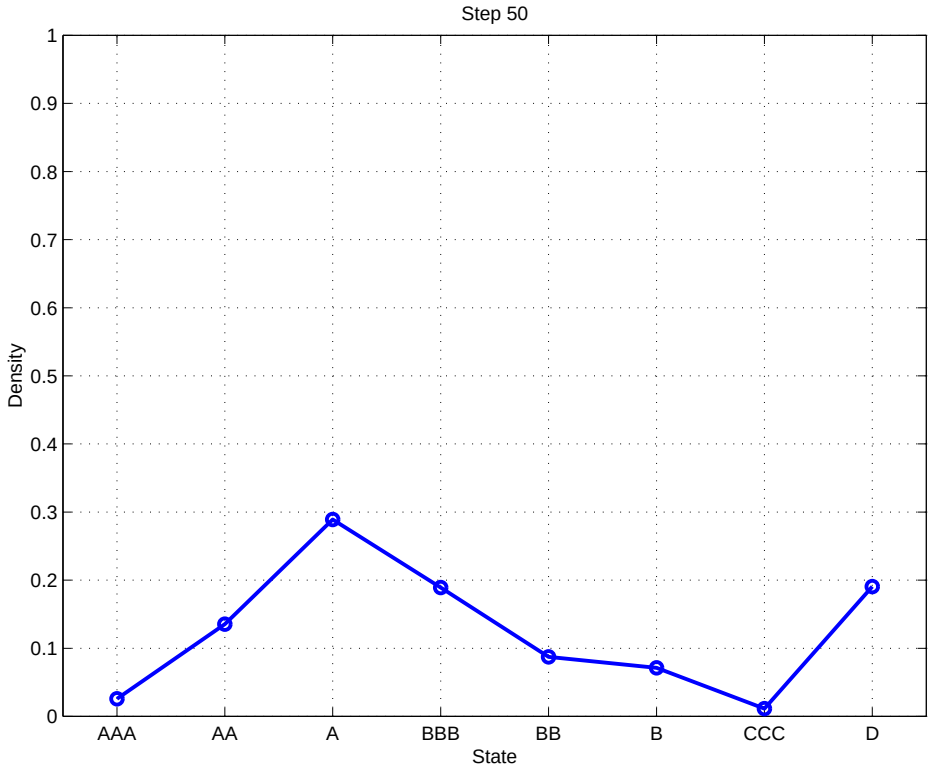
$$\begin{bmatrix} 0.1328 \\ -0.0192 \\ -0.5315 \\ -0.0108 \\ 0.3284 \\ 0.5408 \\ 0.0981 \\ -0.5381 \end{bmatrix}, \begin{bmatrix} -0.0860 \\ -0.3309 \\ -0.1568 \\ 0.3224 \\ 0.3661 \\ 0.4495 \\ 0.0784 \\ -0.6421 \end{bmatrix}, \begin{bmatrix} -0.0148 \\ -0.1126 \\ -0.3049 \\ -0.2308 \\ -0.1190 \\ -0.1047 \\ -0.0170 \\ 0.9030 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1.0000 \end{bmatrix}.$$

Ako za $u^{(0)}$ uzmemo da je jednak prvom jediničnom vektoru e_1, odnosno da je $u^{(0)} = [1\ 0\ \cdots\ 0]^T$, tada su vektori gustoće za korake $k = 0, 50, 100, 200$ prikazani grafički na slikama 22, 23, 24 i 25.

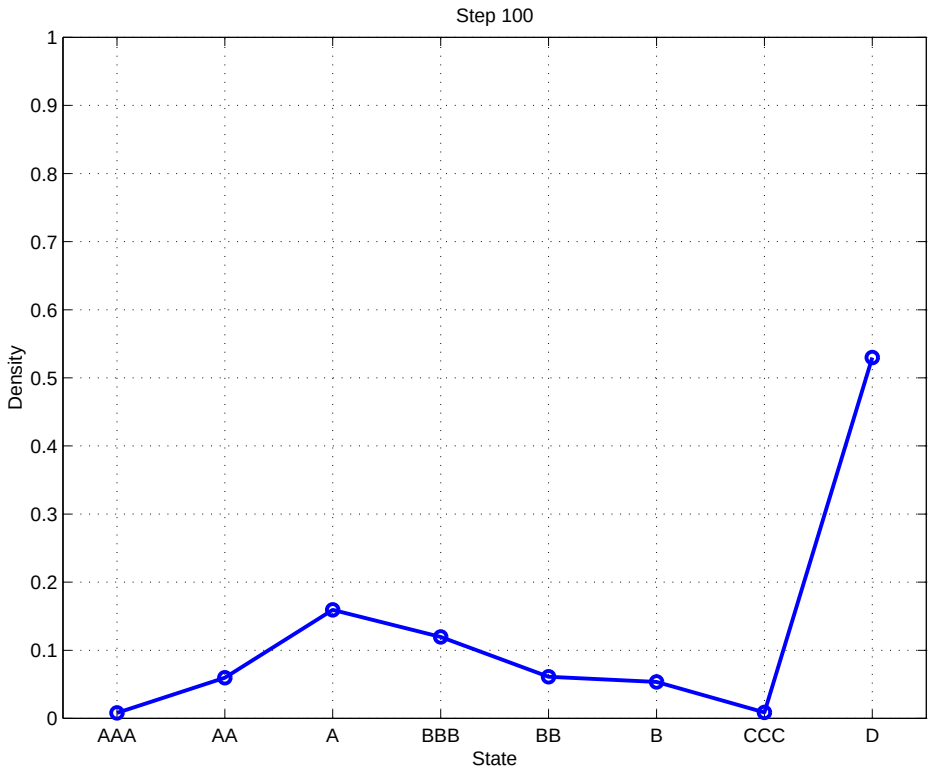


Slika 22: Grafički prikaz vektora gustoće $u^{(0)}$.

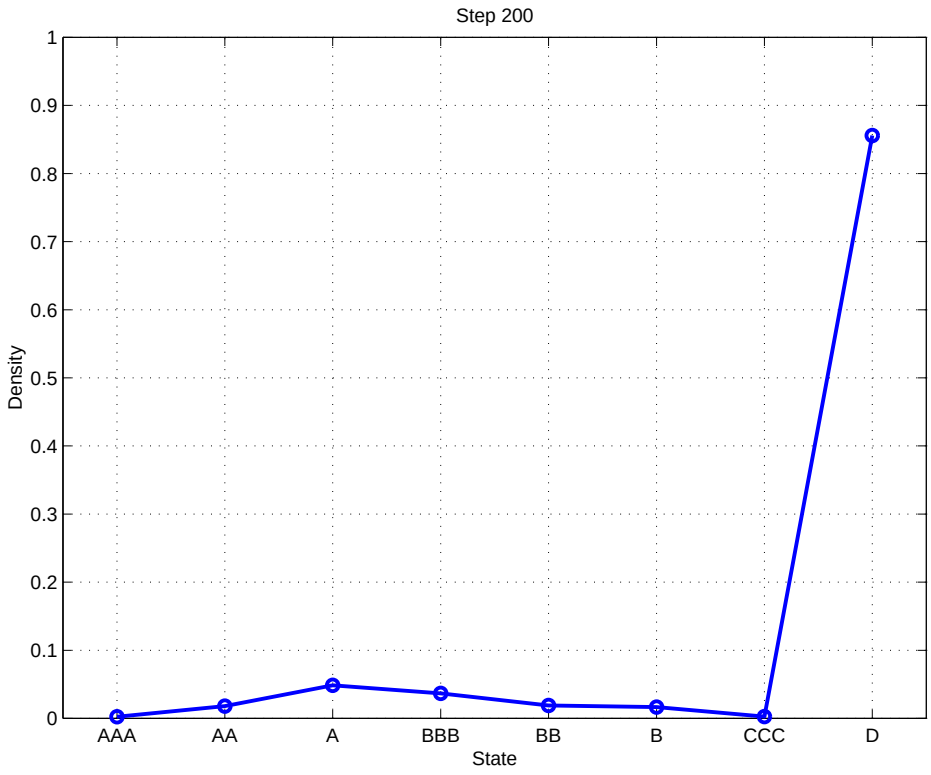
Vidimo da je, prema očekivanom 1 najveća svojstvena vrijednost. Prva sljedeća svojstvena vrijednost je oko 0.9882, što je vrlo blizu 1, i koja ukazuje da će konvergencija prema ravnotežnom stanju biti vrlo spora. Njen svojstveni vektor, osim zadnje komponente, ima najveće komponente u 3. i 4. koordinati. Zbog toga 3. i 4. koordinate od $u^{(k)}$ najsporije padaju.



Slika 23: Grafički prikaz vektora gustoće $u^{(50)}$.

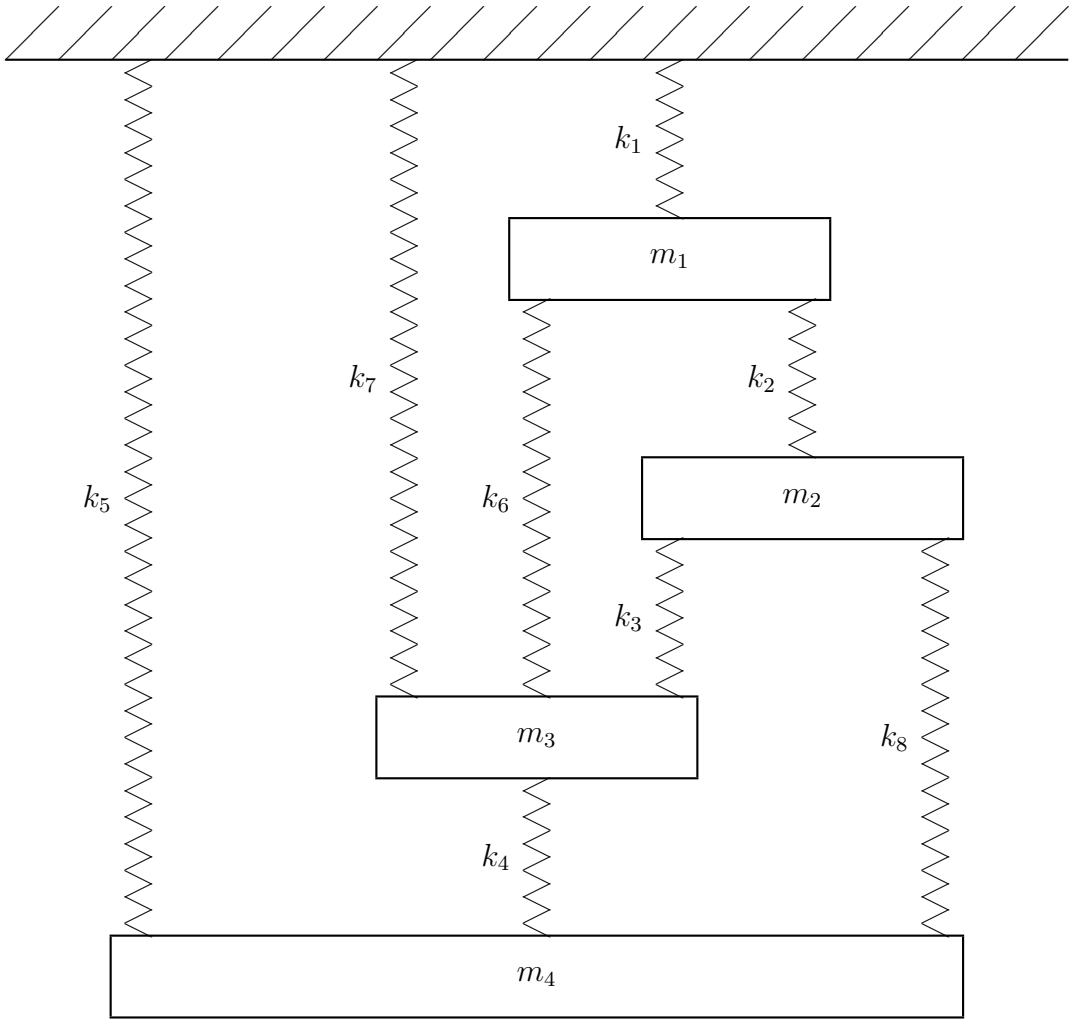


Slika 24: Grafički prikaz vektora gustoće $u^{(100)}$.



Slika 25: Grafički prikaz vektora gustoće $u^{(200)}$.

Primjer 3.2 *Promatramo fizikalni sistem koji se sastoji od tijela različitih masa, povezanih elastičnim oprugama, kao što je prikazano na slici 26. Problem je pronaći slobodne oscilacije ovog sistema. U ovom konkretnom prim-*



Slika 26: Sistem masa povezanih elastičnim oprugama.

jeru imamo četiri tijela masa m_i $i = 1, 2, 3, 4$, i osam opruga krutosti k_l $l = 1, \ldots, 8$. Dalje definirat ćemo sljedeće matrice:

$$M = \begin{bmatrix} m_1 & 0 & 0 & 0 \\ 0 & m_2 & 0 & 0 \\ 0 & 0 & m_3 & 0 \\ 0 & 0 & 0 & m_4 \end{bmatrix},$$

$$K = \begin{bmatrix} k_1+k_2+k_6 & -k_2 & -k_6 & 0 \\ -k_2 & k_2+k_3+k_8 & -k_3 & -k_8 \\ -k_6 & -k_3 & k_3+k_4+k_6+k_7 & -k_4 \\ 0 & -k_8 & -k_4 & k_4+k_5+k_8 \end{bmatrix},$$

pri čemu je u matrici K prikazana interakcija među masama. i-ti redak odgovara i-toj masi, a njegov j-ti stupac odgovara odnosu i-te mase sa j-tom. Na svakom dijagonalnom elementu na poziciji (i,i) nalazi se zbroj krutosti svih opruga vezanih za i-tu masu. Na (i,j)-toj poziciji nalazi se $-k_l$ ukoliko l-ta opruga povezuje i-tu i j-tu masu, ili 0 ako te dvije mase nisu povezane oprugom. Na kraju definirat ćemo vektor

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix},$$

kod kojeg x_i predstavlja vertikalni položaj i-te mase. Iz fizikalnih zakona, položaj masa može se opisati sistemom diferencijalnih jednadžbi

$$\ddot{x} = -M^{-1}Kx.$$

Ako pretpostavimo da je rješenje oblika

$$x = x_0 e^{i\phi t},$$

što je standardni postupak kod rješavanja sistema diferencijalnih jednadžbi, tada za drugu vremensku derivaciju imamo

$$\ddot{x} = -\phi^2 x_0 e^{i\phi t} = -M^{-1}Kx_0 e^{i\phi t}.$$

Sređivanjem dobivamo

$$M^{-1}Kx_0 = \phi^2 x_0,$$

što se svodi na rješavanje problema traženja svojstvenih vrijednosti i svojstvenih vektora matrice $M^{-1}K$. U ovom slučaju svojstvena vrijednost je oblika ϕ^2. Nadalje, možemo uočiti da je matrica K simetrična, ali produkt $M^{-1}K$ gubi to svojstvo. Matrica M je dijagonalna sa pozitivnom dijagonalom, zato je dobro definirana matrica $M^{\frac{1}{2}} = diag(m_1^{\frac{1}{2}}, \ldots, m_4^{\frac{1}{2}})$. Množenjem produkta $M^{-1}K$ matricom $M^{\frac{1}{2}}$ slijeva i matricom $M^{-\frac{1}{2}}$ zdesna, dobit ćemo matricu koja je slična matrici $M^{-1}K$, i oblika

$$A = M^{\frac{1}{2}}(M^{-1}K)M^{-\frac{1}{2}} = M^{-\frac{1}{2}}KM^{-\frac{1}{2}},$$

što pokazuje da je ta matrica i simetrična, i imat će iste svojstvene vrijednosti kao i polazna matrica.

Rješavanju problema svojstvenih vrijednosti simetričnih matrica posvetit ćemo dosta pažnje, zato je važno pokazati da se taj problem često pojavljuje i u praksi.

3.2 Spektralna dekompozicija

Definicija 3.1 *Neka je $A \in \mathbb{C}^{n\times n}$. Skalar $\lambda \in \mathbb{C}$ zove se* **svojstvena vrijednost** *matrice A, ako postoji vektor $x \in \mathbb{C}^n$, $x \neq 0$ takav da je*

$$Ax = \lambda x.$$

Takav vektor x zove se **svojstveni vektor** *od A, koji pripada svojstvenoj vrijednosti λ.*

Napomena 3.1 *Ukoliko za matricu $A = [a_1 \ \dots \ a_n]$ možemo napisati da je $A = SDS^{-1}$, za neku regularnu matricu $S = [s_1 \ \dots \ s_n]$, i $D = diag(d_1, \dots, d_n)$ dijagonalnu matricu tada vrijedi:*

$$AS = DS \quad \Rightarrow \quad As_i = d_i s_i \ i = 1, \dots, n.$$

Dakle, u tom slučaju dijagonalni elementi matrice D predstavljaju svojstvene vrijednosti matrice A, a stupci matrice S predstavljaju svojstvene vektore matrice A. Za rastav $A = SDS^{-1}$ kažemo da je ***spektralna dekompozicija matrice*** *A.*

Definicija 3.2 *Matrica $A \in \mathbb{C}^{n\times n}$ je*

- **normalna** *ako je $A^*A = AA^*$,*
- **hermitska** *ako je $A^* = A$.*

Sljedeći rezultati nam govore o spektralnim dekompozicijama specijalnih matrica.

- Ako je $A \in \mathbb{C}^{n\times n}$ normalna matrica, onda postoji unitarna matrica $U \in \mathbb{C}^{n\times n}$ i dijagonalna matrica $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, takve da je

$$A = U\Lambda U^*.$$

- Ako je $A \in \mathbb{C}^{n\times n}$ hermitska matrica, onda postoji unitarna matrica $U \in \mathbb{C}^{n\times n}$ i dijagonalna matrica $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, pri čemu su $\lambda_i \in \mathbb{R}$ za $i = 1, \dots, n$, takve da je

$$A = U\Lambda U^*.$$

Napomena 3.2 *Analogni rezultati vrijede i za simetrične matrice, samo što je tada matrica U ortogonalna.*

3.3 Spektralna dekompozicija simetričnih matrica

Ovdje ćemo navesti neka osnovna svojstva spektralne dekompozicije simetričnih matrica.

- **(Courant–Fischerov minimax teorem)** Ako je $A \in \mathbb{R}^{n\times n}$ simetrična matrica, i ako njene svojstvene vrijednosti poredamo u nerastućem poretku
$$\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_{n-1}(A) \geq \lambda_n(A),$$
tada vrijedi
$$\lambda_k(A) = \max_{\dim(S)=k} \min_{0\neq y\in S} \frac{y^T Ay}{y^T y}$$
za $k = 1, \ldots, n$. Specijalno je
$$\lambda_{max}(A) = \lambda_1(A) = \max_{0\neq y\in\mathbb{R}^n} \frac{y^T Ay}{y^T y},$$
i
$$\lambda_{min}(A) = \lambda_n(A) = \min_{0\neq y\in\mathbb{R}^n} \frac{y^T Ay}{y^T y}.$$

- **(Weylove nejednakosti)** Ako su A i $A+E$ simetrične $n\times n$ matrice, tada uz iste oznake kao u prethodnom teoremu
$$\lambda_k(A) + \lambda_n(E) \leq \lambda_k(A+E) \leq \lambda_k(A) + \lambda_1(E), \quad k = 1, \ldots, n.$$
Odavde možemo zaključiti da je
$$\max_{i=1,\ldots,n} |\lambda_k(A+E) - \lambda_k(A)| \leq \|E\|_2.$$

- **(Wielandt–Hoffmanov teorem)** Ako su A i $A+E$ simetrične $n\times n$ matrice, tada
$$\sum_{i=1}^{n} [\lambda_i(A+E) - \lambda_i(A)]^2 \leq \|E\|_F^2.$$

- **(Cauchyev teorem ispreplitanja)** Ako sa A_r označimo vodeću $r\times r$ glavnu podmatricu $n\times n$ simetrične matrice A, tada za $r = 1, \ldots, n-1$ vrijede sljedeća svojstva ispreplitanja:
$$\lambda_{j+1}(A_{r+1}) \leq \lambda_j(A_r) \leq \lambda_j(A_{r+1}), \quad j = 1, \ldots, r,$$
i
$$\lambda_{j+n-r}(A) \leq \lambda_j(A_r) \leq \lambda_j(A), \quad j = 1, \ldots, r.$$

3.4 Metode numeričkog računanja spektra

Ako je $A \in \mathbb{R}^{n\times n}$ simetrična matrica, onda tražimo niz ortogonalnih matrica U_1, U_2,... takvih da

$$U_k^T \cdots U_2^T U_1^T A U_1 U_2 \cdots U_k \to \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \quad \text{kad } k \to \infty.$$

Uočimo da se u ovakvim transformacijama čuva simetričnost matrice.

3.4.1 Jacobijeva metoda

Ideja koja stoji iza Jacobijeve metode je sistematsko smanjivanje veličine

$$S(A) = \sqrt{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}^2},$$

koju još nazivamo **normom vandijagonalnih elemenata**.

Najprikladniji izbor za matrice U_i su **Jacobijeve rotacije** $R_i = R_i(p_i, q_i; \phi_i)$. Definirajmo niz simetričnih matrica $A^{(i)} \in \mathbb{R}^{n\times n}$, takav da je,

$$A^{(0)} = A, \qquad A^{(i)} = R_i^T A^{(i-1)} R_i.$$

Za svaki i, biraju se

- pivotni indeksi p_i i q_i, ovisno o **pivotnoj strategiji**,
- kut ϕ_i, takav da poništi elemente sa indeksima p_i i q_i, tj

$$a_{p_i,q_i}^{(i)} = a_{q_i,p_i}^{(i)} = 0.$$

Kada se raspiše djelovanje Jacobijevih rotacija u i-tom koraku, onda se vidi da se mijenjaju samo p_i-ti i q_i-ti stupac i redak. Odnosno

$$\begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & c_i & & -s_i & \\ & & & 1 & & \\ & & s_i & & c_i & \\ & & & & & 1 \end{bmatrix} \cdot \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \end{bmatrix} \cdot \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & c_i & & s_i & \\ & & & 1 & & \\ & & -s_i & & c_i & \\ & & & & & 1 \end{bmatrix} = \begin{bmatrix} * & * & \bullet & * & \bullet & * \\ * & * & \bullet & * & \bullet & * \\ \bullet & \bullet & \bullet & \bullet & 0 & \bullet \\ * & * & \bullet & * & \bullet & * \\ \bullet & \bullet & 0 & \bullet & \bullet & \bullet \\ * & * & \bullet & * & \bullet & * \end{bmatrix}$$

pri čemu su sa $*$ označene stare vrijednosti, a sa $\bullet$ nove.

Ako promatramo samo pivotne indekse, tada imamo

$$\begin{bmatrix} c_i & -s_i \\ s_i & c_i \end{bmatrix} \begin{bmatrix} a_{p_i,p_i}^{(i-1)} & a_{p_i,q_i}^{(i-1)} \\ a_{p_i,q_i}^{(i-1)} & a_{q_i,q_i}^{(i-1)} \end{bmatrix} \begin{bmatrix} c_i & s_i \\ -s_i & c_i \end{bmatrix} = \begin{bmatrix} a_{p_i,p_i}^{(i)} & a_{p_i,q_i}^{(i)} \\ a_{p_i,q_i}^{(i)} & a_{q_i,q_i}^{(i)} \end{bmatrix}. \tag{46}$$

uz uvjet $a_{p_i,q_i}^{(i)} = a_{q_i,p_i}^{(i)} = 0$. Budući da se radi o ortogonalnim transformacijama Frobeniusova norma je sačuvana, odnosno $\|A^{(i)}\|_F = \|A^{(i-1)}\|_F$. Zato imamo

$$[a_{p_i,p_i}^{(i)}]^2 + [a_{q_i,q_i}^{(i)}]^2 = [a_{p_i,p_i}^{(i-1)}]^2 + [a_{q_i,q_i}^{(i-1)}]^2 + 2[a_{p_i,q_i}^{(i-1)}]^2,$$

i

$$\begin{aligned} S(A^{(i)})^2 &= \|A^{(i)}\|_F^2 - \sum_{j=1}^{n} [a_{j,j}^{(i)}]^2 = \\ &= \|A^{(i-1)}\|_F^2 - \sum_{j=1}^{n} [a_{j,j}^{(i-1)}]^2 + ([a_{p_i,p_i}^{(i-1)}]^2 + [a_{q_i,q_i}^{(i-1)}]^2 - [a_{p_i,p_i}^{(i)}]^2 - [a_{q_i,q_i}^{(i)}]^2) = \\ &= S(A^{(i-1)})^2 - 2[a_{p_i,q_i}^{(i-1)}]^2. \end{aligned}$$

Smanjivanjem norme vandijagonalnih elemenata niz $A^{(i)}$ se svakim Jacobijevim korakom približava dijagonalnoj matrici.

Određivanje parametara Jacobijeve metode

1. Ako želimo dijagonalizirati matricu u jednakosti (46), tada imamo

$$0 = a_{p_i,q_i}^{(i)} = a_{p_i,q_i}^{(i-1)}(c_i^2 - s_i^2) + (a_{p_i,p_i}^{(i-1)} - a_{q_i,q_i}^{(i-1)})c_i s_i. \tag{47}$$

Ako je već $a_{p_i,q_i}^{(i-1)} = 0$, tada stavljamo $c_i = 1$ i $s_i = 0$, odnosno $R_i = I$. Inače podijelimo jednakost (47) sa $-c_i^2 a_{p_i,q_i}^{(i-1)}$, pri čemu dobivamo

$$t_i^2 + 2\tau_i t_i - 1 = 0,$$

gdje su

$$\tau_i = \frac{a_{q_i,q_i}^{(i-1)} - a_{p_i,p_i}^{(i-1)}}{2a_{p_i,q_i}^{(i-1)}}, \quad t_i = \frac{s_i}{c_i} = \text{tg } \phi_i.$$

Dakle, dobili smo kvadratnu jednadžbu čije rješenje je tanges kuta kojeg tražimo.

Rješenja jednadžbe (47) dana su sa

$$t_i = -\tau_i \pm \sqrt{\tau_i^2 + 1}.$$

Zbog stabilnosti računa uzimamo rješenje koje je manje po apsolutnoj vrijednosti, i ono je jednako

$$t_i = -\text{sign}(\tau_i)\left(|\tau_i| - \sqrt{\tau_i^2+1}\right) = \frac{\text{sign}(\tau_i)}{|\tau_i| + \sqrt{\tau_i^2+1}},$$

a c_i i s_i možemo izračunati iz formula

$$c_i = \frac{1}{\sqrt{1+t_i^2}}, \qquad s_i = t_i c_i.$$

Odabirom manjeg rješenja t_i osigurali smo da $|\phi_i| \leq \pi/4$, i minimizirali smo razliku $\|A^{(i)} - A^{(i-1)}\|_F$.

2. Kod izračunavanja elementa $a^{(i)}_{p_i,p_i}$, iz (46) slijedi

$$a^{(i)}_{p_i,p_i} = c_i^2 a^{(i-1)}_{p_i,p_i} - 2c_i s_i a^{(i-1)}_{p_i,q_i} + s_i^2 a^{(i-1)}_{q_i,q_i}. \tag{48}$$

Iz (47) možemo izlučiti izraz

$$a^{(i-1)}_{q_i,q_i} = a^{(i-1)}_{p_i,p_i} + \frac{c_i^2 - s_i^2}{c_i s_i} a^{(i-1)}_{p_i,q_i},$$

i uvrstiti ga u (48), čime dobivamo

$$a^{(i)}_{p_i,p_i} = a^{(i-1)}_{p_i,p_i} - t_i a^{(i-1)}_{p_i,q_i}.$$

Slično postupamo i kod računanja elementa $a^{(i)}_{q_i,q_i}$. Iz (46) slijedi

$$a^{(i)}_{q_i,q_i} = s_i^2 a^{(i-1)}_{p_i,p_i} + 2c_i s_i a^{(i-1)}_{p_i,q_i} + c_i^2 a^{(i-1)}_{q_i,q_i}. \tag{49}$$

Iz (47) možemo izlučiti izraz

$$a^{(i-1)}_{p_i,p_i} = a^{(i-1)}_{q_i,q_i} - \frac{c_i^2 - s_i^2}{c_i s_i} a^{(i-1)}_{p_i,q_i},$$

i uvrstiti ga u (49), čime dobivamo

$$a^{(i)}_{q_i,q_i} = a^{(i-1)}_{q_i,q_i} + t_i a^{(i-1)}_{p_i,q_i}.$$

3. Jedino što se još mijenja u matrici $A^{(i)}$ su elementi u pivotnim recima i stupcima. Izračunavanje ostalih elemenata u p_i-tom i q_i-tom stupcu glasi

$$a^{(i)}_{k,p_i} = c_i a^{(i-1)}_{k,p_i} - s_i a^{(i-1)}_{k,q_i} \qquad a^{(i)}_{k,q_i} = s_i a^{(i-1)}_{k,p_i} + c_i a^{(i-1)}_{k,q_i},$$

za $k = 1, \ldots, n$, $k \neq p_i$, $k \neq q_i$, a ostalih elemenata u p_i-tom i q_i-tom retku

$$a^{(i)}_{p_i,k} = c_i a^{(i-1)}_{p_i,k} - s_i a^{(i-1)}_{q_i,k} \qquad a^{(i)}_{q_i,k} = s_i a^{(i-1)}_{p_i,k} + c_i a^{(i-1)}_{q_i,k},$$

za $k = 1, \ldots, n$, $k \neq p_i$, $k \neq q_i$.

4. Preostalo je još objasniti izbor pivotnih indeksa p_i i q_i u svakom koraku algoritma, tj. odabir pivotne strategije. Postoje razne pivotne strategije, od kojih ćemo mi spomenuti samo dvije:

   - **Klasična Jacobijeva metoda** u kojoj se p_i i q_i biraju tako da je $|a_{p_i,q_i}^{(i-1)}|$ maksimalan, tj.

   $$|a_{p_i,q_i}^{(i-1)}| = \max_{j,k=1,\ldots,n,\ j\neq k} |a_{jk}^{(i-1)}|.$$

   - **Jacobijeva metoda sa cikličkom strategijom po recima** u kojoj se ciklusi se sastoje od:

   $$\begin{aligned}(p_i, q_i) \;=\; & (1,2),(1,3),\ldots,(1,n),(2,3),(2,4),\ldots,(2,n),\ldots \\ & \ldots,(n-2,n-1),(n-2,n),(n-1,n),\end{aligned}$$

   odnosno u jednom ciklusu se poništavaju vandijagonalni elementi u sljedećem poretku:

$*$	1	2	3	4	5
1	$*$	6	7	8	9
2	6	$*$	10	11	12
3	7	10	$*$	13	14
4	8	11	13	$*$	15
5	9	12	14	15	$*$

5. Na kraju treba odrediti i uvijet zaustavljanja iteracija. Iteracije se zaustavljaju kada

   $$S(A^{(i)}) \leq tol\|A\|_F,$$

   za neku veličinu tolerancije $tol > 0$, $tol = O(u)$.

Sada sve ovo konačno možemo sklopiti u algoritam, kojeg ćemo napisati za obje pivotne strategije.

Algoritam 3.1 (Klasična Jacobijeva metoda) *Za danu simetričnu matricu $A \in \mathbb{R}^{n\times n}$ i toleranciju $tol > 0$, algoritam matricu A transformira u matricu U^TAU, gdje je U ortogonalna i $S(U^TAU) \leq tol\|A\|_F$.*

while $S(A) > tol\|A\|_F$
begin
Odaberite indekse p i q sa $1 \leq p < q \leq n$, tako da je

$$|a_{p,q}| = \max_{i\neq j} |a_{ij}|;$$

if $a_{pq} \neq 0$
begin
$\tau = \dfrac{a_{qq} - a_{pp}}{2a_{pq}};$
$t = \dfrac{sign(\tau)}{|\tau| + \sqrt{1+\tau^2}};$
$c = \dfrac{1}{\sqrt{1+t^2}};$
$s = tc;$
end
else
begin
$c = 1;$
$s = 0;$
end
$app = a_{pp};\ apq = a_{pq};\ aqq = a_{qq};$
$app = app - t \cdot apq;$
$aqq = aqq + t \cdot apq;$
for $k = 1, \ldots, n$
begin
$pom = a_{kp};$
$a_{kp} = c \cdot pom - s \cdot a_{kq};$
$a_{kq} = s \cdot pom + c \cdot a_{kq};$
$a_{pk} = a_{kp};\ a_{qk} = a_{kq};$
end
$a_{pq} = 0;\ a_{qp} = 0;$
$a_{pp} = app;\ a_{qq} = aqq;$
end

Algoritam 3.2 (Jacobijeva metoda sa cikličkom strategijom po recimaa) *Za danu simetričnu matricu $A \in \mathbb{R}^{n\times n}$ i toleranciju $tol > 0$, algoritam matricu A transformira u matricu U^TAU, gdje je U ortogonalna i $S(U^TAU) \leq tol\|A\|_F$.*

while $S(A) > tol\|A\|_F$
begin
for $p = 1, \ldots, n-1$
begin
for $q = p+1, \ldots, n$
begin
if $a_{pq} \neq 0$
begin
$\tau = \dfrac{a_{qq} - a_{pp}}{2a_{pq}}$;
$t = \dfrac{sign(\tau)}{|\tau| + \sqrt{1+\tau^2}}$;
$c = \dfrac{1}{\sqrt{1+t^2}}$;
$s = tc$;
end
else
begin
$c = 1$;
$s = 0$;
end
$app = a_{pp};\ apq = a_{pq};\ aqq = a_{qq}$;
$app = app - t \cdot apq$;
$aqq = aqq + t \cdot apq$;
for $k = 1, \ldots, n$
begin
$pom = a_{kp}$;
$a_{kp} = c \cdot pom - s \cdot a_{kq}$;
$a_{kq} = s \cdot pom + c \cdot a_{kq}$;
$a_{pk} = a_{kp};\ a_{qk} = a_{kq}$;
end
$a_{pq} = 0;\ a_{qp} = 0$;
$a_{pp} = app;\ a_{qq} = aqq$;
end
end
end

Glavno svojstvo Jacobijeve metode daje sljedeći rezultat.

- Za $A \in \mathbb{R}^{n\times n}$, $A^T = A$, Jacobijeva metoda je globalno konvergentna, tj.
$$\lim_{i\to\infty} S(A^{(i)}) = 0,$$
i
$$\lim_{i\to\infty} A^{(i)} = \Lambda,$$
pri čemu je $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, a λ_i, $i = 1, \ldots, n$ su svojstvene vrijednosti od A. Matrica $U \in \mathbb{R}^{n\times n}$, $U = R_1 R_2 \cdots$ je unitarna, i njeni stupci su svojstveni vektori matrice A.

Zadatak 3.1 *Testirajmo obje Jacobijeve metode na primjeru Risove matrice. Risova matrica $A = [a_{ij}] \in \mathbb{R}^{n\times n}$ je simetrična, i definira se kao*

$$a_{ij} = \frac{1}{2(n-i-j+1.5)}, \qquad i, j = 1, \ldots, n.$$

Poznato je da svojstvene vrijednosti tvore nakupine oko $-\pi/2$ i $\pi/2$.

Napomena 3.3 *Testirajmo zadatak za $n = 10$. Norme vandijagonalnih elemenata za obje varijante Jacobijeve metode su prikazane na slikama 27 i 28.*

Vidimo da vandijagonalni elementi vrlo brzo padaju u nulu, tj. da $S(A^{(i)})$ teži ka nuli. Osim toga Jacobijeva metoda sa cikličkom strategijom izvrši 45 iteracija po ciklusu, to znači da je ukupno izvršila 270 iteracija u 6 ciklusa, što je puno više od 162 koliko je trebalo Klasičnoj Jacobijevoj metode. Međutim, u prvom slučaju smo postigli puno manju normu vandijagonalnih elemenata. Provjerimo sada točnost dobivenih faktorizacija.

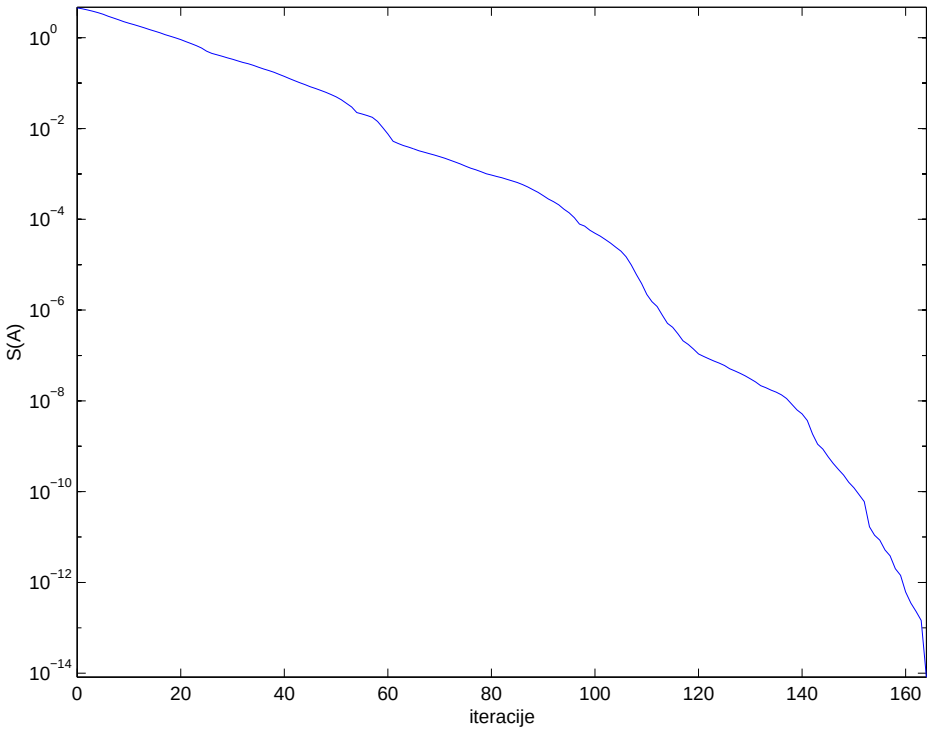
Za klasičnu Jacobijevu metodu imamo

$$\|A - U_K \Lambda_K U_K^T\|_2 \le 1.4256 \cdot 10^{-15} \|A\|_2, \qquad \|U_K^T U_K - I\|_2 \le 1.6953 \cdot 10^{-15},$$

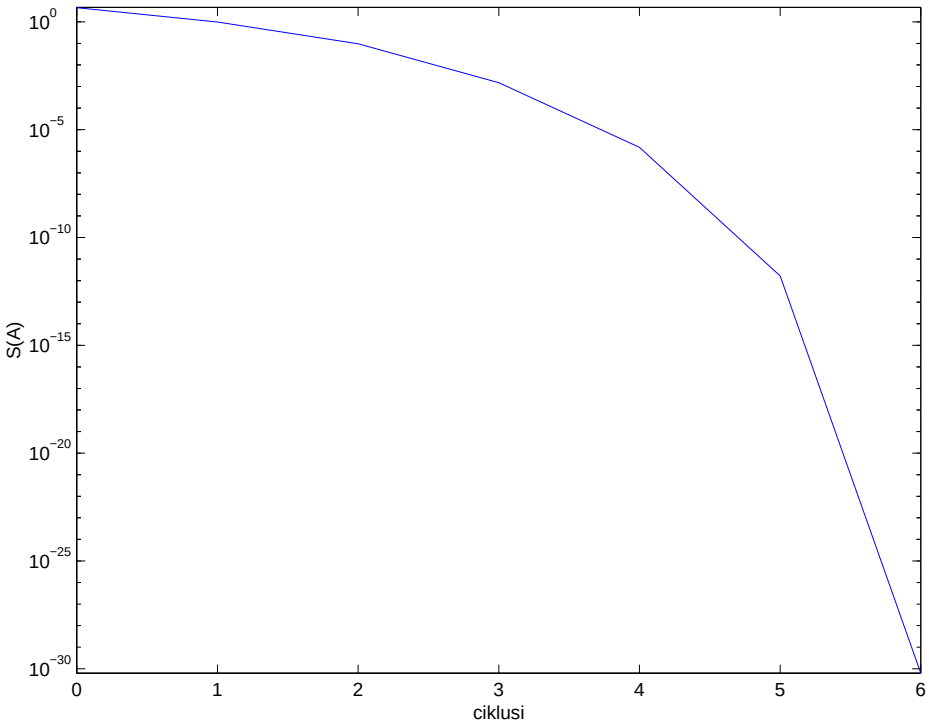
a za Jacobijevu metodu sa cikličkom strategijom

$$\|A - U_C \Lambda_C U_C^T\|_2 \le 3.8901 \cdot 10^{-15} \|A\|_2, \qquad \|U_C^T U_C - I\|_2 \le 3.7673 \cdot 10^{-15}.$$

Obje metode su dale vrlo točne rezultate, ali vidimo da je ciklička metoda malo manje točna od klasične, zbog većeg broja iteracija.



Slika 27: Norme vandijagonalnih elemenata po iteracijama Klasične Jacobijeve metode.



Slika 28: Norme vandijagonalnih elemenata po ciklusima Jacobijeve metode sa cikličkom strategijom po recima.

Primjer 3.3 *Želimo riješiti problem iz primjera 26 za konkretne mase i opruge. Neka su vrijednosti masa i krutosti opruga dane u sljedećim tablicama.*

i	1	2	3	4
$\mathbf{m_i}$	2	5	3	6

i	1	2	3	4	5	6	7	8
$\mathbf{k_i}$	10	9	8	7	6	5	5	5

Tada su matrice M i K dane sa

$$M = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 6 \end{bmatrix}, \qquad K = \begin{bmatrix} 24 & -9 & -5 & 0 \\ -9 & 22 & -8 & -5 \\ -5 & -8 & 25 & -7 \\ 0 & -5 & -7 & 18 \end{bmatrix}.$$

Podsjetimo se, mi želimo riješiti problem svojstvenih vrijednosti

$$M^{-1}Kx_0 = \phi^2 x_0.$$

Pomnožimo li gornju jednadžbu sa $M^{\frac{1}{2}}$, dobit ćemo

$$\begin{aligned} (M^{-\frac{1}{2}}KM^{-\frac{1}{2}})(M^{\frac{1}{2}}x_0) &= \phi^2(M^{\frac{1}{2}}x_0) \\ Au &= \lambda u \end{aligned}$$

pri čemu su

$$A = M^{-\frac{1}{2}}KM^{-\frac{1}{2}}, \quad u = M^{\frac{1}{2}}x_0, \quad \lambda = \phi^2.$$

Prema tome, u našem primjeru je

$$A = \begin{bmatrix} 12.0000 & -2.8460 & -2.0412 & 0 \\ -2.8460 & 4.4000 & -2.0656 & -0.9129 \\ -2.0412 & -2.0656 & 8.3333 & -1.6499 \\ 0 & -0.9129 & -1.6499 & 3.0000 \end{bmatrix},$$

i mi tražimo svojstvene vrijednosti i vektore matrice A.

Primjenom jedne od Jacobijevih metoda dobit ćemo matrice U i Λ, takve da je

$$A \approx U\Lambda U^T,$$

pri čemu su

$$U = \begin{bmatrix} 0.9228 & 0.2354 & 0.1835 & -0.2438 \\ -0.2301 & 0.6226 & -0.4426 & -0.6029 \\ -0.3015 & 0.3894 & 0.8626 & -0.1161 \\ 0.0682 & 0.6367 & -0.1625 & 0.7507 \end{bmatrix},$$

$$\Lambda = \begin{bmatrix} 13.3767 & 0 & 0 & 0 \\ 0 & 1.0983 & 0 & 0 \\ 0 & 0 & 9.2700 & 0 \\ 0 & 0 & 0 & 3.9883 \end{bmatrix}.$$

Dakle, u našem primjeru imamo 4 rješenja koja predstavljaju putanje slobodnih oscilacija, oblika

$$x_i = M^{-\frac{1}{2}} u_i e^{i\sqrt{\lambda_i}t}, \qquad i = 1, 2, 3, 4,$$

gdje je λ_i i-ta svojstvena vrijednost od A, a u_i njen svojstveni vektor, odnosno $\Lambda = diag(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, i $U = [u_1\ u_2\ u_3\ u_4]$. Na kraju dobivamo

$$x_1 = \begin{bmatrix} 0.6525 \\ -0.1029 \\ -0.1741 \\ 0.0278 \end{bmatrix} e^{3.6574it}, \quad x_2 = \begin{bmatrix} 0.1298 \\ -0.1979 \\ 0.4980 \\ -0.0664 \end{bmatrix} e^{3.0447it},$$

$$x_3 = \begin{bmatrix} -0.1724 \\ -0.2696 \\ -0.0670 \\ 0.3065 \end{bmatrix} e^{1.9971it} \quad x_4 = \begin{bmatrix} 0.1665 \\ 0.2784 \\ 0.2248 \\ 0.2599 \end{bmatrix} e^{1.0480it}.$$

3.4.2 Tridijagonalizacija simetrične matrice

Jedan način na koji se može ubrzati iterativna metoda za računanje svojstvenih vrijednosti simetričnih matrica, je ta da se matrica prvo svede na tridijagonalni oblik, relacijama ortogonalne sličnosti koje u svakom koraku čuvaju simetričnost i spektar. Nakon toga se koriste vrlo brze metode, specijalno dizajnirane za računanje svojstvenih vrijednosti tridijagonalne matrice.

Tridijagonalizacija se provodi u konačno mnogo koraka, poništavanjem elemenata u matrici. Općenit proces, za simetričnu matricu $A \in \mathbb{R}^{n\times n}$, je sličan kao kod Jacobijeve metode:

$$A^{(0)} = A, \qquad A^{(i)} = P_i A^{(i-1)} P_i^T,\ i = 1, \ldots, k,$$

gdje je $P_k \in \mathbb{R}^{n\times n}$ ortogonalna matrica. Tada je $[A^{(i)}]^T = A^{(i)}$, $i = 0, \ldots, k$, tj. sve matrice u procesu su ortogonalno slične i simetrične. Na kraju, rezultat je $A^{(k)} = T_n$, kod kojeg je T_n oblika

$$T_n = \begin{bmatrix} d_1 & e_2 & & & \\ e_2 & d_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & e_n \\ & & & e_n & d_n \end{bmatrix}.$$

Kao i kod QR faktorizacije, tridijagonalizacija se može napraviti na dva načina: pomoću Householderovih reflektora, i pomoću Givensovih rotacija.

- **Tridijagonalizacija pomoću Householderovih reflektora**
  Matrice P_i biramo kao Householderove reflektore. Pretpostavimo da je matrica $A^{(i-1)}$ oblika

$$A^{(i-1)} = \left[\begin{array}{c|c|c} & & 0 \\ \multicolumn{2}{c|}{T_i} & \\ \hline & & a_i^T \\ \hline 0 & a_i & B_i \\ & & \end{array}\right],$$

  gdje je T_i $i \times i$ tridijagonalna simetrična matrica, B_i je $(n-i) \times (n-i)$ simetrična matrica, a a_i je $(n-i)$-dimenzionalni vektor, kojem moramo poništiti sve elemente osim prvog.

  Dakle, tražimo Householderov reflektor $\bar{P}_i \in \mathbb{R}^{(n-i)\times(n-i)}$, takav da je

$$\bar{P}_i a_i = -\alpha_i e_1 = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

  Već znamo da je on tada oblika

$$\bar{P}_i = I - \frac{1}{\gamma_i} v_i v_i^T,$$

  uz vrijednosti

$$\begin{aligned} \alpha_i &= \begin{cases} \|a_i\|_2, & (a_i)_1 = a_{i+1,i}^{(i)} \geq 0 \\ -\|a_i\|_2, & (a_i)_1 = a_{i+1,i}^{(i)} < 0 \end{cases} \\ v_i &= a_i + \alpha_i e_1 \in \mathbb{R}^{n-i} \\ \gamma_i &= \|a_i\|_2[\|a_i\|_2 + |(a_i)_1|\ ]. \end{aligned}$$

  Na kraju je

$$P_i = \begin{bmatrix} I_i & 0 \\ 0 & \bar{P}_i \end{bmatrix}.$$

Također vrijedi da je

$$a_i^T \bar{P}_i^T = -\alpha_i e_1^T,$$

odnosno ono što napravimo i-tom stupcu od $A^{(i-1)}$ množeći ga sa P_i slijeva, napravimo isto i i-tom retku množeći ga sa P_i^T zdesna. Važno je primijetiti da množenje matrice $A^{(i-1)}$ sa P_i slijeva neće promijeniti i-ti redak odnosno a_i^T, a i obrnuto, množenje sa P_i^T zdesna neće promijeniti i-i stupac odnosno a_i. Zato je sve jedno kojim poretkom izvršavamo ta množenja. Na kraju i-tog koraka još moramo ažurirati ostatak matrice B_i sa

$$\bar{P}_i B_i \bar{P}_i^T = B_i - v_i w_i^T - w_i v_i^T,$$

uz

$$u_i = \frac{1}{\gamma_i} B_i v_i, \quad p_i = \frac{1}{2\gamma_i} v_i^T u_i, \quad w_i = u_i - p_i v_i.$$

Matrica $A^{(i)} = P_i A^{(i-1)} P_i^T$ će tada imati oblik

$$A^{(i)} = \left[\begin{array}{c|c|c} & & 0 \\ T_i & & \\ & & \\ \hline & -\alpha_i & -\alpha_i \quad 0 \quad \cdots \quad \cdots \quad 0 \\ \hline & 0 & \\ 0 & \vdots & \bar{P}_i B_i \bar{P}_i^T \\ & \vdots & \\ & 0 & \end{array}\right] =$$

$$= \left[\begin{array}{c|c|c} & & 0 \\ T_{i+1} & & \\ \hline & & a_{i+1}^T \\ \hline & a_{i+1} & \\ 0 & & B_{i+1} \\ & & \end{array}\right]$$

Sada sve to možemo sklopiti u algoritam.

Algoritam 3.3 *Za danu simetričnu matricu $A \in \mathbb{R}^{n\times n}$ algoritam matricu A transformira u matricu $T = U^T A U$, pomoću Householderovih reflektora, pri čemu je T tridijagonalna. Matrica U je ortogonalna i dobiva se kao produkt $U = P_1 P_2 \cdots P_{n-2}$, gdje je $P_i = I - \frac{1}{\gamma_i} v_i v_i^T$.*

for $j = 1, \ldots, n-2$

begin

$na = \sqrt{\sum_{i=j+1}^{n} a_{ij}^2}$;

if $a_{j+1,j} > 0$ **then**

begin

$\alpha = na$;

end

else

begin

$\alpha = -na$;

end

$\gamma_j = na(na + |a_{j+1,j}|)$;

for $i = 1, \ldots, j$

begin

$v_i^{(j)} = 0$;

end

$v_{j+1}^{(j)} = a_{j+1,j} + \alpha$;

$a_{j+1,j} = -\alpha$;

$a_{j,j+1} = -\alpha$;

for $i = j+2, \ldots, n$

begin

$v_i^{(j)} = a_{ij}$;

$a_{ij} = 0$;

$a_{ji} = 0$;

end;

for $i = j+1, \ldots, n$

begin

$u_i^{(j)} = \frac{1}{\gamma_j} \sum_{k=j+1}^{n} a_{ik} v_k^{(j)}$;

end

$p_j = \frac{1}{2\gamma_j} \sum_{k=j+1}^{n} v_k^{(j)} u_k^{(j)}$;

for $i = j+1, \ldots, n$
begin
$w_i^{(j)} = u_i^{(j)} - p_j v_i^{(j)}$;
end
for $i = j+1, \ldots, n$
begin
$a_{ii} = a_{ii} - 2w_i^{(j)} v_i^{(j)}$;
for $k = i+1, \ldots, n$
begin
$a_{ik} = a_{ik} - w_i^{(j)} v_k^{(j)} - w_k^{(j)} v_i^{(j)}$;
$a_{ki} = a_{ik}$;
end
end
end;
$T = A$;

- **Tridijagonalizacija pomoću Givensovih rotacija**
Matrice P_i biramo kao Givensove rotacije. Pretpostavimo da je matrica $A^{(i-1)}$ oblika

$$A^{(i-1)} = \left[\begin{array}{c|c|ccccc} & & 0 & & & & \\ \multicolumn{2}{c|}{T_k} & \multicolumn{5}{c}{\dotfill} \\ & & a_{k+1,k}^{(i-1)} & \cdots & a_{l,k}^{(i-1)} & 0 & \cdots & 0 \\ \hline & a_{k+1,k}^{(i-1)} & & & & & & \\ & \vdots & & & & & & \\ 0 & a_{l,k}^{(i-1)} & \multicolumn{5}{c}{B_k^{(i-1)}} \\ & 0 & & & & & & \\ & \vdots & & & & & & \\ & 0 & & & & & & \end{array}\right],$$

gdje je T_k $k \times k$ tridijagonalna simetrična matrica, a $B_k^{(i-1)}$ je $(n-k) \times (n-k)$ simetrična matrica. Želimo poništiti element na poziciji (l,k). Tražimo 2-dimenzionalnu Givensovu rotaciju $\bar{P}_i$ takvu da je

$$\bar{P}_i^T \begin{bmatrix} a_{l-1,k}^{(i-1)} \\ a_{lk}^{(i-1)} \end{bmatrix} = \begin{bmatrix} \alpha_i \\ 0 \end{bmatrix}.$$

Znamo da je $\bar{P}_i$ oblika

$$\bar{P}_i = \begin{bmatrix} c_i & -s_i \\ s_i & c_i \end{bmatrix},$$

uz vrijednosti

$$\begin{aligned} \alpha_i &= \sqrt{(a_{l-1,k}^{(i-1)})^2 + (a_{lk}^{(i-1)})^2} \\ a_{lk}^{(i)} &= a_{kl}^{(i)} = 0 \\ c_i &= \frac{a_{l-1,k}^{(i-1)}}{\alpha_i} \\ s_i &= \frac{a_{lk}^{(i-1)}}{\alpha_i} \end{aligned}$$

Na kraju je

$$P_i = \begin{bmatrix} I_{l-2} & 0 & 0 \\ 0 & \bar{P}_i & 0 \\ 0 & 0 & I_{n-l} \end{bmatrix}.$$

Isto vrijedi da je

$$\begin{bmatrix} a_{l-1,k}^{(i-1)} & a_{lk}^{(i-1)} \end{bmatrix} \bar{P}_i = \begin{bmatrix} \alpha_i & 0 \end{bmatrix},$$

odnosno ono što napravimo u k-tom stupcu od $A^{(i-1)}$ množeći ga sa P_i^T slijeva, napravimo isto i k-tom retku množeći ga sa P_i zdesna. Važno je primijetiti da množenje matrice $A^{(i-1)}$ sa P_i^T slijeva neće promijeniti k-ti redak, a i obrnuto, množenje sa P_i zdesna neće promijeniti k-ti stupac. Zato je sve jedno kojim poretkom izvršavamo ta množenja. Na kraju i-tog koraka još moramo ažurirati ostatak matrice $B_k^{(i-1)}$, s time da se kod nje mijenjaju samo $l-1$-i i l-ti redak i stupac, slično kao kod Jacobijeve metode. Zato imamo

$$\begin{aligned} a_{l-1,l-1}^{(i)} &= c_i^2 a_{l-1,l-1}^{(i-1)} + 2c_i s_i a_{l-1,l}^{(i-1)} + s_i^2 a_{ll}^{(i-1)} \\ a_{l-1,l}^{(i)} &= a_{l-1,l}^{(i-1)}(c_i^2 - s_i^2) + (a_{ll}^{(i-1)} - a_{l-1,l-1}^{(i-1)}) c_i s_i \\ a_{ll}^{(i)} &= s_i^2 a_{l-1,l-1}^{(i-1)} - 2c_i s_i a_{l-1,l}^{(i-1)} + c_i^2 a_{ll}^{(i-1)} \\ a_{l-1,j}^{(i)} &= a_{j,l-1}^{(i)} = c_i a_{j,l-1}^{(i-1)} + s_i a_{jl}^{(i-1)}, \qquad j = k+1, \ldots, n,\ j \neq l-1, l \\ a_{lj}^{(i)} &= a_{jl}^{(i)} = -s_i a_{j,l-1}^{(i-1)} + c_i a_{jl}^{(i-1)}, \qquad j = k+1, \ldots, n,\ j \neq l-1, l \end{aligned}$$

Matrica $A^{(i)} = P_i^T A^{(i-1)} P_i$ će tada imati oblik

$$A^{(i)} = \left[\begin{array}{ccc|ccccc} & & & & & 0 & & \\ & T_k & & & & & & \\ \hline & & & a^{(i)}_{k+1,k} & \cdots & a^{(i)}_{l-1,k} & 0 & \cdots & 0 \\ \hline & & a^{(i)}_{k+1,k} & & & & & & \\ & & \vdots & & & & & & \\ 0 & & a^{(i)}_{l-1,k} & & & B^{(i)}_k & & & \\ & & 0 & & & & & & \\ & & \vdots & & & & & & \\ & & 0 & & & & & & \end{array}\right].$$

Važno je još spomenuti pivotnu strategiju, koja je prema gore opisanome oblika

$$\begin{array}{|cccccc|} \hline * & * & 4 & 3 & 2 & 1 \\ * & * & * & 7 & 6 & 5 \\ 4 & * & * & * & 9 & 8 \\ 3 & 7 & * & * & * & 10 \\ 2 & 6 & 9 & * & * & * \\ 1 & 5 & 8 & 10 & * & * \\ \hline \end{array}.$$

Napokon možemo sve to sklopiti u algoritam.

Algoritam 3.4 *Za danu simetričnu matricu $A \in \mathbb{R}^{n\times n}$ algoritam matricu A transformira u matricu $T = U^T AU$, pomoću Givensovih rotacija, pri čemu je T tridijagonalna. Matrica U je ortogonalna i dobiva se kao produkt $U = P_1 P_2 \cdots P_{(n-1)(n-2)/2}$, gdje je P_i Givensova rotacija.*

for $j = 1, \ldots, n-2$
  begin
    for $i = n, n-1, \ldots j+2$
      begin
        if $|a_{ij}| > |a_{i-1,j}|$
          begin
            $\tau = \frac{a_{i-1,j}}{a_{ij}}$;
            $c = \frac{sign(a_{ij})}{\sqrt{1+\tau^2}}$;
            $s = c\tau$;
          end

else

begin

$\tau = \frac{a_{ij}}{a_{i-1,j}}$;

$s = \frac{sign(a_{i-1,j})}{\sqrt{1+\tau^2}}$;

$c = s\tau$;

end

$a_{i-1,j} = \sqrt{a_{i-1,j}^2 + a_{ij}^2}$;

$a_{j,i-1} = a_{i-1,j}$;

$a_{ij} = 0$;

$a_{ji} = 0$;

$pom = a_{i-1,i-1}$;

$pom2 = a_{i-1,i}$;

$a_{i-1,i-1} = c^2 \cdot pom + 2cs \cdot pom2 + s^2 a_{ii}$;

$a_{i-1,i} = -cs \cdot pom + (c^2 - s^2) \cdot pom2 + csa_{ii}$;

$a_{i,i-1} = a_{i-1,i}$;

$a_{ii} = s^2 \cdot pom - 2cs \cdot pom2 + c^2 a_{ii}$;

for $k = j + 1 \ldots n$

begin

if $k \neq i - 1$ **and** $k \neq i$

begin

$pom = a_{k,i-1}$;

$a_{k,i-1} = c \cdot pom + s \cdot a_{ki}$;

$a_{ki} = -s \cdot pom + c \cdot a_{ki}$;

$a_{i-1,k} = a_{k,i-1};\ a_{ik} = a_{ki}$;

end

end

end

end

$T = A$;

Zadatak 3.2 *Riješimo isti problem kao i u zadatku 3.1, samo što ćemo koristiti oba dvije vrste tridijagonalizacije.*

Napomena 3.4 *Prvo izračunajmo tridijagonalni oblik za matricu* $A \in \mathbb{R}^{10\times 10}$ *pomoću Householderovih reflektora. Dobiveni su tridijagonalna matrica* T_H *i matrica* U_H*, za koje vrijedi*

$$\|A - U_H T_H U_H^T\|_2 \leq 1.6935 \cdot 10^{-15}\|A\|_2, \qquad \|U_H^T U_H - I\|_2 \leq 1.1316 \cdot 10^{-15},$$

što je prilično točno. Ako sada izračunamo spektralnu dekompoziciju matrice T_H*, odnosno*

$$T_H = V_H \Lambda_H V_H^T, \qquad W_H = U_H V_H,$$

Tada je

$$A \approx W_H \Lambda_H W_H^T$$

aproksimacija spektralne dekompozicije matrice A. Za nju imamo

$$\|A - W_H \Lambda_H W_H^T\|_2 \leq 2.1678 \cdot 10^{-10}\|A\|_2, \qquad \|W_H^T W_H - I\|_2 \leq 2.1678 \cdot 10^{-10},$$

što baš nije jako točno. (Velika greška se dogodila kod korištenja Ocatve*-ine funkcije eig pri računanju spektralne dekompozicije matrice* T_H *na koju nemamo utjecaja.)*

Napravimo sada isti račun i za tridijagonalizaciju pomoću Givensovih rotacija. Dobit ćemo tridijagonalnu matricu T_G *i matricu* U_G*, za koje vrijedi*

$$\|A - U_G T_G U_G^T\|_2 \leq 2.1404 \cdot 10^{-15}\|A\|_2, \qquad \|U_G^T U_G - I\|_2 \leq 1.2589 \cdot 10^{-15},$$

što je prilično točno. Ako sada izračunamo spektralnu dekompoziciju matrice T_G*, odnosno*

$$T_G = V_G \Lambda_G V_G^T, \qquad W_G = U_G V_G,$$

Tada je

$$A \approx W_G \Lambda_G W_G^T$$

aproksimacija spektralne dekompozicije matrice A. Za nju imamo

$$\|A - W_G \Lambda_G W_G^T\|_2 \leq 3.0129 \cdot 10^{-15}\|A\|_2, \qquad \|W_G^T W_G - I\|_2 \leq 2.3554 \cdot 10^{-15},$$

što je točo prema očekivanju. Ako sada te rezultate usporedimo sa rezultatima klasične Jacobijeve metode, tada vidimo da je Jacobijeva metoda točnija, ali je sporija, jer joj je potrebno više računskih operacija od računanja spektralne dekompozicije pomoću tridijagonalizacije.