Točnost i numerička stabilnost direktnih metoda za rješavanje sustava linearnih jednadžbi

Šenjug, Diana

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Diana Šenjug

TOČNOST I NUMERIČKA STABILNOST DIREKTNIH METODA ZA RJEŠAVANJE SUSTAVA LINEARNIH JEDNADŽBI

Diplomski rad

Voditelj rada:
Doc. dr. sc. Nela Bosner

Zagreb, Studeni 2018

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Sadržaj **iii**

Uvod **3**

1 Teorija perturbacije **4**
1.1 Opis problema 4
1.2 Analiza po normi 4
1.3 Analiza po komponentama 8
1.4 Skaliranje radi minimizacije broja uvjetovanosti 14
1.5 Numerička stabilnost 17
1.6 Praktične ograde grešaka 18
1.7 Teorija perturbacije po diferencijalnom računu 20

2 Trokutasti sustavi **22**
2.1 Opis 22
2.2 Analiza greške unatrag 23
2.3 Analiza greške unaprijed 25

3 LU faktorizacija i linearne jednadžbe **30**
3.1 Gaussove eliminacije i strategija pivotiranja 30
3.2 LU faktorizacija 32
3.3 Analiza greške 36
3.4 Faktor rasta 40
3.5 Skaliranje i izbor pivotne strategije 41

Bibliografija **43**

Uvod

Glavna tema ovog rada je točnost koju možemo očekivati kada sustav linearnih jednadžbi $Ax = b$, gdje je $A \in \mathbb{R}^{n\times n}$, rješavamo u aritmetici konačne preciznosti na računalu. U prvom poglavlju obrađujemo teoriju perturbacije za sustave linearnih jednadžbi u kojoj se daje odgovor na pitanje koliko je rješenje sustava osjetljivo na perturbacije ulaznih podataka A i b. Kroz klasične rezultate analize po normi vidimo da je broj uvjetovanosti po normi povezan s načinom mjerenja permutacije. Također, tu definiramo i grešku unatrag po normi. Kako bi proveli analizu po komponentama definiramo grešku unatrag po komponentama. Pokazujemo i odnos između komponentnog broja uvjetovanosti i broja uvjetovanosti po normi. Nadalje, radimo analizu kojom vidimo da teorija perturbacije daje povratnu grešku izračunate aproksimacije rješenja, i procjenjuje ogradu za njezinu grešku unaprijed. U drugom poglavlju proučavamo trokutaste sustave jer oni imaju temeljnu ulogu u računanju s matricama. Izvodimo ograde greške unaprijed i unatrag. Treće poglavlje posvetili smo analizi numeričke stabilnosti direktnih metoda za rješavanje sustava, baziranih na LU faktorizaciji. Ta analiza daje povratnu grešku za dobivenu aproksimaciju rješenja preko direktne metode, koja se potom uklapa u teoriju perturbacije za dobivanje greške unaprijed. Rezultat analize diktira koji parametri mogu utjecati na numeričku stabilnost algoritma.
Za lakše razumijevanje rada navedimo definicije i rezultate koji će nam trebati u radu.

Definicija 0.0.1. *Kažemo da je vektor x dual od y ako vrijedi*

$$x^*y = \|x\|_D\|y\| = 1$$

gdje je s $\|x\|_D = \max_{y\neq 0} \frac{|y^*x|}{\|y\|}$ *definirana dualna norma, a* $\|\cdot\|$ *je proizvoljna vektorska norma.*

Definicija 0.0.2. *Norma je apsolutna ako za svaku matricu* $A \in \mathbb{R}^{n\times m}$ *vrijedi* $\| |A| \| = \|A\|$.

Definicija 0.0.3. *Broj uvjetovanosti matrice A dan je s* $\kappa(A) = \|A\| \, \|A^{-1}\|$.

Definicija 0.0.4. *Neka je* $A \in \mathbb{C}^{n\times n}$. *Tada je sa* $\rho(A) = max\{|\lambda| : \lambda \in \sigma(A)\}$ *definiran spektralni radijus matrice A.*

Definicija 0.0.5. *Pseudo inverz od $A \in \mathbb{R}^{m\times n}$ je jedinstvena matrica $A^+ \in \mathbb{R}^{m\times n}$ koja zadovoljava iduća četiri svojstva:*

1. $AA^+A = A$
2. $A^+AA^+ = A^+$
3. $(AA^+)^T = AA^+$
4. $(A^+A)^T = A^+A$

Definicija 0.0.6. *Hölderova p-norma definirana je s*

$$\|x\|_p = \Big(\sum_{i=1}^{n} |x_i|^p\Big)^{1/p} \quad ,za \quad p \geq 1.$$

Za Hölderove p-norme vrijedi i poznata Hölderova nejednakost

$$|x^*y| \leq \|x\|_p \|y\|_p \,, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

Lema 0.0.7. *Ako imamo $|\delta_i| \leq u$ (u je jedinična greška zaokruživanja) i $\rho_i = \pm 1$, za $i = 1 : n$, pri čemu je $nu < 1$, tada vrijedi*

$$\prod_{i=1}^{n} (1 + \delta_i)^{\rho_i} = 1 + \theta_n$$

uz ocjenu

$$|\theta_n| \leq \frac{nu}{1 - nu} := \gamma_n.$$

- Higham je pokazao [10, p. 67] da brojevi γ_n , $n \geq 1$ imaju svojstva:

$$\begin{aligned} \gamma_m + \gamma_n + \gamma_m\gamma_n &\leq \gamma_{m+n} \\ i\gamma_k &\leq \gamma_{ik} \\ \gamma_k + u &\leq \gamma_{k+1} \\ \gamma_k\gamma_j &\leq \gamma_{\min(k,j)} \text{ za } \max(j,k)u \leq 1/2 \end{aligned}$$

Napomena 0.0.8. *Koliko se dvije p-norme vektora mogu razlikovati vidi se po nejednakosti*

$$\|x\|_{p_2} \leq \|x\|_{p_1} \leq n^{(\frac{1}{p_1} - \frac{1}{p_2})} \|x\|_{p_2} \,, \; p_1 \leq p_2$$

Napomena 0.0.9. *Matrična p-norma je norma podređena Hölderovoj p-normi:*

$$||A||_p = \max_{x \neq 0} \frac{||Ax||_p}{||x||_p}\ ,\ p \geq 1.$$

Napomena 0.0.10. *Za* $p = 1, 2, \infty$ *vrijedi* $||A^*||_p = ||A||_q\ ,\ \ \frac{1}{p} + \frac{1}{q} = 1.$

Lema 0.0.11. *Ako uzmemo* $x = e_j$ *u Napomeni 0.0.8, te koristimo Napomenu 0.0.9 i Napomenu 0.0.10, možemo izvesti ograde za* $A \in \mathbb{C}^{m \times n}$

$$\max_j ||A(:, j)||_p \leq ||A||_p \leq n^{1-1/p} \max_j ||A(:, j)||_p$$
$$\max_i ||A(i, :)||_{p/p-1} \leq ||A||_p \leq m^{1/p} \max_i ||A(i, :)||_{p/p-1}.$$

Dokaz. [10, p. 112] □

Poglavlje 1

Teorija perturbacije

1.1 Opis problema

Proučavamo linearan sustav $Ax = b$, gdje je $A \in \mathbb{R}^{n\times n}$. Postoje tri važna pitanja u kontekstu neizvjesnih podataka ili neispravne aritmetike.

- Ako perturbiramo A ili b, koliko se mijenja x, to jest, koliko je osjetljivo rješenje ako perturbiramo podatke?
- Koliko treba perturbirati podatke A i b da bi približno rješenje y bilo jednako točnom rješenju perturbiranog sustava, to jest, kolika je greška unatrag od y ?
- Koliku ogradu bi u praksi trebali računati za grešku unaprijed danog aproksimativnog rješenja?

1.2 Analiza po normi

Za početak ćemo predstaviti klasične rezultate analize po normi. Sa $\|\cdot\|$ označavamo normu vektora i odgovarajuću operatorsku normu. $\kappa(A) = \|A\|\,\|A^{-1}\|$ je broj uvjetovanosti matrice. Kroz rad, matrica E i vektor f su proizvoljni i predstavljaju toleranciju prema kojoj su mjerene perturbacije. Uloga matrice E i vektora f biti će jasnija u analizi po komponentama. Naš prvi teorem će pokazati da imamo "dobro" aproksimativno rješenje ako je rezidual mali.

Teorem 1.2.1 (Rigal, Gaches)**.** *Greška unatrag po normi*

$$\eta_{E,f}(y) = min\Big\{\epsilon : (A + \Delta A)y = b + \Delta b, \quad \|\Delta A\| \leq \epsilon\|E\|, \quad \|\Delta b\| \leq \epsilon\|f\|\Big\} \tag{1.1}$$

dana je s

$$\eta_{E,f}(y) = \frac{||r||}{||E||\,||y|| + ||f||} \tag{1.2}$$

gdje je $r = b - Ay$

Dokaz. Pokažimo da je desna strana (1.2) donja granica za $\eta_{E,f}(y)$:

$$\begin{aligned}(A + \Delta A)y &= b + \Delta b\\ Ay + \Delta Ay &= b + \Delta b\\ \Delta Ay - \Delta b &= b - Ay = r\end{aligned}$$

Djelujemo s normom pa imamo:

$$||r|| = ||\Delta Ay - \Delta b|| \leq ||\Delta A||\,||y|| + ||\Delta b|| \leq \epsilon||E||\,||y|| + \epsilon||f||$$

to jest

$$||b - Ay|| \leq \epsilon(||E||\,||y|| + ||f||),$$

iz čega slijedi:

$$\epsilon \geq \frac{||b - Ay||}{||E||\,||y|| + ||f||}.$$

Sada je

$$\begin{aligned}\eta_{E,f}(y) &= min\Big\{\epsilon : (A + \Delta A)y = b + \Delta b, \quad ||\Delta A|| \leq \epsilon||E||, \quad ||\Delta b|| \leq \epsilon||f||\Big\}\\ &\geq \frac{||r||}{||E||\,||y|| + ||f||}.\end{aligned}$$

Uz $r = b - Ay$ slijedi da je desna strana (1.2) donja granica za $\eta_{E,f}(y)$.

Pokažimo da se ova donja granica dostiže za perturbacije

$$\Delta A_{min} = \frac{||E||\,||y||}{||E||\,||y|| + ||f||} rz^T, \quad \Delta b_{min} = -\frac{||f||}{||E||\,||y|| + ||f||} r \tag{1.3}$$

gdje je z dualni vektor od y.

$$\begin{aligned}||\Delta A_{min}|| &= \frac{||E||\,||y||}{||E||\,||y|| + ||f||}||rz^T|| = \frac{||E||\,||y||}{||E||\,||y|| + ||f||}\max_{v\neq 0}\frac{||rz^Tv||}{||v||} =\\ &= \frac{||E||\,||y||}{||E||\,||y|| + ||f||}||r||\max_{v\neq 0}\frac{|z^Tv|}{||v||} = \frac{||E||\,||r||}{||E||\,||y|| + ||f||}||y||\,||z||_D\\ &= \frac{||r||}{||E||\,||y|| + ||f||}||E||\end{aligned}$$

$$||\Delta b_{min}|| = \frac{||f||}{||E||\,||y|| + ||f||}||r|| = \frac{||r||}{||E||\,||y|| + ||f||}||f||$$

I za $||\Delta A_{\min}||$ i za $||\Delta b_{\min}||$ se vidi da je $\epsilon = \frac{||r||}{||E||\,||y||+||f||}$.
Pokažimo još da vrijedi $(A + \Delta A_{min})y = b + \Delta b_{min}$.

$$\begin{aligned}
Ay + \Delta A_{min}y &= b + \Delta b_{min} \\
Ay + \frac{||E||\,||y||}{||E||\,||y|| + ||f||}rz^T y &= b - \frac{||f||}{||E||\,||y|| + ||f||}r \\
\frac{||E||\,||y||}{||E||\,||y|| + ||f||}r &= b - Ay - \frac{||f||}{||E||\,||y|| + ||f||}r \\
\frac{||E||\,||y||}{||E||\,||y|| + ||f||}r &= r - \frac{||f||}{||E||\,||y|| + ||f||}r \\
\frac{||E||\,||y||}{||E||\,||y|| + ||f||}r &= \Big(1 - \frac{||f||}{||E||\,||y|| + ||f||}\Big)r \\
\frac{||E||\,||y||}{||E||\,||y|| + ||f||}r &= \frac{||E||\,||y|| + ||f|| - ||f||}{||E||\,||y|| + ||f||}r
\end{aligned}$$

□

Ako izaberemo $E = A$ i $f = b$, tada $\eta_{E,f}(y)$ nazivamo relativna greška unatrag po normi.

Teorem 1.2.2. *Neka je $Ax = b$ i $(A + \Delta A)y = (b + \Delta b)$, gdje je $||\Delta A|| \leq \epsilon||E||$ i $||\Delta b|| \leq \epsilon||f||$. Pretpostavimo još da je $\epsilon||A^{-1}||\,||E|| \leq 1$. Tada je*

$$\frac{||x - y||}{||x||} \leq \frac{\epsilon}{1 - \epsilon||A^{-1}||\,||E||}\Big(\frac{||A^{-1}||\,||f||}{||x||} + ||A^{-1}||\,||E||\Big) \tag{1.4}$$

te se ta granica dostiže do na prvi red u ϵ

Dokaz. Pokažimo da granica (1.4) slijedi iz jednadžbe $A(y - x) = \Delta b - \Delta A x + \Delta A(x - y)$. Pomnožimo prethodnu jednadžbu s A^{-1} s lijeva, pa djelujemo s normom i koristimo njena svojstva:

$$\begin{aligned}
y - x &= A^{-1}\Delta b - A^{-1}\Delta A x + A^{-1}\Delta A(x - y) \\
||y - x|| &= ||A^{-1}\Delta b - A^{-1}\Delta A x + A^{-1}\Delta A(x - y)|| \\
||x - y|| &\leq ||A^{-1}||\,||\Delta b|| + ||A^{-1}||\,||\Delta A||\,||x|| + ||A^{-1}||\,||\Delta A||\,||x - y|| \\
||x - y|| &\leq \frac{||A^{-1}||\,||\Delta b|| + ||A^{-1}||\,||\Delta A||\,||x||}{1 - ||A^{-1}||\,||\Delta A||} \\
||x - y|| &\leq \frac{||A^{-1}||\,\epsilon||f|| + ||A^{-1}||\,\epsilon||E||\,||x||}{1 - ||A^{-1}||\,\epsilon||E||}
\end{aligned}$$

Dijeljenjem s $||x||$ i sređivanjem dobivamo

$$\frac{||x-y||}{||x||} \leq \frac{\epsilon}{1-\epsilon||A^{-1}||\,||E||}\Big(\frac{||A^{-1}||\,||f||}{||x||} + ||A^{-1}||\,||E||\Big)$$

Pokažimo još da je dobivena do na prvi red u ϵ za $\Delta A = \epsilon||E||\,||x||wv^T$ i $\Delta b = -\epsilon||f||w$, gdje je $||w|| = 1$, $||A^{-1}w|| = ||A^{-1}||$, a v je dualni vektor od x.

$$\begin{aligned} x - y &= -A^{-1}\Delta b + A^{-1}\Delta A x - A^{-1}\Delta A(x-y) = \\ &= \epsilon||f||A^{-1}w + \epsilon||E||\,||x||A^{-1}w(v^T x) - \epsilon||E||\,||x||A^{-1}w(v^T(x-y)) = \\ &= \epsilon\Big(||f|| + ||E||\,||x|| - ||E||\,||x||(v^T(x-y))\Big)A^{-1}w \end{aligned}$$

Uzmimo normu

$$\begin{aligned} ||x-y|| &= \epsilon\Big(||f|| + ||E||\,||x|| + O(\epsilon)\Big)||A^{-1}|| = \\ &= (\epsilon + O(\epsilon^2))\Big(||A^{-1}||\,||f|| + ||A^{-1}||\,||E||\,||x||\Big) \end{aligned}$$

i podijelimo s $||x||$

$$\frac{||x-y||}{||x||} = (\epsilon + O(\epsilon^2))\Big(\frac{||A^{-1}||\,||f||}{||x||} + ||A^{-1}||\,||E||\Big).$$

□

U oba prethodna teorema je idući broj uvjetovanosti po normi povezan s načinom mjerenja perturbacije
$\kappa_{E,f}(A,x) := \limsup_{\epsilon\to 0}\left\{\frac{||\Delta x||}{\epsilon||x||} : (A+\Delta A)(x+\Delta x) = b + \Delta b,\ ||\Delta A|| \leq \epsilon||E||,\ ||\Delta b|| \leq \epsilon||f||\right\}$

Obzirom da je granica u Teoremu 1.2.2. oštra, imamo

$$\begin{aligned} \kappa_{E,f}(A,x) &= \limsup_{\epsilon\to 0}\left\{\frac{||\Delta x||}{\epsilon||x||} : (A+\Delta A)(x+\Delta x) = b+\Delta b,\ ||\Delta A|| \leq \epsilon||E||,\ ||\Delta b|| \leq \epsilon||f||\right\} \\ &= \limsup_{\epsilon\to 0}\frac{1}{1-\epsilon||A^{-1}||\,||E||}\Big(\frac{||A^{-1}||\,||f||}{||x||} + ||A^{-1}||\,||E||\Big). \end{aligned}$$

Ako pustimo limes, član $\epsilon||A^{-1}||\,||E||$ teži u nulu. Dakle broj uvjetovanosti je

$$\kappa_{E,f}(A,x) = \frac{||A^{-1}||\,||f||}{||x||} + ||A^{-1}||\,||E||. \tag{1.5}$$

Ako izaberemo $E = A$ i $f = b$ imamo

$$\kappa_{A,b}(A, x) = \frac{||A^{-1}||\,||b||}{||x||} + ||A^{-1}||\,||A|| \\ \leq \frac{||A^{-1}||\,||A||\,||x||}{||x||} + ||A^{-1}||\,||A|| = 2||A^{-1}||\,||A|| = 2\kappa(A).$$

S druge strane, očito je $\kappa(A) \leq \kappa_{E,f}(A, x)$ za $E = A$ i $f = b$.
Dakle, vrijedi $\kappa(A) \leq \kappa_{E,f}(A, x) \leq 2\kappa(A)$. Ograda (1.4) može biti neznatno oslabljena kako bi se postigao poznati oblik

$$\frac{||x - y||}{||x||} \leq \frac{2\epsilon\kappa(A)}{1 - \epsilon\kappa(A)}$$

1.3 Analiza po komponentama

Greška unatrag po komponentama definirana je sa

$$\omega_{E,f}(y) = min\Big\{\epsilon : (A + \Delta A)y = b + \Delta b, |\Delta A| \leq \epsilon E, |\Delta b| \leq \epsilon f\Big\} \tag{1.6}$$

gdje za E i f sada pretpostavljamo da imaju nenegativne elemente. Podrazumijeva se da apsolutne vrijednosti i nejednakosti među matricama ili vektorima vrijede po komponentama. U ovoj definiciji greške unatrag svaki element perturbacije mjeren je relativno u odnosu na individualnu toleranciju. Dakle, za razliku od definicije po normi, u potpunosti koristimo $n^2 + n$ parametra u E i f.

Zanima nas sada kako odabrati E i f? Postoje četiri glavna odabira koja su nam interesantna.

- Najčešći izbor tolerancije je $E = |A|$ i $f = |b|$. On doprinosi relativnoj grešci unatrag po komponentama. U ovom slučaju u (1.6.) imamo

  $$a_{ij} = 0 \Rightarrow \Delta a_{ij} = 0 \text{ i } b_i = 0 \Rightarrow \Delta b_i = 0.$$

  Dakle, ako je $\omega_{E,f}(y)$ mali, onda y rješava problem blizak originalnom u smislu komponentne relativne perturbacije, te ima isti raspored netrivijalnih elemenata. Još jedno privlačno svojstvo komponentne relativne greške unatrag je neosjetljivost na skaliranje sustava: ako je $Ax = b$ skalirano tako da dobijemo $(S_1AS_2)(S_2^{-1}x) = S_1b$ gdje su S_1 i S_2 dijagonalne, te y skaliran na $S_2^{-1}y$, tada ω ostaje nepromijenjen. Pokažimo to. Gledajući $(A + \Delta A)y = b + \Delta b$ imamo

  $$S_1(A + \Delta A)S_2 \cdot S_2^{-1}y = S_1b + S_1\Delta b.$$

Pa je sada:

$$|S_1 \Delta A S_2| = |S_1|\,|\Delta A|\,|S_2| \leq \epsilon |S_1|\,|E|\,|S_2| = \epsilon |S_1|\,|A|\,|S_2| = \epsilon |S_1 A S_2|$$
$$|S_1 \Delta b| = |S_1|\,|\Delta b| \leq \epsilon |S_1|\,|f| = \epsilon |S_1|\,|b| = \epsilon |S_1 b|$$

- Retčanu grešku unatrag dobivamo izborom $E = |A|ee^T$, $f = |b|$. Ograda u $\omega_{E,f}$ je sada $|\Delta a_{i,j}| \leq \epsilon\alpha_i$, gdje je α_i 1-norma i-tog retka od A, pa se perturbacije i-tog retka od A računaju u odnosu na normu istog retka. Slično je i za b.

- Stupčanu grešku unatrag formuliramo na sličan način uzimajući $E = ee^T|A|$ i $f = \|b\|_1 e$. Perturbacije j-tog stupca od A računaju se u odnosu na 1-normu istog stupca.

- Prirodni odabir tolerancije je $E = \|A\|ee^T$ i $f = \|b\|e$, za koji je $\omega_{E,f}(y)$ isti kao i normalna greška unatrag $\eta_{E,f}(y)$ do na konstantu. Pokažimo to.
  Iz $|\Delta A| \leq \epsilon E = \epsilon \|A\| ee^T$, odnosno $|\Delta b| \leq \epsilon f = \epsilon \|b\| e$, koristeći definiciju apsolutne norme imamo

$$\|\Delta A\| = \| \,|\Delta A|\, \| \leq \epsilon \|A\|\, \|ee^T\|$$
$$\|\Delta b\| = \| \,|\Delta b|\, \| \leq \epsilon \|b\|\, \|e\|$$

  gdje je $\|ee^T\|$, odnosno $\|e\|$, konstanta.

Kao što imamo formulu u Teoremu 1.2.1. za grešku unatrag po normi, imamo i jednostavnu formulu za $\omega_{E,f}(y)$.

Teorem 1.3.1 (Oettli i Prager)**.** *Greška unatrag po komponentama dana je sa*

$$\omega_{E,f}(y) = \max_i \frac{|r_i|}{(E|y| + f)_i} \tag{1.7}$$

gdje je $r = b - Ay$ i $\zeta/0$ je nula ako je $\zeta = 0$, a ∞ u suprotnom.

Dokaz. Pokažimo da je desna strana (1.7) donja granica za $\omega(y)$:

$$(A + \Delta A)y = b + \Delta b$$
$$Ay + \Delta A y = b + \Delta b$$
$$\Delta A y - \Delta b = b - Ay = r$$

Djelujemo s $|\cdot|$, pa imamo:

$$|r| = |\Delta A y - \Delta b| \leq |\Delta A|\,|y| + |\Delta b| \leq \epsilon E|y| + \epsilon f = \epsilon(E|y| + f)$$

iz čega slijedi

$$\epsilon \geq \frac{|r_i|}{(E|y| + f)_i}.$$

Dakle,

$$\omega_{E,f}(y) \geq \max_i \frac{|r_i|}{(E|y| + f)_i}.$$

Pokažimo da se ova granica dostiže za perturbacije

$$\Delta A = D_1 E D_2, \Delta b = -D_1 f, \tag{1.8}$$

gdje je $D_1 = diag(r_i/(E|y| + f)_i)$ i $D_2 = diag(sign(y_i))$.

$$\begin{aligned} |\Delta A| = |D_1 E D_2| &= \Big[\Big|\frac{r_i}{(E|y| + f)_i} e_{ij} \cdot sign(y_i)\Big|\Big]_{i=1,\dots,n} \\ &= \Big[\frac{|r_i|}{(E|y| + f)_i}|e_{ij}|\Big]_{i=1,\dots,n} \leq \max_i \frac{|r_i|}{(E|y| + f)_i} E \end{aligned}$$

$$\begin{aligned} |\Delta b| = |-D_1 f| &= \Big[\Big|\frac{r_i}{(E|y| + f)_i} f\Big|\Big]_{i=1,\dots,n} = \\ &= \Big[\frac{|r_i|}{(E|y| + f)_i}|f_i|\Big]_{i=1,\dots,n} \leq \max_i \frac{|r_i|}{(E|y| + f)_i} f \end{aligned}$$

I za $|\Delta A_{\min}|$ i za $|\Delta b_{\min}|$ se vidi da je $\epsilon = \max_i \frac{|r_i|}{(E|y|+f)_i}$.
Pokažimo još da vrijedi $(A + \Delta A_{min})y = b + \Delta b_{min}$.

$$\begin{aligned} Ay + \Delta A_{min} y &= b + \Delta b_{min} \\ Ay + D_1 E D_2 y &= b - D_1 f \\ D_1 E D_2 y &= b - Ay - D_1 f \\ \Big[\frac{r_i}{(E\,|y| + f)_i}(E\,|y|)_i\Big]_{i=1,\dots,n} &= \Big[r_i - \frac{r_i}{(E\,|y| + f)_i} f_i\Big]_{i=1,\dots,n} \\ \Big[\frac{r_i}{(E\,|y| + f)_i}(E\,|y|)_i\Big]_{i=1,\dots,n} &= \Big[r_i \frac{(E\,|y| + f)_i - f_i}{(E\,|y| + f)_i}\Big]_{i=1,\dots,n} \end{aligned}$$

□

Idući teorem daje grešku unaprijed koja odgovara grešci unatrag po komponentama.

Teorem 1.3.2. *Neka je* $Ax = b$ *i* $(A + \Delta A)y = b + \Delta b$*, gdje je* $|\Delta A| \leq \epsilon E$ *i* $|\Delta b| \leq \epsilon f$*, i pretpostavimo da je* $\epsilon \| \, |A^{-1}| \, E \, \| < 1$ *gdje je* $\| \cdot \|$ *apsolutna norma. Tada*

$$\frac{\|x - y\|}{\|x\|} \leq \frac{\epsilon}{1 - \epsilon \| \, |A^{-1}| \, E\|} \frac{\| \, |A^{-1}|(E \, |x| + f)\|}{\|x\|}, \tag{1.9}$$

i za ∞*- normu se ova granica dostiže do na prvi red po* ϵ*.*

Dokaz. Pokažimo da granica (1.9) slijedi iz jednadžbe $A(y - x) = \Delta b - \Delta A x + \Delta A(x - y)$. Pomnožimo prethodnu jednadžbu s A^{-1} i primijenimo nejednakost trokuta:

$$\begin{aligned} y - x &= A^{-1}\Delta b - A^{-1}\Delta A x + A^{-1}\Delta A(x - y) \\ |y - x| &\leq |A^{-1}||\Delta b| + |A^{-1}| \, |\Delta A| \, |x| + |A^{-1}| \, |\Delta A| \, |x - y| \\ |y - x| &\leq \epsilon |A^{-1}| f + \epsilon |A^{-1}| \, E \, |x| + \epsilon |A^{-1}| \, E \, |x - y|. \end{aligned}$$

Sada je

$$\|x - y\|(1 - \epsilon \| |A^{-1}|E\|) \leq \epsilon(\| \, |A^{-1}| f + |A^{-1}| \, E \, |x| \, \|).$$

Dijeljenjem s $\|x\|$ i sređivanjem dobivamo

$$\frac{\|x - y\|}{\|x\|} \leq \frac{\epsilon}{1 - \epsilon \| \, |A^{-1}| \, E\|} \frac{\| \, |A^{-1}|(E \, |x| + f)\|}{\|x\|}.$$

Pokažimo još da se za beskonačnu normu ograda dostiže do na prvi red po ϵ, za $\Delta A = \epsilon D_1 E D_2$ i $\Delta b = -\epsilon D_1 f$, gdje je $D_2 = diag(sign(x_i))$ i $D_1 = diag(\zeta_j)$, $\zeta_j = sign(A^{-1})_{kj}$ i $\| \, |A^{-1}|(|E| \, |x| + f)\|_\infty = (|A^{-1}|(|E| \, |x| + f))_k$.

$$\begin{aligned} \left(A^{-1}\Delta A x - A^{-1}\Delta b\right)_k &= \sum_{i=1}^{n} \sum_{j=1}^{n} (A^{-1})_{kj} (\Delta A)_{ji} x_i - \sum_{j=1}^{n} (A^{-1})_{kj} (\Delta b)_j = \\ &= \epsilon(|A^{-1}| \, E \, |x| + |A^{-1}| \, f)_k = \\ &= \epsilon \| \, |A^{-1}| \, E \, |x| + |A^{-1}| \, f\|_\infty. \end{aligned}$$

S druge strane, iz

$$y - x = A^{-1}\Delta b - A^{-1}\Delta A x + A^{-1}\Delta A(x - y)$$

slijedi

$$(A^{-1}\Delta A x - A^{-1}\Delta b)_k = -(y - x - A^{-1}\Delta A(x - y))_k.$$

Sada prema

$$\epsilon|| \, |A^{-1}| \, (E \, |x| + f)||_\infty = ||(I + A^{-1}\Delta A)(x - y)||_\infty = ||x - y||_\infty(1 + O(\epsilon))$$

imamo

$$\begin{aligned} ||x - y||_\infty &= \frac{\epsilon|| \, |A^{-1}| \, (E \, |x| + f)||_\infty}{1 + O(\epsilon)} = \\ &= \epsilon(1 + O(\epsilon))|| \, |A^{-1}| \, (E \, |x| + f)||_\infty = \\ &= (\epsilon + O(\epsilon^2))|| \, |A^{-1}| \, (E \, |x| + f)||_\infty \end{aligned}$$

Dijeljenjem s $||x||_\infty$ dobivamo

$$\frac{||x - y||_\infty}{||x||_\infty} = (\epsilon + O(\epsilon^2))\Big(\frac{|| \, |A^{-1}| \, (E \, |x| + f)||_\infty}{||x||_\infty}\Big).$$

□

Teorem 1.3.2. implicira da je broj uvjetovanosti

$$cond_{E,f}(A, x) := \limsup_{\epsilon \to 0} \Big\{ \frac{||\Delta x||_\infty}{\epsilon ||x||_\infty} : (A + \Delta A)(x + \Delta x) = b + \Delta b, \, |\Delta A| \leq \epsilon |E|, \, |\Delta b| \leq \epsilon |f| \Big\}$$

dan sa

$$cond_{E,f}(A, x) = \frac{|| \, |A^{-1}|(|E| \, |x| + f)||_\infty}{||x||_\infty} \tag{1.10}$$

Broj uvjetovanosti ovisi o x ili ekvivalentno, na desnoj strani ovisi o b. U najgorem slučaju, mjera osjetljivosti koja se odnosi na svaki x je

$$cond_{E,f}(A) = \max_x cond_{E,f}(A, x).$$

Za poseban slučaj kada je $E = |A|$ i $f = |b|$, broj uvjetovanosti je uveo Skeel [1] u obliku :

$$cond(A, x) := \frac{|| \, |A^{-1}| \, |A| \, |x| \, ||_\infty}{||x||_\infty} \tag{1.11}$$

$$cond(A) := cond(A, e) = || \, |A^{-1}| \, |A| \, ||_\infty \leq \kappa_\infty(A) \tag{1.12}$$

Ovi brojevi uvjetovanosti se razlikuju od $cond_{|A|,|b|}(A, x)$ i $cond_{|A|,|b|}(A)$ za najviše faktor 2.

Kako uspoređujemo *cond* s κ ? Obzirom da je *cond*(A) invarijantan na retčano skaliranje $Ax = b \to (DA)x = Db$ (tj. $cond(AD) = || \, |A^{-1}| \, |D^{-1}| \, |D| \, |A| \, ||_\infty = || \, |A^{-1}| \, |A| \, ||_\infty =$

cond(*A*)) gdje je D dijagonalna, *cond*(A) može biti proizvoljno manji od $\kappa_\infty(A)$. Točnije, pokažimo da je

$$\min\left\{\kappa_\infty(DA) : D \text{ dijagonalna}\right\} = cond(A) \tag{1.13}$$

gdje optimalno skaliranje D_R uravnotežuje redove od A, to jest, D_RA ima redove jedinične norme ($D_R|A|e = e$).
Uz

$$\|D_RA\|_\infty = \|D_R|A|e\|_\infty = \|e\|_\infty = 1,$$

za κ_∞ vrijedi

$$\begin{aligned}\kappa_\infty(D_RA) &= \|D_RA\|_\infty\|A^{-1}D_R^{-1}\|_\infty = \|A^{-1}D_R^{-1}\|_\infty = \|\,|A^{-1}D_R^{-1}|\,\|_\infty = \|\,|A^{-1}|D_R^{-1}\|_\infty \\ &= \|\,|A^{-1}|D_R^{-1}e\|_\infty.\end{aligned}$$

S druge strane, jer je

$$|A|e = \begin{bmatrix}\sum_j |a_{ij}| \\ \vdots \\ \sum_j |a_{nj}|\end{bmatrix} = \begin{bmatrix}\frac{1}{D_R(1,1)} \\ \vdots \\ \frac{1}{D_R(n,n)}\end{bmatrix} = D_Re,$$

imamo

$$cond(A) = \|\,|A^{-1}|\,|A|\,\|_\infty = \|\,|A^{-1}|\,|A|e\|_\infty = \|\,|A^{-1}|\,D_R^{-1}e\|_\infty$$

čime smo pokazali (1.13).

Chandrasekaran i Ipsen [2] primijetili su iduću nejednakost. Sa D_R ,definiranim kao gore, je

$$\frac{\kappa_\infty(A)}{\kappa_\infty(D_R)} \le cond(A) \le \kappa_\infty(A). \tag{1.14}$$

Prema tome, *cond*(*A*) može biti puno manji od $\kappa_\infty(A)$ samo kada su redovi od A loše skalirani. Nadalje, ako D_C uravnotežuje stupce od A ($e^T|A|D_C = e^T$) onda

$$\frac{\kappa_1(A)}{n\kappa_\infty(D_C)} \min_j \frac{\|A^{-1}e_j\|_\infty}{\|A^{-1}\|_1} \le cond(A,x) \le \kappa_\infty(A).$$

Ove nejednakosti pokazuju da *cond*(A, x) može biti puno manji od $\kappa_\infty(A)$ samo kada su stupci bilo od A ili A^{-1} loše skalirani.

Proučimo Kahanov [3] primjer. Neka je

$$A = \begin{bmatrix}2 & -1 & 1 \\ -1 & \epsilon & \epsilon \\ 1 & \epsilon & \epsilon\end{bmatrix}, b = \begin{bmatrix}2(1+\epsilon) \\ -\epsilon \\ \epsilon\end{bmatrix}$$

gdje je $0 < \epsilon \ll 1$, pa je $x = [\epsilon, -1, 1]^T$ rješenje sustava $Ax = b$. Broj uvjetovanosti po normi $\kappa_\infty(A)$ je $2(1 + \epsilon^{-1})$, pa je sustav jako osjetljiv na proizvoljne perturbacije u A i b. Štoviše,

$$|A^{-1}||A| = \begin{bmatrix} 1 & \epsilon & \epsilon \\ \frac{2\epsilon+1}{2\epsilon} & 1 & 1 \\ \frac{2\epsilon+1}{2\epsilon} & 1 & 1 \end{bmatrix}$$

pa je $cond(A) = 3 + (2\epsilon)^{-1}$, što sugerira da je sustav također vrlo osjetljiv na komponentne perturbacije za neke desne strane. Međutim, $cond(A, x) = 5/2 + \epsilon$, pa je za ovaj specifični b sustav jako dobro uvjetovan pod komponentnom perturbacijom.

Recimo sada nešto o izboru broja uvjetovanosti. Za linearni sustav je svaki broj uvjetovanosti definiran s obzirom na specifičnu klasu perturbacija. Važno je upotrijebiti dobar broj uvjetovanosti u danom slučaju. Na primjer, ako je $\hat{x}$ izračunato rješenje od $Ax = b$ i ako znamo da je greška unatrag po normi $\eta_{A,b}(\hat{x})$, tada je broj uvjetovanosti $\kappa(A)$, koji se pojavljuje u relevantnoj ogradi greške unaprijed, te nam stoga govori i o točnosti $\hat{x}$. Komponentni broj uvjetovanosti $cond(A, x)$ je relevantan samo ako imamo komponentnu relativnu grešku unatrag $\omega_{|A|,|b|}(\hat{x})$. Promatrajući s druge strane, svaki algoritam ima pridruženu analizu grešaka koja određuje broj uvjetovanosti relevantan za taj algoritam.

1.4 Skaliranje

U prethodnom poglavlju primjetili smo invarijantnosti od *cond(A)* obzirom na retčano skaliranje, što je u suprotnosti sa strogom ovisnošću $\kappa_\infty(A)$ o retčanom skaliranju. Prilika za skaliranje redova ili stupaca od A nastaje u raznim primjenama, pa ćemo sada bolje proučiti učinak skaliranja na broj uvjetovanosti po normi.

Za početak razmatramo jednostrano skaliranje, dajući generalizaciju dobro poznatog rezultata van der Sluis-a [4]. To pokazuje da za jednostrano skaliranje u Hölderovoj p-normi, normiranje redaka ili stupaca je gotovo optimalna strategija. Navodimo rezultat za pravokutne matrice A, za koje definiramo $\kappa_p(A) := \|A\|_p\|A^+\|_p$, gdje je A^+ pseudo inverz od A.

Teorem 1.4.1 (van der Sluis)**.** *Neka je $A \in \mathbb{R}^{m\times n}$,sa $\mathcal{D}_k \subset \mathbb{R}^{k\times k}$ označimo skup regularnih dijagonalnih matrica, te definiramo*

$$D_C := diag(\|A(:, j)\|_p)^{-1},\ D_R := diag(\|A(i, :)\|_p)^{-1}.$$

Tada je

$$\kappa_p(AD_C) \leq n^{1-1/p} \min_{D\in D_n} \kappa_p(AD) \quad ako\ rang(A) = n \tag{1.15}$$

$$\kappa_p(D_RA) \leq m^{1/p} \min_{D\in D_m} \kappa_p(DA) \quad ako\ rang(A) = m. \tag{1.16}$$

Dokaz. Za svaki $A \in \mathbb{R}^{m\times n}$ koristeći Lemu 0.0.11 imamo

$$\max_j \|A(:,j)\|_p \leq \|A\|_p \leq n^{1-1/p} \max_j \|A(:,j)\|_p \tag{1.17}$$

Dakle,

$$\|AD_C\|_p \leq n^{1-1/p} \tag{1.18}$$

Sada, za svaki $D \in \mathcal{D}_n$

$$\begin{aligned} \|D_C^{-1}A^+\|_p = \|D_C^{-1}D \cdot D^{-1}A^+\|_p &\leq \max_j(|d_{jj}|\ \|A(:,j)\|_p)\|D^{-1}A^+\|_p \\ &\leq \|AD\|_p\|D^{-1}A^+\|_p = \kappa_p(AD) \end{aligned} \tag{1.19}$$

koristeći Napomenu 0.0.9 i prvu nejednakost u (1.17). Pomnožimo li (1.18) i (1.19) dobivamo:

$$\begin{aligned} \|AD_C\|_p\|D_C^{-1}A^+\|_p &\leq n^{1-1/p}\kappa_p(AD) \\ \kappa_p(AD_C) &\leq n^{1-1/p}\kappa_p(AD). \end{aligned}$$

Minimizirajući posljedni izraz po D, dobivamo (1.15). Prema Napomeni 0.0.10 je $\kappa_p(DA) = \kappa_q(A^TD)$, gdje $p^{-1} + q^{-1} = 1$, slijedi nejednakost (1.16). □

Za $p = \infty$, (1.16) potvrđuje ono što već znamo iz (1.13) i (1.14) za kvadratne matrice: u ∞-normi, normiranje redaka je optimalna strategija retčanog skaliranja. Slično, za $p = 1$ normiranje stupaca je najbolje stupčano skaliranje po (1.15). Teorem 1.4.1 obično se navodi za 2-normu, za koju se pokaže da normiranje redaka i stupaca daje brojeve uvjetovanosti unutar faktora $\sqrt{m}$ i $\sqrt{n}$, odnosno, od minimuma brojeva uvjetovanosti po 2-normi postignutih za skaliranje po retcima i stupcima.

Idući korolar nam govori da među svim dvostranim dijagonalnim skaliranjima simetrične pozitivno definitne matrice, ono koja daje matrici A jediničnu dijagonalu nije daleko od optimalne.

Korolar 1.4.2 (van der Sluis)**.** *Neka je $A \in \mathbb{R}^{n\times n}$ simetrična pozitivno definitna i neka je $D_* = diag(a_{ii}^{-1/2})$. Tada*

$$\kappa_2(D_*AD_*) \leq n \min_{D\in\mathcal{D}_n} \kappa_2(DAD) \tag{1.20}$$

Dokaz. Neka je $A = R^TR$ faktorizacija Choleskog.

$$\begin{aligned} \kappa_2(DAD) = \|DAD\|_2\|(DAD)^{-1}\|_2 = \|D^TR^TRD\|_2\|D^{-1}R^{-1}R^{-T}D^{-T}\|_2 = \\ = \|RD\|_2^2\|D^{-1}R^{-1}\|_2^2 = \kappa_2(RD)^2. \end{aligned}$$

Primijenimo Teorem 1.4.1 na RD. □

Da li je skaliranje D_* u Korolaru 1.4.2 uopće optimalno? Forsythe i Straus [5] su pokazali da je optimalno ako je A simetrična pozitivno definitna sa svojstvom A (to znači da postoji permutacijska matrica P takva da se PAP^T može izraziti kao blok 2×2 matrica čiji su (1,1) i (2,2) blokovi dijagonalni). Prema tome, na primjer, bilo koja simetrična pozitivno definitna tridijagonalna matrica sa jediničnom dijagonalom je optimalno skalirana.

Primijetimo da koristeći $\max_j \|A(:,j)\|_p \leq \|A\|_p \leq \mu^{1-1/p}\max_j \|A(:,j)\|_p$ na mjestu (1.17), nejednakosti Teorema 1.4.1 i Korolara 1.4.2 mogu ojačati zamjenom m i n sa maksimalnim brojem ne-nula po stupcu i retku.

Slijedi neovisni rezultat za Frobeniusovu normu.

Teorem 1.4.3 (Stewart i Sun)**.** *Neka je $A = [a_1,\ldots,a_n] \in \mathbb{R}^{n\times n}$ regularna sa $B := A^{-1} = [b_1,\ldots,b_n]^T$, te neka je $D_C = diag((\|b_j\|_2/\|a_j\|_2)^{1/2})$. Tada*

$$\sum_j \|a_j\|_2\|b_j\|_2 = \kappa_F(AD_C) = \min_{D\in D_n}\kappa_F(AD).$$

Dokaz. Za $D = diag(d_i) \in \mathcal{D}_n$, te koristeći Cauchy-Schwartz-ovu nejednakost imamo

$$\kappa_F(AD) = \Big(\sum_j d_j^2\|a_j\|_2^2\Big)^{\frac{1}{2}}\Big(\sum_j d_j^{-2}\|b_j\|_2^2\Big)^{\frac{1}{2}} \geq \sum_j \|a_j\|_2\|b_j\|_2.$$

Ako je $d_j\|a_j\|_2 = \alpha d_j^{-1}\|b_j\|_2$ za sve j i neki $\alpha \neq 0$, tada imamo jednakost. Jednakost postoji za $d_j^2 = \|b_j\|_2/\|a_j\|_2$.
Pokažimo $\sum_j \|a_j\|_2\|b_j\|_2 = \kappa_F(AD_C)$.

$$AD_C(:,j) = \sqrt{\frac{\|b_j\|_2}{\|a_j\|_2}}a_j$$

$$(AD_C)^{-1}(j,:) = (D_C^{-1}A^{-1})(j,:) = (D_C^{-1}B)(j,:) = \sqrt{\frac{\|a_j\|_2}{\|b_j\|_2}}b_j$$

Sada je Frobeniusova norma:

$$\|AD_C\|_F^2 = \sum_{i=1}^n\sum_{j=1}^n |AD_C|_{ij}^2 = \sum_{j=1}^n\Big(\sum_{i=1}^n |AD_C|_{ij}^2\Big) = \sum_{j=1}^n \|\sqrt{\frac{\|b_j\|_2}{\|a_j\|_2}}\,a_j\|_2^2 = \sum_{j=1}^n \|a_j\|_2\|b_j\|_2$$

$$\|(AD_C)^{-1}\|_F^2 = \sum_{j=1}^n \|\sqrt{\frac{\|a_j\|_2}{\|b_j\|_2}}b_j\|_2^2 = \sum_{j=1}^n \|a_j\|_2\|b_j\|_2.$$

Dakle,

$$\kappa_F(AD_C) = \|AD_C\|_F\|(AD_C)^{-1}\|_F = \Big(\sum_{j=1}^{n}\|a_j\|_2\|b_j\|_2\Big)^{\frac{1}{2}}\Big(\sum_{j=1}^{n}\|a_j\|_2\|b_j\|_2\Big)^{\frac{1}{2}} = \sum_{j=1}^{n}\|a_j\|_2\|b_j\|_2.$$

□

Kao što smo vidjeli u ovom i prethodnom poglavlju, minimalna vrijednost od $\kappa_\infty(DA)$ je $\| |A^{-1}|\,|A|\,\|_\infty$. Idući rezultat pokazuje da se za dvostrano skaliranje matrica $|A^{-1}||A|$ ponovno pojavljuje u formuli za minimalni broj uvjetovanosti. Matrica je ireducibilna ako se ne može simetrično permutirati do blok trokutastog oblika. Perronov vektor od $B \geq 0$ je nenegativan svojstveni vektor koji odgovara svojstvenoj vrijednosti $\rho(B)$,gdje ρ označava spektralni radijus.

Teorem 1.4.4 (Bauer)**.** *Neka je $A \in \mathbb{R}^{n\times n}$ regularna i pretpostavimo da su $|A||A^{-1}|$ i $|A^{-1}||A|$ ireducibilne. Tada*

$$\min_{D_1,D_2\in D_n}\kappa_\infty(D_1AD_2) = \rho(|A|\,|A^{-1}|). \tag{1.21}$$

Minimum se postiže za $D_1 = diag(x)^{-1}$ i $D_2 = diag(|A^{-1}|x)$,gdje je $x > 0$ desni Perronov vektor od $|A||A^{-1}|$ (pa je $|A||A^{-1}|x = \rho(|A||A^{-1}|)x$).

Dokaz. Bauer [6]. □

Rump [7] je pokazao da (1.21) vrijedi za svaku regularnu matricu A ako se minimum lijeve strane zamijeni sa infimumom.

Za Kahanov primjer iz prošlog poglavlja imamo

$$\rho(|A^{-1}||A|) \approx 2.62 + 1.79\epsilon \ll 3 + (2\epsilon)^{-1} = \| |A^{-1}|\,|A|\,\|_\infty,$$

i u stvari je $\kappa_\infty(DAD) = 3$ za $D = diag(\epsilon^{1/2}, \epsilon^{-1/2}, \epsilon^{-1/2})$, pa je simetrično dvostrano skaliranje gotovo optimalno u ovom slučaju.

1.5 Numerička stabilnost

Greška unatrag, proučena u ovom poglavlju, vodi do definicije numeričke stabilnosti algoritama za rješavanje linearnih sustava. Precizna i točna definicija stabilnosti može se dati, ali postoji toliko mogućnosti kroz toliko različitih problema da bi imenovanje i definiranje svakoga skrenulo pozornost sa nama interesantnog problema. Stoga prihvaćamo neformalni pristup.

Numerička metoda za rješavanje kvadratnog, regularnog linearnog sustava $Ax = b$ je povratno stabilna po normi ako daje izračunato rješenje takvo da $\eta_{A,b}(\hat{x})$ je reda veličine

jedinične greške zaokruživanja. Koliko velik $\eta_{A,b}(\hat{x})/u$ dopuštamo da bude, dok istodobno proglašavamo da je metoda unatrag stabilna, ovisi o kontekstu. Uglavnom je u ovoj definiciji implicitno da je $\eta_{A,x}(\hat{x}) = O(u)$ za sve A i b, te se za metodu koja daje $\eta_{A,b}(\hat{x}) = O(u)$ za određene A i b kaže da je izvedena na povratno stabilan način po normi.

Značaj povratne stabilnosti po normi je ta da izračunato rješenje rješava malo perturbirane probleme, te ako podaci A i b sadrže nesigurnosti ograničene samo po normi, onda $\hat{x}$ može biti točno rješenje problema koje smo željeli riješiti.

Komponentna povratna stabilnost definira se na sličan način: zahtijevamo da komponentna greška unatrag $\omega_{|A|,|b|}(\hat{x})$ bude reda u. Ovo je stroži zahtjev od onog kod povratne stabilnosti po normi. Greške zaokruživanja nastale metodom koja je komponentno povratno stabilna jednake su po veličini i učinku greškama nastalim jednostavnim pretvaranjem podataka A i b u brojeve s pomičnom točkom prije nego započne postupak rješavanja.

Ako je metoda povratno stabilna po normi, onda je po Teoremu 1.2.2 greška unaprijed $\|x-\hat{x}\|/\|x\|$ ograničena višekratnikom broja $\kappa(A)u$. Međutim, metoda može dati rješenje čija je greška unaprijed ograničena na ovaj način bez da je greška unatrag po normi $\eta_{A,b}(\hat{x})$ reda u. Stoga je korisno definirati metodu za koju $\|x - \hat{x}\|/\|x\| = O(cond(A, x)u)$ kao stabilnu unaprijed po normi.

1.6 Praktične ograde grešaka

Pretpostavimo da je izračunata aproksimacija $\hat{x}$ rješenja linearnog sustava $Ax = b$, gdje $A \in \mathbb{R}^{n\times n}$. Koju ogradu greške trebamo računati? Greška unatrag može se točno izračunati iz formula

$$\eta_{E,f}(\hat{x}) = \frac{\|r\|}{\|E\|\,\|\hat{x}\| + \|f\|}$$

$$\omega_{E,f}(\hat{x}) = \max_i \frac{|r_i|}{(E|\hat{x}| + f)_i}\,, \tag{1.22}$$

po cijeni jednog ili dva produkta matrice i vektora, za $r = b - A\hat{x}$ i $E|\hat{x}|$. Jedino pitanje je što učiniti ako je nazivnik toliko mali da uzrokuje overflow ili dijeljenje sa nulom u izrazu za $\omega_{E,f}(\hat{x})$. Ovo se, na primjer, može dogoditi kada $E = |A|$ i $f = |b|$, te za neke i, $a_{ij}x_j = 0$ za sve j kao što je najvjerojatnije kod rijetko popunjenog problema. LAPACKova rutina *xxyRFS* (profinjeno rješenje) primjenjuje iterativno profinjenje u fiksnoj preciznosti u cilju zadovoljenja $\omega_{|A|,|b|} \leq u$. Ako je i-ta komponenta nazivnika u (1.22) manja od safe_{min}/u gdje je safe_{min} najmanji broj takav da $1/\text{safe}_{min}$ odlazi u overflow, tada se dodaje $(n + 1)\text{safe}_{min}$ i-toj komponenti brojnika i nazivnika. Sofisticiranija strategija je predložena za rijetko popunjene probleme od strane Ariolia, Demmela i Duffa [8]. Oni predlažu modificiranje formule (1.22) za $\omega_{|A|,|b|}$ na način da se zamijeni $|b_i|$ sa $\|A(i,:)\|_1\|\hat{x}\|_\infty$ kada je i-ti nazivnik jako mali.

Okrećući se prema pogrešci unaprijed, jedan pristup je procjena greške unaprijed iz Teorema 1.2.2 ili Teorema 1.3.2, sa ϵ jednakim odgovarajućoj grešci unatrag. Budući da je x u (1.9) nepoznat, trebali bismo koristiti modificiranu ogradu

$$\frac{\|x-\hat{x}\|_\infty}{\|\hat{x}\|_\infty} \leq \omega_{E,f}(\hat{x}) \frac{\| \, |A^{-1}| \, (E|\hat{x}| + f)\|_\infty}{\|\hat{x}\|_\infty} \tag{1.23}$$

Ako na raspolaganju imamo specifičan E i f za grešku unatrag, onda je prirodno iskoristiti ih u (1.23). Međutim, veličina ograde na grešku unatrag varira sa E i f, pa je prirodno ispitati koji izbor minimizira granicu.

Lema 1.6.1. *Gornja granica u (1.23) je najmanje velika kao gornja granica u*

$$\frac{\|x-\hat{x}\|_\infty}{\|\hat{x}\|_\infty} \leq \frac{\| \, |A^{-1}| \, \, |r| \, \|_\infty}{\|\hat{x}\|_\infty}, \tag{1.24}$$

a jednake su kada je $E|\hat{x}| + f$ višekratnik od $|r|$

Dokaz. Primijetimo prvo da $r = b - A\hat{x}$ povlači $|x - \hat{x}| \leq |A^{-1}||r|$, što pak povlači (1.24). Sada, za $z \geq 0$ imamo

$$|A^{-1}||r| = |A^{-1}|\Big[z_i \frac{|r_i|}{z_i}\Big]_{i=1,\ldots,n} \leq \max_i \frac{|r_i|}{|z_i|} |A^{-1}|z.$$

Prethodni izraz je jednakost ako je z višekratnik od r. Uzmemo li $z = E|\hat{x}| + f$ dobivamo

$$|A^{-1}| \, |r| \leq \omega_{E,f}(\hat{x}) \, |A^{-1}| \, (E|\hat{x}| + f).$$

Prethodni izraz je jednakost kada je $E|\hat{x}| + f$ višekratnik od $|r|$. Točnost za jednakost je očuvana kada uzmemo ∞-normu, pa slijedi rezultat leme. □

Obzirom da je ograda dobivena uzimajući apsolutne vrijednosti u jednadžbi $x - \hat{x} = A^{-1}r$, jasno je da je ona najmanja moguća takva granica podložna zanemarivanju predznaka u A^{-1} i r. Razumno se zapitati zašto za našu ogradu greške ne uzimamo $\|A^{-1}r\|_\infty / \|\hat{x}\|_\infty$. Jedan razlog je taj što ne možemo točno izračunati r ili $\|A^{-1}r\|$. Umjesto r računamo $\hat{r} = fl(b - A\hat{x})$ i

$$\hat{r} = r + \Delta r \, , |\Delta r| \leq \gamma_{n+1}(|A||\hat{x}| + |b|). \tag{1.25}$$

Stoga je stroga ograda

$$\frac{\|x-\hat{x}\|_\infty}{\|\hat{x}\|_\infty} \leq \frac{\| \, |A^{-1}| \, (|\hat{r}| + \gamma_{n+1}(|A||\hat{x}| + |b|))\|_\infty}{\|\hat{x}\|_\infty}. \tag{1.26}$$

Ta bi se ograda u praksi trebala koristiti umjesto (1.24). S obzirom na LU faktorizaciju od A, ova granica se može jeftino procijeniti bez računanja A^{-1}, i to je napravljeno pomoću LAPACKove xyyRFS rutine. Primijetimo također da ako izračunamo $A^{-1}r$ možemo primjeniti korak iterativnog profinjenja, što može pružiti stabilnije i točnije rješenje.

LAPACKovi rješavači linearnih jednadžbi procjenjuju samo jedan broj uvjetovanosti: standardni broj uvjetovanosti $\kappa_1(A)$ koji daje xxyCON rutina.

1.7 Teorija perturbacije po diferencijalnom računu

Svi perturbacijski rezultati u ovom radu su algebarski izvedeni, bez ijedne uporabe derivacija. Diferencijalni račun se također može koristiti za izvođenje perturbacijskih ograda, često na jednostavan način.

Promotrimo jedan jednostavan primjer. $A(t)x(t) = b(t)$, gdje su $A(t) \in \mathbb{R}^{n\times n}$ i $x(t), b(t) \in \mathbb{R}^n$ neprekidno diferencijabilne funkcije. Diferenciranje daje

$$\dot{A}(t)x(t) + A(t)\dot{x}(t) = \dot{b}(t),$$

ili ako ispustimo argument t

$$\dot{x} = -A^{-1}\dot{A}x + A^{-1}\dot{b}.$$

Uzimajući norme, dobivamo

$$\frac{||\dot{x}||}{||x||} \leq ||A^{-1}||\,||\dot{A}|| + ||A^{-1}||\frac{||\dot{b}||}{||x||} = \kappa(A)\Big(\frac{||\dot{A}||}{||A||} + \frac{||\dot{b}||}{||A||\,||x||}\Big) \tag{1.27}$$

Ova ograda pokazuje da je $\kappa(A)$ ključna veličina u mjerenju osjetljivosti linearnog sustava. Komponentna ograda se na isti način dobije.

Ogradu (1.27) možemo promijeniti u standardniji oblik perturbacijske granice. Odaberimo $A(t) = A + tE$, i $b(t) = b + tf$. Odavde je $\dot{A} = E$, $\dot{b} = f$. Derivirajmo $Ax = b$.

$$\dot{A}x + A\dot{x} = \dot{b}$$

Djelovanjem s A^{-1} imamo

$$\dot{x} = -A^{-1}\dot{A}x + A^{-1}\dot{b}$$

Ponovimo ovaj postupak za idućih nekoliko derivacija, gdje s $A^{(n)}$ i $x^{(n)}$ označimo stupanj derivacije.

$$\begin{aligned} A^{(2)}x + A^{(1)}x^{(1)} + A^{(1)}x^{(1)} + Ax^{(2)} &= b^{(2)} \\ 2A^{(1)}x^{(1)} + Ax^{(2)} = 0 &\Rightarrow x^{(2)} = -2A^{-1}A^{(1)}x^{(1)} \end{aligned}$$

$$\begin{aligned} 2A^{(2)}x^{(1)} + 2A^{(1)}x^{(2)} + A^{(1)}x^{(2)} + Ax^{(3)} &= 0 \\ 3A^{(1)}x^{(2)} + Ax^{(3)} = 0 &\Rightarrow x^{(3)} = -3A^{-1}A^{(1)}x^{(2)} \end{aligned}$$

$$\begin{aligned} 3A^{(2)}x^{(2)} + 3A^{(1)}x^{(3)} + A^{(1)}x^{(3)} + Ax^{(4)} &= 0 \\ 4A^{(1)}x^{(3)} + Ax^{(4)} = 0 &\Rightarrow x^{(4)} = -4A^{-1}A^{(1)}x^{(3)} \end{aligned}$$

$$\vdots$$

Sada imamo

$$\begin{aligned}
x^{(1)} &= -A^{-1}Ex + A^{-1}f \\
x^{(2)} &= -2A^{-1}E(-A^{-1}Ex + A^{-1}f) \\
x^{(3)} &= 6(A^{-1}E)^2(-A^{-1}Ex + A^{-1}f) \\
x^{(4)} &= -24(A^{-1}E)^3(-A^{-1}Ex + A^{-1}f) \\
&\vdots
\end{aligned}$$

Zapišemo li sada $x(\epsilon) = x(0) + \epsilon\dot{x}(0) + O(\epsilon^2)$ i uvrstimo derivacije x-a, imamo:

$$\begin{aligned}
x(\epsilon) - x(0) &= \epsilon x^{(1)}(0) + O((\epsilon)^2) + \dots \\
&= \epsilon\Big(- A^{-1}Ex(0) + A^{-1}f\Big) + \frac{\epsilon^2}{2}(-2A^{-1}E)\Big(- A^{-1}Ex(0) + A^{-1}f\Big) \\
&+ \frac{\epsilon^3}{6}6(A^{-1}E)^2\Big(- A^{-1}Ex(0) + A^{-1}f\Big) \\
&+ \frac{\epsilon^4}{24}(-24)(A^{-1}E)^3\Big(- A^{-1}Ex(0) + A^{-1}f\Big) + \dots \\
&= \epsilon\Big(- A^{-1}Ex(0) + A^{-1}f\Big) - \epsilon^2(A^{-1}E)\Big(- A^{-1}Ex(0) + A^{-1}f\Big) \\
&+ \epsilon^3(A^{-1}E)^2\Big(- A^{-1}Ex(0) + A^{-1}f\Big) \\
&- \epsilon^4(A^{-1}E)^3\Big(- A^{-1}Ex(0) + A^{-1}f\Big) + \dots \\
&= \epsilon\Big(1 - \epsilon A^{-1}E + \epsilon^2(A^{-1}E)^2 - \epsilon^3(A^{-1}E)^3 + \dots\Big)\Big(- A^{-1}Ex(0) + A^{-1}f\Big) \\
&= \epsilon(1 + \epsilon A^{-1}E)^{-1}\Big(- A^{-1}Ex(0) + A^{-1}f\Big)
\end{aligned}$$

Djelovanjem s normom i koristeći njena svojstva, te množeći s $\frac{1}{||x(0)||}$ dobivamo:

$$\frac{||x(\epsilon) - x(0)||}{||x(0)||} \leq \frac{\epsilon}{1 - \epsilon||A^{-1}||\,||E||}\Big(||A^{-1}||\,||E|| + \frac{||A^{-1}||\,||f||}{||x(0)||}\Big),$$

a to je upravo prvobitni oblik ograde (1.4).

Tehnika diferencijalnog računanja je koristan dodatak arsenalu analitičara pogrešaka, ali algebarski pristup je poželjniji za izvođenje rigoroznih perturbacijskih ograda klasičnog oblika.

Poglavlje 2

Trokutasti sustavi

2.1 Opis

Trokutasti sustavi imaju temeljnu ulogu u računanju s matricama. Mnogo metoda građeno je na ideji reduciranja problema na rješavanje jednog ili više trokutastih sustava, uključujući gotovo sve direktne metode za rješavanje linearnih sustava. Na serijskim računalima trokutasti sustavi univerzalno su riješeni standardnim algoritmima za supstituciju unatrag i unaprijed. Za paralelno računanje postoji nekoliko alternativnih metoda.

Analiza greške unatrag za algoritme supstitucije je jednostavna i zaključak je dobro poznat: algoritam je vrlo stabilan. Ponašanje greške unaprijed je s druge strane intrigantno, zbog toga što je greška unaprijed često iznenađujuće mala - daleko manja nego što bi predvidjeli iz broja uvjetovanosti po normi κ, ili čak iz komponentnog broja uvjetovanosti *cond*.

Iduća dva citata naglašavaju veliku preciznost koja se često može opaziti u praksi.
"The solutions of triangular systems are usually computed at high accuracy. This fact... cannot be proved in general, for counter examples exist. However, it is true of many special kinds of triangular matrices and the phenomenon has been observed in many others. The practical consequences of this fact cannot be over-emphasized." - G.W.Stewart
"In practise one almost invariably finds that if L is ill-conditioned, so that $\|L\| \, \|L^{-1}\| \gg 1$, then the computed solution of $Lx = b$ (or the computed inverse) is far more accurate than [standard norm bounds] would suggest." - J.H.Wilkinson
Analiza koju ćemo pokazati u ovom poglavlju daje djelomično objašnjenje za promatranu točnost algoritma supstitucije. Posebno, otkriva tri važna ali ne očita svojstva:

- točnost izračunatog rješenja iz supstitucije ovisi jako o desnoj strani
- trokutasta matrica može biti puno više ili manje loše-uvjetovana nego kad se transponira

- uporaba pivotiranja u LU, QR i faktorizaciji Choleskog može uvelike poboljšati uvjetovanost rezultirajućeg trokutastog sustava

Izvodimo ograde greške unaprijed i unatrag.

2.2 Analiza greške unatrag

Sjetimo se da se za gornje trokutastu matricu $U \in \mathbb{R}^{n\times n}$ sustav $Ux = b$ može riješiti koristeći formulu $x_i = (b_i - \sum_{j=i+1}^{n}(u_{ij}x_j)/u_{ii}$ što daje komponente od x od zadnje prema prvoj.

Algoritam 2.2.1 (povratna supstitucija)**.** *Za regularnu gornje trokutastu matricu $U \in \mathbb{R}^{n\times n}$ ovaj algoritam rješava sustav $Ux = b$.*

$x_n = b_n/u_{nn}$
for $i = n-1 : -1 : 1$ ***do***
  $s = b_i$
  for $j = i+1 : n$ ***do***
    $s = s - u_{ij}x_j$
  end for
  $x_i = s/u_{ii}$
end for
Cijena: n^2 aritmetičkih operacija pomične točke

Nećemo navoditi analogan algoritam za rješavanje donje trokutastog sustava sa supstitucijom unaprijed. Idući rezultati za supstituciju unatrag imaju očitu analogiju za supstituciju unaprijed. U ovom poglavlju sa T ćemo označavati matricu koja može biti gornje ili donje trokutasta. Za analizu greške u supstituciji treba nam iduća lema.

Lema 2.2.2. *Neka je $y = (c - \sum_{i=1}^{k-1} a_i b_i)/b_k$ izračunat u aritmetici s pomičnom točkom prema*

$s = c$
for $i = 1 : k-1$ ***do***
  $s = s - a_i b_i$
end for
$y = s/b_k$

Tada izračunati $\hat{y}$ zadovoljava

$$b_k\hat{y}(1+\theta_k) = c - \sum_{i=1}^{k-1} a_i b_i (1+\theta_i), \tag{2.1}$$

gdje $|\theta_i| \le \gamma_i = iu/(1-iu)$.

Dokaz. Slično kao u analizi koju je napravio N. J. Higham [10, p. 62] može se vidjeti da $\hat{x} := fl(c - \sum_{i=1}^{k-1} a_i b_i)$ zadovoljava

$$\hat{s} = c(1+\delta_i)\cdots(1+\delta_{k-1}) - \sum_{i=1}^{k-1} a_i b_i (1+\epsilon_i)(1+\delta_i)\cdots(1+\delta_{k-1}),$$

gdje su $|\epsilon_i|, |\delta_i| \leq u$. Koristeći

$$fl(x \text{ op } y) = \frac{x \text{ op } y}{1+\delta}\ , \quad |\delta| \leq u\ , \quad \text{op} : +\ , -\ , *\ , /$$

konačno dijeljenje daje $\hat{y} = fl(\hat{s}/b_k) = \hat{s}/(b_k(1+\delta_k))\ , |\delta_k| \leq u$, pa nakon dijeljenja sa $(1+\delta_1)\ldots(1+\delta_{k-1})$ imamo

$$b_k \hat{y} \frac{1+\delta_k}{(1+\delta_1)\ldots(1+\delta_{k-1})} = c - \sum_{i=1}^{k-1} a_i b_i \frac{1+\epsilon_i}{(1+\delta_1)\ldots(1+\delta_{i-1})}. \tag{2.2}$$

Tvrdnju dobivamo koristeći rezultat: Ako je $|\delta_i| \leq u$ i $\rho_i = \pm 1$ za $i = 1:n$, i $nu < 1$, tada

$$\prod_{i=1}^{n} (1+\delta_i)^{\rho_i} = 1 + \theta_n$$

gdje je

$$|\theta_n| \leq \frac{nu}{1-nu} =: \gamma_n.$$

□

Primijetimo da odabiremo određen oblik (2.1), u kojem c nije perturbiran, kako bi dobili rezultate greške unatrag za $Ux = b$ u kojem b nije perturbiran.

Teorem 2.2.3. *Izračunato rješenje $\hat{x}$ iz Algoritma 2.2.1 zadovoljava*

$$(U + \Delta U)\hat{x} = b\ , \quad |\Delta u_{ij}| \leq \begin{cases} \gamma_{n-i+1}|u_{ii}|, & i = j \\ \gamma_{|i-j|}|u_{ij}|, & i \neq j \end{cases}$$

Teorem 2.2.3 vrijedi samo za određen poredak aritmetičkih operacija korištenih u Algoritmu 2.2.1. Rezultat koji vrijedi za svaki poredak je posljedica iduće leme.

Lema 2.2.4. *Ako je $y = (c - \sum_{i=1}^{k-1} a_i b_i)/b_k$ izračunat u aritmetici s pomičnom točkom, tada, bez obzira na redosljed sumacije,*

$$b_k \hat{y}(1+\theta_k^{(0)}) = c - \sum_{i=1}^{k-1} a_i b_i (1+\theta_k^{(i)}),$$

gdje je $|\theta_k^{(i)}| \leq \gamma_k$ za sve i. Ako je $b_k = 1$, dakle ako nema dijeljenja, onda je $|\theta_k^{(i)}| \leq \gamma_{k-1}$ za sve i.

Dokaz. Nicholas J. Higham [10]. □

Teorem 2.2.5. *Neka je trokutasti sustav $Tx = b$, gdje je $T \in \mathbb{R}^{n\times n}$ regularna, riješen za bilo koji poredak. Tada izračunato rješenje $\hat{x}$ zadovoljava*

$$(T + \Delta T)\hat{x} = b \ , \ |\Delta T| \leq \gamma_n |T|.$$

U tehničkim terminima, ovaj rezultat nam govori da $\hat{x}$ ima malu relativnu komponentnu grešku unatrag. Drugim riječima, greška unatrag je najmanja koju smo mogli očekivati.

U većini preostalih analiza grešaka izvodit ćemo rezultate koji kao u Teoremu 2.2.5 ne ovise o poretku aritmetičkih operacija. Rezultati ovog tipa su općenitiji, obično ne sadrže manje informacija i lakši su za izvođenje, od onih koji ovise o poretku. Međutim, važno je shvatiti da *stvarna greška* ovisi o poretku, moguće i jako za određene podatke.

2.3 Analiza greške unaprijed

Iz Teorema 2.2.5 i Teorema 1.3.2 slijedi ograda greške unaprijed

$$\frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} \leq \frac{cond(T, x)\gamma_n}{1 - cond(T)\gamma_n}, \tag{2.3}$$

gdje je

$$cond(T, x) = \frac{\| \, |T^{-1}| \, |T| \, |x| \, \|_\infty}{\|x\|_\infty} \ , \quad cond(T) = \| \, |T^{-1}| \, |T| \, \|_\infty$$

Ova ograda može biti proizvoljno manja od odgovarajuće ograde koja uključuje $\kappa_\infty(T) = \|T\|_\infty \|T^{-1}\|_\infty$, iz razloga objašnjenih u prethodnom poglavlju. Ubuduće zapamtimo da, u terminima klasičnog broja uvjetovanosti $\kappa(T)$, loša uvjetovanost trokutaste matrice proizlazi iz dva moguća izvora: varijacija veličine dijagonalnih elemenata, i retci sa vandijagonalnim elementima koji su veliki u odnosu na dijagonalne elemente. Zanimljivo je da zbog invarijacije na retčano skaliranje $cond(T, x)$ je osjetljiv samo na drugi izvor.

Unatoč zadovoljavajućim svojstvima, $cond(T, x)$ može biti proizvoljno velik. To možemo vidjeti pomoću gornje trokutaste matrice

$$U(\alpha) = (u_{ij}) \ , \ u_{ij} = \begin{cases} 1, & i = j \\ -\alpha, & i < j \end{cases} \tag{2.4}$$

za koju je

$$(U(\alpha)^{-1})_{ij} = \begin{cases} 1, & i = j \\ \alpha(1 + \alpha)^{j-i-1}, & j > i \end{cases} \tag{2.5}$$

Sada imamo $cond(U(\alpha), \mathrm{e}) = cond(U(\alpha)) \sim 2\alpha^{n-1}$ kada $\alpha \to \infty$. Stoga ne možemo tvrditi da su svi trokutasti sustavi riješeni do na visoku preciznost. Ipak, za bilo koju T postoji

bar jedan sustav za koji dobivamo visoku preciznost: $Tx = e_1$ ako je T gornje trokutasta, ili $Tx = e_n$ ako je T donje trokutasta. U oba slučaja je $cond(T, x) = 1$ i rješavanje se svodi na računanje jednog skalarnog recipročnog broja.

Kako bi stekli daljnji uvid razmotrimo posebnu klasu trokutastih matrica. Započnimo s matricom dobivenom određenom standardnom faktorizacijom s pivotiranjem. U rezultatima koje ćemo sada navesti pretpostavili smo da su trokutaste matrice veličine $n \times n$ i regularne, te je $\hat{x}$ izračunato rješenje supstitucije.

Lema 2.3.1. *Pretpostavimo da gornje trokutasta matrica $U \in \mathbb{R}^{n\times n}$ zadovoljava*

$$|u_{ii}| > |u_{ij}| \text{ za sve } j > i \tag{2.6}$$

Tada jedinična gornje trokutasta matrica $W = |U^{-1}|\,|U|$ zadovoljava $w_{ij} \leq 2^{j-i}$ za sve $j > i$ i stoga $cond(U) \leq 2^n - 1$.

Dokaz. Možemo pisati $W = |V^{-1}||V|$ gdje je $V = D^{-1}U$ i $D = diag(u_{ii})$. Matrica V je jedinična gornje trokutasta sa $|v_{ij}| \leq 1$. Pokažimo indukcijom da je $|(V^{-1})_{ij}| \leq 2^{j-i-1}$ za $j > i$. Promatrajmo $V \cdot V^{-1} = I$ i označimo $Z := V^{-1}$. Za elemente od Z imamo

$$\begin{aligned} &|z_{12}|, |z_{23}|, \ldots, |z_{n-1,n}| \leq 1 \\ &|z_{13}|, |z_{24}|, \ldots, |z_{n-2,n}| \leq 2 \\ &\qquad\qquad\vdots \end{aligned}$$

Tvrdnja indukcije: za $k = 1, \ldots, n-1$ vrijedi $|z_{1,k+1}|, \ldots, |z_{n-k,n}| \leq 2^{k-1}$.
Baza: $k = 1$. U $V \cdot V^{-1} = I$ promatramo presjek i-tog retka i *(i+1)*-og stupca:

$$1 \cdot v_{i,i+1} + z_{i,i+1} \cdot 1 = 0.$$

Dakle, $|z_{i,i+1}| = |v_{i,i+1}| \leq 1 \leq 2^{k-1}$.
Korak: Pretpostavimo da tvrdnja vrijedi za neki k. Dokažimo da vrijedi za *k+1*.
U $V \cdot V^{-1} = I$ promatramo presjek i-tog retka i *(i+k+1)*-og stupca:

$$1 \cdot z_{i,i+k+1} + v_{i,i+1} \cdot z_{i+1,i+k+1} + \cdots + v_{i,i+k} \cdot z_{i+k,i+k+1} + v_{i,i+k+1} \cdot 1 = 0$$

Sada koristeći pretpostavku indukcije i $|v_{ij}| \leq 1$ imamo

$$\begin{aligned} |z_{i,i+k+1}| &\leq |v_{i,i+1}|\,|z_{i+1,i+k+1}| + \cdots + |v_{i,i+k}|\,|z_{i+k,i+k+1}| + |v_{i,i+k+1}| \\ &\leq 2^{k-1} + 2^{k-2} + \cdots + 2^1 + 2^0 + 1 \\ &= \frac{1 - 2^k}{1 - 2} + 1 \\ &= -1 + 2^k + 1 \\ &= 2^k = 2^{(k+1)-1}, \end{aligned}$$

te smo gotovi s indukcijom. Obzirom da tvrdnja indukcije vrijedi za svaki k pokazali smo da je $|(V^{-1})_{ij}| \leq 2^{j-i-1}$.
Koristeći $|(V^{-1})_{ij}| \leq 2^{j-i-1}$, za $j > i$, je

$$w_{ij} = \sum_{k=i}^{j} |(V^{-1})_{ik}||v_{kj}| \leq 1 + \sum_{k=i+1}^{j} 2^{k-i-1} \cdot 1 = 2^{j-i}.$$

Prema tome

$$\sum_{j=i}^{n} |w_{ij}| \leq \sum_{j=i}^{n} 2^{j-i} = 2^{n-i+1} - 1.$$

Dakle imamo

$$cond(U) = \| \, |U^{-1}| \, |U| \, \|_\infty = \|W\|_\infty \leq 2^n - 1.$$

□

Teorem 2.3.2. *Pretpostavimo isto kao i u prethodnoj Lemi. Tada izračunato rješenje $\hat{x}$ od $Ux = b$ dobiveno supstitucijom zadovoljava*

$$|x_i - \hat{x}_i| \leq 2^{n-i+1}\gamma_n \max_{j \geq i} |\hat{x}_j|, \quad i = 1 : n.$$

Dokaz. Iz Teorema 2.2.5 imamo

$$|x - \hat{x}| = |U^{-1}\Delta U\hat{x}| \leq \gamma_n |U^{-1}| \, |U| \, |\hat{x}|.$$

Koristeći Lemu 2.3.1 dobivamo

$$|x_i - \hat{x}_i| \leq \gamma_n \sum_{j=i}^{n} w_{ij}|\hat{x}_j| \leq \gamma_n \max_{j \geq i} |\hat{x}_j| \sum_{j=i}^{n} 2^{j-i} \leq 2^{n-i+1}\gamma_n \max_{j \geq i} |\hat{x}_j|.$$

□

Lema 2.3.1 nam pokazuje da za U koji zadovoljava (2.6), cond(U) je ograđen za fiksan n, bez obzira koliko velik je $\kappa(U)$. Ograde za $|x_i - \hat{x}_i|$ u Teoremu 2.3.2, iako velike ako je n velik i i mali, opadaju eksponencijalno s povećanjem i-a. Prema tome, *ranije* izračunate komponente od x se uvijek računaju do visoke točnosti u odnosu na kasnije izračunate elemente.

Analogoni Leme 2.3.1 i Teorema 2.3.2 vrijede za donje trokutastu matricu L koja zadovoljava

$$|l_{ii}| \geq |l_{ij}| \quad \text{za sve } j < i. \tag{2.7}$$

Međutim, primijetimo da ako gornje trokutasta matrica T zadovoljava (2.6) tada T^T nužno ne zadovoljava (2.7). Zapravo, $cond(T^T)$ može biti proizvoljno velik kao što možemo vidjeti primjerom

$$T = \begin{bmatrix} 1 & 1 & 0 \\ 0 & \epsilon & \epsilon \\ 0 & 0 & 1 \end{bmatrix},$$
$$cond(T) = 5, \quad cond(T^T) = 1 + \frac{2}{\epsilon}.$$

Važan zaključak je da trokutasti sustav $Tx = b$ može biti puno više ili manje loše uvjetovan od sustava $T^T y = c$, čak i ako T zadovoljava (2.6).

Lemu 2.3.1 i Teorem 2.3.2 (ili njihove analogone za donje trokutaste matrice) možemo primijeniti na:

- donje trokutaste matrice iz Gaussove eliminacije s parcijalnim pivotiranjem, *rook* pivotiranjem i potpunim pivotiranjem;
- gornje trokutaste matrice iz Gaussove eliminacije s *rook* pivotiranjem i potpunim pivotiranjem;
- gornje trokutaste matrice iz *Cholesky* i *QR* faktorizacije s potpunim pivotiranjem i stupčanim pivotiranjem.

Još posebnija klasa gornje trokutastih matrica od onih koje zadovoljavaju (2.6) je klasa retčano dijagonalno dominantnih matrica: $U \in \mathbb{R}^{n\times n}$ zadovoljava

$$|u_{ii}| \leq \sum_{j=i+1}^{n} |u_{ij}|, \quad i = 1 : n-1.$$

Takve matrice nastaju kao gornje trokutaste matrice u LU faktorizaciji bez pivotiranja od retčano dijagonalno dominantnih matrica i za njih vrijedi jači rezultat od Leme 2.3.1.

Lema 2.3.3. *Ako je gornje trokutasta matrica $U \in \mathbb{R}^{n\times n}$ retčano dijagonalno dominantna, onda cond$(U) \leq 2n - 1$.*

Dokaz. Dokaz slijedi dokaz Leme 2.3.1, uz D i V definirane kao i tamo. Koristeći činjenicu da je U, pa onda i V, retčano dijagonalno dominantna dolazimo do toga da V ima elemente ograničene veličinom 1. Koristeći prethodne činjenice i da je $V \cdot V^{-1} = I$ slijedi da i V^{-1} ima elemente ograničene veličinom 1. Kako je

$$\Big(|V|e\Big)_i = \sum_{j=i}^{n} |v_{ij}| \cdot 1 = 1 + \sum_{j=i+1}^{n} |v_{ij}| \leq 2 \quad \text{, za} \quad i \leq n-1$$
$$\Big(|V|e\Big)_n = v_{nn} = 1,$$

dobivamo

$$\|W\|_\infty = \| \, |W| \, e\|_\infty = \| \, |V^{-1}| \, |V| \, e\|_\infty = \| \, |V^{-1}| \, [2 \; 2 \ldots 2 \; 1]^T\|_\infty = 2n - 1.$$

□

Iz Leme 2.3.3 i (2.2) slijedi da su retčano dijagonalno dominantni gornje trokutasti sustavi riješeni do u suštini savršene relativne točnosti po komponentama.

Poglavlje 3

LU faktorizacija i linearne jednadžbe

3.1 Gaussove eliminacije i strategija pivotiranja

Započnimo dajući standardni opis Gaussovih eliminacija za rješavanje linearnog sustava $Ax = b$, gdje je $A \in \mathbb{R}^{n\times n}$ regularna.

Strategija Gaussovih eliminacija je reduciranje problema kojeg ne možemo riješiti (puni linearni sustav) na onaj koji možemo riješiti (trokutasti sustav) koristeći osnovne retčane operacije. Počevši od $A^{(1)} := A$, $b^{(1)} := b$ i završavajući s gornje trokutastim sustavom $A^{(n)}x = b^{(n)}$, imamo $n - 1$ koraka.

Na početku k-tog koraka originalni sustav je već transformiran u $A^{(k)}x = b^{(k)}$, gdje je

$$A^{(k)} = \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ 0 & A_{22}^{(k)} \end{bmatrix}$$

uz $A_{11}^{(k)} \in \mathbb{R}^{(k-1)\times(k-1)}$ gornje trokutastu. Svrha k-tog koraka eliminacije je dobivanje nule na mjestima ispod dijagonale u k-tom stupcu od $A^{(k)}$. To se postiže operacijama:

$$\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)} , \quad i = k+1 : n,\ j = k+1 : n \\ b_{i}^{(k+1)} &= b_{i}^{(k)} - m_{ik}b_{k}^{(k)} , \quad i = k+1 : n \end{aligned}$$

gdje su $m_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$, $i = k + 1 : n$. Na kraju $(n - 1)$-og koraka imamo gornje trokutasti sustav $A^{(n)}x = b^{(n)}$ koji se rješava povratnom supstitucijom. Za $n \times n$ matricu, cijena Gaussove eliminacije je $2n^3/3$ operacija.

U Gaussovoj metodi eliminacija susrećemo se s dva problema. Prvo, dolazi do sloma metode u slučaju dijeljenja sa nulom ako je $a_{kk}^{(k)} = 0$. Drugo, ako rješavamo sustav u konačnoj preciznosti i neki od multiplikatora m_{ik} je velik, tada postoji mogućnost gubitka značaja: pri oduzimanju $a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)}$ znamenke niskog reda $a_{ij}^{(k)}$ mogu se izgubiti. Gubljenje

tih znamenki može odgovarati relativno velikoj promjeni originalne matrice A. Najjednostavniji primjer ove pojave vidi se na matrici $\begin{bmatrix} \epsilon & 1 \\ 1 & 1 \end{bmatrix}$. Sada je $a_{22}^{(2)} = 1 - 1/\epsilon$ i $fl(a_{22}^{(2)}) = -1/\epsilon$ ako $\epsilon < u$ što bi bio točan odgovor ako promijenimo a_{22} u 0.

Prethodna razmatranja motiviraju strategiju *parcijalnog pivotiranja*. Na početku k-tog koraka k-ti i r-ti redovi su zamijenjeni, gdje

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

Parcijalno pivotiranje osigurava da su multiplikatori *lijepo* ograđeni:

$$|m_{ik}| \le 1 \ , \quad i = k + 1 : n.$$

Skuplju strategiju pivotiranja koja izmjenjuje i retke i stupce nazivamo *potpunom*. Na početku k-tog koraka redovi k i r i stupci k i s su zamijenjeni, gdje

$$|a_{rs}^{(k)}| := \max_{k \le i,j \le n} |a_{ij}^{(k)}|.$$

Ovo zahtijeva ukupno $O(n^3)$ usporedbi, dok parcijalno pivotiranje zahtijeva $O(n^2)$ usporedbi. Obzirom da parcijalno pivotiranje daje dobre rezultate, potpuno pivotiranje koristi se samo u posebnim situacijama.

Manje poznata, ali također zanimljiva strategija pivotiranja je *rook* pivotiranje. Ova metoda u svakom koraku odabire pivotni element srednje veličine između pivotnih elemenata koji bi bili izabrani parcijalnim i potpunim pivotiranjem. Na početku k-tog koraka, redovi k i r i stupci k i s su zamijenjeni, gdje

$$|a_{rs}^{(k)}| = \max_{k \le i \le n} |a_{is}^{(k)}| = \max_{k \le j \le n} |a_{rj}^{(k)}|.$$

Drugim riječima, odabran je pivot koji je najveći u svom stupcu (kao kod parcijalnog pivotiranja) i svom retku. Algoritam za traženje pivotnog elementa glasi:

$s_0 = k$
for $p = 1, 2, \ldots$ **do**
  r_p = retčani indeks prvog elementa najvećeg modula među $\{a_{i,s_{p-1}}\}_{i=k}^{n}$
  if $p > 1$ and $|a_{r_p,s_{p-1}}| = |a_{r_{p-1},s_{p-1}}|$ **then**
    uzmi $a_{r_{p-1},s_{p-1}}$ kao pivota
  end if
  s_p = stupčani indeks prvog elementa najvećeg modula među $\{a_{r_p,j}\}_{j=k}^{n}$
  if $|a_{r_p,s_p}| = |a_{r_p,s_{p-1}}|$ **then**
    uzmi $a_{r_p,s_{p-1}}$ kao pivota
  end if
end for

Iz činjenice da traženje pivotnog elementa odgovara pokretima kule u šahu proizlazi naziv *rook* pivotiranje. Primijetimo da se kod traženje pivota, prethodno razmotreni elementi mogu preskočiti. U daljnjem razmatranju uključit ćemo takvo profinjenje, iako u praksi možda nije potrebno.

Očito traženje pivotnog elementa u *rook* pivotiranju uključuje najmanje dva pute više usporedbi nego kod parcijalnog pivotiranja, te ako je potrebno pretražiti cijelu podmatricu broj usporedbi je isti kao kod potpunog pivotiranja. Foster [9] je pokazao da ako su elementi $\{a_{ij}^{(k)}\}_{i,j=k}^{n}$ nezavisne jednako distribuirane slučajne varijable iz bilo koje neprekidne vjerojatnosne distribucije, da je tada očekivani broj usporedbi u traženju pivotnog elementa za *rook* pivotiranje u koraku k najviše $(n-k)e$ (gdje je $e = \exp(1)$). Ako je ta statistička pretpostavka zadovoljena za svaki k tada je ukupan broj usporedbi ograničen s $(n-1)ne/2$, što je istog reda kao i za parcijalno pivotiranje ($(n-1)n/2$ usporedbi). Numerički eksperimenti pokazuju da je cijena *rook* pivotiranja zaista mali višekratnik cijene parcijalnog pivotiranja i znatno manja od cijene potpunog pivotiranja. Međutim, *rook* pivotiranje može iziskivati $O(n^3)$ usporedbi kao što je ilustrirano bilo kojom matricom oblika

$$\begin{bmatrix} \theta_1 & \theta_2 & & \\ & \theta_3 & \theta_4 & \\ & & \theta_5 & \theta_6 \\ & & & \theta_7 \end{bmatrix}, |\theta_1| < |\theta_2| < \cdot < |\theta_7|,$$

koja zahtijeva $n^3/4 + O(n^2)$ usporedbi.

3.2 LU faktorizacija

Bolji uvid u Guessove eliminacije dobiva se prikazom u matričnoj notaciji. Možemo zapisati:

$$A^{(k+1)} = M_k A^{(k)} := \begin{bmatrix} I_{k-1} & & & & \\ & 1 & & & \\ & -m_{k+1,k} & 1 & & \\ & -m_{k+2,k} & & \ddots & \\ & \vdots & & & \ddots \\ & -m_{n,k} & & & 1 \end{bmatrix} A^{(k)}$$

Matrica M_k se može kompaktno izraziti kao $M_k = I - m_k e_k^T$ gdje je e_k k-ti jedinični vektor i $e_i^T m_k = 0$ za $i \leq k$. Da bi dobili inverz od M_k, promijenimo predznake multiplikatora: $M_k^{-1} = I + m_k e_k^T$. Sveukupno,

$$M_{n-1} M_{n-2} \dots M_1 A = A^{(n)} =: U$$

pa je

$$\begin{aligned} A &= M_1^{-1} M_2^{-1} \dots M_{n-1}^{-1} U \\ &= (I + m_1 e_1^T)(I + m_2 e_2^T) \dots (I + m_{n-1} e_{n-1}^T) U \\ &= (I + \sum_{i=1}^{n-1} m_i e_i^T) U \\ &= \begin{bmatrix} 1 & & & & \\ m_{21} & 1 & & & \\ \vdots & m_{32} & \ddots & & \\ \vdots & \vdots & & \ddots & \\ m_{n1} & m_{n2} & \cdots & m_{n,n-1} & 1 \end{bmatrix} U =: LU. \end{aligned}$$

Zaključujemo da Gaussove eliminacije računaju LU faktorizaciju od A: $A = LU$ gdje je L jedinična donje trokutasta matrica, a U je gornje trokutasta.

Radi jednostavnosti zapisa označimo: $A_k := A(1:k, 1:k)$.

Teorem 3.2.1. *Postoji jedinstvena LU faktorizacija od $A \in \mathbb{R}^{n \times n}$ ako i samo ako je A_k regularna za $k = 1 : n-1$. Ako je A_k singularna za neke $1 \leq k \leq n-1$ onda faktorizacija može postojati, ali nije jedinstvena.*

Dokaz. Pretpostavimo da je A_k regularna za $k = 1 : n-1$. Postojanje LU faktorizacije može se dokazati ispitivanjem koraka Gaussovih eliminacija, ali puno elegantniji dokaz koji također daje jedinstvenost može se dobiti induktivnom graničnom konstrukcijom. Pretpostavimo da A_{k-1} ima jedinstvenu LU faktorizaciju $A_{k-1} = L_{k-1} U_{k-1}$. Tražimo faktorizaciju

$$A_k = \begin{bmatrix} A_{k-1} & b \\ c^T & a_{kk} \end{bmatrix} = \begin{bmatrix} L_{k-1} & 0 \\ l^T & 1 \end{bmatrix} \begin{bmatrix} U_{k-1} & u \\ 0 & u_{kk} \end{bmatrix} =: L_k U_k.$$

Jednadžbe koje moraju biti zadovoljene su $L_{k-1}u = b$, $U_{k-1}^T l = c$ i $a_{kk} = l^T u + u_{kk}$. Matrice L_{k-1} i U_{k-1} su regularne, obzirom da je $0 \neq det(A_{k-1}) = det(L_{k-1})det(U_{k-1})$, pa jednadžbe imaju jedinstveno rješenje, čime završavamo indukciju.

Pokažimo suprotno, pod pretpostavkom da je A regularna. Pretpostavimo da LU faktorizacija postoji. Tada je $A_k = L_k U_k$ za $k = 1 : n$, što daje

$$det(A_k) = det(U_k) = u_{11} \dots u_{kk}. \tag{3.1}$$

Za $k = n$ imamo $0 \neq det(A) = u_{11} \dots u_{nn}$ i stoga je $det(A_k) = u_{11} \dots u_{kk} \neq 0$, $k = 1 : n - 1$. Primjer koji ilustrira da LU faktorizacija može postojati u slučaju singularne matrice A, ali da nije jedinstvena, je faktorizaciju nul-matrice

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

S druge strane, matrica

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

nema LU faktorizaciju, iako je regularna.

□

Izraz $u_{kk} = det(A_k)/det(A_{k-1})$ slijedi iz (3.1). U stvari, svi elementi od L i U mogu se izraziti pomoću determinantne formule:

$$l_{ij} = \frac{det(A([1 : j - 1, i], 1 : j))}{det(A_j)} \text{ , } i \geq j \tag{3.2a}$$

$$u_{ij} = \frac{det(A(1 : i, [1 : i - 1, j]))}{det(A_{i-1})} \text{ , } i \leq j \tag{3.2b}$$

Promatramo li slučaj $n = 4$ lako se vidi učinak parcijalnog pivotiranja. Imamo

$$\begin{aligned} U &= M_3 P_3 M_2 P_2 M_1 P_1 A \text{ , gdje } P_k \text{ zamjenjuje redove k, r } (r \geq k) \text{ ,} \\ &= M_3 \cdot P_3 M_2 P_3 \cdot P_3 P_2 M_1 P_2 P_3 \cdot P_3 P_2 P_1 A \\ &=: M_3' M_2' M_1' PA \end{aligned}$$

gdje je na primjer, $M_1' = P_3 P_2 (I - m_1 e_1^T) P_2 P_3 = I - (P_3 P_2 m_1) e_1^T$. Za $k = 1, 2, 3$, M_k' je isti kao i M_k osim što su multiplikatori izmijenjeni. Stoga, za $n = 4$ Gaussove eliminacije s parcijalnim pivotiranjem (GEPP) primijenjene na A ekvivalentne su Gaussovim eliminacijama bez pivotiranja primijenjenim na retčano permutiranoj matrici PA. Ovaj zaključak vrijedi za svaki n: GEPP računa faktorizaciju $PA = LU$. Slično, Gaussove eliminacije s

rook pivotiranjem ili potpunim pivotiranjem računaju faktorizaciju $PAQ = LU$ gdje su P i Q permutacije.

Iskorištavanje faktorizacije pojednostavljuje i analizu grešaka i praktično rješenje linearnog sustava. Rješenje $Ax = b$ dijeli se na fazu faktorizacije, $PA = LU$ za parcijalno pivotiranje ($O(n^3)$ operacija s pomičnom točkom) i supstitucijsku fazu, gdje se rješavaju trokutasti sustavi $Ly = Pb$, $Ux = y$ ($O(n^2)$ operacija s pomičnom točkom). Ako se rješava više od jednog sustava s istom matricom koeficijenata ali različitim desnim stranama, faktorizacija se može ponovno upotrijebiti uz odgovarajuću uštedu obavljenog posla.

Računanje LU faktorizacije $A = LU$ ekvivalentno je rješavanju jednadžbi:

$$a_{ij} = \sum_{r=1}^{\min(i,j)} l_{ir} u_{rj}.$$

Ove nelinearne jednadžbe lako se rješavaju ako se razmotre u pravilnom poretku. Za općenitost uzmimo $A \in \mathbb{R}^{m\times n} (m \geq n)$, te LU faktorizaciju za $L \in \mathbb{R}^{m\times n}$ i $U \in \mathbb{R}^{n\times n}$ (L je donje trapezoidna: $l_{ij} = 0$ za $i < j$). Pretpostavimo da znamo prvih $k-1$ stupaca od L i prvih $k-1$ redaka od U. Stavimo $l_{kk} = 1$,

$$a_{kj} = l_{k1}u_{1j} + \cdots + l_{k,k-1}u_{k-1,j} + u_{kj} \ , \ j = k : n \tag{3.3}$$

$$a_{ik} = l_{i1}u_{1k} + \cdots + l_{ik}u_{kk} \ , \ i = k+1 : m \tag{3.4}$$

Faktorizaciju možemo riješiti za crveno obojane elemente u k-tom retku od U te zatim u k-tom stupcu od L. Ovaj postupak nazivamo *Doolittle*-ova metoda.

Algoritam 3.2.2. *(Doolittle-ova metoda) Ovaj algoritam računa LU faktorizaciju $A = LU \in \mathbb{R}^{m\times n}$, gdje je $m \geq n$ (pretpostavivši da faktorizacija postoji), Doolittle-ovom metodom.*

for $k = 1 : n$ ***do***
  for $j = k : n$ ***do***
  $(*)$ $u_{kj} = a_{kj} - \sum_{i=1}^{k-1} l_{ki}u_{ij}$
  end for
  for $i = k+1 : m$ ***do***
  $(**)$ $l_{ik} = (a_{ik} - \sum_{j=1}^{k-1} l_{ij}u_{jk})/u_{kk}$
  end for
end for

Cijena: $n^2(m - n/3)$ operacija s pomičnom točkom.

Doolittle-ova metoda je matematički ekvivalentna Gaussovim eliminacijama bez pivotiranja jer u (3.3) imamo

$$a_{kj} - l_{k1}u_{1j} - \cdots - l_{ks}u_{sj} \equiv a_{kj}^{(s+1)} \quad (j > k), \tag{3.5}$$

i slično za (3.4). Da smo odabrali normalizaciju $u_{ii} \equiv 1$, dobili bismo *Crout-ovu* metodu. *Crout-ova* i *Doolittle-ova* metoda su dobre za ručno računanje ili računanje pomoću kalkulatora jer izbjegavaju potrebu za spremanjem srednje količine $a_{ij}^{(k)}$. Također su zanimljive i kada možemo gomilati unutarnje produkte u produljenoj preciznosti.

Jednostavno je ugraditi parcijalno pivotiranje u *Doolittle-ovu* metodu. Međutim, *rook* pivotiranje i potpuno pivotiranje ne mogu biti ugrađeni bez mijenjanja metode.

3.3 Analiza greške

Analiza greške Gaussovih eliminacija je kombinacija grešaka analize unutarnjeg produkta i supstitucije. Kada shvatimo tu činjenicu, analiza postaje jasna. Ključno promatranje koje vodi do ove točke gledišta je ta da sve matematički ekvivalentne varijante Gaussovih eliminacija zadovoljavaju zajedničku ogradu greške. Da bi vidjeli zašto, za početak primijetimo vezu između standardnih Gaussovih eliminacija (kao što smo ju opisali) i *Doolittle-ove* metode opisane s (3.5). Bilo da je unutarnji produkt u (3.5) izračunat kao jedna operacija ili da se računa u više razdvojenih operacija, podržavaju točno iste greške zaokruživanja; to se sve mijenja u trenutku kada se izvrše greške zaokruživanja. Ako unutarnjem produktu dozvolimo drugačiji poredak, tada su stvarne greške zaokruživanja drugačije, ali zajednička ograda vrijedi za bilo koji poredak. Dovoljno je dakle analizirati *Doolittle-ovu* metodu. Također je dovoljno analizirati metodu bez pivotiranja, jer su Gaussove eliminacije sa parcijalnim, *rook* ili potpunim pivotiranjem ekvivalentne Gaussovim eliminacijama bez pivotiranja primijenjenim na permutiranu matricu.

Koraci (*) i (**) iz Algoritmu 3.2.2 su oblika $y = (c - \sum_{i=1}^{k-1} a_i b_i)/b_k$, što smo analizirali u Lemi 2.2.4. Primjenjujući lemu dolazimo do zaključka da bez obzira na poredak unutarnjih produkata, izračunate matrice $\hat{L}$ i $\hat{U}$ zadovoljavaju

$$\left| a_{kj} - \sum_{i=1}^{k-1} \hat{l}_{ki}\hat{u}_{ij} - \hat{u}_{kj} \right| \leq \gamma_k \sum_{i=1}^{k} |\hat{l}_{ki}|\,|\hat{u}_{ij}|\,,\ j \geq k$$

$$\left| a_{ik} - \sum_{j=1}^{k} \hat{l}_{ij}\hat{u}_{jk} \right| \leq \gamma_k \sum_{j=1}^{k} |\hat{l}_{ij}|\,|\hat{u}_{jk}|\,,\ i > k$$

Ove nejednakosti sačinjavaju rezultat greške unatrag za LU faktorizaciju.

Teorem 3.3.1. *Ako se Gaussove eliminacije primijenjene na $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ izvrše do kraja, onda izračunati LU faktori $\hat{L} \in \mathbb{R}^{m\times n}$ i $\hat{U} \in \mathbb{R}^{n\times n}$ zadovoljavaju*

$$\hat{L}\hat{U} = A + \Delta A\,, \quad |\Delta A| \leq \gamma_n |\hat{L}|\,|\hat{U}|. \tag{3.6}$$

Uz malo više truda, rezultat greške unatrag može se dobiti za rješenje $Ax = b$.

Teorem 3.3.2. *Neka je $A \in \mathbb{R}^{n\times n}$ i pretpostavimo da Gaussova eliminacija daje izračunate LU faktore $\hat{L}$ i $\hat{U}$, te izračunato rješenje $\hat{x}$ od $Ax = b$. Tada je*

$$(A + \Delta A)\hat{x} = b\,, \quad |\Delta A| \leq \gamma_{3n}|\hat{L}|\,|\hat{U}|. \tag{3.7}$$

Dokaz. Iz Teorema 3.3.1 imamo $\hat{L}\hat{U} = A + \Delta A_1$, $|\Delta A_1| \leq \gamma_n|\hat{L}|\,|\hat{U}|$. Po Teoremu 2.2.5., supstitucija daje $\hat{y}$ i $\hat{x}$ koji zadovoljavaju

$$\begin{aligned}(\hat{L} + \Delta L)\hat{y} &= b\,, \quad |\Delta L| \leq \gamma_n|\hat{L}| \\ (\hat{U} + \Delta U)\hat{x} &= \hat{y}\,, \quad |\Delta U| \leq \gamma_n|\hat{U}|\end{aligned}$$

Prema tome

$$\begin{aligned}b &= (\hat{L} + \Delta L)(\hat{U} + \Delta U)\hat{x} \\ &= (A + \Delta A_1 + \hat{L}\Delta U + \Delta L\hat{U} + \Delta L\Delta U)\hat{x} \\ &= (A + \Delta A)\hat{x}\,,\end{aligned}$$

gdje je

$$\begin{aligned}|\Delta A| &= |\Delta A_1 + \hat{L}\Delta U + \Delta L\hat{U} + \Delta L\Delta U| \\ &\leq |\Delta A_1| + |\hat{L}|\,|\Delta U| + |\Delta L|\,|\hat{U}| + |\Delta L|\,|\Delta U| \\ &\leq (3\gamma_n + \gamma_n^2)|\hat{L}|\,|\hat{U}| \\ &= \Big(\frac{3nu}{1-nu} + \frac{n^2u^2}{(1-nu)^2}\Big)|\hat{L}||\hat{U}| \\ &= \frac{3nu(1-nu) + n^2u^2}{(1-nu)^2}|\hat{L}|\,|\hat{U}| \\ &= \frac{3nu - 2n^2u^2}{1 - 2nu + n^2u^2}|\hat{L}|\,|\hat{U}| \\ &\leq \frac{3nu}{1-2nu}|\hat{L}||\hat{U}| \\ &\leq \gamma_{3n}\,|\hat{L}|\,|\hat{U}|.\end{aligned}$$

□

Kako interpretiramo Teorem 3.3.2.? Idealno bismo željeli $|\Delta A| \leq u|A|$, što odgovara nesigurnosti uvedenoj sa zaokruživanjem elemenata od A, ali iz razloga što svaki element od A prolazi do n aritmetičkih operacija ne možemo očekivati bolju ogradu od $|\Delta A| \leq c_n u|A|$, gdje je c_n konstanta reda n. Takva ograda vrijedi ako $\hat{L}$ i $\hat{U}$ zadovoljavaju $|\hat{L}||\hat{U}| = |\hat{L}\hat{U}|$, što sigurno vrijedi ako su $\hat{L}$ i $\hat{U}$ nenegativne, jer tada (3.6) daje

$$\begin{aligned}|\hat{L}|\,|\hat{U}| = |\hat{L}\hat{U}| = |A + \Delta A| \leq |A| + \gamma_n|\hat{L}||\hat{U}| \\ \Rightarrow |\hat{L}||\hat{U}| \leq \frac{1}{1-\gamma_n}|A|.\end{aligned} \tag{3.8}$$

Ako ubacimo to u (3.7) dobivamo

$$(A + \Delta A)\hat{x} = b \ , \ |\Delta A| \leq \frac{\gamma_{3n}}{1 - \gamma_n}|A| \quad (\hat{L}, \hat{U} \geq 0). \tag{3.9}$$

Ovaj rezultat kaže da $\hat{x}$ ima malu komponentnu relativnu grešku unatrag.

Jedna klasa matrica koja ima nenegativne LU faktore definirana je na idući način. $A \in \mathbb{R}^{n\times n}$ je potpuno pozitivna (nenegativna) ako je determinanta svake kvadratne podmatrice pozitivna (nenegativna). Posebno, ova definicija zahtijeva da a_{ij} i $det(A)$ budu pozitivne ili nenegativne. Ako je A potpuno nenegativna ona ima LU faktorizaciju $A = LU$ u kojoj su L i U potpuno nenegativne, tako da $L \geq 0$ i $U \geq 0$; štoviše, izračunati faktori $\hat{L}$ i $\hat{U}$ iz Gaussovih eliminacija su nenegativni za dovoljno male vrijednosti jediničnog zaokuživanja u. Inverzi potpuno nenegativnih matrica također imaju svojstvo $|A| = |L||U|$. Primijetimo da je svojstvo matrice ili njenog inverza da bude potpuno nenegativna općenito uništeno pod retčanom permutacijom. Stoga je za potpuno nenegativne matrice i njene inverzne matrice najbolje koristiti Gaussove eliminacije *bez* pivotiranja.

Bitna činjenica koja slijedi iz (3.6) i (3.7) je ta da stabilnost Gaussovih eliminacija nije određena veličinom multiplikatora $\hat{l}_{ij}$ nego veličinom matrice $|\hat{L}||\hat{U}|$. Ta matrica može biti mala kada su multiplikatori veliki, i velika kada su multiplikatori reda 1.

Za daljnje razumijevanje stabilnosti Gaussovih eliminacija promotrimo norme. Za Gaussovu eliminaciju bez pivotiranja, omjer $\| \, |\hat{L}| \, |\hat{U}| \, \|/\|A\|$ može biti proizvoljno velik. Na primjer, za matricu $\begin{bmatrix} \epsilon & 1 \\ 1 & 1 \end{bmatrix}$ omjer je reda ϵ^{-1}. Pretpostavimo da je upotrebljeno parcijalno pivotiranje. Tada je $|l_{ij}| \leq 1$ za sve $i \geq j$, obzirom da je l_{ij} multiplikator. Koristeći relaciju

$$|a_{ij}^{(k+1)}| \leq |a_{ij}^{(k)}| + |a_{kj}^{(k)}| \leq 2 \max |a_{ij}^{(k)}|$$

možemo vidjeti da je

$$|a_{ij}^{(i)}| \leq 2 \max |a_{ij}^{(i-1)}| \leq 2^2 \max |a_{ij}^{(i-2)}| \leq \cdots \leq 2^{i-1} \max |a_{ij}|,$$

pa imamo $|u_{ij}| \leq 2^{i-1} \max |a_{ij}|$. Stoga, za parcijalno pivotiranje, L je mali i U relativno ograničen sa A.

Uobičajeno, analiza greške unatrag za Gaussovu eliminaciju izražena je u terminima *faktora rasta*

$$\rho_n = \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|},$$

što uključuje sve elemente $a_{ij}^{(k)}(k = 1 : n)$ koji se javljaju tijekom eliminacija. Koristeći ogradu $|u_{ij}| = |a_{ij}^{(i)}| \leq \rho_n \max_{i,j} |a_{ij}|$ dobivamo idući klasični teorem.

Teorem 3.3.3 (Wilkinson)**.** *Neka je $A \in \mathbb{R}^{n\times n}$ i pretpostavimo da Gaussove eliminacije s parcijalnim pivotiranjem daju izračunato rješenje $\hat{x}$ od $Ax = b$. Tada je*

$$(A + \Delta A)\hat{x} = b\ ,\ \|\Delta A\|_\infty \leq n^2\gamma_{3n}\rho_n\|A\|_\infty. \tag{3.10}$$

Dokaz. Računamo normu greške iz Teorema 3.3.2:

$$\begin{aligned}
\|\Delta A\|_\infty &= \| \, |\Delta A| \, \|_\infty \leq \gamma_{3n} \| \, |\hat{L}| \, |\hat{U}| \, \|_\infty \\
&= \gamma_{3n} \max_{i=1,\ldots,n} \sum_{j=1}^{n} \left(|\hat{L}| \, |\hat{U}| \right)_{ij} \\
&= \gamma_{3n} \max_{i=1,\ldots,n} \sum_{j=1}^{n} \sum_{k=1}^{\min(i,j)} |\hat{l}_{ik}| \, |\hat{u}_{kj}| \\
&\leq \gamma_{3n} \max_{i=1,\ldots,n} \sum_{j=1}^{n} \sum_{k=1}^{min(i,j)} 1 \cdot \rho_n \max_{s,t} |a_{st}| \\
&\leq \gamma_{3n}\rho_n\|A\|_\infty \max_{i=1,\ldots,n} \sum_{j=1}^{n} \min(i, j) \\
&\leq \gamma_{3n}\rho_n\|A\|_\infty \max_{i=1,\ldots,n} ni \\
&\leq n^2\gamma_{3n}\rho_n\|A\|_\infty
\end{aligned}$$

□

Priznajmo da smo koristili nedopušteni manevar u izvođenju ovog teorema: koristili smo ograde za $\hat{L}$ i $\hat{U}$ koje strogo vrijede samo za precizne L i U. Umjesto toga smo mogli definirati faktor rasta u terminima izračunatog $\hat{a}_{ij}^{(k)}$, ali tada bi svaka ograda za faktor rasta uključivala jediničnu grešku zaokruživanja. Naše kršenje točnosti je bezopasno za svrhe u kojima ćemo koristiti teorem.

Pretpostavka *Wilkinsonovog* teorema da se koristi parcijalno pivotiranje nije nužna: $\| \, |L| \, |U| \, \|_\infty$ se također može ograničiti u terminima faktora rasta za Gaussove eliminacije bez pivotiranja, što možemo vidjeti u idućem rezultatu.

Lema 3.3.4. *Ako je $A = LU \in \mathbb{R}^{n\times n}$ LU faktorizacija dobivena Gaussovim eliminacijama bez pivotiranja, onda je*

$$\| \, |L| \, |U| \, \|_\infty \leq (1 + 2(n^2 - n)\rho_n)\|A\|_\infty$$

Dokaz. Označimo s l_j j-ti stupac od L, te s u_i^T i-ti redak od U. Tada je

$$\tilde{A}^{(k+1)} = \tilde{A}^{(k)} - l_k u_k^T\ ,\ \ k = 1 : n - 1$$

gdje $\tilde{A}^{(k)}$ označava $A^{(k)}$ s nul-elementima u redovima $1 : k-1$. Sada imamo

$$|l_k|\,|u_k^T| = |l_k u_k^T| \leq |\tilde{A}^{(k)}| + |\tilde{A}^{(k+1)}|$$

stoga

$$\begin{aligned}
|L|\,|U| &= \sum_{k=1}^{n} |l_k|\,|u_k^T| \\
&= \sum_{k=1}^{n-1} |l_k|\,|u_k^T| + e_n \cdot |u_{nn}| \cdot e_n^T \\
&\leq \sum_{k=1}^{n-1} \left(|\tilde{A}^{(k)}| + |\tilde{A}^{(k+1)}|\right) + |u_{nn}| e_n e_n^T \\
&\leq |\tilde{A}^{(1)}| + |\tilde{A}^{(2)}| + |\tilde{A}^{(2)}| + |\tilde{A}^{(3)}| + \cdots + |\tilde{A}^{(n-1)}| + |\tilde{A}^{(n)}| + |u_{nn}| e_n e_n^T \\
&= |\tilde{A}^{(1)}| + |\tilde{A}^{(2)}| + |\tilde{A}^{(2)}| + |\tilde{A}^{(3)}| + \cdots + |\tilde{A}^{(n-1)}| + |\tilde{A}^{(n)}| + |\tilde{A}^{(n)}| \\
&= |A| + 2\sum_{k=2}^{n} |\tilde{A}^{(k)}|
\end{aligned}$$

Uzmemo li norme i koristeći $|a_{ij}^{(k)}| \leq \rho_n \max_{i,j} |a_{ij}|$ imamo

$$\| \,|L|\,|U|\,\|_\infty \leq \|A\|_\infty + 2(n-1)n\rho_n \|A\|_\infty = (1 + 2(n-1)n\rho_n)\|A\|_\infty$$

□

Budući da je stabilnost unatrag po normi Gaussovih eliminacija sa ili bez pivotiranja određena faktorom rasta, u nastavku ćemo navesti dva rezultata vezana uz faktora rasta.

3.4 Faktor rasta

Za potrebe idućeg teorema označimo s ρ_n^p faktor rasta parcijalnog pivotiranja.

Teorem 3.4.1 (Higham and Higham)**.** *Sve realne $n \times n$ matrice A za koje je $\rho_n^p(A) = 2^{n-1}$ su oblika*

$$A = DM \begin{bmatrix} T & \vdots & \alpha d \\ 0 & \vdots & \end{bmatrix}$$

gdje je $D = diag(\pm 1)$, M je jedinična donje trokustasta s $m_{ij} = -1$ za $i > j$, T je proizvoljna regularna gornje trokutasta matrica reda $n-1$, $d = (1, 2, 4, \ldots, 2^{n-1})^T$ i α je skalar takav da $\alpha := |a_{1n}| = \max_{i,j} |a_{ij}|$.

Dokaz. Nicholas J. Highmam [10, p. 167]. □

Teorem 3.4.2 (Higham and Higham)**.** *Neka je $A \in \mathbb{C}^{n\times n}$ regularna i stavimo $\alpha = \max_{i,j} |a_{ij}|$, $\beta = \max_{i,j} |(A^{-1})_{ij}|$, i $\theta = (\alpha\beta)^{-1}$. Tada $\theta \leq n$, i za sve permutacijske matrice P i Q takve da PAQ ima LU faktorizaciju, faktor rasta ρ_n za Gaussove eliminacije bez pivotiranja na PAQ zadovoljava $\rho_n \geq \theta$.*

Dokaz. Nicholas J. Highmam [10, p. 168]. □

3.5 Skaliranje i izbor pivotne strategije

Prije rješavanja linearnog sustava $Ax = b$ Gaussovim eliminacijama, slobodni smo skalirati redove i stupce:

$$Ax = b \rightarrow D_1AD_2 \cdot D_2^{-1}x = D_1b, \quad \text{ili} \quad A^{'}y = c, \tag{3.11}$$

gdje su D_1 i D_2 regularne dijagonalne matrice. Primijenimo li Gaussove eleminacije na skalirani sustav $A^{'}y = c$ pa dobivamo x iz $x = D_2y$. Unatoč tomu što je skaliranje već godinama korišteno u programima za Gaussove elminacije, još uvijek nije potpuno jasno kako najbolje izabrati skaliranje, te niti jedan algoritam za skaliranje nema garancije da će uvijek biti zadovoljavajuć. Wilkinsonova opaska: "*We cannot decide whether equations are ill-conditioned without examining the way in which the coefficients were derived*" prilično dobro sumira problem skaliranja.

Posljedica skaliranja u Gaussovim eliminacijama bez pivotiranja jednostavna je za opisati. Ako su elementi od D_1 i D_2 potencije strojne baze β (tako da je skaliranje izvršeno bez pogreške) i Gaussove eliminacije daju $\hat{L}$ i $\hat{U}$ koji zadovoljavaju $A+\Delta A = \hat{L}\hat{U}$ tada Gaussove eliminacije na $A^{'} = D_1AD_2$ daju $D_1\hat{L}D_1^{-1}$ i $D_1\hat{U}D_2$ koji zadovoljavaju $A^{'} + D_1\Delta AD_2 = D_1\hat{L}D_1^{-1} \cdot D_1\hat{U}D_2$. Drugim riječima, greška zaokruživanja u Gaussovim eliminacijama skalira se na isti način kao A. Međutim, kod parcijalnog pivotiranja, izbor pivota je pod utjecajem retčanog skaliranja (iako ne i stupčanog skaliranja), i to na način koji je teško predvidjeti.

Možemo uzeti nezavisni pristup metodi skaliranja, uzimajući u obzir bilo koju metodu za rješavanje $Ax = b$ koja daje rješenje koje zadovoljava

$$\frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} \leq c_n\kappa_\infty(A)u,$$

gdje je c_n konstanta. Za skalirani sustav (3.11) imamo

$$\frac{\|D_2^{-1}(x - \hat{x})\|_\infty}{\|D_2^{-1}x\|_\infty} \leq c_n\kappa_\infty(D_1AD_2)u\ ,$$

pa je prirodno izabrati D_1 i D_2 koji minimiziraju $\kappa_\infty(D_1AD_2)$. Kao što smo vidjeli u Teoremu 1.4.4. minimalna moguća vrijednost je $\rho(|A||A^{-1}|)$. Međutim, stupčano skaliranje (obično) ima nepoželjni učinak mijenjanja norme u kojoj je greška mjerena. Uz samo retčano skaliranje, minimalna vrijednost $\kappa_\infty(D_1A)$ je $cond(A) = \| \, |A^{-1}| \, |A| \, \|_\infty$, koja se dostiže kada D_1A ima retke jedinične 1-norme. Time retčana ravnoteža daje grešku unaprijed ograničenu sa *cond*.

Specijalno za Gaussove eliminacije možemo reći više. Ako je A retčano uravnotežena u 1-normi onda je $|A|e = e$, pa stoga iz Teorema 3.3.2, greška unatrag matrice ΔA zadovoljava

$$|\Delta A| \leq \gamma_{3n}|\hat{L}||\hat{U}| \leq \gamma_{3n}\| \, |\hat{L}| \, |\hat{U}| \, \|_\infty ee^T = \gamma_{3n}\| \, |\hat{L}| \, |\hat{U}| \, \|_\infty |A| ee^T.$$

Drugim riječima, ako su Gaussove eliminacije unatrag stabilne po normi onda je i unatrag retčano stabilna.

Može se pokazati da je greška unaprijed ograničena sa *cond* garantirana za Gaussove eliminacije bez pivotiranja za retčano dijagonalno dominantne matrice, i za Gaussove eliminacije s *rook* pivotiranjem ili potpunim pivotiranjem na proizvoljnoj matrici, pod uvjetom da je $cond(U)$ reda 1. Postoje i druge mogućnosti: ako umjesto A LU faktoriziramo A^T, tada GEPP ima grešku unaprijed ograničenu sa *cond* ako je $cond(L^T)$ reda 1. Uz prikladno skaliranje moguće je postići i bolje. Skeel [11] je pokazao da za $D_1 = diag(|A| \, |x|)^{-1}$, GEPP na A je unatrag stabilna u komponentnom relativnom smislu, i stoga ograda greške unatrag proporcionalna s $cond(A, x) = \| \, |A^{-1}| \, |A| \, |x| \, \|_\infty / \|x\|_\infty$ vrijedi; stvar je u tome da skaliranje ovisi o nepoznatom rješenju x ! Retčano ujednačavanje može se smatrati približavanjem x-u pomoću e u ovom "optimalnom" skaliranju.

Unatoč raznolikim mogućnostima skaliranja i načinima pivotiranja, i obimu situacija u kojima određeni načini pivotiranja mogu imati manju ogradu greške nego što teorija predviđa, u praksi su opći linearni sustavi virtualno uvijek riješeni neskaliranom GEPP. Navedimo tri glavna razloga za to:

1. Većina korisnika smatra da se GEPP u praksi dobro izvodi

2. GEPP ima dugu povijest korištenja u softverskim bibliotekama i paketima, i inercija otežava njegovu zamjenu

3. Greška unaprijed ograničena sa $cond(A, x)$ i mala greška unatrag po komponentama dostižu se primjenom iterativnog profinjenja sa fiksnom preciznošću nakon GEPP.

Bibliografija

[1] Robert D: Skeel. Scaling for numerical stability in Gaussian elimination. *J. Assoc. Comput. Mach.*, 26(3):494-526, 1979

[2] Shivkumar Chandrasekaran and Ilse C.F.Ipsen. On the sensitivity of solution components in linear systems of equations. *SIAM J. Matrix Anal. Appl*, 16(1):93-112, 1995

[3] W.Kahan. Numerical linear algebra. *Canadian Math. Bulletin*, 9:757-801, 1966

[4] A. van der Sluis. Condition numbers and equilibration of matrices. *Numer. Math.*,14:14-23, 1969

[5] George E. Forsythe and E. G. Straus. On best conditioned matrices. *Proc. Amer. Math. Soc.*, 6:340-345, 1955

[6] F. L. Bauer. Optimally scaled matrices. *Numer. Math.*, 5:73-87, 1963

[7] Siegfried M. Rump. Optimal scaling for P-norms and componentwise distance to singularity. *IMA J. Numer. Anal.*, 23(1):1-9, 2003.

[8] M. Ariuli, J.W. Demmel, and I.S. Duff. Solving sparse linear systems with sparse backward error. *SIAM J. Matrix Anal. Appl.*, 10(2):165-190, 1989

[9] Leslie V. Foster. The growth factor and efficiency of Gaussian elimination with rook pivoting. *J. Comput. Appl. Math.*, 86:177-194, 1997.

[10] Nicholas J. Higham. Accuracy and Stability of Numerical Algorithms,Second edition, *SIAM* , Philadelphia, 2002.

[11] Robert D. Skeel. Effect of equilibration on resiudal size for partial pivoting. *SIAM J. Numer. Anal.*, 18(3):449-454, 1981

Sažetak

Glavna tema ovog rada je točnost koju možemo očekivati kada sustav linearnih jednadžbi $Ax = b$, gdje je $A \in \mathbb{R}^{n\times n}$, rješavamo u aritmetici konačne preciznosti na računalu. Kao prvo, obradili smo teoriju perturbacije za sustave linearnih jednadžbi u kojoj se daje odgovor na pitanje koliko je rješenje sustava osjetljivo na perturbacije ulaznih podataka A i b. Nadalje ta teorija daje povratnu grešku izračunate aproksimacije rješenja, i procjenjuje ogradu za njezinu grešku unaprijed. Nakon toga posvetili smo se analizi numeričke stabilnosti direktnih metoda za rješavanje sustava, baziranih na LU faktorizaciji. Ta analiza daje povratnu grešku za dobivenu aproksimaciju rješenja preko direktne metode, koja se potom uklapa u teoriju perturbacije za dobivanje greške unaprijed. Rezultat analize diktira koji parametri mogu utjecati na numeričku stabilnost algoritma.

Summary

The main subject of this thesis is accuracy we can expect when we solve the system of linear equations $Ax = b$, where $A \in \mathbb{R}^{n\times n}$, in finite precision arithmetic on a computer. First, we have elaborated the perturbation theory for linear equation systems in which we describe how sensitive the system solution is to the perturbation of input data A i b. Further, this theory gives a backward error of the approximation of the solution, and estimates the bound for its forward error. After that we are devoted to the analysis of the numerical stability of direct methods for solving the system, based on LU factorization. This analysis gives a backward error to the resulting approximation of the solution via a direct method, which is then inserted in the perturbation theory to get the forward error. The result of the analysis dictates which parameters can affect the numerical stability of the algorithm.

Životopis

Diana Šenjug rođena je u Zagrebu dana 04.02.1991. Svoje djetinjstvo provela je u Velikoj Gorici gdje je pohađala osnovnu školu Eugena Kumičića i osnovnu glazbenu školu Franje Lučića. Ljubav prema matematici rodila se u ranim danima te bira svoje školovanje nastaviti u XV gimnaziji, gdje sve razrede završava s odličnim uspjehom. Želja za daljnjim razvijanjem u području matematike usmjerava ju na Prirodoslovno matematički fakultet u Zagrebu gdje titulu prvostupnika matematike stječe 2015. godine nakon čega upisuje diplomski studij Primijenjene matematike.