Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku
Preddiplomski studij matematike

Slobodan Jelić

GAUSSOVA METODA ZA RJEŠAVANJE SUSTAVA LINEARNIH JEDNADŽBI

Završni rad

Osijek, 2007.

Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku
Preddiplomski studij matematike

Slobodan Jelić

Gaussova metoda za rješavanje sustava linearnih jednadžbi

Završni rad

Voditelj: Prof. dr. sc. Ninoslav Truhar

Osijek, 2007.

Sažetak. U radu ćemo proučavati rješavanje sustava linearnih jednadžbi $Ax = b$ pomoću Gaussovih eliminacija. Ono se provodi LU dekompozicijom matrice sustava $A = LU$ te uzastopnim rješavanjem trokutastih sustava $Ly = b$ i $Ux = y$. Zbog činjenice da za neke regularne matrice LU dekompozicija ne postoji, proučavat ćemo LU dekompoziciju s djelomičnim pivotiranjem.

Ključne riječi: sustav linearnih jednadžbi, aritmetika s pomičnim zarezom, matrica, vektor, Gaussove eliminacije, trokutasti sustav, supstitucija unaprijed, supstitucija unazad, LU dekompozicija, djelomično pivotiranje, permutacijska matrica, permutacijski vektor, povratna analiza greške

Abstract. In this paper we will study a solving of system of linear equations $Ax = b$ by Gaussian elimination method. It includes LU factorization of matrix $A = LU$ and consecutive solving of triangular systems $Ly = b$ and $Ux = y$. Because of existence of regular matrices which don't have LU factorization, we will also study LU factorization with partial pivoting.

Key words: system of linear equations, floating point arithmetic, matrix, vector, Gaussian eliminations, triangular system, forward substitution, backward substitution, LU factorization, partial pivoting, matrix of permutation, vector of permutation, backward error analysis

Sadržaj

1. Uvod

Gotovo da ne postoji područje matematike gdje nije potrebno riješiti sustave linearnih jednadžbi. U praksi se javljaju različiti tipovi sustava koje općenito nije lako riješiti pa je u tu svrhu neizostavna uporaba računala. Kako je računalo ograničenih mogućnosti u smislu aritmetike kojom se služi (što će biti opisano u uvodnom poglavlju) prirodno se javila potreba za proučavanje i poboljšavanje algoritama koji se koriste za rješavanje različitih problema, a koji se izvode na računalima. Nemogućnost izračunavanja egzaktnog rješenja nametnula je potrebu određivanja pogreške dobivene aproksimacije.

Rad je tematski usmjeren na većinu navedenih problema koji su vezani za *Gaussovu metodu*, jednu od najstarijih direktnih metoda za rješavanje sustava linearnih jednadžbi. Vremenom su se javljale i određene modifikacije ove metode koje su svakako predstavljale napredak u odnosu na prethodne. U ovom radu bavit ćemo se dvijema Gaussovim metodama: standardnom Gaussovom metodom eliminacija bez pivotiranja i Gaussovom metodom eliminacija s djelomičnim pivotiranjem. Za stvaranje potrebnog matematičkog aparata na osnovi kojeg se konstruiraju algoritmi, potrebno je uvesti i definirati osnovne pojmove.

Tako je u prvom poglavlju objašnjen način na koji se sustav linearnih jednadžbi može zapisati u odgovarajućem matričnom zapisu. Zbog izvođenja algoritama na računalu, kratko je objašnjena i *aritmetika s pomičnim zarezom (eng. Floating Point Arithmetic)* te neki standardi koji se najčešće koriste. Prvo poglavlje sadrži analizu pogreški zaokruživanja koje se javljaju prilikom računanja skalarnog produkta. U drugom dijelu objašnjene su metode za rješavanje specijalne vrste sustava koji se nazivaju trokutasti sustavi. Metoda Gaussovih eliminacija podrazumijevat će uporabu takvih algoritama pa ih je nemoguće izostaviti. Treći dio uvodi nas u problematiku samog rada. Ovdje je objašnjen postupak Gaussovih eliminacija kao i njegova primjena u računanju *LU dekompozicije* matrice sustava. Detaljno je opisana metoda Gaussovih eliminacija s djelomičnim pivotiranjem kroz sve verzije algoritma. Izložena materija praćena je odgovarajućim teorijskim podlogama koje su navedene u obliku teorema i propozicija. Cilj četvrtog poglavlja je analiza pogreški koje se javljaju tijekom izvođenja prethodno navedenih algoritama.

1.1. Sustav linearnih jednadžbi

Pod pojmom **sustava linearnih jednadžbi**[1] podrazumijevamo skup jednadžbi sljedećeg oblika

$$\begin{array}{ccccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \ldots\ldots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & + & \ldots\ldots & + & a_{2n}x_n & = & b_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & + & \ldots\ldots & + & a_{3n}x_n & = & b_3 \\ \vdots & & \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & a_{m3}x_3 & + & \ldots\ldots & + & a_{mn}x_n & = & b_n \end{array} \tag{1.1}$$

pri čemu skalare a_{ij} za $1 \leq i \leq m$ i $1 \leq j \leq n$ nazivamo **koeficijentima jednadžbe**, skalare b_i za $1 \leq i \leq m$ **slobodnim članovima**, a x_j za $1 \leq j \leq n$ **nepoznanicama**. Evidentno je da se sustav (1.1) može kraće zapisati kao

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \qquad 1 \leq i \leq m \tag{1.2}$$

što nas motivira za uvođenje matričnog zapisa sustava (1.1).

Definicija 1.1 *Neka je $A \in \mathbb{R}^{m\times n}$ matrica u kojoj se na presjeku i-tog retka i j-tog stupca nalazi $a_{ij} \in \mathbb{R}$ iz (1.1) za $1 \leq i \leq m$ i $1 \leq j \leq n$, tj.*

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \ldots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \ldots & a_{mn} \end{bmatrix} \tag{1.3}$$

Za $x \in \mathbb{R}^n$ uzmimo da je vektor nepoznanica x_j za $1 \leq j \leq n$ te neka je $b \in \mathbb{R}^m$ vektor slobodnih članova b_i za $1 \leq i \leq m$. Iz jednakosti (1.2) i definicije matričnog množenja slijedi da se SLJ iz (1.1) može zapisati kao

$$Ax = b \tag{1.4}$$

pri čemu je A **matrica sustava**, *x* **vektor rješenja**, *a b* **vektor desne strane**.

Primjedba 1.1 *U sljedećim razmatranjima zadržat ćemo se samo na tzv.* ***Cramerovim sustavima*** *kod kojih je $m = n$ (tj. broj jednadžbi jednak je broju nepoznanica) i matrica sustava regularna. Takvi sustavi imaju jedinstveno rješenje. Dodatna razmatranja u pogledu egzistencije i broja rješenja sustava linearnih jednadžbi izlaze iz okvira ovog rada pa ih preskačemo.*

[1] dalje u tekstu SLJ

1.2. Aritmetika s pomičnim zarezom

Memorijski registri računala mogu pohraniti samo konačno mnogo bitova te je konačan i skup realnih brojeva koji se mogu pohraniti u računalu. Što je s realnim brojevima čije se vrijednosti ne mogu točno prikazati? Odgovor na to pitanje daju sljedeći pojmovi.

Definicija 1.2 *Za $\mathbb{F} \subset \mathbb{R}$ kažemo da je* **sustav brojeva s pomičnom zarezom** *ako $\forall x \in \mathbb{F}$ vrijedi*

$$x = 0.z_1z_2z_3 \ldots\ldots z_p \cdot b^e \tag{1.5}$$

pri čemu je $m = 0.z_1z_2z_3 \ldots\ldots z_p$ **mantisa** *broja x, gdje su $z_i \in \{0, 1, 2, \ldots\ldots, b-1\}$, $1 \leq i \leq p$* **znamenke** *broja x, $b \geq 2$* **baza**, *$e \in \mathbb{Z}$* **eksponent**. *Za broj p kažemo da je* **preciznost sustava brojeva** $\mathbb{F}$. *Nadalje, za svaki realan broj x takav da je $x \in \mathbb{F}$ kažemo da je* **reprezentabilan** *u sustavu brojeva s pomičnim zarezom $\mathbb{F}$.*

Definicija 1.3 *Za $x \in \mathbb{F}$ kažemo da je* **normaliziran** *ako je $z_1 \neq 0$, tj. ako je najznačajnija znamenka različita od nule.*

Neka od svojstava sustava brojeva s pomičnim zarezom dana su sljedećom propozicijom.

Propozicija 1.1 *Neka je $\mathbb{F} \subset \mathbb{R}$ sustav brojeva s pomičnim zarezom, p preciznost tog sustava, b baza tog sustava, a e eksponent takav da je $e_{min} \leq e \leq e_{max}$. Tada vrijede sljedeće tvrdnje:*

i) Za svaki normalizirani broj $y \in \mathbb{F}$ vrijedi da je $b^{e_{min}-1} \leq |y| \leq b^{e_{max}}(1 - b^{-p})$

ii) Udaljenost između jedan i prvog većeg reprezentabilnog broja iznosi $\varepsilon_m = b^{1-p}$

Dokaz:
i) Kako je $y \in \mathbb{F}$ možemo ga prikazati kao

$$y = \pm 0.z_1z_2z_3 \ldots\ldots z_p \cdot b^e = (z_1b^{-1} + z_2b^{-2} + z_3b^{-3} + \ldots\ldots + z_pb^{-p})b^e$$

Uz pretpostavku da je y normaliziran, sigurno znamo da je $z_1 \neq 0$. Dakle, najmanji broj po apsolutnoj vrijednosti kojeg ćemo označiti s y_{min} možemo konstruirati tako da na prvu najznačajniju znamenku z_1 stavimo najmanji broj iz skupa $\{0, 1, 2, \ldots\ldots, b-1\}$ koji je različit od nule, dok za sve ostale znamenke stavimo nule. Prirodno je da za eksponent od y_{min} uzmemo najmanji eksponent e_{min}. Sada imamo

$$y_{min} = 0.z_100 \ldots\ldots 0 \cdot b^{e_{min}} = z_1 \cdot b^{e_{min}-1} = b^{e_{min}-1}$$

Analogno, za maksimalni element po apsolutnoj vrijednosti y_{max} uzimamo da je

$$\begin{aligned} y_{max} &= 0.z_1z_2z_3\ldots\ldots z_p \cdot b^e = (z_1b^{-1} + z_2b^{-2} + z_3b^{-3} + \ldots\ldots + z_pb^{-p})b^{e_{max}} \\ &= ((b-1)b^{-1} + (b-1)b^{-2} + (b-1)b^{-3} + \ldots\ldots + (b-1)b^{-p})b^{e_{max}} \\ &= b^{e_{max}}(1-b^{-p}) \end{aligned}$$

Očito je $y_{min} \leq |y| \leq y_{max}$ čime je dokazana tvrdnja *i)*
ii) Neka je $x_1 = 1$ i prvi veći broj $x_2 = 0.1000\ldots01 \cdot b^1$. Tada je

$$\varepsilon_m = |x_2 - x_1| = 0.00\ldots01 \cdot b^1 = b^{1-p}$$

□

Primjedba 1.2 *Broj ε_m iz Propozicije 1.1 naziva se* **strojni epsilon**[2].

Iz same definicije skupa brojeva s pomičnim zarezom vidi se da je nemoguće sve realne brojeve egzaktno zapisati u računalu. Poznato je da su to brojevi $\sqrt{2}$, e, π, itd. koji su iracionalni, ali i neki racionalni brojevi ne mogu biti egzaktno predstavljeni (npr. broj $\frac{1}{10}$ u sustavu s bazom 2). Takav problem rješava tzv. **zaokruživanje**[3] i sastavni je dio aritmetike s pomičnim zarezom.

Definicija 1.4 *Neka je $\mathbb{F} \subset \mathbb{R}$ sustav brojeva s pomičnim zarezom. Preslikavanje $fl : \mathbb{R} \to \mathbb{F}$ zovemo* **zaokruživanje** *ako vrijedi da je $fl(x) = x_F$ pri čemu je $x_F \in \mathbb{F}$ najbliži broju $x \in \mathbb{R}$. Nadalje, ako je $|fl(x)| > max\{|y| : y \in \mathbb{F}\}$, kažemo da je došlo do* **prekoračenja**[4]*, a ukoliko je $0 < |fl(x)| < min\{|y| : 0 \neq y \in \mathbb{F}\}$, kažemo da je došlo do* **potkoračenja**[5].

Sljedeći teorem veoma je važan za proučavanje **grešaka zaokruživanja**[6] koje se javljaju prilikom izvođenja algoritama na računalu.

Teorem 1.1 *Neka je $x \in \mathbb{R}$ takav da postoje x_-,$x_+ \in \mathbb{F}$ sa svojstvom da je $x_- \leq x \leq x_+$. Tada je*

$$fl(x) = x(1+\delta), \quad |\delta| < u \tag{1.6}$$

pri čemu je $u = \frac{1}{2}b^{1-p}$

[2]eng. machine epsilon
[3]eng. rounding
[4]eng. overflow
[5]eng. underflow
[6]eng. roundoff error

Dokaz:
Pretpostavimo da je $x > 0$ (analogno se pokazuje za $x < 0$). Prikažimo broj x kao u (1.5) na sljedeći način:

$$x = 0.z_1z_2z_3\ldots\ldots z_p \cdot b^e = (z_1b^{p-1} + z_2b^{p-2} + \cdots\cdots + z_p^0)b^{e-p}$$

Uvedemo li oznaku $\mu = z_1b^{p-1} + z_2b^{p-2} + \cdots\cdots + z_p^0$ broj x možemo zapisati kao $x = \mu \cdot b^{e-p}$ gdje je $b^{p-1} \leq \mu \leq b^p - 1$. Po pretpostavci teorema postoje x_-,$x_+ \in \mathbb{F}$ takvi da je $x_- \leq x \leq x_+$. Očito je tada $x_- = \lfloor\mu\rfloor b^{e-p}$ i $x_+ = \lceil\mu\rceil b^{e-p}$ (gdje je $\lfloor\mu\rfloor$ najveći cijeli broj manji od μ, a $\lceil\mu\rceil$ najmanji cijeli broj veći od μ). Sada imamo

$$|fl(x) - x| \leq \frac{|x_+ - x_-|}{2} \leq \frac{b^{e-p}}{2}$$

Odatle slijedi da je

$$\left|\frac{fl(x) - x}{x}\right| \leq \frac{1}{2}\mu^{-1} = \frac{1}{2}b^{1-p}$$

Uzmimo da je $u = \frac{1}{2}b^{1-p}$. Vratimo li se na nejednakost iz (1.6), vidimo da je

$$\delta = \frac{fl(x) - x}{x}$$

iz čega slijedi da je $|\delta| \leq u$. Jednakost vrijedi ukoliko je $\mu = b^{1-p}$, inače vrijedi stroga nejednakost. □

Primjedba 1.3 *Broj u iz teorema 1.1 nazivamo* **maksimalna relativna pogreška zaokruživanja**[7].

Osim načina reprezentacije realnog broja u računalu, važno je opisati s kojom točnošću se izvode osnovne aritmetičke operacije. Kada računalo izračuna određeni rezultat, pravi pogrešku prilikom pohrane tog rezultata u spremnik, no i sami operandi prije izvođenja operacije bili su aproksimirani. Model aritmetike osnovnih računskih operacija dan je sljedećim teoremom.

Teorem 1.2 *Neka je* $\mathbb{F} \subset \mathbb{R}$ *sustav brojeva s pomičnim zarezom i* $fl : \mathbb{R} \to \mathbb{F}$ *zaokruživanje. Za* $x, y \in \mathbb{R}$, $y \neq 0$ *vrijedi da je*

$$fl(x \circledast y) = (x \circledast y)(1 + \delta), \quad |\delta| \leq u, \quad \circledast \in \{+, -, \cdot, /\} \tag{1.7}$$

Najčešće korišteni model aritmetike s pomičnim zarezom u mlađim generacijama računala je **IEEE aritmetika dvostruke preciznosti**[8]. Taj model za mantisu koristi 52 bita

[7] eng. maximal relative roundoff error
[8] eng. IEEE double precision

(tj. $p = 53$), za eksponent 11 bita, a predznak zauzima 1 bit. Odatle se može zaključiti da je $e_{min} = -1021$ i $e_{max} = 1024$. Za najmanji reprezentabilan broj uzimamo da je $2^{-1022} \approx 2.2 \cdot 10^{-308}$, a najveći reprezentabilan broj uzimamo da je $2^{1024} \approx 1.8 \cdot 10^{308}$. Relativna pogreška zaokruživanja može maksimalno iznositi $u = 2^{-53} \approx 1.11 \cdot 10^{-16}$, a $\varepsilon_m = 2^{-52} \approx 2.22 \cdot 10^{-16}$. MATLAB koristi IEEE aritmetiku dvostruke preciznosti.

1.3. Ocjene pogreške zaokruživanja

Za **analizu pogreške unazad**[9] Gaussove metode za rješavanje SLJ trebat će nam neki pomoćni rezultati koje ćemo prikazati u ovom poglavlju. Dokazat ćemo stabilnost skalarnog produkta u odnosu na pogrešku zaokruživanja u smislu ocjene pogreške koja nastaje tom prilikom. Prvi rezultat koji će nam trebati glasi:

Lema 1.1 *Neka je* $|\delta_i| \leq u$ *i* $\rho_i = \pm 1$ *za* $1 \leq i \leq n$. *Ako je* $nu < 1$ *onda je*

$$\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} = 1 + \Theta_n, \text{ gdje je } |\Theta_n| \leq \frac{nu}{1-nu} := \gamma_n \tag{1.8}$$

Dokaz:
Dokaz ćemo provesti matematičkom indukcijom (samo za slučaj kada je $\rho_i = 1$). Baza indukcije. Dokažimo da tvrdnja vrijedi za $n = 1$.

$$\prod_{i=1}^{1}(1+\delta_1) = 1 + \delta_1, \quad |\Theta_1| = |\delta_1| \leq u$$

Pretpostavka indukcije. Pretpostavimo da tvrdnja vrijedi za $n = k$, tj. da je

$$\prod_{i=1}^{k}(1+\delta_i)^{\rho_i} = 1 + \Theta_k \quad \text{, gdje je} \tag{1.9}$$

$$|\Theta_k| \leq \frac{ku}{1-ku} := \gamma_k \tag{1.10}$$

$$\text{pri čemu je} \quad |\delta_i| \leq u, \quad \rho_i = 1 \quad \text{za} \quad 1 \leq i \leq k \quad \text{i} \quad ku < 1$$

Korak indukcije. Dokažimo da ista tvrdnja vrijedi za $n = k + 1$. Iz (1.9) imamo

$$\prod_{i=1}^{k+1}(1+\delta_i) = \prod_{i=1}^{k}(1+\delta_i)(1+\delta_{k+1}) = (1+\Theta k)(1+\delta_{k+1}) := 1 + \Theta_{k+1} \tag{1.11}$$

pri čemu uzimamo da je

$$\Theta_{k+1} = \delta_{k+1} + \Theta_k(1+\delta_k) \tag{1.12}$$

[9] eng. Backward Error Analysis

Odredimo sada ogradu za (1.12). Iz relacija (1.10), (1.12) i nejednakosti trokuta slijedi

$$\begin{aligned}|\Theta_{k+1}| &= |\delta_{k+1} + \Theta_k(1+\delta_k)| \leq |\delta_{k+1}| + |\Theta_k||(1+\delta_k)| \\ &\leq |\delta_{k+1}| + |\Theta_k|(1+|\delta_k|) \leq u + \frac{ku}{1-ku}(1+u) \\ &\leq \frac{(k+1)u}{1-(k+1)u} := \gamma_{k+1}\end{aligned}$$

čime je lema u potpunosti dokazana. □

Koristeći lemu 1.1 lako se može ocjeniti pogreška zaokruživanja prilikom računanja skalarnog produkta, što je pokazano u sljedećem teoremu.

Teorem 1.3 *Neka su x, $y \in \mathbb{R}^n$ vektori. Tada za relativnu pogrešku zaokruživanja prilikom izračunavanja skalarnog produkta vektora x i y vrijedi sljedeća ocjena*

$$fl(x^T y) = \sum_{i=1}^{n} x_i y_i (1+\Theta_i), \quad |\Theta_i| \leq \gamma_n \tag{1.13}$$

pri čemu je γ_n ocjena iz Leme 1.1.

Dokaz:
Dokaz provodimo induktivno. U tu svrhu sa s_i označimo i-tu parcijalnu sumu

$$s_i = \sum_{k=1}^{i} x_k y_k, \quad 1 \leq i \leq n$$

Za $n = 1$ očito vrijedi tvrdnja (1.13), jer je

$$\tilde{s_1} = fl(x_1 \cdot y_1) = x_1 y_1 (1+\delta_1), \quad |\Theta_1| = |\delta_1| \leq u < \gamma_1 \tag{1.14}$$

Radi ilustracije, dokazat ćemo tvrdnju za $n = 2$:

$$fl(x_1 y_1 + x_2 y_2) = fl(\tilde{s_1} + x_2 y_2 (1+\delta_2)) = (\tilde{s_1} + x_2 y_2 (1+\delta_2))(1+\delta_3) \tag{1.15}$$

pri čemu je $\tilde{s_1}$ aproksimacija prve parcijalne sume. Uvrštavanjem (1.14) u (1.15) dobivamo da je

$$fl(\sum_{i=1}^{2} x_i y_i) = x_1 y_1 (1+\delta_1)(1+\delta_3) + x_2 y_2 (1+\delta_2)(1+\delta_3) \tag{1.16}$$

Uzmemo li da je $1 + \delta \equiv 1 + \delta_i$ za $1 \leq i \leq 3$ dobivamo da je

$$fl(\sum_{i=1}^{2} x_i y_i) = x_1 y_1 (1+\delta)^2 + x_2 y_2 (1+\delta)^2$$

što po lemi 1.1 daje ocjenu $|\delta| \leq \gamma_2$. Induktivno se može pokazati da je

$$\tilde{s_n} = fl(x^T y) = x_1 y_1 (1+\delta)^n + x_2 y_2 (1+\delta)^n \cdot \cdots \cdot + x_n y_n (1+\delta)^2$$

Koristeći lemu 1.1 opet možemo zaključiti da je

$$fl(x^T y) = \sum_{i=1}^{n} x_i y_i (1+\Theta_i), \quad |\Theta_i| \leq \gamma_n$$

čime je tvrdnja teorema dokazana. □

2. Trokutasti sustavi

U ovom poglavlju proučavat ćemo tzv. **trokutaste sustave linearnih jednadžbi**. Takav naziv potječe od oblika matrice sustava koja je u takvim sustavima gornja, odnosno donja trokutasta matrica. Pretpostavimo da je matrica $A \in \mathbb{R}^{n\times n}$ u (1.3) gornja trokutasta matrica, odnosno da je $a_{ij} = 0$ za $i > j$. Tada zbog (1.4) matrični zapis sustava (1.1) glasi

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2n} \\ 0 & 0 & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix} \tag{2.1}$$

iz čega slijedi da SLJ ima oblik

$$\begin{array}{rcrcrcrcl} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \dots\dots & + & a_{1n}x_n & = & b_1 \\ & & a_{22}x_2 & + & a_{23}x_3 & + & \dots\dots & + & a_{2n}x_n & = & b_2 \\ & & & & a_{33}x_3 & + & \dots\dots & + & a_{3n}x_n & = & b_3 \\ & & & & & & & & \vdots & & \vdots \\ & & & & & & & & a_{nn}x_n & = & b_n \end{array} \tag{2.2}$$

Iz posljednje jednakosti u (2.2) možemo izračunati x_n.

$$x_n = \frac{b_n}{a_{nn}}, \quad a_{nn} \neq 0$$

Znajući x_n, iz pretposljednje jednakosti u (2.2) može se izračunati x_{n-1}.

$$x_{n-1} = \frac{1}{a_{n-1,n-1}}(b_{n-1} - a_{n-1,n}x_n)$$

Tako nastavljajući postupak možemo izračunati sve x_i za $1 \leq i \leq n$, pomoću sljedećeg izraza:

$$x_i = \frac{1}{a_{ii}}(b_i - \sum_{j=i+1}^{n} a_{ij}x_j), \quad 1 \leq i \leq n, \quad a_{ii} \neq 0 \tag{2.3}$$

Na osnovi prethodnog razmatranja možemo napisati pseudokod tzv. algoritma **supstitucija unazad**[10].

[10] eng. Backward Substitution

Algoritam 2.1 (Supstitucije unazad)

Ulaz: *Regularna gornja trokutasta matrica* $U \in \mathbb{R}^{n\times n}$, *vektor desne strane* $b \in \mathbb{R}^n$
Izlaz: *Vektor rješenja* $x \in \mathbb{R}^n$, *takav da je* $Ux = b$
1 : $x_n = b_n/u_{nn}$
2 : **Za** $i = n-1, n-2, \ldots\ldots, 1$
3 : $\quad x_i = (b_i - \sum_{j=i+1}^{n} u_{ij}x_j)/u_{ii}$
4 : **Kraj Za**

Sličnu situaciju imamo i kada je matrica sustava $A \in \mathbb{R}^{n\times n}$ donja trokutasta matrica, tj. kada je $a_{ij} = 0$ za $i < j$. U tom slučaju SLJ ima sljedeći oblik

$$\begin{array}{lllllllll} a_{11}x_1 & & & & & & & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & & & & & = & b_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & & & = & b_3 \\ \vdots & & \vdots & & \vdots & & & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & a_{n3}x_3 & + \ \ldots\ldots \ + & a_{nn}x_n & = & b_n \end{array}$$

Iz prve jednadžbe slijedi da je

$$x_1 = \frac{b_1}{a_{11}}, \quad a_{11} \neq 0$$

Iz druge jednadžbe možemo izračunati x_2, tj.

$$x_2 = \frac{1}{a_{22}}(b_2 - a_{21}x_1).$$

Općenito

$$x_i = \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j), \quad 1 \leq i \leq n, \quad a_{ii} \neq 0$$

Algoritam koji rješava sustav linearnih jednadži čija je matrica sustava donja trokutasta, naziva se **supstitucije unaprijed**[11].

Algoritam 2.2 (Supstitucije unaprijed)

Ulaz: *Regularna donja trokutasta matrica* $L \in \mathbb{R}^{n\times n}$, *vektor desne strane* $b \in \mathbb{R}^n$
Izlaz: *Vektor rješenja* $x \in \mathbb{R}^n$, *takav da je* $Lx = b$
1 : $x_1 = b_1/l_{11}$
2 : **Za** $i = 2, 3, \ldots\ldots, n$
3 : $\quad x_i = (b_i - \sum_{j=1}^{i-1} l_{ij}x_j)/l_{ii}$
4 : **Kraj Za**

[11] eng. Forward Substitution

3. Gaussove eliminacije

3.1. Gaussove eliminacije bez pivotiranja

Standardna *Gaussova metoda eliminacija bez pivotiranja*[12], sastoji se od $(n-1)$ koraka u kojima vršimo elementarne transformacije nad matricom sustava $A \in \mathbb{R}^{n\times n}$ i vektorom desne strane $b \in \mathbb{R}^n$. Zbog toga uvodimo sljedeće oznake:

$$\begin{array}{ll} A^{(1)} = A & \text{matrica prije izvođenja eliminacija} \\ A^{(2)} & \text{matrica nakon 1. koraka eliminacije} \\ \vdots & \vdots \\ \vdots & \vdots \\ A^{(k)} & \text{matrica nakon } (k-1)\text{-og koraka eliminacije} \end{array}$$

Osnovni cilj *Gaussovih eliminacija* jeste da se u k-tom koraku ponište svi subdijagonalni elementi u k-tom stupcu. Kako bi definirali k-ti korak eliminacije, najprije pogledajmo transformirani oblik matrice A i vektora b nakon $(k-1)$-og koraka.

$$A^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1k}^{(1)} & \dots & a_{1j}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \dots & a_{2k}^{(2)} & \dots & a_{2j}^{(2)} & \dots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \dots & a_{3k}^{(3)} & \dots & a_{3j}^{(3)} & \dots & a_{3n}^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{kk}^{(k)} & \dots & a_{kj}^{(k)} & \dots & a_{kn}^{(k)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{ik}^{(k)} & \dots & a_{ij}^{(k)} & \dots & a_{in}^{(k)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{nk}^{(k)} & \dots & a_{nj}^{(k)} & \dots & a_{nn}^{(k)} \end{bmatrix}, b^{(k)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \vdots \\ b_k^{(k)} \\ \vdots \\ b_j^{(k)} \\ \vdots \\ b_n^{(k)} \end{bmatrix} \tag{3.1}$$

Nakon $(k-1)$-og koraka u lijevom gornjem kutu matrice $A^{(k)}$ nalazi se gornja trokutasta matrica formata $(k-1)\times(k-1)$. Kako bi poništili sve subdijagonalne elemente u k-tom stupcu, k-ti redak množimo sa $-\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$, za $i = k+1 : n$, uz pretpostavku da je $a_{kk}^{(k)} \neq 0$, i dodajemo ga i-tom retku, za $i = k+1 : n$. Nakon toga dobivamo da je

$$a_{ij}^{(k+1)} = \begin{cases} 0 & i = k+1 : n \quad j = k \\ a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)} & i,j = k+1 : n \end{cases} \tag{3.2}$$

Na isti način transformiramo vektor b

$$b_i^{(k+1)} = b_i^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} b_k^{(k)}, \quad i = k+1 : n \tag{3.3}$$

[12]Dalje u tekstu označena je kao GEBP

Nakon $(n-1)$-og koraka dolazimo do gornje trokutastog sustava

$$A^{(n)}x = b^{(n)}, \tag{3.4}$$

koji se lako rješava pomoću algoritma 2.1.

3.2. LU dekompozicija

LU dekompozicija je rastav matrice $A \in \mathbb{R}^{n\times n}$ na produkt jedinične donje trokutaste matrice $L \in \mathbb{R}^{n\times n}$ (tj. matrice koja je donja trokutasta a na dijagonali ima sve jedinice) i gornje trokutaste matrice $U \in \mathbb{R}^{n\times n}$. Prije nego li konstruiramo algoritam za računanje takvog rastava, moramo provjeriti njegovu egzistenciju, tj. nužne i dovoljne uvjete za postojanje takvog rastava.

Definicija 3.1 *Glavna j-ta podmatrica matrice $A \in \mathbb{R}^{n\times n}$ je matrica $A_j \in \mathbb{R}^{j\times j}$ sa svojstvom da je $(A_j)_{kl} = a_{kl}$ za $1 \le k, l \le j$, gdje je $1 \le j \le n$.*

Teorem 3.1 *Neka je $A \in \mathbb{R}^{n\times n}$. Tada su sljedeće dvije tvrdnje ekvivalentne:*

i) Postoje jedinstvena jedinična donja trokutasta matrica $L \in \mathbb{R}^{n\times n}$ i regularna gornja trokutasta matrica $U \in \mathbb{R}^{n\times n}$ takve da je $A = LU$.

ii) Sve glavne podmatrice od A su regularne

Dokaz:
Dokažimo najprije da iz i) slijedi ii). Pretpostavimo da postoje jedinična donja trokutasta matrica $L \in \mathbb{R}^{n\times n}$ i regularna gornja trokutasta matrica $U \in \mathbb{R}^{n\times n}$ takve da je $A = LU$. To možemo zapisati kao

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix} = \begin{bmatrix} L_{11}U_{11} & L_{11}U_{12} \\ L_{21}U_{11} & L_{21}U_{12} + L_{22}U_{22} \end{bmatrix} \tag{3.5}$$

gdje su $A_{11}, L_{11}, U_{11} \in \mathbb{R}^{j\times j}$, pri čemu je $A_{11} = L_{11}U_{11}$ glavna podmatrica od A. Prema Binnet-Cauchyjevom teoremu slijedi da je

$$\det(A_{11}) = \det(L_{11}U_{11}) = \det(L_{11})\det(U_{11}) = 1 \cdot \det(U_{11}) = 1 \cdot \prod_{i=1}^{j} u_{ii} \neq 0$$

jer je $u_{ii} \neq 0$ za $1 \le i \le j$ zbog pretpostavke da je U_{11} regularna. Budući je $\det A_{11} \neq 0$, slijedi da je svaka glavna podmatrica regularna.
Dokažimo sada da iz ii) slijedi i). Neka je svaka glavna podmatrica od A regularna. Dokaz ćemo provesti indukcijom s obzirom na red matrice A. Za $n = 1$ očito postoji

takva dekompozicija jer ako je $A = [a]$ tada za traženu dekompoziciju možemo uzeti da je $A = 1 \cdot a$ pri čemu je $L = [1]$ i $U = [a]$. Kako je $a \neq 0$ po pretpostavci, slijedi da je U regularna. Pretpostavimo da takva dekompozicija postoji za matricu $A \in \mathbb{R}^{(n-1)\times(n-1)}$ u kojoj je svaka glavna podmatrica regularna. Pitamo se postoji li LU-dekompozicija iz tvrdnje i) za n-dimenzionalnu matricu A_n koja se iz matrice A_{n-1} dobije proširenjem na sljedeći način:

$$A_n = \begin{bmatrix} A_{n-1} & b \\ c^T & a_{nn} \end{bmatrix} = \begin{bmatrix} L_{n-1} & 0 \\ l^T & 1 \end{bmatrix} \begin{bmatrix} U_{n-1} & u \\ 0 & u_{nn} \end{bmatrix} =: L_n U_n \tag{3.6}$$

gdje su $c, b \in \mathbb{R}^{n-1}$ i $a_{nn} \in \mathbb{R}$.
Treba pokazati da postoje l,$u \in \mathbb{R}^{n-1}$ i $u_{nn} \in \mathbb{R}$ takvi da vrijedi matrična jednadžba (3.6), tj.

$$L_{n-1}u = b \tag{3.7}$$

$$U_{n-1}^T l = c \tag{3.8}$$

$$a_{nn} = l^T u + u_{nn} \tag{3.9}$$

Kako su L_{n-1} i U_{n-1} regularne prema pretpostavci, jednadžbe (3.7), (3.8) i (3.9) imaju jedinstvena rješenja pa matrica A_n ima jedinstvenu LU dekompoziciju.

Primjedba 3.1 *Ukoliko postoji barem jedna singularna podmatrica A_k, $1 \leq k \leq n$ matrice A, tada ne postoji LU-dekompozicija opisana u tvrdnji i) iz Teorema 3.1. Iz LU dekompozicije matrice*

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

vidimo da l može biti bilo koji realan broj, pa LU dekompozicija nije jedinstvena.

LU dekompozicija matrice A može se dobiti primjenom Gaussovih eliminacija. Ovdje će taj postupak biti opisan kroz odgovarajuća matrična množenja iako prilikom implementacije algoritama na računalu izbjegavamao matrično množenje. Prvi korak se može zapisati na sljedeći način:

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ -\frac{a_{21}^{(1)}}{a_{11}^{(1)}} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}^{(1)}}{a_{11}^{(1)}} & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ -\frac{a_{21}^{(1)}}{a_{11}^{(1)}} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}^{(1)}}{a_{11}^{(1)}} & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \\ \vdots \\ b_n^{(1)} \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{bmatrix}$$

odnosno

$$A^{(2)} = L_1 A^{(1)} \quad b^{(2)} = L_1 b^{(1)}$$

gdje je

$$L_1 = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -m_2^{(1)} & 1 & 0 & \dots & 0 \\ -m_3^{(1)} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -m_n^{(1)} & 0 & 0 & \dots & 1 \end{bmatrix}, \quad m_i^{(1)} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \quad i = 2 : n.$$

Nakon $(n-1)$-og koraka dobivamo sljedeće

$$A^{(n)} = L_{n-1} L_{n-2} \cdot \cdots \cdot L_2 L_1 A \quad b^{(n)} = L_{n-1} L_{n-2} \cdot \cdots \cdot L_2 L_1 b$$

gdje je

$$L_k = \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & -m_{k+1}^{(k)} & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & -m_n^{(k)} & 0 & \dots & 1 \end{bmatrix}, \quad m_i^{(k)} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \quad i = k+1 : n. \tag{3.10}$$

Iz prethodno opisanog postupka slijedi da je matrica $A^{(n)}$ gornja trokutasta matrica pa ćemo uzeti da je $U = A^{(n)}$. Dakle, sada imamo da je

$$U = L_{n-1} L_{n-2} \cdot \cdots \cdot L_2 L_1 A \tag{3.11}$$

Uzastopnim množenjem lijeve strane matricama $L_{n-1}^{-1}, L_{n-2}^{-1}, \dots\dots, L_1^{-1}$ jednakost (3.11) prijelazi u

$$A = L_1^{-1} L_2^{-1} \cdot \cdots \cdot L_{n-2}^{-1} L_{n-1}^{-1} U$$

Neka je

$$L = L_1^{-1} L_2^{-1} \cdot \cdots \cdot L_{n-2}^{-1} L_{n-1}^{-1}, \tag{3.12}$$

tada je

$$A = LU$$

Teorem 3.2 *Neka je $L \in \mathbb{R}^{n \times n}$ definirana relacijom (3.12). Tada je L jedinična donja trokutasta matrica s elementima $(L)_{ij} = m_{ij}, \quad i > j$, različitim od nule.*

Dokaz:
Uočimo da je

$$L_k = I - m^{(k)} e_k^T$$

gdje je

$$m^{(k)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ m_{k+1}^{(k)} \\ \vdots \\ m_n^{(k)} \end{bmatrix}, \qquad e_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

gdje je e_k k-ti vektor kanonske baze vektorskog prostora $\mathbb{R}^n$ nad poljem $\mathbb{R}$. Nadalje, kako je

$$(I - m^{(k)}e_k^T)(I + m^{(k)}e_k^T) = I = (I + m^{(k)}e_k^T)(I - m^{(k)}e_k^T)$$

vidimo da je

$$L_k^{-1} = I + m^{(k)}e_k^T, \quad k = 1 : n-1 \tag{3.13}$$

Koristeći jednakost (3.13) može se pokazati sljedeće

$$\begin{aligned} L =& L_1^{-1}L_2^{-1}\cdot\cdots\cdot L_{n-2}^{-1}L_{n-1}^{-1} = (I + m^{(1)}e_1^T)(I + m^{(2)}e_2^T)\cdot\cdots\cdot(I + m^{(n-1)}e_{n-1}^T) \\ &= I + \sum_{\sigma_k \subseteq P_k, 1\le k\le n} m^{(i_1)}e_{i_1}^T m^{(i_2)}e_{i_2}^T \cdot\cdots\cdot m^{(i_k)}e_{i_k}^T \end{aligned} \tag{3.14}$$

gdje je $\sigma_k = \{i_1, \ldots\ldots i_k\}$ k-člani skup indeksa takav da je $i_1 < i_2 < \ldots\ldots < i_k$, a $P_k = \binom{[n-1]}{k}$ skup svih k-članih podskupova od skupa $[n-1] = \{1, \ldots\ldots, n-1\}$. Kako je $e_i^T m^{(k)} = 0$ za $i < k$, slijedi da je $m^{(i_1)}e_{i_1}^T m^{(i_2)}e_{i_2}^T \cdot\cdots\cdot m^{(i_k)}e_{i_k}^T = 0$, $\forall k \ge 2$. Iz prethodnog razmatranja vidimo da je

$$L = L_1^{-1}L_2^{-1}\cdot\cdots\cdot L_{n-2}^{-1}L_{n-1}^{-1} = I + m^{(1)}e_1^T + m^{(2)}e_2^T + \ldots\ldots m^{(n-1)}e_{n-1}^T$$

$$= \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 \\ m_2^{(1)} & 0 & 0 & \ldots & 0 \\ m_3^{(1)} & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_n^{(1)} & 0 & \ldots & 0 & 0 \end{bmatrix} + \ldots + \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & m_n^{(n-1)} & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ m_2^{(1)} & 1 & 0 & \ldots & 0 \\ m_3^{(1)} & m_3^{(2)} & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_n^{(1)} & m_n^{(2)} & \ldots & m_n^{(n-1)} & 1 \end{bmatrix}, \quad gdje \quad je \quad m_i^{(k)} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \quad i < k$$

□

3.3. Gaussove eliminacije s djelomičnim pivotiranjem

3.3.1. Problemi u rješavanju linearnih sustava

Definicija 3.2 *Element $a_{kk}^{(k)}$ iz (3.1) nazivamo **pivotni element** ili **pivot** u k-tom koraku Gaussovih eliminacija*

U prethodno navedenom opisu *GEBP*-a pretpostavljali smo da je $a_{kk}^{(k)} \neq 0$ i samo je pod tim uvjetom bilo moguće izvesti algoritam GEBP-a. Ako se u koraku eliminacije pojavi nula kao pivotni element, algoritam se prekida iako je možda matrica sustava regularna. Naime, postoje regularne matrice čiji su pivoti nula. Npr. linearni sustav

$$\begin{bmatrix} 0 & 3 & 1 \\ 1 & 2 & 3 \\ 4 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \\ 7 \end{bmatrix}$$

ima riješenje $\begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}$, no metodom Gaussovih eliminacija bez pivotiranja ne možemo ga riješiti, jer je prvi pivotni element jednak nuli.

Drugi tip problema javlja se zbog izvođenja algoritma *GEBP* na računalu koje koristi aritmetiku s pomičnim zarezom. Sljedeći primjer ilustrira jedan takav slučaj. Neka je $\varepsilon > 0$ takav da je $\varepsilon < \frac{1}{2^{p+1}}$ gdje je p preciznost sustava brojeva s pomičnim zarezom (vidi 1.2.). Treba rješiti sustav

$$\begin{aligned} \varepsilon x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \end{aligned} \tag{3.15}$$

koji u matričnom zapisu izgleda ovako

$$\begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{3.16}$$

Primjenom *GEBP* sustav (3.16) svodimo na gornji trokutasti sustav

$$\begin{bmatrix} \varepsilon & 1 \\ 0 & 1-\varepsilon^{-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2-\varepsilon^{-1} \end{bmatrix}$$

čije rješenje iznosi

$$x_2 = \frac{1-2\varepsilon}{1-\varepsilon}, \quad x_1 = \frac{1}{\varepsilon}(1-x_2) = \frac{1}{1-\varepsilon}$$

iz čega vidimo da je točno rješenje sustava blisko vektoru $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Međutim, kada pogledamo kako će algoritam biti izveden na računalu, dolazimo do sljedećih rezultata

$$\begin{aligned} fl(1-\frac{1}{\varepsilon}) &= -\frac{1}{\varepsilon} \\ fl(2-\frac{1}{\varepsilon}) &= -\frac{1}{\varepsilon} \end{aligned}$$

Dakle, nakon Gaussovih eliminacija provedenih na računalu, dolazi do pogreški zaokruživanja što daje sustav

$$\begin{bmatrix} \varepsilon & 1 \\ 0 & -\varepsilon^{-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -\varepsilon^{-1} \end{bmatrix}$$

čije je rješenje $x_2 = 1$ i $x_1 = 0$, a to je značajna pogreška u odnosu na pravo rješenje. Pogledajmo što bi se dogodilo nakon zamjene redaka matrice sustava i vektora desne strane (to možemo napraviti jer takva zamjena ne utječe na konačni rezultat)

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

Nakon primjene Gaussovih eliminacija dolazimo do sustava

$$\begin{bmatrix} 1 & 1 \\ 0 & 1-\varepsilon \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1-2\varepsilon \end{bmatrix} \tag{3.17}$$

čije je rješenje $x_2 = \frac{1-2\varepsilon}{1-\varepsilon} \approx 1$ i $x_1 = 2 - x_2 \approx 1$. Kao što vidimo, zamjenom redaka izbegli smo neželjenu numeričku nestabilnost do koje dolazi zbog malog pivota. Ovako jednostavan primjer daje ideju za formiranje mnogo efikasnijeg algoritma nego što su *GEBP*.

3.3.2. Permutacijska matrica

Kao što smo vidjeli u primjeru (3.15) do pravog rješenja došli smo tek nakon zamjene redaka matrice sustava i vektora desne strane. Općenito, zamjene redaka možemo predstaviti množenjem s lijeva posebnom grupom matrica koje se nazivaju permutacijske matrice.

Definicija 3.3 *Neka je dana permutacija* $\sigma : \{1, 2, 3 \ldots\ldots n\} \to \{1, 2, 3 \ldots\ldots n\}$. *Permutacijska matrica* P_σ *pridružena permutaciji* σ *jeste matrica definirana formulom*

$$(P_\sigma)_{ij} = \delta_{\sigma_{(i)j}}$$

Iz same definicije slijedi da permutacijsku matricu dobijemo zamjenom redaka jedinične matrice, tako da svaki redak i svaki stupac ima samo jednu jedinicu a sve ostalo su nule.

Lema 3.1 *Neka su* P_σ, P_{σ_1} *i* P_{σ_2} *permutacijske matrice pridružene permutacijama* σ,σ_1 *i* σ_2, *te neka je* $A \in \mathbb{R}^{n\times n}$ *matrica n-tog reda. Tada vrijede sljedeće tvrdnje:*

i) Množenjem s lijeva matrice A s matricom P_σ *dobivamo matricu* $P_\sigma A$ *čiji su retci permutirani, a množenjem s desna matrice A s matricom* P_σ *dobivamo matricu* AP_σ *čiji su stupci permutirani.*

ii) Matrica P_σ je ortogonalna.

iii) $\det(P_\sigma) = \pm 1$

iv) $P_{\sigma_1} \cdot P_{\sigma_2}$ je također permutacijska matrica i vrijedi

$$P_{\sigma_1} \cdot P_{\sigma_2} = P_{\sigma_1 \circ \sigma_2}$$

Definicija 3.4 *Neka je $\tau : \{1, 2, 3 \ldots\ldots n\} \to \{1, 2, 3 \ldots\ldots n\}$ permutacija zadana na sljedeći način*

$$\tau(k) = \begin{cases} i & , k{=}j \\ j & , k{=}i \\ k & , \textit{inače} \end{cases} \tag{3.18}$$

Permutacijsku matricu T pridruženu permutaciji τ zovemo ***transpozicijska matrica****.*

Propozicija 3.1 *Svaka transpozicijska matrica je sama sebi inverz.*

Dokaz:
Iz definicije 3.4 slijedi da je $T^T = T$. Primjenom tvrdnje *ii*) iz leme 3.1 zaključujemo da je $T^{-1} = T^T$ iz čega slijedi da je

$$T = T^{-1}.$$

□

3.3.3. Osnovna ideja pivotiranja

Problemi opisani u točki 3.3.1. motiviraju uvođenje nove strategije rješavanja linearnih sustava. U algoritmu *GEBP* nismo vodili računa o izboru pivotnog elementa jer smo pretpostavljali da je različit od nule. **Gaussove eliminacije s djelomičnim pivotiranjem**[13] su algoritam rješavanja SLJ kod kojeg se u svakom koraku vodi računa o izboru pivotnog elementa. Pretpostavimo da smo izvršili $(k-1)$-u eliminaciju i da u k-tom koraku moramo poništiti sve elemente k-tog stupca ispod dijagonale.

$$\begin{array}{cc} & \begin{bmatrix} x & x & x & x & x & x & x & x & x & x \\ & x & x & x & x & x & x & x & x & x \\ & & x & x & x & x & x & x & x & x \\ & & & x & x & x & x & x & x & x \\ k & \to & & & y & y & y & y & y & y \\ & & & & * & * & * & * & * & * \\ & & & & * & * & * & * & * & * \\ & & & & * & * & * & * & * & * \\ r & \to & & & z & z & z & z & z & z \\ & & & & * & * & * & * & * & * \end{bmatrix} \end{array}$$

[13] Dalje u tekstu je označen kao GEDP

Prije formiranja multiplikatora m_{ik} biramo $a_{rk}^{(k)}$ takav da je

$$|a_{rk}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}| \tag{3.19}$$

a nakon toga vršimo zamjenu r-tog i k-tog retka. Time postižemo da za svaki multiplikator m_{ik} vrijedi

$$m_{ik} = -\frac{a_{ik}^{(k)}}{a_{rk}^{(k)}}, \quad i = k+1 : n, \quad | \, m_{ik} \, | \le 1 \tag{3.20}$$

Nakon formiranja multiplikatora eliminacija se nastavlja kao i kod *GEBP*-a.

3.3.4. PLU dekompozicija i algoritam GEDP

Sada ćemo uz prethodno navedenu ideju pivotiranja detaljnije opisati modificirani algoritam koji podrazumijeva zamjenu redaka u svakom koraku (osim ako je pivotni element po apsolutnoj vrijednosti veći od svih subdijagonalnih elemenata). Ovdje je veoma važno naglasiti da zamjena redaka u matrici sustava mora biti praćena odgovarajućim zamjenama elemenata u vektoru desne strane. Tako ćemo npr. u k-tom koraku eliminacija transformiranu matricu A i vektor desne strane b označiti sa $[A^{(k)}, b^{(k)}]$. Do uključujući k-tog koraka, početna matrica A i vektor b transformirani su k puta na sljedeći način

$$[A^{(k+1)}, b^{(k+1)}] = L_k T_k L_{k-1} T_{k-1} \ldots\ldots L_1 T_1 [A, b]$$

gdje je L_i, za $1 \le i \le k$ definirana kao u (3.10), a T_i, za $1 \le i \le k$, transpozicijska matrica koja vrši zamjenu i-tog retka s r-tim retkom gdje je $i \le r \le n$. Izbor indeksa r vrši se po kriteriju iz (3.19). Budući da ukupno ima $(n-1)$ koraka, u zadnjem koraku $A^{(n)}$ bit će gornja trokutasta matrica.

$$[A^{(n)}, b^{(n)}] = L_{n-1} T_{n-1} L_{n-2} T_{n-2} \ldots\ldots L_1 T_1 [A, b]$$

Koristeći propoziciju 3.1 možemo učiniti sljedeće

$$\begin{aligned}[A^{(n)}, b^{(n)}] = {} & L_{n-1} T_{n-1} L_{n-2} T_{n-1} T_{n-1} T_{n-2} L_{n-3} T_{n-2} T_{n-1} T_{n-1} T_{n-2} \ldots\ldots \\ & L_1 T_2 T_3 \cdot\cdots\cdot T_{n-2} T_{n-1} T_{n-1} T_{n-2} \cdot\cdots\cdot T_3 T_2 T_1 [A, b]\end{aligned}$$

što nakon grupiranja daje

$$\begin{aligned}[A^{(n)}, b^{(n)}] = {} & L_{n-1}(T_{n-1} L_{n-2} T_{n-1})(T_{n-1} T_{n-2} L_{n-3} T_{n-2} T_{n-1})(T_{n-1} T_{n-2} \ldots\ldots \\ & L_2 \cdot\cdots\cdot T_{n-2} T_{n-1})(T_{n-1} T_{n-2} T_{n-3} \cdot\cdots\cdot T_2 L_1 T_2 T_3 \cdot\cdots\cdot \\ & T_{n-2} T_{n-1})(T_{n-1} T_{n-2} \cdot\cdots\cdot T_3 T_2 T_1)[A, b]\end{aligned} \tag{3.21}$$

Uvedemo li oznake

$$\widetilde{L}_i = P_i L_i P_i^T$$

gdje je

$$P_i = T_{n-1}T_{n-2}\cdot\cdots\cdot T_{i+1}, \quad \text{za} \quad 1 \le i \le n-1, \tag{3.22}$$

relacija (3.21) prelazi u oblik

$$[A^{(n)}, b^{(n)}] = \widetilde{L}_{n-1}\widetilde{L}_{n-2}\cdot\cdots\cdot\widetilde{L}_1[\widetilde{A}, \tilde{b}] \tag{3.23}$$

pri čemu je $\widetilde{A} = PA$ i $\tilde{b} = Pb$, gdje je $P = T_{n-1}T_{n-2}\cdot\cdots\cdot T_1$. Uzmemo li da je

$$L = (\widetilde{L}_{n-1}\widetilde{L}_{n-2}\cdot\cdots\cdot\widetilde{L}_1)^{-1} = \widetilde{L}_1^{-1}\widetilde{L}_2^{-1}\cdot\cdots\cdot\widetilde{L}_{n-1}^{-1} \tag{3.24}$$

relaciju (3.23) možemo pisati kao

$$[A^{(n)}, b^{(n)}] = L^{-1}[PA, Pb]$$

Kako je $A^{(n)}$ gornja trokutasta matrica, možemo je označiti sa U, nakon čega dobivamo da je $U = L^{-1}PA$ i $b^{(n)} = L^{-1}Pb$, što množenjem s lijeva matricom L daje $PA = LU$ i $Pb = Lb^{(n)}$.

Lema 3.2 *Matrica L iz (3.24) je jedinična donja trokutasta matrica s elementima* $(L)_{ij}$*, gdje je* $i > j$*, koji su po modulu manji od* 1*.*

Dokaz: Iz relacije (3.22) vidimo da je P_i matrica permutacije koja permutira retke sa indeksom od $(i+1)$ do $(n-1)$ (gdje je $1 \le i \le n-1$), a reci sa indeksima od 1 do i ostaju nepermutirani. To znači da je

$$P_i = \begin{bmatrix} I_i & 0 \\ 0 & S_{n-i} \end{bmatrix}$$

gdje je I_i jedinična matrica i-tog reda, a S_{n-i} permutacijska matrica $(n-i)$-tog reda. Za matricu L_i uzimamo da je ista kao u teoremu 3.2, tj. $L_i = I - m^{(i)}e_i^T$ gdje je

$$m^{(i)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ m_{i+1}^{(i)} \\ \vdots \\ m_n^{(i)} \end{bmatrix}, \qquad e_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Tada je

$$\begin{aligned} \widetilde{L}_i &= P_iL_iP_i^T = P_i(I - m^{(i)}e_i^T)P_i^T = (P_i - P_im^{(i)}e_i^T)P_i^T \\ &= P_iP_i^T - P_im^{(i)}e_i^TP_i^T = I - P_im^{(i)}(P_ie_i)^T \end{aligned}$$

Kako je $P_i e_i = e_i$ slijedi da je

$$L_i = I - P_i m^{(i)} e_i^T \tag{3.25}$$

gdje je

$$P_i m^{(i)} = \begin{bmatrix} I_i & 0 \\ 0 & S_{n-i} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ m_{i+1}^{(i)} \\ \vdots \\ m_n^{(i)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ S_{n-i} m_i \end{bmatrix} \qquad m_i = \begin{bmatrix} m_{i+1}^{(i)} \\ \vdots \\ \vdots \\ m_n^{(i)} \end{bmatrix}.$$

Iz (3.25) slijedi da $\widetilde{L}_i$ ima isti oblik kao i L_i u teoremu 3.2 samo što su subdijagonalni elementi ispermutirani permutacijskom matricom $S_{n-i} \in \mathbb{R}^{n-i}$. Dokaz se provodi analogno teoremu 3.2. Inverz matrice $\widetilde{L}_i^{-1}$ jednak je

$$\widetilde{L}_i^{-1} = I + (P_i m^{(i)}) e_i^T, \qquad i = 1, 2, \ldots\ldots, n-1$$

Množenjem tih matrica dobivamo

$$\begin{aligned} L &= \widetilde{L}_1^{-1} \widetilde{L}_2^{-1} \cdot \cdots\cdots \widetilde{L}_{n-1}^{-1} \\ &= I + (P_1 m^{(1)}) e_1^T + (P_2 m^{(2)}) e_2^T + \cdots\cdots + (P_{n-2} m^{(n-2)}) e_{n-2}^T + m^{(n-1)} e_{n-1}^T, \end{aligned}$$

iz čega se vidi da je L donja trokutasta matrica. Budući da smo tijekom eliminacija na pivotno mjesto dovodili element koji je po apsolutnoj vrijednosti najveći u tom stupcu, možemo zaključiti da su svi multiplikatori m_{ik} iz (3.20) koji se nalaze u donjem trokutu od L, po apsolutnoj vrijednosti manji od 1. □

Rezultate prethodnih razmatranja prikazat ćemo sljedećim teoremom:

Teorem 3.3 *Neka je A regularna matrica reda n. Tada postoji matrica permutacije P takva da postoji LU dekompozicija matrice PA, odnosno da je $LU = PA$. Pri tome su u L svi elementi ispod dijagonale po apsolutnoj vrijednosti manji od 1.*

Dakle, prema teoremu 3.3 postoji PLU dekompozicija regularne matrice $A \in \mathbb{R}^{n \times n}$ iz čega slijedi da možemo konstruirati odgovarajući algoritam.

Algoritam 3.1 (LU dekompozicija s djelomičnim pivotiranjem)

Ulaz: *Regularna matrica* $A \in \mathbb{R}^{n\times n}$
Izlaz: $L, U, P \in \mathbb{R}^{n\times n}$ *takve da je* $PA = LU$, *gdje je* L *jedinična donja trokutasta matrica,* U *regularna gornja trokutasta matrica i* P *permutacijska matrica*

$U = A, \quad L = I, \quad P = I$
Za $k = 1, \ldots\ldots, n-1$:
  Pronaći $i \in \{k, \ldots\ldots, n\}$ *takvo da je* $|u_{ik}^{(k)}| = \max_{k\le j\le n} |u_{jk}^{(k)}|$
  Zamjeniti redak $(u_{k,k}^{(k)}, \ldots\ldots, u_{k,n}^{(k)})$ *s retkom* $(u_{i,k}^{(k)}, \ldots\ldots, u_{i,n}^{(k)})$
  Zamjeniti redak $(l_{k,1}, \ldots\ldots, l_{k,k-1})$ *s retkom* $(l_{i,1}, \ldots\ldots, l_{i,k-1})$
  Zamjeniti redak $(p_{k,1}, \ldots\ldots, p_{k,n})$ *s retkom* $(p_{i,1}, \ldots\ldots, p_{i,n})$
    Za $j = k+1, \ldots\ldots, n$
      $l_{j,k} = u_{j,k}^{(k)} / u_{k,k}^{(k)}$
      $(u_{j,k}^{(k+1)}, \ldots\ldots, u_{j,n}^{(k+1)}) = (u_{j,k}^{(k)}, \ldots\ldots, u_{j,n}^{(k)}) - l_{j,k}(u_{k,k}^{(k)}, \ldots\ldots, u_{k,n}^{(k)})$
    Kraj Za
Kraj Za

Osnovni cilj LU dekompozicije matrice A jeste rješavanje sustava linearnih jednadžbi. Kao što je već poznato, takav rastav matrice svodi problem rješavanja SLJ na uzastopno rješavanje dva trokutasta sustava što je opisano sljedećim algoritmom.

Algoritam 3.2 (Gaussove eliminacije s djelomičnim pivotiranjem)

Ulaz: *Regularna matrica sustava* $A \in \mathbb{R}^{n\times n}$, *vektor desne strane* $b \in \mathbb{R}^n$
Izlaz: *Vektor rješenja sustava* $x \in \mathbb{R}^n$

Pronaći $PA = LU$ *koristeći* ***Algoritam 3.1***
Riješiti sustav $Ly = P$ *koristeći* ***Algoritam 2.2***
Riješiti sustav $Ux = y$ *koristeći* ***Algoritam 2.1***

3.3.5. Algoritam GEDP bez stvarne zamjene redaka

Kao što smo vidjeli, u algoritmu 3.1 bilo je potrebno izvesti zamjenu redaka unutar matrice U, L i P. Međutim, algoritam LU dekompozicije s djelomičnim pivotiranjem može se provesti bez fizičke zamjene redaka. To postižemo uvođenjem n-dimenzionalnog per-

mutacijskog vektora p

$$p = (1, 2, 3 \ldots\ldots, n)$$

Pomoću tako definiranog vektora, retku matrice A pristupit ćemo pomoću elemenata vektora p, tj. umjesto da direktno pristupamo elementu a_{ij}, pristupit ćemo mu indirektno kao $a_{p_i j}$, gdje je p_i i-ti element vektora p. Umjesto fizičke zamjene k-tog i l-tog retka, vršit ćemo zamjenu k-tog i l-tog elementa vektora p.

Algoritam 3.3 (LU dekompozicija s djelomičnim pivotiranjem bez stvarne zamjene redaka)

Ulaz: *Regularna matrica* $A \in \mathbb{R}^{n\times n}$
Izlaz: $A \in \mathbb{R}^{n\times n}$, *permutacijski vektor* p

Za $i = 1, 2, \ldots\ldots, n$
  $p_i = i$
Kraj Za
Za $k = 1, 2, \ldots\ldots, n-1$
  Pronaći $j \in \{k, \ldots\ldots, n\}$ *takav da je* $|a_{p_j k}| = \max_{k\le i\le n} |a_{p_i k}|$
  Ako je $a_{p_j k} = 0$
    Ispiši: *Matrica sustava je singularna*
  Kraj Ako
  Zamjeniti p_k *i* p_j
  Za $i = k+1, k+2, \ldots\ldots, n$
    $z = a_{p_i k}/a_{p_k k}$
    $a_{p_i k} = z$
    Za $j = k+1, k+2, \ldots\ldots, n$
      $a_{p_i j} = a_{p_i j} - z a_{p_k j}$
    Kraj Za
  Kraj Za
Kraj Za

Nakon izvršenja prethodno opisanog algoritma, dobivamo permutiranu matricu A u čijem se donjem trokutu bez dijagonale nalaze elementi matrice L, a u gornjem trokutu sa dijagonalom elementi gornje trokutaste matrice U. Sljedećim algoritmom opisano je rješavanje linearnog sustava nakon provedenog algoritma LU dekompozicije bez fizičke zamjene redaka.

Algoritam 3.4 (Gaussove eliminacije s djelomičnim pivotiranjem bez stvarne zamjene redaka)

Ulaz: *Matrica A nakon provedene LU dekompozicije iz Algoritma 3.3, vektor permutacije p, vektor desne strane b*
Izlaz: *Vektor rješenja x*

1 : **Za** $k = 1, 2, \ldots\ldots, n-1$
2 : **Za** $i = k+1, k+2, \ldots\ldots, n$
3 : $b_{p_i} = b_{p_i} - a_{p_i k} b_{p_k}$
4 : ***Kraj Za***
5 : ***Kraj Za***
6 : $x_n = b_{p_n} / a_{p_n n}$
7 : **Za** $i = n-1, n-2, \ldots\ldots, 1$
8 : $temp = 0$
9 : ***Za*** $j = i+1, i+2, \ldots\ldots, n$
10 : $temp = temp + a_{p_i j} x_j$
11 : ***Kraj Za***
12 : $x_i = (b_{p_i} - temp) / a_{p_i i}$
13 : ***Kraj Za***

4. Analiza pogreške Gaussove metode eliminacija

4.1. Povratna analiza greške algoritma GEBP

Analiza povratne pogreške bit će provedena samo na algoritmu Gaussovih eliminacija bez pivotiranja. Važno je reći da su ove tzv. *a priori* granice znatno veće od stvarnih pogreški, ali one nam daju sigurnu granicu pogreške koja se ne može premašiti.

Teorem 4.1 *Neka je $A \in \mathbb{R}^{n\times n}$. Pretpostavimo da su $\widetilde{L}$ i $\widetilde{U}$ matrice dobivene LU dekompozicijom iz Algoritma 3.1 te neka je $\tilde{x}$ rješenje linearnog sustava $Ax = b$. Tada vrijedi sljedeća ocjena*

$$(A + \Delta A)\tilde{x} = b, \text{ gdje je } |\Delta A| \leq (3\gamma_n + \gamma_n^2)|\widetilde{L}||\widetilde{U}|,$$

gdje je γ_n broj iz (1.8). Drugim riječima, algoritam GEBP je povratno stabilan.

Dokaz: Nakon provedenih Gaussovih eliminacija dobivamo LU dekopoziciju matrice A

$$A = \widetilde{L}\widetilde{U}$$

takvu da je

$$a_{jk} = \sum_{i=1}^{j} l_{ji}u_{ik}$$

(jer su matrice $\widetilde{L}$ i $\widetilde{U}$ trokutaste). Iz činjenice da je $l_{ii} = 1$ za $1 \leq i \leq n$, lako je uočiti da je za $j \leq k$

$$u_{jk} = a_{jk} - \sum_{i=1}^{j-1} l_{ji}u_{ik} \tag{4.1}$$

a za $j > k$ (tj. za donji trokut)

$$l_{jk} = \frac{a_{jk} - \sum_{i=1}^{k-1} l_{ji}u_{ik}}{u_{kk}}. \tag{4.2}$$

Promotrimo što se događa prilikom računanja suma u (4.1) i (4.2) koje možemo zamisliti kao računanje skalarnog produkta. Prema teoremu 1.3 vidimo da je

$$u_{jk} = \left(a_{jk} - \sum_{i=1}^{j-1} l_{ji}u_{ik}(1 + \delta_i)\right)(1 + \delta')$$

gdje je $|\delta_i| \leq \gamma_{j-1}$ i $\delta' \leq u$.
Izlučivanjem a_{jk} iz prethodne jednakosti (i uzimajući u obzir da je $l_{jj} = 1$) vidimo da je

$$\begin{aligned} a_{jk} &= \tfrac{1}{1+\delta'}u_{jk}l_{jj} + \textstyle\sum_{i=1}^{j-1} l_{ji}u_{ik}(1 + \delta_i) \\ &= \textstyle\sum_{i=1}^{j} l_{ji}u_{ik} + \sum_{i=1}^{j} l_{ji}u_{ik}\delta_i \\ &\equiv \textstyle\sum_{i=1}^{j} l_{ji}u_{ik} + E_{jk} \end{aligned}$$

gdje je $|\delta_i| \leq \gamma_{j-1}$ i $1 + \delta_j \equiv \frac{1}{1+\delta'}$.
Sada moramo ograditi E_{jk}. To ćemo učiniti na sljedeći način

$$|E_{jk}| = \left| \sum_{i=1}^{j} l_{ji} u_{ik} \delta_i \right| \leq \sum_{i=1}^{j} |l_{ji}||u_{ik}|\gamma_n = \gamma_n (|L||U|)_{jk},$$

gdje $|A|$ definiramo tako da je $|A|_{ij} = |a_{ij}|$ za neku matricu $A = [a_{ij}] \in \mathbb{R}^{n \times n}$.

Koristeći taj rezultat i činjenicu da je $A = \widetilde{L}\widetilde{U} + E$, možemo zaključiti da je $|E| \leq \gamma_n |\widetilde{L}||\widetilde{U}|$. Primjenjujući matričnu normu na dobivenu nejednakost dobivamo da je $\|E\| \leq \| |\widetilde{L}| \| \| |\widetilde{U}| \|$ (Frobeniusova, beskonačna i 1-norma ne ovise o predznaku matričnih elemenata pa kod takvih normi možemo izostaviti znak apsolutne vrijednosti). Nakon provedene LU-dekompozicije, rješavamo sustave $Ly = b$ i $Ux = y$ pomoću supstitucija unaprijed i unazad. Zbog povratne stabilnosti algoritama 2.1 i 2.2 možemo zaključiti sljedeće

$$\begin{aligned} (\widetilde{L} + \Delta L)\tilde{y} &= b \quad \text{, gdje je } |\Delta L| \leq \gamma_n |\widetilde{L}| \\ (\widetilde{U} + \Delta U)\tilde{x} &= \tilde{y} \quad \text{, gdje je } |\Delta U| \leq \gamma_n |U| \end{aligned} \tag{4.3}$$

Kombinirajući prethodne rezultate dobivamo da je

$$\begin{aligned} b &= (\widetilde{L} + \Delta L)\tilde{y} = (\widetilde{L} + \Delta L)(\widetilde{U} + \Delta U)\tilde{x} = (\widetilde{L}\widetilde{U} + \Delta L\widetilde{U} + \widetilde{L}\Delta U + \Delta L \Delta U)\tilde{x} \\ &= (A - E + \Delta L\widetilde{U} + \widetilde{L}\Delta U + \Delta L \Delta U)\tilde{x} = (a + \Delta A)\tilde{x} \end{aligned} \tag{4.4}$$

gdje je $\Delta A = -E + \Delta L\widetilde{U} + \widetilde{L}\Delta U + \Delta L \Delta U$. Sada još moramo odrediti ogradu za ΔA

$$\begin{aligned} |\Delta A| &= | - E + \Delta L\widetilde{U} + \widetilde{L}\Delta U + \Delta L \Delta U| \leq |E| + |\Delta L\widetilde{U}| + |\widetilde{L}\Delta U| + |\Delta L \Delta U| \\ &\leq |E| + |\Delta L||\widetilde{U}| + |\widetilde{L}||\Delta U| + |\Delta L||\Delta U| \\ &\leq 3\gamma_n |\widetilde{L}||\widetilde{U}| + \gamma_n^2 |\widetilde{L}||\widetilde{U}| \\ &= (3\gamma_n + \gamma_n^2)|\widetilde{L}||\widetilde{U}| \end{aligned}$$

□

4.2. Važnost pivotiranja i elementarna analiza greške

Kao što je pokazano u točki 3.3.1., algoritam *GEBP* za rješavanje SLJ kod nekih sustava može dati rješenje koje značajno odstupa od pravog. Sada ćemo preciznije odrediti koji su to sustavi i na primjeru ilustrirati važnost pivotiranja. U tu svrhu definirat ćemo nekoliko pojmova.

Definicija 4.1 ***Matrična norma*** *na $\mathbb{R}^{n \times n}$ je preslikavanje $\| \cdot \| : \mathbb{R}^{n \times n} \to \mathbb{R}$ za koje vrijedi:*

i) $\|A\| \geq 0$ *za sve* $A \in \mathbb{R}^{n\times n}$ *i* $\|A\| = 0$ *ako i samo ako je* $A = 0$

ii) $\|\lambda A\| = |\lambda|\|A\|$ *za sve* $\lambda \in \mathbb{R}$, *i* $A \in \mathbb{R}^{n\times n}$

iii) $\|A + B\| \leq \|A\| + \|B\|$ *za sve* $A, B \in \mathbb{R}^{n\times n}$

iv) $\|AB\| \leq \|A\|\|B\|$ *za sve* $A, B \in \mathbb{R}^{n\times n}$

Primjer 4.1 *Može se pokazati da je preslikavanje* $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ *zadano izrazom*

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

zadovoljava sva svojstva iz definicije 4.1. Takvo preslikavanje zovemo matrična norma inducirana vektorskom normom ∞ *(vidi [3]).*

Definicija 4.2 *Neka je* $A \in \mathbb{R}^{n\times n}$ *regularna matrica i* $b \in \mathbb{R}^n$ *te neka je* $x \in \mathbb{R}^n$ *točno rješenje sustava*

$$Ax = b$$

a $\tilde{x}$ *približno rješenje dobiveno nekom numeričkom metodom. Vektor definiran izrazom*

$$r = b - A\tilde{x}$$

naziva se ***rezidualni vektor*** *ili* ***rezidual****.*

Ocjenimo sada relativnu pogrešku pomoću reziduala. Budući da je

$$x - \tilde{x} = A^{-1}b - \tilde{x} = A^{-1}(b - A\tilde{x}) = A^{-1}r$$

koristeći svojstvo *iv)* za matrične norme iz definicije 4.1, vidimo da je

$$\frac{\|x - \tilde{x}\|}{\|x\|} = \frac{\|A^{-1}r\|}{\|x\|} \leq \|A^{-1}\|\frac{\|r\|}{\|x\|} \tag{4.5}$$

Kako je $Ax = b$, koristeći svojstvo *iv)* dobivamo $\|b\| \leq \|A\|\|x\|$ što daje

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|} \tag{4.6}$$

Iz (4.5) i (4.6) slijedi da je

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|r\|}{\|b\|} \tag{4.7}$$

Na sličan način se može pokazati da vrijedi

$$\frac{\|x - \tilde{x}\|}{\|x\|} \geq \frac{1}{\|A\|\|A^{-1}\|}\frac{\|r\|}{\|b\|} \tag{4.8}$$

što na koncu daje donju i gornju ogradu za relativnu pogrešu izračunatog rješenja

$$\frac{1}{\|A\|\|A^{-1}\|}\frac{\|r\|}{\|b\|} \leq \frac{\|x-\tilde{x}\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|r\|}{\|b\|} \tag{4.9}$$

Kao što vidimo, veličina relativne pogreške rješenja SLJ ne ovisi samo o normi rezidualnog vektora već i o veličini broja $\|A\|\|A^{-1}\|$.

Definicija 4.3 ***Uvjetovanost matrice*** *A s oznakom $\kappa(A)$ definiramo na sljedeći način:*

$$\kappa(A) = \begin{cases} \|A\|\|A^{-1}\|, & \textit{ako je } A \textit{ regularna} \\ +\infty, & \textit{inače} \end{cases}$$

Iz (4.9) vidimo da veličina relativne pogreške ovisi o uvjetovanosti matrice sustava A. Što je uvjetovanost matrice A veća, veća je i relativna pogreška izračunatog rješenja. Za sustave linearnih jednadžbi kažemo da su loše uvjetovani ako je uvjetovanost matrice sustava velika.

Primjer 4.2 *Neka je*

$$A = \begin{bmatrix} 8\cdot 10^{-7} & 1 & 0 & 1 & 0 & 1 \\ 0 & -4\cdot 10^{-10} & 1 & 1 & 1 & 0 \\ -3\cdot 10^{-17} & 1 & 0 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 10^{-15} & 0 \\ 1 & 1 & -1 & -1 & 1 & 10^{-11} \\ 1 & 1 & 1 & 0 & 6\cdot 10^{-8} & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 3+8\cdot 10^{-7} \\ 3-4\cdot 10^{-10} \\ -3\cdot 10^{-17} \\ 10^{-15} \\ 1+10^{-11} \\ 4+6\cdot 10^{-8} \end{bmatrix}$$

Oderdimo rješenje sustava

$$Ax = b.$$

Kondicija matrice sustava iz primjera 4.2 iznosi $\kappa(A) = 10.674$ pa ne možemo reći da je sustav loše uvjetovan. Rješenje tog sustava je $x = (1,1,1,1,1,1)^T$. Primjenom algoritma $GEBP$ dobivamo da rješenje iznosi

$$\tilde{x} = (-0.74496, 3.28571, 2.42857, -1.00000, 1.57143, 0.71429)^T$$

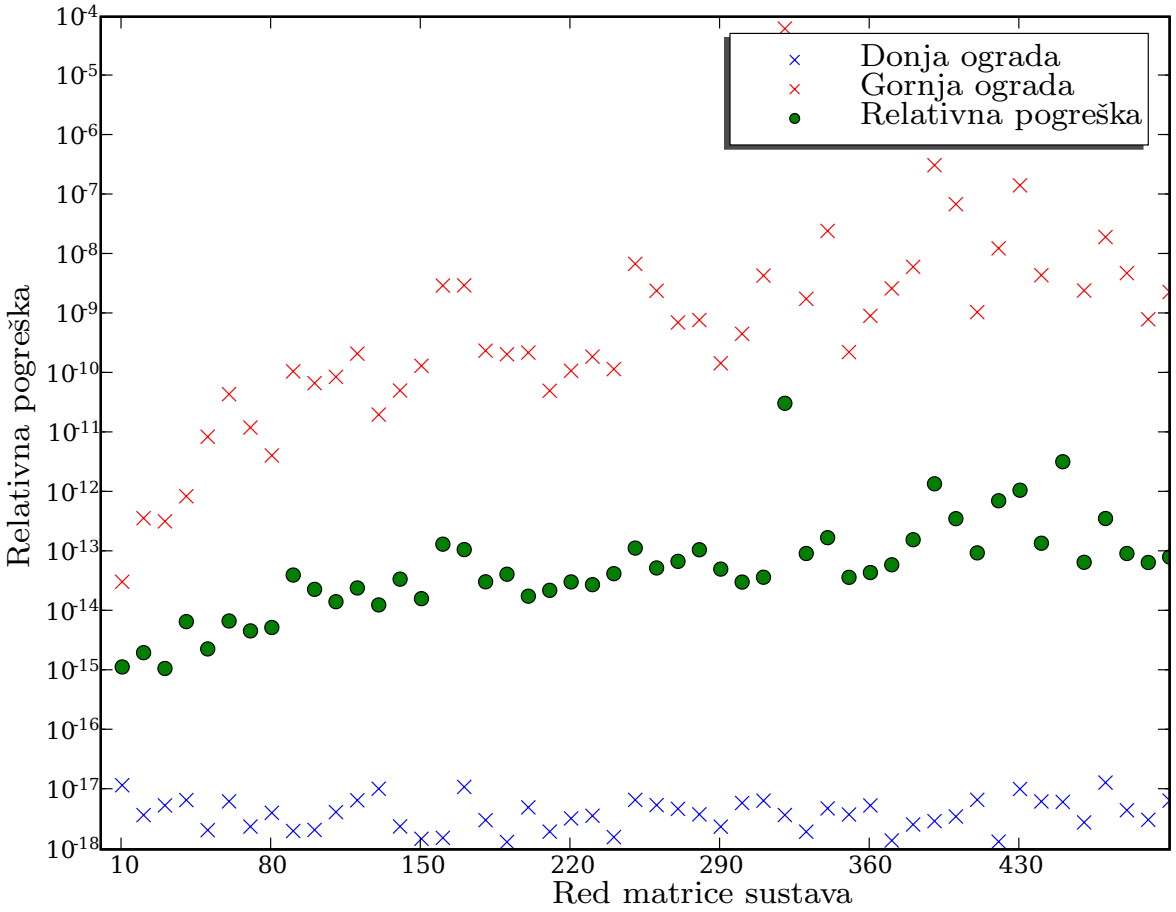
pri čemu norma rezidualnog vektora iz definicije 4.2 iznosi 2.4559. Iz činjenice da relativna pogreška dobivenog rješenja iznosi 0.83267 (pri čemu gornja ograda iz (4.9) iznosi 4.4309, a donja 0.038893) vidimo da je rješenje dobiveno algoritmom $GEBP$ netočno.

Problem koji se javlja prilikom izvođenja algoritma $GEBP$ sličan je problemu opisanom u 3.3.1..

Primjernom algoritma 3.2 dobivamo rješenje

$$\tilde{x} = (1.00000, 1.00000, 1.00000, 1.00000, 1.00000, 1.00000)$$

Slika 4.1: Relativna pogreška rješenja sustava SLJ s donjom i gornjom ogradom iz (4.9)



čija relativna pogreška iznosi $5.2271 \cdot 10^{-16}$ što je značajno manje od pogreške dobivene rješavanjem sustava pomoću $GEBP$.

U primjeru 4.2 vidjeli smo da za pojedine sustave linearnih jednadžbi algoritam $GEDP$ daje točnije rješenje od algoritma $GEBP$ što ukazuje na važnost pivotiranja. Kako bi prikazali veličine relativne pogreške kod izračunavanja rješenja SLJ pomoću algoritma 3.2, koristimo rezultat (4.9) na nizu slučajno generiranih SLJ. Na slici 4.1 prikazani su rezultati provedene analize.

5. Zaključak

Zbog izvesne numeričke stabilnosti algoritam *GEDP* pokazao se mnogo robusnijim i pouzdanijim od algoritma bez pivotiranja. Algoritam *GEDP* se gotovo najčešće koristi za rješavanje SLJ jer zahtjeva isti broj operacija kao i običan algoritam, samo što se vrši dodatno permutiranje redaka koje u slučaju velikog sustava može znatno usporiti proces rješavanja. Rješenje tog problema daje nam algoritam bez fizičke zamjene redaka koji je prikazan u točki (3.3.5.). Osnovni cilj ovog rada bio je prikazati algoritam Gaussovih eliminacija za rješavanje općih (tj. "gustih") sustava linearnih jednadžbi zajedno sa svim poteškoćama koje se javljaju prilikom njegove implementacije na računalu.

Literatura

[1] James W. Demmel, *Applied Numerical Linear Algebra*, SIAM Press, Philadelphia, 1997.

[2] P.E. Gill, W. Murray, M.H. Wright, *Numerical Linear Algebra and Optimization, Vol.1*, Addison-Wesley, Redwood City, CA, 1991.

[3] Gene H. Golub, Charles F. Van Loan, *Matrix computations*, The Johns Hopkins University Press, Baltimore and London, 1996.

[4] Nicholas J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM Press, Philadelphia, 1996.

[5] Rudolf Scitovski, *Numerička matematika*, Odjel za matematiku, Osijek, 2004.

[6] Ninoslav Truhar, *Numerička linearna algebra (interna skripta Odjela za matematiku)*, Osijek, 2006.

[7] Jochen Voss, *Numerical Linear Algebra*, University of Warwick, December 2004.