Numerička Matematika

Studeni 2010.

Sadržaj

Poglavlje 1

Svojstvene vrijednosti

1.1 Karakteristični polinom

Algebarska definicija svojstvene vrijednosti λ matrice A kao skalara za kojeg je, s nekim vektorom x različitim od nul–vektora, ispunjeno $\mathsf{A}x = \lambda x$ povlači elegantnu algebarsku karakterizaciju svojstvenih vrijednosti kao nultočaka određenog polinoma.

Kako $(\mathsf{A} - \lambda\mathsf{I})x = \mathbf{0}$ i $x \neq \mathbf{0}$ povlače da $\mathsf{A} - \lambda\mathsf{I}$ ima netrivijalnu jezgru tj. da je singularna, odmah je $\det(\lambda\mathsf{I} - \mathsf{A}) = 0$. Iz definicije determinante je odmah jasno da je $\det(\lambda\mathsf{I} - \mathsf{A})$ polinom stupnja n u varijabli λ,

$$\det\left(\begin{pmatrix} \lambda-a_{11} & -a_{12} & \cdots & -a_{1,n-1} & -a_{1n} \\ -a_{21} & \lambda-a_{22} & \cdots & -a_{2,n-1} & -a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -a_{n-1,1} & & & \lambda-a_{n-1,n-1} & -a_{n-1,n} \\ -a_{n1} & \cdots & \cdots & -a_{n,n-1} & \lambda-a_{nn} \end{pmatrix}\right) = \lambda^n - \operatorname{trag}(\mathsf{A})\lambda^{n-1} + \cdots + (-1)^n\det(\mathsf{A})$$

Definicija 1.1.1. Za $\mathsf{A} \in \mathbb{C}^{n\times n}$ je karakteristični polinom od A definiran s $\chi_{\mathsf{A}}(\lambda) = \det(\lambda\mathsf{I} - \mathsf{A})$.

Dakle, pitanje egzistencije svojstvenih vrijednosti matrice možemo jednostavno odgovoriti – polinom stupnja n uvijek ima n općenito kompleksnih svojstvenih vrijednosti, računato s kratnostima.

Teorem 1.1.1. *Matrica $\mathsf{A} \in \mathbb{C}^{n\times n}$ ima n svojstvenih vrijednosti koje su općenito kopleksni brojevi i koje brojimo kao nultočke $\lambda_1, \lambda_2, \ldots, \lambda_n$ karakterističnog polinoma $\chi_{\mathsf{A}}(\lambda)$, zajedno sa kratnostima. Vrijedi formula* $\det(\lambda\mathsf{I} - \mathsf{A}) = \prod_{i=1}^{n}(\lambda - \lambda_i)$. *Specijalno je* $\operatorname{trag}(\mathsf{A}) = \sum_{i=1}^{n}\lambda_i$, $\det(\mathsf{A}) = (-1)^n\prod_{i=1}^{n}\lambda_i$.

Definicija 1.1.2. Neka je λ svojstvena vrijedost od A i neka je α njena kratnost kao nultočke karakterističnog polinoma χ_{A}. Tada je α algebarska kratnost od λ.

Korolar 1.1.2. *Neka je $\mathsf{A} \in \mathbb{F}^{n\times n}$, sa svojstvenim vrijednostima $\lambda_1, \lambda_2, \ldots, \lambda_n$. Iz Teorema 1.1.1 odmah slijedi:*

- *Slične matrice A i $\mathsf{S}^{-1}\mathsf{AS}$ imaju isti karakteristični polinom pa i iste svojstvene vrijednosti.*
- *Matrice A i A^{T} imaju iste svojstvene vrijednosti. Svojstvene vrijednosti matrica $\overline{\mathsf{A}}$ i A^* su $\overline{\lambda}_1, \overline{\lambda}_2, \ldots, \overline{\lambda}_n$.*
- *Ako je A regularna, onda su svi $\lambda_i \neq 0$ i svojstvene vrijednosti od A^{-1} su $1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_n$.*

Propozicija 1.1.3. *$\chi_{\mathsf{A}}(\cdot)$ je neprekidna funkcija od A.*

Propozicija 1.1.4. *Ako su $\lambda_1, \ldots, \lambda_k$ k međusobno različitih svojstvenih vrijednosti od A, i ako su $x_1, \ldots, x_k$ pripadni svojstveni vektori, onda su $x_1, \ldots, x_k$ linearno nezavisni.*

Definicija 1.1.3. Neka je λ svojstvena vrijedost od A i neka je γ dimenzija jezgre matrice $\mathsf{A} - \lambda\mathsf{I}$. Tada je γ geometrijska kratnost od λ.

1.1.1 Hamilton–Cayleyev teorem

Za $n \times n$ matricu A sigurno možemo odrediti polinom $p(\zeta) = \sum_{j=0}^{s} \beta_j \zeta^j$ nekog stupnja $s \leq n^2$ tako da je $p(\mathsf{A}) = \mathbf{0}$. To slijedi iz činjenice da je niz vektora

$$\mathsf{I}, \mathsf{A}, \mathsf{A}^2, \ldots, \mathsf{A}^n, \ldots, \mathsf{A}^{n^2}$$

nužno linearno zavisan nad $\mathbb{F}$.

Teorem 1.1.5. *Svaka matrica poništava svoj karakteristični polinom, $\chi_{\mathsf{A}}(\mathsf{A}) = \mathbf{0}$.*

Dokaz: Neka je A dimenzije $n \times n$. Karakteristični polinom ćemo zapisati u obliku

$$\chi_A(\lambda) = \lambda^n + \sum_{j=0}^{n-1} \alpha_j \lambda^j.$$

Iz definicije adjunkte je $\mathrm{adj}(\lambda\mathsf{I} - \mathsf{A})(\lambda\mathsf{I} - \mathsf{A}) = \det(\lambda\mathsf{I} - \mathsf{A})\mathsf{I}$, pri čemu je, s nekim konstantnim matricama B_j (ovisnim o A)

$$\mathrm{adj}(\lambda\mathsf{I} - \mathsf{A}) = \sum_{j=0}^{n-1} \lambda^j \mathsf{B}_j.$$

Usporedbom koeficijenata uz potencije λ^j u relaciji

$$\sum_{j=0}^{n-1} \lambda^j \mathsf{B}_j(\lambda\mathsf{I} - \mathsf{A}) \equiv \sum_{j=0}^{n-1} \lambda^{j+1}\mathsf{B}_j - \sum_{j=0}^{n-1} \lambda^j \mathsf{B}_j\mathsf{A} = (\lambda^n + \sum_{j=0}^{n-1} \alpha_j \lambda^j)\mathsf{I}$$

dobijemo niz jednakosti

$$\begin{array}{rcrcr} \mathsf{B}_{n-1} & & & = & \mathsf{I} \\ \mathsf{B}_{n-2} & - & \mathsf{B}_{n-1}\mathsf{A} & = & \alpha_{n-1}\mathsf{I} \\ \mathsf{B}_{n-3} & - & \mathsf{B}_{n-2}\mathsf{A} & = & \alpha_{n-2}\mathsf{I} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathsf{B}_0 & - & \mathsf{B}_1\mathsf{A} & = & \alpha_1\mathsf{I} \\ & & -\mathsf{B}_0\mathsf{A} & = & \alpha_0\mathsf{I} \end{array} \tag{1.1.1}$$

Ako te jednakosti sada pomnožimo s desna redom potencijama A^n, A^{n-1}, A^{n-2},..., $\mathsf{A}^1, \mathsf{A}^0 \equiv \mathsf{I}$ i odvojeno zbrojimo sve lijeve i sve desne strane, dobijemo $\mathbf{0} = \chi_A(\mathsf{A})$. ⊞

Komentar 1.1.1. A dijagonalizabilna, $\chi_A(A) = S\chi_A(\Lambda)S^{-1} = 0$.
$A = \lim_{\epsilon\to 0} A_\epsilon$, A_ϵ dijagonalizabilne.
$\chi_A(A) = \lim_{\epsilon\to 0} \chi_{A_\epsilon}(A_\epsilon)$

Korolar 1.1.6. *Neka je* A *regularna matrica. Tada je* A^{-1} *polinom u matrici* A*, stupnja najviše* $n-1$*.*

Korolar 1.1.7. *Neka je* $p(\zeta) = \sum_{j=0}^{m} \alpha_j \zeta^j$ *polinom stupnja* $m > n$*. Matricu* $p(\mathsf{A})$ *možemo izračunati računanjem u* A *vrijednosti polinoma stupnja najviše* $n-1$*.*

Dokaz: Dijeljenjem polinoma $p(\cdot)$ s $\chi_A(\cdot)$ dobijemo $p(\zeta) = q(\zeta)\chi_A(\zeta) + r(\zeta)$, pri čemu je r stupnja najviše $n-1$ i vrijedi $p(A) = r(A)$. ⊞

1.1.2 Minimalni polinom

Vidimo da iz činjenice da A poništava određeni polinom stupnja n slijedi korisna formula koja A^{-1} izražava kao polinom stupnja najviše $n-1$. Također, računanje u A polinoma bilo kojeg stupnja se svodi na računanje polinoma stupnja najviše $n-1$. Iz dokaza je jasno da se se ova dva (a ima i drugih koje nismo spomenuli) računa pojednostavljuju ako je stupanj polinoma kojeg A poništava manji. Zato je važno znati koji je to polinom minimalnog stupnja kojeg A poništava.

Već znamo da A poništava polinom stupnja n. Jednostavno je zaključiti da mora postojati jedinstven minimalni stupanj $m \leq n$ takav da postoji polinom stupnja m kojeg A poništava i da ne postoji polinom stupnja manjeg od m koji izračunat u A daje nul matricu. Jednostavno u nizu $\mathsf{I}, \mathsf{A}, \mathsf{A}^2, \ldots$ tražimo prvu potenciju m za koju su $\mathsf{I}, \mathsf{A}, \mathsf{A}^2, \ldots, \mathsf{A}^m$ linearno zavisni, $\sum_{j=0}^{m} \mu_j \mathsf{A}^j = \mathbf{0}$, gdje bez smanjenja općenitosti možemo uzeti $\mu_m = 1$.

Propozicija 1.1.8. *Ako je* p *polinom minimalnog stupnja* m *za kojeg je* $p(A) = \mathbf{0}$*, onda* p *dijeli bez ostatka svaki polinom* f *stupnja* $k > m$ *za kojeg je* $f(A) = \mathbf{0}$*.*

1.1.3 Matrica pratilica polinoma

Vidjeli smo da je pridruživanje (karakterističnog) polinoma matrici omogućilo da se problem (svojstvenih vrijednosti) zadan na matricama elegantno prebaci u teoriju

polinoma. Sada ćemo uspostaviti vezu i u drugom smjeru – polinomu ćemo pridružiti matricu čiji je upravo on karakteristični polinom. Motivacija za takvu vezu je jasna, otvaramo mogućnost da teorijske rezultate i numeričke tehnike za računanje svojstvenih vrijednosti matrica primijenimo na nultočke polinoma. Neka je

$$p(\lambda) = \lambda^n + \sum_{j=0}^{n-1} \alpha_j \lambda^j$$

proizvoljan polinom stupnja n, normiran tako da ima jedinični koeficijent uz vodeću potenciju. U slučaju $n = 1$, $p(\lambda) = \lambda + \alpha_0$, je

$$\lambda + \alpha_0 = \det(\big(\lambda + \alpha_0\big)) = \det(1 - \big(-\alpha_0\big)).$$

Ako je $n = 2$, $p(\lambda) = \alpha_0 + \alpha_1\lambda + \lambda^2 = \alpha_0 + \lambda(\lambda + \alpha_1)$,

$$\alpha_0 + \lambda(\lambda + \alpha_1) = \det(\begin{pmatrix} \lambda & \alpha_0 \\ -1 & \lambda + \alpha_1 \end{pmatrix}) = \det(\lambda \mathsf{I}_2 - \begin{pmatrix} 0 & -\alpha_0 \\ 1 & -\alpha_1 \end{pmatrix})$$

Općenito, iz $\lambda^n + \sum_{j=0}^{n-1} \alpha_j \lambda^j = \lambda(\lambda^{n-1} + \sum_{j=1}^{n-1} \alpha_j \lambda^{j-1}) + \alpha_0$ vidimo da gornju proceduru možemo nastaviti za $n = 3, 4, \ldots$.

Propozicija 1.1.9. *Karakteristični polinom matrice*

$$\mathsf{F}_n = \begin{pmatrix} 0 & 0 & \cdots & 0 & -\alpha_0 \\ 1 & 0 & \ddots & \vdots & -\alpha_1 \\ 0 & \ddots & \ddots & 0 & \vdots \\ \vdots & \ddots & 1 & 0 & -\alpha_{n-2} \\ 0 & \cdots & 0 & 1 & -\alpha_{n-1} \end{pmatrix}$$

je ujedno i njen minimalni polinom i glasi $\chi_{\mathsf{F}_n}(\lambda) = \lambda^n + \sum_{j=0}^{n-1} \alpha_j \lambda^j$. *Matricu* F_n *zovemo Frobeniusova matrica pratilica polinoma.*

Dokaz: Lako provjerimo, razvojem determinante po zadnjem stupcu, da je zbilja $\det(\lambda \mathsf{I} - \mathsf{F}_n) = \lambda^n + \sum_{j=0}^{n-1} \alpha_j \lambda^j$. Nadalje, ako je $q(\lambda) = \sum_{j=0}^{\ell} \beta_j \lambda^j$ polinom stupnja $\ell < n$ ($\beta_\ell \neq 0$), pokazat ćemo da je $q(\mathsf{F}_n) \neq \mathbf{0}$. Primijetimo da je $\mathsf{F}_n^j e_1 = e_{j+1}$, $j = 1, \ldots, n-1$. Dakle,

$$q(\mathsf{F}_n)e_1 = \sum_{j=0}^{\ell} \beta_j \mathsf{F}_n^j e_1 = (\beta_0, \beta_1, \ldots, \beta_\ell, \underbrace{0, \ldots, 0}_{n-\ell-1})^{\mathrm{T}} \neq \mathbf{0}.$$

$\boxplus$

S obzirom na specijalnu strukturu matrice F_n, pokušat ćemo izvesti .....

Dakle, imamo matricu A i njen karakteristični polinom $\chi_A(\cdot)$, te F_n za koju je $\chi_{\mathsf{F}_n}(\cdot) = \chi_{\mathsf{A}}(\cdot)$

Stavimo $\hat{\mathsf{F}}_{n-1} = \mathsf{F}_n(2:n, 2:n)$. Vrijede

$$\alpha_0\mathsf{I} + \alpha_1\mathsf{F}_n + \cdots + \alpha_{n-1}\mathsf{F}_n^{n-1} + \mathsf{F}_n^n = \mathbf{0} \tag{1.1.2}$$

$$\alpha_1\mathsf{I} + \alpha_2\hat{\mathsf{F}}_{n-1} + \cdots + \alpha_{n-1}\hat{\mathsf{F}}_{n-1}^{n-2} + \hat{\mathsf{F}}_{n-1}^{n-1} = \mathbf{0} \tag{1.1.3}$$

Odmah je

$$\alpha_{n-1} = -\mathrm{trag}(\mathsf{F}_n) = -\mathrm{trag}(\hat{\mathsf{F}}_{n-1}) = -\mathrm{trag}(\mathsf{A}) = -(\lambda_1 + \cdots + \lambda_n). \tag{1.1.4}$$

Propozicija 1.1.10. *Vrijedi* $\mathrm{trag}(\mathsf{F}_n^k) = \mathrm{trag}(\hat{\mathsf{F}}_{n-1}^k)$, $k = 1, \ldots, n-1$.

Dokaz: Matricu F_n napišimo u obliku $\mathsf{F}_n = G_n + E_n$, gdje je $E_n = e_2e_1^{\mathrm{T}}$. Uočimo da je $G_n(2:n, 2:n) = \hat{\mathsf{F}}_{n-1}$ i $E_n^2 = \mathbf{0}$. Sada u jednakosti

$$\mathsf{F}_n^k = (G_n + E_n)^k = \sum_{j=0}^{k} \binom{k}{j} G_n^j E_n^{k-j} = G_n^k + kG_n^{k-1}E_n$$

uzmemo trag:

$$\mathrm{trag}(\mathsf{F}_n^k) = \mathrm{trag}(G_n^k) + k\mathrm{trag}(G_n^{k-1}E_n),$$

gdje je $\mathrm{trag}(G_n^{k-1}E_n) = \mathrm{trag}(G_n^{k-1}e_2e_1^{\mathrm{T}}) = e_1^{\mathrm{T}}G_n^{k-1}e_2$. Pokazat ćemo da je u matrici G_n^{k-1} element na poziciji $(1,2)$ jednak nuli. Vrijedi

$$G_n^{k-1} = G_nG_n^{k-2} = \begin{pmatrix} 0 & \alpha_n e_{n-1}^{\mathrm{T}} \\ \mathbf{0} & \hat{\mathsf{F}}_{n-1} \end{pmatrix} \begin{pmatrix} 0 & \star \\ \mathbf{0} & \hat{\mathsf{F}}_{n-1}^{k-2} \end{pmatrix}, \quad e_{n-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \in \mathbb{R}^{n-1}.$$

Kako je $k-2 \leq n-3$, a matrica $\hat{\mathsf{F}}_{n-1}$ je $(n-1)\times(n-1)$ Hessenbergova, potencija $\hat{\mathsf{F}}_{n-1}^{k-2}$ na poziciji $(n-1, 1)$ ima nulu. Dakle, $(G_n^{k-2})_{n,2} = 0$, pa je $(G_n^{k-1})_{12} = 0$. Dakle, $\mathrm{trag}(\mathsf{F}_n^k) = \mathrm{trag}(G_n^k) = \mathrm{trag}(\hat{\mathsf{F}}_{n-1}^k)$. $\boxplus$

Sada uočimo: Iz relacije (1.1.2) je

$$\alpha_0 = -\frac{1}{n}(\alpha_1\mathrm{trag}(\mathsf{F}_n) + \alpha_2\mathrm{trag}(\mathsf{F}_n^2) + \cdots + \alpha_{n-1}\mathrm{trag}(\mathsf{F}_n^{n-1}) + \mathrm{trag}(\mathsf{F}_n^n)), \tag{1.1.5}$$

a iz (1.1.3) i Propozicije 1.1.10 je

$$\alpha_1 = -\frac{1}{n-1}(\alpha_2 \mathrm{trag}(\mathsf{F}_n)+\alpha_3 \mathrm{trag}(\mathsf{F}_n^2)+\cdots+\alpha_{n-1}\mathrm{trag}(\mathsf{F}_n^{n-2})+\mathrm{trag}(\mathsf{F}_n^{n-1})). \quad (1.1.6)$$

Drugim riječima, ako znamo $\alpha_2, \ldots, \alpha_{n-1}$, onda pomoću tih vrijednosti i tragova potencija od F_n možemo izračunati α_1. Na isti način, ako imamo $\alpha_1, \ldots, \alpha_{n-1}$, onda α_0 računamo eksplicitnom formulom (1.1.5).

Na prvi pogled, prethodno razmatranje nema smisla jer je F_n zadana pomoću koeficijenata $\alpha_0, \ldots, \alpha_{n-1}$ koje sada kao pokušavamo izračunati pomoću tragova potencija od F_n. Ključna karika su relacije

$$\mathrm{trag}(\mathsf{F}_n^k) = \mathrm{trag}(\mathsf{A}^k),\ k = 0, 1, 2, 3, \ldots \quad (1.1.7)$$

Prisjetimo se kako smo rekurzivno konstruirali F_n. Matrica $\hat{\mathsf{F}}_{n-1}$ je iste strukture kao i F_n i pratilica je polinoma stupnja $n-1$. Ako u njoj uočimo podmatricu $\hat{\mathsf{F}}_{n-2} = \hat{\mathsf{F}}_{n-1}(2:n-1, 2:n-1) = \mathsf{F}_n(3:n, 3:n)$, onda je

$$\alpha_2 \mathsf{I} + \alpha_3 \hat{\mathsf{F}}_{n-2} + \cdots + \alpha_{n-1}\hat{\mathsf{F}}_{n-2}^{n-3} + \hat{\mathsf{F}}_{n-2}^{n-2} = \mathbf{0}$$

i još je, kao u Propoziciji 1.1.10,

$$\mathrm{trag}(\mathsf{F}_n^k) = \mathrm{trag}(\hat{\mathsf{F}}_{n-1}^k) = \mathrm{trag}(\hat{\mathsf{F}}_{n-2}^k),\ \ k = 1, \ldots, n-2.$$

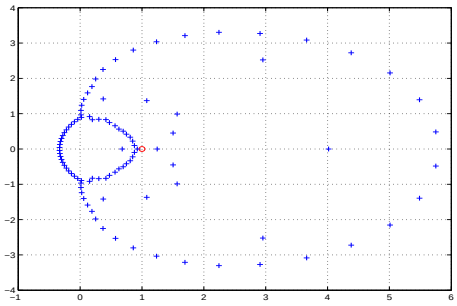
Dakle

$$\alpha_2 = -\frac{1}{n-2}(\alpha_3 \mathrm{trag}(\mathsf{F}_n)+\alpha_4 \mathrm{trag}(\mathsf{F}_n^2)+\cdots+\alpha_{n-1}\mathrm{trag}(\mathsf{F}_n^{n-3})+\mathrm{trag}(\mathsf{F}_n^{n-2})) \quad (1.1.8)$$

Teorem 1.1.11. *Koeficijente svojstvenog polinoma .. matrice* A *možemo računati sljedećim formulama:*

$$\begin{aligned} \alpha_{n-1} &= -\mathrm{trag}(\mathsf{A}) \\ \alpha_k &= -\frac{1}{n-k}\Big(\sum_{j=1}^{n-k-1} \alpha_{k+j}\mathrm{trag}(\mathsf{A}^j) + \mathrm{trag}(\mathsf{A}^{n-k})\Big),\ \ k = n-2, n-3, \ldots, 2, 1, 0. \end{aligned}$$

Primjer 1.1.1. Uzmimo da je A 100×100 jedinična matrica. Pomoću Teorema 1.1.11 ćemo izračunati njen karakteristični polinom i zatim njegove nultočke odrediti koristeći funkciju roots iz programskog paketa Matlab. Na Slici 1.1 je crvenim krugom označena jedinica (100–struka svojstvena vrijednost od I_{100}) a plavi plusevi su izračunate nultočke svojstvenog polinoma.



Slika 1.1:

1.2 Schurova dekompozicija

Spektralni elementi matrice A (svojstvene vrijednosti i svojstveni vektori) se jednostavno prenose na sličnu matricu $\mathsf{B} = \mathsf{S}^{-1}\mathsf{AS}$: Ako je $\mathsf{Bx} = \lambda\mathsf{x}$, onda je $\mathsf{ASx} = \lambda\mathsf{Sx}$. Odavde je jasno da je jedna razumna metoda za računanje svojstvenih vrijednosti i vektora matrice A traženje transformacija sličnosti koje će dati matricu B jednostavnije spektralne strukture (za koju je lako naći svojstvene vrijednosti i pripadne vektore). Pri tome, unitarna (ortogonalna) transformacija sličnosti S ima prednost jer se inverz matrice transformacija lako računa ($\mathsf{S}^{-1} = \mathsf{S}^*$), a u primjenama, kada moramo računati s neizbježnim perturbacijama, je važno da tranformacija sličnosti te perturbacije ne povećava. Nadalje unitarna transformacija sličnosti $\mathsf{B} = \mathsf{S}^*\mathsf{AS}$ čuva i druga važna svojstva: A je normalna, hertmitska, anti–hermitska, unitarna ako i samo ako je B, redom, normalna, hertmitska, anti–hermitska, unitarna.

Dakle, prirodno je pitanje koliko najviše možemo unitarnom transformacijom sličnosti pojednostaviti matricu $\mathsf{A} \in \mathbb{C}^{n\times n}$. Odgovor je tzv. Schurova dekompozicija koja je jedan od osnovnih alata u analiziranju i numeričkom rješavanju problema svojstvenih vrijednosti.

Teorem 1.2.1. *Neka je $\mathsf{A} \in \mathbb{C}^{n\times n}$ i neka su $\lambda_1, \ldots, \lambda_n$ svojstvene vrijednosti od A u proizvoljnom poretku. Postoji unitarna matrica U i gornje trokutasta matrica T tako da je $\mathsf{A} = \mathsf{UTU}^*$ i $\mathsf{T}_{ii} = \lambda_i$, $i = 1, \ldots, n$. Ako je $\mathsf{A} \in \mathbb{R}^{n\times n}$ i ako su sve svojstvene vrijednosti od A realne, onda je T također realna i U se može odabrati realna ortogonalna. Zapis $\mathsf{A} = \mathsf{UTU}^*$ zovemo Schurova dekompozicija od A, a matrica T se zove Schurova forma od A.*

Dokaz: Uzmimo svojstvenu vrijednost λ_1 i pripadni svojstveni vektor u_1 tako da

je $\mathsf{A}\mathsf{u}_1 = \lambda_1\mathsf{u}_1$, $\|\mathsf{u}_1\|_2 = 1$. Lako odredimo matricu $\hat{\mathsf{U}}_1 \in \mathbb{C}^{n\times(n-1)}$ tako da je $\mathsf{U}_1 = \begin{pmatrix}\mathsf{u}_1 & \hat{\mathsf{U}}_1\end{pmatrix}$ unitarna. (To je standardna dopuna do baze u unitarnom prostoru, koju najlakše dobijemo algoritamski iz QR faktorizacije matrice u_1.) Vrijedi

$$\mathsf{U}_1^*\mathsf{A}\mathsf{U}_1 = \begin{pmatrix}\mathsf{u}_1^*\\ \hat{\mathsf{U}}_1^*\end{pmatrix}\begin{pmatrix}\mathsf{A}\mathsf{u}_1 & \mathsf{A}\hat{\mathsf{U}}_1\end{pmatrix} = \begin{pmatrix}\lambda_1 & \mathsf{u}_1^*\mathsf{A}\hat{\mathsf{U}}_1\\ \mathbf{0} & \mathsf{A}_2\end{pmatrix},\quad \mathsf{A}_2 = \hat{\mathsf{U}}_1^*\mathsf{A}\hat{\mathsf{U}}_1.$$

Iz $\det(\mathsf{A}-\lambda\mathsf{I}_n) = (\lambda_1-\lambda)\det(\mathsf{A}_2-\lambda\mathsf{I}_{n-1})$ slijedi da su $\lambda_2,\ldots,\lambda_n$ svojstvene vrijednosti matrice A_2. Ako je $n>2$, uzimamo λ_2 i pripadni svojstveni vektor u_2 tako da je $\mathsf{A}_2\mathsf{u}_2 = \lambda_2\mathsf{u}_2$, $\|\mathsf{u}_2\|_2 = 1$. Kao i s λ_1, u_2 dopunimo do unitarne matrice $\mathsf{U}_2 = \begin{pmatrix}\mathsf{u}_2, \hat{\mathsf{U}}_2\end{pmatrix} \in \mathbb{C}^{(n-1)\times(n-1)}$, te izračunamo da je

$$\mathsf{U}_2^*\mathsf{A}_2\mathsf{U}_2 = \begin{pmatrix}\mathsf{u}_2^*\\ \hat{\mathsf{U}}_2^*\end{pmatrix}\begin{pmatrix}\mathsf{A}_2\mathsf{u}_2 & \mathsf{A}_2\hat{\mathsf{U}}_2\end{pmatrix} = \begin{pmatrix}\lambda_2 & \mathsf{u}_2^*\mathsf{A}_2\hat{\mathsf{U}}_2\\ \mathbf{0} & \mathsf{A}_3\end{pmatrix},\quad \mathsf{A}_3 = \hat{\mathsf{U}}_2^*\mathsf{A}_2\hat{\mathsf{U}}_2.$$

Ako ova dva koraka napravimo zajedno, imamo

$$\begin{pmatrix}1 & \mathbf{0}\\ \mathbf{0} & \hat{\mathsf{U}}_2^*\end{pmatrix}\mathsf{U}_1^*\mathsf{A}\underbrace{\mathsf{U}_1\begin{pmatrix}1 & \mathbf{0}\\ \mathbf{0} & \hat{\mathsf{U}}_2\end{pmatrix}}_{\text{unitarna}} = \left(\begin{array}{c|cc}\lambda_1 & \multicolumn{2}{c}{\mathsf{u}_1^*\mathsf{A}\hat{\mathsf{U}}_1}\\ \hline 0 & \lambda_2 & \mathsf{u}_2^*\mathsf{A}_2\hat{\mathsf{U}}_2\\ \hline \mathbf{0} & \mathbf{0} & \mathsf{A}_3\end{array}\right)$$

Kao i ranije, zaključujemo da su svojstvene vrijednosti matrice A_3 upravo $\lambda_3,\ldots,\lambda_n$. Sada je jasno da formalni dokaz završavamo matematičkom indukcijom po n.

Nadalje, ako je matrica A realna i ako su joj sve svojstvene vrijednosti realne, onda je jasno da u prethodnom induktivnom dokazu u svakom koraku možemo odabrati realan svojstveni vektor i sve matrice transformacija realne ortogonalne. ⊞

Dakle, svaka kvadratna matrica A je unitarno slična trokutastoj matrici, $\mathsf{A} = \mathsf{U}\mathsf{T}\mathsf{U}^*$. U praksi se ova sličnost, baš kao i u dokazu teorema, pojavljuje kao transformacija $\mathsf{A}\mapsto\mathsf{T} = \mathsf{U}^*\mathsf{A}\mathsf{U}$, tj. zadanu matricu A se "napada" unitarnim transformacijama sličnosti sve dok ne postane trokutasta. Valja primijetiti da konstrukcija opisana u dokazu prethodnog teorema nije jako praktična jer direktno koristi svojstvene vrijednosti i vektore (zato je sama redukcija na trokutasti oblik gotova nakon $n-1$ koraka) koji u stvarnoj primjeni nisu dostupni i nije ih lako izračunati. Kako ćemo poslije vidjeti, numeričko računanje Schurove dekompozicije se svodi na beskonačan niz transformacija sličnosti koje sustavno i strpljivo reduciraju elemente ispod glavne dijagonale i osiguravaju trokutastu formu tek u limesu.

Komentar 1.2.1. Primijetimo da se iz Schurove forme za A jednostavno dobije Schurova forma za A^* i to na sljedeći način: Neka je π permutacija na $\{1,\ldots,n\}$, $\pi(i) = n-i+1$ i Π matrica te permutacije, $\Pi\mathbf{e}_i = \mathbf{e}_{\pi(i)}$. Tada je Schurova forma od A^* dana s $\mathsf{A}^* = (\mathsf{U}\Pi)(\Pi\mathsf{T}^*\Pi)\Pi\mathsf{U}^*$, gdje je $\mathsf{U}\Pi$ unitarna a $\Pi\mathsf{T}^*\Pi$ gornje trokutasta.

Korolar 1.2.2. *Trokutasta matrica* T *u Schurovoj dekompoziciji* $\mathsf{A} = \mathsf{UTU}^*$ *je dijagonalna ako i samo ako je matrica* A *normalna, što je također ekvivalentno sa* $\|\mathsf{A}\|_F^2 = \sum_{j=1}^n |\lambda_j(\mathsf{A})|^2$. *Specijalno vrijede sljedeći spektralni teoremi:*

- *Schurova forma hermitske matrice je realna dijagonalna matrica.*
- *Schurova forma anti–hermitske matrice je dijagonalna sa čisto imaginarnim dijagonalnim elementima.*
- *Schurova forma unitarne matrice je dijagonalna sa* $|\lambda_j| = 1$, $j = 1,\ldots,n$.

Dokaz: Neka je $\mathsf{AA}^* = \mathsf{A}^*\mathsf{A}$ i $\mathsf{A} = \mathsf{UTU}^*$. Lako se provjeri da $\mathsf{T} = \mathsf{U}^*\mathsf{AU}$ mora biti također normalna. Nadalje, ako je T gornje trokutasta i $\mathsf{TT}^* = \mathsf{T}^*\mathsf{T}$, onda se lako zaključi da je T nužno dijagonalna. Nadalje, primijetimo da je $\|\mathsf{A}\|_F = \|\mathsf{T}\|_F$ i $\mathsf{T}_{jj} = \lambda_j(\mathsf{A})$, te da je $\|\mathsf{T}\|_F^2 = \sum_{j=1}^n |\lambda_j(\mathsf{A})|^2$ ako i samo ako je T dijagonalna. Spektralni teoremi slijede iz već spomenute činjenice da Schurova forma, osim što je trokutasta, nasljeđuje svojstva hermitičnosti, anti–hermitičnosti i unitarnosti. ⊞

Kratnost svojstvenih vrijednosti je važno spektralno svojstvo matrice, sa cijelim nizom netrivijalnih posljedica. Pri tome su višestruke svojstvene vrijednosti višestruko problematične. Na primjer, svojstveni vektor jednostruke svojstvene vrijednosti je određen do na množenje netrivijalnim skalarom (pripadni svojstveni potprostor je jednodimenzionalan), dok npr. svojstvenoj vrijednosti algebarske kratnosti dva pripada jedan takav svojstveni vektor ako je njena geometrijska kratnost jedan, ili je svaki netrivijalni vektor iz dvodimenzionalnog potprostora svojstveni vektor (geometrijska kratnost te svojstvene vrijednosti je onda jednaka dva).

Važno je uočiti da je višestrukost svojstvene vrijednosti osjetljivo svojstvo – lako ga je izgubiti. Na primjer, uzmimo

$$\mathsf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \det(\lambda\mathsf{I}_2 - \mathsf{A}) = \lambda^2 - (a+b)\lambda + ab - cd$$

sa svojstvenim vrijednostima

$$\lambda_{1,2} = \frac{a+b \pm \sqrt{(a+b)^2 - 4(ab-cd)}}{2}.$$

Matrica A će imati dvostruku svojstvenu vrijednost $\lambda_1 = \lambda_2 = (a+b)/2$ ako i samo ako je $\Delta \equiv (a+b)^2 - 4(ab-cd) = 0$, tj. ako je diskriminanta svojstvenog polinoma jednaka nuli. Jasno je da se proizvoljno malim promjenama koeficijenata a, b, c, d diskriminantu Δ može iz trivijalne $\Delta = 0$ pretvoriti u netrivijalnu (različitu od nule) vrijednost. Sljedeći korolar Schurovog teorema pokazuje da su matrice sa višestrukim svojstvenim vrijednostima nigdje gust skup (interior zatvarača im je prazan) u $\mathbb{C}^{n\times n}$.

Korolar 1.2.3. *Matrice sa jednostrukim svojstvenim vrijednostima su gust podskup u $\mathbb{C}^{n\times n}$. U proizvoljnoj ε okolini svake matrice A se nalazi matrica $\tilde{\mathsf{A}}$ sa jednostrukim svojstvenim vrijednostima. Specijalno su dijagonalizabilne matrice gust dio od $\mathbb{C}^{n\times n}$. Pri tome, ako je A normalna, hermitska, anti–hermitska, unitarna, se matrica $\tilde{\mathsf{A}}$ može odabrati također, redom, normalna, hermitska, anti–hermitska, unitarna.*

Dokaz: Neka A ima s različitih svojstvenih vrijednosti $\lambda_1, \ldots, \lambda_s$, sa algebarskim kratnostima $\alpha(\lambda_1), \ldots, \alpha(\lambda_s)$, i neka je $\gamma = \min_{i\neq j} |\lambda_i - \lambda_j|$. Netrivijalan slučaj je $s < n$.

Neka je $\varepsilon > 0$ proizvoljan. U Schurovoj dekompoziciji $\mathsf{A} = \mathsf{U}\mathsf{T}\mathsf{U}^*$ svih $\alpha(\lambda_i)$ dijagonalnih elemenata za koje je $\mathsf{T}_{jj} = \lambda_i$ malim promjenama $\epsilon_j^{(i)}$, pri čemu je $|\epsilon_j^{(i)}| < \min\{\varepsilon/\sqrt{n}, \gamma/2\}$, možemo pretvoriti u $\alpha(\lambda_i)$ različitih elemenata $\tilde{\mathsf{T}}_{jj}$. Kada to napravimo za sve svojstvene vrijednosti λ_i, dobijemo n međusobno različitih vrijednosti $\tilde{\mathsf{T}}_{11}, \ldots, \tilde{\mathsf{T}}_{nn}$. Neka je $\tilde{\mathsf{T}}$ matrica dobivena od T zamjenom T_{ii} s $\tilde{\mathsf{T}}_{ii}$, $i = 1, \ldots, n$. Sigurno je $\|\mathsf{T} - \tilde{\mathsf{T}}\|_F < \varepsilon$. Ako definiramo $\tilde{\mathsf{A}} = \mathsf{U}\tilde{\mathsf{T}}\mathsf{U}^*$, onda $\tilde{A}$ ima n međusobno različitih svojstvenih vrijednosti i $\|\mathsf{A} - \tilde{\mathsf{A}}\|_F = \|\mathsf{T} - \tilde{\mathsf{T}}\|_F < \varepsilon$. Naravno, možemo konstruirati beskonačno mnogo matrica $\tilde{\mathsf{A}}$ koje zadovoljavaju ovu konstrukciju. Sada još uočimo da:

- Ako je A normalna, onda su T i $\tilde{\mathsf{T}}$ dijagonalne, pa je i $\tilde{\mathsf{A}}$ normalna.
- Ako je A hermitska (anti–hermitska), onda je T dijagonalna sa dijagonalnim elementima na realnoj (imaginarnoj) osi i opisana varijacija dijagonalnih elemenata se očito može provesti tako da $\tilde{\mathsf{T}}$ bude dijagonalna hermitska (anti–hermitska) i $\tilde{\mathsf{A}}$ hermitska (anti–hermitska).
- Ako je A unitarna, onda, zaključujući na isti način, vidimo da $\tilde{\mathsf{A}}$ može biti odabrana da bude unitarna.

$\boxplus$

1.2.1 Realna Schurova forma

Schurova dekompozicija je koristan i moćan alat u numeričkom računanju jer unitarnom sličnošću matricu prevodi u jednostavniji oblik (gornje trokutastu). Vrlo često u primjenama radimo sa realnim matricama koje općenito imaju kompleksne svojstvene vrijednosti, ali su i ulazni podaci i očekivani rezultati realni. Ako numerički algoritam zbog kompleksnog spektra u nekom momentu 'napusti' realno polje i počne se izvoditi nad $\mathbb{C}$ onda se povećava potrebni memorijski prostor u računalu (treba dvostruko više memorije za svaki kompleksni broj), povećava se broj računskih operacija (jedno kompleksno množenje sadrži[1] četiri realna množenja, te po jedno zbrajanje i oduzimanje), a zbog grešaka zaokruživanja će izračunato rješenje vjerojatno imati netrivijalnu imaginarnu komponentu.

Zbog toga je korisno imati realnu inačicu Schurove forme, baziranu na jednostavnoj ideji: Kako u kompleksne svojstvene vrijednosti realne matrice dolaze u parovima kompleksno–konjugiranih brojeva, onda svaki kompleksno–konjugirani par na dijagonali od T možemo isporučiti skriven kao spektar realne 2×2 matrice.

Teorem 1.2.4. *Neka je* $\mathsf{A} \in \mathbb{R}^{n \times n}$. *Neka* A *ima* r *realnih svojstvenih vrijednosti i* c *kompleksno konjugiranih parova. Tada postoji realna ortogonalna matrica* U *i blok gornje trokutasta matrica* T, *blok dimenzije* $(r+c) \times (r+c)$, *tako da je*

$$\mathsf{U}^T \mathsf{A} \mathsf{U} = \begin{pmatrix} \mathsf{T}_{[11]} & \mathsf{T}_{[12]} & \cdots & \mathsf{T}_{[1k]} & \cdots & \mathsf{T}_{[1,r+c]} \\ & \mathsf{T}_{[22]} & \mathsf{T}_{[23]} & \cdots & \cdots & \mathsf{T}_{[2,r+c]} \\ & & \ddots & \ddots & \cdots & \vdots \\ & & & \mathsf{T}_{[kk]} & \cdots & \mathsf{T}_{[k,r+c]} \\ & & & & \ddots & \vdots \\ & & & & & \mathsf{T}_{[r+c,r+c]} \end{pmatrix}. \tag{1.2.1}$$

Pri tome su dijagonalni blokovi $\mathsf{T}_{[kk]}$ *dimenzija* 1×1 *(takvih ima točno* r*) ili* 2×2 *(takvih je točno* c*). Trivijalni* 1×1 *blokovi su realne svojstvene vrijednosti od* A, *a svojstvene vrijednosti svakog* 2×2 *bloka su jedan par kompleksno konjugiranih svojstvenih vrijednosti od* A.

Dokaz: Slijedimo dokaz Teorema 1.2.1. Za svaku realnu svojstvenu vrijednost možemo ponoviti isti postupak – realnim svojstvenim vrijednostima pripadaju realni svojstveni vektori i redukcija se provodi pomoću ortogonalnih matrica. Kompleksno konjugirane parove svojstvenih vrijednosti ćemo obrađivati malo drugačije. Na primjer, neka u prvom koraku imamo kompleksno konjugirani par $\lambda_{1,2} = \alpha \pm \mathrm{i}\beta$

[1] Moguće je x mnozenja ??? referenca

svojstvenih vrijednosti. Ako je $\mathsf{A}(x+\mathrm{i}y) = (\alpha+\mathrm{i}\beta)(x+\mathrm{i}y)$, $x+\mathrm{i}y \neq \mathbf{0}$, onda konjugiranjem dobijemo da je $x-\mathrm{i}y$ svojstveni vektor koji pripada $\alpha-\mathrm{i}\beta$. Lako pokažemo da su realni vektori x i y linearno nezavisni, te da je $\mathsf{A}\begin{pmatrix} x & y \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix}\left(\begin{smallmatrix} \alpha & \beta \\ -\beta & \alpha \end{smallmatrix}\right)$. Spektar realne 2×2 matrice $\left(\begin{smallmatrix} \alpha & \beta \\ -\beta & \alpha \end{smallmatrix}\right)$ je kompleksno konjugirani par $\lambda_{1,2} = \alpha \pm \mathrm{i}\beta$, a x i y razapinju dvodimenzonalni invarijantni potprostor. U njemu QR faktorizacijom

$$\begin{pmatrix} x & y \end{pmatrix} = \begin{pmatrix} \mathsf{u}_1 & \mathsf{u}_2 \end{pmatrix} R,\ R = \begin{pmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{pmatrix}$$

izračunamo ortonormiranu bazu. Matricu $\begin{pmatrix} \mathsf{u}_1 & \mathsf{u}_2 \end{pmatrix}$ dopunimo do ortogonalne $\mathsf{U}_1 = \begin{pmatrix} \mathsf{u}_1 & \mathsf{u}_2 & \hat{\mathsf{U}}_1 \end{pmatrix}$ i izračunamo

$$\mathsf{U}_1^{\mathrm{T}}\mathsf{A}\mathsf{U}_1 = \begin{pmatrix} R\left(\begin{smallmatrix} \alpha & \beta \\ -\beta & \alpha \end{smallmatrix}\right)R^{-1} & \left(\begin{smallmatrix} \mathsf{u}_1 \\ \mathsf{u}_2 \end{smallmatrix}\right)\mathsf{A}\hat{\mathsf{U}}_1 \\ \mathbf{0} & \hat{\mathsf{U}}_1^{\mathrm{T}}\mathsf{A}\hat{\mathsf{U}}_1 \end{pmatrix}$$

Postupak nastavljamo s matricom $\hat{\mathsf{U}}_1^{\mathrm{T}}\mathsf{A}\hat{\mathsf{U}}_1$. Na ovaj način u svakom koraku obradimo ili jednu realnu svojstvenu vrijednosti ili kompleksno konjugurani par, što rezultira 1×1 ili 2×2 dijagonalnim blokom $\mathsf{T}_{[ii]}$ u (1.2.1). ⊞

Korolar 1.2.5. *Matrica $\mathsf{A} \in \mathbb{R}^{n\times n}$ je normalna ako i samo ako postoji realna ortogonalna matrica U i blok dijagonalna matrica T tako da je*

$$\mathsf{U}^T\mathsf{A}\mathsf{U} = \begin{pmatrix} \mathsf{T}_{[11]} & & \\ & \ddots & \\ & & \mathsf{T}_{[r+c,r+c]} \end{pmatrix}.$$

Pri tome su dijagonalni blokovi $\mathsf{T}_{[jj]}$ ili 1×1 (takvih ima r, po jedan za svaku realnu svojstvenu vrijednost) ili 2×2 oblika $\mathsf{T}_{[jj]} = \left(\begin{smallmatrix} \alpha & \beta \\ -\beta & \alpha \end{smallmatrix}\right)$ (po jedan za svaki od c kompleksno konjugiranih parova $\alpha\pm\mathrm{i}\beta$ svojstvenih vrijednosti od A). U specijalnim slučajevima normalnih matrica imamo

- A *je simetrična, $\mathsf{A} = \mathsf{A}^T$, ako i samo ako su svi dijagonalni blokovi 1×1, tj. $\mathsf{A} = \mathsf{U}\Lambda\mathsf{U}^T$, gdje $\Lambda = \operatorname{diag}(\lambda_i)_{i=1}^n$ sadrži svojstvene vrijednosti, a odgovarajući stupci od U su svojstveni vektori.*
- A *je antisimetrična, $\mathsf{A} = -\mathsf{A}^T$, ako i samo ako su svi 1×1 dijagonalni blokovi jednaki nuli, a svi 2×2 dijagonalni blokovi su oblika ...*

Korolar 1.2.6. *U proizvoljnoj ε okolini svake realne matrice A se nalazi realna matrica $\tilde{\mathsf{A}}$ sa jednostrukim svojstvenim vrijednostima. Pri tome, ako je A normalna, simetrična, anti–simetrična, ortogonalna, se matrica $\tilde{\mathsf{A}}$ može odabrati također, redom, normalna, simetrična, anti–simetrična, ortogonalna.*

1.3 Rezolventa

Svojstvene vrijednosti su, zajedno sa pripadnim vektorima, važne funkcije matrice.

Ako kompleksan broj ξ nije svojstvena vrijednost od A, onda je $\xi\mathsf{I}-\mathsf{A}$ regularna.

Definicija 1.3.1. Rezolventni skup matrice $\mathsf{A} \in \mathbb{C}^{n\times n}$ je skup $\varrho(\mathsf{A}) = \mathbb{C} \setminus \mathfrak{S}(\mathsf{A})$. Matrična (operatorska) funkcija $\mathrm{R}(\xi, \mathsf{A})$ koja je na $\varrho(\mathsf{A})$ definirana s

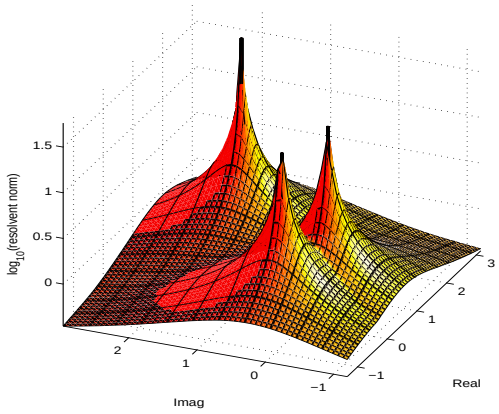
$$\mathrm{R}(\xi, \mathsf{A}) = (\xi\mathsf{I} - \mathsf{A})^{-1}, \quad \xi \in \varrho(\mathsf{A}),$$

se zove rezolventa of A.

Rezolventa je racionalna funkcija, tj. njeni alementi su racionalne funkcije od ξ, sa polovima u svojstvenim vrijednostima od A, što direktno slijedi iz relacije

$$\mathrm{R}(\xi, \mathsf{A}) = (\xi\mathsf{I} - \mathsf{A})^{-1} = \frac{\mathrm{adj}(\xi\mathsf{I} - \mathsf{A})}{\det(\xi\mathsf{I} - \mathsf{A})} = \frac{\mathrm{adj}(\xi\mathsf{I} - \mathsf{A})}{\chi_{\mathsf{A}}(\xi)}. \tag{1.3.1}$$

Primjer 1.3.1. Na Slici 1.2 je prikazana norma rezolvente slučajne 3×3 kompleksne matrice.



Slika 1.2:

Propozicija 1.3.1. *Rezolventa zadovoljava sljedeće dvije jednakosti:*

- *Za proizvoljne* $\xi_1, \xi_2 \in \varrho(\mathsf{A})$ *vrijedi*

$$\mathrm{R}(\xi_1, \mathsf{A}) - \mathrm{R}(\xi_2, \mathsf{A}) = -(\xi_1 - \xi_2)\mathrm{R}(\xi_1, \mathsf{A})\mathrm{R}(\xi_2, \mathsf{A}).$$

- *Ako su* $\mathsf{A}, \mathsf{B} \in \mathbb{C}^{n\times n}$ *i* $\xi \in \varrho(\mathsf{A}) \bigcap \varrho(\mathsf{B})$*, onda je*

$$\mathrm{R}(\xi, \mathsf{B}) - \mathrm{R}(\xi, \mathsf{A}) = \mathrm{R}(\xi, \mathsf{A})(\mathsf{B} - \mathsf{A})\mathrm{R}(\xi, \mathsf{B}).$$

Dokaz: Za dokaz prve jednakosti treba krenuti od $(\xi_1\mathsf{I} - \mathsf{A}) - (\xi_2\mathsf{I} - \mathsf{A}) = (\xi_1 - \xi_2)\mathsf{I}$ a za drugu treba uzeti $(\xi\mathsf{I} - \mathsf{A}) - (\xi\mathsf{I} - \mathsf{B}) = \mathsf{B} - \mathsf{A}$. ⊞

Iz definicije je jasno da je $\varrho(\mathsf{A})$ otvoren skup u kompleksnoj ravnini i da je za svaki $\xi \in \varrho(\mathsf{A})$ rezolventa dobro definirana u okolini točke ξ. Sljedeća lema opisuje kako možemo varirati i ξ i A.

Lema 1.3.2. *Neka je* $\xi \in \varrho(\mathsf{A})$*. Tada postoje* $\delta_\xi > 0$ *i* $\delta_\mathsf{A} > 0$ *takvi da je* $\tilde{\xi} \in \varrho(\tilde{\mathsf{A}})$ *za sve* $\tilde{\xi}$*,* $\tilde{\mathsf{A}}$ *za koje je* $|\tilde{\xi} - \xi| < \delta_\xi$ *i* $\|\tilde{\mathsf{A}} - \mathsf{A}\| < \delta_\mathsf{A}$*.*

Dokaz: Stavimo $\tilde{\xi}\mathsf{I} - \tilde{\mathsf{A}} = \xi\mathsf{I} - \mathsf{A} + (\tilde{\xi} - \xi)\mathsf{I} - (\tilde{\mathsf{A}} - \mathsf{A})$. Kako $\mathrm{R}(\xi, \mathsf{A}) = (\xi\mathsf{I} - \mathsf{A})^{-1}$ postoji po pretpostavci, osiguranje dovoljnih uvjeta za postojanje $(\tilde{\xi}\mathsf{I} - \tilde{\mathsf{A}})^{-1}$ možemo dobiti koristeći perturbacijske ocjene za inverz. Prema (**??**), dovoljno je da je $\|\mathrm{R}(\xi, \mathsf{A})((\tilde{\xi} - \xi)\mathsf{I} - (\tilde{\mathsf{A}} - \mathsf{A}))\| < 1$ a to je osigurano ako je

$$|\tilde{\xi} - \xi|\|\mathsf{I}\| + \|\tilde{\mathsf{A}} - \mathsf{A}\| < \frac{1}{\|\mathrm{R}(\xi, \mathsf{A})\|}.$$

Svaki par pozitivnih brojeva δ_ξ, δ_A sa svojstvom $\delta_\xi\|\mathsf{I}\| + \delta_\mathsf{A} < 1/\|\mathrm{R}(\xi, \mathsf{A})\|$ ima traženo svojstvo. ⊞

Propozicija 1.3.3. *Neka je* $\mathcal{O}$ *otvorena okolina spektra* $\mathfrak{S}(\mathsf{A})$ *matrice* A*. Tada je* $\sup_{\xi\in\mathbb{C}\setminus\mathcal{O}} \|(\xi\mathsf{I} - \mathsf{A})^{-1}\| < \infty$*.*

Dokaz: Uzmimo $\mathcal{O}' \subseteq \mathcal{O}$ otvorenu i ograničenu okolinu (ako je $\mathcal{O}$ ograničen, onda $\mathcal{O}' = \mathcal{O}$) i neka je $r > 0$ radijus zatvorenog kruga koji sadrži $\mathcal{O}'$. Neka je $\mathcal{D}$ zatvoren krug radijusa $r_\mathcal{D} = \max\{r, 2\|A\|\}$. Uočimo da je za dovoljno veliki $|\xi|$

$$(\xi\mathsf{I} - \mathsf{A})^{-1} = \frac{1}{\xi}\sum_{k=0}^{\infty}\frac{\mathsf{A}^k}{\xi^k}, \quad \|(\xi\mathsf{I} - \mathsf{A})^{-1}\| \le \frac{1}{|\xi| - \|\mathsf{A}\|}.$$

Dakle, $\sup\limits_{|\xi|>r_\mathcal{D}} \|(\xi\mathsf{I} - \mathsf{A})^{-1}\| < 1/\|\mathsf{A}\|$, dok je, zbog svojstava neprekidne funkcije na kompaktnom skupu, $\max\limits_{\xi\in\mathcal{D}\setminus\mathcal{O}'} \|(\xi\mathsf{I} - \mathsf{A})^{-1}\| < \infty$. ⊞

1.4 Neprekidnost svojstvenih vrijednosti

Sada ćemo svojstvene vrijednosti matrice promotriti kao funkcije matrice (njenih elemenata) i pokazati da su te funkcije neprekidne. Najjednostavniji dokaz te tvrdnje je preko karakterizacije svojstvenih vrijednosti kao nultočaka pripadnog karakterističnog polinoma i poznate činjenice da su nultočke polinoma neprekidne funkcije njegovih koeficijenata. Kako je jasno da su koeficijenti svojstvenog polinoma $\kappa_{\mathsf{A}}(\zeta)$ neprekidne funkcije elemenata od A, odmah slijedi, kompozicijom neprekidnih preslikavanja, da su svojstvene vrijednosti od A neprekidne funkcije od A. Sada ćemo to precizno formulirati i dokazati.

U prostoru $\mathbb{C}^{n\times n}$ ćemo koristiti proizvoljnu matričnu normu $\|\cdot\|$. Sve međusobno različite svojstvene vrijednosti od $\mathsf{A}\in\mathbb{C}^{n\times n}$ ćemo označiti s $\lambda_1,\ldots,\lambda_\ell$, pri čemu je α_j algebarska kratnost od λ_j.

Prvi dokaz neprekidnosti spektra ćemo dati koristeći svojstva rezolvente.

Teorem 1.4.1. *Neka je $\mathcal{O}$ otvorena okolina spektra $\mathfrak{S}(\mathsf{A})\subset\mathbb{C}$. Tada postoji $\delta>0$ takav da za svaku matricu $\tilde{\mathsf{A}}$, $\|\tilde{\mathsf{A}}-\mathsf{A}\|<\delta$ povlači da je $\mathfrak{S}(\tilde{\mathsf{A}})\subset\mathcal{O}$.*

Dokaz: Dovoljno je uzeti da je $\mathcal{O}$ unija otvorenih krugova sa centrima u svojstvenim vrijednostima od A. Iz $\mathfrak{S}(\mathsf{A})\subset\mathcal{O}$ slijedi da je $\mathbb{C}\setminus\mathcal{O}\subset\varrho(\mathsf{A})$. Dovoljno je pokazati da je $\mathbb{C}\setminus\mathcal{O}\subset\varrho(\tilde{\mathsf{A}})$ za sve $\tilde{\mathsf{A}}$ iz neke δ–okoline od A. Slično kao u Lemi **??**, za $\xi\in\mathbb{C}\setminus\mathcal{O}$ možemo staviti

$$\xi\mathsf{I}-\tilde{\mathsf{A}}=\xi\mathsf{I}-\mathsf{A}-(\tilde{\mathsf{A}}-\mathsf{A})=(\xi\mathsf{I}-\mathsf{A})(\mathsf{I}-(\xi\mathsf{I}-\mathsf{A})^{-1}(\tilde{\mathsf{A}}-\mathsf{A}))$$

pa je za $\xi\in\varrho(\tilde{\mathsf{A}})$ dovoljno da je $\|(\xi\mathsf{I}-\mathsf{A})^{-1}\|\|\tilde{\mathsf{A}}-\mathsf{A}\|<1$.

Kako je na zatvorenom skupu $\mathbb{C}\setminus\mathcal{O}$ rezolventa od A neprekidna, sa vrijednosti nula u beskonačnosti, supremum $\omega=\sup_{\xi\in\mathbb{C}\setminus\mathcal{O}}\|(\xi\mathsf{I}-\mathsf{A})^{-1}\|$ je konačan. Zaključujemo da je $\xi\in\varrho(\tilde{\mathsf{A}})$ čim je $\|\tilde{\mathsf{A}}-\mathsf{A}\|<\delta$, gdje je $\delta\equiv 1/\omega>0$. Dakle, $\|\tilde{\mathsf{A}}-\mathsf{A}\|<\delta$ povlači $\mathbb{C}\setminus\mathcal{O}\subset\varrho(\tilde{\mathsf{A}})$, pa je $\mathfrak{S}(\tilde{\mathsf{A}})\subset\mathcal{O}$. $\boxplus$

U Teoremu 1.4.1 možemo za okolinu $\mathcal{O}$ uzeti uniju $\mathcal{O}_\varepsilon$ diskova s centrima u svojstvenim vrijednostima $\lambda_1,\ldots,\lambda_\ell$, i radijusom ε. Kada ε smanjujemo ($\varepsilon\to 0$), ti se diskovi sažimaju i postepeno se počinju razdvajati dok u jednom momentu, za dovoljno mali ε ne postanu ℓ međusobno disjunktnih diskova s centrima u $\lambda_1,\ldots,\lambda_\ell$. Prema Teoremu 1.4.1, postoje odgovarajuće $\delta(\varepsilon)$–okoline od A tako da je spektar svake $\tilde{\mathsf{A}}$ iz te okoline sadržan u $\mathcal{O}_\varepsilon$. Sada tu tvrdnju želimo još precizirati. Ako α–struku svojstvenu vrijednost "stavimo pod povećalo" i matricu A malom perturbacijom $\delta\mathsf{A}$ pomjerimo u $\tilde{\mathsf{A}}=\mathsf{A}+\delta\mathsf{A}$, želimo vidjeti što se dešava sa α kopija svojstvene vrijednosti polazne, neperturbirane, matrice.

Kako su svojstvene vrijednosti matrice nultočke pripadnog svojstvenog polinoma $\kappa_{\mathsf{A}}(\zeta) = \det(\zeta\mathsf{I} - \mathsf{A})$, promatrat ćemo nultočke funkcije

$$\kappa_* : \mathbb{C} \times \mathbb{C}^{n\times n} \longrightarrow \mathbb{C}, \;\; \kappa_*(\zeta, \mathsf{X}) = \det(\zeta\mathsf{I} - \mathsf{X})$$

u okolini točaka (λ_j, A), $j = 1, \ldots, \ell$.

Propozicija 1.4.2. *Neka je $f : \mathcal{U} \longrightarrow \mathbb{C}$ holomorfna na otvorenom skupu $\mathcal{U} \subseteq \mathbb{C}$ i neka je $\overline{D}(z, r) \subset \mathcal{U}$. Pretpostavimo da u otvorenom disku $D(z, r)$ f ima nultočke $z_1, \ldots, z_\ell$, sa kratnostima redom $n_1, \ldots, n_\ell$. Ako je f različita od nule na ∂D, onda je*

$$\frac{1}{2\pi\mathfrak{i}} \oint_{|\zeta - z| = r} \frac{f'(\zeta)}{f(\zeta)} d\zeta = \sum_{j=1}^{\ell} n_j.$$

Lema 1.4.3. *Odaberimo proizvoljnu svojstvenu vrijednost λ_j matrice A, kratnosti α_j, i definirajmo $\gamma_j = \min_{k\neq j} |\lambda_j - \lambda_k|$. (Primijetimo da γ_j mjeri udaljenost od λ_j do najbližeg susjeda u spektru od A.) Tada za svaki $\varepsilon \in (0, \gamma_j)$ postoji $\delta_j(\varepsilon) > 0$ takav da, za svaku matricu $\mathsf{X} \in \mathbb{C}^{n\times n}$, $\|\mathsf{X} - \mathsf{A}\| < \delta_j(\varepsilon)$ povlači da X u disku $\mathfrak{D}_j(\varepsilon) = \{z \in \mathbb{C} \;:\; |z - \lambda_j| < \varepsilon\}$ ima točno α_j svojstvenih vrijednosti.*

Dokaz: Oko λ_j nacrtajmo kružnicu $K_j = \{\zeta \in \mathbb{C} \;:\; |\zeta - \lambda_j| = \varepsilon\}$, radijusa $\varepsilon < \gamma_j$ i proučimo djelovanje $\kappa_*(\cdot, \cdot)$ na $K_j \times \{\mathsf{A}\}$. Kako K_j sigurno ne sadrži niti jednu svojstvenu vrijednost od A, $\kappa_*(\zeta, \mathsf{A}) \neq 0$ za svaku točku $\zeta \in K_j$. Tada je, zbog neprekidnosti, κ_* različita od nule na cijeloj otvorenoj okolini

$$\mathfrak{C}_\zeta = \{(z, X) \in \mathbb{C} \times \mathbb{C}^{n\times n} \;:\; |z - \zeta| < r_\zeta, \;\; \|\mathsf{A} - \mathsf{X}\| < r_{\mathsf{A},\zeta}\}.$$

Očito je $K_j \times \{\mathsf{A}\} \subset \bigcup_{\zeta\in K_j} \mathfrak{C}_\zeta$, pa zbog kompaktnosti možemo $K_j \times \{\mathsf{A}\}$ prekriti s konačno mnogo takvih okolina $\mathfrak{C}_{\zeta_1}, \ldots, \mathfrak{C}_{\zeta_p}$. Ako od svih pripadnih kugala oko A uzmemo najmanji radijus, $r_0 = \min\limits_{i=1:p} r_{\mathsf{A},\zeta_i}$, onda je za svaki $\zeta \in K_j$ i svaku matricu X koja zadovoljava $\|\mathsf{X} - \mathsf{A}\| < r_0$ zadovoljeno $\kappa_*(\zeta, \mathsf{X}) \equiv \kappa_{\mathsf{X}}(\zeta) \neq 0$. Dakle, za svaku takvu X je prema Propoziciji 1.4.2 broj nultočki od $\kappa_{\mathsf{X}}(\cdot)$ unutar kruga K_j jednak[2]

$$\xi_j(\mathsf{X}) = \frac{1}{2\pi\mathfrak{i}} \oint_{K_j} \frac{\kappa_*'(\zeta, \mathsf{X})}{\kappa_*(\zeta, \mathsf{X})} d\zeta. \tag{1.4.1}$$

Podintegralna funkcija u (1.4.1) je neprekidna funkcija za sve X, $\|\mathsf{X} - \mathsf{A}\| < r_0$, pa je i $\xi_j(\mathsf{X})$ neprekidna funkcija $\mathsf{X} \ni \mathbb{C}^{n\times n} \mapsto \xi_j(\mathsf{X}) \in \mathbb{N}_0$. Odavde slijedi da

[2]Vidi Teorem **??** u §**??**.

$\xi_j(\cdot)$ mora biti konstanta na cijeloj r_0–okolini od A. Kako je λ_j jedina svojstvena vrijednost od A unutar kružnice K_j, mora biti $\xi_j(\mathsf{A}) = \alpha_j$, pa je tvrdnja dokazana s $\delta_j(\varepsilon) = r_0$. $\boxplus$

Lema 1.4.3 nam govori da se određenom promjenom matrice A njena α_j–struka svojstvena vrijednost razbije na nekoliko svojstvenih vrijednosti čiji je zbroj algebarskih kratnosti točno α_j, i koje ostaju u okolini određenoj s udaljenosti γ_j od λ_j do ostalih vrijednosti u $\mathfrak{S}(\mathsf{A})$. Ako lemu primijenimo istovremeno na cijeli spektar $\mathfrak{S}(\mathsf{A})$, pri čemu ε odaberemo tako da zadani diskovi $\mathfrak{D}_j(\varepsilon)$ budu disjunktni, dobijemo sljedeći teorem:

Teorem 1.4.4. *Neka su $\lambda_1, \ldots, \lambda_\ell$ sve međusobno različite svojstvene vrijednosti od A, i neka je α_j algebarska kratnost od λ_j. Tada za svaki $\varepsilon > 0$ postoji $\delta > 0$ takav da, za svaku matricu $\mathsf{X} \in \mathbb{C}^{n\times n}$, $\|\mathsf{X}-\mathsf{A}\| < \delta$ povlači da su sve svojstvene vrijednosti od X sadržane u uniji ε–krugova s centrima u λ_j, $\bigcup_{j=1}^{\ell}\{z \in \mathbb{C} \ : \ |z - \lambda_j| < \varepsilon\}$. Pri tome, ako je $\varepsilon < \frac{1}{2}\min_{i\neq k}|\lambda_i - \lambda_k|$, onda svaki ε–krug oko λ_j sadrži točno α_j svojstvenih vrijednosti od X.*

Tvrdnju prethodnog teorema možemo iskazati i na uobičajeni formalni način. Kako svaka matrica A ima n (općenito kompleksnih) svojstvenih vrijednosti računato s kratnostima, zgodno je tih n brojeva promatrati kao neuređenu n–torku tj. sa proizvoljnim poretkom svojstvenih vrijednosti. Reći ćemo da su dvije takve neuređene n–torke ekvivalentne ako se jedna može dobiti iz druge promjenom poretka elemenata, tj. primjenom permutacije $\sigma \in \mathbf{S}_n$. Lako se provjeri da je time definirana relacija ekvivalencije na skupu $\mathbb{C}^{1\times n}$. Klasu ekvivalencije n–torke $\lambda = (\lambda_1, \ldots, \lambda_n)$ ćemo označiti s $[\lambda]$, a cijeli kvocijentni prostor s $\mathbb{C}^{1\times n}|_{\mathbf{S}_n}$. Ako $[\lambda]$ označava svojstvene vrijednosti od A, onda koristimo i oznaku $[\lambda(\mathsf{A})]$. Udaljenost dvije klase ekvivalencije definiramo pomoću funkcije optimalnog sparivanja na sljedeći način:

Definicija 1.4.1. Za $[\lambda], [\tilde{\lambda}] \in \mathbb{C}^{1\times n}|_{\mathbf{S}_n}$ definiramo funkciju udaljenosti s

$$\varpi_n([\lambda], [\tilde{\lambda}]) = \min_{\sigma\in\mathbf{S_n}} \max_{i=1:n} |\lambda_i - \tilde{\lambda}_{\sigma(i)}|. \tag{1.4.2}$$

Lako se uvjerimo da je $\varpi_n(\cdot,\cdot)$ metrika na $\mathbb{C}^{1\times n}|_{\mathbf{S}_n}$. Uz ove oznake, tvrdnju o neprekidnosti spektra 'u točki' A možemo formalno zapisati kao

$$(\forall\varepsilon > 0)\ \ (\exists\delta(\varepsilon) > 0)\ \ (\forall \mathsf{X} \in \mathbb{C}^{n\times n})$$
$$(\|\mathsf{X} - \mathsf{A}\| < \delta(\varepsilon) \Longrightarrow \varpi_n([\lambda(\mathsf{X})], [\lambda(\mathsf{A})]) < \varepsilon).$$

Ako imamo matričnu funkciju koja neprekidno ovisi o kompleksnim parametrima, onda vrijedi

Korolar 1.4.5. *Neka je* $\Omega \subseteq \mathbb{C}$ *i* $\mathsf{A}(\cdot) : \Omega \to \mathbb{C}^{n\times n}$ *neprekidna funkcija,* $z \mapsto \mathsf{A}(z)$. *Tada je preslikavanje*

$$\mathbb{C} \supseteq \Omega \ni z \mapsto [\lambda(\mathsf{A}(z))] \in \mathbb{C}^{1\times n}|_{\mathbf{s}_n} \tag{1.4.3}$$

neprekidno na Ω. *Tvrdnja vrijedi i za* $\mathsf{A}(\cdot) : \Omega \subseteq \mathbb{C}^k \to \mathbb{C}^{n\times n}$, *tj. za neprekidnu funkciju od* $k \geq 1$ *varijabli.*

U vezi s neprekidnom funkcijom (1.4.3) se postavlja sljedeće prirodno pitanje:
Da li je moguće definirati n neprekidnih funkcija $\lambda_1(z), \ldots, \lambda_n(z)$ *na* Ω *tako da, za svaki* $z \in \Omega$, *n–torka* $(\lambda_1(z), \ldots, \lambda_n(z))$ *čini spektar od* $\mathsf{A}(z)$*?*
Ako takva n–torka postoji onda je zovemo neprekidnom realizacijom spektra od (1.4.3).

Primjer 1.4.1. Neka je $\Omega \subset \mathbb{C}$ otvorena okolina nule, npr. $\Omega = \{z \in \mathbb{C} \,:\, |z| < 2\}$ i

$$\mathsf{A}(z) = \begin{pmatrix} 0 & 1 \\ z & 0 \end{pmatrix}, \quad z \in \Omega.$$

Za svaki z je spektar od $\mathsf{A}(z)$ jednak $(\sqrt{z}, -\sqrt{z})$. Uz gornje oznake je

$$[\lambda(\mathsf{A}(z))] = \{(\sqrt{z}, -\sqrt{z}), (-\sqrt{z}, \sqrt{z})\}, \quad z \in \Omega.$$

Stavimo $\lambda_1(z) = \sqrt{z}$, $\lambda_2(z) = -\sqrt{z}$, tj. $\lambda_1(\cdot)$ i $\lambda_2(\cdot)$ su dvije grane kompleksnog kvadratnog korjena i $(\lambda_1(z), \lambda_2(z))$ je traženi par funkcija. Ostaje samo provjeriti neprekidnost.
Ponašanje korjena je najlakše opisati u polarnim koordinatama. Odaberimo $(-\pi, \pi]$ za glavnu vrijednost argumenta. Ako stavimo $z = r\mathbf{e}^{\mathbf{i}\phi} = r(\cos\phi + \mathbf{i}\sin\phi)$, $\phi \in (-\pi, \pi]$, onda je npr. $\lambda_1(z) = \sqrt{r}(\cos\frac{\phi}{2} + \mathbf{i}\sin\frac{\phi}{2})$. Sada odaberimo točku $z_0 \in \Omega$ oblika $z_0 = r_0\mathbf{e}^{\mathbf{i}\pi}$ i njoj dvije bliske točke $z_+ = r_0\mathbf{e}^{\mathbf{i}\phi_+}$, $z_- = r_0\mathbf{e}^{\mathbf{i}\phi_-}$, gdje je $\phi_+ = \pi - \delta$, $\phi_- = -\pi + \delta$ s proizvoljno malim $\delta > 0$. Uočimo da je $\lambda_1(z_0) = \sqrt{r_0}\mathbf{i}$, $\lambda_1(z_+) \approx \sqrt{r_0}\mathbf{i}$, ali $\lambda_1(z_-) \approx -\sqrt{r_0}\mathbf{i}$. Dakle sve točke iz presjeka Ω s negativnom realnom osi su točke prekida funkcije $\lambda_1(z)$. Isto vrijedi i za $\lambda_2(z)$.

Komentar 1.4.1. Valja uočiti ....

Primjer 1.4.1 pokazuje da općenito nije moguće realizirati spektar neprekidne matrične funkcije $\mathsf{A}(z) : \Omega \to \mathbb{C}^{n\times n}$ pomoću n neprekidnih funkcija. Vidjeli smo da je u ovom primjeru problem nastao zbog prekida kompleksnog korjena. Sljedeća propozicija pokazuje da u slučaju realnog spektra ($\mathfrak{S}(\mathsf{A}(z)) \subset \mathbb{R}$, $z \in \Omega$) nema prepreke neprekidnoj realizaciji.

Propozicija 1.4.6. *Neka je* $\Omega \subseteq \mathbb{C}$ *i* $\mathsf{A}(\cdot) : \Omega \to \mathbb{C}^{n\times n}$ *neprekidna funkcija,* $z \mapsto \mathsf{A}(z)$*, te neka su svojstvene vrijednosti od* $\mathsf{A}(z)$ *realne za sve* $z \in \Omega$*. (Na primjer, ako su sve matrice* $\mathsf{A}(z)$ *hermitske, onda spektar ostaje realan.) Tada postoji* n *neprekidnih funkcija* $\lambda_1(z), \ldots, \lambda_n(z)$ *na* Ω *tako da je, za svaki* $z \in \Omega$*,* $(\lambda_1(z), \ldots, \lambda_n(z))$ *spektar od* $\mathsf{A}(z)$*.*

Dokaz: Za svaki $z \in \Omega$, svojstvene vrijednosti od $\mathsf{A}(z)$ uredimo u rastućem nizu i definirajmo $\lambda_i(z)$ kao i–tu svojstvenu vrijednost od $\mathsf{A}(z)$. Za svaki z su dakle svojstvene vrijednosti od $\mathsf{A}(z)$ dane s $\lambda_1(z) \leq \cdots \leq \lambda_n(z)$. Ostaje pokazati da su sve funkcije $\lambda_i(z)$ neprekidne. Odaberimo proizvoljni $z_0 \in \Omega$ i z iz male okoline od z_0. Sada uočimo da je

$$\varpi_n([\lambda(\mathsf{A}(z))], [\lambda(\mathsf{A}(z_0))]) = \max_{i=1:n} |\lambda_i(z) - \lambda_i(z_0)|, \tag{1.4.4}$$

pa neprekidnost svih funkcija $\lambda_i(\cdot)$ slijedi primjenom Korolara 1.4.5. Iz dokaza je jasno da tvrdnja vrijedi i za $\Omega \subseteq \mathbb{C}^k$, $k > 1$. $\boxplus$

Komentar 1.4.2. Za dokazati relaciju (1.4.4) je dovoljno uočiti da se za n–torke $\alpha_1 \leq \cdots \leq \alpha_n$ i $\beta_1 \leq \cdots \leq \beta_n$ vrijednost $\max_{i=1:n} |\alpha_i - \beta_i|$ ne može smanjiti ako elementarnom permutacijom zamijenimo pozicije β_k i β_j.

Sljedeća sekcija opisuje jednu drugu važnu situaciju u kojoj je također moguća neprekidna realizacija.

1.4.1 Neprekidnost spektra duž neprekidnog puta

Sada promatramo neprekidnu matričnu funkciju definiranu na realnom intervalu. Pokazuje se da ovakva restrikcija domene osigurava neprekidnu realizaciju spektra. Preciznije, ako je $\mathsf{A} : \mathcal{I} \subseteq \mathbb{R} \to \mathbb{C}^{n\times n}$, $t \mapsto \mathsf{A}(t)$, i ako $[\lambda(t)]$ označava spektar od $\mathsf{A}(t)$, onda neprekidna realizacija slijedi iz sljedećeg teorema.

Teorem 1.4.7. *Neka je* $\mathcal{I} \subseteq \mathbb{R}$ *interval i neka je* $[\lambda(\cdot)] : \mathcal{I} \to \mathbb{C}^{1\times n}|_{\mathbf{S}_n}$ *neprekidna funkcija,* $t \mapsto [\lambda(t)]$*. Tada postoji* n *neprekidnih kompleksnih funkcija na* $\mathcal{I}$*,* $\lambda_1(\cdot), \ldots, \lambda_n(\cdot) : \mathcal{I} \to \mathbb{C}$*, sa svojstvom da je za svaki* $t \in \mathcal{I}$*,* $(\lambda_1(t), \ldots, \lambda_n(t)) \in [\lambda(t)]$*.*

Dokaz: Tvrdnja je trivijalna ako je $n = 1$. Tada se svaka klasa $[\lambda(t)]$ sastoji od samo jednog elementa $\lambda_1(t)$ koji definira traženu neprekidnu funkciju, $t \mapsto \lambda_1(t)$. Neka je sada $n > 1$. Dokaz će biti baziran na matematičkoj indukciji –

n dimenzionalni slučaj će biti dokazan uz pretpostavku da tvrdnja vrijedi za sve dimenzije strogo manje od n. Radi jasnoće ćemo neke elemente koraka indukcije ilustrirati za $n = 2$.

Neka je $\mathcal{J} \subseteq \mathcal{I}$ skup svih $t \in \mathcal{I}$ sa svojstvom da se klasa $[\lambda(t)]$ sastoji od samo jedne n–torke istih vrijednost, $[\lambda(t)] = \{(\lambda_*(t), \lambda_*(t), \ldots, \lambda_*(t))\}$. Ako je $\mathcal{J} = \mathcal{I}$, onda je problem riješen kao trivijalni slučaj: $\lambda_*(t)$ je neprekidna na $\mathcal{I}$ i $\lambda_1(t) = \cdots = \lambda_n(t) \equiv \lambda_*(t)$, $t \in \mathcal{I}$, su tražene funkcije.

Inače je $\mathcal{K} = \mathcal{I} \setminus \mathcal{J}$ neprazan. Odaberimo proizvoljan $t_0 \in \mathcal{K}$. U pripadnoj klasi $[\lambda(t_0)]$ možemo odabrati predstavnika

$$(\lambda_1(t_0), \ldots, \lambda_{n_1}(t_0), \lambda_{n_1+1}(t_0), \ldots, \lambda_n(t_0))$$

sa svojstvom da se niti jedna od n_1 vrijednosti n_1–torke $\hat{\lambda}(t_0) = (\lambda_1(t_0), \ldots, \lambda_{n_1}(t_0))$ ne nalazi među preostalih $n_2 = n - n_1$ vrijednosti n_2–torke $\breve{\lambda}(t_0) = (\lambda_{n_1+1}(t_0), \ldots, \lambda_n(t_0))$. (Budući svaki predstavnik klase ima barem dvije međusobno različite vrijednosti, dovoljno je sve kopije jedne od njih uzeti kao prvih n_1 elemenata. U slučaju $n = 2$ jednostavno imamo $\lambda_1(t_0) \neq \lambda_2(t_0)$.) Kako je $[\lambda(\cdot)]$ neprekidna, onda možemo odrediti $\delta > 0$ takav da se, za sve $t \in (t_0 - \delta, t_0 + \delta) \bigcap \mathcal{I}$, $[\lambda(t)]$ može na isti način kao i $[\lambda(t_0)]$ reprezentirati n–torkom koja se sastoji od dva disjunktna dijela $\hat{\lambda}(t) = (\lambda_1(t), \ldots, \lambda_{n_1}(t))$, $\breve{\lambda}(t) = (\lambda_{n_1+1}(t), \ldots, \lambda_n(t))$. (U slučaju $n = 2$, imamo da je $\lambda_1(t) \neq \lambda_2(t)$ za sve t iz navedene okoline.) Znači da je $\mathcal{K}$ otvoren u $\mathcal{I}$ (u smislu relativne topologije na $\mathcal{I}$).

Nadalje, koristeći ponovo neprekidnost, znamo da δ možemo odabrati dovoljno mali da vrijedi i

$$\varpi_n([\lambda(t)], [\lambda(t_0)]) = \max\{\varpi_{n_1}([\hat{\lambda}(t)], [\hat{\lambda}(t_0)]), \varpi_{n_2}([\breve{\lambda}(t)], [\breve{\lambda}(t_0)])\}$$

na cijelom $\mathcal{O}_{t_0} = (t_0 - \delta, t_0 + \delta) \bigcap \mathcal{I}$. To znači da su $[\hat{\lambda}(t)]$ i $[\breve{\lambda}(t)]$ neprekidne na $\mathcal{O}_{t_0}$. Po pretpostavci indukcije (jer su $n_1, n_2 < n$) za njih na $\mathcal{O}_{t_0}$ postoje neprekidne realizacije $(\hat{\lambda}_1^*(t), \ldots, \hat{\lambda}_{n_1}^*(t))$ i $(\breve{\lambda}_{n_1+1}^*(t), \ldots, \breve{\lambda}_n^*(t))$ koje, uzete zajedno kao n neprekidnih funkcija, za svaki $t \in \mathcal{O}_{t_0}$ reprezentiraju klasu $[\lambda(t)]$. Dakle, problem je riješen na $\mathcal{O}_{t_0}$.

Sada uočimo da neprekidne realizacije $(\lambda_1^{(i)}(t), \ldots, \lambda_n^{(i)}(t))$, $i = 0, 1$, na $\mathcal{O}_{t_0}$ i $\mathcal{O}_{t_1}$ sa $\mathcal{O}_{t_0} \bigcap \mathcal{O}_{t_1} \neq \emptyset$ možemo "zalijepiti zajedno" i dobiti neprekidnu realizaciju na $\mathcal{O}_{t_0} \bigcup \mathcal{O}_{t_1}$. Neka je $\mathcal{O}_{t_0}$ lijevo od $\mathcal{O}_{t_1}$ (tj. $\inf(\mathcal{O}_{t_0}) \leq \inf(\mathcal{O}_{t_1}) < \sup(\mathcal{O}_{t_0}) \leq \sup(\mathcal{O}_{t_1})$) i neka je $t_\bullet \in \mathcal{O}_{t_0} \bigcap \mathcal{O}_{t_1}$ proizvoljna točka iz presjeka. Kako obadvije realizacije u točki $t_\bullet$ daju predstavnika klase $[\lambda(t_\bullet)]$, onda je, za neku permutaciju $\sigma \in \mathbf{S}_n$, $\lambda_j^{(0)}(t_\bullet) = \lambda_{\sigma(j)}^{(1)}(t_\bullet)$. Lako se vidi da je jedna tražena neprekidna realizacija

na $\mathcal{O}_{t_0} \bigcup \mathcal{O}_{t_1}$ dana s

$$\mathcal{O}_{t_0} \bigcup \mathcal{O}_{t_1} \ni t \mapsto \lambda_j^{(01)}(t) = \begin{cases} \lambda_j^{(0)}(t), & \text{za } t \leq t_\bullet \\ \lambda_{\sigma(j)}^{(1)}(t) & \text{za } t > t_\bullet \end{cases} \tag{1.4.5}$$

Skup $\mathcal{K}$, kao otvoren skup u $\mathcal{I}$, je unija od najviše prebrojivo intervala $\mathcal{K}_k$, $k = 1, 2, \ldots$, pri čemu na svakom $\mathcal{K}_k$ upravo opisanom metodom ljepljenja (1.4.5) možemo konstruirati neprekidnu realizaciju $(\lambda_1^{(k)}(t), \ldots, \lambda_n^{(k)}(t))$.

Tražena realizacija je

$$\mathcal{I} \ni t \mapsto \lambda_j^*(t) = \begin{cases} \lambda_j^{(k)}(t), & \text{za } t \in \mathcal{K}_k \\ \lambda_*(t), & \text{za } t \in \mathcal{J},\ [\lambda(t)] = \{(\lambda_*(t), \ldots, \lambda_*(t))\} \end{cases} \tag{1.4.6}$$

⊞

1.5 Spektralni radijus

Svojstvene vrijednosti su važna funkcija kvadratnih matrica i često su upravo svojstvene vrijednosti ključ za mnoge zaključke i u teorijskim razmatranjima i u praktičnim numeričkima algoritmima. Na primjer, važno je znati koliko najviše mogu biti velike apsolutne vrijednosti svojstvenih vrijednosti zadane matrice.

Definicija 1.5.1. Spektralni radijus $\mathrm{spr}(\mathsf{A})$ matrice $\mathsf{A} \in \mathbb{M}_n$ je definiran s $\mathrm{spr}(\mathsf{A}) = \max_{\lambda \in \mathfrak{S}(A)} |\lambda|$.

Iz definicije je jasno da je spektralni radijus neprekidna funckija matrice. Spektralni radijus na neki način mjeri veličinu matrice tako da joj izmjeri spektar. Ako na primjer sve svojstvene vrijednosti složimo u vektor $\overrightarrow{\lambda} = (\lambda_1, \ldots, \lambda_n)^{\mathrm{T}}$, onda je $\mathrm{spr}(\mathsf{A}) = \|\overrightarrow{\lambda}\|_\infty$. Lako je pokazati da $\mathrm{spr}(\cdot)$ nije norma na $\mathbb{M}_n$. Na primjer, ako uzmemo $A = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, $B = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)$, vidimo da je $\mathrm{spr}(A) = \mathrm{spr}(B) = 0$ (a $A \neq \mathbf{0} \neq B$) te da je $\mathrm{spr}(A + B) = 1 > \mathrm{spr}(A) + \mathrm{spr}(B)$. Nije norma, ali je blisko povezan sa normama na $\mathbb{M}_n$.

Teorem 1.5.1. *Za proizvoljnu matričnu normu $\|\cdot\|$ na $\mathbb{M}_n$ i svaku matricu $\mathsf{A} \in \mathbb{M}_n$ vrijedi* $\mathrm{spr}(\mathsf{A}) \leq \|\mathsf{A}\|$. *Nadalje, za svaku matricu $\mathsf{A} \in \mathbb{M}_n$ i svaki $\epsilon > 0$ postoji inducirana matrična norma $|||\cdot|||$ za koju je $|||\mathsf{A}||| \leq \mathrm{spr}(\mathsf{A}) + \epsilon$.*

Dokaz: Neka je λ proizvoljna svojstvena vrijednost od A, sa pripadnim svojstvenim vektorom v. Ako definiramo $n \times n$ matricu $\mathsf{V} \neq \mathbf{0}$ tako da za njenih n stupaca

stavimo n kopija vektora v, onda je $\mathsf{AV} = \lambda\mathsf{V}$. Uzimanjem norme je $|\lambda|\|\mathsf{V}\| = \|\mathsf{AV}\| \leq \|\mathsf{A}\|\|\mathsf{V}\|$, tj. $|\lambda| \leq \|\mathsf{A}\|$.

Neka je $\mathsf{U}^*\mathsf{AU} = \mathsf{T}$ Schurova forma marice A. Uzmimo proizvoljan $\epsilon > 0$ i stavimo $D_{\tilde{\epsilon}} = \mathrm{diag}(1, \tilde{\epsilon}, \tilde{\epsilon}^2, \ldots, \tilde{\epsilon}^{n-1})$, pri čemu ćemo $\tilde{\epsilon} \in (0, \epsilon)$ odabrati poslije. Matrica $\mathsf{T}_{\tilde{\epsilon}} = D_{\tilde{\epsilon}}^{-1}\mathsf{T}D_{\tilde{\epsilon}} = D_{\tilde{\epsilon}}^{-1}\mathsf{U}^*\mathsf{AU}D_{\tilde{\epsilon}}$ se od matrice T razlikuje samo na pozicijama (i, j) u strogo gornjem trokutu ($i < j$) gdje je $(\mathsf{T}_{\tilde{\epsilon}})_{ij} = (\mathsf{T})_{ij}\tilde{\epsilon}^{j-i}$. Definirajmo vektorsku normu $\|x\| = \|(\mathsf{U}D_{\tilde{\epsilon}})^{-1}x\|_\infty$. Pripadna operatorska norma je

$$\begin{aligned}
|||\mathsf{A}||| &= \max_{x\neq\mathbf{0}} \frac{\|\mathsf{A}x\|}{\|x\|} = \max_{x\neq\mathbf{0}} \frac{\|(\mathsf{U}D_{\tilde{\epsilon}})^{-1}\mathsf{A}x\|_\infty}{\|(\mathsf{U}D_{\tilde{\epsilon}})^{-1}x\|_\infty} = \max_{y\neq\mathbf{0}} \frac{\|(\mathsf{U}D_{\tilde{\epsilon}})^{-1}\mathsf{AU}D_{\tilde{\epsilon}}y\|_\infty}{\|y\|_\infty} \\
&= \max_{y\neq\mathbf{0}} \frac{\|\mathsf{T}_{\tilde{\epsilon}}\|_\infty}{\|y\|_\infty} = |||\mathsf{T}_{\tilde{\epsilon}}|||_\infty = \max_{i=1:n} \sum_{j=i}^{n} |(\mathsf{T}_{\tilde{\epsilon}})_{ij}| = \max_{i=1:n} \sum_{j=i}^{n} |(\mathsf{T})_{ij}|\tilde{\epsilon}^{j-i} \\
&\leq \mathrm{spr}(\mathsf{A}) + \epsilon, \quad \text{npr. za } 0 < \tilde{\epsilon} < \epsilon/|||\mathsf{T}|||_\infty.
\end{aligned}$$

⊞

Teorem 1.5.2. *Neka je $\|\cdot\|$ proizvoljna matrična norma na $\mathbb{M}_n$. Tada je za svaku matricu $\mathsf{A} \in \mathbb{M}_n$*

$$\mathrm{spr}(\mathsf{A}) = \lim_{k\to\infty} \|\mathsf{A}^k\|^{1/k}.$$

Dokaz: Iz relacija $\mathrm{spr}(\mathsf{A})^k = \mathrm{spr}(\mathsf{A}^k) \leq \|\mathsf{A}^k\|$ odmah slijedi $\mathrm{spr}(\mathsf{A}) \leq \|\mathsf{A}^k\|^{1/k}$. Sada uzmimo proizvoljan $\epsilon > 0$ i definirajmo $\tilde{\mathsf{A}} = (\mathrm{spr}(\mathsf{A}) + \epsilon)^{-1}\mathsf{A}$. Kako je $\mathrm{spr}(\tilde{\mathsf{A}}) < 1$, vrijedi $\lim_{k\to\infty} \|\tilde{\mathsf{A}}^k\| = 0$, pa postoji indeks k_0 tako da je za sve $k > k_0$ $\|\tilde{\mathsf{A}}^k\| < 1$. Drugim riječima, za sve $k > k_0$ je $\|\mathsf{A}^k\|^{1/k} < \mathrm{spr}(\mathsf{A}) + \epsilon$ i tvrdnja je dokazana. ⊞

Teorem 1.5.3. *Neka su $\mathsf{A} \in \mathbb{M}_n, \mathsf{B} \in \mathbb{R}^{n\times n}$. Vrijedi $\mathrm{spr}(\mathsf{A}) \leq \mathrm{spr}(|\mathsf{A}|)$. Nadalje, ako je $|\mathsf{A}| \leq \mathsf{B}$, onda je $\mathrm{spr}(|\mathsf{A}|) \leq \mathrm{spr}(\mathsf{B})$.*

Dokaz: Uočimo da vrijedi $|\mathsf{A}^k| \leq |\mathsf{A}|^k \leq \mathsf{B}^k$, pa uzimanjem Frobenijusove norme dobijemo

$$\|\mathsf{A}^k\|_{\mathrm{F}}^{1/k} \leq \||\mathsf{A}|^k\|_{\mathrm{F}}^{1/k} \leq \|\mathsf{B}^k\|_{\mathrm{F}}^{1/k}.$$

Sada puštanjem $k \to \infty$ i primjenom Teorema 1.5.2 slijedi tvrdnja. ⊞

Teorem 1.5.4. *Za proizvoljnu matricu $\mathsf{A} \in \mathbb{M}_n$ vrijedi:*

$$\lim_{k\to\infty} \mathsf{A}^k = \mathbf{0} \Longleftrightarrow \mathrm{spr}(\mathsf{A}) < 1.$$

Dokaz: Neka je $\lim_{k\to\infty} \mathsf{A}^k = \mathbf{0}$. Za Schurovu trokutastu formu $\mathsf{T} = \mathsf{U}^*\mathsf{A}\mathsf{U}$ tada očito vrijedi $\lim_{k\to\infty} \mathsf{T}^k = \mathbf{0}$, pri čemu su na dijagonali od T^k k–te potencije svojstvenih vrijednosti od A. Kako su svi ti brojevi u limesu ($k \to \infty$) jednaki nuli, nužno je $\mathrm{spr}(\mathsf{A}) < 1$.

Neka je sada $\mathrm{spr}(\mathsf{A}) < 1$. Prema Teoremu 1.5.1 postoji norma $\|\cdot\|$ u kojoj je $\|\mathsf{A}\| < 1$. No, tada je $\|\mathsf{A}^k\| \leq \|\mathsf{A}\|^k \to 0$ ($k \to \infty$), tj. $\lim_{k\to\infty} \mathsf{A}^k = \mathbf{0}$. ⊞

Teorem 1.5.5. *Neka je* $\mathsf{A} \in \mathbb{C}^{n\times n}$. *Ako je* $\mathrm{spr}(\mathsf{A}) < 1$ *onda je* $\mathsf{I} - \mathsf{A}$ *regularna matrica i*

$$(\mathsf{I} - \mathsf{A})^{-1} = \sum_{k=0}^{\infty} \mathsf{A}^k. \tag{1.5.1}$$

Obratno, ako $\sum_{k=0}^{\infty} \mathsf{A}^k$ *konvergira, onda je njegova suma* $(\mathsf{I} - \mathsf{A})^{-1}$ *i* $\mathrm{spr}(\mathsf{A}) < 1$.

Dokaz: Svojstvene vrijednosti od $\mathsf{I} - \mathsf{A}$ su oblika $1 - \lambda$, gdje λ označava svojstvenu vrijednost od A. Kako je $|\lambda| \leq \mathrm{spr}(\mathsf{A}) < 1$, nula ne može biti u spektru od $\mathsf{I} - \mathsf{A}$. Iz očite relacije $(\mathsf{I} - \mathsf{A})(\mathsf{I} + \mathsf{A} + \mathsf{A}^2 + \cdots + \mathsf{A}^m) = \mathsf{I} - \mathsf{A}^{m+1}$ lako izvedemo

$$(\mathsf{I} - \mathsf{A})^{-1} - (\mathsf{I} + \mathsf{A} + \mathsf{A}^2 + \cdots + \mathsf{A}^m) = (\mathsf{I} - \mathsf{A})^{-1}\mathsf{A}^{m+1},$$

pri čemu desna strana konvergira u nulu kada $m \to \infty$, tj. u proizvoljnoj normi $\|\cdot\|$ je

$$\|(\mathsf{I} - \mathsf{A})^{-1} - (\mathsf{I} + \mathsf{A} + \mathsf{A}^2 + \cdots + \mathsf{A}^m)\| \leq \|(\mathsf{I} - \mathsf{A})^{-1}\| \underbrace{\|\mathsf{A}^{m+1}\|}_{\to 0\ (m\to\infty)}.$$

Ovime je dokazano (1.5.1). Sada pretpostavimo da $\sum_{k=0}^{\infty} \mathsf{A}^k$ konvergira. Tada je nužno $\lim_{k\to\infty} \mathsf{A}^k = \mathbf{0}$ i prema Teoremu 1.5.4 je $\mathrm{spr}(\mathsf{A}) < 1$, i očito je suma reda jednaka $(\mathsf{I} - \mathsf{A})^{-1}$. ⊞

Lema 1.5.6. *Neka je* $\mathsf{A} \in \mathbb{M}_n$ *i neka su* λ, x, y *svojstvena vrijednost sa pripadnim lijevim i desnim svojstvenim vektorom,* $\mathsf{A}x = \lambda x$, $\mathsf{A}^{\mathrm{T}}y = \lambda y$, $x^{\mathrm{T}}y = 1$. *Stavimo* $\mathsf{L} = xy^T$. *Tada je*

$$(\mathsf{A} - \lambda\mathsf{L})^m = \mathsf{A}^m - \lambda^m\mathsf{L}. \tag{1.5.2}$$

Nadalje ako su svojstvene vrijednosti $\lambda_1, \ldots, \lambda_n$ *od* A *numerirane tako da je* $|\lambda_1| = \mathrm{spr}(\mathsf{A}) > \max_{i=2:n} |\lambda_i|$, *onda je*

$$\lim_{m\to\infty} (\frac{1}{\lambda_1}\mathsf{A})^m = \mathsf{L}. \tag{1.5.3}$$

Dokaz: Lako se provjeri da je, za svaki $m \in \mathbb{N}$, $\mathsf{L}^m = \mathsf{L}$, te da je $\mathsf{AL} = \mathsf{LA} = \lambda\mathsf{L}$. Matematičkom indukcijom se lako pokaže (1.5.2). Za drugu tvrdnju, uočimo da dijeljenjem (1.5.2) s $\lambda^m \equiv \lambda_1^m$ dobijemo

$$(\frac{1}{\lambda_1}\mathsf{A})^m = \mathsf{L} + (\frac{1}{\lambda_1}\mathsf{A} - \mathsf{L})^m,$$

pri čemu je

$$\mathrm{spr}(\frac{1}{\lambda_1}\mathsf{A} - \mathsf{L}) = \frac{\mathrm{spr}(\mathsf{A} - \lambda_1\mathsf{L})}{\mathrm{spr}(\mathsf{A})} = \frac{\max_{i=2:n}|\lambda_i|}{\mathrm{spr}(\mathsf{A})} < 1, \text{ tj. } \lim_{m\to\infty}(\frac{1}{\lambda_1}\mathsf{A} - \mathsf{L})^m = \mathbf{0}.$$

⊞

1.6 Lociranje spektra

Ponekad je dovoljno samo približno locirati svojstvene vrijednosti matrice. Na primjer, za zaključak da je matrica regularna, dovoljno je na neki način zaključiti da nula nije u njenom spektru. U nekoj drugoj situaciji može biti važna npr. činjenica da su sve svojstvene vrijednosti po modulu manje od jedan ili da su sve u lijevoj otvorenoj kompleksnoj poluravnini tj. sa strogo negativnim realnim dijelovima.

1.6.1 Geršgorinovi krugovi

Teorem 1.6.1. *Neka je $\mathsf{A} \in \mathbb{C}^{n\times n}$ i neka su $\lambda_1, \ldots, \lambda_n$ njene svojstvene vrijednosti. Za $i = 1, \ldots, n$ definirajmo Geršgorinove krugove*

$$\mathcal{G}_i = \{z \in \mathbb{C} \,:\, |z - a_{ii}| \le \rho_i\}, \quad \rho_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \tag{1.6.1}$$

Tada vrijedi:

- ∗ *Sve svojstvene vrijednosti od A su sadržane u uniji Geršgorinovih krugova, $\mathfrak{S}(\mathsf{A}) \subseteq \bigcup_{i=1}^{n} \mathcal{G}_i$.*
- ⋆ *Ako je unija $\mathcal{G}_{i_1\cdots i_k} = \bigcup_{j=1}^{k} \mathcal{G}_{i_j}$ nekih k Geršgorinovih krugova disjunktna s unijom preostalih $n-k$ krugova, onda se u $\mathcal{G}_{i_1\cdots i_k}$ nalazi točno k svojstvenih vrijednosti od A (računato s algebarskim kratnostima).*

Dokaz: Dokažimo prvo da je svaka svojstvena vrijednost u nekom krugu. Neka je $\mathsf{A}\mathsf{v} = \lambda_i \mathsf{v}$, $\mathsf{v} \in \mathbb{C}^n \setminus \{\mathsf{0}\}$, te neka je $|\mathsf{v}_j| = \|\mathsf{v}\|_\infty$. Sada j–tu jednadžbu u relaciji $\mathsf{A}\mathsf{v} = \lambda_i \mathsf{v}$ transformiramo u

$$\lambda_i - a_{jj} = \sum_{\substack{k=1 \\ k\neq j}}^{n} a_{jk} \frac{\mathsf{v}_k}{\mathsf{v}_j}, \quad \text{odakle slijedi } \; |\lambda_i - a_{jj}| \leq \rho_j, \quad \text{tj. } \; \lambda_i \in \mathcal{G}_j. \tag{1.6.2}$$

Pokažimo sada drugu tvrdnju. Neka $\mathcal{G}^c_{i_1\cdots i_k}$ označava uniju preostalih $n-k$ krugova i neka je $d = \min\{|z_1 - z_2| : z_1 \in \mathcal{G}_{i_1\cdots i_k}, z_2 \in \mathcal{G}^c_{i_1\cdots i_k}\}$. Zbog zatvorenosti i disjunknosti je $d > 0$. Matricu A napišimo kao zbroj $\mathsf{A} = \mathsf{D} + \mathsf{N}$, gdje je $\mathsf{D} = \mathrm{diag}(a_{11}, \ldots, a_{nn})$. Neka je $\mathsf{N} \neq \mathbf{0}$ (inače je tvrdnja trivijalna). Definirajmo neprekidno preslikavanje

$$\mathsf{A}(t) : [0,1] \longrightarrow \mathbb{C}^{n\times n}, \quad \mathsf{A}(t) = (1-t)\mathsf{D} + t\mathsf{A} \equiv \mathsf{D} + t\mathsf{N}. \tag{1.6.3}$$

Očito je $\mathsf{A}(0) = \mathsf{D}$, $\mathsf{A}(1) = \mathsf{A}$. Ako prvu tvrdnju primijenimo na matrice $\mathsf{A}(t)$ dobijemo da je, za svaki $t \in [0,1]$, $\mathfrak{S}(\mathsf{A}(t)) \subseteq \bigcup_{i=1}^n \mathcal{G}_i(t)$, gdje je $\mathcal{G}_i(t) = \{z \in \mathbb{C} : |z - a_{ii}| \leq t\rho_i\} \subseteq \mathcal{G}_i(1) = \mathcal{G}_i$. Pri tome, za sve $\mathsf{A}(t)$, $\mathcal{G}_{i_1\cdots i_k}(t) = \bigcup_{j=1}^k \mathcal{G}_{i_j}(t)$ ostaje disjunktna sa unijom $\mathcal{G}_{i_1\cdots i_k}(t)^c$ preostalih $n-k$ krugova.

Primijetimo da $\mathcal{G}_{i_1\cdots i_k}(0) = \bigcup_{j=1}^k \{a_{i_j}\}$ sadrži točno k svojstvenih vrijednosti od $\mathsf{A}(0) = \mathsf{D}$; to su upravo dijagonalni elementi s indeksima $i_1, \ldots, i_k$. Preostaje ove degenerirane krugove neprekidno "napuhati" do $\mathcal{G}_{i_1\cdots i_k}(1)$ i pri tome pažljivo pratiti broj svojstvenih vrijednosti.

Prvo Teorem 1.4.4 primijenimo na $\mathsf{A}(0) = \mathsf{D}$ sa proizvoljnim dovoljno malim $\epsilon > 0$. Ovdje dovoljno mali znači da je $\epsilon < \min_i \rho_i$, i da su krugovi radijusa ϵ, s centrima u međusobno različitim dijagonalnim elementima od D, međusobno disjunktni. Neka je $\|\cdot\|$ proizvoljna norma u $\mathbb{C}^{n\times n}$. Postoji $\delta > 0$ takav da svaka matrica X za koju je $\|\mathsf{A}(0) - \mathsf{X}\| < \delta$ ima točno k svojstvenih vrijednosti u uniji diskova $\bigcup_{j=1}^k \{z \in \mathbb{C} : |z - a_{i_j i_j}| < \epsilon\}$. Analogno se točno $n-k$ svojstvenih vrijednosti od X nalazi u uniji ϵ–krugova oko preostalih $n-k$ dijagonalnih elemenata. Ako dobivenu δ okolinu presiječemo s $\mathsf{A}(t)$, $0 \leq t \leq 1$, onda to specijalno vrijedi za matrice X oblika $\mathsf{A}(t)$, $0 \leq t < \tau \equiv \min\{\delta/\|\mathsf{N}\|, 1\}$. Sada na svaku takvu $\mathsf{A}(t)$ primijenimo prvu tvrdnju, $\mathfrak{S}(\mathsf{A}(t)) \subseteq \bigcup_{i=1}^n \mathcal{G}_i(t) = \mathcal{G}_{i_1\cdots i_k}(t) \bigcup \mathcal{G}_{i_1\cdots i_k}(t)^c$, pri čemu je za $t \leq 1$ $\mathcal{G}_{i_1\cdots i_k}(t) \bigcap \mathcal{G}_{i_1\cdots i_k}(t)^c = \emptyset$ i $\bigcup_{j=1}^k \{z \in \mathbb{C} : |z - a_{i_j i_j}| < \epsilon\} \bigcap \mathcal{G}_{i_1\cdots i_k}(t)^c = \emptyset$. Zaključujemo da, za sve $t \in [0, \tau)$, $\mathcal{G}_{i_1\cdots i_k}(t)$ sadrži točno k svojstvenih vrijednosti od $\mathsf{A}(t)$.

Neka je τ_* supremum svih $\tau \leq 1$ s navedenim svojstvom. Pokažimo da $\mathcal{G}_{i_1\cdots i_k}(\tau_*)$ sadrži točno k svojstvenih vrijednosti od $\mathsf{A}(\tau_*)$. Odaberimo dovoljno mali $\epsilon > 0$,

na primjer $\epsilon < d$. Kako je spektar od $\mathsf{A}(\tau_*)$ neprekidan, postoji δ okolina matrice $\mathsf{A}(\tau_*)$ takva da su spektri svih matrica iz te okoline u ϵ okolini od $\mathfrak{S}(\mathsf{A}(\tau_*))$ u smislu Teorema 1.4.4. Očito za svaki t za kojeg je $|t - \tau_*| < \delta/\|\mathsf{N}\|$ vrijedi da $\mathsf{A}(t)$ pripada toj δ okolini. Iz svojstva supremuma τ_* slijedi da možemo odabrati $t \in (\tau_* - \delta/\|\mathsf{N}\|, \tau_*)$ takav da matrica $\mathsf{A}(t)$ u $\mathcal{G}_{i_1\cdots i_k}(t)$ ima točno k svojstvenih vrijednosti, računato s kratnostima, a cijeli spektar joj je u ϵ okolini od $\mathfrak{S}(\mathsf{A}(\tau_*))$ u smislu Teorema 1.4.4. Kako svo vrijeme sve svojstvene vrijednosti pripadaju uniji dvije disjunktne familije Geršgorinovih krugova, međusobno udaljene za $d > 0$, te kako je $\epsilon < d$, jasno je sljedeće:

- Da bi svojstvena vrijednost od $\mathsf{A}(t)$ koja je u krugovima $\mathcal{G}_{i_1\cdots i_k}(t)$ bila u ϵ okolini neke svojstvene vrijednosti λ matrice $\mathsf{A}(\tau_*)$, onda λ i sama mora pripadati krugovima $\mathcal{G}_{i_1\cdots i_k}(\tau_*)$. (Inače je u $\mathcal{G}^c_{i_1\cdots i_k}(\tau_*)$ koji je od cijelog $\mathcal{G}_{i_1\cdots i_k}$ udaljen za $d > \epsilon$.)

- Iz prethodnog slijedi da $\mathsf{A}(\tau_*)$ u $\mathcal{G}_{i_1\cdots i_k}(\tau_*)$ mora imati barem k svojstvenih vrijednosti (s kratnostima).

- Prethodna dva zaključka vrijede i za $\mathcal{G}^c_{i_1\cdots i_k}(t)$ i $\mathcal{G}^c_{i_1\cdots i_k}(\tau_*)$, pa $\mathsf{A}(\tau_*)$ u $\mathcal{G}^c_{i_1\cdots i_k}(\tau_*)$ mora imati barem $n - k$ svojstvenih vrijednosti (s kratnostima).

- Kako je ukupan broj svojstvenih vrijednosti s kratnostima točno n i kako je $\mathcal{G}_{i_1\cdots i_k}(\tau_*) \bigcap \mathcal{G}^c_{i_1\cdots i_k}(\tau_*) = \emptyset$, slijedi da $\mathcal{G}_{i_1\cdots i_k}(\tau_*)$ sadrži točno k svojstvenih vrijednosti $\lambda_{j_1}(\tau_*), \ldots, \lambda_{j_k}(\tau_*)$ od $\mathsf{A}(\tau_*)$.

Ostaje pokazati da je $\tau_* = 1$. Ako bi bio $\tau_* < 1$, onda bismo uzeli $t \in (\tau_*, \min(1, \tau_* + \delta/\|\mathsf{N}\|))$ i, analogno prethodnim razmatranjima, zaključili da onih k svojstvenih vrijednosti od $\mathsf{A}(t)$ koje su u ϵ okolini svojstvenih vrijednosti $\lambda_{j_1}(\tau_*), \ldots, \lambda_{j_k}(\tau_*)$ zapravo moraju pripadati $\mathcal{G}_{i_1\cdots i_k}(t)$. No, tada τ_* nije supremum. $\boxplus$
Kako je $\mathfrak{S}(\mathsf{A}) = \mathfrak{S}(\mathsf{A}^{\mathrm{T}})$, dokazanu tvrdnju možemo primijeniti na matricu A^{T}, pa vrijedi i

$$\mathfrak{S}(\mathsf{A}) \subseteq \bigcup_{i=1}^{n} \mathcal{G}'_i, \quad \mathcal{G}'_i = \{z \in \mathbb{C} \ : \ |z - a_{ii}| \leq \rho'_i\}, \quad \rho'_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ji}|.$$

Tako presjek dvije unije krugova može dati precizniju informaciju. Naravno da područje dobiveno pomoću Geršgorinovih krugova možemo presijeći sa bilo kojim drugim za koje znamo da sadrži $\mathfrak{S}(\mathsf{A})$. Na primjer ako znamo da matrica ima samo realne svojstvene vrijednosti, dobivenu uniju Geršgorinovih krugova ćemo presijeći s realnom osi.

Očito je da će informacija dobivena Geršgorinovim krugovima biti preciznija ako su im radijusi manji (u usporedbi s modulima centara krugova). Takvu situaciju imamo kod matrica u kojima dijagonalni elementi po modulu dominiraju ostacima svojih redaka ili stupaca.

Definicija 1.6.1. Kažemo da je $\mathsf{A} \in \mathbb{C}^{n\times n}$ *strogo dijagonalno dominantna po retcima* ako je

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

Kažemo da je $\mathsf{A} \in \mathbb{C}^{n\times n}$ *strogo dijagonalno dominantna po stupcima* ako vrijedi $|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ji}|$, $i = 1, \ldots, n$.

Odmah zaključujemo da su matrice koje su strogo dijagonalno dominantne (po retcima ili po stupcima) nužno regularne.

Informaciju dobivenu pomoću Geršgorinovih krugova možemo i na druge načine kombinirati sa nekim drugim poznatim svojstvima. Na primjer, svojstvene vrijednosti realne matrice su ili realne ili dolaze u kompleksno–konjugiranim parovima, pa možemo zaključiti:

Korolar 1.6.2. *Ako je* A *realna matrica, onda svaki izolirani Geršgorinov krug (onaj koji je disjunktan sa svim ostalim) sadrži točno jednu i to realnu svojstvenu vrijednost.*

Ponekada jednostavnom transformacijom sličnosti prije primjene Teorema 1.6.1 možemo dobiti precizniju informaciju. Najjednostavniji primjer je sličnost $\mathsf{B} = \mathsf{D}^{-1}\mathsf{A}\mathsf{D}$, gdje je D dijagonalna matrica.

Primjer 1.6.1. Neka je

1.6.2 Cassinijevi ovali

Ideja Geršgorinovih krugova je jednostavna i elegantna i naravno da je inspirirala cijeli niz sličnih ocjena. Za ilustraciju, spomenut ćemo Cassinijeve ovale koji se dobiju kada u dokazu Teorema 1.6.1 u svojstvenom vektoru promatramo dvije po modulu najveće komponente.

Propozicija 1.6.3. *Neka je* $\mathsf{A} \in \mathbb{C}^{n\times n}$, $\rho_i = \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|$, $i = 1, \ldots, n$. *Tada je*

$$\mathfrak{S}(\mathsf{A}) \subseteq \bigcup_{j<\ell} \mathcal{C}_{j\ell}, \quad \text{gdje je za } j \neq \ell \;\; \mathcal{C}_{j\ell} = \{z \in \mathbb{C} \;:\; |z - a_{jj}||z - a_{\ell\ell}| \leq \rho_j \rho_\ell\}\,.$$

Skupovi $\mathcal{C}_{j\ell}$ *se zovu Cassinijevi ovali.*

Dokaz: Neka je $\mathsf{A}\mathsf{v} = \lambda_i \mathsf{v}$, $\mathsf{v} \in \mathbb{C}^n \setminus \{\mathsf{0}\}$, te neka je $|\mathsf{v}_j| = \|\mathsf{v}\|_\infty \geq |\mathsf{v}_\ell| \geq |\mathsf{v}_k|$, $k \neq j, \ell$. Ako je $\mathsf{v}_\ell \neq 0$ lako dobijemo da vrijedi

$$|\lambda_i - a_{jj}| \leq \sum_{\substack{k=1\\ k\neq j}}^{n} |a_{jk}| \frac{|\mathsf{v}_\ell|}{|\mathsf{v}_j|}, \quad |\lambda_i - a_{\ell\ell}| \leq \sum_{\substack{k=1\\ k\neq \ell}}^{n} |a_{\ell k}| \frac{|\mathsf{v}_j|}{|\mathsf{v}_\ell|},$$

odakle je odmah $|\lambda_i - a_{jj}||\lambda_i - a_{\ell\ell}| \leq \rho_j \rho_\ell$, tj. $\lambda_i \in \mathcal{C}_{j\ell} = \mathcal{C}_{\ell j}$. Ako je $\mathsf{v}_\ell = 0$, onda je $\mathsf{v} = \mathsf{v}_j \mathbf{e}_j$ pa je $\lambda_i = a_{jj} \in \mathcal{C}_{j\ell}$. ⊞

Primjer 1.6.2. Neka je ...

1.6.3 Skalirani Geršgorinovi krugovi

Vidjeli smo da Teorem 1.6.1 daje dobre procjene svojstvenih vrijednosti ako je matrica strogo dijagonalno dominantna. Barlow i Demmel [1] su uočili da se svojstvo dijagonalne dominantnosti može shvatiti i u jednom širem smislu i uveli su pojam skalirane dijagonalne dominantnosti.

Definicija 1.6.2. Matrica $\mathsf{A} \in \mathbb{C}^{n\times n}$ je γ*–dijagonalno dominantna* u normi $\|\cdot\|$ ako je možemo rastaviti na $\mathsf{A} = D + N$, gdje je D dijagonalna, N ima nule na dijagonali i vrijedi $\|N\| \leq \gamma \min_i |D_{ii}|$ sa $\gamma \in [0, 1)$.

Definicija 1.6.3. Kažemo da je matrica $\mathsf{A} \in \mathbb{C}^{n\times n}$ γ*–skalirano dijagonalno dominantna* u normi $\|\cdot\|$ (kratko pišemo γ–s.d.d) ako je možemo napisati kao produkt $A = D_1 B D_2$, gdje su D_1, D_2 dijagonalne matrice, $|B_{jj}| = 1$ za $j = 1, \ldots, n$, i B je γ–dijagonalno dominantna u normi $\|\cdot\|$.

Primjer 1.6.3. Matrica

$$\mathsf{A} = \begin{pmatrix} 10^9 & -10^2 \\ 10^2 & 1 \end{pmatrix} = \begin{pmatrix} 10^5 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -10^{-3} \\ 10^{-2} & 1 \end{pmatrix} \begin{pmatrix} 10^4 & 0 \\ 0 & 1 \end{pmatrix}$$

je u normi $\|\cdot\|_\infty$ γ–s.d.d sa $\gamma = 0.01$.

Sljedeća propozicija daje novi skup krugova, koji se može koristiti kao dodatak Teoremu 1.6.1 i to posebno u slučaju γ–s.d.d. matrica sa $\gamma \ll 1$ kada bitno bolje locira po modulu najmanje svojstvene vrijednosti.

Propozicija 1.6.4. *Neka je* $\mathsf{A} \in \mathbb{C}^{n\times n}$ *γ–skalirano dijagonalno dominantna u normi* $\|\cdot\|_\infty$ *ili* $\|\cdot\|_1$. *Tada je*

$$\mathfrak{S}(\mathsf{A}) \subseteq \bigcup_{i=1}^{n} \mathcal{BD}_i, \;\; \mathcal{BD}_i = \{z \in \mathbb{C} \;:\; |z - a_{ii}| \leq \gamma |a_{ii}|\}\,. \tag{1.6.4}$$

Ako je unija $\mathcal{BD}_{i_1\cdots i_k} = \bigcup_{j=1}^{k} \mathcal{BD}_{i_j}$ *nekih k krugova disjunktna s unijom preostalih* $n-k$ *krugova, onda se u* $\mathcal{BD}_{i_1\cdots i_k}$ *nalazi točno k svojstvenih vrijednosti od* A *(računato s algebarskim kratnostima).*

Dokaz: Dovoljno je tvrdnju dokazati za $\|\cdot\|_\infty$ i dokazano primijeniti na A^{T}. Uzmimo proizvoljnu svojstvenu vrijednost λ_i matrice A. Tada je $\mathsf{A} - \lambda_i \mathsf{I}$ singularna, pa je i $B - \lambda_i D_1^{-1} D_2^{-1}$ singularna. Neka je $x \neq \mathbf{0}$ netrivijalni vektor iz njene jezgre i neka je $|x_j| = \|x\|_\infty > 0$. Sada u jednakosti $(B - \lambda_i D_1^{-1} D_2^{-1})x = \mathbf{0}$ pročitamo j–ti redak,

$$\sum_{k=1}^{n} b_{jk} x_k - \lambda_i x_j \frac{1}{(D_1)_{jj}(D_2)_{jj}} = 0,$$

pa ga napišemo u ekvivalentnom obliku $\lambda - a_{jj} = (D_1)_{jj}(D_2)_{jj} \displaystyle\sum_{\substack{k=1\\k\neq j}}^{n} b_{jk} \frac{x_k}{x_j}$ i uzimanjem apsolutne vrijednosti dobijemo $|\lambda - a_{jj}| \leq \gamma |a_{jj}|$.

Za drugi dio tvrdnje je, nakon iskustva s Teoremom 1.6.1, jasno da mora vrijediti i da možemo, uz jednostavnu prilagodbu, iskoristiti dokaz Teorema 1.6.1. Ali, to nećemo napraviti. Umjesto toga ćemo napraviti novi, drugačiji, dokaz koristeći neprekidnu realizaciju spektra duž segmenta (1.6.3). (Naravno, ista tehnika onda može poslužiti i za novi dokaz Teorema 1.6.1.)

Prema Teoremu 1.4.7, postoji n neprekidnih funkcija $\lambda_1(t), \ldots, \lambda_n(t)$ čije vrijednosti u bilo kojem trenutku t čine spektar of $\mathsf{A}(t)$. Lako se vidi da je za svaki $t \in [0,1]$ matrica $\mathsf{A}(t)$ $t\gamma$–s.d.d. i da za nju vrijedi (1.6.4). Pri tome su krugovi koji prekrivaju spektar matrice $\mathsf{A}(t)$ dani s $\mathcal{BD}_i(t) = \{z \in \mathbb{C} \;:\; |z - a_{ii}| \leq t\gamma |a_{ii}|\}$, te vrijedi $\mathcal{BD}_i(t) \subseteq \mathcal{BD}_i(1) \equiv \mathcal{BD}_i$. Dakle je za svaki t i unija $\bigcup_{j=1}^{k} \mathcal{BD}_{i_j}(t)$ disjunktna s $\bigcup_{j=k+1}^{n} \mathcal{BD}_{i_j}$ (pa onda i s $\bigcup_{j=k+1}^{n} \mathcal{BD}_{i_j}(t)$).

U trenutku $t = 0$ tvrdnja vrijedi: $\mathsf{A}(0)$ u uniji $\bigcup_{j=1}^{k} \mathcal{BD}_{i_j}(0)$ ima točno k svojstvenih vrijednosti $\lambda_{i_j}(0) = a_{i_j i_j}$, $j = 1, \ldots, k$. Sada pustimo patametar t da se

neprekidno mijenja $0 \rightsquigarrow 1$ i pratimo k neprekidnih krivulja $\lambda_{i_j}(t)$, $j = 1, \ldots, k$. One cijelo vrijeme moraju biti u uniji dvije disjunktne komponente $\bigcup_{j=1}^{k} \mathcal{BD}_{i_j}$ i $\bigcup_{j=k+1}^{n} \mathcal{BD}_{i_j}$. Kako su na početku puta, u $t = 0$, krenule iz prve komponente, intuitivno je jasno da neprekidnom promjenom u trenutku $t = 1$ nisu mogle završiti u drugoj – naime, onda bi, zbog neprekidnosti, u nekom momentu morale biti van te unije, što je nemoguće. Dakle, $\bigcup_{j=1}^{k} \mathcal{BD}_{i_j}$ sadrži barem k svojstvenih vrijednosti od A. Na isti način $\bigcup_{j=k+1}^{n} \mathcal{BD}_{i_j}$ sadrži barem $n-k$ svojstvenih vrijednosti od A.

Sada još ostaje formalno dokazati ono što je bilo intuitivno jasno. Neka je $d = \inf\{|\xi - \zeta| \ : \ \xi \in \bigcup_{j=1}^{k} \mathcal{BD}_{i_j}, \zeta \in \bigcup_{j=k+1}^{n} \mathcal{BD}_{i_j}\}$. Zbog zatvorenosti i disjunktnosti je $d > 0$. Funkcija $f_{i_j}(t) = \inf\{|z - \lambda_{i_j}(t)| \ : \ z \in \bigcup_{j=k+1}^{n} \mathcal{BD}_{i_j}\}$ je neprekidna i $f_{i_j}(0) \geq d$. Ako bi bilo $\lambda_{i_j}(1) \in \bigcup_{j=k+1}^{n} \mathcal{BD}_{i_j}$, onda bi vrijedilo $f_{i_j}(1) = 0$ pa bi zbog neprekidnosti (Weierstrassov teorem) postojala vrijednost $t_! \in (0,1)$ za koju je $f_{i_j}(t_!) = d/2$. Tada bi vrijedilo $\lambda_{i_j}(t_!) \notin \bigcup_{i=1}^{n} \mathcal{BD}_i(t_!)$ – kontradikcija. ⊞

Primjer 1.6.4. Neka je ...

Na sličan način možemo dobiti skalirane Cassinijeve ovale.

Propozicija 1.6.5. *Neka je* $\mathsf{A} \in \mathbb{C}^{n\times n}$ *γ–skalirano dijagonalno dominantna u normi* $\|\cdot\|_\infty$ *ili* $\|\cdot\|_1$. *Tada je*

$$\mathfrak{S}(\mathsf{A}) \subseteq \bigcup_{j<\ell} \left\{ z \in \mathbb{C} \ : \ |z - a_{jj}||z - a_{\ell\ell}| \leq \gamma^2 |a_{jj}||a_{\ell\ell}| \right\}.$$

Primjer 1.6.5. Neka je ...

Poglavlje 2

Ireducibilne i nenegativne matrice

2.1 Ireducibilne matrice

Ponekad raspored nula u matrici A inducira strukturu koja je kombinatornog tipa ali koja za posljedicu ima niz netrivijalnih svojstava matrice, posebno njenih svojstvenih vrijednosti.

Definicija 2.1.1. Kažemo da je $\mathsf{A} \in \mathbb{C}^{n\times n}$, $n \geq 2$, reducibilna ako postoji matrica permutacije P i $r \in \{1, \ldots, n-1\}$ tako da je

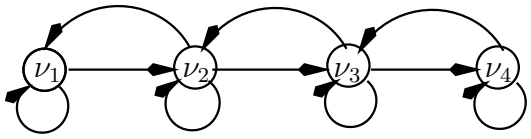
$$\mathsf{P}^{\mathrm{T}}\mathsf{A}\mathsf{P} = \begin{pmatrix} \tilde{\mathsf{A}}_{[11]} & \tilde{\mathsf{A}}_{[12]} \\ \mathbf{0} & \tilde{\mathsf{A}}_{[22]} \end{pmatrix}, \quad \tilde{\mathsf{A}}_{[11]} \in \mathbb{C}^{r\times r}. \tag{2.1.1}$$

Za $n = 1$ je A reducibilna ako je $\mathsf{A} = (0)$. Matrica A je ireducibilna ako nije reducibilna.

U kontekstu rješavanja sustava linearnih jednadžbi $\mathsf{A}\mathsf{x} = \mathsf{b}$, renumeriranjem nepoznanica i jednadžbi matricom permutacije P dobijemo sustav $\mathsf{P}^{\mathrm{T}}\mathsf{A}\mathsf{P}(\mathsf{P}^{\mathrm{T}}\mathsf{x}) = \mathsf{P}^{\mathrm{T}}\mathsf{b}$. U slučaju reducibilnosti možemo postići blok–trokutasti oblik (2.1.1) i particijom vektora $\tilde{\mathsf{x}} = \mathsf{P}^{\mathrm{T}}\mathsf{x}$, $\tilde{\mathsf{b}} = \mathsf{P}^{\mathrm{T}}\mathsf{b}$ dobijemo

$$\begin{pmatrix} \tilde{\mathsf{A}}_{[11]} & \tilde{\mathsf{A}}_{[12]} \\ \mathbf{0} & \tilde{\mathsf{A}}_{[22]} \end{pmatrix} \begin{pmatrix} \tilde{\mathsf{x}}_{[1]} \\ \tilde{\mathsf{x}}_{[2]} \end{pmatrix} = \begin{pmatrix} \tilde{\mathsf{b}}_{[1]} \\ \tilde{\mathsf{b}}_{[2]} \end{pmatrix}.$$

Time je polazni problem dimenzije n reduciran na dva manje dimenzije: $\tilde{\mathsf{A}}_{[22]}\tilde{\mathsf{x}}_{[2]} = \tilde{\mathsf{b}}_{[2]}$ i $\tilde{\mathsf{A}}_{[11]}\tilde{\mathsf{x}}_{[1]} = \tilde{\mathsf{b}}_{[1]} - \tilde{\mathsf{A}}_{[12]}\tilde{\mathsf{x}}_{[2]}$.



$$\mathsf{A} = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}$$

Slika 2.1: Matrica A i pripadni graf $\Gamma(\mathsf{A})$. Uočimo da je $\Gamma(\mathsf{A})$ jako povezan.

Ako je neka od matrica $\tilde{\mathsf{A}}_{[11]}$, $\tilde{\mathsf{A}}_{[22]}$ reducibilna onda je na isti način možemo permutacijskom transformacijom sličnosti svesti na blok–trokutasti oblik. Postupak primjenjujemo rekurzivno sve dok na kraju ne dobijemo blok–trokutastu matricu

$$\mathsf{P}^T\mathsf{A}\mathsf{P} = \begin{pmatrix} \mathsf{A}_{[11]} & \mathsf{A}_{[12]} & \cdots & \mathsf{A}_{[1k]} & \cdots & \mathsf{A}_{[1,\ell]} \\ & \mathsf{A}_{[22]} & \mathsf{A}_{[23]} & \cdots & \cdots & \mathsf{A}_{[2,\ell]} \\ & & \ddots & \ddots & \cdots & \vdots \\ & & & \mathsf{A}_{[kk]} & \cdots & \mathsf{A}_{[k,\ell]} \\ & & & & \ddots & \vdots \\ & & & & & \mathsf{A}_{[\ell\ell]} \end{pmatrix}. \tag{2.1.2}$$

u kojoj su svi dijagonalni blokovi ireducibilne matrice ili 1×1 nul–matrice.

Definicija 2.1.2. Blok gornje trokutasta matrica sa ireducibilnim dijagonalnim blokovima u relaciji (2.1.1) se zove ireducibilna normalna forma matrice reducibilne A.

Ireducibilnost matrice je svojstvo kombinatorne prirode i kao takvoga ga možemo prirodnije karakterizirati preko jednog kombinatornog objekta izvedenog iz matrice.

Definicija 2.1.3. Za matricu $\mathsf{A} \in \mathbb{M}_n$ vežemo usmjereni graf $\Gamma(\mathsf{A})$ koji ima n vrhova $\nu_1, \nu_2, \ldots, \nu_n$ i u kojem postoji usmjereni brid $\overrightarrow{\nu_i\nu_j}$ ako i samo ako je $a_{ij} \neq 0$. Put u Γ od vrha ν_k do vrha ν_ℓ je niz usmjerenih bridova $\overrightarrow{\nu_k\nu_{i_1}}, \overrightarrow{\nu_{i_1}\nu_{i_2}}, \ldots, \overrightarrow{\nu_{i_{j-1}}\nu_\ell}$, pri čemu je j duljina puta. Kažemo da je $\Gamma(\mathsf{A})$ jako povezan ako su svaka dva njegova vrha povezana nekim putem.

Propozicija 2.1.1. *Matrica $\mathsf{A} \in \mathbb{C}^{n\times n}$ je ireducibilna ako i samo ako je njen graf $\Gamma(\mathsf{A})$ jako povezan.*

Dokaz: Neka je A reducibilna i neka s nekom permutacijom P vrijedi (2.1.1). Stavimo $\tilde{\mathsf{A}} = \mathsf{P}^T\mathsf{A}\mathsf{P}$. Grafovi $\Gamma(\mathsf{A})$ i $\Gamma(\tilde{\mathsf{A}})$ se razlikuju samo u oznakama čvorova, označimo ove druge sa $\tilde{\nu}_1, \ldots, \tilde{\nu}_n$. Ako bismo tražili put u $\Gamma(\tilde{\mathsf{A}})$ koji počinje u nekom

čvoru iz $\tilde{\mathcal{V}}_1 = \{\tilde{\nu}_{r+1}, \ldots, \tilde{\nu}_n\}$ a završava u nekom od čvorova iz $\tilde{\mathcal{V}}_2 = \{\tilde{\nu}_1, \ldots, \tilde{\nu}_r\}$, onda bi na tom putu morao postojati brid $\overrightarrow{\tilde{\nu}_i\tilde{\nu}_j}$ sa $\tilde{\nu}_i \in \tilde{\mathcal{V}}_2$, $\tilde{\nu}_j \in \tilde{\mathcal{V}}_2$, tj. morao bi biti $\tilde{a}_{ij} \neq 0$ – ali to je nemoguće jer je taj dio matrice $\tilde{\mathsf{A}}$ jednak nuli. Dakle, $\Gamma(\tilde{\mathsf{A}})$ pa niti $\Gamma(\mathsf{A})$ nije jako povezan.

Sada pretpostavimo da $\Gamma(\mathsf{A})$ nije jako povezan. To znači da postoje dva vrha ν_k i ν_ℓ iz skupa vrhova takva da ne postoji put od ν_k do ν_ℓ. Sada skup $\mathcal{V}$ svih vrhova particionirajmo na sljedeći način: u $\mathcal{V}_1$ neka sadrži ν_ℓ i sve vrhove koji su povezani sa ν_ℓ, a $\mathcal{V}_2 = \mathcal{V} \setminus \mathcal{V}_1$. Sada uvedemo renumeraciju vrhova tako da prvo izlistamo one iz $\mathcal{V}_1$, a zatim iz $\mathcal{V}_2$. Neka je $|\mathcal{V}_1| = r$. Sa odgovarajućom permutacijom P transformiramo A u $\tilde{\mathsf{A}} = \mathsf{P}^T\mathsf{A}\mathsf{P}$. Ako bi za neki $i \in \{r+1, \ldots, n\}$ i $j \in \{1, \ldots, r\}$ bilo $\tilde{a}_{ij} \neq 0$, onda bi to značilo da je neki vrh iz $\mathcal{V}_2$ povezan bridom sa nekim vrhom iz $\mathcal{V}_1$, a time onda i putem sa ν_ℓ. No, to je kontradikcija sa definicijom skupa $\mathcal{V}_2$ pa je $\tilde{\mathsf{A}}$ reducibilna. ⊞

Propozicija 2.1.2. *Neka je* $\mathbb{M}_n \ni \mathsf{A} \geq \mathbf{0}$. *Tada je* A *ireducibilna ako i samo ako je* $(\mathsf{I} + \mathsf{A})^{n-1} > \mathbf{0}$.

Teorem 2.1.3. *(Olga Taussky) Neka je* $\mathsf{A} \in \mathbb{C}^{n\times n}$ *ireducibilna matrica i neka je* $\lambda \in \mathfrak{S}(\mathsf{A})$ *njena svojstvena vrijednost sa svojstvom da nije u unutrašnjosti niti jednog Geršgorinovog kruga. Tada je* λ *sadržana u presjeku svih kružnica* $\partial\mathcal{G}_i$, $i = 1, \ldots, n$.

Dokaz: Neka je v pripadni svojstveni vektor, $\mathsf{A}\mathsf{v} = \lambda\mathsf{v}$, normiran tako da je $\|\mathsf{v}\|_\infty = 1$, te neka je $|\mathsf{v}_r| = \|\mathsf{v}\|_\infty$. Slično kao u dokazu Teorema 1.6.1 zaključujemo da je

$$|\lambda - a_{rr}| \leq \sum_{\substack{j=1\\ j\neq r}}^{n} |a_{rj}||\mathsf{v}_j| \leq \sum_{\substack{j=1\\ j\neq r}}^{n} |a_{rj}| = \rho_r, \tag{2.1.3}$$

a kako λ nije u unutrašnjosti niti jednog kruga, mora biti $|\lambda - a_{rr}| = \rho_r$, tj. $\lambda \in \partial\mathcal{G}_r$. Sada pokažimo da je za proizvoljan $p \neq r$ također $\lambda \in \partial\mathcal{G}_p$. Koristeći ireducibilnost matrice A znamo da postoji niz elemenata $a_{rr_1}, a_{r_1r_2}, \ldots, a_{r_\ell p}$, svi različiti od nule, koji trasira put od r–tog do p–tog čvora u $\Gamma(\mathsf{A})$. Nadalje, kako su u (2.1.3) sve nejednakosti zapravo jednakosti, iz $a_{rr_1} \neq 0$ slijedi $|\mathsf{v}_{r_1}| = 1$. To znači da možemo provesti isto zaključivanje ali s r_1 umjesto r i dobiti $\lambda \in \partial\mathcal{G}_{r_1}$ i (jer je $a_{r_1r_2} \neq 0$) $|v_{r_2}| = 1$. Sada je jasno da na ovaj način, korak po korak, dolazimo do p–tog čvora i zaključka $\lambda \in \partial\mathcal{G}_p$. ⊞

Definicija 2.1.4. Matrica $\mathsf{A} \in \mathbb{M}_n$ je dijagonalno dominantna (po retcima) ako vrijedi $|a_{ii}| \geq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|$, $i = 1, \ldots, n$.

Definicija 2.1.5. Matrica $\mathsf{A} \in \mathbb{M}_n$ je ireducibilno dijagonalno dominantna (po retcima) ako je ireducibilna, dijagonalno dominantna i za barem jedan indeks i vrijedi $|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$.

Korolar 2.1.4. *Ako je* A *ireducibilno dijagonalno dominantna, onda je regularna.*

Dokaz: Dovoljno je dokazati da $0 \notin \mathfrak{S}(\mathsf{A})$. Pretpostavimo da to nije istina i odmah uočimo da zbog dijagonalne dominantnosti nula ne može biti u unutrašnjosti niti jednog Geršgorinovog kruga, pa prema Teoremu 2.1.3 mora biti u presjeku svih kružnica, $\bigcap_{i=1}^{n} \partial\mathcal{G}_i$, tj. $|a_{ii}| = \rho_i$ za sve $i = 1, \ldots, n$. No, to je u kontradikciji sa slabom dijagonalnom dominantnosti jer je $|a_{ii}| > \rho_i$ za barem jedan indeks i. $\boxplus$

2.2 Nenegativne i pozitivne matrice

U dosta primjena se javljaju matrice sa realnim nenegativnim ili pozitivnim elementima, tj. $\mathsf{A} \geq \mathbf{0}$ ($a_{ij} \geq 0$ za sve i, j), ili $\mathsf{A} > \mathbf{0}$ ($a_{ij} > 0$ za sve i, j). Na primjer, Markovljevi lanci u statistici prirodno generiraju matrice sa nenegativnim elementima koji imaju značenje vjerojatnosti. Za nenegativne matrice je razvijen moćan teorijski alat, tzv. Perron–Frobenijusova teorija čiji počeci su u radovima Perrona i Frobenijusa, početkom dvadesetog stoljeća.

Nas ovdje zanimaju spektralna svojstva nenegativnih matrica, a posebno po modulu dominantne svojstvene vrijednosti i pripadni svojstveni vektori.

Propozicija 2.2.1. *Neka je* $\mathsf{A} \in \mathbb{R}^{n\times n}$ *nenegativna matrica. Tada vrijedi*[1]

$$\min_{i=1:n} \sum_{j=1}^{n} a_{ij} \leq \operatorname{spr}(\mathsf{A}) \leq \max_{i=1:n} \sum_{j=1}^{n} a_{ij}. \tag{2.2.1}$$

Nadalje, za svaki vektor $x > \mathbf{0}$ *je*

$$\min_{i=1:n} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij}x_j \leq \operatorname{spr}(\mathsf{A}) \leq \max_{i=1:n} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij}x_j. \tag{2.2.2}$$

Ako je za neki $x > \mathbf{0}$ *i* $\alpha, \beta \geq 0$ *zadovoljeno* $\alpha x \leq \mathsf{A}x \leq \beta x$*, onda je* $\alpha \leq \operatorname{spr}(\mathsf{A}) \leq \beta$*. Također,* $\alpha x < \mathsf{A}x$ *povlači* $\alpha < \operatorname{spr}(\mathsf{A})$*, i* $\mathsf{A}x < \beta x$ *povlači* $\operatorname{spr}(\mathsf{A}) < \beta$*.*

[1] Kako su svojstva nenegativnosti i pozitivnosti invarijantna na transponiranje i kako A i A^{T} imaju isti spektar, dokazane tvrdnje se mogu primijeniti na A^{T} i dobiju se analogni rezultati. Mi, zbog jednostavnosti, nećemo davati odgovarajuće rezultate za transponiranu matricu ali primamo na znanje da vrijede.

Dokaz: Gornja ograda za $\mathrm{spr}(\mathsf{A})$ slijedi iz $\mathrm{spr}(\mathsf{A}) \leq |||\mathsf{A}|||_\infty = \max_{i=1:n} \sum_{j=1}^n a_{ij}$. Stavimo $\mu = \min_{i=1:n} \sum_{j=1}^n a_{ij}$ i definirajmo novu matricu B na sljedeći način: Ako je $\mu = 0$, onda $\mathsf{B} = \mathbf{0}$, inače $b_{ij} = a_{ij}\mu / \sum_{j=1}^n a_{ij}$ za sve i, j. Odmah vidimo da je $\mathbf{0} \leq \mathsf{B} \leq \mathsf{A}$ pa je prema Teoremu 1.5.3 $\mathrm{spr}(\mathsf{B}) \leq \mathrm{spr}(\mathsf{A})$. Nadalje, za svaki redak i je $\sum_{j=1}^n b_{ij} = \mu$ pa je $\mu = |||\mathsf{B}|||_\infty$. Kako je $\mathsf{B}\mathbf{1} = \mu\mathbf{1}$, zaključujemo $\mu = \mathrm{spr}(\mathsf{B}) \leq \mathrm{spr}(\mathsf{A})$. (Ovdje $\mathbf{1}$ označava vektor kojem su svih n komponenti jednake jedinici.) Ovime smo dokazali (2.2.1). Za dokaz nejednakosti (2.2.2) treba od vektora $x > \mathbf{0}$ napraviti dijagonalnu matricu $X = \mathrm{diag}(x)$ i (2.2.1) primijeniti na matricu $X^{-1}\mathsf{A}X$.

Preostale tvrdnje sada lagano slijede. Ako je $\alpha x \leq \mathsf{A}x$, onda dijeljenjem s x_i ($x_i > 0$ po pretpostavci) imamo, koristeći (2.2.2),

$$\alpha \leq \min_{i=1:n} \frac{1}{x_i} \sum_{j=1}^n a_{ij} x_j \leq \mathrm{spr}(\mathsf{A}).$$

Na isti način dobijemo $\beta \geq \max_{i=1:n} \frac{1}{x_i} \sum_{j=1}^n a_{ij} x_j \geq \mathrm{spr}(\mathsf{A})$. Što se tiče strogih nejednakosti, uočimo da za $\alpha x < \mathsf{A}x$ možemo naći $\tilde{\alpha} > \alpha$ tako da je $\tilde{\alpha} x \leq \mathsf{A}x$. Dakle, $\alpha < \tilde{\alpha} \leq \mathrm{spr}(\mathsf{A})$. $\boxplus$

Korolar 2.2.2. *Neka je $\mathsf{A} \in \mathbb{R}^{n\times n}$ nenegativna matrica. Ako A ima pozitivan svojstveni vektor $\mathsf{v} > \mathbf{0}$ onda je pripadna svojstvena vrijednost jednaka* $\mathrm{spr}(\mathsf{A})$.

Dokaz: Uočimo da iz $\mathsf{A}\mathsf{v} = \lambda\mathsf{v}$, $\mathsf{v} > \mathbf{0}$, slijedi da je λ realna i nenegativna svojstvena vrijednosti, a iz $\lambda\mathsf{v} \leq \mathsf{A}\mathsf{v} \leq \lambda\mathsf{v}$ je $\lambda \leq \mathrm{spr}(\mathsf{A}) \leq \lambda$. $\boxplus$

Lema 2.2.3. *Neka je $\mathsf{A} \in \mathbb{R}^{n\times n}$ pozitivna matrica. Ako je $\mathsf{A}\mathsf{x} = \lambda\mathsf{x}$, $x \neq \mathbf{0}$, i $|\lambda| = \mathrm{spr}(\mathsf{A})$ tada postoji $\phi \in \mathbb{R}$ sa kojim je $\mathbf{e}^{\mathrm{i}\phi}\mathsf{x} = |\mathsf{x}| > \mathbf{0}$.*

Sljedeći teorem je centralni rezultat teorije pozitivnih matrica:

Teorem 2.2.4. *(Perron) Neka je $\mathsf{A} \in \mathbb{R}^{n\times n}$ pozitivna matrica. Tada vrijedi:*

- $\mathrm{spr}(\mathsf{A}) > 0$ *i* $\mathrm{spr}(\mathsf{A})$ *je svojstvena vrijednosti od* A *algebarske kratnosti jedan, pri čemu se pripadni svojstveni vektor* v *može odabrati sa realnim pozitivnim komponentama,* $\mathsf{A}\mathsf{v} = \mathrm{spr}(\mathsf{A})\mathsf{v}$, $\mathsf{v} > \mathbf{0}$.

- $\mathrm{spr}(\mathsf{A})$ *je jedina svojstvena vrijednost koja postiže maksimalnu apsolutnu vrijednost, tj.* $(\forall \lambda \in \mathfrak{S}(\mathsf{A}))\ (\lambda \neq \mathrm{spr}(\mathsf{A}) \Longrightarrow |\lambda| < \mathrm{spr}(\mathsf{A}))$.

- *Neka su* $\mathsf{u} > \mathbf{0}$, $\mathsf{v} > \mathbf{0}$ *lijevi i desni svojstveni vektor koji pripadaju* $\mathrm{spr}(\mathsf{A})$. *Tada je* $\lim_{m\to\infty}(\mathsf{A}/\mathrm{spr}(\mathsf{A}))^m = \mathsf{v}\mathsf{u}^{\mathrm{T}}$.

Dokaz: Iz propozicije 2.2.1 odmah slijedi $\mathrm{spr}(\mathsf{A}) > 0$. Neka je sada $\lambda \in \mathfrak{S}(\mathsf{A})$ svojstvena vrijednost za koju je $|\lambda| = \mathrm{spr}(\mathsf{A})$, te neka je $\mathsf{x} \neq \mathbf{0}$ pripadni svojstveni vektor, $\mathsf{A}\mathsf{x} = \lambda\mathsf{x}$. Uzimanjem apsolutne vrijednosti dobijemo $\mathrm{spr}(\mathsf{A})|\mathsf{x}| \leq \mathsf{A}|\mathsf{x}|$, pa je vektor $\mathsf{y} = \mathsf{A}|\mathsf{x}| - \mathrm{spr}(\mathsf{A})|\mathsf{x}|$ nenegativan, $\mathsf{y} \geq \mathbf{0}$. Ako je $\mathsf{y} = \mathbf{0}$, $\mathrm{spr}(\mathsf{A})$ je svojstvena vrijednost i $\mathsf{v} \equiv |\mathsf{x}| \neq \mathbf{0}$ je pripadni svojstveni vektor i $\mathsf{v} = \mathrm{spr}(\mathsf{A})^{-1}\mathsf{A}\mathsf{v} > \mathbf{0}$. Zapravo mora biti $\mathsf{y} = \mathbf{0}$. Ako bi bio $\mathsf{y} \neq \mathbf{0}$, onda bi vrijedilo $\mathsf{A}\mathsf{y} > \mathbf{0}$ i vektor $\mathsf{z} = \mathsf{A}|\mathsf{x}|$ bi zadovoljavao $\mathsf{z} > \mathbf{0}$ pa bismo imali $\mathsf{A}\mathsf{z} - \mathrm{spr}(\mathsf{A})\mathsf{z} > \mathbf{0}$, tj. $\mathsf{A}\mathsf{z} > \mathrm{spr}(A)\mathsf{z}$. No, to bi prema Propoziciji 2.2.1 povlačilo $\mathrm{spr}(\mathsf{A}) > \mathrm{spr}(\mathsf{A})$, kontradikciju.

Neka je sada λ svojstvena vrijednost sa $|\lambda| = \mathrm{spr}(\mathsf{A})$, $\mathsf{A}\mathsf{x} = \lambda\mathsf{x}$, $x \neq \mathbf{0}$. Prema Lemi 2.2.3 je, sa nekim $\phi \in \mathbb{R}$, $\mathbf{e}^{\mathrm{i}\phi}\mathsf{x} = |\mathsf{x}| > \mathbf{0}$, i $\mathsf{A}|\mathsf{x}| = \lambda|\mathsf{x}|$, pa je prema Korolaru 2.2.2 $\lambda = \mathrm{spr}(\mathsf{A})$.

Pokažimo da je $\lambda = \mathrm{spr}(\mathsf{A})$ geometrijske kratnosti jedan. Neka je $\mathsf{v} > \mathbf{0}$ pripadni pozitivni svojstveni vektor i neka je u neki drugi svojstveni vektor koji pripada $\mathrm{spr}(\mathsf{A})$, te neka je $\mathbf{e}^{\mathrm{i}\phi}\mathsf{u} = |\mathsf{u}|$. Definirajmo

$$r = |\mathsf{u}| - \xi\mathsf{v}, \text{ gdje je } \xi = \min_{i=1:n} \frac{|\mathsf{u}_i|}{\mathsf{v}_i}.$$

Lako se uvjerimo da je r svojstveni vektor, $\mathsf{A}r = \mathrm{spr}(\mathsf{A})r$, $r \geq \mathbf{0}$ i da je $r_i = 0$ za barem jedan indeks i. Primijetimo da $r \neq \mathbf{0}$ povlači $\mathsf{A}r > \mathbf{0}$ i $r > \mathbf{0}$ što je kontradikcija. Dakle, $|\mathsf{u}| = \xi\mathsf{v}$, tj. $\mathsf{u} = \mathbf{e}^{-\mathrm{i}\phi}\xi\mathsf{v}$. $\boxplus$

Nenegativnu matricu možemo uvijek napisati kao limes pozitivnih matrica i to otvara prostor da neke rezultate s pozitivnih proširimo na nenegativne matrice. Ipak valja biti oprezan jer neka važna svojstva matrice kao npr. rang ili stroga pozitivnost spektralnog radijusa nisu nužno sačuvana u limesu.

Teorem 2.2.5. *Neka je* $\mathsf{A} \in \mathbb{R}^{n\times n}$ *nenegativna matrica. Tada je* $\mathrm{spr}(\mathsf{A})$ *svojstvena vrijednost od* A *i pripadni svojstveni vektor* v *možemo odabrati da bude nenegativan,* $\mathsf{A}\mathsf{v} = \mathrm{spr}(\mathsf{A})\mathsf{v}$, $\mathsf{v} \geq \mathbf{0}$, $\mathsf{v} \neq \mathbf{0}$.

Dokaz: Uzmimo strogo padajući niz (ϵ_k) strogo pozitivnih brojeva i sa $\lim_{k\to\infty}\epsilon_k = 0$, te definirajmo niz pozitivnih matrica $\mathsf{A}(\epsilon_k) = \mathsf{A} + \epsilon_k\mathbf{1}\cdot\mathbf{1}^{\mathrm{T}}$. Prema prvoj tvrdnji Teorema 2.2.4 možemo pisati, za svaki ϵ_k,

$$\mathsf{A}(\epsilon_k)\mathsf{v}(\epsilon_k) = \mathrm{spr}(\mathsf{A}(\epsilon_k))\mathsf{v}(\epsilon_k), \text{ gdje je } \mathsf{v}(\epsilon_k) > \mathbf{0}, \|\mathsf{v}(\epsilon_k)\|_1 = 1. \tag{2.2.3}$$

Niz svojstvenih vektora je sadržan u kompaktnom skupu (jedinična sfera u normi $\|\cdot\|_1$) pa svakako ima konvergentan podniz, neka je $\lim_{j\to\infty} \mathsf{v}(\epsilon_{k_j}) = \mathsf{v}$. Očito je $\mathsf{v} \geq \mathbf{0}$ i $\|\mathsf{v}\|_1 = 1$. Za odgovarajući podniz (ϵ_{k_j}) vrijedi $\epsilon_{k_j} > \epsilon_{k_{j+1}}$, $j = 1, 2, \ldots$ i $\lim_{j\to\infty} \epsilon_{k_j} = 0$. Kako je $\mathsf{A}(\epsilon_{k_j}) > \mathsf{A}(\epsilon_{k_{j+1}})$, pomoću Teorema 1.5.3 zaključujemo da je $\mathrm{spr}(\mathsf{A}(\epsilon_{k_j})) \geq \mathrm{spr}(\mathsf{A}(\epsilon_{k_{j+1}})) \geq \mathrm{spr}(\mathsf{A})$ pa stoga postoji $\rho = \lim_{j\to\infty} \mathrm{spr}(\mathsf{A}(\epsilon_{k_j}))$ i $\rho \geq \mathrm{spr}(\mathsf{A})$. Ako sada u (2.2.3) uzmemo podniz $j \mapsto k_j$ i onda limes, dobijemo $\mathsf{Av} = \rho\mathsf{v}$, odakle je i $\rho \leq \mathrm{spr}(\mathsf{A})$, tj. $\rho = \mathrm{spr}(\mathsf{A})$. $\boxplus$

Nenegativne matrice tek uz dodatan uvjet ireducibilnosti uživaju neka svojstva pozitivnih matrica.
Sljedeće dvije propozicije su koristan tehnički alat.

Propozicija 2.2.6. *Ako je* B *proizvoljna glavna podmatrica ireducibilne nenegativne matrice* A *onda je* $\mathrm{spr}(\mathsf{B}) < \mathrm{spr}(\mathsf{A})$.

Propozicija 2.2.7. *Neka je* $\chi_{\mathsf{A}}(\lambda) = \det(\lambda\mathsf{I} - \mathsf{A})$ *karakteristični polinom proizvoljne kompleksne* $n \times n$ *matrice* A. *Tada je*

$$\frac{d}{d\lambda}\chi_{\mathsf{A}}(\lambda) = \sum_{i=1}^{n} \det(\lambda\mathsf{I}_{n-1} - \mathsf{A}_{]i,i[}),$$

gdje $\mathsf{A}_{]i,i[}$ *označava* $(n-1)\times(n-1)$ *podmatricu od* A, *dobivenu izbacivanjem* i*–tog retka i* i*–tog stupca.*

Slijedi osnovni teorem za nenegativne ireducibilne matrice.

Teorem 2.2.8. *Neka je* $\mathsf{A} \in \mathbb{R}^{n\times n}$ *nenegativna i ireducibilna matrica. Tada vrijedi:*

- $\mathrm{spr}(\mathsf{A}) > 0$ *i* $\mathrm{spr}(\mathsf{A})$ *je svojstvena vrijednosti od* A *algebarske kratnosti jedan, pri čemu se pripadni svojstveni vektor* v *može odabrati sa realnim pozitivnim komponentama,* $\mathsf{Av} = \mathrm{spr}(\mathsf{A})\mathsf{v}$, $\mathsf{v} > \mathbf{0}$.

- *Spektralni radijus je strogo rastuća funkcija: Ako povećamo bilo koji element od* A *i dobijemo* $\tilde{\mathsf{A}} \geq \mathsf{A}$ *i* $\tilde{\mathsf{A}} \neq \mathsf{A}$, *onda je* $\mathrm{spr}(\tilde{\mathsf{A}}) > \mathrm{spr}(\mathsf{A})$.

- *Ako u spektru od* A *točno* ℓ *svojstvenih vrijednosti ima modul jednak* $\mathrm{spr}(\mathsf{A})$, *onda su one jednake*

$$\mathrm{spr}(\mathsf{A})\mathbf{e}^{2\pi\mathrm{i}j/\ell},\ j = 0, \ldots, \ell - 1.$$

Dokaz: U Teoremu 2.2.5 smo pokazali da je $\mathrm{spr}(\mathsf{A})$ svojstvena vrijednost od A sa pripadnim nenegativnim svojstvenim vektorom v. Budući je A ireducibilna, ne može imati nul redak pa je zbog Propozicije 2.2.1 nužno $\mathrm{spr}(\mathsf{A}) > 0$.

Dokažimo da je $\mathrm{spr}(\mathsf{A})$ algebarski jednostruka svojstvena vrijednost, tj. jednostruka nultočka karakterističnog polinoma. Vratimo se na Propoziciju 2.2.7. Za bilo koji $\mathbb{R} \ni t \geq \mathrm{spr}(\mathsf{A})$ je po sili Propozicije 2.2.6 nužno $\det(t\mathsf{I}_{n-1} - \mathsf{A}_{]i,i[}) \neq 0$. Kako je $\lim_{t\to\infty} \det(t\mathsf{I}_{n-1} - \mathsf{A}_{]i,i[}) = \infty$, mora biti $\det(\mathrm{spr}(\mathsf{A})\mathsf{I}_{n-1} - \mathsf{A}_{]i,i[}) > 0$, pa je $\chi'_{\mathsf{A}}(\mathrm{spr}(\mathsf{A})) > 0$.

Pokažimo strogu monotonost spektralnog radijusa. Neka je $\mathsf{Av} = \mathrm{spr}(\mathsf{A})\mathsf{v}$, $\mathsf{v} > \mathbf{0}$. Iz pretpostavke je $\tilde{\mathsf{A}}\mathsf{v} > \mathsf{Av} = \mathrm{spr}(\mathsf{A})\mathsf{v}$, pa je, prema Propoziciji 2.2.1, $\mathrm{spr}(\tilde{\mathsf{A}}) > \mathrm{spr}(\mathsf{A})$. ⊞

Definicija 2.2.1. Jedinstveni pozitivni svojstveni vektor $\mathsf{v} > 0$, $\sum_{i=1}^{n} \mathsf{v}_i = 1$, koji pripada svojstvenoj vrijednosti $\mathrm{spr}(\mathsf{A})$ nenegativne ireducibilne matrice A se zove Perronov vektor od A.

Definicija 2.2.2. Nenegativna ireducibilna matrica A koja ima samo jednu svojstvenu vrijednost čiji modul je jednak $\mathrm{spr}(\mathsf{A})$ se zove primitivna matrica.

Dio I

Numeričko rješavanje sustava jednadžbi

Poglavlje 3

Iterativne metode za $\mathsf{Ax} = \mathsf{b}$

3.1 Uvod i motivacija

Jedan od osnovnih problema numeričke matematike je rješavanje linearnih sustava jednadžbi. Uovom poglavlju ćemo istraživati metode za rješavanje kvadratnih $n \times n$ sustava, tj. sustava od n jednadžbi sa n nepoznanica,

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

Matrica $\mathsf{A} = \left(a_{ij}\right)_{i,j=1}^{n}$ je *matrica sustava*, a njeni elementi su *koeficijenti sustava*. Vektor $\mathsf{b} = \left(b_i\right)_{i=1}^{n}$ je *vektor desne strane* sustava. Treba odrediti *vektor nepoznanica* $x = \left(x_i\right)_{i=1}^{n}$ tako da vrijedi $\mathsf{Ax} = \mathsf{b}$.
Kako znamo iz linearne algebre, za teorijsku matematiku je rješavanje sustava $\mathsf{Ax} = \mathsf{b}$ gotovo trivijalan problem, posebno u slučaju kada je matrica sustava kvadratna i regularna. Rješenje x je dano formulom $\mathsf{x} = \mathsf{A}^{-1}\mathsf{b}$ u kojoj je A^{-1} inverzna matrica od A ($\mathsf{AA}^{-1} = \mathsf{A}^{-1}\mathsf{A} = \mathsf{I}$). Pri tome postoje eksplicitne formule i za elemente matrice A^{-1} i za samo rješenje x. Osim toga, svima dobro poznata Gaussova metoda eliminacija dolazi do rješenja u $O(n^3)$ elementarnih operacija[1]. Dakle, situacija je potpuno jasna: rješenje $\mathsf{x} = \mathsf{A}^{-1}\mathsf{b}$ postoji, i to samo jedno, i

[1]Ovdje elementarna operacija označava zbrajanje, oduzimanje, množenje ili dijeljenje.

znamo jednostavan algoritam koji to rješenje eksplicitno računa koristeći konačno mnogo samo jednostavnih aritmetičkih operacija.

Na žalost, konačna aritmetika računala niti te jednostavne operacije ne može izvršavati egzaktno, pa Gaussovim eliminacijama, koje su jednostavno konačan niz formula koje vode rješenju, rješenje linearnog sustava $\mathsf{Ax} = \mathsf{b}$ općenito ne možemo izračunati apsolutno točno.

Također, u nekim primjenama niti točno rješenje $\mathsf{x} = \mathsf{A}^{-1}\mathsf{b}$ nije puno bolje od neke dovoljno dobre aproksimacije $\tilde{x}$, gdje obično $\tilde{x}$ zadovoljava sustav $(A+\delta A)\tilde{x} = b + \delta b$, blizak polaznom. Razlog je u činjenici da su u primijenjenim znanostima i inženjerstvu podaci A, b dobiveni mjerenjima fizikalnih veličina (dakle, s neizbježnim pogreškama mjernih instrumenata) ili su zadani formulama (npr. $a_{ij} = \int\limits_{\Omega} \psi_i(x)\psi_j(x)dx$ gdje integral računamo numerički i to u strojnoj aritmetici, dakle s neizbježnom pogreškom) pa je zapravo $A = A_0 + \delta A_0$, $b = b_0 + \delta b_0$, gdje su A_0, b_0 idealni i nama nedostupni podaci a δA_0, δb_0 inicijalna pogreška u podacima. Dakle, čak i kada bismo egzaktno riješili $Ax = b$, imali bismo $(A_0+\delta A_0)x = b_0+\delta b_0$.

Nadalje, u praksi moramo biti svjesni da je stroj (računalo) omeđen ne samo u pitanju numeričke točnosti nego i u još dva važna aspekta: raspoloživi memorijski prostor i vrijeme izvršavanja. Moderne primjene matematike zahtijevaju rješenja sustava velikih dimenzija, npr. $n > 10^5$. Ponekad je to rješenje samo mali dio u kompleksnijem računu i potrebno ga je računati više (stotina ili tisuća) puta, s različitim desnim stranama i/ili različitim matricama koeficijenata.

Lako se uvjeriti da je u takvim primjerima proces Gaussovih eliminacija često praktično neupotrebljiv. Jer, matrica dimenzije $n = 10^5$ zahtijeva $n^2 = 10^{10}$ lokacija u memoriji, svaka barem 4 bajta (veličina reprezentacije realnog broja u jednostrukoj preciznosti). Dakle, moguće je da samo spremanje matrice koeficijenata u memoriju računala predstavalja poteškoću – ponekad matricu držimo na vanjskoj, sporoj memoriji (datoteka na disku) i onda možemo dijelove matrice učitavati dio po dio u radnu memoriju. Same eliminacije trebaju $O(n^3)$ aritmetičkih operacija, ako je $n = 10^5$ i ako možemo računati brzinom npr. 10^9 operacija u sekundi, te ako uračunamo vrijeme transporta podataka iz memorije u procesor i natrag, vidimo da s praktične strane stvar nije baš jednostavna kao što je reći da rješenje postoji, jedinstveno je, zadano je izrazom $x = A^{-1}b$ i postoji konačan algoritam koji ga računa.

U puno važnih primjena je matrica sustava A velike dimenzije, ali je *rijetko popunjena*. To znači da je velika većina elemenata od A jednaka nuli, a elementi koji nisu nula su obično pravilno raspoređeni po matrici ili čak imaju i pravilno raspoređene numeričke vrijednosti. U svakom retku je broj elemenata koji nisu

nula unaprijed poznat (kao i pozicije gdje se ti elementi nalaze) i broj takvih elemenata je puno manji od dimenzije n. To znači da je računanje produkta Av složenosti puno manje od $2n^2 - n$, obično je $O(n)$.
Pogledajmo sada nekoliko primjera.

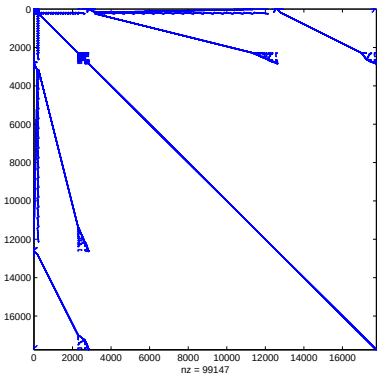
Primjer 3.1.1. Za ilustraciju, pogledajmo kolekciju matrica *Matrix Market* koju na adresi `http://math.nist.gov/MatrixMarket/` nudi američki *National Institute of Standards and Technology* (NIST). Za učitavanje matrica koje su spremljene u određenom formatu trebamo Matlab funkciju `mmread` koja se nalazi u rubrici "Software–Matrix Market I/O – in Matlab". Datoteku `mmread.m` treba spremiti u radni direktorij. Matrice u kolekciji su podijeljene u nekoliko grupa. Čitatelju savjetujemo da malo istraži cijelu kolekciju i tako stekne osjećaj koliko je širok spektar primjena u kojima se pojavljuje problem rješavanja sustava $Ax = b$. Ako odaberemo npr. "Browse–by collection–Harwell-Boeing Collection–SHERMAN–SHERMAN5", kopiramo `sherman5.mtx.gz`, te ju 'unzippamo', dobijemo datoteku `sherman5.mtx`. U Matlabu onda s `A=mmread('sherman5.mtx')` učitamo matricu A koja se pojavljuje u istraživanjima vezanim za eksploataciju nafte. Ako ispišemo varijablu A (`>> A`) vidimo da je spremljena kao lista $((i, j),\ a_{ij})$ elemenata koji nisu nula. Samu strukturu matrice (pozicije netrivijalnih elemenata) možemo vizualizirati s `spy(A)`. Naredbom `[L,U]=lu(A)` dobijemo LU faktorizaciju čiju strukturu vidimo na Slici 3.1.

Slika 3.1: Rijetko popunjena 3312×3312 matrica A =**sherman5** iz kolekcije *Matrix Market*, i njeni trokutasti faktori u LU faktorizaciji. Točkice pokazuju pozicije elemenata u matrici koji su različiti od nule. Uočimo da A ima 20793 netrivijalna elementa, dok ih L i U imaju 409063 i 592791.

Vidimo da samo matrica U ima preko 28 puta više netrivijalnih elemenata, što znači da trebamo toliko puta više memorije za spremiti matricu U kako bismo je mogli iskoristiti u supstitucijama unatrag. Naravno, dimenzija $n = 3312$ je zapravo mala. Preporučamo čitatelju da ovaj pokus ponovi s matricama veće dimenzije. Na primjer, matrica `memplus` iz kolekcije `top ten` je dimenzije $n = 17758$, i naredba `[L,U]=lu(A)` je nakon nekog vremena i dosta buke ventilatora računala vratila

```
>> [L,U]=lu(A);
??? Error using ==> lu Out of memory. Type HELP MEMORY for your
options.
```

Valja reći da niti ovu dimenziju n ne smatramo velikom. S druge strane, ako



Slika 3.2: Rijetko popunjena 17758×17758 matrica A =**memplus** iz kolekcije *Matrix Market*.

u ovom primjeru zadamo desnu stranu b i pokušamo riješiti $Ax = b$ i provjeriti rezidual $\|b - Ax\|_2$, dobijemo jako brzo

```
>> x=A\b;
>> norm(b-A*x)/norm(b)

ans =

  3.7032e-013
```

Dakle, izračunati x egzaktno rješava $Ax = b + \delta b$, gdje je $\|\delta b\|_2/\|b\|_2 \lessapprox 4 \cdot 10^{-13}$. Operacijom `x=A\b` smo bez problema dobili rješenje, dok LU faktorizacija nije izvediva jer nemamo dovoljno memorije za spremiti trokutaste faktore – ako pogledamo profil matrice A ne Slici 3.2 i zamislimo proces Gaussovih eliminacija jasno je da se, već pri eliminaciji elemenata u prvom stupcu, elementi prvog retka 'razmažu' po cijeloj matrici tako da trokutasti faktori u LU faktorizaciji imaju velik broj netrivijalnih elemenata.

3.2 Primjeri

Prije nego počnemo sa opisom i analizom metoda, pogledat ćemo par primjera koji ilustriraju kako nastaje sustav linearnih jednadžbi s posebno strukturiranom matricom koeficijenata. Velika klasa primjera dolazi iz numeričkog rješavanja običnih i parcijalnih diferencijalnih jednadžbi.

Primjer 3.2.1. Promotrimo slijedeći rubni problem:

$$-\frac{d^2}{dx^2}u(x) = f(x), \quad 0 < x < 1, \tag{3.2.1}$$

$$u(0) = \alpha;\ u(1) = \beta. \tag{3.2.2}$$

Rješenje u problema (3.2.1, 3.2.2) ćemo aproksimirati na skupu od konačno mnogo točaka iz segmenta $[0, 1]$. Odaberimo prirodan broj n i definirajmo

$$h = \frac{1}{n+1}, \quad x_i = ih, \quad i = 0, \ldots, n+1. \tag{3.2.3}$$

Sada promatramo vrijednosti $u_i = u(x_i)$, $i = 0, \ldots, n+1$. Iz uvjeta (3.2.2) je odmah $u_0 = \alpha$, $u_{n+1} = \beta$. U unutarnjim čvorovima x_i, $i = 1, \ldots, n$, primjenom Taylorovog teorema je

$$u(x_{i+1}) = u(x_i) + u'(x_i)h + \frac{u''(x_i)}{2}h^2 + \frac{u'''(x_i)}{6}h^3 + \frac{u^{(4)}(\hat{x}_i)}{24}h^4, \tag{3.2.4}$$

$$u(x_{i-1}) = u(x_i) - u'(x_i)h + \frac{u''(x_i)}{2}h^2 - \frac{u'''(x_i)}{6}h^3 + \frac{u^{(4)}(\check{x}_i)}{24}h^4 \tag{3.2.5}$$

Zbrajanjem jednadžbi (3.2.4) i (3.2.5) dobijemo

$$u_{i+1} + u_{i-1} = 2u_i + u''(x_i)h^2 + \underbrace{(u^{(4)}(\hat{x}_i) + u^{(4)}(\check{x}_i))\frac{h^4}{24}}_{\epsilon_i}, \tag{3.2.6}$$

$$\text{pa je } -u''(x_i) = \frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} - \epsilon_i,\ i = 1, \ldots, n. \tag{3.2.7}$$

Dakle, zadana diferencijalna jednadžba u n unutarnjih čvorova mreže glasi

$$\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i + \epsilon_i, \quad i = 1, \ldots, n. \tag{3.2.8}$$

Za $i = 1$ i $i = n$ možemo iskoristiti zadana rubne uvjete pa te dvije jednažbe prelaze u $2u_1 - u_2 = h^2\hat{f}_1 + h^2\epsilon_1$, $-u_{n-1} + 2u_n = h^2\hat{f}_n + h^2\epsilon_n$, $\hat{f}_1 = f_1 + \alpha/h^2$,

$\hat{f}_n = f_n + \beta/h^2$. Matrično to možemo zapisati kao

$$\underbrace{\begin{pmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{pmatrix}}_{\mathsf{T}_n} \underbrace{\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{n-2} \\ u_{n-1} \\ u_n \end{pmatrix}}_{u} = h^2 \underbrace{\begin{pmatrix} \hat{f}_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{n-2} \\ f_{n-1} \\ \hat{f}_n \end{pmatrix}}_{f} + h^2 \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_{n-2} \\ \epsilon_{n-1} \\ \epsilon_n \end{pmatrix}}_{e} \tag{3.2.9}$$

U jednadžbi $\mathsf{T}_n u = h^2 f + h^2 e$ član e ne poznamo pa ga zanemarujemo, tj. pokušat ćemo riješiti $\mathsf{T}_n \tilde{u} = h^2 f$. Osjetljivijim matematičkim dušama se ovo zanemarivanje članova koje ne znamo može s pravom učiniti nedopustivim. Naravno, naknadno treba pokazati da pod određenim uvjetima, za $n \to \infty$, $\tilde{u}$ po volji dobro aproksimira u. Mi ovaj čas nećemo ulaziti u ta razmatranja jer nam je primarni cilj primjerom pokazati kako diskretizacijom diferencijalne jednadžbe nastaje algebarski sustav jednadžbi. I više, upravo smo vidjeli da ćemo rješavati ono što možemo, $\mathsf{T}_n \tilde{u} = h^2 f$, a ne ono što bi bilo idealno, $\mathsf{T}_n u = h^2 f + h^2 e$. *To je česta situacija u primjenama – desna strana sustava s kojom računamo inicijalno sadrži pogreške koje su nastale u fazi formiranja sustava.*

Tek za ilustraciju, napravimo par koraka analize ove situacije, koristeći jednostavan račun. Pomoću §2.1 (Korolar 2.1.4) zaključujemo da je matrica T_n regularna. Neka je $\tilde{u} = h^2 \mathsf{T}_n^{-1} f$. Imamo

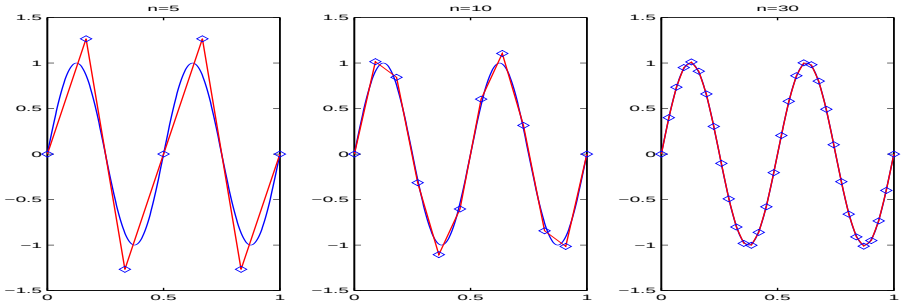
$$u - \tilde{u} = \mathsf{T}_n^{-1}(h^2 f + h^2 e) - h^2 \mathsf{T}_n^{-1} f = h^2 \mathsf{T}_n^{-1} e, \tag{3.2.10}$$

pa je

$$\frac{\|u - \tilde{u}\|_2}{\|\tilde{u}\|_2} \leq \frac{h^2 |||\mathsf{T}_n^{-1}|||_2}{h^2 \|\mathsf{T}_n^{-1} f\|_2} \|e\|_2. \tag{3.2.11}$$

Mi naravno ne možemo točno odrediti niti $\tilde{u}$ jer računamo u aritmetici sa konačnom preciznosti. Neka je $\hat{u}$ izračunata aproksimacija vektora $\tilde{u}$. Pitanje je koliku točnost aproksimacije treba imati $\hat{u}$. Ako je $r = h^2 f - \mathsf{T}_n \hat{u}$ rezidual, onda $\hat{u}$ egzaktno rješava $\mathsf{T}_n \hat{u} = h^2 f - r$, dok idealni u rješava $\mathsf{T}_n u = h^2 f + h^2 e$. Otvara se puno pitanja na koje odgovore daju odgovarajuće teorije u numeričkoj matematici.

Tek da ilustriramo, uzmimo npr. $f(x) = 16\pi^2 \sin 4\pi x$, za kojeg znamo točno rješenje $u(x) = \sin 4\pi x$. Uzet ćemo $n = 5$, $n = 10$ i $n = 30$ i vidjeti koliko su dobre dobivene aproksimacije. Rezultati su dani na Slici 3.3.



Slika 3.3: Diskretne aproksimacije rješenja dobivene s $n = 5$, $n = 10$ i $n = 30$

Primjer 3.2.2. Idemo sada sve ponoviti u dvije dimenzije. Koristimo priliku da pomalo uvodimo uobičajenu notaciju koja je u numeričkom rješavanju diferencijalnih jednadžbi nužna da bi se sve kompliciranije tehničke konstrukcije moglo teorijski analizirati i praktično koristiti.

Promatramo Poissonovu dvodimenzionalnu parcijalnu diferencijalnu jednadžbu

$$\begin{aligned} -u_{xx}(x,y) - u_{yy}(x,y) &= f(x,y), \quad (x,y) \in \Omega \\ u(x,y) &= 0, \quad (x,y) \in \partial\Omega \end{aligned}$$

gdje je $\Omega = \{(x,y) \ : \ 0 < x, y < 1\} \subset \mathbb{R}^2$ jedinični kvadrat, a $\partial\Omega$ njegov rub.

Diskretizaciju konstruiramo na sljedeći način: Odaberimo $n \in \mathbb{N}$ i stavimo

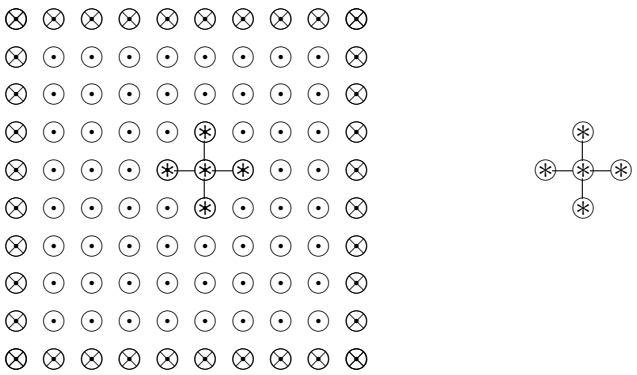
$$\begin{aligned} h &= \frac{1}{n+1}, \quad x_i = ih, \quad y_j = jh, \quad i,j = 0,1,2,\ldots,n+1 \\ \Omega_h &= \{(x_i,y_j) \ : \ i,j = 1,\ldots,n\}, \\ \partial\Omega_h &= \{(x_i,0),(x_i,1),(0,y_j),(1,y_j) \ : \ i,j = 0,1,\ldots,n+1\} \end{aligned}$$

Zamislimo diskretne točke $\Omega_h \bigcup \partial\Omega_h$ bačene kao mreža na zatvoreni kvadrat $\Omega \bigcup \partial\Omega$. Stavimo $u_{ij} = u(x_i,y_j)$, $i,j = 0,1,\ldots,n+1$. Sada $-u_{xx}(x_i,y_j)$ i $-u_{yy}(x_i,y_j)$ lako aproksimiramo centralnim diferencijama.

Koristeći prethodni primjer, za $(x_i,y_j) \in \Omega_h$ je

$$\begin{aligned} -u_{xx}(x_i,y_j) - u_{yy}(x_i,y_j) &= \frac{4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}}{h^2} \\ &+ e_{ij} \end{aligned}$$

gdje je e_{ij} pogreška diskretizacije koju možemo samo ocijeniti s $O(h^2)$. Zato e_{ij} zanemarimo (činimo pogrešku diskretizacije) i, s $f_{ij} = f(x_i,y_j)$, rješavamo sustav



Slika 3.4: $-\triangle u(x_i, y_j) \approx \frac{4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}}{h^2}$.

linearnih jednadžbi

$$4v_{ij} - v_{i-1,j} - v_{i+1,j} - v_{i,j-1} - v_{i,j+1} = h^2 f_{ij}, \quad i, j = 1, \ldots, n \tag{3.2.12}$$

$$v_{0j} = v_{n+1,j} = v_{i0} = v_{i,n+1} = 0 \tag{3.2.13}$$

čije rješenje $V = (v_{ij})_{i,j=1}^{n}$ aproksimira $U = (u_{ij})$. Primijetimo da je prirodno vrijednosti u_{ij}, v_{ij} držati u matrici jer to odražava 2D strukturu. Ipak, za operativno računanje je koristan i generički zapis "$Ax = b$".

Elemente matrice V stavimo u vektor stupac po stupac, isto napravimo s $F = (f_{ij})$:

$$V = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ \vdots & \vdots & \cdots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{pmatrix} \mapsto v = \begin{pmatrix} v_{11} \\ \vdots \\ v_{n1} \\ v_{12} \\ \vdots \\ v_{n2} \\ \vdots \\ v_{1n} \\ \vdots \\ v_{nn} \end{pmatrix} \equiv \mathrm{vec}(V)$$

Ovdje linearni operator $\mathrm{vec}(\cdot)$ poistovjećuje vektorske prostore $\mathbb{R}^{n\times n}$ i $\mathbb{R}^{n^2}$. Sada naš sustav jednadžbi možemo zapisati kao

$$\mathsf{T}_{\otimes n} v = h^2 \mathrm{vec}(F)$$

gdje je

$$\mathsf{T}_{\otimes n} = \left(\begin{array}{cccc|cccc|cccc|cccc}
4 & -1 & & & -1 & & & & & & & & & & & \\
-1 & 4 & \ddots & & & -1 & & & & & & & & & & \\
 & \ddots & \ddots & -1 & & & \ddots & & & & & & & & & \\
 & & -1 & 4 & & & & -1 & & & & & & & & \\
\hline
-1 & & & & 4 & -1 & & & \ddots & & & & & & & \\
 & -1 & & & -1 & 4 & \ddots & & & \ddots & & & & & & \\
 & & \ddots & & & \ddots & \ddots & -1 & & & \ddots & & & & & \\
 & & & -1 & & & -1 & 4 & & & & \ddots & & & & \\
\hline
 & & & & \ddots & & & & \ddots & & & & -1 & & & \\
 & & & & & \ddots & & & & \ddots & & & & -1 & & \\
 & & & & & & \ddots & & & & \ddots & & & & \ddots & \\
 & & & & & & & \ddots & & & & \ddots & & & & -1 \\
\hline
 & & & & & & & & -1 & & & & 4 & -1 & & \\
 & & & & & & & & & -1 & & & -1 & 4 & \ddots & \\
 & & & & & & & & & & \ddots & & & \ddots & \ddots & -1 \\
 & & & & & & & & & & & -1 & & & -1 & 4
\end{array}\right) \tag{3.2.14}$$

Par komentara:

- $\begin{pmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix} = \mathsf{T}_n + 2\mathsf{I}_n$
- $\mathsf{T}_{\otimes n} = \mathsf{I}_n \otimes \mathsf{T}_n + \mathsf{T}_n \otimes \mathsf{I}_n$, gdje je $\otimes$ Kroneckerov produkt.
- Ako uzmemo da korak subdivizije h bude samo reda veličine 10^{-2} trebamo uzeti $n = 100$ i dobivamo $n^2 = 10000$ jednadžbi i isto toliko nepoznanica. Ako je problem trodimenzionalan, onda broj nepoznanica postaje $n^3 = 10^6$.
- U Matlab–u se $\mathsf{T}_{\otimes n}$ dobije pomoću `delsq(numgrid('S',n+2))`.

Definicija 3.2.1. Kažemo da je realna matrica A TST matrica (tridijagonalna simetrična Toeplitzova) ako je tridijagonalna, simetrična i konstantna duž svojih dijagonala,

$$\mathsf{A} = \begin{pmatrix} \alpha & \beta & & \\ \beta & \ddots & \ddots & \\ & \ddots & \ddots & \beta \\ & & \beta & \alpha \end{pmatrix}$$

Propozicija 3.2.1. *Neka je* A *TST matrica reda* n*. Tada su njene svojstvene vrijednosti dane formulama*

$$\lambda_i = \alpha + 2\beta \cos \frac{i\pi}{n+1},\ i = 1, \ldots, n.$$

Pripadni ortonormalni vektori $\mathsf{v}_1, \ldots, \mathsf{v}_n$ *su dani formulama*

$$\mathsf{v}_{ji} = \sqrt{\frac{2}{n+1}} \sin \frac{ji\pi}{n+1},\ \ j = 1, \ldots, n. \tag{3.2.15}$$

(Ovdje v_{ji} *označava* j*–tu komponentu od* v_i*.) Sve TST matrice međusobno komutiraju.*

Korolar 3.2.2. *Svojstvene vrijednosti matrice* T_n *su*

$$\lambda_i = 2(1 - \cos \frac{i\pi}{n+1}),\ i = 1, \ldots, n,$$

a pripadni svojstveni vektori su dani formulama (3.2.15).

Definicija 3.2.2. Matricu $S = \mathsf{I}_m \otimes A + B \otimes \mathsf{I}_n$ zovemo Kroneckerov zbroj matrica $A \in \mathbb{M}_n$ i $B \in \mathbb{M}_n$, u oznaci $S = A \oplus B$.

Propozicija 3.2.3. *Ako su* $Au_j = \alpha_j u_j$*,* $Bv_i = \beta_i v_i$ *svojstvene vrijednosti i vektori, onda su svojstvene vrijednosti od* $A \otimes B$ *svi produkti* $\alpha_i \cdot \beta_j$*, a pripadni svojstveni vektori su* $u_i \otimes v_j$*.*

Dokaz: Koristimo svojstva Kroneckerovog produkta. Lako se provjeri da općenito vrijedi $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. Sada imamo

$$\begin{aligned}(A \otimes B)(u_i \otimes v_j) &= (Au_i) \otimes (Bv_j) = (\alpha_i u_i) \otimes (\beta_j v_j) \\ &= (\alpha_i \beta_j)(u_i \otimes v_j).\end{aligned}$$

⊞

Propozicija 3.2.4. *Ako su* $Au_j = \alpha_j u_j$*,* $Bv_i = \beta_i v_i$ *svojstvene vrijednosti i vektori, onda su svojstvene vrijednosti od* $A \oplus B$ *svi zbrojevi* $\alpha_j + \beta_i$*, a pripadni svojstveni vektori su* $v_i \otimes u_j$*.*

Dokaz:

$$\begin{aligned}&(\mathsf{I} \otimes A + B \otimes \mathsf{I})(v_i \otimes u_j) = (\mathsf{I} \otimes A)(v_i \otimes u_j) + (B \otimes \mathsf{I})(v_i \otimes u_j) \\ &= (v_i \otimes \alpha_j u_j) + (\beta_i v_i \otimes u_j) = \alpha_j v_i \otimes u_j + \beta_i v_i \otimes u_j \\ &= (\alpha_j + \beta_i)(v_i \otimes u_j).\end{aligned}$$

⊞

Korolar 3.2.5. *Svojstvene vrijednosti matrice* $\mathsf{T}_{\otimes n} = \mathsf{I}_n \otimes \mathsf{T}_n + \mathsf{T}_n \otimes \mathsf{I}_n$ *su*

$$\begin{aligned}\lambda_{ij} &= 4 - 2(\cos\frac{i\pi}{n+1} + \cos\frac{j\pi}{n+1}) \\ &= 4(\sin^2\frac{i\pi}{2(n+1)} + \sin^2\frac{j\pi}{2(n+1)}),\ i,j = 1,\ldots,n,\end{aligned}$$

tj. $\lambda_{ij} = \lambda_i + \lambda_j$*, gdje su* $\lambda_1,\ldots,\lambda_n$ *svojstvene vrijednosti matrice* T_n*.*

Dokaz:
⊞

3.3 Gaussove eliminacije i trokutaste faktorizacije

Metoda Gaussovih eliminacija je svakako najstariji, najjednostavniji i najpoznatiji algoritam za rješavanje sustava linearnih jednadžbi $Ax = b$. Ideja je jednostavna: Za riješiti sustav

$$\begin{array}{rcrcl} 2x_1 & - & x_2 & = & 1 \\ -x_1 & + & 2x_2 & = & 1 \end{array}$$

dovoljno je primijetiti da zbog prve jednadžbe vrijedi $x_1 = \frac{1}{2}(1 + x_2)$, pa je druga jednadžba

$$-\underbrace{\frac{1}{2}(1+x_2)}_{x_1} + 2x_2 = 1,\ \text{ tj. }\ \frac{3}{2}x_2 = \frac{3}{2},\ \text{ tj. }\ x_2 = 1,$$

odakle je $x_1 = 1$. Kažemo da smo x_1 *eliminirali* iz druge jednadžbe.
Ova ideja se lako generalizira na dimenziju $n > 1$, gdje sustavno eliminiramo neke nepoznanice iz nekih jednadžbi. Pokazuje se da takav algoritam ima dosta zanimljivu strukturu i da ga se može ekvivalentno zapisati u terminima matričnih operacija. Kvalitativno novi moment u analizi metode eliminacija nastaje kada sam proces eliminacija interpretiramo kao faktorizaciju matrice sustava A na produkt trokutastih matrica.

3.3.1 Matrični zapis metode eliminacija

Primjer 3.3.1. Riješimo sljedeći sustav jednadžbi:

$$\begin{array}{rcrcrcr} 5x_1 & + & x_2 & + & 4x_3 & = & 19 \\ 10x_1 & + & 4x_2 & + & 7x_3 & = & 39 \\ -15x_1 & + & 5x_2 & - & 9x_3 & = & -32 \end{array} \equiv \underbrace{\begin{pmatrix} 5 & 1 & 4 \\ 10 & 4 & 7 \\ -15 & 5 & -9 \end{pmatrix}}_{A=(a_{ij})_{i,j=1}^{3}} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \underbrace{\begin{pmatrix} 19 \\ 39 \\ -32 \end{pmatrix}}_{b=(b_i)_{i=1}^{3}}. \tag{3.3.1}$$

Koristimo metodu supstitucija, odnosno eliminacija: prvo iz prve jednadžbe izrazimo x_1 pomoću x_2 i x_3, te to uvrstimo u zadnje dvije jednadžbe, koje postaju dvije jednadžbe sa dvije nepoznanice (x_2 i x_3). Dobijemo

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

pa druga jednadžba sada glasi

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

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

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

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

$$2x_2 - x_3 = 1.$$

Ako ovu transformaciju sustava zapišemo matrično, imamo

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

Nepoznanicu x_1 eliminiramo iz zadnje jednadžbe ako prvu pomnožimo s

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

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

$$\underbrace{\begin{pmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ -15 & 5 & -9 \end{pmatrix}}_{A^{(1)}} \mapsto \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}}_{L^{(3,1)}} \underbrace{\begin{pmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ -15 & 5 & -9 \end{pmatrix}}_{A^{(1)}} = \underbrace{\begin{pmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ 0 & 8 & 3 \end{pmatrix}}_{A^{(2)} = (a_{ij}^{(2)})_{ij=1}^{3}} .$$

Vektor desne strane je u ove dvije transformacije promijenjen u

$$\underbrace{\begin{pmatrix} 19 \\ 39 \\ -32 \end{pmatrix}}_{b} \mapsto \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{L^{(2,1)}} \begin{pmatrix} 19 \\ 39 \\ -32 \end{pmatrix} = \underbrace{\begin{pmatrix} 19 \\ 1 \\ -32 \end{pmatrix}}_{b^{(1)}} \mapsto \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix}}_{L^{(3,1)}} \begin{pmatrix} 19 \\ 1 \\ -32 \end{pmatrix} = \underbrace{\begin{pmatrix} 19 \\ 1 \\ 25 \end{pmatrix}}_{b^{(2)}} .$$

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

$$\begin{array}{rcrcrcl} 5x_1 & + & x_2 & + & 4x_3 & = & 19 \\ & & 2x_2 & - & 1x_3 & = & 1 \\ & & 8x_2 & - & 3x_3 & = & 25, \end{array} \tag{3.3.2}$$

u kojem su druga i treća jednadžba sustav od dvije jednadžbe sa dvije nepoznanice. Očito je da rješenje $x = (x_1, x_2, x_3)^\tau$ sustava (3.3.1) zadovoljava i sustav (3.3.2). Obratno, ako trojka x_1, x_2, x_3 zadovoljava (3.3.2), onda množenjem prve jednadžbe u (3.3.2) s 2 i zatim pribrajanjem drugoj jednadžbi, dobijemo drugu jednadžbu sustava (3.3.1). Na sličan način iz prve i treće jednadžbe sustava (3.3.2) rekonstruiramo treću jednadžbu polaznog sustava (3.3.1). U tom smislu kažemo da sus sustavi (3.3.1) i (3.3.2) ekvivalentni: imaju isto rješenje.

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

Jasno je da je sustav (3.3.2) jednostavniji od polaznog. Zato sada nastavljamo s primjenom iste strategije: iz treće jednadžbe eliminiramo x_2 tako što drugu jednadžbu pomnožimo s

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

i pribrojimo je trećoj. Tako treća jednadžba postaje

$$7x_3 = 21,$$

a cijeli sustav ima oblik

$$\begin{array}{rcrcrcl} 5x_1 & + & x_2 & + & 4x_3 & = & 19 \\ & & 2x_2 & - & x_3 & = & 1\,. \\ & & & & 7x_3 & = & 21 \end{array} \tag{3.3.3}$$

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

$$\underbrace{\begin{pmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ 0 & 8 & 3 \end{pmatrix}}_{A^{(2)}} \mapsto \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -4 & 1 \end{pmatrix}}_{L^{(3,2)}} \underbrace{\begin{pmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ 0 & 8 & 3 \end{pmatrix}}_{A^{(2)}} = \underbrace{\begin{pmatrix} 5 & 1 & 4 \\ 0 & 2 & -1 \\ 0 & 0 & 7 \end{pmatrix}}_{A^{(3)}} \tag{3.3.4}$$

i transformaciju vektora desne strane

$$\underbrace{\begin{pmatrix} 19 \\ 1 \\ 25 \end{pmatrix}}_{b^{(2)}} \mapsto \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -4 & 1 \end{pmatrix}}_{L^{(3,2)}} \underbrace{\begin{pmatrix} 19 \\ 1 \\ 25 \end{pmatrix}}_{b^{(2)}} = \underbrace{\begin{pmatrix} 19 \\ 1 \\ 21 \end{pmatrix}}_{b^{(3)} = (b_i^{(3)})_{i=1}^3} \tag{3.3.5}$$

Sustav (3.3.3), koji je ekvivalentan polaznom, lako riješimo:

1. Iz treće jednadžbe je $x_3 = \dfrac{21}{7} = 3$;
2. Iz druge jednadžbe je $x_2 = \dfrac{1}{2}(1 + x_3) = 2$;
3. Iz prve jednadžbe je $x_1 = \dfrac{1}{5}(19 - x_2 - 4x_3) = 1.$

Jednostavna provjera potvrđuje da je sa ovim x_1, x_2, x_3 riješen polazni sustav (3.3.1).

Analizirajmo postupak rješavanja u prethodnom primjeru. Relacija

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

zaslužuje posebnu pažnju. Matrica $A^{(3)}$ je gornje trokutasta, a produkt $L^{(3,2)}L^{(3,1)}L^{(2,1)}$ je donje trokutasta matrica. Dakle, polaznu matricu A smo množenjem s lijeva donje trokutastom matricom načinili gornje trokutastom. To možemo pročitati i ovako:

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

gdje je L donje trokutasta matrica. Lako provjerimo da je

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

Dakle, matricu A smo napisali kao produkt donje trokutaste i gornje trokutaste matrice, $A = LA^{(3)}$. Gornje trokutastu matricu u ovom kontekstu obično označavamo s $U = A^{(3)}$, pa je A rastavljen na produkt $A = LU$. Govorimo o LU *faktorizaciji* matrice A. Uočimo da je računanje produkta koji definira matricu L jednostavno: Inverze od $L^{(2,1)}$, $L^{(3,1)}$, $L^{(3,2)}$ dobijemo samo promjenom predznaka netrivijalnih elemenata u donjem trokutu, a cijeli produkt je jednostavno stavljanje tih elemenata na odgovarajuće pozicije u matrici L. Sada još primijetimo da je relacija (3.3.3) zapravo linearni sustav

$$Ux = b^{(3)}, \;\; \text{gdje je} \;\; b^{(3)} = L^{(3,2)}L^{(3,1)}L^{(2,1)}b = L^{-1}b.$$

Jasno, $x = A^{-1}b = (LU)^{-1}b = U^{-1}L^{-1}b$.
Dakle, u terminima matrice A i vektora b, linearni sustav u primjeru 3.3.1 je riješen metodom koja se sastoji od tri glavna koraka:

1. *Matricu sustava A faktorizirati u obliku $A = LU$, gdje je L donje trokutasta, a U gornje trokutasta matrica.*

2. *Rješavanjem donje trokutastog sustava $Ly = b$ odrediti vektor $y = L^{-1}b$.*

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

Ovakav zapis metode opisane u primjeru 3.3.1 ima niz prednosti:

– Operacije su iskazane u terminima matrice A i desne strane b, a ne u terminima izražavanja neke nepoznanice pomoću ostalih. Umjesto "x_1 izrazimo

pomoću $x_2, x_3, \ldots$" i sl., operacije izražavamo jezikom operacija sa matricama i vektorima. To omogućuje jednostavnu i sustavnu primjenu opisane metode na sustav sa proizvoljnim brojem nepoznanica. Sam linearni sustav je u računalu pohranjen kao matrica koeficijenata A i vektor desne strane b. Dakle, ovakav zapis metode eliminacija je prirodan.

– Ponekad u primjenama rješavamo nekoliko linearnih sustava sa istom matricom A, ali sa nizom različitih desnih strana b. Vidimo da je u tom slučaju transformacije na matrici A dovoljno napraviti jednom (prvi korak u gornjem zapisu metode), a zatim za različite desne strane provesti samo zadnja dva koraka.

3.3.2 Trokutasti sustavi: rješavanje supstitucijama naprijed i unazad

Trokutasti sustavi jednadžbi su laki za riješiti. Pogledajmo na primjer donje trokutasti sustav $Lx = b$ dimenzije $n = 4$:

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

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

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

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

Algoritam 3.3.1. Rješavanje linearnog sustava jednadžbi $Lx = b$ sa regularnom donje trokutastom matricom $L \in \mathbf{R}^{n\times n}$.

/* Supstitucije naprijed za $Lx = b$ */
$x_1 = \dfrac{b_1}{\ell_{11}}$;
za $i = 2, \ldots, n$ {
$$x_i = \frac{1}{\ell_{ii}}(b_i - \sum_{j=1}^{i-1} \ell_{ij}x_j) \text{ ; }\}$$

Prebrojimo operacije u gornjem algoritmu:

- Dijeljenja: n ;
- Množenja: $1 + 2 + \cdots + (n-1) = \dfrac{1}{2}n(n-1)$;
- Zbrajanja i oduzimanja: $1 + 2 + \cdots + (n-1) = \dfrac{1}{2}n(n-1)$.

Dakle, ukupna složenost je $O(n^2)$.
Gornje trokutaste sustave rješavamo na sličan način. Ako je sustav $Ux = b$ oblika

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

onda, polazeći od zadnje jednadžbe unazad, imamo

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

Ovakav postupak zovemo *supstitucije unazad*.

Algoritam 3.3.2. Rješavanje linearnog sustava jednadžbi $Ux = b$ sa regularnom gornje trokutastom matricom $U \in \mathbf{R}^{n\times n}$.

/* Supstitucije unazad za $Ux = b$ */
$x_n = \dfrac{b_n}{u_{nn}}$;
za $i = n-1, \ldots, 1$ {
$$x_i = \frac{1}{u_{ii}}(b_i - \sum_{j=i+1}^{n} u_{ij}x_j) \ ; \}$$

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

3.3.3 LU faktorizacija

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

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

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

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

[2] U primjeru 3.3.1 smo zbog jednostavnosti poništavali koeficijente jedan po jedan.

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

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

Objasnimo oznake koje koristimo za elemente matrice $A^{(1)}$. Općenito, elementi od $A^{(1)}$ su označeni s $a_{ij}^{(1)}$, $1 \leq i, j \leq n$. Međutim, elementi prvog retka u $A^{(1)}$ su jednaki prvom retku u A, $a_{1j}^{(1)} = a_{1j}$, $1 \leq j \leq n$, pa smo to eksplicitno naznačili u zapisu matrice $A^{(1)}$.
Primijetimo da je transformaciju $A \mapsto A^{(1)}$ moguće izvesti samo ako je

$$a_{11} \neq 0. \tag{3.3.6}$$

Također, lako se uvjerimo da je

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

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

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{a_{21}}{a_{11}} & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{22}^{(1)} \end{pmatrix}.$$

Jednostavno, dobili smo faktorizaciju vodeće 2×2 podmatrice od A. Uvjet za izvod ove faktorizacije je bio (3.3.6). Stavimo

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

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

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

Vrijedi

$$A^{(2)} \equiv L^{(2)}A^{(1)} = L^{(2)}L^{(1)}A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} & a_{45}^{(2)} \\ 0 & 0 & a_{53}^{(2)} & a_{54}^{(2)} & a_{55}^{(2)} \end{pmatrix}. \tag{3.3.7}$$

(Uočimo da oznake u relaciji (3.3.7) naglašavaju da je u matrici $A^{(2)} = \left(a_{ij}^{(2)}\right)_{i,j=1}^{n}$ prvi redak jednak prvom retku od A, a drugi redak jednak drugom retku matrice $A^{(1)}$.) Ako sada u relaciji $A = (L^{(1)})^{-1}(L^{(2)})^{-1}A^{(2)}$ izračunamo produkt $(L^{(1)})^{-1}(L^{(2)})^{-1}$ dobijemo

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \dfrac{a_{21}}{a_{11}} & 1 & 0 & 0 & 0 \\ \dfrac{a_{31}}{a_{11}} & \dfrac{a_{32}^{(1)}}{a_{22}^{(1)}} & 1 & 0 & 0 \\ \dfrac{a_{41}}{a_{11}} & \dfrac{a_{42}^{(1)}}{a_{22}^{(1)}} & 0 & 1 & 0 \\ \dfrac{a_{51}}{a_{11}} & \dfrac{a_{52}^{(1)}}{a_{22}^{(1)}} & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & a_{44}^{(2)} & a_{45}^{(2)} \\ 0 & 0 & a_{53}^{(2)} & a_{54}^{(2)} & a_{55}^{(2)} \end{pmatrix}, \tag{3.3.8}$$

odakle zaključujemo da vrijedi

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

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

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

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

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

i vrijedi

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

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

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

te da je

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

Ponovo zaključujemo na isti način: definiramo

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

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

$$L^{(4)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -\dfrac{a_{54}^{(3)}}{a_{44}^{(3)}} & 1 \end{pmatrix}, \quad (L^{(4)})^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \dfrac{a_{54}^{(3)}}{a_{44}^{(3)}} & 1 \end{pmatrix}. \tag{3.3.9}$$

Lako provjerimo da vrijedi

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

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

$$A = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{a_{21}}{a_{11}} & 1 & 0 & 0 & 0 \\ \frac{a_{31}}{a_{11}} & \frac{a_{32}^{(1)}}{a_{22}^{(1)}} & 1 & 0 & 0 \\ \frac{a_{41}}{a_{11}} & \frac{a_{42}^{(1)}}{a_{22}^{(1)}} & \frac{a_{43}^{(2)}}{a_{33}^{(2)}} & 1 & 0 \\ \frac{a_{51}}{a_{11}} & \frac{a_{52}^{(1)}}{a_{22}^{(1)}} & \frac{a_{53}^{(2)}}{a_{33}^{(2)}} & \frac{a_{54}^{(3)}}{a_{44}^{(3)}} & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & a_{25}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & a_{35}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} & a_{45}^{(3)} \\ 0 & 0 & 0 & 0 & a_{55}^{(4)} \end{pmatrix}}_{U}. \tag{3.3.10}$$

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

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

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

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

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

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

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

Algoritam 3.3.3. Računanje LU faktorizacije matrice A.

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

Sljedeći teorem i formalno dokazuje egzistenciju i jedinstvenost LU faktorizacije.

Teorem 3.3.1. *Neka je* $A \in \boldsymbol{R}^{n\times n}$ *i neka su determinante glavnih podmatrica* $A(1:k, 1:k)$ *različite od nule za* $k = 1, 2, \ldots, n-1$. *Tada postoji donje trokutasta matrica* L *sa jedinicama na dijagonali i postoji gornje trokutasta matrica* U, *tako da vrijedi* $A = LU$. *Ako takva faktorizacija* $A = LU$ *postoji i ako je još i matrica* A *regularna, onda je faktorizacija jedinstvena: postoji točno jedna matrica* L *i točno jedna matrica* U *sa ovim svojstvima. Tada je i* $\det(A) = \prod_{i=1}^{n} u_{ii}$.

Dokaz: Dokažimo prvo jedinstvenost LU faktorizacije. Neka postoje dvije takve faktorizacije,

$$A = LU = L'U'.$$

Ako je A regularna onda su i L, U, L', U' također regularne matrice pa vrijedi

$$L^{-1}L' = U(U')^{-1}$$

U gornjoj relaciji imamo jednakost donje trokutaste i gornje trokutaste matrice – znači da na obe strane jednakosti stoje dijagonalne matrice. Nadalje, L i L' po pretpostavci imaju jedinice na dijagonali, a zbog činjenice da se na dijagonali produkta donje trokutastih matrica nalaze produkti dijagonalinih elemenata matrica koje se množe su na dijagonali od $L^{-1}L'$ jedinice. Dakle, $L^{-1}L' = I$, tj. $L = L'$. Tada je i $U = U'$.
Dokažimo sada egzistenciju LU faktorizacije. Induktivni dokaz je zapravo već skiciran u opisu računanja faktorizacije 5×5 matrice. Pogledajmo kako uvjeti

teorema omogućuju prelaz sa $A^{(k)}$ na $A^{(k+1)}$, gdje je

$$A^{(k)} = L^{(k)} \cdots L^{(1)} A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \cdots & \cdots & a_{1,n-1} & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & \cdots & \cdots & a_{2,n-1}^{(1)} & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & \cdots & \cdots & a_{3,n-1}^{(2)} & a_{3n}^{(2)} \\ 0 & 0 & 0 & \ddots & \cdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & & \ddots & a_{kk}^{(k-1)} & a_{k,k+1}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ 0 & 0 & & & 0 & a_{k+1,k+1}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ \vdots & \vdots & & & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & & 0 & a_{n,k+1}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}.$$

Kako je produkt $(L^{(k)} \cdots L^{(1)})^{-1}$ donje trokutasta matrica sa jedinicama na dijagonali, zaključujemo da je

$$\det(A(1:k+1, 1:k+1)) = a_{11} a_{22}^{(1)} a_{33}^{(2)} \cdots a_{kk}^{(k-1)} a_{k+1,k+1}^{(k)} \neq 0.$$

Odavde je i $a_{k+1,k+1}^{(k)} \neq 0$ pa možemo definirati matricu $L^{(k+1)}$ koja će poništiti elemente ispod dijagonale u $(k+1)$–om stupcu i dati $A^{(k+1)} = L^{(k+1)} A^{(k)}$. Jasno je da nakon konačno koraka dobijemo matricu $A^{(n-1)}$ koja je gornje trokutasta. □

Komentar 3.3.1. Primijetimo, ako je A regularna i ako ima LU faktorizaciju, onda su nužno i sve glavne podmatrice $A(1:k, 1:k)$ regularne. To slijedi iz činjenice da je

$$\det(A(1:k, 1:k)) = \prod_{i=1}^{k} u_{ii}, \quad k = 1, \ldots, n.$$

3.3.4 LU faktorizacija sa pivotiranjem

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

Primjer 3.3.2. Neka je matrica sustava $Ax = b$ dana s

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

Ova matrica je regularna, $\det(A) = -1$, pa sustav uvijek ima rješenje, ali A očito nema LU faktorizaciju. Jer,

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

povlači da je $\ell_{11}u_{11} = 0$, pa je $\ell_{11} = 0$ ili $u_{11} = 0$, a tada je $1 = \ell_{11}u_{12} = 0$ ili $1 = \ell_{21}u_{11} = 0$.
S druge strane, matrica A reprezentira linearni sustav

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

koji uvijek ima rješenje $x_1 = b_2 - b_1$, $x_2 = b_1$, i kojeg možemo ekvivalento zapisati kao[3]

$$\begin{array}{rcrcl} x_1 & + & x_2 & = & b_2 \\ 0x_1 & + & x_2 & = & b_1 \end{array}.$$

Matrica ovog sustava je

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

i očito ima jednostavnu LU faktorizaciju sa $L = I$, $U = A'$. Vezu između A i A' lako opišemo matrično:

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

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

Da bismo ilustrirali kako zamjenama redaka uvijek možemo dobiti LU faktorizaciju, vratimo se našem 5×5 primjeru i pogledajmo npr. relacije (3.3.7), (3.3.8):

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

[3]Zamjena redoslijeda jednadžbi ne mijenja sustav.

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

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

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

Sada možemo definirati matrice

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

i postići

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

Primijetimo da je treći redak od $A^{(3)}$ jednak petom retku od $A^{(2)}$. Za sljedeći korak eliminacija provjeravamo vrijednost $a_{44}^{(3)}$. Ako je $a_{44}^{(3)} \neq 0$, postupamo kao i ranije, tj. definiramo matricu $L^{(4)}$ kao u relaciji (3.3.9). Ako je $a_{44}^{(3)} = a_{54}^{(3)} = 0$, onda možemo staviti $L^{(4)} = I$, jer je u tom slučaju $A^{(3)}$ već gornje trokutasta. Neka je $a_{44}^{(3)} = 0$, ali $a_{54}^{(3)} \neq 0$, tako da $L^{(4)}$ nije definirana. Lako provjerimo da permutacijska matrica

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

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

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

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

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

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

Dakle, $P^{(4)}$ možemo *prebaciti s lijeve na desnu stranu od* $L^{(3)}$, ako u $L^{(3)}$ ispermutiramo elemente ispod dijagonale u trećem stupcu. Tako dobivena matrica $\tilde{L}^{(3)}$ ima istu strukturu kao i $L^{(3)}$. Na isti način je $P^{(3)}L^{(2)}L^{(1)} = \tilde{L}^{(2)}\tilde{L}^{(1)}P^{(3)}$ i $P^{(4)}\tilde{L}^{(2)}\tilde{L}^{(1)} = \tilde{\tilde{L}}^{(2)}\tilde{\tilde{L}}^{(1)}P^{(4)}$ pa je

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

tj.

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

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

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

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

$$U = A^{(n-1)} = L^{(n-1)}P^{(n-1)}(\cdots(\underbrace{L^{(3)}P^{(3)}\,\underbrace{(L^{(2)}P^{(2)}\,\underbrace{(L^{(1)}P^{(1)}A)}_{A^{(1)}})}_{A^{(2)}}}_{A^{(3)}})\cdots), \tag{3.3.11}$$

i $P = P^{(n-1)}P^{(n-2)}\cdots P^{(2)}P^{(1)}$, gdje neke od permutacija $P^{(k)}$ mogu biti jednake identitetama (jediničnim matricama).
Ilustrirajmo opisanu proceduru jednim numeričkim primjerom.

Primjer 3.3.3. Neka je

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

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

permutacija $P^{(1)}$, gdje je

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

Sada definiramo

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

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

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

Sljedeći korak eliminacija glasi

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

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

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

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

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

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

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

Dakle, možemo zaključiti sljedeće:

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

Preciznije, vrijedi sljedeći teorem :

Teorem 3.3.2. *Neka je $A \in \boldsymbol{R}^{n\times n}$ proizvoljna matrica. Tada postoji permutacija P takva da Gaussove eliminacije daju LU faktorizaciju $PA = LU$ matrice PA. Matrica $L = \left(\ell_{ij}\right)$ je donje trokutasta sa jedinicama na dijagonali, a U je gornje trokutasta. Pri tome, ako je P produkt od p inverzija, vrijedi da je* $\det(A) = (-1)^p \prod_{i=1}^{n} u_{ii}$.

Ako su matrice $P^{(k)}$ odabrane tako da vrijedi

$$|(P^{(k)}A^{(k-1)})_{kk}| = \max_{k\le j\le n} |(P^{(k)}A^{(k-1)})_{jk}|$$

onda je

$$\max_{1\le k\le n} \max_{1\le i,j\le n} |(L^{(k)})_{ij}| = \max_{1\le i,j\le n} |\ell_{ij}| = 1.$$

U tom slučaju faktorizaciju $PA = LU$ zovemo LU faktorizacijom sa (standardnim) pivotiranjem redaka.

Dokaz: Za početak, primijetimo da za matricu

$$L^{(k)} = \begin{pmatrix} I_k & 0 \\ 0 \quad v & I_{n-k} \end{pmatrix}, \quad v = \begin{pmatrix} \ell_{k+1,k}^{(k)} \\ \vdots \\ \ell_{nk}^{(k)} \end{pmatrix}$$

i permutaciju $\Pi \in \mathcal{S}_n$ oblika

$$\Pi = \begin{pmatrix} I_k & 0 \\ 0 & \hat{\Pi}_{n-k} \end{pmatrix}, \quad \hat{\Pi} \in \mathcal{S}_{n-k}$$

vrijedi

$$\Pi L^{(k)} = \begin{pmatrix} I_k & 0 \\ 0 \quad \hat{\Pi}_{n-k} v & \hat{\Pi}_{n-k} \end{pmatrix} = \begin{pmatrix} I_k & 0 \\ 0 \quad \hat{\Pi}_{n-k} v & I_{n-k} \end{pmatrix} \Pi = \tilde{L}^{(k)} \Pi.$$

Nadalje, svaka permutacija Π oblika

$$\Pi = \begin{pmatrix} I_m & 0 \\ 0 & \hat{\Pi}_{n-m} \end{pmatrix}, \quad m > k, \quad \hat{\Pi} \in \mathcal{S}_{n-m}$$

je trivijalno i oblika

$$\Pi = \begin{pmatrix} I_k & 0 \\ 0 & \tilde{\Pi}_{n-k} \end{pmatrix}, \quad \tilde{\Pi}_{n-k} = \begin{pmatrix} I_{m-k} & 0 \\ 0 & \hat{\Pi}_{n-m} \end{pmatrix} \in \mathcal{S}_{n-k},$$

pa je množenje analogno slučaju $m = k$. Kratko kažemo da "Π prolazi kroz $L^{(k)}$". Nadalje, jasno je da u svakom koraku možemo odrediti permutaciju $P^{(k)}$ tako da postoji donje trokutasta transformacija $L^{(k)}$ sa jedinicama na dijagonali za koju $L^{(k)} P^{(k)} A^{(k-1)}$ ima sve nule ispod dijagonale u k–tom stupcu. Dakle, kao u relaciji (3.3.11), možemo postići da je $U = A^{(n-1)}$ gornje trokutasta matrica. U produktu

$$U = L^{(n-1)} P^{(n-1)} L^{(n-2)} P^{(n-2)} L^{(n-3)} P^{(n-3)} L^{(n-4)} P^{(n-4)} \cdots L^{(2)} P^{(2)} L^{(1)} P^{(1)} A$$

je $P^{(k+1)}$ oblika

$$P^{(k+1)} = \begin{pmatrix} I_k & 0 \\ 0 & \hat{\Pi}_{n-k} \end{pmatrix}, \quad \hat{\Pi} \in \mathcal{S}_{n-k},$$

što znači da $P^{(n-1)}$ prolazi kroz $L^{(n-2)}$, produkt $P^{(n-1)} P^{(n-2)}$ prolazi kroz $L^{(n-3)}$, produkt $P^{(n-1)} P^{(n-2)} P^{(n-3)}$ prolazi kroz $L^{(n-4)}$ itd.
Ako stavimo $P = P^{(n-1)} P^{(n-2)} \cdots P^{(2)} P^{(1)}$, onda je

$$U = \tilde{L}^{(n-1)} \tilde{L}^{(n-2)} \cdots \tilde{L}^{(2)} \tilde{L}^{(1)} P A,$$

odakle kao i ranije dobijemo $PA = LU$. Jasno je da strategija odabira permutacija iz iskaza teorema osigurava da su svi elementi od L po apsolutnoj vrijednosti najviše jednaki jedinici. □

3.3.5 Numerička svojstva Gaussovih eliminacija

U prethodnim sekcijama smo se Gaussovim eliminacijama bavili u okvirima linearne algebre. Preciznije, nismo razmatrali praktične detalje realizacije izvedenih algoritama na računalu. Zapravo, termin *praktični detalji* bi trebalo čitati kao **problemi**. Zašto?
Računalo je ograničen, konačan stroj. Imamo ograničenu količinu memorijskog prostora u kojem možemo držati polazne podatke, međurezultate i rezultate računanja.[4] Umjesto skupa realnih brojeva $\mathbf{R}$ imamo njegovu aproksimaciju pomoću konačno mnogo strojnih brojeva (strojni brojevi su zapravo konačan skup razlomaka) što znači da računske operacije ne možemo izvršavati niti točno niti rezultat možemo po volji dobro aproksimirati.
Za one čitatelje koji nisu svladali osnove numeričkih operacija linearne algebre na računalu, kao i za one koji taj materijal žele ponoviti, osnovne činjenice su dane u dodatku u sekciji 3.3.9. Preporučamo da čitatelj svakako baci pogled na tu sekciju prije nastavka čitanja ovog materijala.
Praktično je odvojeno analizirati LU faktorizaciju i rješenje trokutastog sustava. Počinjemo sa LU faktorizacijom, gdje nas očekuje niz zanimljivih zaključaka.

3.3.6 Analiza LU faktorizacije. Važnost pivotiranja.

Prije nego pređemo na numeričku analizu algoritma, pogledajmo kako ga možemo implementirati na računalu s minimalnim korištenjem dodatnog memorijskog prostora. Prisjetimo se našeg 5×5 primjera i relacije (3.3.10):

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

[4] Svaka operacija zahtijeva izvjesno vrijeme izvršavanja pa je ukupno trajanje algoritma također važan faktor. U ovoj sekciji ćemo prvenstveno analizirati problem točnosti.

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

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

Sve matrice $A^{(k)}$, $k = 1, 2, \ldots, n-1$ su pohranjene u istom $n \times n$ polju koje na početku sadrži matricu $A \equiv A^{(0)}$. Na ovaj način zapis algoritma 3.3.3 postaje još jednostavniji i elegantniji.

Algoritam 3.3.4. Računanje LU faktorizacije matrice A bez dodatne memorije.

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

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

greške zaokruživanja.

Analiza grešaka zaokruživanja je ponekad tehnički kompliciranа. Ono što je važno uočiti je da cilj takve analize nije jednostavno tehničko prebrojavanje svih grešaka zaokruživanja nego izvođenje složenijih i dubljih zaključaka o numeričkoj stabilnosti algoritma i o pouzdanosti korištenja dobivenih rezultata.

Da bismo dobili ideju o kvaliteti izračunate faktorizacije, analizirat ćemo primjer faktorizacije 4×4 matrice

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}.$$

Izračunate aproksimacije matrica $L = \big(\ell_{ij}\big)$ i $U = \big(u_{ij}\big)$ ćemo označiti s $\tilde{L} = \big(\tilde{\ell}_{ij}\big)$ i $\tilde{U} = \big(\tilde{u}_{ij}\big)$. Kao u opisu algoritam za računanje LU faktorizacije u sekciji 3.3.3, koristit ćemo matrice $L^{(i)}$ i transformacije oblika $A^{(i)} = L^{(i)}A^{(i-1)}$, $i = 1, \ldots, n-1$. Izračunate aproksimacije označavamo s $\tilde{L}^{(i)}$ i $\tilde{A}^{(i)}$.
Primijetimo da je prvi redak matrice $\tilde{U}$ jednak prvom retku polazne matrice A,

$$\tilde{U} = \begin{pmatrix} \tilde{u}_{11} & \tilde{u}_{12} & \tilde{u}_{13} & \tilde{u}_{14} \\ 0 & \tilde{u}_{22} & \tilde{u}_{23} & \tilde{u}_{24} \\ 0 & 0 & \tilde{u}_{33} & \tilde{u}_{34} \\ 0 & 0 & 0 & \tilde{u}_{44} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & \tilde{u}_{22} & \tilde{u}_{23} & \tilde{u}_{24} \\ 0 & 0 & \tilde{u}_{33} & \tilde{u}_{34} \\ 0 & 0 & 0 & \tilde{u}_{44} \end{pmatrix}, \quad \tilde{u}_{1j} = a_{1j}, \quad j = 1, \ldots, 4.$$

Sada umjesto matrica $L^{(1)}$ i $A^{(1)} = L^{(1)}A$ imamo izračunate matrice

$$\begin{aligned} \tilde{L}^{(1)} &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\tilde{\ell}_{21} & 1 & 0 & 0 \\ -\tilde{\ell}_{31} & 0 & 1 & 0 \\ -\tilde{\ell}_{41} & 0 & 0 & 1 \end{pmatrix} \\ \tilde{A}^{(1)} &= \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22} \ominus \tilde{\ell}_{21} \odot \tilde{u}_{12} & a_{23} \ominus \tilde{\ell}_{21} \odot \tilde{u}_{13} & a_{24} \ominus \tilde{\ell}_{21} \odot \tilde{u}_{14} \\ 0 & a_{32} \ominus \tilde{\ell}_{31} \odot \tilde{u}_{12} & a_{33} \ominus \tilde{\ell}_{31} \odot \tilde{u}_{13} & a_{34} \ominus \tilde{\ell}_{31} \odot \tilde{u}_{14} \\ 0 & a_{42} \ominus \tilde{\ell}_{41} \odot \tilde{u}_{12} & a_{43} \ominus \tilde{\ell}_{41} \odot \tilde{u}_{13} & a_{44} \ominus \tilde{\ell}_{41} \odot \tilde{u}_{14} \end{pmatrix} \\ &= \begin{pmatrix} \tilde{u}_{11} & \tilde{u}_{12} & \tilde{u}_{13} & \tilde{u}_{14} \\ 0 & \tilde{u}_{22} & \tilde{u}_{23} & \tilde{u}_{24} \\ 0 & \star & \star & \star \\ 0 & \star & \star & \star \end{pmatrix}, \quad \tilde{u}_{2j} = a_{2j} \ominus \tilde{\ell}_{21} \odot \tilde{u}_{1j}, \quad j = 2, 3, 4. \end{aligned}$$

Ovdje smo sa $\star$ označili one elemente koje ćemo mijenjati u sljedećem koraku. Primijetimo da su u prva dva retka matrice $\tilde{A}^{(1)}$ već izračunata prva dva retka matrice $\tilde{U}$. U sljedećem koraku računamo

$$\begin{aligned}
\tilde{L}^{(2)} &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -\tilde{\ell}_{32} & 1 & 0 \\ 0 & -\tilde{\ell}_{42} & 0 & 1 \end{pmatrix} \\
\tilde{A}^{(2)} &= \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & \tilde{u}_{22} & \tilde{u}_{23} & \tilde{u}_{24} \\ 0 & 0 & (a_{33} \ominus \tilde{\ell}_{31} \odot \tilde{u}_{13}) \ominus \tilde{\ell}_{32} \odot \tilde{u}_{23} & (a_{34} \ominus \tilde{\ell}_{31} \odot \tilde{u}_{14}) \ominus \tilde{\ell}_{32} \odot \tilde{u}_{24} \\ 0 & 0 & (a_{43} \ominus \tilde{\ell}_{41} \odot \tilde{u}_{14}) \ominus \tilde{\ell}_{42} \odot \tilde{u}_{23} & (a_{44} \ominus \tilde{\ell}_{41} \odot \tilde{u}_{14}) \ominus \tilde{\ell}_{42} \odot \tilde{u}_{24} \end{pmatrix} \\
&= \begin{pmatrix} \tilde{u}_{11} & \tilde{u}_{12} & \tilde{u}_{13} & \tilde{u}_{14} \\ 0 & \tilde{u}_{22} & \tilde{u}_{23} & \tilde{u}_{24} \\ 0 & 0 & \tilde{u}_{33} & \tilde{u}_{34} \\ 0 & 0 & \star & \star \end{pmatrix}, \quad \tilde{u}_{3j} = (a_{3j} \ominus \tilde{\ell}_{31}\tilde{u}_{1j}) \ominus \tilde{\ell}_{32} \odot \tilde{u}_{2j}, \quad j = 3, 4.
\end{aligned}$$

I, u zadnjem koraku je ostala transformacija

$$\tilde{L}^{(3)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\tilde{\ell}_{43} & 1 \end{pmatrix},$$

koja primjenom na $\tilde{A}^{(2)}$ daje i preostali element matrice $\tilde{U}$,

$$\tilde{u}_{44} = ((a_{44} \ominus \tilde{\ell}_{41} \odot \tilde{u}_{14}) \ominus \tilde{\ell}_{42} \odot \tilde{u}_{24}) \ominus \tilde{\ell}_{43} \odot \tilde{u}_{34}.$$

Uočavamo da se elementi u_{ij} računaju prema formuli

$$u_{ij} = a_{ij} - \sum_{m=1}^{i-1} \ell_{im} u_{mj}, \quad 2 \le i \le n, \quad i \le j \le n,$$

pri čemu je $u_{1j} = a_{1j}$, $1 \le j \le n$. Ovu formulu je lako provjeriti raspisivanjem produkta $A = LU$ po elementima. U našem algoritmu, zbog grešaka zaokruživanja, vrijedi

$$\tilde{u}_{ij} = (\cdots((a_{ij} \ominus \tilde{\ell}_{i1} \odot \tilde{u}_{1j}) \ominus \tilde{\ell}_{i2} \odot \tilde{u}_{2j}) \ominus \cdots) \ominus \tilde{\ell}_{i,i-1} \odot \tilde{u}_{i-1,j}. \tag{3.3.12}$$

Formula (3.3.12) je samo specijalan slučaj računanja općenitog izraza oblika

$$s = v_1 w_1 \pm v_2 w_2 \pm v_3 w_3 \pm \cdots \pm v_p w_p,$$

pri čemu se koristi algoritam

$\tilde{u}_{ij} = a_{ij}$;
za $m = 1, \ldots, i-1$ {
$\quad \tilde{u}_{ij} = \tilde{u}_{ij} \ominus \tilde{\ell}_{im} \odot \tilde{u}_{mj}$; }

Koristeći Propoziciju 3.3.8, zaključujemo da postoje ξ_{ij}, ζ_{ijm} tako da je u (3.3.12)

$$\tilde{u}_{ij} = a_{ij}(1 + \xi_{ij}) - \sum_{m=1}^{i-1} \tilde{\ell}_{im} \tilde{u}_{mj}(1 + \zeta_{ijm}). \tag{3.3.13}$$

Pri tome je za sve i, j, m

$$|\xi_{ij}|,\ |\zeta_{ijm}| \leq \frac{n\boldsymbol{\varepsilon}}{1 - n\boldsymbol{\varepsilon}}.$$

Relaciju (3.3.13) možemo pročitati i kao

$$a_{ij} = \sum_{m=1}^{i} \tilde{\ell}_{im} \tilde{u}_{mj} + \delta a_{ij},\ \ \delta a_{ij} = \sum_{m=1}^{i} \tilde{\ell}_{im} \tilde{u}_{mj} \zeta_{ijm} - \xi_{ij} a_{ij}, \tag{3.3.14}$$

gdje smo $\tilde{u}_{ij}$ napisali kao $\tilde{\ell}_{ii} \tilde{u}_{ij}(1 + \zeta_{iji})$, uz $\tilde{\ell}_{ii} = 1$ i $\zeta_{iji} = 0$.
Time smo dokazali sljedeći teorem.

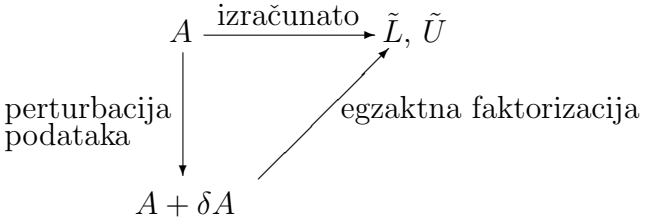
Teorem 3.3.3. *Neka je algoritam 3.3.4 primijenjen na matricu* $A \in \boldsymbol{R}^{n \times n}$ *i neka su uspješno izvršene sve njegove operacije. Ako su* $\tilde{L}$ *i* $\tilde{U}$ *izračunati trokutasti faktori, onda je*

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

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

Komentar 3.3.2. Rezultat teorema zaslužuje poseban komentar. Naša analiza nije dala odgovor na pitanje koliko su $\tilde{L}$ i $\tilde{U}$ daleko od točnih matrica L i U. Umjesto toga, dobili smo zaključak da $\tilde{L}$ i $\tilde{U}$ čine egzaktnu LU faktorizaciju matrice $A + \delta A$. Drugim riječima, ako bismo A promijenili u $A + \delta A$ i zatim uzeli egzaktnu faktorizaciju, dobili bismo upravo $\tilde{L}$ i $\tilde{U}$. Ovu situaciju možemo ilustrirati komutativnim dijagramom na slici 3.5. Dobiveni rezultat je u praksi od izuzetne važnosti.

Jer, često je u primjenama nemoguće raditi sa egzaktnim podacima – matrica A može biti rezultat mjerenja ili nekih prethodnih proračuna, dakle netočna. Ako je egzaktna (nepoznata) matrica $\hat{A}$ i $A = \hat{A} + \delta\hat{A}$, onda je $\tilde{L}\tilde{U} = \hat{A} + \delta\hat{A} + \delta A$ i $LU = \hat{A} + \delta\hat{A}$. Ako su δA i $\delta\hat{A}$ usporedivi po veličini, onda možemo u mnogim primjenama $\tilde{L}$ i $\tilde{U}$ smatrati jednako dobrim kao i L i U.



Slika 3.5: *Komutativni dijagram LU faktorizacije u aritmetici konačne preciznosti. Izračunati rezultat je ekvivalentan egzaktnom računu sa promijenjenim polaznim podacima.*

Iz prethodne analize je jasno da je δA mala ako produkt $|\tilde{L}||\tilde{U}|$ nije prevelik u usporedbi s $|A|$. To na žalost nije osigurano u LU faktorizaciji. Sljedeći primjer pokazuje numeričku nestabilnost algoritma.

Primjer 3.3.4. Neka je α mali parametar, $|\alpha| \ll 1$, i neka je matrica A definirana s

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

U egzaktnom računanju imamo

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

pa je LU faktorizacija matrice A dana s

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

Pretpostavimo sada da ovaj račun provodimo na računalu u aritmetici sa 8 decimalnih znamenki, tj. sa točnosti $\boldsymbol{\varepsilon} \approx 10^{-8}$. Neka je $|\alpha| < \boldsymbol{\varepsilon}$, npr. neka je $\alpha = 10^{-10}$. Kako je problem jednostavan, vrijedi $\tilde{\ell}_{21} = \ell_{21}(1+\epsilon_1)$, $|\epsilon_1| \leq \boldsymbol{\varepsilon}$, $\tilde{u}_{11} = u_{11}$, $\tilde{u}_{12} = u_{12}$ i

$$\tilde{u}_{22} = 1 \ominus 1 \oslash \alpha = -1 \oslash \alpha = -\frac{1}{\alpha}(1+\epsilon_1).$$

Primijetimo da je

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

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

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

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

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

Njihova rješenja su

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

Vidimo da se x_1 i $\tilde{x}_1$ potpuno razlikuju. Zaključujemo da Gaussove eliminacije mogu biti numerički nestabilne – dovoljna je jedna greška zaokruživanja " u krivo vrijeme na krivom mjestu" pa da dobiveni rezultat bude potpuno netočan.

Komentar 3.3.3. I ovaj primjer zaslužuje komentar. Vidimo da katastrofalno velika greška nije uzrokovana akumuliranjem velikog broja grešaka zaokruživanja. Cijeli problem je u samo jednoj aritmetičkoj operaciji (pri računanju $\tilde{u}_{22}$) koja je zapravo izvedena jako točno, sa malom greškom zaokruživanja. Cilj numeričke analize algoritma je da otkrije moguće uzroke nestabilnosti, objasni fenomene vezane za numeričku nestabilnost i ponudi rješenja za njihovo uklanjanje.
Primijetimo da je nestabilnost ilustrirana u primjeru u skladu sa teoremom 3.3.3. Naime, ako izračunamo $|\tilde{L}||\tilde{U}|$ dobijemo

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

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

Jasno nam je da je, zbog nenegativnosti matrica $|\tilde{L}|$ i $|\tilde{U}|$, mali produkt $|\tilde{L}||\tilde{U}|$ moguć samo ako su elementi od $\tilde{L}$ i $\tilde{U}$ mali po apsolutnoj vrijednosti. Pogledajmo nastavak primjera 3.3.4.

Primjer 3.3.5. Neka je A matrica iz primjera 3.3.4. Zamijenimo joj poredak redaka,

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

LU faktorizacija matrice $A' = LU$ je

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

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

$$\tilde{L} = \begin{pmatrix} 1 & 0 \\ \alpha & 1 \end{pmatrix}, \quad \tilde{U} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

i vrijedi

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

Primijetimo i da je produkt

$$|\tilde{L}||\tilde{U}| = \begin{pmatrix} 1 & 1 \\ |\alpha| & 1+\alpha \end{pmatrix}$$

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

Iz prethodnih primjera i diskusija je jasno da standardno pivotiranje redaka, koje osigurava da su u matrici L svi elementi po apsolutnoj vrijednosti najviše jednaki jedinici[5], doprinosi numeričkoj stabilnosti. Naime, lako se vidi sa su tada i svi elementi matrice $|\tilde{L}|$ manji ili jednaki od jedan. U tom slučaju veličina produkta $|\tilde{L}||\tilde{U}|$ bitno ovisi o elementima matrice $\tilde{U}$. S druge strane, elementi matrice $\tilde{U}$ su dobiveni iz matrica $\tilde{A}^{(k)}$, $k = 0, 1, \ldots, n-1$, pa je broj

$$\rho = \frac{\max\limits_{i,j,k} \tilde{a}_{ij}^{(k)}}{\max\limits_{ij} a_{ij}} \tag{3.3.15}$$

dobra mjera za relativni rast (u odnosu na A) elemenata u produktu $|\tilde{L}||\tilde{U}|$. Broj ρ zovemo faktor rasta elemenata u LU faktorizaciji i definiran je bez obzira da li koristimo pivotiranje redaka. Primijetimo da u analizi grešaka zaokruživanja pivotiranje ne predstavlja dodatnu tehničku poteškoću, pa odmah možemo iskazati sljedeći teorem.

Teorem 3.3.4. *Neka je LU faktorizacija $n \times n$ matrice A izračunata sa pivotiranjem redaka u aritmetici sa relativnom točnosti ε i neka su $\tilde{L}$ i $\tilde{U}$ dobivene aproksimacije za L i U. Ako je pri tome korištena permutacija P, onda je*

$$\tilde{L}\tilde{U} = P(A + \delta A), \quad |\delta A| \leq \frac{2n\varepsilon}{1 - 2n\varepsilon} P^\tau |\tilde{L}||\tilde{U}|.$$

Specijalno je, bez obzira na pivotiranje,

$$\|\delta A\|_F \leq O(n^3)\varepsilon\rho\|A\|_F.$$

Dokaz: Nakon ponovnog čitanja dokaza teorema 3.3.2 bi trebalo biti jasno da permutacije prolaze kroz elementarne transformacije $\tilde{L}^{(i)}$ neovisno o točnosti računanja (egzaktno ili do na greške zaokruživanja). Dakle, možemo zaključiti da čak i računanje faktorizacije s pivotiranjem na stroju odgovara računu bez pivotiranja ali sa polaznom matricom $A' = PA$. Sada primjenom teorema 3.3.3 dobijemo da vrijedi

$$\tilde{L}\tilde{U} = A' + \delta A', \quad |\delta A'| \leq \frac{2n\varepsilon}{1 - 2n\varepsilon} |\tilde{L}||\tilde{U}|.$$

[5] Vidi teorem 3.3.2.

Kako je $A' + \delta A' = P(A + P^\tau \delta A')$, stavljanjem $\delta A = P^\tau \delta A'$ dobivamo tvrdnju teorema. Primijetimo i da je, bez obzira da li pivotiramo retke ili ne,

$$\|\delta A\|_F \leq \frac{2n\varepsilon}{1-2n\varepsilon}\sqrt{\frac{n(n+1)}{2}}\sqrt{\frac{n(n+1)}{2}}\rho\|A\|_F.$$

□

Razlika u numeričkoj stabilnosti koju donosi pivotiranje redaka je bolje ponašanje parmetra ρ, tj. pivotiranjem možemo osigurati umjeren rast elemenata u toku LU faktorizacije. U primjeru 3.3.4 smo vidjeli da u LU faktorizaciji bez pivotiranja rast elemenata tokom faktorizacije može biti po volji veliki. Sljedeća propozicija pokazuje da u slučaju pivotiranja redaka faktor ρ ima gornju ogradu koja je funkcija samo dimenzije problema.

Propozicija 3.3.5. *Ako LU faktorizaciju računamo s pivotiranjem redaka u aritmetici sa maksimalnom greškom zaokruživanja ε, onda je*

$$\rho \leq 2^{n-1}(1+\varepsilon)^{2(n-1)}.$$

Specijalno je u slučaju egzaktnog računanja $\rho \leq 2^{n-1}$.

Dokaz: Dokaz ostavljamo čitatelju za vježbu. □

Primijetimo da je gornja ograda za ρ reda veličine 2^n, što brzo raste kao funkcija od n. Postoje primjeri na kojima se ta gornja ograda i dostiže. Ipak, iskustvo iz prakse govori da su takvi primjeri rijetki i da je LU faktorizacija sa pivotiranjem redaka dobar algoritam za rješavanje sustava linearnih jednadžbi. Možemo zaključiti i preporučiti sljedeće:

- ♣ *Gaussove eliminacije, odnosno LU faktorizaciju, valja u praksi uvijek raditi sa pivotiranjem redaka.*

3.3.7 Analiza numeričkog rješenja trokutastog sustava

Kako smo vidjeli u sekciji 3.3.2, trokutaste sustave rješavamo jednostavnim i elegantnim supstitucijama naprijed ili unazad. Ta jednostavnost se odražava i na dobra numerička svojstva supstitucija, kada ih provedemo na računalu. Sljedeća propozicija opisuje kvalitetu numerički izračunatog rješenja trokutastog sustava jednadžbi.

Propozicija 3.3.6. *Neka je T donje (gornje) trokutasta matrica reda n i neka je sustav $Tv = d$ riješen supstitucijama naprijed (unazad) kako je opisano u sekciji 3.3.2. Ako je $\tilde{v}$ rješenje dobiveno primjenom strojne aritmetike preciznosti $\boldsymbol{\varepsilon}$, onda postoji donje (gornje) trokutasta matrica δT tako da vrijedi*

$$(T + \delta T)\tilde{v} = d, \;\; |\delta T| \leq \eta_{\triangleright}|T|, \;\; 0 \leq \eta_{\triangleright} \leq \frac{n\boldsymbol{\varepsilon}}{1 - n\boldsymbol{\varepsilon}}.$$

Dokaz: Dokaz zbog jednostavnosti provodimo samo za donje trokutastu matricu T. Pretpostavljamo da se i–ta komponenta rješenja za $i > 1$ računa na sljedeći način:

$\tilde{v}_i = T_{i1} \odot \tilde{v}_1$;
za $j = 2, \ldots, i-1$ {
$\quad \tilde{v}_i = \tilde{v}_i \oplus T_{ij} \odot \tilde{v}_j$; }
$\tilde{v}_i = (d_i \ominus \tilde{v}_i) \oslash T_{ii}$.

Primjenom pravila strojne aritmetike dobijemo[6] $\tilde{v}_1 = (d_1/T_{11})(1 + \epsilon_1)$, $|\epsilon_1| \leq \boldsymbol{\varepsilon}$, te za $i = 2, \ldots, n$

$$\tilde{v}_i = \frac{d_i - \sum_{j=1}^{i-1} T_{ij}(1 + \zeta_j)\tilde{v}_j}{\frac{T_{ii}}{(1 + \epsilon_{1,i})(1 + \epsilon_{2,i})}}, \;\; |\zeta_j| \leq \frac{(i-1)\boldsymbol{\varepsilon}}{1 - (i-1)\boldsymbol{\varepsilon}}, \;\; |\epsilon_{1,i}| \leq \boldsymbol{\varepsilon}, \;\; |\epsilon_{2,i}| \leq \boldsymbol{\varepsilon}.$$

□

Komentar 3.3.4. Koliko god da je prethodni rezultat tehnički jednostavan, valja naglasiti da je zaključak o točnosti rješenja trokutastog sustava važan: izračunato rješenje zadovoljava trokutasti sustav sa matricom koeficijenata koja se po elementima malo razlikuje od zadane. Pojednostavljeno govoreći, ako radimo sa $\boldsymbol{\varepsilon} \approx 10^{-8}$ i ako je $n = 1000$, onda izračunati vektor $\tilde{v}$ zadovoljava $\tilde{T}\tilde{v} = d$, gdje se elemeni od $\tilde{T}$ i T poklapaju u barem 5 decimalnih znamenki (od 8 na koliko je zadana matrica T).

3.3.8 Točnost izračunatog rješenja sustava

Sada nam ostaje napraviti kompoziciju dobivenih rezultata i ocijeniti koliko točno možemo na računalu riješiti linearni sustav $Ax = b$ u kojem smo izračuali LU faktorizaciju $PA = LU$ i supstitucijama naprijed i unazad izračunali $x = U^{-1}(L^{-1}(Pb))$.

[6] Vidi sekciju 3.3.9.

Kako smo vidjeli u prethodnim sekcijama permutacija P se može (za potrebe analize) odmah primjeniti na polazne podatke i numeričku analizu možemo provesti bez pivotiranja. Kako to pojednostavljuje oznake, mi ćemo pretpostaviti da su na polazne podatke A i b već primijenjene zamjene redaka, tako da su formule jednostavno $A = LU$ i $x = U^{-1}(L^{-1}b)$).
Neka su $\tilde{L}$ i $\tilde{U}$ izračunate trokutaste matrice, gdje je $\tilde{L}\tilde{U} = A + \delta A$, kao u teoremu 3.3.3. Izračunato rješenje $\tilde{y}$ sustava $\tilde{L}y = b$ zadovoljava (prema propoziciji 3.3.6)

$$(\tilde{L} + \delta\tilde{L})\tilde{y} = b, \;\; |\delta\tilde{L}| \leq \frac{n\varepsilon}{1 - n\varepsilon}|\tilde{L}|.$$

Na isti način rješenje $\tilde{x}$ sustava $\tilde{U}x = \tilde{y}$ zadovoljava

$$(\tilde{U} + \delta\tilde{U})\tilde{x} = \tilde{y}, \;\; |\delta\tilde{U}| \leq \frac{n\varepsilon}{1 - n\varepsilon}|\tilde{U}|.$$

Dakle, $(\tilde{L} + \delta\tilde{L})(\tilde{U} + \delta\tilde{U})\tilde{x} = b$, tj.

$$\begin{aligned} (A + \delta A + E)\tilde{x} = b, \;\; E &= \tilde{L}\delta\tilde{U} + \delta\tilde{L}\tilde{U} + \delta\tilde{L}\delta\tilde{U}, \\ |E| &\leq |\tilde{L}||\delta\tilde{U}| + |\delta\tilde{L}||\tilde{U}| + |\delta\tilde{L}||\delta\tilde{U}|. \end{aligned}$$

Time smo dokazali sljedeći teorem.

Teorem 3.3.7. *Neka je $\tilde{x}$ rješenje regularnog $n \times n$ sustava jednadžbi $Ax = b$, dobiveno Gaussovim eliminacijama sa pivotiranjem redaka. Tada postoji perturbacija ΔA za koju vrijedi*

$$(A + \Delta A)\tilde{x} = b, \;\; |\Delta A| \leq \frac{5n\varepsilon}{1 - 2n\varepsilon}P^{\tau}|\tilde{L}||\tilde{U}|.$$

Ovdje je P permutacija koja realizira zamjenu redaka. Također pretpostavljamo da je $2n\varepsilon < 1$.

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

Algoritam 3.3.5. Rješavanje trokutastog sustava jednadžbi $Ax = b$ Gaussovim eliminacijama bez dodatne memorije.

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

3.3.9 Dodatak: Osnove matričnog računa na računalu

Na računalu općenito ne možemo egzaktno izvršavati aritmetičke operacije. Rezultat zbrajanja, oduzimanja, množenja ili dijeljenja dva strojna broja x i y je po definiciji strojni broj koji je najbliži egzaktnom zbroju, razlici, umnošku, odnosno kvocijentu x i y. Pri tome je relativna greška tako izvedenih operacija manja ili jednaka polovini najvećeg relativnog razmaka dva susjedna strojna broja. Na primjer, u standardnoj jednostrukoj preciznosti (32–bitna reprezentacija) je relativni razmak susjednih brojeva omeđen s 2^{-23} pa je relativna greška aritmetičkih operacija najviše $\boldsymbol{\varepsilon} \approx 10^{-8}$.
Navedena pravila za izvršavanje elementarnih aritmetičkih operacija lako zapišemo na sljedeći način:

$$\begin{aligned}
\text{zbrajanje: } x \oplus y &= (x+y)(1+\epsilon_1), \quad |\epsilon_1| \le \boldsymbol{\varepsilon} \\
\text{oduzimanje: } x \ominus y &= (x-y)(1+\epsilon_2), \quad |\epsilon_2| \le \boldsymbol{\varepsilon} \\
\text{množenje: } x \odot y &= xy(1+\epsilon_3), \quad |\epsilon_3| \le \boldsymbol{\varepsilon} \\
\text{dijeljenje: } x \oslash y &= \frac{x}{y}(1+\epsilon_4), \quad |\epsilon_4| \le \boldsymbol{\varepsilon}, \quad y \neq 0.
\end{aligned}$$

Ove relacije vrijede ako su rezultati navedenih operacija po apsolutnoj vrijednosti

u inervalu (μ, M) gdje je npr. u 32–bitnoj reprezentaciji $\mu = 2^{-126} \approx 10^{-38}$ najmanji a $M = (1 + 2^{-1} + \cdots + 2^{-23})2^{127} \approx 10^{38}$ najveći normalizirani strojni broj. (U dvostrukoj preciznosti (64–bitna reprezentacija brojeva) je $\mu \approx 10^{-308}$, $M \approx 10^{308}$.) Analiza za rezultate izvan intervala (μ, M) je nešto složenija pa je nećemo raditi.
Kako na računalu izgledaju osnovne operacije linearne algebre? Lako se uvjerimo da je većina operacija (skalarni produkt, norma, linearne kombinacije, matrične operacije) bazirana na računanju

$$s = \sum_{i=1}^{m} x_i y_i,$$

gdje su x_i, y_i skalari (realni ili kompleksni brojevi ili njihove aproksimacije na računalu). Ako s računamo na standardan način, u računalu dobijemo, npr. sa $m = 4$, izraz oblika

$$\tilde{s} = ((((x_1 \odot y_1) \oplus x_2 \odot y_2) \oplus x_3 \odot y_3) \oplus x_4 \odot y_4).$$

Sustavnom primjenom osnovnih svojstava aritmetike na stroju, lako se provjeri da je

$$\begin{aligned}
\tilde{s} &= (((x_1y_1(1+\epsilon_1) + x_2y_2(1+\epsilon_2))(1+\xi_2) + x_3y_3(1+\epsilon_3))(1+\xi_3) \\
&\quad +x_4y_4(1+\epsilon_4))(1+\xi_4) \\
&= x_1y_1 \underbrace{(1+\epsilon_1)(1+\xi_2)(1+\xi_3)(1+\xi_4)}_{1+\zeta_1} + x_2y_2 \underbrace{(1+\epsilon_2)(1+\xi_2)(1+\xi_3)(1+\xi_4)}_{1+\zeta_2} \\
&\quad +x_3y_3 \underbrace{(1+\epsilon_3)(1+\xi_3)(1+\xi_4)}_{1+\zeta_3} + x_4y_4 \underbrace{(1+\epsilon_4)(1+\xi_4)}_{1+\zeta_4} = \sum_{i=1}^{m=4} x_iy_i(1+\zeta_i),
\end{aligned}$$

gdje su sve vrijednosti ϵ_i, ξ_i po modulu manje od $\boldsymbol{\varepsilon}$. Sada je jasno kako bi izgledala formula za proizvoljan broj od m sumanada. Primijetimo da $1+\zeta_k$ možemo ocijeniti s

$$1 - m\boldsymbol{\varepsilon} \leq 1 + \zeta_k \leq \frac{1}{1 - m\boldsymbol{\varepsilon}}, \text{ tj. vrijedi } |\zeta_k| \leq \frac{m\boldsymbol{\varepsilon}}{1 - m\boldsymbol{\varepsilon}}, \quad k = 1, 2, \ldots, m.$$

Propozicija 3.3.8. *Neka su* $x_1, \ldots, x_m$, $y_1, \ldots, y_m$ *brojevi u računalu,* $m \geq 1$. *Ako vrijednost* $s = \sum_{i=1}^{m} x_i y_i$ *računamo kao*

$\tilde{s} = x_1 \odot y_1$;
za $i = 2, \ldots, m$ {
$\quad \tilde{s} = \tilde{s} \oplus x_i \odot y_i$; }

onda postoje brojevi ζ_i, $i = 1, \ldots, m$, *tako da vrijedi*

$$\tilde{s} = \sum_{i=1}^{m} x_i y_i (1 + \zeta_i), \quad |\zeta_i| \leq \frac{m\boldsymbol{\varepsilon}}{1 - m\boldsymbol{\varepsilon}}, \quad i = 1, 2, \ldots, m.$$

Dokaz: Dokaz smo već skicirali na primjeru $m = 4$. Očito je formalni dokaz najlakše izvesti matematičkom indukcijom po m. Dovoljno je primijetiti da je u koraku indukcije

$$\begin{aligned} \tilde{s} \oplus x_{m+1} \odot y_{m+1} &= (\tilde{s} + x_{m+1} y_{m+1}(1 + \omega_1))(1 + \omega_2) \\ &= \tilde{s}(1 + \omega_2) + x_{m+1} y_{m+1}(1 + \omega_1)(1 + \omega_2), \quad |\omega_1| \leq \boldsymbol{\varepsilon}, \quad |\omega_2| \leq \boldsymbol{\varepsilon}, \end{aligned}$$

te da je $1 - (m+1)\boldsymbol{\varepsilon} \leq (1 - \boldsymbol{\varepsilon})(1 - m\boldsymbol{\varepsilon})$ i

$$\frac{1 + \boldsymbol{\varepsilon}}{1 - m\boldsymbol{\varepsilon}} \leq 1 + \frac{(m+1)\boldsymbol{\varepsilon}}{1 - (m+1)\boldsymbol{\varepsilon}}.$$

□

3.4 Teorija perturbacija za linearne sustave

Iz prethodnih razmatranja je jasno da u primjenama rijetko možemo izračunati egzaktno rješenje sustava $Ax = b$. Jer, i formiranje samog sustava (računanje koeficijenata sustava i desne strane) i njegovo rješavanje na računalu uzrokuju greške. Analizom tih grešaka dolazimo do zaključka da izračunata aproksimacija rješenja $\tilde{x} = x + \delta x$ zadovoljava tzv. perturbirani sustav, $(A + \delta A)(x + \delta x) = b + \delta b$. Sada se postavlja pitanje kako ocijeniti veličinu greške $\delta x = \tilde{x} - x$, ako je poznata informacija o veličini grešaka δA i δb.
U primjeru 3.3.4 smo vidjeli da čak i mala perturbacija δA može potpuno promijeniti rješenje x. Kako mi u realnoj primjeni ne znamo točno rješenje, cilj nam je otkriti kako možemo iz matrice A i vektora b dobiti ne samo (što je moguće bolju) aproksimaciju $\tilde{x}$, nego i procjenu koliko je ta aproksimacija dobra.
Za početak teorijske analize, promotrimo jednostavniji slučaj u kojem je $\delta b = 0$. Dakle, jedina perturbacija je ona koja A promijeni u $A + \delta A$. Zbog jednostavnosti

ćemo promatrati samo (inače, važan) slučaj u kojem je matrica koeficijenata $A+\delta A$ i dalje regularna, pa je $x + \delta x$ jedinstveno određen.
Iz jednakosti $A + \delta A = A(I + A^{-1}\delta A)$ vidimo da je regularnost matrice $A + \delta A$ osigurana ako je $I + A^{-1}\delta A$ regularna. Uvjet pod kojim možemo garantirati regularnost matrice $I + A^{-1}\delta A$ daje sljedeća propozicija.
Koristeći ovu propoziciju, regularnost matrice $A + \delta A$ obično osiguravamo tako da zahtijevamo da je $\|A^{-1}\delta A\| < 1$. Izbor matrične norme $\|\cdot\|$ ovisi o konkretnoj situaciji, npr. o tipu informacije o δA ili o teorijskim rezultatima koje koristimo u analizi. Neka je matrična norma jednaka Frobeniusovoj normi, $\|\cdot\| = \|\cdot\|_F$,

$$\|X\|_F = \sqrt{\sum_{i,j=1}^{n} |X_{ij}|^2} = \sqrt{\mathrm{trag}(X^\tau X)}.$$

Informacija o perturbaciji δA je važan faktor u razvoju analize. Neka je na primjer zadano (poznato) da je

$$\epsilon \equiv \frac{\|\delta A\|_F}{\|A\|_F} \ll 1$$

mali broj, tj. da je perturbacija *mala po normi*. Regularnost matrice $A + \delta A$ je osigurana ako je npr.

$$\|A^{-1}\|_F\|\delta A\|_F = \epsilon(\|A\|_F\|A^{-1}\|_F) < 1, \ \text{ tj. } \ \epsilon < \frac{1}{\|A\|_F\|A^{-1}\|_F}.$$

Tada je $\|A^{-1}\delta A\|_F < 1$ i $(A + \delta A)^{-1} = (I + A^{-1}\delta A)^{-1}A^{-1}$, pa $\tilde{x} = (A + \delta A)^{-1}b$ možemo pisati kao

$$\tilde{x} = (I + A^{-1}\delta A)^{-1}A^{-1}b = (I + A^{-1}\delta A)^{-1}x, \ \text{ tj. } \ (I + A^{-1}\delta A)\tilde{x} = x.$$

Znači, $x - \tilde{x} = A^{-1}\delta A\tilde{x}$, pa je

$$\|x - \tilde{x}\|_2 \leq \|A^{-1}\delta A\|_F\|\tilde{x}\|_2.$$

Kako je $\|A^{-1}\delta A\|_F \leq \|A^{-1}\|_F\|\delta A\|_F$, dobijamo

$$\frac{\|x - \tilde{x}\|_2}{\|\tilde{x}\|_2} \leq \|A^{-1}\|_F\|A\|_F\frac{\|\delta A\|_F}{\|A\|_F} = \epsilon\|A^{-1}\|_F\|A\|_F. \tag{3.4.1}$$

Relacija (3.4.1) pokazuje da relativna greška u izračunatom rješenju $\tilde{x}$ može biti uvećana najviše sa faktorom $\kappa_F(A) = \|A^{-1}\|_F\|A\|_F$ u odnosu na relativnu promjenu $\epsilon = \|\delta A\|_F/\|A\|_F$ u polaznoj matrici A.

3.4.1 Perturbacije male po normi

Sljedeći teorem daje potpuni opis greške ako je perturbacija dana po normi. Općenito ćemo promatrati i δA i δb, a mjerenja perturbacija će biti u proizvoljnoj vektorskoj normi $\|\cdot\|$ i pripadnoj matričnoj normi $\|\cdot\|$.

Teorem 3.4.1. *Neka je $Ax = b$, $(A+\delta A)(x+\delta x) = b+\delta b$, gdje je $\|\delta A\| \leq \epsilon\|A\|$, $\|\delta b\| \leq \epsilon\|b\|$. Ako je $\epsilon\|A^{-1}\|\|A\| < 1$, onda je*

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\epsilon}{1-\epsilon\|A^{-1}\|\|A\|}\left(\frac{\|A^{-1}\|\|b\|}{\|x\|} + \|A^{-1}\|\|A\|\right) \leq 2\frac{\epsilon\|A^{-1}\|\|A\|}{1-\epsilon\|A^{-1}\|\|A\|}.$$

Pri tome postoje perturbacije δA i δb za koje je gornja nejednakost skoro dostignuta. Preciznije, postoje δA i δb tako da je $\|\delta A\| = \epsilon\|A\|$, $\|\delta b\| = \epsilon\|b\|$, te

$$\frac{\|\delta x\|}{\|x\|} \geq \frac{\epsilon}{1+\epsilon\|A^{-1}\|\|A\|}\left(\frac{\|A^{-1}\|\|b\|}{\|x\|} + \|A^{-1}\|\|A\|\right).$$

Dokaz: Iz pretpostavki teorema je

$$\delta x = A^{-1}\delta b - A^{-1}\delta A x - A^{-1}\delta A\delta x, \tag{3.4.2}$$

pa uzimanjem norme dobijemo

$$\|\delta x\| \leq \epsilon\|A^{-1}\|\|b\| + \epsilon\|A^{-1}\|\|A\|\|x\| + \epsilon\|A^{-1}\|\|A\|\|\delta x\|,$$

odakle, rješavanjem nejednakosti po $\|\delta x\|$ slijedi tvrdnja.
Da bismo konstruirali perturbacije za koje dobivena nejednakost skoro postaje jednakost, pogledajmo desnu stranu jednakosti (3.4.2). Vrijedi

$$\|\delta x\| \geq \|A^{-1}\delta b - A^{-1}\delta A x\| - \|A^{-1}\delta A\delta x\|.$$

Pokušajmo odrediti perturbacije δA, δb tako da vrijedi

$$\|A^{-1}\delta b - A^{-1}\delta A x\| = \epsilon\|A^{-1}\|\|b\| + \epsilon\|A^{-1}\|\|A\|\|x\|.$$

Dakle, δA i δb treba odabrati tako da je $\|\delta A\| \leq \epsilon\|A\|$, $\|\delta b\| \leq \epsilon\|b\|$,

$$\|A^{-1}\delta b\| = \epsilon\|A^{-1}\|\|b\|, \quad \|A^{-1}\delta A x\| = \epsilon\|A^{-1}\|\|A\|\|x\|,$$

te da je norma razlike $A^{-1}\delta b - A^{-1}\delta A x$ jednaka sumi normi vektora. Ovaj zadnji uvjet znači da $A^{-1}\delta b$ i $-A^{-1}\delta A x$ moraju biti kolinearni.
Ako je u jedinični vektor za kojeg je $\|A^{-1}u\| = \|A^{-1}\|$, onda $\delta b = \epsilon\|b\|u$ zadovoljava $\|\delta b\| = \epsilon\|b\|$ i $\|A^{-1}\delta b\| = \epsilon\|A^{-1}\|\|b\|$. Sada stavimo $\delta A = \epsilon\|A\|uv^\tau$, gdje je $v \in \mathbf{R}^n$ vektor kojeg ćemo odrediti da postignemo željene relacije:

(i) $\|\delta A\| = \epsilon\|A\| \max\limits_{z\neq 0} \dfrac{\|uv^\tau z\|}{\|z\|} = \epsilon\|A\| \max\limits_{z\neq 0} \dfrac{|v^\tau z|}{\|z\|}$ treba postati $\|\delta A\| = \epsilon\|A\|$;

(ii) $\|A^{-1}\delta Ax\| = \epsilon\|A\|\|A^{-1}\||v^\tau x|$ treba postati $\|A^{-1}\delta Ax\| = \epsilon\|A\|\|A^{-1}\|\|x\|$.

Dakle, treba nam vektor v sa svojstvom da je, za sve z, $|v^\tau z| \leq \|z\|$, te da je $|v^\tau x| = \|x\|$. Postojanje takvog vektora je rezultat Hahn–Banachovog teorema: takav vektor v postoji. Dakle, konstruirali smo δA i δb za koje je

$$\|\delta x\| \geq \epsilon\|A^{-1}\|\|b\| + \epsilon\|A^{-1}\|\|A\|\|x\| - \epsilon\|A^{-1}\|\|A\|\|\delta x\|,$$

čime je dokaz druge tvrdnje teorema završen. □

Vidimo da teorem 3.4.1 iz zadane informacije o veličini perturbacija po normi ($\|\delta A\| \leq \epsilon\|A\|$, $\|\delta b\| \leq \epsilon\|b\|$) izvodi optimalnu[7] ocjenu iz koje se jasno vidi da je broj

$$\kappa(A) = \|A^{-1}\|\|A\| \tag{3.4.3}$$

odlučujući faktor u donošenju suda o numeričkoj kvaliteti izračunate aproksimacije $\tilde{x} = x + \delta x$ sustava $Ax = b$. Pravilo je jednostavno:

♣ *Ako je relativna greška (po normi) u podacima najviše ϵ, onda se relativna greška u rješenju ponaša kao $\kappa(A)\epsilon$.*

3.4.2 Rezidualni vektor i stabilnost

Postoji jedan jednostavan i koristan način kako prosuditi kvalitetu aproksimacije $\tilde{x}$ rješenja sustava $\mathsf{Ax} = \mathsf{b}$. Radi se o *rezidualnom vektoru*

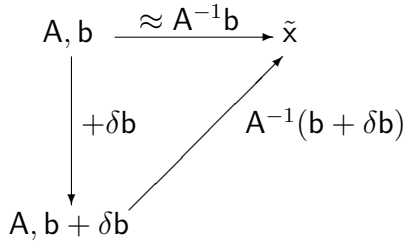
$$r = \mathsf{b} - \mathsf{A}\tilde{x}. \tag{3.4.4}$$

Kako je $\mathsf{b} - \mathsf{Ax} = \mathbf{0}$, jasno je da bi za dobru aproksimaciju $\tilde{x}$ pripadni rezidual trebao biti mali po normi. Ako relaciju (3.4.4) pročitamo kako

$$\mathsf{A}\tilde{x} = \mathsf{b} - r, \text{ tj. kao } \mathsf{A}\tilde{x} = \mathsf{b} + \delta\mathsf{b}, \text{ gdje je } \delta\mathsf{b} = -r,$$

onda vidimo da je $\tilde{x}$ egzaktno rješenje sustava koji je dobiven iz originalnog sustava promjenom desne strane u $\mathsf{b} + \delta\mathsf{b}$. Tu situaciju obično ilustriramo komutativnim dijagramom kao na Slici 3.6. Na neki način, perturbacija $\delta\mathsf{b} = -r$ opravdava $\tilde{x}$

[7]Ovdje pod optimalnosti podrazumijevamo činjenicu da je gornju ogradu za $\|\delta x\|/\|x\|$ nemoguće bitno poboljšati.



Slika 3.6: Približno rješenje sustava kao egzaktno rješenje perturbiranog sustava.

i, ako je $\xi \equiv \|r\|/\|\mathsf{b}\|$ dovoljno mali broj (relativno prema jedinici), daje nam argument da $\tilde{x}$ koristimo kao prihvatljivu aproksimaciju rješenja x.

Zašto? Recimo da su naša polazna matrica A i vektor b rezultati mjerenja ili nekih prethodnih proračuna.

Teorem 3.4.2. *Neka je $\tilde{x}$ aproksimacija rješenja sustava $\mathsf{Ax} = \mathsf{b}$ i neka je*

$$\beta(\tilde{x}) = \min\{\epsilon \geq 0 \ : \ (\mathsf{A} + \delta\mathsf{A})\tilde{x} = \mathsf{b} + \delta\mathsf{b}, \ \ \|\delta\mathsf{A}\| \leq \epsilon\|\mathsf{A}\|, \ \ \|\delta\mathsf{b}\| \leq \epsilon\|\mathsf{b}\|\}.$$

Tada je

$$\beta(\tilde{x}) = \frac{\|\mathsf{b} - \mathsf{A}\tilde{x}\|}{\|\mathsf{A}\|\|\tilde{x}\| + \|\mathsf{b}\|}.$$

Dokaz: Neka je $\mathsf{r} = \mathsf{b} - \mathsf{A}\tilde{x}$ rezidualni vektor. Ako je $\epsilon \geq 0$ takav da postoje $\delta\mathsf{A}$, $\delta\mathsf{b}$ takvi da je $\|\delta\mathsf{A}\| \leq \epsilon\|\mathsf{A}\|$, $\|\delta\mathsf{b}\| \leq \epsilon\|\mathsf{b}\|$, $(\mathsf{A} + \delta\mathsf{A})\tilde{x} = \mathsf{b} + \delta\mathsf{b}$, onda vrijedi

$$r = \delta\mathsf{A}\tilde{x} - \delta\mathsf{b}, \ \text{ pa je } \ \|r\| \leq \epsilon(\|\mathsf{A}\|\|\tilde{x}\| + \|\mathsf{b}\|), \ \text{ tj. } \ \epsilon \geq \underline{\epsilon} \equiv \frac{\|\mathsf{b} - \mathsf{A}\tilde{x}\|}{\|\mathsf{A}\|\|\tilde{x}\| + \|\mathsf{b}\|}.$$

Dakle, znamo donju ogradu za veličinu perturbacije koja opravdava $\tilde{x}$. Sada preostaje konstruirati perturbaciju koja dostiže tu vrijednost. Stavimo

$$\delta\mathsf{b} = -\frac{\|\mathsf{b}\|}{\|\mathsf{A}\|\|\tilde{x}\| + \|\mathsf{b}\|} r.$$

Očito je $\|\delta\mathsf{b}\| = \underline{\epsilon}\|\mathsf{b}\|$. Odredimo $\delta\mathsf{A}$ tako da je $\|\delta\mathsf{A}\| = \underline{\epsilon}\|\mathsf{A}\|$ i

$$(\mathsf{A} + \delta\mathsf{A})\tilde{x} = \mathsf{b} - \frac{\|\mathsf{b}\|}{\|\mathsf{A}\|\|\tilde{x}\| + \|\mathsf{b}\|} r, \ \text{ tj. } \ \delta\mathsf{A}\tilde{x} = \frac{\|\mathsf{A}\|\|\tilde{x}\|}{\|\mathsf{A}\|\|\tilde{x}\| + \|\mathsf{b}\|} r.$$

Definirajmo

$$\delta\mathsf{A} = \frac{\|\mathsf{A}\|}{\|\mathsf{A}\|\|\tilde{x}\| + \|\mathsf{b}\|} r v^{\mathrm{T}},$$

gdje je v vektor sa svojstvima

$$v^{\mathrm{T}}\tilde{x} = \|\tilde{x}\| \text{ i za sve } z \text{ vrijedi } |v^{\mathrm{T}}z| \leq \|z\|.$$

Takav vektor v postoji po Hahn–Banachovom teoremu i lako provjerimo da δA ima sva tražena svojstva. □

3.4.3 Perturbacije po elementima

Najpreciznija ocjena perturbacije u matrici A je kada imamo informaciju o perturbaciji svakog njenog elementa, tj. svakog koeficijenta u sustavu jednadžbi. Ako je $A = (a_{ij})_{i,j=1}^n$ i $A + \delta A = (a_{ij} + \delta a_{ij})_{i,j=1}^n$, onda je takva ocjena dana relacijama

$$|\delta a_{ij}| \leq \varepsilon |a_{ij}|, \quad i, j = 1, 2, \ldots, n,$$

gdje je $0 \leq \varepsilon \ll 1$. Ove nejednakosti jednostavno zapisujemo kao $|\delta A| \leq \varepsilon |A|$, tj. apsolutne vrijednosti matrica i nejednakost među matricama shvaćamo po elementima. Na isti način pišemo $|\delta b| \leq \varepsilon |b|$. Primijetimo da ovakve perturbacije ($|\delta A| \leq \varepsilon |A|$, $|\delta b| \leq \varepsilon |b|$) čuvaju strukturu u smislu da nule u matrici A i vektoru b ostaju nepromijenjene. Nadalje, ove perturbacije su neizbježne pri pohranjivanju podataka u računalo.

Teorem 3.4.3. *Neka je $Ax = b$ i $(A + \delta A)(x + \delta x) = b + \delta b$, gdje je*

$$|\delta A| \leq \varepsilon |A|, \quad |\delta b| \leq \varepsilon |b|.$$

Uzmimo proizvoljnu apsolutnu vektorsku normu $\|\cdot\|$ i njenu induciranu matričnu normu, također označenu s $\|\cdot\|$. Neka je $\varepsilon \| |A^{-1}||A| \| < 1$. Tada vrijedi

$$\frac{\|\delta x\|}{\|x\|} \leq \varepsilon \frac{\| |A^{-1}||A||x| + |A^{-1}||b| \|}{(1 - \varepsilon \| |A^{-1}||A| \|)\|x\|}. \tag{3.4.5}$$

Nadalje, postoje perturbacije δA i δb takve da je $|\delta A| = \varepsilon |A|$, $|\delta b| = \varepsilon |b|$ i da za rješenje $x + \delta x = (A + \delta A)^{-1}(b + \delta b)$ vrijedi

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \geq \varepsilon \frac{\| |A^{-1}||A||x| + |A^{-1}||b| \|_\infty}{(1 + \varepsilon \| |A^{-1}||A| \|_\infty)\|x\|_\infty}. \tag{3.4.6}$$

Dokaz: Prije svega, uvjet $\varepsilon \| |A^{-1}||A| \| < 1$ osigurava da je $A + \delta A$ regularna matrica, pa je $x + \delta x$ jedinstveno određen. Sada lako provjerimo da vrijedi

$$\delta x = -A^{-1}\delta A(x + \delta x) + A^{-1}\delta b, \tag{3.4.7}$$

pa primjenom nejednakosti trokuta (mnogokuta) dobijemo da je

$$\begin{aligned} |\delta x| &\leq |A^{-1}||\delta A||x| + |A^{-1}||\delta A||\delta x| + |A^{-1}||\delta b| \\ &\leq \varepsilon|A^{-1}||A||x| + \varepsilon|A^{-1}||A||\delta x| + \varepsilon|A^{-1}||b| \end{aligned}$$

pa je

$$\|\delta x\| \leq \varepsilon(\||A^{-1}||A||x| + |A^{-1}||b|\|) + \varepsilon\||A^{-1}||A|\|\|\delta x\|$$

Neka je $m \in \{1, 2, \ldots, n\}$ odabran tako da je

$$(|A^{-1}||A||x| + |A^{-1}||b|)_m = \||A^{-1}||A||x| + |A^{-1}||b|\|_\infty.$$

Definirajmo dijagonalne matrice

$$D_1 = \operatorname{diag}(\operatorname{sign}((A^{-1})_{mi}))_{i=1}^n, \quad D_2 = \operatorname{diag}(\operatorname{sign}(x_i))_{i=1}^n,$$

i perturbacije $\delta A = \varepsilon D_1|A|D_2$, $\delta b = -\varepsilon D_1|b|$. Sada lako računamo

$$\begin{aligned} (A^{-1}\delta A x - A^{-1}\delta b)_m &= \sum_{j=1}^n \sum_{i=1}^n (A^{-1})_{mj}(\delta A)_{ji} x_i - \sum_{j=1}^n (A^{-1})_{mj}\delta b_j \\ &= \varepsilon(|A^{-1}||A||x| + |A^{-1}||b|)_m \\ &= \varepsilon\||A^{-1}||A||x| + |A^{-1}||b|\|_\infty. \end{aligned}$$

S druge strane, iz relacije 3.4.7 lako izvedemo da je

$$(A^{-1}\delta A x - A^{-1}\delta b)_m = -(\delta x + A^{-1}\delta A \delta x)_m,$$

pa je

$$\varepsilon\||A^{-1}||A||x| + |A^{-1}||b|\|_\infty \leq \|\delta x\|_\infty + \varepsilon\||A^{-1}||A|\|_\infty \|\delta x\|_\infty.$$

Korolar 3.4.4. *Relativni faktor osjetljivosti je dan relacijom*

$$\lim_{\varepsilon \to 0} \sup \left\{ \frac{\dfrac{\|\delta x\|_\infty}{\|x\|_\infty}}{\varepsilon} \;:\; (A + \delta A)(x + \delta x) = b + \delta b,\ |\delta A| \leq \varepsilon|A|,\ |\delta b| \leq \varepsilon|b| \right\}$$

$$= \frac{\||A^{-1}||A||x| + |A^{-1}||b|\|_\infty}{\|x\|_\infty}.$$

Primijetimo da je

$$\||A^{-1}||A||x|\|_\infty \le \||A^{-1}||A||x| + |A^{-1}||b|\|_\infty \le 2\||A^{-1}||A||x|\|_\infty.$$

Zato kao koeficijent osjetljivosti možemo koristiti veličinu

$$\kappa_\infty(A, x) = \frac{\||A^{-1}||A||x|\|_\infty}{\|x\|_\infty}.$$

Teorem 3.4.5. *Neka je $\tilde{x}$ izračunata aproksimacija rješenja sustava $Ax = b$ i neka je $r = b - A\tilde{x}$ izračunati rezidual. Vrijedi*

$$\min\{\varepsilon \ge 0 \ : \ (A + \delta A)\tilde{x} = b + \delta b, \ |\delta A| \le \varepsilon|A|, \ |\delta b| \le \varepsilon|b|\} = \max_i \frac{|r_i|}{(|A||\tilde{x}| + |b|)_i}$$

Optimalna perturbacija polaznih podataka koja reproducira izračunato rješenje je dana s

$$\delta A = D_1|A|D_2, \ \ \delta b = -D_1|b|,$$

gdje je $D_1 = \operatorname{diag}\left(\frac{|r_i|}{(|A||\tilde{x}| + |b|)_i}\right)_{i=1}^n$, $D_2 = \operatorname{diag}(\operatorname{sign}(\tilde{x}_i))_{i=1}^n$.

Primjer 3.4.1. U ovom primjeru istražujemo stabilnost Gaussovih eliminacija po elementima matrice. Neka su $\alpha \neq 0$ i $\beta \neq 0$ zadani i neka je

$$A = \begin{pmatrix} \alpha & \beta \\ \alpha & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ \ x = A^{-1}b = \begin{pmatrix} 0 \\ \dfrac{1}{\beta} \end{pmatrix}.$$

Trokutasta LU faktorizacija matrice A je

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

Vektor $y = L^{-1}b = \left(\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right)$ je izračunat bez greške Ako pogledamo proces supstitucija unazad, egzaktne formule $x_2 = 1/\beta$, $x_1 = 0$, prelaze u

$$\begin{aligned} \tilde{x}_2 &= \frac{1}{\beta}(1 + \xi_1), \\ \tilde{x}_1 &= \frac{1}{\alpha}(1 - \beta\tilde{x}_2(1 + \xi_2))(1 + \xi_3)(1 + \xi_4) = \frac{1}{\alpha}(-\xi_1 - \xi_2 - \xi_1\xi_2)(1 + \xi_3)(1 + \xi_4), \end{aligned}$$

gdje su $\xi_1, \xi_2, \xi_3, \xi_4$ male greške reda veličine strojne točnosti. Primijetimo da općenito ne možemo garantirati $\tilde{x}_1 = 0$. Ako je β strojni broj takav da je $\beta \odot (1 \oslash \beta) \neq 1$, bit će $\tilde{x}_1 \neq 0$. Na primjer u IEEE aritmetici je takav broj npr. $\beta = 4.057062130620955e{-}001$, pri čemu je $\beta \odot (1 \oslash \beta) = 9.999999999999999e{-}001$. (Čitatelju za vježbu ostavljamo da pokuša naći još takvih brojeva.)
Pokušajmo sada izračunato rješenje $\tilde{x}_1, \tilde{x}_2$ interpretirati kao egzaktno rješenje sustava $(A + \delta A)\tilde{x} = b + \delta b$, gdje su elementi matrice δA oblika $a_{ij}(1 + \epsilon_{ij})$, a elementi od δb su $b_i(1 + \varepsilon_i)$. Drugim riječima, treba odrediti ϵ_{ij}, ε_i, $i, j = 1, \ldots, n$, tako da vrijedi

$$\begin{pmatrix} \alpha(1+\epsilon_{11}) & \beta(1+\epsilon_{12}) \\ \alpha(1+\epsilon_{21}) & 0 \end{pmatrix} \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} = \begin{pmatrix} 1+\varepsilon_1 \\ 0 \end{pmatrix}. \tag{3.4.8}$$

Ako je $\tilde{x}_1 \neq 0$, jasno je da je za zadovoljavanje druge jednadžbe u gornjem sustavu nužno uzeti $\epsilon_{21} = -1$, što znači da element a_{21} treba promijeniti u nulu. Znači da imamo veliku promjenu elementa, $\delta a_{21} = -a_{21}$, tj. $(\delta A)_{22} = -\alpha$, pa je i

$$\frac{\|\delta A\|_F}{\|A\|_F} \geq \frac{|\alpha|}{\sqrt{3}|\alpha|} > \frac{1}{2}.$$

Iz prethodnog primjera zaključujemo da bez obzira na točnost koju koristimo u računanju na stroju, općenito ne možemo garantirati da će izračunato rješenje $\tilde{x}$ biti točno rješenje sustava $\tilde{A}\tilde{x} = \tilde{b}$ u kojem su $\tilde{A}$ i $\tilde{b}$ nastali malim relativnim perturbacijama koeficijenata u A i b.

3.4.4 Dodatak: Udaljenost matrice do skupa singularnih matrica

Neka je A $n \times n$ regularna matrica. Pokušajmo joj naći najbližu singularnu matricu, pri čemu mjerimo u nekoj matričnoj normi $\|\cdot\|$. Neka je $A + \Delta A$ singularna. Iz $A + \Delta A = A(I + A^{-1}\Delta A)$ slijedi da je $I + A^{-1}\Delta A$ singularna. (I više, $A + \Delta A$ i $I + A^{-1}\Delta A$ su istog ranga.) Tada je nužno $\|A^{-1}\Delta A\| \geq 1$. Naime, $\|A^{-1}\Delta A\| < 1$ povlači (zbog propozicije **??**) da je $I + A^{-1}\Delta A$ regularna. Dakle, $1 \leq \|A^{-1}\Delta A\| \leq \|A^{-1}\|\|\Delta A\|$, pa je

$$\|\Delta A\| \geq \frac{1}{\|A^{-1}\|}.$$

Time smo dokazali:

Propozicija 3.4.6. *Neka je A regularna matrica i $\|\cdot\|$ proizvoljna matrična norma. Tada je*

$$\inf_{\det(X)=0} \|A - X\| \geq \frac{1}{\|A^{-1}\|}.$$

3.5 Jacobijeva, Gauss–Seidelova i SOR metoda

Jacobijeva, Gauss–Seidelova i SOR metoda spadaju u klasične iterativne metode. Povijesno, najvažniji period razvoja tih metoda je počeo pedesetih godina dvadesetog stoljeća, u kontekstu numeričkog rješavanja eliptičkih parcijalnih diferencijalnih jednadžbi, kada počinje intenzivno i sustavno korištenje računala u numeričkoj matematici. Fundamentalne doprinose teoriji ovih metoda je dao David M. Young u svojoj doktorskoj disertaciji 1950. godine u kojoj je detaljno opisana SOR metoda. Detaljna analiza ovih metoda se može naći npr. u knjigama Davida Younga [3] i Richarda Varge [2].

Iako spadaju u klasične i dobro istražene metode, sve tri su još uvijek predmet interesa i istraživanja. Jedan razlog je jednostavnost tih metoda i kao takve su izvrstan materijal za uvod u teoriju i praksu iterativnih metoda za sustave jednadžbi. Drugi razlog je da su uistinu efikasne za rješavanje određene klase problema, a mogu se koristiti i kao komponente u sofisticiranijim metodama. Vrijedi i spomenuti da je proučavanje tih metoda motiviralo razvoj netrivijalnih i elegentnih rezultata u teoriji matrica.

3.5.1 Opis metodâ

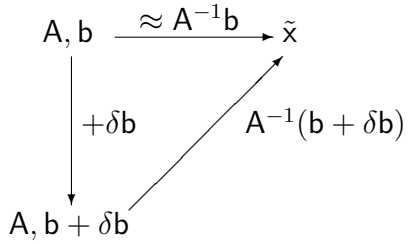
Pije nego počnemo s opisom metoda, prisjetimo se diskusije iz primjera u prethodnoj sekciji. Ako rješavamo sustav $\mathsf{A}x = \mathsf{b}$, $\det(\mathsf{A}) \neq 0$, te ako smo svjesni da ćemo umjesto egzaktnog rješenja x morati koristiti neku aproksimaciju $\tilde{\mathsf{x}}$, moramo imati način procjene kvalitete aproksimacije. Jedan način je, očito, izračunati normu razlike $\delta\mathsf{x} = \tilde{\mathsf{x}} - \mathsf{x}$, ali to nije izvedivo jer ne znamo x. Drugi način je da izračunamo rezidual

$$\mathsf{r} = \mathsf{b} - \mathsf{A}\tilde{\mathsf{x}}, \tag{3.5.1}$$

pa $\tilde{\mathsf{x}}$ prihvatimo kao dobru aproksimaciju ako je, u nekoj normi, $\epsilon \equiv \|\mathsf{r}\|/\|\mathsf{b}\|$ dovoljno mali. Takav kriterij možemo opravdati činjenicom da je iz (3.5.1)

$$\mathsf{A}\tilde{\mathsf{x}} = \tilde{\mathsf{b}} \equiv \mathsf{b} - \mathsf{r}, \ \frac{\|\tilde{\mathsf{b}} - \mathsf{b}\|}{\|\mathsf{b}\|} = \frac{\|\mathsf{r}\|}{\|\mathsf{b}\|} = \epsilon.$$

Kažemo da $\tilde{\mathsf{x}}$ egzaktno rješava sustav koji je blizak zadanom, sa kontroliranom razlikom u desnoj strani sustava (vektoru b).
Prethodna diskusija nas motivira da potražimo i drugačije pristupe za rješavanje linearnog sustava $\mathsf{Ax} = \mathsf{b}$. Primijetimo da ne moramo nužno težiti pronalaženju



Slika 3.7: Približno rješenje sustava kao egzaktno rješenje perturbiranog sustava.

egzaktnog rješenja – u uvjetima konačne strojne aritmetike se niti Gaussove eliminacije ne mogu izvesti egzaktno pa smo prinuđeni raditi s aproksimacijama. Dakle, želimo *dovoljno dobru* aproksimaciju $\tilde{\mathsf{x}} \approx \mathsf{A}^{-1}\mathsf{b}$. Zato ima smisla pokušati konstruirati niz $x^{(0)}, x^{(1)}, \ldots, x^{(k)}, \ldots$ vektora sa sljedećim svojstvima:

(i) Za svaki k je formula za računanje $x^{(k)}$ jednostavna i matrica A se koristi samo kao funkcijski potprogram koji računa $v \mapsto f(\mathsf{A})v$, gdje je v vektor a $f(\mathsf{A})$ označava A, A^*, A^{T} ili neki dio od A (npr. dijagonalni dio od A, gornji ili donji trokut od A i sl.).

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

Nabrojana svojstva su namjerno za sada dana u nepreciznoj formi. Detalji, koji ovise o konkretnom problemu i o konkretnom načinu konstruiranja niza $(x^{(k)})$, će biti dani malo kasnije.

Napišimo matricu A kao $\mathsf{A} = \mathsf{M} - \mathsf{N}$, gdje je M regularna matrica i $\mathsf{M}^{-1}\mathsf{A} \approx \mathsf{I}$. Polazni sustav napišimo kao

$$\mathsf{M}\mathsf{x} = \mathsf{N}\mathsf{x} + b, \quad \text{tj. } \mathsf{x} = \underbrace{\mathsf{M}^{-1}\mathsf{N}}_{\mathsf{F}}\mathsf{x} + \underbrace{\mathsf{M}^{-1}\mathsf{b}}_{\mathsf{c}}. \tag{3.5.2}$$

To je problem fiksne točke, $\mathsf{x} = \mathsf{F}\mathsf{x} + \mathsf{c}$, sa $\mathsf{F} = \mathsf{M}^{-1}\mathsf{N} = \mathsf{I} - \mathsf{M}^{-1}\mathsf{A}$, pa je prirodno pokušati jednostavne iteracije

$$x^{(k+1)} = \mathsf{F}x^{(k)} + \mathsf{c}. \tag{3.5.3}$$

Primijetimo da gornje iteracije možemo ekvivalentno pisati i kao

$$x^{(k+1)} = x^{(k)} + \mathsf{M}^{-1}\mathsf{r}_k, \text{ gdje je } \mathsf{r}_k = \mathsf{b} - \mathsf{A}x^{(k)}. \tag{3.5.4}$$

Intuitivno, ako je M odabrana tako da znamo efikasno koristiti M^{-1} i tako da je $\mathsf{M}^{-1} \approx \mathsf{A}^{-1}$ onda je

$$x^{(k+1)} = x^{(k)} + \mathsf{M}^{-1}\mathsf{r}_k \approx x^{(k)} + \mathsf{A}^{-1}\mathsf{r}_k = x^{(k)} + \mathsf{x} - x^{(k)} = \mathsf{x}$$

Komentar 3.5.1. Kažemo da je M prekondicioner za A i pri tome mislimo na činjenicu da M^{-1} aproksimira A^{-1}. Ako imamo dobar izbor matrice M, onda polazni sustav možemo zamijeniti ekvivalentnim $(\mathsf{M}^{-1}\mathsf{A})\mathsf{x} = \mathsf{M}^{-1}\mathsf{b}$, čija matrica koeficijenata $\mathsf{M}^{-1}\mathsf{A}$ se ponaša puno bolje od A.

Igra se sastoji u tome kako odabrati rastav $\mathsf{A} = \mathsf{M} - \mathsf{N}$ koji će za neka klase problema osigurati konvergenciju. Za Jacobijevu, Gauss–Seidelovu i SOR metodu su ti rastavi izvedeni iz sljedeće reprezentacije matrice A:

$$\mathsf{A} = \mathsf{D}(\mathsf{I} - \mathsf{L} - \mathsf{U}),\ \mathsf{D} = \operatorname{diag}(\mathsf{A}),\ \begin{array}{l}\mathsf{L} \text{ strogo donje trokutasta,}\\ \mathsf{U} \text{ strogo gornje trokutasta.}\end{array} \tag{3.5.5}$$

Koristit ćemo i rastav $\mathsf{A} = \mathsf{D} - \hat{\mathsf{L}} - \hat{\mathsf{U}}$, $\hat{\mathsf{L}} = \mathsf{DL}$, $\hat{\mathsf{U}} = \mathsf{DU}$.

3.5.1.1 Jacobijeva metoda

Jacobijevu metodu definiramo za matricu $\mathsf{A} \in \mathbb{M}_n$ za koju je $a_{ii} \neq 0$, $i = 1, \ldots, n$. Za matricu M ćemo uzeti samo dijagonalu, $\mathsf{M} = \operatorname{diag}(\mathsf{A})$, pa je $\mathsf{N} = \mathsf{M} - \mathsf{A}$ izvandijagonalni dio od $-\mathsf{A}$. Pripadnu matricu $\mathsf{F} = \mathsf{D}^{-1}(\mathsf{D} - \mathsf{A})$ ćemo označiti s J, gdje je u terminima (3.5.5)

$$\mathsf{J} = \mathsf{L} + \mathsf{U}. \tag{3.5.6}$$

Jacobijeve iteracije $x^{(k+1)} = \mathsf{J}x^{(k)} + \mathsf{D}^{-1}\mathsf{b}$ na nivou elemenata glase:

$$x_i^{(k+1)} = \frac{1}{a_{ii}}(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j^{(k)}) \tag{3.5.7}$$

Algoritam 3.5.1. Jacobijeva metoda: $x^{(k+1)} = \mathsf{J}x^{(k)} + \mathsf{D}^{-1}\mathsf{b}$

$x = \text{Jacobi}(\mathsf{A}, \mathsf{b}, x^{(0)}, k_{\max})$

for $k = 1, 2, \ldots, k_{\max}$
  for $i = 1, 2, 3, \ldots, n$
    $x_i^{(k)} = \frac{1}{a_{ii}}(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j^{(k-1)});$
  end
  if $x^{(k)}$ *"dovoljno dobar"*
    $x = x^{(k)}$; *return*
  end
end

3.5.1.2 Gauss–Seidelova metoda

Gledajući ključnu formulu (3.5.7) u Jacobijevoj metodi primjećujemo da su u momentu računanja $x_i^{(k+1)}$ već poznate vrijednosti $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$. Kako očekujemo da naše iteracije konvergiraju, razumno je očekivati da su $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$ bolje vrijednosti od $x_1^{(k)}, \ldots, x_{i-1}^{(k)}$. Zato uvodimo sljedeću formulu za računanje $x^{(k+1)}$:

$$x_i^{(k+1)} = \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}),\ i = 1, \ldots, n. \tag{3.5.8}$$

Drugim riječima, uvijek koristiti najsvježije podatke. Da bismo ovaj novi proces napisali u generičkom obliku (3.5.3), primijetimo da iz (3.5.8) slijedi

$$\sum_{j=1}^{i} a_{ij}x_j^{(k+1)} = b_i - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)},\ i = 1, \ldots, n.$$

U terminima reprezentacije (3.5.5) to zvuči kao

$$\begin{aligned} \mathsf{D}(\mathsf{I} - \mathsf{L})x^{(k+1)} &= \mathsf{D}\mathsf{U}x^{(k)} + \mathsf{b},\ \text{tj.} \\ x^{(k+1)} &= \mathsf{G}x^{(k)} + \underbrace{(\mathsf{I} - \mathsf{L})^{-1}\mathsf{D}^{-1}\mathsf{b}}_{\mathsf{g}},\ \ \mathsf{G} = (\mathsf{I} - \mathsf{L})^{-1}\mathsf{U}. \end{aligned} \tag{3.5.9}$$

Iteracije (3.5.9) definiraju Gauss–Seidelovu metodu.

3.5.1.3 Metode SOR i JOR

Ideja SOR metode je jednostavna: U Gauss–Seidelovu metodu ubaciti jedan slobodan parametar $\omega \in \mathbb{R}$ i onda ga pokušati naštimati tako da dobijemo bržu konvergenciju. Pri tome zadržavamo osnovnu ideju Gauss–Seidelove metode – uvijek koristiti najsvježije podatke:

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}),\ i = 1, \ldots, n. \tag{3.5.10}$$

Matrično, vidimo da vrijedi

$$\begin{aligned} \mathsf{D}(\mathsf{I} - \omega\mathsf{L})x^{(k+1)} &= ((1 - \omega)\mathsf{D} + \omega\mathsf{D}\mathsf{U})x^{(k)} + \omega\mathsf{b},\ \text{tj.} \\ x^{(k+1)} &= \mathsf{S}x^{(k)} + \underbrace{\omega(\mathsf{I} - \omega\mathsf{L})^{-1}\mathsf{D}^{-1}\mathsf{b}}_{s}, \end{aligned} \tag{3.5.11}$$

$$\mathsf{S} \equiv \mathsf{S}_\omega = (\mathsf{I} - \omega\mathsf{L})^{-1}((1 - \omega)\mathsf{I} + \omega\mathsf{U}). \tag{3.5.12}$$

Iteracije (3.5.11) definiraju metodu SOR(ω). Očito je SOR(1) jednak Gauss–Seidelovoj metodi. Primijetimo da možemo ekvivalentno pisati

$$\mathsf{M}x^{(k+1)} = \mathsf{N}x^{(k)} + \mathsf{b}, \;\; \mathsf{M} = \frac{1}{\omega}\mathsf{D} - \hat{\mathsf{L}}, \;\; \mathsf{N} = (\frac{1}{\omega} - 1)\mathsf{D} + \hat{\mathsf{U}}. \tag{3.5.13}$$

Na isti način možemo definirati metodu JOR(ω), sa matricom iteracija

$$\mathsf{J}_\omega = \omega\mathsf{J} + (1-\omega)\mathsf{I}. \tag{3.5.14}$$

Naravno, JOR(1) je Jacobijeva metoda.

3.5.1.4 SSOR metoda

U dosta primjena je matrica A hermitska, specijalno i realna simetrična. Poželjno je da M kao prekondicioner također bude simetrična, kao i da matrica koja generira iteracije ima neka od svojstava koja slijede iz simetrije, npr. realne svojstvene vrijednosti.

Cilj nam je "simetrizirati" SOR metodu. Primijetimo da prijelaz sa Jacobijeve na Gauss–Seidelovu i SOR metodu možemo napraviti i tako da nepoznanice računamo od zadnje prema prvoj.

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k+1)}), \; i = n, \ldots, 1. \tag{3.5.15}$$

Matrično, vidimo da vrijedi

$$\begin{aligned} \mathsf{D}(\mathsf{I} - \omega\mathsf{U})x^{(k+1)} &= ((1-\omega)\mathsf{D} + \omega\mathsf{DL})x^{(k)} + \omega\mathsf{b}, \text{ tj.} \\ x^{(k+1)} &= \mathsf{S}^{\downarrow}x^{(k)} + \underbrace{\omega(\mathsf{I} - \omega\mathsf{U})^{-1}\mathsf{D}^{-1}\mathsf{b}}_{s^{\downarrow}}, \qquad (3.5.16) \\ \mathsf{S}^{\downarrow} \equiv \mathsf{S}_\omega^{\downarrow} &= (\mathsf{I} - \omega\mathsf{U})^{-1}((1-\omega)\mathsf{I} + \omega\mathsf{L}). \qquad (3.5.17) \end{aligned}$$

Sada kombiniramo ove dvije iteracije

$$\begin{aligned} x^{(k+1/2)} &= \mathsf{S}_\omega x^{(k)} + s \qquad (3.5.18) \\ x^{(k+1)} &= \mathsf{S}_\omega^{\downarrow}x^{(k+1/2)} + s^{\downarrow} = \mathsf{S}_\omega^{\downarrow}(\mathsf{S}_\omega x^{(k)} + s) + s^{\downarrow} \qquad (3.5.19) \end{aligned}$$

Ako prijelaz s $x^{(k)}$ na $x^{(k+1)}$ želimo prikazati u obliku $\mathsf{M}x^{(k+1)} = \mathsf{N}x^{(k)} + \mathsf{b}$. Za to je dovoljno u (3.5.19) otkriti faktor uz b. Matrica koja množi b je

$$\begin{aligned}
\mathsf{M}^{-1} &= (\mathsf{I}-\omega\mathsf{U})^{-1}((1-\omega)\mathsf{I}-\omega\mathsf{L})\omega(\mathsf{I}-\omega\mathsf{L})^{-1}\mathsf{D}^{-1} + \omega(\mathsf{I}-\omega\mathsf{U})^{-1}\mathsf{D}^{-1} \\
&= (\frac{1}{\omega}\mathsf{D}-\hat{\mathsf{U}})^{-1}\left\{\mathsf{I}+\left[(\frac{1}{\omega}-1)\mathsf{D}+\hat{\mathsf{L}}\right](\frac{1}{\omega}\mathsf{D}-\hat{\mathsf{L}})^{-1}\right\} \\
&= (\frac{2}{\omega}-1)(\frac{1}{\omega}\mathsf{D}-\hat{\mathsf{U}})^{-1}\mathsf{D}(\frac{1}{\omega}\mathsf{D}-\hat{\mathsf{L}})^{-1} \text{ pa je} \\
\mathsf{M} &= \frac{\omega}{2-\omega}(\frac{1}{\omega}\mathsf{D}-\hat{\mathsf{L}})\mathsf{D}^{-1}(\frac{1}{\omega}\mathsf{D}-\hat{\mathsf{U}}),\ \ \mathsf{N}=\mathsf{M}-\mathsf{A}.
\end{aligned}$$

Ako je $\mathsf{A} = \mathsf{A}^*$, onda je $\hat{\mathsf{U}} = \hat{\mathsf{L}}^*$, pa je $\mathsf{M} = \mathsf{M}^*$. Lako se vidi da opisane iteracije generira matrica $\mathsf{S}_\omega^{\downarrow\uparrow} = \mathsf{S}_\omega^{\downarrow}\mathsf{S}_\omega$, te da je

$$\begin{aligned}
\mathsf{S}_\omega^{\downarrow\uparrow} &= \mathsf{M}^{-1}\mathsf{N} = \mathsf{I}-\mathsf{M}^{-1}\mathsf{A} = \mathsf{I}-\frac{2-\omega}{\omega}(\frac{1}{\omega}\mathsf{D}-\hat{\mathsf{U}})^{-1}\mathsf{D}(\frac{1}{\omega}\mathsf{D}-\hat{\mathsf{L}})^{-1}\mathsf{A} && (3.5.20)\\
&= \mathsf{I}-\omega(2-\omega)(\mathsf{D}-\omega\hat{\mathsf{U}})^{-1}\mathsf{D}(\mathsf{D}-\omega\hat{\mathsf{L}})^{-1}\mathsf{A} && (3.5.21)\\
&= \mathsf{I}-\omega(2-\omega)(\mathsf{I}-\omega\mathsf{U})^{-1}(\mathsf{I}-\omega\mathsf{L})^{-1}\mathsf{D}^{-1}\mathsf{A}. && (3.5.22)
\end{aligned}$$

Ovime smo opisali SSOR, simetriziranu inačicu SOR metode.

Propozicija 3.5.1. *Ako je* A *hermitska matrica sa pozitivnim dijagonalnim elementima, onda matrica* $\mathsf{M}^{-1}\mathsf{N}$ *u SSOR metodi ima realne svojstvene vrijednosti.*

Dokaz: Dovoljno je dokazati da je $(1/\omega\mathsf{D}-\hat{\mathsf{U}})^{-1}\mathsf{D}(1/\omega\mathsf{D}-\hat{\mathsf{L}})^{-1}\mathsf{A}$ slična hermitskoj matrici. Kako je D pozitivno definitna i $\hat{\mathsf{U}} = \hat{\mathsf{L}}^*$, lako ostvarimo sličnost s $\mathsf{D}^{1/2}(1/\omega\mathsf{D}-\hat{\mathsf{L}})^{-1}\mathsf{A}(1/\omega\mathsf{D}-\hat{\mathsf{L}})^{-*}\mathsf{D}^{1/2}$. ⊞

3.5.2 Konvergencija Jacobijeve i Gauss–Seidelove metode

Dakle, konvergenciju jedne cijele klase iterativnih metoda možemo proučavati kao konvergenciju iteracije fiksne točke oblika (3.5.3), gdje je konvergencija pojedine metode ovisna o njenoj konkretnoj matrici F. Počnimo sa jednim dovoljnim uvjetom konvergencije iteracija (3.5.3):

Lema 3.5.2. *Neka je* $|||\cdot|||$ *operatorska norma inducirana vektorskom normom* $\|\cdot\|$. *Ako je* $|||\mathsf{F}||| < 1$, *onda iteracije* $x^{(k+1)} = \mathsf{F}x^{(k)} + c$ *konvergiraju za svaki početni* $x^{(0)}$.

Dokaz: Slijedimo klasičnu shemu: pokazat ćemo da je niz $x^{(k)}$ Cauchyjev, odakle slijedi konvergencija i to očito vektoru x za kojeg je $x = \mathsf{F}x + c$. Prvo uočimo da je

$$\begin{aligned} \|x^{(k+1)} - x^{(k)}\| &\leq |||\mathsf{F}|||\|x^{(k)} - x^{(k-1)}\| \leq |||\mathsf{F}|||^2\|x^{(k-1)} - x^{(k-2)}\| \\ &\leq \cdots \leq |||\mathsf{F}|||^k\|x^{(1)} - x^{(0)}\|, \end{aligned}$$

odakle teleskopskom sumom dobijemo za proizvoljne k, ℓ

$$\begin{aligned} \|x^{(k+\ell)} - x^{(k)}\| &= \|\sum_{j=0}^{\ell-1}(x^{(k+j+1)} - x^{(k+j)})\| \leq \sum_{j=0}^{\ell-1}|||\mathsf{F}|||^{k+j}\|x^{(1)} - x^{(0)}\| \\ &\leq \frac{|||\mathsf{F}|||^k}{1 - |||\mathsf{F}|||}\|x^{(1)} - x^{(0)}\| \longrightarrow 0 \ (k \to \infty). \end{aligned}$$

Dakle, niz je Cauchyjev i postoji $\mathsf{x} = \lim_{k\to\infty} x^{(k)}$, pri čemu je jasno $\mathsf{x} = \mathsf{F}\mathsf{x} + c$. Korisno je procijeniti ponašanje razlike između iteracija i fiksne točke. Oduzimanjem relacija

$$\begin{aligned} \mathsf{x} &= \mathsf{F}\mathsf{x} + c \\ x^{(k)} &= \mathsf{F}x^{(k-1)} + c \ \text{ dobijemo } \ \mathsf{x} - x^{(k)} = \mathsf{F}(\mathsf{x} - x^{(k-1)}), \end{aligned}$$

pa je $\|\mathsf{x} - x^{(k)}\| \leq |||\mathsf{F}|||\|\mathsf{x} - x^{(k-1)}\| \leq \cdots \leq |||\mathsf{F}|||^k\|\mathsf{x} - x^{(0)}\|$. ⊞

Potpuni opis stanja stvari je dan teoremom o nužnim i dovoljnim uvjetima:

Teorem 3.5.3. *Iteracije* $x^{(k+1)} = \mathsf{F}x^{(k)} + c$ *konvergiraju fiksnoj točki* x *za svaki početni* $x^{(0)}$ *ako i samo ako je* $\mathrm{spr}(\mathsf{F}) < 1$.

Dokaz: Ako je $\mathrm{spr}(\mathsf{F}) < 1$ onda po Teoremu 1.5.1 postoji operatorska norma tako da je $|||\mathsf{M}||| < 1$, pa konvergencija slijedi iz Leme 3.5.2. S druge strane, neka je $\mathrm{spr}(\mathsf{F}) \geq 1$, i neka je $|\lambda| = \mathrm{spr}(\mathsf{F})$, $\mathsf{F}\mathsf{v} = \lambda\mathsf{v}$, $\mathsf{v} \neq \mathbf{0}$. Neka je $x^{(0)} = \mathsf{x} - \mathsf{v}$. Tada je $\mathsf{x} - x^{(k)} = \mathsf{F}^k(\mathsf{x} - x^{(0)}) = \lambda^k(\mathsf{x} - x^{(0)})$, tj. $\|\mathsf{x} - x^{(k)}\| = |\lambda|^k\|\mathsf{x} - x^{(0)}\|$ ne konvergira u nulu kada $k \to \infty$. ⊞

Dakle, u slučaju Jacobijeve i Gauss–Seidelove metode nam preostaje proučiti spektralne radijuse njihovih matrica. Za očekivati je da će trebati dodatni uvjeti kako bi odgovarajući spektralni radijus bio manji od jedan. Sljedeći teorem je jedan od ključnih za važne klase problema.

Teorem 3.5.4. *(Stein i Rosenberg) Neka je* $\mathsf{J} = \mathsf{L} + \mathsf{U}$ *nenegativna matrica, pri čemu je* L *strogo donje trokutasta, a* U *strogo gornje trokutasta matrica. Neka je* $\mathsf{G} = (\mathsf{I} - \mathsf{L})^{-1}\mathsf{U}$. *Tada vrijedi samo jedna od sljedeće četiri tvrdnje:*

1. $\mathrm{spr}(\mathsf{J}) = \mathrm{spr}(\mathsf{G}) = 0$.

2. $\mathrm{spr}(\mathsf{J}) = \mathrm{spr}(\mathsf{G}) = 1$.

3. $0 < \mathrm{spr}(\mathsf{G}) < \mathrm{spr}(\mathsf{J}) < 1$.

4. $1 < \mathrm{spr}(\mathsf{J}) < \mathrm{spr}(\mathsf{G})$.

Dokaz: Jasno je da su L i U nenegativne. Sada uočimo da je $\mathsf{G} \geq \mathbf{0}$, zato što je

$$(\mathsf{I} - \mathsf{L})^{-1} = \mathsf{I} + \mathsf{L} + \mathsf{L}^2 + \cdots + \mathsf{L}^{n-1} \geq \mathbf{0}.$$

Kao nenegativna matrica, G ima nenegativan svojstveni vektor v koji pripada svojstvenoj vrijednosti $\lambda = \mathrm{spr}(\mathsf{G})$, $(\mathsf{I} - \mathsf{L})^{-1}\mathsf{U}\mathsf{v} = \lambda\mathsf{v}$. Ovu relaciju možemo zapisati i kao $(\lambda\mathsf{L} + \mathsf{U})\mathsf{v} = \lambda\mathsf{v}$. Uočimo da je G uvijek reducibilna.

Ako pretpostavimo da je J ireducibilna, onda možemo zaključiti da je $\lambda > 0$. Ako bi bilo $\mathrm{spr}(\mathsf{G}) = 0$, onda bi ireducibilna normalna forma $\mathsf{F} = \Pi^{\mathrm{T}}\mathsf{G}\Pi$ matrice $\mathsf{G} = \sum_{j=0}^{n-1} \mathsf{L}^j\mathsf{U}$ morala biti strogo gornje trokutasta matrica. Pokažimo da to nije moguće: Neka je $\mu = \mathrm{spr}(\mathsf{J}) > 0$, sa pripadnim svojstvenim vektorom $\mathsf{w} > \mathbf{0}$, $(\mathsf{L} + \mathsf{U})\mathsf{w} = \mu\mathsf{w}$. Množenjem s $(\mathsf{I} - \mu^{-1}\mathsf{L})^{-1}$ i jednostavnom algebarskom manipulacijom dobijemo $(\mathsf{I} - \mu^{-1}\mathsf{L})^{-1}\mathsf{U}\mathsf{w} = \mu\mathsf{w}$. Kako G i $(\mathsf{I} - \mu^{-1}\mathsf{L})^{-1}\mathsf{U} = \sum_{j=0}^{n-1} \mu^{-j}\mathsf{L}^j\mathsf{U}$ imaju točno isti raspored nula, $\Pi^{\mathrm{T}}(\mathsf{I} - \mu^{-1}\mathsf{L})^{-1}\mathsf{U}\Pi$ bi također morala biti strogo gornje trokutasta, što očito nije moguće.

Tada je $\mathsf{S} \equiv \lambda\mathsf{L} + \mathsf{U}$ također nenegativna ireducibilna matrica. Iz Propozicije 2.1.2 znamo da je $(\mathsf{I} + \mathsf{S})^{n-1} > \mathbf{0}$, pa kombiniranjem s $(\mathsf{I} + \mathsf{S})^{n-1}\mathsf{v} = (1 + \lambda)^{n-1}\mathsf{v}$ odmah vidimo da je $\mathsf{v} > 0$. Sada primjenom drugog dijela Propozicije 2.2.1 zaključujemo da je $\mathrm{spr}(\lambda\mathsf{L} + \mathsf{U}) = \lambda$.

Na isti način, koristeći relaciju $(\mathsf{L} + (1/\lambda)\mathsf{U})\mathsf{v} = \mathsf{v}$ dobijemo da je $\mathrm{spr}(\mathsf{L} + (1/\lambda)\mathsf{U}) = 1$. Sada je sve spremno za konačno zaključivanje (uz pretpostavku da je J ireducibilna):

- Ako je $\mathrm{spr}(\mathsf{J}) \equiv \mathrm{spr}(\mathsf{L} + \mathsf{U}) = 1$, onda gledajući u upravo dokazanu relaciju

$$\mathrm{spr}(\mathsf{L} + \frac{1}{\mathrm{spr}(\mathsf{G})}\mathsf{U}) = 1 \tag{3.5.23}$$

  zaključujemo da bi $\mathrm{spr}(\mathsf{G}) \neq 1$, u kombinaciji sa strogom monotonosti spektralnog radijusa (druga tvrdnja u Teoremu 2.2.8), vodilo na kontradikciju. Dakle $\mathrm{spr}(\mathsf{G}) = 1$.

- Slično zaključujemo ako je $0 < \text{spr}(\mathsf{J}) < 1$. Ako bi bio $1/\text{spr}(\mathsf{G}) < 1$, onda bi $\text{spr}(\mathsf{J}) < 1$, stroga monotonost i relacija (3.5.23) bile u kontradikciji. Dakle, $1/\text{spr}(\mathsf{G}) > 1$ tj. $0 < \text{spr}(\mathsf{G}) < 1$. Nadalje, pokazali smo ranije da je

$$\text{spr}(\text{spr}(\mathsf{G})\mathsf{L} + \mathsf{U}) = \text{spr}(\mathsf{G}), \tag{3.5.24}$$

  pa je zbog monotonosti $\text{spr}(\mathsf{G}) < \text{spr}(1\mathsf{L}+\mathsf{U}) = \text{spr}(\mathsf{J})$. Sve zajedno, pokazali smo da je $0 < \text{spr}(\mathsf{G}) < \text{spr}(\mathsf{J}) < 1$.

- Neka je $\text{spr}(\mathsf{J}) > 1$. Ako bi vrijedilo $1/\text{spr}(\mathsf{G}) > 1$, onda bi vrijedilo $\mathsf{L} + \frac{1}{\text{spr}(\mathsf{G})}\mathsf{U} \geq \mathsf{L} + \mathsf{U}$ i za barem jedan element bi vrijedila stroga nejednakost pa bi (3.5.23) i monotonost spektralnog radijusa implicirali $\text{spr}(\mathsf{L} + \mathsf{U}) < 1$, što je suprotno pretpostavci. Dakle, $\text{spr}(\mathsf{G}) > 1$. Ali tada je zbog (3.5.24) $\text{spr}(\mathsf{G}) > \text{spr}(\mathsf{L} + \mathsf{U}) = \text{spr}(\mathsf{J})$.

Ovime je teorem dokazan za slučaj ireducibilne nenegativne matrice J (koji isključuje mogućnost $\text{spr}(\mathsf{J}) = 0$). Sada promotrimo slučaj kada je J reducibilna:

- Neka je $\text{spr}(\mathsf{J}) = 0$. Ako bismo konstruirali ireducibilnu normalnu formu (2.1.2) od J, onda bi s nekom permutacijom P imali $P^T\mathsf{J}P$ jednaku strogo gornje trokutastoj matrici. (Jer je spektralni radijus od J jednak nuli, ne mogu se dobiti netrivijalni ireducibilni blokovi.) Sada lako provjerimo da je $P^T\mathsf{G}P$ također strogo gornje trokutasta, pa je $\text{spr}(\mathsf{G}) = 0$.

- Neka je $\text{spr}(\mathsf{J}) > 0$.

⊞

Teorem 3.5.5. *Neka je* $\mathsf{A} \in \mathbb{M}_n$ *strogo dijagonalno domimantna ili ireducibilno dijagonalno dominantna matrica. Tada su i Jacobijeva i Gauss–Seidelova metoda konvergentne sa svakom početnom iteracijom.*

Dokaz: Promotrimo prvo Jacobijevu metodu, $x^{(k+1)} = \mathsf{J}x^{(k)} + \mathsf{c}$, $\mathsf{J}_{ij} = (\boldsymbol{\delta}_{ij} - 1)a_{ij}/a_{ii}$. Ako je A strogo dijagonalno dominantna, onda odmah zaključujemo da vrijedi $\text{spr}(\mathsf{J}) \leq \text{spr}(|\mathsf{J}|) \leq \max_{i=1:n} \sum_{j=1}^{n} |\mathsf{J}_{ij}| < 1$ (Svi Geršgorinovi krugovi su sa centrom u ishodištu i radijusom strogo manjim od jedan.)

Ako je A ireducibilno dijagonalno dominantna, onda je $|\mathsf{J}|$ također ireducibilna, vrijedi $\max_{i=1:n} \sum_{j=1}^{n} |\mathsf{J}_{ij}| \leq 1$ i za barem jedan indeks i_* je $\sum_{j=1}^{n} |\mathsf{J}_{i_*j}| < 1$. Ako bi $|\mathsf{J}|$ imala svojstvenu vrijednost λ apsolutne vrijednosti jedan, onda λ ne bi mogla biti unutarnja točka niti jednog Geršgorinovog kruga. Tada bi prema Teoremu

2.1.3 sve Geršgorinove kružnice morale prolaziti kroz λ, što bi bilo u kontradikciji sa strogom nejednakosti u retku s indeksom i_*. Dakle, opet imamo $\mathrm{spr}(\mathsf{J}) \leq \mathrm{spr}(|\mathsf{J}|) < 1$. Time je dokazana konvergencija Jacobijeve metode.

Kod Gauss–Seidelove metode $x^{(k+1)} = \mathsf{G}x^{(k)} + \mathsf{d}$ je $\mathsf{G} = (\mathsf{I} - \mathsf{L})^{-1}\mathsf{U}$. Uočimo da je $(\mathsf{I} - \mathsf{L})^{-1} = \sum_{k=0}^{n-1} \mathsf{L}^k$ te je

$$|(\mathsf{I} - \mathsf{L})^{-1}| \leq \sum_{k=0}^{n-1} |\mathsf{L}|^k = (\mathsf{I} - |\mathsf{L}|)^{-1},$$

odakle slijedi $|\mathsf{G}| \leq \tilde{\mathsf{G}} \equiv (\mathsf{I} - |\mathsf{L}|)^{-1}|\mathsf{U}|$. Dovoljno je pokazati da je $\mathrm{spr}(\tilde{\mathsf{G}}) \leq \mathrm{spr}(|\mathsf{J}|)$. Kako je $|\mathsf{J}| = |\mathsf{L}| + |\mathsf{U}|$, nalazimo se u uvjetima Teorema 3.5.4, a kako smo već pokazali da je $\mathrm{spr}(|\mathsf{J}|) < 1$, u tom teoremu je istinita prva ili treća tvrdnja. U svakom slučaju je $\mathrm{spr}(\tilde{\mathsf{G}}) \leq \mathrm{spr}(|\mathsf{J}|) < 1$, pa Gauss–Seidelova metoda konvergira. ⊞

Važna klasa problema su sustavi jednadžbi sa hermitskom (ili simetričnom) pozitivno definitnom matricom koeficijenata. Za proučavanje konvergencije su korisne razne ekvivalentne karakterizacije nužnog i dovoljnog uvjeta.

Teorem 3.5.6. *Neka je* A *pozitivno definitna i neka je za sustav* $\mathsf{Ax} = \mathsf{b}$ *konstruirana iterativna metoda* $x^{(k+1)} = \mathsf{F}x^{(k)} + c$. *Stavimo* $\mathsf{Q} = \mathsf{A}(\mathsf{I} - \mathsf{F})^{-1}$ *i* $\mathsf{H} = \mathsf{Q} + \mathsf{Q}^* - \mathsf{A}$. *Tada je* $\mathrm{spr}(\mathsf{F}) < 1$ *ako i samo ako je* H *pozitivno definitna.*

Dokaz: Stavimo $\mathsf{G} = \mathsf{A}^{1/2}\mathsf{F}\mathsf{A}^{-1/2}$. Kako je $\mathsf{F} = \mathsf{I} - \mathsf{Q}^{-1}\mathsf{A}$, imamo $\mathsf{G} = \mathsf{I} - \mathsf{A}^{1/2}\mathsf{Q}^{-1}\mathsf{A}^{1/2}$. Sada lako provjerimo da je

$$\mathsf{G}\mathsf{G}^* = \mathsf{I} - \mathsf{A}^{1/2}\mathsf{Q}^{-1}\mathsf{H}\mathsf{Q}^{-*}\mathsf{A}^{1/2}.$$

Ako je H pozitivno definitna, onda je i $\mathsf{A}^{1/2}\mathsf{Q}^{-1}\mathsf{H}\mathsf{Q}^{-*}\mathsf{A}^{1/2}$ pozitivno definitna pa je

$$\|\mathsf{G}\|_2 = \sqrt{\mathrm{spr}(\mathsf{G}\mathsf{G}^*)} = \sqrt{1 - \lambda_{\min}(\mathsf{A}^{1/2}\mathsf{Q}^{-1}\mathsf{H}\mathsf{Q}^{-*}\mathsf{A}^{1/2})} < 1.$$

Dakle, $\|\mathsf{A}^{1/2}\mathsf{F}\mathsf{A}^{-1/2}\|_2 = \mathrm{spr}(\mathsf{F}) < 1$. Možemo zaključivati i u drugom smjeru: Ako je $\mathrm{spr}(\mathsf{F}) < 1$, onda je $\lambda_{\min}(\mathsf{A}^{1/2}\mathsf{Q}^{-1}\mathsf{H}\mathsf{Q}^{-*}\mathsf{A}^{1/2}) > 0$, pa je H pozitivno definitna. ⊞

Teorem 3.5.7. *Neka je* A *hermitska regularna matrica sa pozitivnim dijagonalnim elementima. Tada je metoda JOR*(ω) *konvergentna ako i samo su* A *i* $2\omega^{-1}\mathsf{D} - \mathsf{A}$ *pozitivno definitne.*

Dokaz: Primjenjujemo Teorem 3.5.6: U metodi JOR(ω) je $\mathsf{F} = \mathsf{J}_\omega = \omega\mathsf{J} + (1-\omega)\mathsf{I}$. (Za $\omega = 1$ imamo Jacobijevu metodu.) Odmah vidimo da je $\mathsf{A}(\mathsf{I} - \mathsf{J}_\omega)^{-1} = \omega^{-1}\mathsf{D}$. Sada pozitivna definitnost matrica A i $2\omega^{-1}\mathsf{D} - \mathsf{A}$, po Teoremu 3.5.6, povlači $\mathrm{spr}(\mathsf{J}_\omega) < 1$, tj. konvergenciju metode JOR(ω).

Obratno, neka JOR(ω) konvergira, tj. neka je $\mathrm{spr}(\mathsf{J}_\omega) < 1$. Ako su $\lambda_1, \ldots, \lambda_n$ svojstvene vrijednosti od J, onda $\mathrm{spr}(\mathsf{J}_\omega) < 1$ znači da je

$$-1 < \omega\lambda_i + 1 - \omega < 1,\ i = 1, \ldots, n.$$

Kako je $\lambda_{\min} \leq 0$ (jer je $\mathrm{trag}(\mathsf{J}) = 0$), odmah vidimo da je $\omega > 0$, a nakon toga i

$$1 - \frac{2}{\omega} < \lambda_i < 1,\ i = 1, \ldots, n.$$

Zbog $1 - \lambda_i > 0$ je matrica

$$\mathsf{D}^{-1/2}\mathsf{A}\mathsf{D}^{-1/2} = \mathsf{D}^{-1/2}\mathsf{D}(\mathsf{I} - \mathsf{J})\mathsf{D}^{-1/2} = \mathsf{I} - \mathsf{D}^{1/2}\mathsf{J}\mathsf{D}^{-1/2}$$

pozitivno definitna, pa je i A pozitivno definitna. Na isti način, zbog $1 - \lambda_i < 2\omega^{-1}$, je

$$\mathsf{D}^{-1/2}(2\omega^{-1}\mathsf{D} - \mathsf{A})\mathsf{D}^{-1/2} = 2\omega^{-1}\mathsf{I} - \mathsf{D}^{1/2}(\mathsf{I} - \mathsf{J})\mathsf{D}^{-1/2}$$

pozitivno definitna pa je i $2\omega^{-1}\mathsf{D} - \mathsf{A}$ pozitivno definitna. ⊞

Propozicija 3.5.8. *Neka je* A *hermitska matrica sa pozitivnim dijagonalnim elementima i* $\mathsf{D} = \mathrm{diag}(\mathsf{A})$. *Za realni parametar* $\omega \neq 0$ *je* $2\omega^{-1}\mathsf{D} - \mathsf{A}$ *pozitivno definitna ako i samo ako je*

$$0 < \omega < \frac{2}{1 - \lambda_{\min}(\mathsf{J})}, \tag{3.5.25}$$

gdje je $\lambda_{\min}(\mathsf{J}) \leq 0$ *najmanja svojstvena vrijednost od* J.

Dokaz: Neka je $2\omega^{-1}\mathsf{D} - \mathsf{A}$ pozitivno definitna. Koristeći kongruenciju, lako zaključimo pozitivnu definitnost matrice

$$2\omega^{-1}\mathsf{I} - \mathsf{D}^{-1/2}\mathsf{A}\mathsf{D}^{-1/2} = (2\omega^{-1} - 1)\mathsf{I} + \mathsf{D}^{1/2}\mathsf{J}\mathsf{D}^{-1/2}.$$

Njene sve svojstvene vrijednosti su oblika $\alpha = 2\omega^{-1} - 1 + \lambda$, gdje λ prolazi spektrom od J. Prisjetimo se da J ima realan spektar, te da je njena najmanja svojstvena vrijednost $\lambda_{\min}$ manja ili jednaka od nule, najveća veća ili jednaka od nule (zbog $\mathrm{trag}(\mathsf{J}) = 0$). Kako je $\alpha > 0$, odmah slijedi (3.5.25). Još uočimo da izvedene relacije možemo čitati i tako da iz pretpostavljenih relacija (3.5.25) zaključimo $\alpha > 0$ tj. da je $2\omega^{-1}\mathsf{D} - \mathsf{A}$ pozitivno definitna. ⊞

3.5.3 Konvergencija SOR metode

Sada ćemo dati kratki uvod u teoriju konvergencije SOR metode. Vidjet ćemo kako prilično jednostavna shema SOR metode zahtijeva netrivijalna razmatranja. Valja uočiti i to kako su razvoj i analiza metode motivirani konkretnim klasama matrica koje se javljaju pri diskretizaciji diferencijalnih jednadžbi.

Teorem 3.5.9. *(Kahan) Neka je S_ω matrica SOR metode s parametrom ω. Tada je $\mathrm{spr}(\mathsf{S}_\omega) \geq |\omega - 1|$. Dakle, za konvergenciju SOR(ω) je nužno da je $\omega \in (0, 2)$.*

Dokaz: Vrijedi

$$\begin{aligned} \chi_{\mathsf{S}_\omega}(\lambda) &= \det(\lambda\mathsf{I} - \mathsf{S}_\omega) = \det((\mathsf{I} - \omega\mathsf{L})(\lambda\mathsf{I} - \mathsf{S}_\omega)) \\ &= \det((\lambda + \omega - 1)\mathsf{I} - \omega\lambda\mathsf{L} - \omega\mathsf{U}), \end{aligned}$$

pa računanjem karakterističnog polinoma u nuli dobijemo

$$\chi_{\mathsf{S}_\omega}(0) = (-1)^n \prod_{i=1}^{n} \lambda_i(\mathsf{S}_\omega) = \det((\omega - 1)\mathsf{I} - \omega\mathsf{U}) = (\omega - 1)^n,$$

odakle je jasno da mora biti $\max_i |\lambda_i(\mathsf{S}_\omega)| \geq |\omega - 1|$, te da metoda ne može biti konvergentna ako $\alpha \notin (0, 2)$. ⊞

Sljedeći teorem garantira konvergenciju metode SOR(ω) za jednu veliku klasu matrica koje se često javljaju u primjenama – pozitivno definitne matrice.

Teorem 3.5.10. *(Ostrowski, Reich) Neka je A hermitska matrica sa pozitivnim dijagonalnim elementima. Tada je $\mathrm{spr}(\mathsf{S}_\omega) < 1$ i metoda SOR(ω) je konvergentna ako i samo ako je A pozitivno definitna i $\omega \in (0, 2)$. Specijalno je za $\omega = 1$ Gauss–Seidelova metoda konvergentna za pozitivno definitnu A.*

Dokaz: Kao i kod JOR metode, dovoljnost uvjeta slijedi iz Teorema 3.5.6. Prvo izračunamo da je

$$\mathsf{Q} = \mathsf{A}(\mathsf{I} - \mathsf{S}_\omega)^{-1} = \mathsf{D}(\mathsf{I} - \mathsf{L} - \mathsf{U})((\mathsf{I} - \omega\mathsf{L})^{-1}\omega(\mathsf{I} - \mathsf{L} - \mathsf{U}))^{-1} = \omega^{-1}\mathsf{D} - \mathsf{DL},$$

a zatim $\mathsf{H} = \mathsf{Q} + \mathsf{Q}^* - \mathsf{A} = (2\omega^{-1} - 1)\mathsf{D}$. Očito je H pozitivno definitna za $\omega \in (0, 2)$, te je tada $\mathrm{spr}(\mathsf{S}_\omega) < 1$, tj. metode SOR($\omega$) je konvergentna.

Dokaz nužnosti navedenih uvjeta je tehnički i za sada nije uključen u ovaj materijal. ⊞

Kako smo višestruko naglašavali, u primjenama numeričke matematike nas osim konvergencije zanima i efikasnost metode, tj. kako brzo, ili kako još brže doći blizu željenog cilja (limesa niza aproksimacija). U kontekstu SOR(ω) metode se onda prirodno postavlja pitanje optimalnog izbora parametra ω. Odgovor opet treba tražiti u specijalnim slučajevima matrica sa strukturom.

3.5.4 Svojstvo $\mathbb{A}$ i konzistentni uređaj

Definicija 3.5.1. Kažemo da matrica A ima svojstvo $\mathbb{A}$ ako postoji permutacija P tako da je $\mathsf{P}^{\mathrm{T}}\mathsf{AP}$ oblika

$$\mathsf{P}^{\mathrm{T}}\mathsf{AP} = \begin{pmatrix} \mathsf{A}_{[11]} & \mathsf{A}_{[12]} \\ \mathsf{A}_{[21]} & \mathsf{A}_{[22]} \end{pmatrix}, \text{ gdje su } \mathsf{A}_{[11]}, \mathsf{A}_{[22]} \text{ dijagonalne matrice.} \tag{3.5.26}$$

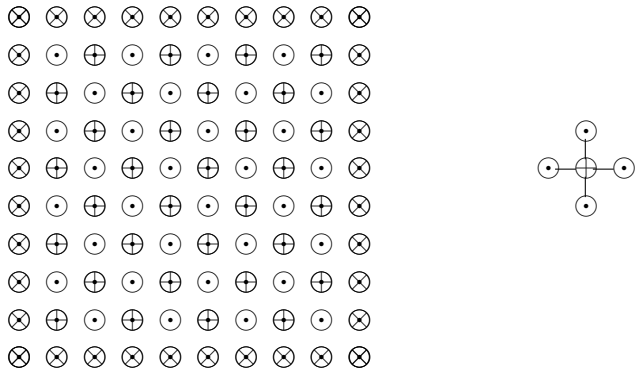
Primjer 3.5.1. Matrica $\mathsf{T}_{\otimes n}$ u (3.5.27) ima svojstvo $\mathbb{A}$. Proučimo to u primjeru $n = 4$.

$$T_{\otimes 4} = \left(\begin{array}{cccc|cccc|cccc|cccc} 4 & -1 & & & -1 & & & & & & & & & & & \\ -1 & 4 & -1 & & & -1 & & & & & & & & & & \\ & -1 & 4 & -1 & & & -1 & & & & & & & & & \\ & & -1 & 4 & & & & -1 & & & & & & & & \\ \hline -1 & & & & 4 & -1 & & & -1 & & & & & & & \\ & -1 & & & -1 & 4 & -1 & & & -1 & & & & & & \\ & & -1 & & & -1 & 4 & -1 & & & -1 & & & & & \\ & & & -1 & & & -1 & 4 & & & & -1 & & & & \\ \hline & & & & -1 & & & & 4 & -1 & & & -1 & & & \\ & & & & & -1 & & & -1 & 4 & -1 & & & -1 & & \\ & & & & & & -1 & & & -1 & 4 & -1 & & & -1 & \\ & & & & & & & -1 & & & -1 & 4 & & & & -1 \\ \hline & & & & & & & & -1 & & & & 4 & -1 & & \\ & & & & & & & & & -1 & & & -1 & 4 & -1 & \\ & & & & & & & & & & -1 & & & -1 & 4 & -1 \\ & & & & & & & & & & & -1 & & & -1 & 4 \end{array}\right) \tag{3.5.27}$$

Sjetimo se kako smo dobili $\mathsf{T}_{\otimes n}$ – matrica je bila rezultat naše odluke da kvadratnu shemu nepoznanica $(v_{ij})_{i,j=1}^{n}$ preslikamo u jedan vektor stupac $\mathsf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_{n^2})^{\mathrm{T}}$ na točno određen način, te da istu strategiju primijenimo na odgovarajuće vrijednosti $(h^2 f_{ij})_{i,j=1}^{n}$ desne strane jednadžbe. Ali to je tek jedan od $n^2!$ mogućih načina. Naravno, elementarna mudrost veli da je za očekivati da su neki odabiri za neke stvari bolji od nekih drugih.

Ako smo koristili bilo koju drugu strategiju, rezultat je oblika $\mathsf{y} = \mathsf{P}^{\mathrm{T}}\mathsf{x}$, gdje je P neka permutacijska matrica. Dakle, bilo koji drugi poredak nepoznanica bi umjesto linearnog sustava $\mathsf{T}_{\otimes n}\mathsf{x} = \mathsf{b}$ dao novi sustav $(\mathsf{P}^{\mathrm{T}}\mathsf{T}_{\otimes n}\mathsf{P})\mathsf{y} = (\mathsf{P}^{\mathrm{T}}\mathsf{b})$. Ako bi $\mathsf{P}^{\mathrm{T}}\mathsf{T}_{\otimes n}\mathsf{P}$ imala blok strukturu (3.5.26) sa dijagonalnim blokovima dimenzija n_1 i n_2, to bi značilo da u množenju $(\mathsf{P}^{\mathrm{T}}\mathsf{T}_{\otimes n}\mathsf{P})\mathsf{y}$, koje reprezentira lijeve strane relacija (3.2.12), svaka od varijabli $\mathsf{y}_1, \ldots, \mathsf{y}_{n_1}$ ulazi u linearne kombinacije samo sa $\mathsf{y}_{n_1+1}, \ldots, \mathsf{y}_n$. Slično bi $\mathsf{y}_{n_1+1}, \ldots, \mathsf{y}_n$ bile kombinirane samo sa $\mathsf{y}_1, \ldots, \mathsf{y}_{n_1}$. To znači

da do odgovarajuće permutacije možemo doći promatrajući jednadžbe (3.2.12) i shemu na Slici 3.11. Uočavamo karakteristične veze među varijablama v_{ij} i to nam daje ideju da čvorove diskretne mreže obojimo u crvene ($\odot$) i crne ($\oplus$) na sljedeči način: kao i ranije, krenemo od pozicije $(1,1)$ u gornjem lijevom kutu i čvorove obilazimo odozgo prema dolje, stupac po stupac, i redom ih bojimo crveno–crno–crveno–..., vidi Sliku 3.8. Vidimo da je u jednadžbama (3.2.12) svaki crveni (crni) čvor povezan samo sa crnim (crvenim) i rubnim čvorovima. Dakle,



Slika 3.8: Crveno–crni uređaj (red–black ordering).

ako je y konstruiran tako da su mu prvih $n_1 = \lceil n^2/2 \rceil$ komponenti svi crveni čvorovi, nakon kojih slijede svi crni, jasno je da je matrica $\mathsf{P}^{\mathrm{T}}\mathsf{T}_{\otimes n}\mathsf{P}$ strukture kao u (3.5.26). Ta se struktura neće pokvariti ako crvene i crne čvorove odvojeno i neovisno permutiramo, dakle imamo na raspolaganju $\lceil n^2/2 \rceil ! \lfloor n^2/2 \rfloor !$ mogućih matrica oblika (3.5.26).[8]

Zašto je svojstvo $\mathbb{A}$ važno? Zato što to svojstvo imaju matrice iz važnih primjena i zato što implicira jedno drugo važno svojstvo:

Teorem 3.5.11. *Neka matrica* A *ima svojstvo* $\mathbb{A}$ *i neka su joj svi dijagonalni elementi različiti od nule. Tada postoji permutacija* P *tako da u dekompoziciji* $\mathsf{P}^{\mathrm{T}}\mathsf{AP} = \mathsf{D}(\mathsf{I} - \mathsf{L} - \mathsf{U})$ *(*D *dijagonalna,* L *strogo donje trokutasta,* U *strogo gornje trokutasta) svojstvene vrijednosti matrica* $\mathsf{J}_\alpha = \alpha\mathsf{L} + \alpha^{-1}\mathsf{U}$*,* $\alpha \in \mathbb{C} \setminus \{0\}$*, ne ovise o* α*.*

[8]Ovdje se opet valja pozvati na ranije spomenutu elementarnu mudrost, ali to odgađamo za neku drugu priliku.

Dokaz: Za dva proizvoljna parametra $0 \neq \alpha_1 \neq \alpha_2 \neq 0$ su J_{α_1} i J_{α_2} slične matrice. Kako vrijedi (3.5.26), možemo pisati $\mathsf{P}^{\mathrm{T}}\mathsf{AP} = \mathsf{D}(\mathsf{I} - \mathsf{L} - \mathsf{U})$ sa

$$\mathsf{D} = \begin{pmatrix} \mathsf{A}_{[11]} & \mathbf{0} \\ \mathbf{0} & \mathsf{A}_{[22]} \end{pmatrix}, \ \mathsf{L} = -\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathsf{A}_{[22]}^{-1}\mathsf{A}_{[21]} & \mathbf{0} \end{pmatrix}, \ \mathsf{U} = -\begin{pmatrix} \mathbf{0} & \mathsf{A}_{[11]}^{-1}\mathsf{A}_{[12]} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

i vrijedi

$$\mathsf{J}_{\alpha_2} = -\begin{pmatrix} \mathbf{0} & \alpha_2^{-1}\mathsf{A}_{[11]}^{-1}\mathsf{A}_{[12]} \\ \alpha_2\mathsf{A}_{[22]}^{-1}\mathsf{A}_{[21]} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathsf{I} & \mathbf{0} \\ \mathbf{0} & \frac{\alpha_2}{\alpha_1}\mathsf{I} \end{pmatrix} \mathsf{J}_{\alpha_1} \begin{pmatrix} \mathsf{I} & \mathbf{0} \\ \mathbf{0} & \frac{\alpha_1}{\alpha_2}\mathsf{I} \end{pmatrix}.$$

⊞

Definicija 3.5.2. Neka je matrica A rastavljena s $\mathsf{A} = \mathsf{D}(\mathsf{I} - \mathsf{L} - \mathsf{U})$, gdje je D dijagonalna, L strogo donje trokutasta, U strogo gornje trokutasta. Ako za $\alpha \neq 0$ svojstvene vrijednosti matrice $\mathsf{J}_\alpha = \alpha\mathsf{L} + \alpha^{-1}\mathsf{U}$ ne ovise o α onda kažemo da je A konzistentno uređena matrica.

Korolar 3.5.12. *Ako* A *ima svojstvo* $\mathbb{A}$ *onda postoji permutacija* P *tako da je* $\mathsf{P}^{\mathrm{T}}\mathsf{AP}$ *konzistentno uređena.*

Teorem 3.5.13. *Neka je* A *konzistentno uređena i* $\omega \neq 0$*. Tada vrijedi*

1. *Svojstvene vrijednosti matrice* $\mathsf{J} = \mathsf{L} + \mathsf{U}$ *dolaze u* $\pm$ *parovima:*

$$\xi \in \mathfrak{S}(\mathsf{J}) \Leftrightarrow -\xi \in \mathfrak{S}(\mathsf{J}).$$

2. *Ako je* ξ *svojstvena vrijednost od* J *i* λ *rješenje jednadžbe*

$$(\lambda + \omega - 1)^2 = \lambda\omega^2\xi^2 \tag{3.5.28}$$

onda je λ *svojstvena vrijednost od* S_ω*. Obratno, ako je* $\lambda \neq 0$ *svojstvena vrijednost od* S_ω *i* ξ *zadovoljava (3.5.28), onda je* ξ *svojstvena vrijednost od* J*.*

Dokaz: Prema Teoremu 3.5.11 $\mathsf{J} = \mathsf{J}_1$ i $\mathsf{J}_{-1} = -\mathsf{J}$ imaju iste svojstvene vrijednosti, što dokazuje prvu tvrdnju. Ako $\lambda = 0$ rješava (3.5.28), onda je $\omega = 1$ i $\mathsf{S}_1 = (\mathsf{I} - \mathsf{L})^{-1}\mathsf{U}$ očito ima nulu u spektru. Neka je sada $\lambda \neq 0$ rješenje od (3.5.28). Kako

možemo raditi sa $\pm\xi$, bez smanjenja općenitosti možemo staviti $\lambda+\omega-1 = \sqrt{\lambda}\omega\xi$. Korištenjem relacije koju smo dokazali u Teoremu 3.5.9 imamo

$$\begin{aligned}\chi_{\mathsf{S}_\omega}(\lambda) &= \det((\lambda+\omega-1)\mathsf{I} - \omega\lambda\mathsf{L} - \omega\mathsf{U}) \\ &= \det((\lambda+\omega-1)\mathsf{I} - \omega\sqrt{\lambda}(\sqrt{\lambda}\mathsf{L} + \frac{1}{\sqrt{\lambda}}\mathsf{U})) \\ &= (\sqrt{\lambda}\omega)^n \det(\xi\mathsf{I} - (\sqrt{\lambda}\mathsf{L} + \frac{1}{\sqrt{\lambda}}\mathsf{U})) = 0, \text{ jer je } \pm\xi \in \mathfrak{S}(\mathsf{J}) = \mathfrak{S}(\mathsf{J}_{\sqrt{\lambda}}).\end{aligned}$$

Dakle, λ je svojstvena vrijednost od S_ω. Obratno, ako je $\lambda \neq$ svojstvena vrijednost of S_ω i ako vrijedi (3.5.28) onda je $\lambda+\omega-1 = \pm\sqrt{\lambda}\omega\xi$ pri čemu je dovoljno promotriti ξ za kojeg je $\lambda+\omega-1 = \sqrt{\lambda}\omega\xi$. No, tada su gornje relacije i dalje valjane, s time da jednakost nuli daje pretpostavka $\lambda \in \mathfrak{S}(\mathsf{S}_\omega)$, a zaključak je da je $\xi \in \mathfrak{S}(\mathsf{J}_{\sqrt{\lambda}}) = \mathfrak{S}(\mathsf{J}) = \mathfrak{S}(-\mathsf{J})$. $\boxplus$

Sljedeći korolar otkriva zanimljivu informaciju: Gauss–Seidelova metoda je "duplo brža" od Jacobijeve, ako za mjeru brzine konvergencije uzmemo spektralni radijus.

Korolar 3.5.14. *Ako je* A *konzistentno uređena, onda je* $\mathrm{spr}(\mathsf{G}) = \mathrm{spr}(\mathsf{J})^2$.

Kako SOR(ω) uključuje Gauss–Seidelovu metodu kao specijalni slučaj $\omega = 1$, pitanje je da li možemo odrediti optimalni parametar ω, u smislu da rezultira najmanjim mogućim spektralnim radijusom. Odgovor je u sljedećem teoremu:

Teorem 3.5.15. *Neka je* A *konzistentno uređena matrica i neka su sve svojstvene vrijednosti matrice* J *realne i* $\mathrm{spr}(\mathsf{J}) < 1$. *Tada je optimalni parametar* ω_* *dan s*

$$\omega_* = \frac{2}{1+\sqrt{1-\mathrm{spr}(\mathsf{J})^2}}, \tag{3.5.29}$$

$$\mathrm{spr}(\mathsf{S}_{\omega_*}) = \min_{\omega\in(0,2)} \mathrm{spr}(\mathsf{S}_\omega) = \omega_* - 1 = \frac{\mathrm{spr}(\mathsf{J})^2}{(1+\sqrt{1-\mathrm{spr}(\mathsf{J})^2})^2}. \tag{3.5.30}$$

Dokaz:
$\boxplus$

Primjer 3.5.2. Pogledajmo kako izgleda izbor optimalnog parametra u slučaju matrice $\mathsf{T}_{\otimes n}$. Neka je A dobivena iz $\mathsf{T}_{\otimes n}$ prelaskom npr. na crveno–crni uređaj. Prvo primijetimo da prelazom na konzistentni uređaj dobijemo matricu J Jacobijeve metode (definiranu pomoću A) koja je slična onoj koju bismo izveli iz $\mathsf{T}_{\otimes n}$. Dakle, primjenom Korolara 3.2.5 dobijemo da su svojstvene vrijednosti od J oblika

$$\lambda_{ij}(\mathsf{J}) = \frac{1}{2}(\cos\frac{i\pi}{n+1} + \cos\frac{j\pi}{n+1}),\ 1 \le i,j \le n.$$

Odavde je $\mathrm{spr}(\mathsf{J}) = \cos\frac{\pi}{n+1}$, a Korolar 3.5.14 odmah daje $\mathrm{spr}(\mathsf{G}) = \cos^2\frac{\pi}{n+1}$. Iz Teorema 3.5.15 je

$$\omega_* = \frac{2}{1+\sin\frac{\pi}{n+1}}, \tag{3.5.31}$$

$$\mathrm{spr}(\mathsf{S}_{\omega_*}) = \frac{\cos^2\frac{\pi}{n+1}}{(1+\sin\frac{\pi}{n+1})^2} = \frac{1-\sin\frac{\pi}{n+1}}{1+\sin\frac{\pi}{n+1}}. \tag{3.5.32}$$

3.5.5 Konvergencija SSOR metode

Teorem 3.5.16. *Neka je* A *hermitska matrica sa pozitivnim dijagonalnim elementima. U metodi SSOR*(ω)*, za svaki* $\omega \in \mathbb{R}$*, matrica* $\mathsf{S}_\omega^{\downarrow\uparrow}$ *je slična hermitskoj pozitivno semidefinitnoj matrici pa ima realne nenegativne svojstvene vrijednosti. Ako je* A *pozitivno definitna, i ako je* $\omega \in (0,2)$*, onda je*

$$\mathrm{spr}(\mathsf{S}_\omega^{\downarrow\uparrow}) = \|\mathsf{A}^{1/2}\mathsf{S}_\omega^{\downarrow\uparrow}\mathsf{A}^{-1/2}\|_2 = \|\mathsf{A}^{1/2}\mathsf{S}_\omega\mathsf{A}^{-1/2}\|_2^2 < 1,$$

pa SSOR(ω) *konvergira. Vrijedi i obrat: Ako SSOR*(ω) *konvergira, onda je* $\omega \in (0,2)$ *i* A *je pozitivno definitna.*

Dokaz: Prvo izračunajmo da je

$$\begin{aligned}\mathsf{S}_\omega &= (\mathsf{I}-\omega\mathsf{L})^{-1}(\omega\mathsf{U}+(1-\omega)\mathsf{I}) = \mathsf{I}-\omega(\mathsf{D}-\omega\hat{\mathsf{L}})^{-1}\mathsf{A}\\ \mathsf{S}_\omega^{\downarrow} &= (\mathsf{I}-\omega\mathsf{U})^{-1}(\omega\mathsf{L}+(1-\omega)\mathsf{I}) = \mathsf{I}-\omega(\mathsf{D}-\omega\hat{\mathsf{U}})^{-1}\mathsf{A}\end{aligned}$$

Sjetimo se, $\mathsf{A} = \mathsf{D} - \hat{\mathsf{L}} - \hat{\mathsf{U}}$, gdje je, zbog $\mathsf{A} = \mathsf{A}^*$, $\hat{\mathsf{U}} = \hat{\mathsf{L}}^*$. To znači da u gornjim relacijama vrijedi $(\mathsf{A}^{1/2}\mathsf{S}_\omega\mathsf{A}^{-1/2})^* = \mathsf{A}^{1/2}\mathsf{S}_\omega^{\downarrow}\mathsf{A}^{-1/2}$, pa je

$$\mathsf{A}^{1/2}\mathsf{S}_\omega^{\downarrow\uparrow}\mathsf{A}^{-1/2} = \mathsf{A}^{1/2}\mathsf{S}_\omega^{\downarrow}\mathsf{A}^{-1/2}(\mathsf{A}^{1/2}\mathsf{S}_\omega^{\downarrow}\mathsf{A}^{-1/2})^* \succeq 0.$$

Dakle, $\|\mathsf{A}^{1/2}\mathsf{S}_\omega^{\downarrow\uparrow}\mathsf{A}^{-1/2}\|_2 = \mathrm{spr}(\mathsf{S}_\omega^{\downarrow\uparrow}) = \|\mathsf{A}^{1/2}\mathsf{S}_\omega\mathsf{A}^{-1/2}\|_2^2$. No, prema Teoremu ?? je $\|\mathsf{A}^{1/2}\mathsf{S}_\omega\mathsf{A}^{-1/2}\|_2 = \mathrm{spr}(\mathsf{S}_\omega) < 1$, jer je SOR$(\omega)$ konvergentna metoda za $\omega \in (0,2)$ i $\mathsf{A} \succ 0$.

Obrat nećemo dokazivati. ⊞

3.6 Polinomijalno ubrzanje konvergencije

Neka je s $x_{i+1} = \mathsf{F}x_i + c$ dana metoda za rješavanje sustava $\mathsf{Ax} = \mathsf{b}$. Sjetimo se, u svakom koraku je $x_{i+1} - x = \mathsf{F}(x_i - x)$ i konvergencija za svaki početni x_0 je osigurana pretpostavkom $\mathrm{spr}(\mathsf{F}) < 1$.

Nakon m koraka imamo aproksimacije $x_0, x_1, \ldots, x_m$ i možemo postaviti dosta razumno pitanje: Da li je možda neka linearna kombinacija $y_m = \sum_{i=0}^{m} \eta_i^{(m)} x_i$ svih ovih aproksimacija puno bolja od svake posebno? Kako odrediti koeficijente $\eta_i^{(m)}$? Prvo uočimo da je uvjet $\sum_{i=0}^{m} \eta_i^{(m)} = 1$ prirodan jer bi u idealnoj situaciji $x_0 = x_1 = \ldots = x_m = x$ moralo vrijediti $y_m = x$. Sada računamo pogrešku:

$$\begin{aligned} y_m - x &= \sum_{i=1}^{m} \eta_i^{(m)} x_i - \sum_{i=1}^{m} \eta_i^{(m)} x = \sum_{i=1}^{m} \eta_i^{(m)} (x_i - x) \\ &= \sum_{i=1}^{m} \eta_i^{(m)} \mathsf{F}^i (x_0 - x) = p_m(\mathsf{F})(x_0 - x) \end{aligned}$$

gdje je $p_m(t) = \sum_{i=0}^{m} \eta_i^{(m)} t^i$ polinom stupnja najviše m, sa svojstvom $p_m(1) = 1$.

Na primjer, ako je $\chi_{\mathsf{F}}(t)$ karakteristični polinom od F, onda je, zbog $\mathrm{spr}(\mathsf{F}) < 1$, $\chi_{\mathsf{F}}(1) \neq 0$ (jedinica nije svojstvena vrijednost) i $p_n(t) = \chi_{\mathsf{F}}(t)/\chi_{\mathsf{F}}(1)$ dobro definiran, $p_n(1) = 1$ i $p_n(\mathsf{F}) = \mathbf{0}$, pa je $y_n - x = \mathbf{0}$. Dakle, vidimo da je u principu moguće naći 'zlatnu linearnu kombinaciju' koja daje rješenje. Naravno, ono što je u principu moguće nije uvijek i praktično izvodivo. Jer, nije lako naći svojstveni polinom, ako i izračunamo koeficijente oni su zbog osjetljivosti na perturbacije i konačne aritmetike vjerojatno većinom pogrešni, a čak i kad bismo ih imali točno, pitanje je koliko je sve efikasno ako je $m = n$.

Praktični pristup je direktna konstrukcija polinoma p_m za kojeg se $p_m(\mathsf{F})$ lako računa i da matrica $p_m(\mathsf{F})$ ima mali spektralni radijus. Općenita konstrukcija je tehnički zahtjevna, mi ćemo proučiti jedan specijalni slučaj i suzdrat ćemo se od ulaženja u strogu teoriju. Glavni cilj je ilustrirati ideje.

Pretpostavimo da je spektar od F realan i da imamo $r \in (0, 1)$ za kojeg znamo da $[-r, r]$ sadrži sve svojstvene vrijednosti od F. Konstruirat ćemo polinome $p_m(t)$ sa svojstvima

- p_m je stupnja točno m i $p_m(1) = 1$;
- $\max_{t \in [-r,r]} |p_m(t)|$ je najmanji među svim polinomima stupnja najviše m koji u $t = 1$ imaju vrijednost jedan.

Drugi uvjet odmah poziva Čebiševljeve polinome prve vrste,

$$T_n(t) = \begin{cases} \cos n \arccos t, & |t| \leq 1 \\ \cosh n \cosh^{-1} t, & |t| \geq 1. \end{cases}$$

Sljedeći teorem je samo kratak podsjetnik:

Teorem 3.6.1. *Čebiševljevi polinomi imaju sljedeća svojstva:*

1. *Vrijedi $T_0(t) \equiv 1$, $T_1(t) = t$, a za $n \geq 1$ je*

$$T_{n+1}(t) = 2tT_n(t) - T_{n-1}(t). \tag{3.6.1}$$

Odavde se jasno vidi da su T_n polinomi sa specijalnom strukturom. Na primjer, $T_2(t) = 2t^2 - 1$, $T_3(t) = 4t^3 - 3t$, $T_4(t) = 8t^4 - 8t^2 + 1$. Nadalje, ako je n paran (neparan) onda $T_n(t)$ ima samo parne (neparne) potencije od t, pa je parna (neparna) funkcija, $T_n(-t) = (-1)^n T_n(t)$. Koeficijent uz n–tu potenciju u $T_n(t)$ je 2^{n-1}.

2. *Sve nultočke od T_n su međusobno različite i realne. Dane su formulama*

$$t_j = \cos\frac{(2j+1)\pi}{2n}, \quad j = 0, \ldots, n-1. \tag{3.6.2}$$

3. *Vrijedi $\max\limits_{t\in[-1,1]} |T_n(t)| = 1$. Pri tome je $T_n(s_j) = (-1)^j$ za*

$$s_j = \cos\frac{j\pi}{n}, \quad j = 0, \ldots, n. \tag{3.6.3}$$

4. *Od svih polinoma stupnja n i s vodećim koeficijentom 1, polinom $\hat{T}_n(t) = 2^{1-n}T_n(t)$ ima na segmentu $[-1, 1]$ najmanju maksimalnu apsolutnu vrijednost, jednaku 2^{1-n}.*

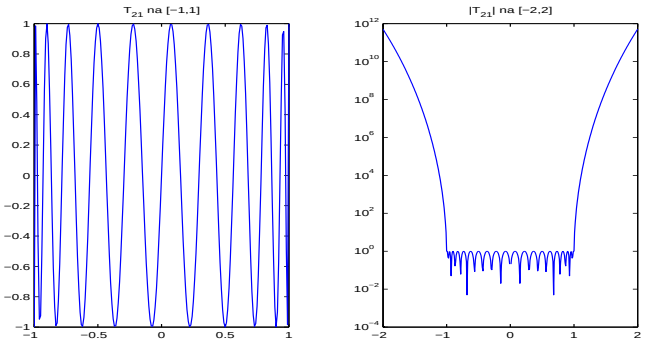
Dakle, ako tražimo polinome koji bi trebali biti mali po modulu na $[-r, r]$, onda su Čebiševljevi polinomi najbolji kandidati. Kako mi imamo dodatni uvjet $p_m(1) = 1$, definiramo

$$p_m(t) = \frac{T_m(t/r)}{T_m(1/r)}.$$

Za $t \in [-r, r]$ je

$$|p_m(t)| \leq \frac{1}{|T_m(1/r)|} = \frac{1}{|T_m(1+q)|}, \quad \frac{1}{r} \equiv 1 + q, q > 0.$$

Slika 3.9 ilustrira zašto je $|p_m(t)|$ mala vrijednost za $t \in [-r, r]$. Vidimo i da manji r implicira veći q pa i veći $|T_m(1 + q)|$, tj. manji $1/|T_m(1 + q)|$. Efekt je jači s većim m.



Slika 3.9: Čebiševljev polinom T_{21} na $[-1, 1]$ i $|T_{21}|$ na [-2,2]

.

Sada imamo

$$\begin{aligned}
y_m - \mathsf{x} &= p_m(\mathsf{F})(x_0 - \mathsf{x}) = \{ \text{ stavimo } \mu_m = \frac{1}{T_m(1/r)} \} \\
&= \mu_m T_m(\mathsf{F}/r)(x_0 - \mathsf{x}) = \{ \text{ rekurzija (3.6.1)} \} \\
&= \mu_m \left[2\frac{\mathsf{F}}{r} T_{m-1}(\mathsf{F}/r)(x_0 - \mathsf{x}) - T_{m-2}(\mathsf{F}/r)(x_0 - \mathsf{x}) \right] \\
&= \mu_m \left[2\frac{\mathsf{F}}{r} \frac{p_{m-1}(\mathsf{F})(x_0 - \mathsf{x})}{\mu_{m-1}} - \frac{p_{m-2}(\mathsf{F})(x_0 - \mathsf{x})}{\mu_{m-2}} \right] \\
&= \mu_m \left[2\frac{\mathsf{F}}{r} \frac{y_{m-1} - \mathsf{x}}{\mu_{m-1}} - \frac{y_{m-2} - \mathsf{x}}{\mu_{m-2}} \right] \Longrightarrow \\
y_m &= \frac{2\mu_m}{\mu_{m-1}} \frac{\mathsf{F}}{r} y_{m-1} - \frac{\mu_m}{\mu_{m-2}} y_{m-2} + d_m, \text{ gdje je} \\
d_m &= \mathsf{x} - \frac{2\mu_m}{\mu_{m-1}} \frac{\mathsf{F}}{r} \mathsf{x} + \frac{\mu_m}{\mu_{m-2}} \mathsf{x} = \mathsf{x} - \frac{2\mu_m}{\mu_{m-1}} \frac{x - c}{r} + \frac{\mu_m}{\mu_{m-2}} \mathsf{x} \\
&= \mu_m \underbrace{\left(\frac{1}{\mu_m} - \frac{2}{r\mu_{m-1}} + \frac{1}{\mu_{m-2}} \right)}_{T_m(1/r) - \frac{2}{r} T_{m-1}(1/r) + T_{m-2}(1/r)} \mathsf{x} + \frac{2\mu_m}{r\mu_{m-1}} c = \frac{2\mu_m}{r\mu_{m-1}} c
\end{aligned}$$

Znači da možemo direktno računati vektore y_m, bez da računamo iteracije x_i. Time smo dobili sljedeći općenit zapis ubrzanja iteracija $x_{i+1} = \mathsf{F}x_i + c$ pomoću Čebiševljevih polinoma:

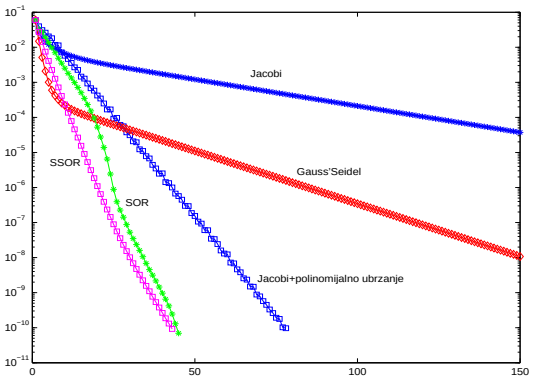
$[y_m, m] = \text{ČEBIŠEV}(\mathsf{F}, c, x_0, r, m_{\max})$

$\mu_0 = 1;\ \mu_1 = r;$
$y_0 = x_0\ ;$
$y_1 = \mathsf{F}x_0 + c;$
for $m = 2, 3, \ldots, m_{\max}$

$$\mu_m = \frac{1}{\frac{2}{r\mu_{m-1}} - \frac{1}{\mu_{m-2}}};$$

$$y_m = \frac{2\mu_m}{r\mu_{m-1}}\mathsf{F}y_{m-1} - \frac{\mu_m}{\mu_{m-2}}y_{m-2} + \frac{2\mu_m}{r\mu_{m-1}}c\ ;$$

if y_m *"dovoljno dobar"* STOP.
end



Slika 3.10: Polinomijalno ubrzanje Jacobijeve metode pomoću Čebiševljevih polinoma i usporedba brzine konvergencije sa SOR, SSOR i Gauss–Seidelovom metodom.

3.7 Krilovljevi potprostori

Često se rješenja jednadžbe $\mathsf{Ax} = \mathsf{b}$ mogu dobro aproksimirati iz pogodno odabranih potprostora čije su dimenzije puno manje od dimenzije prostora. To vrijedi kako za problem dan u beskonačnodimenzionalnom prostoru (npr. parcijalne diferencijalna jednadžba $-\triangle u = f$) čije rješenje aproksimiramo iz konačnodimenzionalnih potprostora (npr. metodom konačnih elemenata), tako i za konačnodimenzionalne probleme velike dimenzije koje opet rješavano iz niždedimenzionalnih potprostora. Takvi potprostori, da bi bili uporabljivi u razvoju numeričkih algoritama, moraju

biti jednostavni za generirati u rastućem nizu,[9] te svojom strukturom moraju osigurati dovoljno kvalitetnu informaciju o objektima koje želimo izračunati. Jedna klasa takvih potprostora su Krilovljevi potprostori, koji su istovremeno i teorijski alat i osnova za mnoge numeričke algoritme.

3.7.1 Motivacija

Počet ćemo s dva primjera u kojima valja uočiti kako se u različitim situacijama prirodno pojavi niz vektora $b, \mathsf{A}b, \mathsf{A}^2b, \ldots$

Primjer 3.7.1. Promotrimo sustav jednadžbi $\mathsf{A}x = b$, s regularnom $n \times n$ matricom A. Iz Hamilton–Cayle–vog teorema je $\mathsf{A}^{-1} = p(\mathsf{A})$ gdje je p polinom stupnja najviše $n-1$, $p(\xi) = \sum_{i=0}^{k} \alpha_i \xi^i$. (Ovdje je $k+1$ najviše jednako stupnju minimalnog polinoma od A.) Dakle, rješenje sustava $\mathsf{A}x = b$ je zapisivo u obliku

$$x = \mathsf{A}^{-1}b = p(\mathsf{A})b = \sum_{i=0}^{k} \alpha_i \mathsf{A}^i b, \tag{3.7.1}$$

tj. kao linearna kombinacija vektora $b, \mathsf{A}b, \mathsf{A}^2b, \ldots, \mathsf{A}^kb$. Naravno, iako jednostavna, formula (3.7.1) nije praktična za stvarno računanje rješenja. Naime, koeficijente svojstvenog odn. minimalnog polinoma nije jednostavno dobiti s zadovoljavajućom točnosti jer su osjetljive funkcije matričnih elemenata. To znači da elegantne algebarske formule za te koeficijente u uvjetima konačne aritmetike nisu osobito korisne.

Primjer 3.7.2. Promotrimo diskretni linearni sistem sa stanjima $x(k) \in \mathbb{R}^n$ u diskretnim vremenima $k \cdot \Delta t$, $k = 0, 1, 2, 3, \ldots$

$$\begin{aligned} x(k+1) &= \mathsf{A}x(k) + \mathsf{b}u(k), \quad x(0) = x_0 && (3.7.2)\\ y(k) &= \mathsf{c}^T x(k), \quad k = 0, 1, 2, 3, \ldots && (3.7.3) \end{aligned}$$

gdje su $\mathsf{A} \in \mathbb{R}^{n\times n}$, $\mathsf{b}, \mathsf{c} \in \mathbb{R}^n$. Vrijednosti $u(k)$ predstavljaju kontrolu (ulaz) u k–tom koraku, a $y(k)$ je izlaz. Postavlja se sljedeće pitanje: Da li je moguće za bilo koje inicijalno stanje x_0 odabrati niz kontrolnih ulaza $u(0), u(1), \ldots, u(n-1)$ tako da $x(n)$ bude jednak bilo kojem zadanom vektoru iz $\mathbb{R}^n$? Drugim riječima, da li sistem možemo u najviše n koraka kontrolirano, pomoću $u(\cdot)$, prevesti iz bilo kojeg

[9] Ovdje se misli u smislu inkluzije, ili s rastućim dimenzijama

u bilo koje drugo zadano stanje? (Ako da, kažemo da je sistem *kontrolabilan.*) Sada uočimo da je

$$\begin{aligned} x(1) &= \mathsf{A}x(0) + \mathsf{b}u(0), \\ x(2) &= \mathsf{A}(\mathsf{A}x(0) + \mathsf{b}u(0)) + \mathsf{b}u(1) = \mathsf{A}^2x(0) + [\mathsf{b}, \mathsf{Ab}] \begin{pmatrix} u(1) \\ u(0) \end{pmatrix}, \dots \\ x(n) &= \mathsf{A}x(n-1) + \mathsf{b}u(n-1) = \mathsf{A}^nx(0) + [\mathsf{b}, \mathsf{Ab}, \dots, \mathsf{A}^{n-1}b] \begin{pmatrix} u(n-1) \\ u(n-2) \\ \vdots \\ u(0) \end{pmatrix}. \end{aligned}$$

Ako stavimo $\mathsf{K} = [\mathsf{b}, \mathsf{Ab}, \dots, \mathsf{A}^{n-1}b]$, onda je traženi niz kontrolnih ulaza $u_{\rightsquigarrow}$ rješenje sustava $\mathsf{K}u_{\rightsquigarrow} = x(n) - \mathsf{A}^nx(0)$. Ako je K regularna, onda se kontrola $u_{\rightsquigarrow}$ može odrediti za bilo koji par $x(0), x(n) \in \mathbb{R}^n$.

3.7.2 Definicija i osnovna svojstva

Definicija 3.7.1. Neka je $\mathsf{A} \in \mathbb{C}^{n\times n}$ i $b \in \mathbb{C}^n \setminus \{0\}$. Krilovljeva matrica reda i je definirana s[10] $\mathsf{K}_i \equiv \mathsf{K}_i(\mathsf{A}, b) = [b, \mathsf{A}b, \dots, \mathsf{A}^{i-1}b]$, a i–ti Krilovljev potprostor $\mathcal{K}_i$ je definiran kao slika od K_i, $\mathcal{K}_i \equiv \mathcal{K}_i(\mathsf{A}, b) = \mathfrak{R}(\mathsf{K}_i)$.

Primijetimo da odmah iz definicije slijedi da je $\mathcal{K}_i(\mathsf{A}, b) = \mathcal{K}_i(\alpha\mathsf{A} - \sigma\mathsf{I}, b)$ za proizvoljne skalare $\alpha \neq 0$ i σ. Nekoliko daljnjih važnih svojstava je dano u sljedećoj propoziciji.

Definicija 3.7.2. Kažemo da je $n \times n$ matrica H u Hessenbergovoj formi (Hessenbergova matrica) ako je $\mathsf{H}_{ij} = 0$ za $i > j + 1$.

Definicija 3.7.3. Za $n \times n$ Hessenbergovu matricu H kažemo da je strogo Hessenbergova ako je $\mathsf{H}_{j+1,j} \neq 0$ za sve $j = 1, \dots, n-1$.

Propozicija 3.7.1. *Za svaki indeks $i = 1, 2, \dots$ je* $\dim(\mathcal{K}_i) \leq i$. *Nadalje,* $\mathsf{A}\mathcal{K}_i \subseteq \mathcal{K}_{i+1}$, *i postoji indeks $\ell \leq n$ za kojeg je*

$$\mathcal{K}_1 \subsetneq \mathcal{K}_2 \subsetneq \cdots \subsetneq \mathcal{K}_i \subsetneq \mathcal{K}_{i+1} \subsetneq \cdots \subsetneq \mathcal{K}_\ell = \mathcal{K}_{\ell+1}, \quad \mathsf{A}\mathcal{K}_\ell \subseteq \mathcal{K}_\ell.$$

Potprostor $\mathcal{K}_\ell$ je najmanji A–invarijantni potprostor koji sadrži vektor b. Osim toga,

[10] Ako se A i b podrazumijevaju iz konteksta, onda ovisnost Krilovljevih potprostora o A i b ne navodimo eksplicitno i jednostavno koristimo oznake K_i, $\mathcal{K}_i$.

- *Ako je* Q_ℓ *ortonormirana baza dobivena QR faktorizacijom matrice* K_ℓ*, onda je u toj bazi* $\mathsf{P}_{\mathcal{K}_\ell}\mathsf{A}_{\downarrow\mathcal{K}_\ell}$ *reprezentiran gornje Hessenbergovom matricom* $\hat{\mathsf{H}}_\ell = \mathsf{Q}_\ell^*\mathsf{A}\mathsf{Q}_\ell$ *s* $(\hat{\mathsf{H}}_\ell)_{i+1,i} \neq 0$*,* $i = 1, \ldots, \ell - 1$*. Ako je* $\mathsf{A} = \mathsf{A}^*$*, onda je* $\hat{\mathsf{H}}_\ell$ *tridijagonalna.*

- *Ako je* A *regularna, onda* $\mathcal{K}_\ell$ *sadrži* $x = \mathsf{A}^{-1}b$*, i vrijedi formula* $\mathsf{A}^{-1}b = \|b\|_2\mathsf{Q}_\ell(\hat{\mathsf{H}}_\ell^{-1}\mathbf{e}_1)$*.*

- *Sa svakim proširenjem matrice* Q_ℓ *do unitarne matrice* $\mathsf{Q} = [\mathsf{Q}_\ell, \mathsf{Q}_\ell^\perp]$ *je* $\mathsf{Q}^*\mathsf{A}\mathsf{Q}$ 2×2 *blok gornje trokutasta matrica s* $\hat{\mathsf{H}}_\ell$ *na poziciji* $(1,1)$*. Specijalno je svaka svojstvena vrijednost od* $\hat{\mathsf{H}}_\ell$ *ujedno i svojstvena vrijednost od* A*.*

Dokaz: Primijetimo da je $\mathsf{A}\mathsf{K}_\ell = \mathsf{K}_\ell\hat{\mathsf{A}}$, pri čemu je $\hat{\mathsf{A}}$ $\ell \times \ell$ matrica. Ako je A regularna, onda je $\dim(\mathcal{K}_\ell) = \operatorname{rang}(\mathsf{K}_\ell) = \ell$ i $\hat{\mathsf{A}}$ je također regularna i očito je

$$\hat{\mathsf{A}} = (\mathsf{K}_\ell^*\mathsf{K}_\ell)^{-1}(\mathsf{K}_\ell^*\mathsf{A}\mathsf{K}_\ell). \tag{3.7.4}$$

Kako je i $\mathsf{A}^{-1}\mathsf{K}_\ell = \mathsf{K}_\ell\hat{\mathsf{A}}^{-1}$, čitanjem prvog stupca u ovoj relaciji vidimo da je $\mathsf{A}^{-1}b$ linearna kombinacija stupaca od K_ℓ, pri čemu su koeficijenti te kombinacije dani u prvom stupcu od $\hat{\mathsf{A}}^{-1}$. Nadalje, lako se uvjerimo da je u bazi K_ℓ operator $\mathsf{P}_{\mathcal{K}_\ell}\mathsf{A}_{\downarrow\mathcal{K}_\ell}$ reprezentiran upravo matricom $\hat{\mathsf{A}}$. (Podsjećamo da je ortogonalni projektor $\mathsf{P}_{\mathcal{K}_\ell}$ reprezentiran matricom $\mathsf{K}_\ell(\mathsf{K}_\ell^*\mathsf{K}_\ell)^{-1}\mathsf{K}_\ell^*$.)

Ako bismo sada željeli izračunati $x = \mathsf{A}^{-1}b$ ili svojstvene vrijednosti od $\hat{\mathsf{A}}$, formula za $\hat{\mathsf{A}}$ sugerira da reprezentacija od $\mathcal{K}_\ell$ pomoću K_ℓ nije najbolji izbor. Pogledajmo zašto: Ako je $\mathsf{K}_\ell = \mathsf{Q}_\ell\mathsf{R}_\ell$ QR faktorizacija, onda je $\mathcal{K}_\ell = \mathfrak{R}(\mathsf{Q}_\ell)$ i $\mathsf{A}\mathsf{Q}_\ell = \mathsf{Q}_\ell(\mathsf{R}_\ell\hat{\mathsf{A}}\mathsf{R}_\ell^{-1})$, gdje je $\hat{\mathsf{H}}_\ell = \mathsf{R}_\ell\hat{\mathsf{A}}\mathsf{R}_\ell^{-1} = \mathsf{Q}_\ell^*\mathsf{A}\mathsf{Q}_\ell$. Ako je $\mathsf{Q}_\ell = [\mathsf{q}_1, \ldots, \mathsf{q}_\ell]$, onda je $\mathsf{q}_1 = b/\|b\|_2$, i za svaki $i < \ell$ je $\mathsf{Q}_i = [\mathsf{q}_1, \ldots, \mathsf{q}_i]$ ortonormirana baza za $\mathcal{K}_i$. Kako je $\mathsf{A}\mathcal{K}_i \subsetneq \mathcal{K}_{i+1}$, odmah zaključujemo da je $\hat{\mathsf{H}}_\ell$ gornje Hessenbergova matrica s $(\hat{\mathsf{H}}_\ell)_{i+1,i} \neq 0$. Ako je $\mathsf{A} = \mathsf{A}^*$, onda je i $\hat{\mathsf{H}}_\ell = \hat{\mathsf{H}}_\ell^*$ očito tridijagonalna. Naravno, vrijedi i $\mathsf{A}^{-1}\mathsf{Q}_\ell = \mathsf{Q}_\ell\hat{\mathsf{H}}_\ell^{-1}$ pa je

$$\mathsf{A}^{-1}b = \|b\|_2\mathsf{Q}_\ell(\hat{\mathsf{H}}_\ell^{-1}\mathbf{e}_1). \tag{3.7.5}$$

Neka je sada $\ell < n$ i $\mathsf{Q} = [\mathsf{Q}_\ell, \mathsf{Q}_\ell^\perp]$ $n \times n$ unitarna, gdje jedno moguće proširenje

$\mathsf{Q}_\ell^\perp$ možemo uzeti iz potpune QR faktorizacije $\mathsf{K}_\ell = \mathsf{Q}\begin{pmatrix}\mathsf{R}_\ell \\ \mathsf{O}\end{pmatrix}$. Tada je

$$\mathsf{H} = \mathsf{Q}^*\mathsf{A}\mathsf{Q} = \left(\begin{array}{c|c} \mathsf{Q}_\ell^*\mathsf{A}\mathsf{Q}_\ell & \mathsf{Q}_\ell^*\mathsf{A}\mathsf{Q}_\ell^\perp \\ \hline \mathsf{O} & (\mathsf{Q}_\ell^\perp)^*\mathsf{A}\mathsf{Q}_\ell^\perp \end{array}\right) = \left(\begin{array}{cccc|ccc} * & * & \star & \star & \star & \cdots & \star \\ * & * & * & \star & \star & \cdots & \star \\ 0 & * & * & * & \star & \cdots & \star \\ 0 & 0 & * & * & \star & \cdots & \star \\ \hline & & & & \bullet & \cdots & \bullet \\ & & \mathsf{O} & & \vdots & & \vdots \\ & & & & \bullet & \cdots & \bullet \end{array}\right)$$

gdje su u slučaju $\mathsf{A} = \mathsf{A}^*$ svi elementi označeni s $\star$ po sili simetrije jednaki nuli. $\boxplus$

Vratimo se našem sustavu $\mathsf{Ax} = \mathsf{b}$. U idealnoj situaciji bismo imali $\mathcal{K}_\ell = \mathcal{K}_{\ell+1}$, i to s dimenzijom ℓ koja je puno manja od n. Vidimo da smo rješenje raspisali u ortonormiranoj bazi, a koeficijenti su dobiveni iz matrice $\hat{\mathsf{H}}_\ell$ koja je s polaznom matricom vezana ortogonalnom transformacijom. Naravno, u konačnoj aritmetici ne možemo očekivati da je taj uvjet ispunjen (čak i da izračunati potprostor zadovoljava uvjet, nije sigurno da li bismo to uspjeli otkriti) sve do zadnjeg koraka. Zašto bi takav jedan potprostor sadrzavao dobru informaciju o $\mathsf{A}^{-1}\mathsf{b}$? Ako je to slučaj, kako to možemo znati te kako onda tu informaciju pretočiti u dobru aproksimaciju rješenja? Odgovori na ova pitanja nisu uvijek jednoznačni i različitim pristupima se dolazi do različitih metoda. Naravno, ti pristupi tj. metode ovise o dodatnim informacijama, npr. da li je A simetrična (možda i pozitivno definitna) ili opća nesimetrična, realna ili kompleksna matrica.

3.8 Arnoldijev algoritam

Iz diskusije u §3.7 nam je jasno da je od interesa Krilovljeve potprostore imati zadane u ortonormiranim bazama. Ako nam trebaju baze za $\mathcal{K}_i$, $i = 1, \ldots, m$, gdje je $m < n$ zadan indeks, onda je svakako jedan način da se formira niz matrica $\mathsf{K}_1, \ldots, \mathsf{K}_m$ i rekurzivno izračunaju pripadne QR faktorizacije. Ako je

$\mathsf{K}_i = [\mathsf{k}_1, \ldots, \mathsf{k}_i] = \mathsf{Q}_i\mathsf{R}_i$ već izračunata, onda imamo

$$\begin{aligned}
\mathsf{K}_{i+1} &= [\mathsf{K}_i, \mathsf{k}_{i+1}] = [\mathsf{Q}_i\mathsf{R}_i, \mathsf{k}_{i+1}] = [\mathsf{Q}_i, \mathsf{k}_{i+1}] \begin{pmatrix} \mathsf{R}_i & 0 \\ 0 & 1 \end{pmatrix} \\
&= [\mathsf{Q}_i, \mathsf{k}_{i+1} - \mathsf{Q}_i\mathsf{Q}_i^*\mathsf{k}_{i+1}] \begin{pmatrix} \mathsf{R}_i & \mathsf{Q}_i^*\mathsf{k}_{i+1} \\ 0 & 1 \end{pmatrix} \\
&= [\mathsf{Q}_i, \mathsf{q}_{i+1}] \begin{pmatrix} \mathsf{R}_i & v_i \\ 0 & \gamma_{i+1} \end{pmatrix}, \quad \text{uz} \quad v_i = \mathsf{Q}_i^*\mathsf{k}_{i+1}, \quad \gamma_{i+1} = \|\mathsf{k}_{i+1} - \mathsf{Q}_i v_i\|_2.
\end{aligned}$$

Dakle, nakon formiranja $\mathsf{k}_{i+1} = \mathsf{A}\mathsf{k}_i$, izračuna se $v_i = \mathsf{Q}_i^*\mathsf{k}_{i+1}$, $\tilde{\mathsf{k}}_{i+1} = \mathsf{k}_{i+1} - \mathsf{Q}_i v_i$, $\gamma_{i+1} = \|\tilde{\mathsf{k}}_{i+1}\|_2$. Ako je $\gamma_{i+1} \neq 0$, onda je $\mathsf{q}_{i+1} = \tilde{\mathsf{k}}_{i+1}/\gamma_{i+1}$ i $\mathsf{Q}_{i+1} = [\mathsf{Q}_i, \mathsf{q}_{i+1}]$ je tražena baza. Vidimo da zajedno s formiranjem matrica K_i primjenom Gram–Schmidtovog algoritma dobivamo pripadne ortonormirane baze. Primijetimo da su te baze u principu jedinstveno određene.

Propozicija 3.8.1. *Ako je* $\mathsf{Q}_m = (\mathsf{q}_1, \mathsf{q}_2, \ldots, \mathsf{q}_m)$ *ortonormalna matrica takva da je* $\mathsf{Q}_i = (\mathsf{q}_1, \ldots, \mathsf{q}_i)$ *baza za* $\mathcal{K}_i$, $i = 1, \ldots, m$, *onda je* Q_m *jedinstvena do na množenje s desna dijagonalnom ortogonalnom (unitarnom) matricom.*

Sada uočimo jedno važno svojstvo niza $\mathsf{b}, \mathsf{A}\mathsf{b}, \ldots$. Neka je A dijagonalizabilna matrica, tj. neka ima n linearno nezavisnih svojstvenih vektora $\mathsf{v}_1, \ldots, \mathsf{v}_n$ i neka su pripadne svojstvene vrijednosti indeksirane tako da je $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. Neka je također $|\lambda_1| > |\lambda_2|$. Koristeći relacije $\mathsf{A}\mathsf{v}_j = \lambda_j\mathsf{v}_j$ i reprezentaciju $\mathsf{b} = \beta_1\mathsf{v}_1 + \beta_2\mathsf{v}_2 + \cdots + \beta_n\mathsf{v}_n$ dobijemo

$$\begin{aligned}
\mathsf{A}^k\mathsf{b} &= \beta_1\lambda_1^k\mathsf{v}_1 + \beta_2\lambda_2^k\mathsf{v}_2 + \cdots + \beta_n\lambda_n^k\mathsf{v}_n \\
&= \lambda_1^k \left(\beta_1\mathsf{v}_1 + \beta_2 \underbrace{\left(\frac{\lambda_2}{\lambda_1}\right)^k}_{\to 0}\mathsf{v}_2 + \cdots + \beta_n \underbrace{\left(\frac{\lambda_n}{\lambda_1}\right)^k}_{\to 0}\mathsf{v}_n \right)
\end{aligned}$$

Dakle, ako je $\beta_1 \neq 0$, vektori $\mathsf{A}^k\mathsf{b}$ će s rastućim k zatvarati sve manje kuteve s linearnom ljuskom svojstvenog vektora v_1. Ovaj zaključak je osnova za metodu potencija za računanje dominantnog svojstvenog para λ_1, v_1.

Dakle, stupci od K_m su dio niza kojeg formira metoda potencija pa je za očekivati da će sa rastućim m kutevi među novoformiranim stupcima postajati sve manji što rezultira povećanjem osjetljivosti QR faktorizacije na neizbježne pogreške računanja u konačnoj aritmetici. Zato računanje matrica K_i i njihovih

QR faktorizacija može rezultirati numerički ortonormalnim bazama u krivim potprostorima. (Zamislite dva skoro paralelna vektora i ravninu koju razapinju; uočite kako se sve može promijeniti ta ravnina ako se vektori malo promijene.)

Umjesto toga, pokušat ćemo generirati isti niz potprostora s pripadnim bazama, ali tako da izbjegnemo mehanizam metode potencija. Drugim riječima, rekurzivnu strukturu $\mathsf{b}, \mathsf{Ab}, \mathsf{A}^2\mathsf{b}, \ldots$ ćemo zamijeniti jednom drugom rekurzijom u kojoj su efekti metode potencija manje izraženi, ali tako da opet dobijemo rekurzivno generiran niz ortonormalnih baza za $\mathcal{K}_1 \subset \mathcal{K}_2 \subset \cdots$

Propozicija 3.8.2. *Neka je* $\mathcal{K}_i \subsetneq \mathcal{K}_{i+1}$ *te neka su* $\mathsf{Q}_i = (\mathbf{q}_1, \ldots, \mathbf{q}_i)$ *i* $\mathsf{Q}_{i+1} = (\mathsf{Q}_i, \mathbf{q}_{i+1})$ *odgovarajuće baze. Tada je* $\mathcal{K}_{i+1}$ *generiran s* $(\mathsf{Q}_i, \mathsf{Aq}_i)$ *Dakle, za svaki i je* $\mathcal{K}_{i+1} = \mathfrak{R}((\mathbf{q}_1, \mathsf{Aq}_1, \mathsf{Aq}_2, \ldots, \mathsf{Aq}_i))$.

Dokaz: Vrijedi $\mathbf{q}_i \in \mathcal{K}_i \setminus \mathcal{K}_{i-1}$, tj. $\mathbf{q}_i = \sum_{j=1}^{i} \gamma_j \mathbf{k}_j$, pri čemu je $\gamma_i \neq 0$. (Primijetimo da $\mathcal{K}_i \subsetneq \mathcal{K}_{i+1}$ povlači i $\mathcal{K}_{i-1} \subsetneq \mathcal{K}_i$.) Očito je $\mathsf{Aq}_i \in \mathcal{K}_{i+1}$. Kada bi bio i $\mathsf{Aq}_i \in \mathcal{K}_i$ onda bi vrijedila relacija $\sum_{j=1}^{i} \gamma_j \mathbf{k}_{j+1} = \sum_{j=1}^{i} \beta_j \mathbf{k}_j$ koja bi implicirala neistinitu tvrdnju $\mathcal{K}_i = \mathcal{K}_{i+1}$. Dakle $\mathcal{K}_{i+1} = \mathfrak{R}((\mathsf{Q}_i, \mathsf{Aq}_i))$ $\boxplus$

To znači da pri prelazu s baze Q_i u $\mathcal{K}_i$ na bazu Q_{i+1} u $\mathcal{K}_{i+1}$ trebamo QR faktorizaciju matrice $(\mathsf{Q}_i, \mathsf{Aq}_i)$, što se svodi na Gram–Schmidtovu ortogonalizaciju vektora Aq_i u odnosu na Q_i.

$$v = (\mathsf{I} - \mathsf{Q}_i \mathsf{Q}_i^*)\mathsf{Aq}_i, \quad \mathbf{q}_{i+1} = \frac{1}{\|v\|_2} v.$$

Ortogonalizaciju $v = \mathsf{Aq}_i - \mathsf{Q}_i(\mathsf{Q}_i^* \mathsf{Aq}_i)$ obično implementiramo u obliku modificiranog Gram–Schmidtovog algoritma, i tada cijeli postupak dobivanja ortonormirane baze u Krilovljevom potprostoru zovemo Arnoldijev algoritam.

Algoritam 3.8.1. Algoritam ARNOLDI(A, b, m) za zadane $\mathsf{A} \in \mathbb{C}^{n \times n}$ i $b \in \mathbb{C}^n \setminus \{0\}$ računa ortonormirane baze $\mathsf{Q}_i = [\mathsf{q}_1, \ldots, \mathsf{q}_i]$ za $\mathcal{K}_i$, te Hessenbergove matrice $\mathsf{H}_{1:i,1:i}$ za koje je

$$\mathsf{AQ}_i = \mathsf{Q}_{i+1}\mathsf{H}_{1:i+1,1:i}, \quad \mathsf{Q}_i^*\mathsf{AQ}_i = \mathsf{H}_{1:i,1:i}, \quad i = 1, \ldots, m. \tag{3.8.1}$$

Ako je za neki $\ell \leq m$, $\mathcal{K}_\ell = \mathcal{K}_{\ell+1}$, onda algoritam završava u koraku $i = \ell$ i vraća vrijednost ℓ. Inače završava u zadanom m–tom koraku i stavlja $\ell = m$.

$[\mathsf{Q}, \mathsf{H}, \ell] = \text{ARNOLDI}(\mathsf{A}, b, m)$

$\mathsf{q}_1 = b/\|b\|_2$
for $i = 1, \ldots, m$
  $v = \mathsf{A}\mathsf{q}_i$
  for $j = 1, \ldots, i$
    $h_{ji} = \mathsf{q}_j^* v$; $v = v - \mathsf{q}_j h_{ji}$
  end
  $h_{i+1,i} = \|v\|_2$
  if $h_{i+1,i} = 0$ *then*
    $\ell = i$; $\mathsf{Q} = [\mathsf{q}_1, \ldots, \mathsf{q}_\ell]$;
    $\mathsf{H} = (h_{ij})_{(\ell+1)\times\ell}$; STOP
  end_if
  $\mathsf{q}_{i+1} = v/h_{i+1,i}$;
end
$\ell = m$;
$\mathsf{Q} = [\mathsf{q}_1, \ldots, \mathsf{q}_m]$; $\mathsf{H} = (h_{ij})_{(m+1)\times m}$.

i–ti korak:

$$\mathsf{H}_{1:i+1,1:i} = \left(\begin{array}{cccc} * & * & * & \star \\ * & * & * & \star \\ 0 & * & * & \star \\ 0 & 0 & * & \star \\ \hline & & & \circledast \end{array}\right)$$

$\circledast = h_{i+1,i}$
$\star = h_{ji},\ j = 1, \ldots, i$

3.8.1 Reortogonalizacija

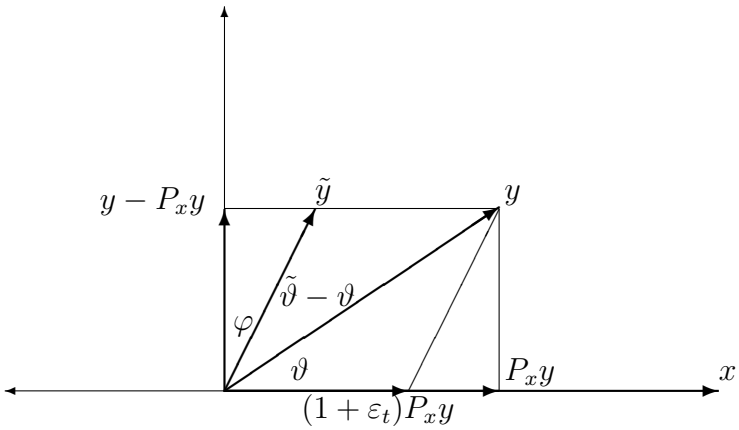
Arnoldijev algoritam je baziran na Gram–Schmidtovoj ortogonalizaciji, ali niti modificirani algoritam (iako bolji od klasičnog) ne garantira numeričku ortogonalnost vektora dobivenih u konačnoj aritmetici. Ovdje pod numeričkom ortogonalnosti podrazumijevamo situaciju u kojoj je, za $i \neq j$, $|\mathsf{q}_i^* \mathsf{q}_j| \leq O(\varepsilon)$, gdje je ε relativna preciznost strojne aritmetike ($\varepsilon \approx 10^{-8}$ u jednostrukoj i $\varepsilon \approx 10^{-16}$ u dvostrukoj preciznosti)

Zapravo se lako mogu konstruirati primjeri u kojima dobivena baza nije numerički ortogonalna. Jedno rješenje je potpuno napustiti ideju korištenja Gram–Schmidtovog algoritma i preći na elementarne ortogonalne transformacije, npr. Householderove reflektore. Drugo, manje radikalno, rješenje je tzv. reortogonalizacija.

Primjer 3.8.1. Neka su $x, y \in \mathbb{R}^m$, $m \geq 2$. Gram–Schmidtov algoritam mijenja vektor y u $y - P_x y$, gdje je $P_x y = (x^\tau y/x^\tau x)x$. (Vektor $y - P_x y$ treba još normirati tako da bude jedinične euklidske norme. Kako ćemo u ovom primjeru analizirati samo kuteve među vektorima, to normiranje ispuštamo zbog jednostavnosti.) Neka je koeficijent $x^\tau y/x^\tau x$ izračunat s malom greškom,

$$t = \frac{x^\tau y}{x^\tau x}(1 + \varepsilon_t), \quad \tilde{y} = y - tx, \tag{3.8.2}$$

i neka je $\boldsymbol{\varepsilon}_t$ jedina greška u cijelom računu. Vidimo da cijelu analizu možemo napraviti u ravnini razapetoj s x i y. Promotrimo sliku 3.11. Očito je



Slika 3.11: Gram–Schmidtova ortogonalizacija sa perturbacijom.

$$\tan\varphi = \frac{|\boldsymbol{\varepsilon}_t| \|P_x y\|_2}{\|y - P_x y\|_2} = \frac{|\varepsilon_t|}{\frac{\|y - P_x y\|_2}{\|P_x y\|_2}} = \frac{|\varepsilon_t|}{\tan\vartheta}, \quad \tan\tilde{\vartheta} = \frac{\tan\vartheta}{|\varepsilon_t|}. \tag{3.8.3}$$

Vidimo da $\tilde{y}$ nije okomit na x i da odstupanje može biti značajno ako je kut između x i y mali. Kako se u računanju na stroju greška ε_t ne može izbjeći, jasno nam je da Gram–Schmidtova ortogonalizacija može biti numerički nestabilna. Izračunati vektori nisu nužno niti približno ortogonalni, ali je pokušaj ortogonalizacije ipak uspio povećati kut među njima. To znači da ponavljanje koraka Gram–Schmidtove ortogonalizacije ali s vektorima x i $\tilde{y}$ daje bolju numeričku ortogonalnost. To ponavljanje Gram–Schmidtovog postupka se zove reortogonalizacija.

3.8.2 Hessenbergova forma

Teorem 3.8.3. *Neka je* $\mathsf{A} \in \mathbb{C}^{n\times n}$. *Postoje* $n \times n$ *unitarna matrica* Q *i Hessenbergova matrica* H, *tako da je* $\mathsf{A} = \mathsf{QHQ}^*$. *Ako je* A *realna matrica, onda* Q *možemo odabrati da bude realna ortogonalna, a* H *realna Hessenbergova. Ako je u dekompoziciji* $\mathsf{A} = \mathsf{QHQ}^*$ *matrica* H *strogo Hessenbergova, onda je ta dekompozicija jedinstveno određena u sljedećem smislu: Ako je* $\mathsf{A} = \tilde{\mathsf{Q}}\tilde{\mathsf{H}}\tilde{\mathsf{Q}}^*$ *također dekompozicija s unitarnom* $\tilde{\mathsf{Q}}$ *i Hessenbergovom* $\tilde{\mathsf{H}}$, *onda* $\tilde{\mathsf{q}}_1 = \mathbf{e}^{\mathrm{i}\phi_1}\mathsf{q}_1$ *povlači* $\tilde{\mathsf{Q}} = \mathsf{Q}\Phi$, *gdje je* $\Phi = \mathrm{diag}(\mathbf{e}^{\mathrm{i}\phi_k})_{k=1}^n$. *U slučaju realnih dekompozicija realne matrice* A *su svi* $\mathbf{e}^{\mathrm{i}\phi_k} \in \{-1, 1\}$.

Dokaz:

Prvo uočimo da $\mathsf{H} = \mathsf{Q}^*\mathsf{AQ}$ i $\tilde{\mathsf{H}} = \tilde{\mathsf{Q}}^*\mathsf{A}\tilde{\mathsf{Q}}$, zajedno sa $\mathsf{q}_1 = e^{\mathrm{i}\phi_1}\tilde{\mathsf{q}}_1$, povlače $\mathsf{h}_{11} = \tilde{\mathsf{h}}_{11}$. Čitanjem prvih stupaca u $\mathsf{AQ} = \mathsf{QH}$ i $\mathsf{A}\tilde{\mathsf{Q}} = \tilde{\mathsf{Q}}\tilde{\mathsf{H}}$ dobijemo

$$\mathsf{Aq}_1 = \mathsf{q}_1\mathsf{h}_{11} + \mathsf{q}_2\mathsf{h}_{21}, \;\; \mathsf{Aq}_1\mathbf{e}^{\mathrm{i}\phi_1} = \mathsf{q}_1\mathsf{h}_{11}\mathbf{e}^{\mathrm{i}\phi_1} + \tilde{\mathsf{q}}_2\tilde{\mathsf{h}}_{21},$$

odakle slijedi $\mathsf{q}_2\mathsf{h}_{21} = \mathbf{e}^{-\mathrm{i}\phi_1}\tilde{\mathsf{q}}_2\tilde{\mathsf{h}}_{21}$. Kako je po pretpostavci $\mathsf{h}_{21} \neq 0$, zaključujemo da je $|\tilde{\mathsf{h}}_{21}| = |\mathsf{h}_{21}|$, i $\tilde{\mathsf{q}}_2 = \mathsf{q}_2\frac{\mathsf{h}_{21}}{\tilde{\mathsf{h}}_{21}}\mathbf{e}^{\mathrm{i}\phi_1} \equiv \mathsf{q}_2\mathbf{e}^{\mathrm{i}\phi_2}$. Odavde je $\tilde{\mathsf{h}}_{22} = \mathsf{h}_{22}$, $\tilde{\mathsf{h}}_{21} = \mathsf{h}_{21}\mathbf{e}^{\mathrm{i}(\phi_1-\phi_2)}$, $\tilde{\mathsf{h}}_{12} = \mathsf{h}_{12}\mathbf{e}^{\mathrm{i}(\phi_2-\phi_1)}$. Nastavljamo induktivno: pretpostavimo da smo za $m < n$ vektora dobili $\tilde{\mathsf{q}}_j = \mathsf{q}_j\mathbf{e}^{\mathrm{i}\phi_j}$, $j = 1, \ldots, m$. Sada iz relacija $\mathsf{q}_{m+1}\mathsf{h}_{m+1,m} = \mathsf{Aq}_m - \sum_{j=1}^{m}\mathsf{q}_j\mathsf{h}_{jm}$ i

$$\tilde{\mathsf{q}}_{m+1}\tilde{\mathsf{h}}_{m+1,m} = \mathsf{A}\tilde{\mathsf{q}}_m - \sum_{j=1}^{m}\tilde{\mathsf{q}}_j\tilde{\mathsf{h}}_{jm} = \mathsf{Aq}_m\mathbf{e}^{\mathrm{i}\phi_m} - \sum_{j=1}^{m}\mathsf{q}_j\mathbf{e}^{\mathrm{i}\phi_j}\mathsf{h}_{jm}\mathbf{e}^{\mathrm{i}(\phi_m-\phi_j)}$$

zaključujemo da je $\tilde{\mathsf{q}}_{m+1}\tilde{\mathsf{h}}_{m+1,m} = \mathsf{q}_{m+1}\mathsf{h}_{m+1,m}\mathbf{e}^{\mathrm{i}\phi_m}$. Ostatak slijedi kao i u slučaju vektora $\tilde{\mathsf{q}}_2$: $\mathsf{h}_{m+1,m} \neq 0$ povlači $\tilde{\mathsf{h}}_{m+1,m} \neq 0$ i zaključujemo da je $\tilde{\mathsf{q}}_{m+1} = \mathsf{q}_{m+1}\frac{\mathsf{h}_{m+1,m}}{\tilde{\mathsf{h}}_{m+1,m}}\mathbf{e}^{\mathrm{i}\phi_m} \equiv \mathsf{q}_{m+1}\mathbf{e}^{\mathrm{i}\phi_{m+1}}$. ⊞

3.9 Lanczosev algoritam

Kako smo već vidjeli u §3.7, u slučaju $\mathsf{A} = \mathsf{A}^*$, u Arnoldijevom algoritmu i specijalno u relaciji (3.8.1) matrice $\mathsf{H}_{1:i,1:i}$ moraju za sve i biti hermitske tridijagonalne. Nadalje, kako su u Arnoldijevom algoritmu elementi ispod glavne dijagonale u $\mathsf{H}_{1:i,1:i}$ uvijek realni, u hermitskom slučaju $\mathsf{H}_{1:i,1:i}$ mora biti realna simetrična tridijagonalna matrica. Zato ćemo, da naglasimo tridijagonalnu strukturu, matrice $\mathsf{H}_{1:i,1:i}$ i $\mathsf{H}_{1:i+1,1:i}$ označavati s $\mathsf{T}_{1:i,1:i}$, $\mathsf{T}_{1:i+1,1:i}$, gdje je

$$\mathsf{T}_{1:i,1:i} = \begin{pmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \beta_2 & \ddots & \ddots & \\ & & \ddots & \alpha_{i-1} & \beta_{i-1} \\ & & & \beta_{i-1} & \alpha_i \end{pmatrix}, \mathsf{T}_{1:i+1,1:i} = \left(\begin{array}{ccccc} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \beta_2 & \ddots & \ddots & \\ & & \ddots & \alpha_{i-1} & \beta_{i-1} \\ & & & \beta_{i-1} & \alpha_i \\ \hline & & & & \beta_i \end{array}\right). \tag{3.9.1}$$

Zbog tridijagonalne strukture matricu $\mathsf{T}_{1:i,1:i}$ u algoritmima uvijek reprezentiramo s dva niza skalara $(\alpha_j)_{j=1}^{i}$, $(\beta_j)_{j=1}^{i-1}$.

Pogledajmo sada kako se Algoritam 3.8.1 pojednostavljuje u hermitskom slučaju. Specijalno je, u prvom koraku, $\alpha_1 \equiv h_{11} = \mathsf{q}_1^*\mathsf{A}\mathsf{q}_1 \in \mathbb{R}$. Dalje je $\beta_1 \equiv h_{21} = \|\mathsf{A}\mathsf{q}_1 - \alpha_1\mathsf{q}_1\|_2$, a zbog simetrije je i $h_{21} = h_{12}$. Za svaki indeks i, petlja $j = 1, \ldots, i$ ima netrivijalne koeficijente h_{ji} samo za $j = i-1$, te $j = i$. Pri tome je $h_{i-1,i} = h_{i,i-1} = \beta_{i-1}$ (poznato iz prethodnog koraka), pa se ortogonalizacija vektora $\mathsf{A}\mathsf{q}_i$ u odnosu na q_{i-1} svodi na računanje vektora $v = \mathsf{A}\mathsf{q}_i - \beta_{i-1}\mathsf{q}_{i-1}$. Još jedna ortogonalizacija, u odnosu na q_i se svodi na $\alpha_i = \mathsf{q}_i^* v$ i $v - \alpha_i\mathsf{q}_i$. Dobiveno pojednostavljenje Arnoldijevog algoritma zovemo Lanczosev algoritam.

Algoritam 3.9.1. Algoritam LANCZOS(A, b, m) za zadane $\mathsf{A} = \mathsf{A}^* \in \mathbb{C}^{n\times n}$ i $b \in \mathbb{C}^n \setminus \{0\}$ računa ortonormirane baze $\mathsf{Q}_i = [\mathsf{q}_1, \ldots, \mathsf{q}_i]$ za $\mathcal{K}_i$, te realne simetrične tridijagonalne matrice $\mathsf{T}_{1:i,1:i}$ za koje je

$$\mathsf{A}\mathsf{Q}_i = \mathsf{Q}_{i+1}\mathsf{T}_{1:i+1,1:i}, \quad \mathsf{Q}_i^*\mathsf{A}\mathsf{Q}_i = \mathsf{T}_{1:i,1:i}, \quad i = 1, \ldots, m. \tag{3.9.2}$$

Ako je za neki $\ell \le m$, $\mathcal{K}_\ell = \mathcal{K}_{\ell+1}$, onda algoritam završava u koraku $i = \ell$ i vraća vrijednost ℓ. Inače završava u zadanom m–tom koraku i stavlja $\ell = -1$.

$[\mathsf{Q}, \mathsf{T}, \ell] =$ LANCZOS(A, b, m)

$\mathsf{q}_1 = b/\|b\|_2$; $\mathsf{q}_0 = 0$; $\beta_0 = 0$
for $i = 1, \ldots, m$
  $v = \mathsf{A}\mathsf{q}_i - \beta_{i-1}\mathsf{q}_{i-1}$
  $\alpha_i = \mathsf{q}_i^* v$; $v = v - \alpha_i\mathsf{q}_i$
  $\beta_i = \|v\|_2$
  if $\beta_i = 0$ *then*
    $\ell = i$; $\mathsf{Q} = [\mathsf{q}_1, \ldots, \mathsf{q}_\ell]$;
    $\alpha = (\alpha_j)_{j=1}^{\ell}$, $\beta = (\beta_j)_{j=1}^{\ell}$; STOP
  end_if
  $\mathsf{q}_{i+1} = v/\beta_i$
end
$\mathsf{Q} = [\mathsf{q}_1, \ldots, \mathsf{q}_m]$;
$\alpha = (\alpha_j)_{j=1}^{m}$, $\beta = (\beta_j)_{j=1}^{m}$; $\ell = -1$.

i–ti korak:

$$\mathsf{T}_{1:i+1,1:i} = \begin{pmatrix} * & * & & \\ * & * & * & \\ & * & * & * \\ & & * & * \\ \hline & & & \circledast \end{pmatrix}$$

$\beta_i = \circledast$

Primijetimo da u i–tom koraku Lanczosevog algoritma računamo samo sa tri vektora q_{i-1}, q_i, q_{i+1}, koji su vezani u tročlanu rekurziju

$$\mathsf{A}\mathsf{q}_i = \beta_{i-1}\mathsf{q}_{i-1} + \alpha_i\mathsf{q}_i + \beta_i\mathsf{q}_{i+1}. \tag{3.9.3}$$

Istovremeno je prikaz kompresije matrice (operatora) A na $\mathcal{K}_i$ u bazi Q_i dan tridijagonalnom matricom $\mathsf{T}_{1:i,1:i}$ tj. sa $2i-1$ skalara $(\alpha_j)_{j=1}^{i}$, $(\beta_j)_{j=1}^{i-1}$.

3.10 Metoda GMRES

Metoda generaliziranih minimalnih reziduala (GMRES) formira niz Krilovljevih potprostora $\mathcal{K}_i$ tako da Arnoldijevim algoritmom računa pripadne ortonormirane baze Q_i, i u i–tom koraku se aproksimacija x_i traži na linearnoj mnogostrukosti $\mathsf{x}_0 + \mathcal{K}_i$. Koristeći ortonormiranu bazu Q_i, x_i tražimo u obliku $\mathsf{x}_i = \mathsf{x}_0 + \mathsf{Q}_i\mathsf{y}_i$, pri čemu je $\mathsf{y}_i \in \mathbb{C}^i$ odabran tako da je euklidska norma reziduala $\mathsf{r}_i = \mathsf{b} - \mathsf{A}\mathsf{x}_i$ minimalna.

Pitanje je, koji Krilovljevi potprostori su dobar izbor za određivanje korekcije koja bi inicijalni x_0 pomaknula bliže rješenju $\mathsf{x} = \mathsf{A}^{-1}\mathsf{b}$? Sjetimo se da su pogreška aproksimacije $\mathsf{e}_0 = \mathsf{x} - \mathsf{x}_0$ i rezidual r_0 vezani korekcijskom jednadžbom $\mathsf{A}\mathsf{e}_0 = \mathsf{r}_0$ pa je $\mathsf{x}_0 + \mathsf{A}^{-1}\mathsf{r}_0 = \mathsf{x}$. Slijedi da je korekciju $\mathsf{A}^{-1}\mathsf{r}_0$ dobro tražiti u nizu potprostora $\mathcal{K}_i$ generiranih s $\mathsf{r}_0, \mathsf{A}\mathsf{r}_0, \mathsf{A}^2\mathsf{r}_0, \ldots$

Propozicija 3.10.1. *Optimalni izbor vektora* y_i *koji daje korekciju* $\mathsf{Q}_i\mathsf{y}_i \in \mathcal{K}_i$ *s kojom* $\mathsf{x}_i = \mathsf{x}_0 + \mathsf{Q}_i\mathsf{y}_i$ *ima minimalni rezidual je rješenje problema najmanjih kvadrata*

$$\min_{\mathsf{y}\in\mathbb{C}^i} \|\|\mathsf{r}_0\|_2\mathbf{j}_1 - \mathsf{H}_{1:i+1,1:i}\,\mathsf{y}\|_2, \tag{3.10.1}$$

gdje je $\mathbf{j}_1 = \big(1, 0, \cdots, 0\big)^*$.

Dokaz: Prvo uočimo da je za $\mathsf{x}_i = \mathsf{x}_0 + \mathsf{Q}_i\mathsf{y}$ pripadni rezidual jednak $\mathsf{r}_i = \mathsf{b} - \mathsf{A}\mathsf{x}_i = \mathsf{r}_0 - \mathsf{A}\mathsf{Q}_i\mathsf{y}$. Nadalje, iz Arnoldijevog algoritma je $\mathsf{A}\mathsf{Q}_i = \mathsf{Q}_{i+1}\mathsf{H}_{1:i+1,1:i}$, i pri tome je prvi stupac u Q_i jednak $\mathsf{q}_1 = \mathsf{r}_0/\|\mathsf{r}_0\|_2$, tj. $\mathsf{r}_0 = \|\mathsf{r}_0\|_2\mathsf{Q}_{i+1}\mathbf{j}_1$. Dakle, vrijedi

$$\|\mathsf{r}_i\|_2 = \|\mathsf{r}_0 - \mathsf{A}\mathsf{Q}_i\mathsf{y}\|_2 = \|\mathsf{Q}_{i+1}(\|\mathsf{r}_0\|_2\mathbf{j}_1 - \mathsf{H}_{1:i+1,1:i}\mathsf{y})\|_2 = \|\|\mathsf{r}_0\|_2\mathbf{j}_1 - \mathsf{H}_{1:i+1,1:i}\mathsf{y}\|_2.$$

⊞

U problemu najmanjih kvadrata (3.10.1) je matrica $\mathsf{H}_{1:i+1,1:i}$ gornje Hessenbergova i računa se rekurzivno, stupac po stupac, a sam problem ima sljedeću strukturu:

$$\|\mathsf{r}_0\|_2\mathbf{j}_1 - \mathsf{H}_{1:i+1,1:i}\,\mathsf{y} = \begin{pmatrix}\star\\0\\0\\0\\0\end{pmatrix} - \left(\begin{array}{ccc|c} * & * & * & * \\ \circledast & * & * & * \\ 0 & \circledast & * & * \\ 0 & 0 & \circledast & * \\ \hline 0 & 0 & 0 & \circledast \end{array}\right)\begin{pmatrix}\bullet\\\bullet\\\bullet\\\bullet\end{pmatrix}$$

Opišimo sada rješenje problema (3.10.1). Neka je $\mathsf{H}_{1:i+1,1:i} = \mathsf{U}_{i+1}\mathsf{T}_{1:i+1,1:i}$ QR faktorizacija s $\mathsf{T}_{1:i+1,1:i} = \left(\begin{smallmatrix}\mathsf{T}_i\\0\end{smallmatrix}\right)$ i gornje trokutastom $i \times i$ matricom T_i i unitarnom $(i+1)\times(i+1)$ matricom U_{i+1}. Ovu QR faktorizaciju računamo tako da

U_{i+1}^* konstruiramo kao produkt Givensovih rotacija koje množe $\mathsf{H}_{1:i+1,1:i}$ s lijeva i poništavaju elemente ispod dijagonale ($\circledast$). Na primjer transformacija $\mathsf{G}_{12}\mathsf{H}_{1:i+1,1:i}$ je oblika

$$\begin{pmatrix} c_1 & s_1 & 0 & 0 & 0 \\ -\overline{s_1} & c_1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \left(\begin{array}{ccc|c} * & * & * & * \\ \circledast & * & * & * \\ 0 & \circledast & * & * \\ 0 & 0 & \circledast & * \\ \hline 0 & 0 & 0 & \circledast \end{array}\right) = \left(\begin{array}{ccc|c} \boxtimes & * & * & * \\ \mathbf{0} & * & * & * \\ 0 & \circledast & * & * \\ 0 & 0 & \circledast & * \\ \hline 0 & 0 & 0 & \circledast \end{array}\right),$$

gdje su parametri rotacije G_{12} određeni iz uvjeta

$$\begin{pmatrix} c_1 & s_1 \\ -\overline{s_1} & c_1 \end{pmatrix} \begin{pmatrix} \mathsf{h}_{11} \\ \mathsf{h}_{21} \end{pmatrix} = \begin{pmatrix} \sqrt{|\mathsf{h}_{11}|^2 + |\mathsf{h}_{21}|^2} \\ \mathbf{0} \end{pmatrix}; \quad \begin{aligned} s_1 &= \frac{\mathsf{h}_{11}}{|\mathsf{h}_{11}|} \frac{\overline{\mathsf{h}_{21}}}{\sqrt{|\mathsf{h}_{11}|^2 + |\mathsf{h}_{21}|^2}} \\ c_1 &= \frac{|\mathsf{h}_{11}|}{\sqrt{|\mathsf{h}_{11}|^2 + |\mathsf{h}_{21}|^2}} \end{aligned} . \quad (3.10.2)$$

Naravno, ako je u gornjim formulama $\mathsf{h}_{11} = 0$, trivijalno uzimamo $c_1 = 0$, $s_1 = 1$. Ako radimo nad $\mathbb{R}$, onda se formule pojednostave na očit način. Primijetimo da je nova pozicija $(1,1)$ (označena s $\boxtimes$) sigurno različita od nule ako je $\mathsf{h}_{21} \neq 0$. Sada $\mathsf{G}_{23}\mathsf{G}_{12}\mathsf{H}_{1:i+1,1:i}$ izgleda

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & c_2 & s_2 & 0 & 0 \\ 0 & -\overline{s_2} & c_2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \left(\begin{array}{ccc|c} \boxtimes & * & * & * \\ 0 & * & * & * \\ 0 & \circledast & * & * \\ 0 & 0 & \circledast & * \\ \hline 0 & 0 & 0 & \circledast \end{array}\right) = \left(\begin{array}{ccc|c} \boxtimes & * & * & * \\ \mathbf{0} & \boxplus & * & * \\ 0 & \mathbf{0} & * & * \\ 0 & 0 & \circledast & * \\ \hline 0 & 0 & 0 & \circledast \end{array}\right)$$

gdje su c_2, s_2 odabrani da poniše poziciju $(3,2)$, analogno s (3.10.2). Pri tome $\mathsf{h}_{32} \neq 0$ garantira da je $(2,2)$ pozicija $\boxplus$ u produktu $\mathsf{G}_{23}\mathsf{G}_{12}\mathsf{H}_{1:i+1,1:i}$ različita od nule. U sljedećem koraku transformiramo treći i četvrti redak, $\mathsf{G}_{34}\mathsf{G}_{23}\mathsf{G}_{12}\mathsf{H}_{1:i+1,1:i}$ se dobije kao

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & c_3 & s_3 & 0 \\ 0 & 0 & -\overline{s_3} & c_3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \left(\begin{array}{ccc|c} \boxtimes & * & * & * \\ \mathbf{0} & \boxplus & * & * \\ 0 & \mathbf{0} & * & * \\ 0 & 0 & \circledast & * \\ \hline 0 & 0 & 0 & \circledast \end{array}\right) = \left(\begin{array}{ccc|c} \boxtimes & * & * & * \\ \mathbf{0} & \boxplus & * & * \\ 0 & \mathbf{0} & \boxdot & * \\ 0 & 0 & \mathbf{0} & * \\ \hline 0 & 0 & 0 & \circledast \end{array}\right)$$

i vrijede analogni zaključci kao u prethodnim koracima. Ovdje još uočimo da je u ovom koraku završila QR faktorizacija vodeće $i \times (i-1)$ podmatrice $\mathsf{H}_{1:i,1:i-1}$

od $\mathsf{H}_{1:i+1,1:i}$. To znači da, ako smo u prethodnom $(i-1)$–vom koraku već bili izračunali QR faktorizaciju matrice $\mathsf{H}_{1:i,1:i-1}$, u ovom i–tom taj dio posla ne treba opet ponavljati – jedino što u $(i-1)$–vom koraku nismo mogli napraviti je transformirati i–ti stupac od $\mathsf{H}_{1:i+1,1:i}$ i to jer tada nije bio niti izračunat. Znači, kada u i–tom koraku Arnoldijevim algoritmom odredimo novi stupac, onda prethodne rotacije treba primijeniti u istom poretku na taj novi stupac. Nakon toga ostaje još pozicija $(j+1, j)$ koju poništimo jednom dodatnom Givensovom rotacijom – u našem malom primjeru $(i = 4)$ to znači $\mathsf{T}_{1:i+1,1:i} = \mathsf{G}_{45}\mathsf{G}_{34}\mathsf{G}_{23}\mathsf{G}_{12}\mathsf{H}_{1:i+1,1:i}$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & c_4 & s_4 \\ 0 & 0 & 0 & -\overline{s_4} & c_4 \end{pmatrix} \left(\begin{array}{ccc|c} \boxtimes & * & * & * \\ \mathbf{0} & \boxplus & * & * \\ 0 & \mathbf{0} & \boxdot & * \\ 0 & 0 & \mathbf{0} & * \\ \hline 0 & 0 & 0 & \circledast \end{array}\right) = \left(\begin{array}{ccc|c} \boxtimes & * & * & * \\ \mathbf{0} & \boxplus & * & * \\ 0 & \mathbf{0} & \boxdot & * \\ 0 & 0 & \mathbf{0} & \boxminus \\ \hline 0 & 0 & 0 & \mathbf{0} \end{array}\right) \equiv \begin{pmatrix} \mathsf{T}_i \\ 0 \end{pmatrix}.$$

Time smo dobili traženu QR faktorizaciju:

$$\underbrace{\mathsf{G}_{i,i+1}\mathsf{G}_{i-1,1}\cdots\mathsf{G}_{23}\mathsf{G}_{12}}_{\mathsf{U}_{i+1}^*}\mathsf{H}_{1:i+1,1:i} = \begin{pmatrix} \mathsf{T}_i \\ 0 \end{pmatrix}$$

Također smo zaključili da će svi dijagonalni elementi u T_i biti različiti od nule ako su u matrici $\mathsf{H}_{1:i+1,1:i}$ svi elementi ispod glavne dijagonale (pozicije $(k+1, k)$) različiti od nule. Pretpostavimo u momentu da je to ispunjeno, dakle $\det(\mathsf{T}_i) \neq 0$. Stavimo $\rho_0 = \|\mathsf{r}_0\|_2$. Slijedi

$$\begin{aligned} \min_{\mathsf{y}} \|\rho_0\mathbf{j}_1 - \mathsf{H}_{1:i+1,1:i}\,\mathsf{y}\|_2 &= \min_{\mathsf{y}} \|\mathsf{U}_{i+1}\underbrace{\mathsf{U}_{i+1}^*\rho_0\mathbf{j}_1}_{\mathsf{f}^{(i+1)}} - \mathsf{U}_{i+1}\mathsf{T}_{1:i+1,1:i}\,\mathsf{y}\|_2 \\ &= \min_{\mathsf{y}} \|\mathsf{f}^{(i+1)} - \begin{pmatrix} \mathsf{T}_i\mathsf{y} \\ 0 \end{pmatrix}\|_2 \\ &= \min_{\mathsf{y}} \sqrt{\|\mathsf{f}_{1:i}^{(i+1)} - \mathsf{T}_i\mathsf{y}\|_2^2 + |\mathsf{f}_{i+1}^{(i+1)}|^2} \end{aligned}$$

pa je optimalni y očito dan s $\mathsf{y}_i = \mathsf{T}_i^{-1}\mathsf{f}_{1:i}^{(i+1)}$ i minimalna vrijednost reziduala je $|\mathsf{f}_{i+1}^{(i+1)}|$. Ovdje uočimo kako se računa $\mathsf{f}^{(i+1)}$:

$$\mathsf{f}^{(\mathrm{i+1})} = \mathsf{G}_{i,i+1}\mathsf{G}_{i-1,1}\cdots\mathsf{G}_{23}\mathsf{G}_{12}\begin{pmatrix} \rho_0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathsf{G}_{i,i+1}\mathsf{G}_{i-1,1}\cdots\mathsf{G}_{23}\begin{pmatrix} c_1\rho_0 \\ -\overline{s_1}\rho_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \cdots$$

Dakle, ako stavimo $\mathsf{x}_i = \mathsf{x}_0 + \mathsf{Q}_i\mathsf{y}_i$, onda je $\|\mathsf{b} - \mathsf{A}\mathsf{x}_i\|_2 = |\mathsf{f}_{i+1}^{(i+1)}|$ i to je najmanji rezidual kojeg možemo postići koristeći samo informaciju iz $\mathcal{K}_i$. Ako to nije dovoljno dobro, prelazimo na $\mathcal{K}_{i+1}$ – Arnoldijevim algorithmom se izračuna novi stupac i time odredimo $\mathsf{H}_{1:i+2,1:i+1}$ kojoj je $\mathsf{H}_{1:i+1,1:i}$ vodeća $(i+1) \times i$ podmatrica. Rotacije koje smo primijenili pri računanju QR faktorizacije matrice $\mathsf{H}_{1:i+1,1:i}$ primijenimo na novi stupac i jednom dodatnom rotacijom poništimo $\mathsf{h}_{i+2,i+1}$ itd.

Ostaje još vidjeti što ako je, za neki indeks i, $\mathsf{h}_{i+1,i} = 0$. Prvo, u Arnoldijevom algoritmu to znači da se ne može formirati novi smjer q_{i+1} i da algoritam staje. Tada je $\mathsf{AQ}_i = \mathsf{Q}_i\mathsf{H}_{1:i,1:i}$ i matrica $\mathsf{H}_{1:i,1:i} = \mathsf{Q}_i^*\mathsf{AQ}_i$ je regularna ako je A regularna. Za $\mathsf{x}_i = \mathsf{x}_0 + \mathsf{Q}_i\mathsf{y}_i$ je pripadni rezidual jednak

$$\mathsf{r}_i = \mathsf{b} - \mathsf{A}\mathsf{x}_i = \mathsf{r}_0 - \mathsf{AQ}_i\mathsf{y}_i = \mathsf{r}_0 - \mathsf{Q}_i\mathsf{H}_{1:i,1:i}\mathsf{y} = \mathsf{Q}_i(\rho_0\mathbf{j}_1 - \mathsf{H}_{1:i,1:i}\mathsf{y}_i)$$

i jasno je da s $y_i = \rho\mathsf{H}_{1:i,1:i}^{-1}\mathbf{j}_1$ postižemo $\mathsf{r}_i = \mathbf{0}$. To se u prethodnom računu automatski realizira jer $\mathsf{h}_{i+1,i} = 0$ implicira $\mathsf{G}_{i,i+1} = \mathsf{I}_{i+1}$ i $\mathsf{f}_{i+1}^{(i+1)} = 0$. Dakle, kada Arnoldijev algoritam mora prekinuti izvršavanje, za GMRES to znači da je otkrio egzaktnu korekciju $\mathsf{A}^{-1}\mathsf{r}_0$, tj. rješenje $\mathsf{x} = \mathsf{A}^{-1}\mathsf{b}$.

$[\mathsf{Q}, \mathsf{H}, \ell] = \text{ARNOLDI}(\mathsf{A}, b, m, \mathsf{x}_0)$

$\mathsf{x}_0 =$ inicijalna aproksimacija ; $\mathsf{r}_0 = \mathsf{b} - \mathsf{A}\mathsf{x}_0$;
for $i = 1, \ldots, m$
  $v = \mathsf{A}\mathsf{q}_i$
  for $j = 1, \ldots, i$
    $h_{ji} = \mathsf{q}_j^* v$; $v = v - \mathsf{q}_j h_{ji}$
  end
  $h_{i+1,i} = \|v\|_2$
  for $j = 1, \ldots, i-1$
    $\begin{pmatrix} \mathsf{h}_{ji} \\ \mathsf{h}_{j+1,i} \end{pmatrix} = \begin{pmatrix} c_j & s_j \\ -\overline{s_j} & c_j \end{pmatrix} \begin{pmatrix} \mathsf{h}_{ji} \\ \mathsf{h}_{j+1,i} \end{pmatrix}$
  end
  if $\mathsf{h}_{ii} \neq 0$ *then*
    $c_i = \dfrac{|\mathsf{h}_{ii}|}{\sqrt{|\mathsf{h}_{ii}|^2 + |\mathsf{h}_{i+1,1}|^2}}$; $s_i = \dfrac{\mathsf{h}_{ii}}{|\mathsf{h}_{ii}|} \dfrac{\overline{\mathsf{h}_{i+1,i}}}{\sqrt{|\mathsf{h}_{ii}|^2 + |\mathsf{h}_{i+1,1}|^2}}$;
  else
    $c_i = 0$; $s_i = 1$;
  end_if
  $\mathsf{h}_{ii} = c_i \mathsf{h}_{ii} + s_i \mathsf{h}_{i+1,i}$; $\mathsf{h}_{i+1,i} = 0$;
  $\begin{pmatrix} \mathsf{f}_i \\ \mathsf{f}_{i+1} \end{pmatrix} = \begin{pmatrix} c_i & s_i \\ -\overline{s_i} & c_i \end{pmatrix} \begin{pmatrix} \mathsf{f}_i \\ 0 \end{pmatrix}$;
  if $|\mathsf{f}_{i+1}| \leq \varepsilon$ *then*
    $\mathsf{y} = \mathsf{H}_{1:i,1:i}^{-1} \mathsf{f}_{1:i}$; $\mathsf{x}_i = \mathsf{x}_0 + \mathsf{Q}_i \mathsf{y}$; STOP.
  end_if
  $\mathsf{q}_{i+1} = v / h_{i+1,i}$
end
$\mathsf{Q} = [\mathsf{q}_1, \ldots, \mathsf{q}_m]$; $\mathsf{H} = (h_{ij})_{(m+1)\times m}$; $\ell = -1$.

3.10.1 Konvergencija GMRES metode

Sada želimo proučiti mehanizam koji osigurava brzu redukciju reziduala.

U k–tom koraku imamo $\mathsf{x}_k = \mathsf{x}_0 + \mathsf{z}_k$, gdje je

$$\mathsf{z}_k = \sum_{j=0}^{k-1} \zeta_j \mathsf{A}^j \mathsf{r}_0 \in \mathcal{K}_k = \mathcal{L}(\mathsf{r}_0, \mathsf{A}\mathsf{r}_0, \ldots, \mathsf{A}^{k-1}\mathsf{r}_0)$$

korekcija, i pripadni rezidual $\mathsf{r}_k = \mathsf{b} - \mathsf{A}\mathsf{x}_k = \mathsf{r}_0 - \mathsf{A}\mathsf{z}_k$. Primijetimo da možemo pisati $\mathsf{A}\mathsf{z}_k = q(A)\mathsf{r}_0$, gdje je $q(t) = \sum_{j=1}^{k} \zeta_{j-1} t^j \in \mathcal{P}_k$. Ovdje $\mathcal{P}_k$ označava polinome stupnja najviše k.

Propozicija 3.10.2. *U k–tom koraku GMRES metode je*

$$\|\mathsf{r}_k\|_2 = \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \|p(\mathsf{A})\mathsf{r}_0\|_2.$$

Dokaz: Stavimo $p(x) = 1 - q(x)$. Odmah je $p(0) = 1$, $\mathsf{r}_k = p(\mathsf{A})\mathsf{r}_0$ i jasno je da variranjem korekcije z_k po $\mathcal{K}_k$ svi mogući reziduali su oblika $p(\mathsf{A})\mathsf{r}_0$, $p \in \mathcal{P}_k$, $p(0) = 1$. Isto tako, svakom takvom polinomu p odgovara korekcija koja reproducira $\mathsf{r}_k = p(\mathsf{A})\mathsf{r}_0$. $\boxplus$

Dakle, redukcija reziduala ovisi o ponašanju određene klase polinoma u varijabli A. Ostatak ove diskusije ćemo provesti u pojednostavljenim uvjetima i pretpostaviti da je A diagonalizabilna, $\mathsf{A} = \mathsf{S}\Lambda\mathsf{S}^{-1}$, $\Lambda = \mathrm{diag}(\lambda_i)_{i=1}^n$. Tada je

$$p(\mathsf{A}) = \mathsf{S}p(\Lambda)\mathsf{S}^{-1} = \mathsf{S}\begin{pmatrix} p(\lambda_1) & & \\ & \ddots & \\ & & p(\lambda_n) \end{pmatrix}\mathsf{S}^{-1}$$

pa vrijedi

$$\begin{aligned} \|p(\Lambda)\|_2 &= \max_{i=1:n} |p(\lambda_i)| \\ \|p(\mathsf{A})\|_2 &\leq \|\mathsf{S}\|_2\|\mathsf{S}^{-1}\|_2\|p(\Lambda)\|_2 = \kappa_2(\mathsf{S})\|p(\Lambda)\|_2 \\ \|p(\mathsf{A})\mathsf{r}_0\|_2 &\leq \kappa_2(\mathsf{S})\|p(\Lambda)\|_2\|\mathsf{r}_0\|_2 \end{aligned}$$

Odavde je

$$\|\mathsf{r}_k\|_2 = \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \|p(\mathsf{A})\mathsf{r}_0\|_2 \leq \kappa_2(\mathsf{S}) \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \|p(\Lambda)\|_2$$

pa je

$$\frac{\|\mathsf{r}_k\|_2}{\|\mathsf{r}_0\|_2} \leq \kappa_2(\mathsf{S}) \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \max_{i=1:n} |p(\lambda_i)| \leq \kappa_2(\mathsf{S}) \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \max_{z\in D} |p(z)|$$

gdje je $D \subset \mathbb{C}$ skup koji sadrži sve svojstvene vrijednosti od A. Ako je A normalna, onda je S unitarna i $\kappa_2(S) = 1$.

Vidimo da brzina redukcije reziduala ovisi o ponašanju određenih polinoma na spektru matrice A. To je polazna točka teorijske analize GMRES metode u koju ovajčas nećemo ulaziti. Tek jedan primjer: Neka je $\zeta_0 \in \mathbb{C}$, $\alpha \in \mathbb{R}$, $0 < \alpha < |\zeta_0|$ i $D = \{\zeta \in \mathbb{C} \ : \ |\zeta - \zeta_0| < \alpha\}$. Tada polinom $p(\zeta) = (1 - \zeta/\zeta_0)^k$ ima svojstvo $p(0) = 1$ i

$$\text{za } \zeta \in D, \ \ |p(\zeta)| = \left|\frac{\zeta_0 - \zeta}{\zeta_0}\right|^k \leq \left(\frac{\alpha}{|\zeta_0|}\right)^k.$$

Dakle, ako spektar od A možemo prekriti diskom D s centrom u ζ_0 i radijusom α koji je dosta manji od $|\zeta_0|$, onda će konvergencija GMRES metode garantirano biti brza.

Sada je prirodna ideja da se polazni sistem $\mathsf{Ax} = \mathsf{b}$ zamijeni ekvivalentnim (kažemo: *prekondicioniranim*) sustavom $(\mathsf{MA})\mathsf{x} = (\mathsf{Mb})$ pri čemu je M regularna matrica odabrana tako da nova matrica koeficijenata MA ima određenu distribuciju svojstvenih vrijednosti. Na primjer matrica M (koju zovemo *prekondicioner*) može biti neka gruba aproksimacija za A^{-1}. U dosta važnih primjena se uz konstrukciju matrice A može napraviti i dobar prekondicioner, ali to nije uvijek lagan posao. U praksi GMRES kao i sve ostale iterativne metode dobro funkcioniraju samo uz odgovarajući prekondicioner, tako da praktično gotovo nikada implementacije u komercijalnom, industrijskom, softwareu nisu samo ono što piše u knjigama. Isto tako, ponekad niti ne funkcioniraju. Razvoj dobrih iterativnih metoda i odgovarajućih tehnika prekondicioniranja je aktivno područje istraživanja.

3.11 Biortogonalni Lanczosev algoritam

Tročlana rekurzija (3.9.3) i tridijagonalna reprezentacija kompresije u Lanczosevom algoritmu su bitni faktori u razvoju algoritama, posebno kada se radi o matricama velike dimenzije. S druge strane, u Arnoldijevom algoritmu u i–tom koraku sudjeluju svi vektori $\mathsf{q}_1, \ldots, \mathsf{q}_i, \mathsf{q}_{i+1}$, a kompresija matrice $\mathsf{A} \in \mathbb{C}^{n\times n}$ na $\mathcal{K}_i \subseteq \mathbb{C}^n$ je $i \times i$ Hessenbergova matrica ($(i^2 + 3i)/2 - 1$ parametara). Da dobijemo osjećaj zašto je to važno, dovoljno je spomenuti da sa npr. $n = 10^6$ u dvostrukoj preciznosti (8 byte–ova za prikaz realnog broja) trebamo 8 Mb memorije za spremiti samo jedan vektor. Za tisuću takvih vektora trebamo 8 Gb memorije.

Dio II

Numeričko rješavanje problema svojstvenih vrijednosti

Poglavlje 4

Svojstvene vrijednosti

4.1 Numeričke metode

Sada ćemo proučavati metode za numeričko računanje svojstvenih vrijednosti i pripadnih svojstvenih vektora. Počet ćemo od najjednostavnijih ideja i razvijati ih do najsofisticiranijih algoritama.

Prije nego što krenemo na opis metoda, trebamo naučiti kako procijeniti kvalitetu izračunatih aproksimacija. Zato u sljedećoj sekciji dajemo nekoliko jednostavnih rezultata na tu temu, uz komentar da su to tek najjednostavniji elementi dosta komplicirane teorije perturbacija.

4.1.1 Rayleighev kvocijent

Ako je (λ, s) svojstveni par matrice A, onda je iz $\mathsf{As} = \lambda\mathsf{s}$ očito $\lambda = (\mathsf{s}^*\mathsf{As})/(\mathsf{s}^*\mathsf{s})$. Drugim riječima, svojstvenu vrijednost lako dobijemo iz pripadnog svojstvenog vektora. Ako imamo samo aproksimaciju svojstvenog vektora, $x \approx \mathsf{s}$, postavlja se pitanje koja je najbolja informacija o pripadnoj svojstvenoj vrijednosti, tj. kako odrediti najbolju aproksimaciju $\rho \approx \lambda$? Naravno, pitanje je i što znači *najbolja aproksimacija*?

Ako želimo da par (ρ, x) dobro aproksimira neki svojstveni par (λ, s) (za kojeg je $\mathsf{As} - \lambda\mathsf{s} = \mathbf{0}$), onda je jedan razuman kriterij da je rezidual $r = \mathsf{A}x - \rho x$ najmanji mogući u nekoj normi. Drugi kriterij može biti npr. da je (ρ, x) egzaktan svojstveni par neke matrice $\mathsf{A} + \delta A$ koja je blizu zadanoj matrici A. Naravno, bilo bi idealno znati kolika je razlika $|\lambda - \rho|$, ali je jasno da se za takav rezultat moramo malo više potruditi jer uključuje nepoznatu vrijednost λ.

Idemo korak po korak.

Teorem 4.1.1. *Za $x \neq \mathbf{0}$ i proizvoljni $\rho \in \mathbb{C}$ je (ρ, x) svojstveni par matrice $\mathsf{A} - \frac{r}{x^*x}x^*$, gdje je $r = \mathsf{A}x - \rho x$. Matrica $\delta\mathsf{A} = -\frac{r}{x^*x}x^*$ ima normu $\|\delta\mathsf{A}\|_2 = \|r\|_2/\|x\|_2$. Pri tome je $\|r\|_2$ minimalna ako je $\rho = \frac{x^*\mathsf{A}x}{x^*x}$.*

Dokaz: Lako je provjeriti da je

$$(\mathsf{A} - \frac{r}{x^*x}x^*)x = \rho x \;\text{ i }\; \| - \frac{r}{x^*x}x^*\|_2 = \frac{\|r\|_2}{\|x\|_2}.$$

Iz torema o projekciji je $\|\mathsf{A}x - \rho x\|_2$ minimalno ako je $\mathsf{A}x - \rho x$ okomit na x pa odmah dobijemo

$$\rho x = \frac{xx^*}{x^*x}\mathsf{A}x, \;\text{ tj. }\; \rho = \frac{x^*\mathsf{A}x}{x^*x}.$$

⊞

Definicija 4.1.1. Za $\mathsf{A} \in \mathbb{C}^{n\times n}$ i $x \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ definiramo Rayleighev kvocijent $\rho = \rho(\mathsf{A}, x) = \frac{x^*\mathsf{A}x}{x^*x}$.

Teorem 4.1.1 na neki način opravdava da u slučaju malog reziduala $r = \mathsf{A}x - \rho x$ skalar ρ možemo uzeti kao aproksimaciju neke svojstvene vrijednosti od A. Samo opravdanje je u činjenici da je ρ svojstvena vrijednost matrice $\mathsf{A} + \delta\mathsf{A}$ koja je blizu A ako je $\|r\|_2$ mali broj. Naravno, važno je i znati koliko je ρ daleko od neke svojstvene vrijednosti matrice A. Sljedeći teorem daje takvu informaciju.

Teorem 4.1.2. *(Bauer–Fike) Neka je A dijagonalizabilna, $\mathsf{A} = \mathsf{S}\Lambda\mathsf{S}^{-1}$. Ako je ρ svojstvena vrijednost matrice $\mathsf{A} + \delta\mathsf{A}$ onda je*

$$\min_{i=1:n} |\lambda_i - \rho| \leq \|\mathsf{S}\|\|\mathsf{S}^{-1}\|\|\delta\mathsf{A}\|$$

sa svakom matričnom normom $\|\cdot\|$ za koju je norma dijagonalne matrice maksimalna apsolutna vrijednost dijagonalnih elemenata, npr. $\|\cdot\|_2$, $\|\cdot\|_1$, $\|\cdot\|_\infty$.

Dokaz: Ako je ρ svojstvena vrijednost i od A onda tvrdnja očito vrijedi. Inače su $\mathsf{A} - \rho\mathsf{I}$ i $\Lambda - \rho\mathsf{I}$ regularne. Po pretpostavci je $\mathsf{A} + \delta\mathsf{A} - \rho\mathsf{I}$ singularna, pa je i $\Lambda + \mathsf{S}^{-1}\delta\mathsf{A}\mathsf{S} - \rho\mathsf{I}$ također singularna. Iz rastava

$$\Lambda + \mathsf{S}^{-1}\delta\mathsf{A}\mathsf{S} - \rho\mathsf{I} = (\Lambda - \rho\mathsf{I})(\mathsf{I} + (\Lambda - \rho\mathsf{I})^{-1}\mathsf{S}^{-1}\delta\mathsf{A}\mathsf{S})$$

onda nužno slijedi $\|(\Lambda - \rho\mathsf{I})^{-1}\mathsf{S}^{-1}\delta\mathsf{A}\mathsf{S}\| \geq 1$ u bilo kojoj matričnoj normi $\|\cdot\|$. Dakle, ako norma ima svojstvo opisano u iskazu teorema, vrijedi

$$1 \leq \|\mathsf{S}^{-1}\|\|\mathsf{S}\|\|\delta\mathsf{A}\|\|(\Lambda - \rho\mathsf{I})^{-1}\| = \|\mathsf{S}^{-1}\|\|\mathsf{S}\|\|\delta\mathsf{A}\|\frac{1}{\min_{i=1:n}|\lambda_i - \rho|}.$$

⊞

Komentar 4.1.1. Teoremi 4.1.1 i 4.1.2 ilustriraju dva aspekta analize numeričkih algoritama. U Teoremu 4.1.1 se uz aproksimaciju ρ, x računa i jedna mjera njihove kvalitete (rezidual r), te se dokazuje da je izračunata aproksimacija egzaktni rezultat (egzaktni svojstveni par) za matricu $\mathsf{A} + \delta\mathsf{A}$. Veličina perturbacije $\delta\mathsf{A}$ je određena veličinom reziduala $\|r\|$. Kažemo otprilike ovako: *Da, ρ možda nije svojstvena vrijednost od A, ali će biti ako A malo promijenimo, tj. ρ je svojstvena vrijednost od $\mathsf{A} + \delta\mathsf{A}$*. Dakle, činjenicu da ρ nije rješenje koje tražimo nego samo aproksimacija "ublažavamo" tako da dokazujemo da mala promjena polaznih podataka tu aproksimaciju pretvara u točno rješenje. Vidimo kako smo u Teoremu 4.1.1 rezidual r ugurali natrag u polazne podatke i to na način da je sa novim podacima par ρ, x postao svojstveni par. Kažemo da smo napravili *analizu greške unatrag* (*backward error analysis*). Valja naglasiti da ova analiza ne daje informaciju koliko je aproksimacija daleko od točnog rješenja. Kako točno rješenje ne znamo, onda procjenu udaljenosti izračunate aproksimacije do rješenja dobivamo teorijskim analizama promjene funkcije koja nas zanima (svojstvene vrijednosti) ako joj promijenimo argumente (matricu A). To je takozvana teorija perturbacija i Teorem 4.1.2 je jedan jednostavan primjer takvih razmatranja.

4.1.2 Metoda potencija

Metoda potencija je najjednostavnija metoda za računanje svojstvenih vrijednosti i vektora. Jednostavnom primjenom matrice A na polazni vektor x se generira niz vektora koji pod određenim uvjetima daje informaciju o najvećoj po modulu svojstvenoj vrijednosti i pripadnom vektoru.

Nekoliko je elementarnih ideja koje motiviraju metodu potencija. Za početak, promotrimo problem svojstvenih vrijednosti za matricu $\mathsf{A} = uv^*$, gdje su $u, v \in \mathbb{C}^n$, $u, v \neq \mathbf{0}$. Kako je $\mathsf{A}u = (v^*u)u$, vidimo da je $\lambda_1 = v^*u$ svojstvena vrijednost s pripadnim svojstvenim vektorom u. Kako je ortogonalni komplement od $\mathcal{L}(v)$ jezgra matrice A, vidimo da je ostatak spektra nula, $\lambda_2 = \cdots = \lambda_n = 0$.

Sada zamislimo da trebamo naći svojstveni vektor i svojstvenu vrijednost od A i da znamo da je matrica A ranga jedan, tj. da je oblika $\mathsf{A} = uv^*$, pri čemu su vektori

u, v nepoznati. Uzmimo proizvoljan vektor $x \neq \mathbf{0}$ i izračunajmo $y = \mathsf{A}x = u(v^*x)$. Ako je $y = \mathbf{0}$ onda je x jedan svojstveni vektor svojstvene vrijednosti nula. Ako je $y \neq \mathbf{0}$, onda je y kolinearan s vektorom u (svojstvenim vektorom jedine netrivijalne svojstvene vrijednosti). Dakle je y i sam svojstveni vektor, $\mathsf{A}y = \lambda_1 y$ i svojstvena vrijednost λ_1 se lako izračuna kao $\lambda_1 = y^*\mathsf{A}y/y^*y$.

Općenito imamo $n \times n$ matricu nepoznatog ranga, ali se pod određenim uvjetima može jednostavno dobiti informacija o jednom svojstvenom paru. Naime, ako je $\mathsf{A} = \mathsf{S}\Lambda\mathsf{S}^{-1}$, gdje je

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & \lambda_n \end{pmatrix}, \quad |\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|, \tag{4.1.1}$$

onda A^k možemo rastaviti na zbroj

$$\mathsf{A}^k = \mathsf{S}\Lambda^k\mathsf{S}^{-1} = \underbrace{\mathsf{S}\begin{pmatrix} \lambda_1^k & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}\mathsf{S}^{-1}}_{\mathsf{B}^k} + \underbrace{\mathsf{S}\begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2^k & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1}^k & 0 \\ 0 & 0 & \cdots & 0 & \lambda_n^k \end{pmatrix}\mathsf{S}^{-1}}_{\mathsf{C}^k}$$

pri čemu je A^k s rastućim k sve bliže matrici ranga jedan:

$$\frac{\|\mathsf{A}^k - \mathsf{B}^k\|_2}{\|\mathsf{A}^k\|_2} \leq \|\mathsf{S}\|_2\|\mathsf{S}^{-1}\|_2 \left|\frac{\lambda_2}{\lambda_1}\right|^k \to 0 \ \ (k \to \infty)$$

Dakle, ako je k dovoljno veliki, A^k je blizu matrice ranga jedan i $\mathsf{A}^k x$ bi trebao dati dobru informaciju o svojstvenom vektoru matrice A.

Sličan zaključak možemo dobiti i na sljedeći način: Kako je matrica A po pretpostavci dijagonalizabilna, njeni svojstveni vektori $\mathsf{s}_1, \ldots, \mathsf{s}_n$ (stupci matrice S) čine bazu u $\mathbb{C}^n$ pa bilo koji x možemo napisati kao linearnu kombinaciju $x = \xi_1\mathsf{s}_1 + \xi_2\mathsf{s}_2 + \cdots + \xi_n\mathsf{s}_n$. Sada lako izračunamo

$$\begin{aligned} \mathsf{A}x &= \xi_1\lambda_1\mathsf{s}_1 + \xi_2\lambda_2\mathsf{s}_2 + \cdots + \xi_n\lambda_n\mathsf{s}_n \\ \mathsf{A}^k x &= \xi_1\lambda_1^k\mathsf{s}_1 + \xi_2\lambda_2^k\mathsf{s}_2 + \cdots + \xi_n\lambda_n^k\mathsf{s}_n \\ &= \lambda_1^k(\xi_1\mathsf{s}_1 + \xi_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\mathsf{s}_2 + \cdots + \xi_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\mathsf{s}_n) \end{aligned}$$

Kako za sve $i > 1$ $(\lambda_i/\lambda_1)^k \to 0$ kada $k \to \infty$, vidimo da $\mathsf{A}^k x$ sve više "naginje" smjeru svojstvenog vektora s_1, pod uvjetom da je $\xi_1 \neq 0$. Da $\mathsf{A}^k x$ postaje paralelan s s_1 se lako vidi iz

$$\|\frac{\mathsf{A}^k x}{\lambda_1^k} - \xi_1 \mathsf{s}_1\|_2 \leq \left|\frac{\lambda_2}{\lambda_1}\right|^k (|\xi_2|\|\mathsf{s}_2\|_2 + \cdots + |\xi_n|\|\mathsf{s}_n\|_2) \to 0.$$

Aproksimacije obično uzimamo normirane u nekoj normi, npr. računamo

$$\begin{aligned} y^{(k)} = \frac{\mathsf{A}^k x}{\|\mathsf{A}^k x\|} &= \frac{\lambda_1^k(\xi_1\mathsf{s}_1 + \xi_2\left(\frac{\lambda_2}{\lambda_1}\right)^k \mathsf{s}_2 + \cdots + \xi_n\left(\frac{\lambda_n}{\lambda_1}\right)^k \mathsf{s}_n)}{|\lambda_1^k|\|\xi_1\mathsf{s}_1 + \xi_2\left(\frac{\lambda_2}{\lambda_1}\right)^k \mathsf{s}_2 + \cdots + \xi_n\left(\frac{\lambda_n}{\lambda_1}\right)^k \mathsf{s}_n\|} \\ &\approx \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{\xi_1}{|\xi_1|}\frac{\mathsf{s}_1}{\|\mathsf{s}_1\|} \equiv \eta_k \frac{\mathsf{s}_1}{\|\mathsf{s}_1\|} \end{aligned}$$

Odavde je jasno da niz $y^{(k)}$ nije nužno konvergentan, ali je sa dovoljno velikim k, $y^{(k)}$ uvijek blizu nekog svojstvenog vektora svojstvene vrijednosti λ_1. Sjetimo se, budući je λ_1 jednostruka svojstvena vrijednost, pripada joj jednodimenzionalni svojstveni potprostor, tj. svojstveni vektor je određen do na množenje skalarom različitim od nule. Zato je prirodnije konvergenciju mjeriti pomoću kuteva $\theta(y^{(k)}, \mathsf{s}_1) \in [0, \pi/2]$,

$$\cos\theta(y^{(k)}, \mathsf{s}_1) = \frac{|\mathsf{s}_1^* y^{(k)}|}{\|\mathsf{s}_1\|_2\|y^{(k)}\|_2} \longrightarrow 1, \quad \text{tj. } \theta(y^{(k)}, \mathsf{s}_1) \longrightarrow 0 \ \ (k \to \infty)$$

Dakle, iako vektori $y^{(k)}$ ne konvergiraju nekom fiksnom svojstvenom vektoru, linearne ljuske vektora $y^{(k)}$ (kao jednodimenzionalni potprostori) konvergiraju linearnoj ljusci od s_1 u smislu da niz kuteva $\theta(y^{(k)}, \mathsf{s}_1)$ konvergira u nulu.
Brzina konvergencije je određena kvocijentom $|\lambda_2|/|\lambda_1|$.

Algoritam 4.1.1. Metoda potencija za računanje svojstvenog vektora jedinstvene apsolutno dominantne svojstvene vrijednosti. Ulazna matrica A ima svojstvene vrijednosti $\lambda_1, \lambda_2, \ldots, \lambda_n$, numerirane tako da je $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$.

POTENCIJE($\mathsf{A}, x^{(0)}$)

$y^{(0)} = \dfrac{x^{(0)}}{\|x^{(0)}\|}$
$k = 0$
ponavljaj
  $x^{(k+1)} = \mathsf{A}y^{(k)}$
  $y^{(k+1)} = \dfrac{x^{(k+1)}}{\|x^{(k+1)}\|}$
  $k = k + 1$
do konvergencije

Komentar 4.1.2. Normiranje $y^{(k+1)} = x^{(k+1)}/\|x^{(k+1)}\|$ je posebno važno kada računamo na računalu u konačnoj aritmetici jer vektori $\mathsf{A}^k x$, ovisno o spektru matrice A, mogu postati jako veliki ili jako mali tako da ih se ne može reprezentirati u zadanom formatu na računalu.

4.1.2.1 Analiza metode potencija u općem slučaju

Prethodnu analizu smo napravili u specijalnom slučaju dijagonalizabile matrice A u kojoj je dominantna po modulu svojstvena vrijednost strogo veća od ostalih, $|\lambda_1| > \max_{i=2:n} |\lambda_i|$. Naravno da je važno znati što se dešava ako te pretpostavke nisu ispunjene.

– –

4.1.3 Inverzne iteracije

Promotrimo sada ponovo dijagonalizabilnu matricu A kojoj je najmanja po modulu svojstvena vrijednost različita od nule i dobro odvojena od preostalih svojstvenih vrijednosti,

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_{n-1}| > |\lambda_n| > 0$$

Recimo da trebamo λ_n, zajedno sa pripadnim svojstvenim vektorom. Odmah možemo iskoristiti jednostavnu ideju: Svojstvene vrijednosti od A^{-1} su $1/\lambda_i$, $i = 1, \ldots, n$, i vrijedi

$$\left|\frac{1}{\lambda_n}\right| > \left|\frac{1}{\lambda_{n-1}}\right| \geq \cdots \geq \left|\frac{1}{\lambda_2}\right| \geq \left|\frac{1}{\lambda_1}\right|.$$

Dakle, metodom potencija možemo aproksimirati dominantni svojstveni vektor od A^{-1}, koji je zapravo svojstveni vektor od A sa svojstvenom vrijednosti λ_n. Primjena metode potencija na A^{-1} s ciljem dobivanja najmanje po modulu svojstvene

vrijednosti se zove *metoda inverznih iteracija*. Sada, za razliku od metode potencija, u k–tom koraku imamo $x^{(k+1)} = \mathsf{A}^{-1}y^{(k)}$, tj. svaki novi vektor $x^{(k+1)}$ se dobije rješavanjem linearnog sustava $\mathsf{A}x^{(k+1)} = y^{(k)}$. Brzina konvergencije je određena kvocijentom

$$\frac{|\lambda_{n-1}^{-1}|}{|\lambda_n^{-1}|} = \left|\frac{\lambda_n}{\lambda_{n-1}}\right|.$$

Sada, kada smo upoznali metodu potencija i inverzne iteracija, sljedeći koraci su motivirani sljedećim dosta razumnim pitanjima:

- Da li možemo na neki način poboljšati brzinu konvergencije? Primijetimo da je svaki korak inverznih iteracija zahtjevan jer rješavanje linearnog sustava jednadžbi u svakom koraku i s jako velikom dimenzijom n to nije jednostavno.
- Da li možemo na sličan način aproksimirati svojstvene parove koji nisu ekstremni, tj. da li možemo aproksimirati bilo koji λ_i i pripadni vektor? Što ako trebamo sve svojstvene vrijednosti?

Komentar 4.1.3. Kako to često biva, jednostavne ideje, strpljivo kombinirane i malo po malo dograđivane isto tako jednostavnim elementima će na kraju dati elegantan i efikasan numerički algoritam. Primijetimo da za sada radimo u specijalnim uvjetima – pretpostavljamo da je A dijagonalizabilna, da trebamo samo jedan svojstveni par, i da je tražena svojstvena vrijednost dobro odvojena od ostalih. To je zapravo prirodan pristup problemu – pojednostavimo ga dodatnim pretpostavkama (od kojih neke možda čak jako pojednostavljuju problem) i u tim uvjetima studiramo rješenje. Nakon što smo riješili pojednostavljeni problem i dobili nove spoznaje, malo po malo mičemo dodane pretpostavke.

Nastavljamo s jednostavnim idejama. Prelaz sa metode potencija na metodu inverznih iteracija je motiviran ekvivalentnosti relacija $\mathsf{A}\mathsf{s}_i = \lambda_i\mathsf{s}_i$ i $\mathsf{A}^{-1}\mathsf{s}_i = (1/\lambda_i)\mathsf{s}_i$. Sada uočimo jednostavne transformacije matrice A i svojstvenih vrijednosti,

$$(\mathsf{A} - \sigma\mathsf{I})\mathsf{s}_i = (\lambda_i - \sigma)\mathsf{s}_i \tag{4.1.2}$$

$$(\mathsf{A} - \sigma\mathsf{I})^{-1}\mathsf{s}_i = \frac{1}{\lambda_i - \sigma}\mathsf{s}_i. \ (\sigma \notin \mathfrak{S}(\mathsf{A})) \tag{4.1.3}$$

Znači, ako primijenimo metodu inverznih iteracija na $\mathsf{A} - \sigma\mathsf{I}$, imat ćemo brzinu konvergencije određenu s

$$\gamma(\sigma) = \left|\frac{\lambda_n - \sigma}{\lambda_{n-1} - \sigma}\right|$$

Ako je σ puno bliže ciljanoj vrijednosti λ_n nego bilo kojoj drugoj svojstvenoj vrijednosti, onda je $\gamma(\sigma) \ll 1$ i imamo brzu konvergenciju. Vidimo i više: ako je σ puno bliže nekoj svojstvenoj vrijednosti λ_j nego bilo kojoj drugoj svojstvenoj vrijednosti, onda je $\lambda_j - \sigma$ najmanja po modulu svojstvena vrijednost od $\mathsf{A} - \sigma\mathsf{I}$ pa će inverzne iteracije konvergirati k s_j, a pripadni Rayleighevi kvocijenti k λ_j.

Algoritam 4.1.2. Metoda inverznih iteracija za računanje svojstvenog vektora i pripadne svojstvene vrijednosti koja je najbliža zadanom parametru σ. Ulazna matrica A ima svojstvene vrijednosti $\lambda_1, \lambda_2, \ldots, \lambda_n$

INVERZNE ITERACIJE($\mathsf{A}, x^{(0)}, \sigma$)

$$y^{(0)} = \frac{x^{(0)}}{\|x^{(0)}\|}$$
$k = 0$
ponavljaj
$$x^{(k+1)} = (\mathsf{A} - \sigma\mathsf{I})^{-1} y^{(k)}$$
$$y^{(k+1)} = \frac{x^{(k+1)}}{\|x^{(k+1)}\|}$$
$k = k + 1$
do konvergencije

Naravno, odmah se nameću dodatna pitanja:

- Kako naći σ koji je blizu nepoznate vrijednosti λ_j?
- Ako imamo tako dobar $\sigma \approx \lambda_j$, onda je $\mathsf{A} - \sigma\mathsf{I}$ blizu singularne matrice $\mathsf{A} - \lambda_j\mathsf{I}$ pa je linearni sustav $(\mathsf{A} - \sigma\mathsf{I})x^{(k+1)} = y^{(k)}$ koji definira $x^{(k+1)}$ skoro singularan što znači da izračunati $x^{(k+1)}$ može biti netočan.
- Što znači "do konvergencije", tj. kada zaustaviti iteracije?

4.1.4 Iteracije potprostora

Vratimo se na metodu potencija i razmislimo o poopćenju koje bi moglo aproksimirati više od jedne svojstvene vrijednosti. Dakle, umjesto jednog jednodimenzionalnog svojstvenog potprostora bismo želji ℓ–dimenzionalni invarijantni potprostor $\mathcal{V}$ reprezentiran kao slika matrice $\mathsf{V} \in \mathbb{C}^{n\times\ell}$ takve da je $\mathsf{V}^*\mathsf{V} = \mathsf{I}_\ell$ (stupci od V su ortonormirana baza za $\mathcal{V}$). Da je $\mathcal{V}$ A–invarijantan po definiciji znači da je $\mathsf{A}\mathcal{V} \subseteq \mathcal{V}$. Matrično, to zapisujemo kao $\mathsf{AV} = \mathsf{VM}$, gdje je $\mathsf{M} = \mathsf{V}^*\mathsf{AV}$ matrični Rayleighev kvocijent.

Sjetimo se, invarijantni potprostori su važni u rješavanju problema svojstvenih vrijednosti: Ako matricu V, čiji stupci razapinju A–invarijantan potprostor, dopunimo do unitarne $(\mathsf{V}, \mathsf{V}_\perp)$ onda, koristeći invarijantnost, transformacija sličnosti

$$\begin{pmatrix} \mathsf{V}^* \\ \mathsf{V}_\perp^* \end{pmatrix} \mathsf{A} \begin{pmatrix} \mathsf{V} & \mathsf{V}_\perp \end{pmatrix} = \begin{pmatrix} \mathsf{V}^*\mathsf{AV} & \mathsf{V}^*\mathsf{AV}_\perp \\ \mathsf{V}_\perp^*\mathsf{AV} & \mathsf{V}_\perp^*\mathsf{AV}_\perp \end{pmatrix} = \begin{pmatrix} \mathsf{M} & \mathsf{V}^*\mathsf{AV}_\perp \\ \mathbf{0} & \mathsf{V}_\perp^*\mathsf{AV}_\perp \end{pmatrix}$$

odmah daje rezultat da su svojstvene vrijednosti od M ujedno i svojstvene vrijednosti od A. Naime, ako je $\mathsf{M}y = \lambda y$, onda $\mathsf{AV} = \mathsf{VM}$ povlači $\mathsf{A}(\mathsf{V}y) = \mathsf{VM}y = \lambda(\mathsf{V}y)$

Trivijalno poopćenje metode potencija je da jednostavno umjesto polaznog vektora $x^{(0)} \neq \mathbf{0}$ uzmemo ℓ linearno neovisnih vektora, $x^{(1,0)}, \ldots, x^{(\ell,0)}$ poslažemo ih u $n \times \ell$ matricu $X^{(0)} = (x^{(1,0)} \cdots x^{(\ell,0)})$ i u k–toj iteraciji računamo $X^{(k)} = \mathsf{A}^k X^{(0)}$. Dakle, to je ℓ istovremenih iteracija metode potencija i javlja se očit problem – stupci matrice $X^{(k)}$ postaju sve lošija baza za njenu sliku i svi konvergiraju istom vekoru (npr. ako je $|\lambda_1|$ jedinstvena dominantna svojstvena vrijednost u limesu imamo ℓ vektora kolinearnih s s_1). Jednostavna ideja kako to spriječiti je da $X^{(k)}$ zamijenimo matricom koja ima istu sliku i stupce koji su ortonormirana baza te slike. Nadalje, same iteracije ćemo shvatiti kao generiranje niza potprostora koji su zadani kao slike odgovarajućih matrica (pa su stupci tih matrica baze za odgovarajuće potprostore).

Algoritam 4.1.3. Iteracije potprostora za računanje invarijantnog potprostora razapetog s ℓ svojstvenih vektora od ℓ apsolutno dominantnih svojstvenih vrijednosti. Ulazna matrica A ima svojstvene vrijednosti $\lambda_1, \lambda_2, \ldots, \lambda_n$, numerirane tako da je $|\lambda_1| \geq \cdots \geq |\lambda_\ell| > |\lambda_{\ell+1}| \geq \cdots \geq |\lambda_n|$.

ITERACIJE POTPROSTORA($\mathsf{A}, X^{(0)}$)

$X^{(0)} = \mathsf{Y}^{(0)} R^{(0)}$ (QR faktorizacija, $(\mathsf{Y}^{(0)})^* \mathsf{Y}^{(0)} = \mathsf{I}_\ell$)
$k = 0$
ponavljaj
  $X^{(k+1)} = \mathsf{A}\mathsf{Y}^{(k)}$
  $X^{(k+1)} = \mathsf{Y}^{(k+1)} R^{(k+1)}$ (QR faktorizacija)
  $k = k + 1$
do konvergencije

Komentar 4.1.4. Iz formula u Algoritmu 4.1.3 je jasno da uzimanjem prvih $\ell' < \ell$ stupaca u matricama $X^{(k)}$ i $\mathsf{Y}^{(k)}$ dobijemo identičan proces ali s ℓ'–dimenzionalnim potprostorima. Dakle, algoritam istovremeno izvodi ℓ iteracija.

Komentar 4.1.5. QR faktorizacija $X^{(k+1)} = \mathsf{Y}^{(k+1)}R^{(k+1)}$ ima zadatak osigurati da je slika matrice $X^{(k+1)}$ reprezentirana u dobroj bazi, te spriječiti efekte metode potencija. Taj korak možemo realizirati npr. modificiranim Gram–Schmidtovim algoritmom i to ne nužno u svakom koraku.

4.1.4.1 Skica dokaza konvergencije

Radi jednostavnosti, pretpostavimo da je A regularna dijagonalizabilna matrica sa spektralnom dekompozicijom $\mathsf{A} = \mathsf{S}\Lambda\mathsf{S}^{-1}$, gdje su svojstvene vrijednosti na dijagonali od Λ numerirane tako da je $|\lambda_1| \geq \cdots \geq |\lambda_\ell| > |\lambda_{\ell+1}| \geq \cdots \geq |\lambda_n| > 0$.

Ako je $\ell = 1$ imamo običnu metodu potencija, pa zato uzmimo $\ell > 1$. Zbog regularnosti, za slike izračunatih matrica vrijedi $\mathfrak{R}(\mathsf{Y}^{(\mathsf{k})}) = \mathfrak{R}(\mathsf{X}^{(\mathsf{k})}) = \mathsf{A}\mathfrak{R}(\mathsf{Y}^{(\mathsf{k-1})}) = \cdots = \mathsf{A}^k\mathfrak{R}(\mathsf{Y}^{(0)})$. Pogledajmo strukturu matrice $\mathsf{A}^k\mathsf{Y}^{(0)} \equiv \mathsf{S}\Lambda^k\mathsf{S}^{-1}\mathsf{Y}^{(0)}$:

$$
\begin{aligned}
\mathsf{S}\Lambda^k\mathsf{S}^{-1}\mathsf{Y}^{(0)} &= \mathsf{S}\begin{pmatrix} {\lambda_1}^k & & & & & & \\ & \ddots & & & & & \\ & & \lambda_{\ell-1}^k & & & & \\ & & & \lambda_\ell^k & & & \\ & & & & \lambda_{\ell+1}^k & & \\ & & & & & \ddots & \\ & & & & & & \lambda_n^k \end{pmatrix}\mathsf{S}^{-1}\mathsf{Y}^{(0)} \\
&= \mathsf{S}\begin{pmatrix} D_1^k & \mathbf{0} \\ \mathbf{0} & D_2^k \end{pmatrix}\begin{pmatrix} \Xi_0 \\ F_0 \end{pmatrix},\ D_1 = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_\ell \end{pmatrix} \\
&\qquad D_2 = \begin{pmatrix} \lambda_{\ell+1} & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix},\ \begin{aligned} \Xi_0 &= (\mathsf{S}^{-1}\mathsf{Y}^{(0)})(1:\ell, 1:\ell) \\ F_0 &= (\mathsf{S}^{-1}\mathsf{Y}^{(0)})(\ell+1:n, 1:\ell) \end{aligned}
\end{aligned}
$$

Pretpostavimo da je $\det(\Xi_0) \neq 0$ i matricu S particionirajmo s $\mathsf{S} = \begin{pmatrix} \mathsf{S}_1 & \mathsf{S}_2 \end{pmatrix}$, gdje S_1 sadrži prvih ℓ stupaca (svojstvenih vektora za $\lambda_1, \ldots, \lambda_\ell$). Stavimo $\mathcal{S} = \mathfrak{R}(\mathsf{S}_1)$, $\mathcal{X}_k = \mathfrak{R}(\mathsf{X}^{(k)})$. Želimo dokazati da $\mathcal{X}_k \longrightarrow \mathcal{S}$ kada $k \to \infty$.

Iz prethodnih relacija lagano dobijemo $\mathsf{S}\Lambda^k\mathsf{S}^{-1}\mathsf{Y}^{(0)} = \mathsf{S}_1 D_1^k \Xi_0 + \mathsf{S}_2 D_2^k F_0$ i

$$
\mathsf{S}\Lambda^k\mathsf{S}^{-1}\mathsf{Y}^{(0)}\Xi_0^{-1}D_1^{-k} = \mathsf{S}_1 + \mathsf{S}_2 D_2^k F_0 \Xi_0 D_1^{-k}, \tag{4.1.4}
$$

gdje lako provjerimo da je $\lim_{k\to\infty} \mathsf{S}_2 D_2^k F_0 \Xi_0 D_1^{-k} = \mathbf{0}$. Drugim riječima, za svaki dovoljno veliki k u potprostoru $\mathcal{X}_k$ možemo naći bazu u kojoj je ℓ vektora koji su u $1 \leftrightsquigarrow 1$ korespondenciji po volji blizu stupcima matrice S_1 koji su baza u $\mathcal{S}$.

4.1.5 QR metoda

Vidjeli smo da u slučaju dobro odvojenih svojstvenih vrijednosti $|\lambda_i| > |\lambda_{i+1}|$, tj.

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_{n-1}| > |\lambda_n| \tag{4.1.5}$$

iteracije potprostorima (Algoritam 4.1.3) u limesu računaju unitarnu matricu koja daje Schurovu formu od A. Međutim, čak i kada bismo imali uvjet (4.1.5), ostaje činjenica da su iteracije ekstremno skupe: u svakom koraku se množe dvije $n \times n$ matrice i računa QR faktorizacija $n \times n$ matrice. To znači da je cijena svakog koraka iteracija $O(n^3)$ aritmetičkih operacija, pri čemu je za dobru aproksimaciju spektra obično potrebno n ili više iteracija.

Sada je ideja te iteracije zamijeniti ekvivalentnim procesom kojeg se dade dalje pojednostaviti (u smislu broja računskih operacija u jednom koraku) i poboljšati u smislu konvergencije.

Algoritam 4.1.4. Ulazna matrica A ima apsolutno dobro odvojene svojstvene vrijednosti $\lambda_1, \lambda_2, \ldots, \lambda_n$, numerirane tako da vrijedi (4.1.5).

QR ITERACIJE($\mathsf{A}, X^{(0)}$)
$\mathsf{A}^{(1)} = \mathsf{A}$ $k = 1$ ponavljaj $\quad \mathsf{A}^{(k)} = \mathsf{Q}^{(k)}\mathsf{R}^{(k)}$ (QR faktorizacija) $\quad \mathsf{A}^{(k+1)} = \mathsf{R}^{(k)}\mathsf{Q}^{(k)}$ $\quad k = k + 1$ do konvergencije

Teorem 4.1.3. *Matrice izračunate u Algoritmu 4.1.4 imaju sljedeća svojstva:*

- *Za svaki k je $\mathsf{A}^{(k+1)} = (\mathsf{Q}^{(k)})^*\mathsf{A}^{(k)}\mathsf{Q}^{(k)}$, tj. algoritam generira niz unitarno sličnih matrica.*
- *Za svaki k je $\mathsf{A}^{(k+1)} = (\mathsf{Q}^{(1)}\mathsf{Q}^{(2)}\cdots\mathsf{Q}^{(k)})^*\mathsf{A}(\mathsf{Q}^{(1)}\mathsf{Q}^{(2)}\cdots\mathsf{Q}^{(k)})$.*
- *Ako definiramo $\mathsf{Q}^{[1:k]} = \mathsf{Q}^{(1)}\mathsf{Q}^{(2)}\cdots\mathsf{Q}^{(k)}$ i $\mathsf{R}^{[1:k]} = \mathsf{R}^{(k)}\mathsf{R}^{(k-1)}\cdots\mathsf{R}^{(1)}$, onda je $\mathsf{A}^k = \mathsf{Q}^{[1:k]}\mathsf{R}^{[1:k]}$ QR faktorizacija potencije A^k.*

Dokaz:

- Vrijedi $\mathsf{A}^{(k+1)} = \mathsf{R}^{(k)}\mathsf{Q}^{(k)} = (\mathsf{Q}^{(k)})^*\mathsf{Q}^{(k)}\mathsf{R}^{(k)}\mathsf{Q}^{(k)} = (\mathsf{Q}^{(k)})^*\mathsf{A}^{(k)}\mathsf{Q}^{(k)}$.

- Induktivno, koristeći prethodnu tvrdnju, imamo :

$$\mathsf{A}^{(k+1)} = (\mathsf{Q}^{(k)})^* \mathsf{A}^{(k)} \mathsf{Q}^{(k)} = (\mathsf{Q}^{(k)})^* (\mathsf{Q}^{(k-1)})^* \mathsf{A}^{(k-1)} \mathsf{Q}^{(k-1)} \mathsf{Q}^{(k)}.$$

- Tvrdnja postaje jasna odmah nakon proučavanja strukture prvih nekoliko potencija. Tako je

$$\begin{aligned}
\mathsf{A}^2 &= \mathsf{Q}^{(1)} \underbrace{\mathsf{R}^{(1)}\mathsf{Q}^{(1)}}_{\mathsf{A}^{(2)}=\mathsf{Q}^{(2)}\mathsf{R}^{(2)}} \mathsf{R}^{(1)} = \mathsf{Q}^{(1)}\mathsf{Q}^{(2)}\mathsf{R}^{(2)}\mathsf{R}^{(1)} \\
\mathsf{A}^3 &= \mathsf{Q}^{(1)}\mathsf{R}^{(1)}\mathsf{Q}^{(1)}\mathsf{Q}^{(2)}\mathsf{R}^{(2)}\mathsf{R}^{(1)} = \mathsf{Q}^{(1)}\mathsf{Q}^{(2)} \underbrace{\mathsf{R}^{(2)}\mathsf{Q}^{(2)}}_{\mathsf{A}^{(3)}=\mathsf{Q}^{(3)}\mathsf{R}^{(3)}} \mathsf{R}^{(2)}\mathsf{R}^{(1)} \\
&= \mathsf{Q}^{(1)}\mathsf{Q}^{(2)}\mathsf{Q}^{(3)}\mathsf{R}^{(3)}\mathsf{R}^{(2)}\mathsf{R}^{(1)}
\end{aligned}$$

⊞

Sada ćemo pokazati da su u QR algoritmu skriveni i metoda iteracija potprostora i inverzne iteracije.

Propozicija 4.1.4. *Neka je* A *regularna matrica. Uz oznake iz Teorema 4.1.3 vrijedi:*

- *Za svaki* $\ell = 1, \ldots, n$*, prvih* ℓ *stupaca matrice* $\mathsf{Q}^{[1:k]}$ *u svakom koraku* k *razapinju isti potprostor kao u metodi iteracija potprostora primijenjenoj na* A *sa početnom iteracijom* $\mathsf{X}^{(0)} = \mathsf{I}_n(:, 1:\ell)$*.*

- *Za svaki* $\ell = 1, \ldots, n$*, zadnjih* ℓ *stupaca matrice* $\mathsf{Q}^{[1:k]}$ *u svakom koraku* k *razapinju isti potprostor kao u metodi inverznih iteracija primijenjenoj na* A^* *sa početnom iteracijom* $\mathsf{X}^{(0)} = \mathsf{I}_n(:, 1:\ell)$*.*

Dokaz: Iz druge tvrdnje Teorema 4.1.3 slijedi $\mathsf{Q}^{[1:k-1]}\mathsf{A}^{(k)} = \mathsf{A}\mathsf{Q}^{[1:k-1]}$ pa je

$$\mathsf{Q}^{[1:k]}\mathsf{R}^{[1:k]} = \mathsf{Q}^{[1:k-1]}\mathsf{Q}^{(k)}\mathsf{R}^{(k)}\mathsf{R}^{[1:k-1]} = \mathsf{Q}^{[1:k-1]}\mathsf{A}^{(k)}\mathsf{R}^{[1:k-1]} = \mathsf{A}\mathsf{Q}^{[1:k-1]}\mathsf{R}^{[1:k-1]}$$

Odavde je $\mathsf{Q}^{[1:k]}\mathsf{R}^{(k)} = \mathsf{A}\mathsf{Q}^{[1:k-1]}$, tj. za svaki $\ell = 1, \ldots, n$ imamo QR faktorizaciju koja predstavlja jedan korak metode potencija

$$\mathsf{Q}^{[1:k]}(:, 1:\ell)\ \mathsf{R}^{(k)}(1:\ell, 1:\ell) = \mathsf{A}\ \mathsf{Q}^{[1:k-1]}(:, 1:\ell),$$

a inicijalno je $\mathsf{Q}^{(1)}\mathsf{R}^{(1)} = \mathsf{A}$, tj. $\mathsf{Q}^{(1)}(:, 1:\ell)\ \mathsf{R}^{(1)}(1:\ell, 1:\ell) = \mathsf{A}\ \mathsf{I}_n(:, 1:\ell)$.

Za dokaz druge tvrdnje, prvo relaciju $\mathsf{Q}^{[1:k]}\mathsf{R}^{(k)} = \mathsf{A}\mathsf{Q}^{[1:k-1]}$ invertiranjem i adjungiranjem transformiramo u

$$\mathsf{Q}^{[1:k]} = \mathsf{A}^{-*}\mathsf{Q}^{[1:k-1]}(\mathsf{R}^{(k)})^{*}.$$

Sada se prisjetimo Komentara 1.2.1 i permutacije $\pi(i) = n - i + 1$. Ako je Π matrica te permutacije, onda relacija

$$(\mathsf{Q}^{[1:k]}\Pi)(\Pi(\mathsf{R}^{(k)})^{*}\Pi)^{-1} = \mathsf{A}^{-*}(\mathsf{Q}^{[1:k-1]}\Pi)$$

predstavlja jedan korak inverznih iteracija potprostora s matricom A^{*}, pa čitanjem prvih ℓ stupaca dobijemo tvrdnju. ⊞

Komentar 4.1.6. Ako su svojstvene vrijednosti od A uređene po apsolutnoj vrijednsoti u padajući niz tako da je $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_{n-1}| > |\lambda_n| > 0$, onda znamo da metoda inverznih iteracija (primijenjena na A) daje niz koji konvergira svojstvenom vektoru od λ_n. To vrijedi i u slučaju matrice A^{*} jer su njene svojstvene vrijednosti $\overline{\lambda}_i$, $i = 1, \ldots, n$; specijalno je tada u limesu metode inverznih iteracija jedinični vektor u_n za kojeg je $\mathsf{A}^{-*}\mathsf{u}_n = (1/\overline{\lambda}_n)\mathsf{u}_n$, tj. $\mathsf{u}^{*}\mathsf{A} = \lambda_n \mathsf{u}_n^{*}$. (Po definiciji je u_n *lijevi* svojstveni vektor od A i svojstvene vrijednosti λ_n.) U QR algoritmu to znači, prema Propoziciji 4.1.4, da zadnji stupac u nizu matrica $\mathsf{Q}^{[1:k]}$ zapravo konvergira k u_n, brzinom koju određuje kvocijent $|\lambda_n|/|\lambda_{n-1}|$. Odatle slijedi da zadnji redak u nizu $\mathsf{A}^{(k)}$ u limesu postaje $\begin{pmatrix} 0 & \cdots & 0, \lambda_n \end{pmatrix}$. Slično se može zaključivati iz veze QR algoritma i metode potencija ako je dominantna po apsolutnoj vrijednosti svojstvena vrijednost dobro odvojena od ostali. Ipak, valja odmah naglasiti da je veza sa metodom inverznih iteracija važnija jer u tom slučaju znamo da pomacima (*shift*ovima) možemo ubrzati konvergenciju.

4.1.5.1 Konvergencija QR iteracija

Dokaz konvergencije QR algoritma je netrivijalan, s dosta otvorenih problema. Do danas ne postoji dokaz globalne konvergencije za opće kvadratne matrice, zapravo je poznato da postoje skupovi matrica na kojima divergira.

Globalna konvergencija se može dokazati uz dodatne pretpostavke o svojstvenim vrijednostima ili strukturi matrice (npr. hermitičnost).

Teorem 4.1.5. *Neka je $\mathsf{A} \in \mathbb{C}^{n\times n}$ regularna matrica sa svojstvenim vrijednostima (4.1.5). Tada niz matrica $\mathsf{A}^{(k)}$ izračunat u Algoritmu 4.1.4 konvergira u sljedećem*

smislu: Postoje dijagonalne unitarne matrice $\Phi^{(k)}$ *takve da je*

$$\lim_{k\to\infty}(\Phi^{(k)})^*\mathsf{A}^{(k+1)}\Phi^{(k)} = \begin{pmatrix} \lambda_1 & * & \cdots & \cdots & * \\ 0 & \lambda_2 & * & \cdots & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot \\ \vdots & \vdots & \ddots & \lambda_{n-1} & * \\ 0 & 0 & \cdots & 0 & \lambda_n \end{pmatrix} = \mathsf{Q}^*\mathsf{A}\mathsf{Q},$$

gdje je $\mathsf{Q} = \lim_{k\to\infty}\mathsf{Q}^{(1)}\mathsf{Q}^{(2)}\cdots\mathsf{Q}^{(k)}\Phi^{(k)}$.

Dokaz: U dokazu koristimo oznake iz Algoritma 4.1.4 i Teorema 4.1.3. Sam dokaz ćemo napraviti u dva prolaza: prvo ćemo tvrdnju dokazati uz još jednu *dodatnu pretpostavku* koja će olakšati tehnički dio posla, a u drugom prolazu ćemo provesti dokaz bez te dodatne pretpostavke.

Iz pretpostavki teorema slijedi da A ima samo jednostruke svojstvene vrijednosti pa je dijagonalizabilna, $\mathsf{A} = \mathsf{S}\Lambda\mathsf{S}^{-1}$, $\Lambda = \mathrm{diag}(\lambda_i)_{i=1}^n$. Neka je $\mathsf{S} = \mathsf{Q}\mathsf{R}$ QR faktorizacija matrice S i neka je $\mathsf{S}^{-1} = \mathsf{L}\mathsf{T}$ LU faktorizacija od S^{-1}. Primijetimo da ovdje koristimo *dodatnu pretpostavku da ta* LU *faktorizacija postoji tj. da su sve vodeće minore u* S^{-1} *različite od nule.*

Sada se prisjetimo treće tvrdnje Teorema 4.1.3: matrica A^k ima QR faktorizaciju $\mathsf{A}^k = \mathsf{Q}^{[1:k]}\mathsf{R}^{[1:k]}$. S druge strane, vrijedi

$$\mathsf{A}^k = \mathsf{S}\Lambda^k\mathsf{S}^{-1} = \mathsf{Q}\mathsf{R}\Lambda^k\mathsf{L}\mathsf{T} = \mathsf{Q}\mathsf{R}(\Lambda^k\mathsf{L}\Lambda^{-k})\Lambda^k\mathsf{T}. \tag{4.1.6}$$

Kako je L donje trokutasta matrica s jediničnom dijagonalom, za elemente strogo donjeg trokuta vrijedi

$$(\Lambda^k\mathsf{L}\Lambda^{-k})_{ij} = \mathsf{L}_{ij}\left(\frac{\lambda_i}{\lambda_j}\right)^k \longrightarrow 0 \ (k\to\infty) \tag{4.1.7}$$

jer je za $i > j$, $|\lambda_i| < |\lambda_j|$. Znači da možemo pisati $\Lambda^k\mathsf{L}\Lambda^{-k} = \mathsf{I} + E_k$, $E_k \to \mathbf{0}$ za $k\to\infty$. Dakle, imamo

$$\mathsf{A}^k = \mathsf{Q}\mathsf{R}(\Lambda^k\mathsf{L}\Lambda^{-k})\Lambda^k\mathsf{T} = \mathsf{Q}\mathsf{R}(\mathsf{I}+E_k)\Lambda^k\mathsf{T} = \mathsf{Q}(\mathsf{I}+\underbrace{\mathsf{R}E_k\mathsf{R}^{-1}}_{F_k})\mathsf{R}\Lambda^k\mathsf{T}$$

gdje $F_k = \mathsf{R}E_k\mathsf{R}^{-1} \to \mathbf{0}$. Ako napišemo QR faktorizaciju matrice $\mathsf{I}+F_k$, $\mathsf{I}+F_k = Q_{F_k}R_{F_k}$, gdje R_{F_k} ima pozitivne dijagonalne elemente, onda je, zbog neprekidnosti, $\lim_{k\to\infty}R_{F_k} = \mathsf{I}$, $\lim_{k\to\infty}Q_{F_k} = \mathsf{I}$. Dakle, možemo pisati

$$\mathsf{A}^k = \mathsf{Q}^{[1:k]}\mathsf{R}^{[1:k]} = (\mathsf{Q}Q_{F_k})(R_{F_k}\mathsf{R}\Lambda^k\mathsf{T}) \equiv (\text{unitarna})(\text{gornje trokutasta}) \tag{4.1.8}$$

što su zapravo dvije QR faktorizacije matrice A^k. Kako je QR faktorizacija u biti jedinstvena, znamo da postoji dijagonalna unitarna matrica $\Phi^{(k)} = \mathrm{diag}(\mathbf{e}^{\mathrm{i}\varphi_j^{(k)}})_{j=1}^n$ tako da je $\mathsf{Q}^{[1:k]} = \mathsf{Q}Q_{F_k}(\Phi^{(k)})^*$, $\mathsf{R}^{[1:k]} = \Phi^{(k)}R_{F_k}\mathsf{R}\Lambda^k\mathsf{T}$. Slijedi da

$$\mathsf{Q}^{[1:k]}\Phi^{(k)} - \mathsf{Q} = \mathsf{Q}(Q_{F_k} - \mathsf{I}) = \delta\mathsf{Q}^{(k)} \longrightarrow \mathbf{0} \text{ (kada } k \to \infty) \tag{4.1.9}$$

Dakle, akumulirani produkti $\mathsf{Q}^{[1:k]} = \mathsf{Q}^{(1)}\mathsf{Q}^{(2)}\cdots\mathsf{Q}^{(k)}$ s $k \to \infty$ do na skaliranje unitarnom dijagonalnom matricom konvergiraju k Q. Uočimo da matrica Q transformira A u Schurovu formu, $\mathsf{Q}^*\mathsf{A}\mathsf{Q} = \mathsf{R}\Lambda\mathsf{R}^{-1}$. Sada koristeći drugu tvrdnju Teorema 4.1.3 i relaciju (4.1.9) imamo da je

$$\begin{aligned}(\Phi^{(k)})^*\mathsf{A}^{(k+1)}\Phi^{(k)} &= (\Phi^{(k)})^*(\mathsf{Q}^{[1:k]})^*\mathsf{A}\mathsf{Q}^{[1:k]}\Phi^{(k)} = (\mathsf{Q}+\delta\mathsf{Q}^{(k)})^*\mathsf{A}(\mathsf{Q}+\delta\mathsf{Q}^{(k)}) \\ &= \mathsf{Q}^*\mathsf{A}\mathsf{Q} + \underbrace{(\delta\mathsf{Q}^{(k)})^*\mathsf{A}\mathsf{Q} + \mathsf{Q}^*\mathsf{A}\delta\mathsf{Q}^{(k)} + (\delta\mathsf{Q}^{(k)})^*\mathsf{A}\delta\mathsf{Q}^{(k)}}_{\longrightarrow\mathbf{0}} \\ &\longrightarrow \mathsf{R}\Lambda\mathsf{R}^{-1} = \begin{pmatrix} \lambda_1 & * & \cdots & \cdots & * \\ 0 & \lambda_2 & * & \cdots & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot \\ \vdots & \vdots & \ddots & \lambda_{n-1} & * \\ 0 & 0 & \cdots & 0 & \lambda_n \end{pmatrix}.\end{aligned}$$

Sada ćemo maknuti dodatnu pretpostavku o postojanju LU faktorizacije $\mathsf{S}^{-1} = \mathsf{L}\mathsf{T}$. Primijetimo da je taj uvjet došao isključivo kao posljedica načina dokazivanja. Ako želimo slijediti isti dokaz, onda postojanje LU faktorizacije možemo osigurati ako dozvolimo permutacije redaka, tj. uvijek možemo odrediti permutacijsku matricu P tako da $P\mathsf{S}^{-1}$ ima LU faktorizaciju, napišimo je opet $P\mathsf{S}^{-1} = \mathsf{L}\mathsf{T}$. Iz $\mathsf{A} = \mathsf{S}\Lambda\mathsf{S}^{-1} = \mathsf{S}P^{\mathrm{T}}(P\Lambda P^{\mathrm{T}})P\mathsf{S}^{-1}$ vidimo da se postojanje LU faktorizacije može osigurati jednostavnom ponovnom numeracijom svojstvenih vrijednosti i pripadnih svojstvenih vektora. Stavimo $\hat{\Lambda} = P\Lambda P^{\mathrm{T}}$, $\hat{\mathsf{S}} = \mathsf{S}P^{\mathrm{T}}$ i pripadne faktorizacije označimo s $\hat{\mathsf{S}} = \hat{\mathsf{Q}}\hat{\mathsf{R}}$, $\hat{\mathsf{S}}^{-1} = \hat{\mathsf{L}}\hat{\mathsf{T}}$. Sada (4.1.6) postaje

$$\mathsf{A}^k = \hat{\mathsf{S}}\hat{\Lambda}^k\hat{\mathsf{S}}^{-1} = \hat{\mathsf{Q}}\hat{\mathsf{R}}\hat{\Lambda}^k\hat{\mathsf{L}}\hat{\mathsf{T}} = \hat{\mathsf{Q}}\hat{\mathsf{R}}(\hat{\Lambda}^k\hat{\mathsf{L}}\hat{\Lambda}^{-k})\hat{\Lambda}^k\hat{\mathsf{T}}. \tag{4.1.10}$$

s time da ne možemo zaključivati jednostavno kao u (4.1.7), jer dijagonalni elementi u $\hat{\Lambda}$ više nisu nužno u padajućem poretku po apsolutnim vrijednostima. Trebamo se dodatno potruditi – jedna mogućnost je u izboru permutacije P. Ideju ćemo ilustrirati u dimenziji $n = 5$. Zamislimo proces Gaussovih eliminacija na matrici S^{-1} i prvi moment kada je pivotni element eliminacije jednak nuli, npr.

$$\begin{pmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & 0 & \spadesuit & \spadesuit & \spadesuit \\ 0 & 0 & \diamondsuit & \diamondsuit & \diamondsuit \\ 0 & \clubsuit & \clubsuit & \clubsuit & \clubsuit \\ 0 & \heartsuit & \heartsuit & \heartsuit & \heartsuit \end{pmatrix}, \quad \begin{array}{l}\text{pivotni element } (2,2) \text{ jednak nuli} \\ \text{element na } (4,2) \text{ različit od nule}\end{array}$$

Sada retke permutiramo na sljedeći način:

$$\begin{pmatrix} \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 0 & \clubsuit & \clubsuit & \clubsuit & \clubsuit \\ 0 & 0 & \spadesuit & \spadesuit & \spadesuit \\ 0 & 0 & \diamondsuit & \diamondsuit & \diamondsuit \\ 0 & \heartsuit & \heartsuit & \heartsuit & \heartsuit \end{pmatrix}, \text{ permutacija } \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 4 & 2 & 3 & 5 \end{pmatrix}$$

Općenito, ako je u j–tom koraku pivotni element u Gaussovim eliminacijama (pozicija (j,j)) jednak nuli, tražimo prvi od elemenata na pozicijama $(j+1,j),(j+2,j),\ldots(n,j)$ koji je različit od nule. Takav sigurno postoji jer je S^{-1} regularna matrica; neka je njegova pozicija (j_*,j). Sada permutaciju redaka $i \mapsto p(i)$ definiramo tako da djeluje samo na retke s indeksima $j, j+1, \ldots, j_*$ i to kao *cirkularni pomak (circular shift)*

$$p \;:\; \begin{pmatrix} 1 & \ldots & j-1 & j & j+1 & j+2 & \ldots & j_*-1 & j_* & j_*+1 & \ldots & n \\ 1 & \ldots & j-1 & j_* & j & j+1 & \ldots & j_*-2 & j_*-1 & j_*+1 & \ldots & n \end{pmatrix}$$

Ako nakon ove zamjene redaka pogledamo korak Gaussovih eliminacija, odmah je jasno da je $\hat{\mathsf{L}}_{j+1,j} = \hat{\mathsf{L}}_{j+2,j} = \ldots = \hat{\mathsf{L}}_{j_*,j} = 0$. Nadalje, nakon primjene permutacije p na svojstvene vrijednosti, poredak (4.1.5) je narušen samo unutar dijela indeksa, naime

$$|\lambda_1| > \cdots > |\lambda_{j-1}| > \underbrace{|\lambda_{p(j)}| < |\lambda_{p(j+1)}| > |\lambda_{p(j+2)}| > \cdots > |\lambda_{p(j_*)}|}_{\text{promjena uređaja}} > \cdots > |\lambda_{p(n)}|$$

dok je stroga monotonost ostala sačuvana u preostalih $n-j$ vrijednosti,

$$|\lambda_{p(j+1)}| > |\lambda_{p(j+2)}| > \cdots > |\lambda_{p(j_*)}| > \cdots > |\lambda_{p(n)}|.$$

Sada, analogno relaciji (4.1.7) promatramo

$$(\hat{\Lambda}^k \hat{\mathsf{L}} \hat{\Lambda}^{-k})_{ij} = \hat{\mathsf{L}}_{ij} \left(\frac{\lambda_{p(i)}}{\lambda_{p(j)}} \right)^k, \; i = j+1, \ldots, n. \tag{4.1.11}$$

Pri tome odmah uočavamo da za $i = j+1, \ldots, j_*$ imamo nepovoljnu nejednakost $|\lambda_{p(i)}| > |\lambda_{p(j)}|$, ali stvar spašava činjenica da je za te indekse $\hat{\mathsf{L}}_{ij} = 0$. Za preostale indekse $i = j_*+1, \ldots, n$ je $|\lambda_{p(i)}| > |\lambda_{p(j)}|$ pa zaključujemo kao u (4.1.7), pri čemu se argumentacija ne mijenja ako se svojstvene vrijednosti $\lambda_{j_*+1}, \ldots, \lambda_n$ naknadno

permutira (da bi se, kao u ovom j–tom koraku, pronašao netrivijalan pivotni element.) Dakle, na isti način možemo obraditi i sljedeću zamjenu redaka u nekom koraku $j_1 > j$ i pokazati da $(\hat{\Lambda}^k \hat{\mathsf{L}} \hat{\Lambda}^{-k})_{ij_1} \to 0$, $i = j_1 + 1, \ldots n$, za $k \to \infty$.

Zaključujemo da je $\hat{\Lambda}^k \hat{\mathsf{L}} \hat{\Lambda}^{-k} = \mathsf{I} + \hat{E}_k$, s $\hat{E}_k \to \mathbf{0}$ za $k \to \infty$. To znači da dokaz možemo završiti na isti način kao i ranije, s time da u limesu imamo na dijagonali trokutaste forme svojstvene vrijednosti u poretku koji je određen opisanim permutacijama redaka. ⊞

Sljedeći korak u studiranju konvergencije je da smanjimo polazne zahtjeve, tj. da "olabavimo" pretpostavku (4.1.5). Konkretno, neka bude dozvoljeno da A ima nekoliko svojstvenih vrijednosti koje su jednake po apsolutnoj vrijednosti, ali da takve svojstvene vrijednosti imaju samo linearne elementarne divizore.

4.1.5.2 QR iteracije s Hessenbergovim matricama

Vidjeli smo da QR iteracije imaju dosta bogatu i elegantnu strukturu i da daju praktični algoritam za računanje Schurove dekompozicije. Ipak, jedan korak iteracija je dosta složen i za praktičnu primjenu algoritma treba napraviti netrivijalne modifikacije. Nadalje, sjetimo se da smo u prethodnim analizama radili pod pretpostavkom da je matrica A regularna i da je u nekim dokazima ta regularnost bila bitan element dokaza. Svakako je poželjno imati algoritam, sa pripadnom analizom, bez obzira na to da li je nula svojstvena vrijednost ili ne.

Primijenit ćemo strategiju koja je česta u numeričkoj linearnoj algebri: Inicijalno matricu A dozvoljenom klasom transformacija transformiramo u matricu H na kojoj će QR iteracije biti izvedene puno efikasnije i koja će omogućiti jednostavno uklanjanje poteškoća vezanih za singularnost polazne matrice. Kako nam je cilj Schurova forma $\mathsf{A} \mapsto \mathsf{T} = \mathsf{U}^*\mathsf{A}\mathsf{U}$, T gornje trokutasta, U unitarna, onda je dozvoljena klasa transformacija unitarna sličnost. Preciznije, ako je $\mathsf{H} = (\mathsf{U}^{(0)})^*\mathsf{A}\mathsf{U}^{(0)}$ unitarna sličnost, te ako je $\mathsf{T} = (\mathsf{U}^{(1)})^*\mathsf{H}\mathsf{U}^{(1)}$ Schurova forma od H, onda $\mathsf{U} = \mathsf{U}^{(0)}\mathsf{U}^{(1)}$ daje Schurovu formu $\mathsf{T} = \mathsf{U}^*\mathsf{A}\mathsf{U}$ matrice A. Pri tome prva transformacija koja računa matricu H treba biti bazirana na transformaciji s konačno koraka, a struktura matrice H mora biti takva da je svaki korak Algoritma 4.1.4 primijenjenog na H puno efikasniji nego s A. Sada ćemo pokazati da je Hessenbergova forma (Sekcija **??**) dobar izbor matrice H.

Propozicija 4.1.6. *Neka je* $\mathsf{H} = \mathsf{Q}\mathsf{R}$ *QR faktorizacija Hessenbergove matrice* H. *Vrijedi:*

- *Matrice* Q *i* $\mathsf{R}\mathsf{Q}$ *su također Hessenbergove.*

- *Ako je* H *strogo Hessenbergova i singularna, onda je* $\mathsf{R}_{nn} = 0$ *i* $\mathsf{R}_{jj} \neq 0$ *za* $j = 1, \ldots, n-1$.
- *Ako je* H *strogo Hessenbergova, onda su* Q *i* R *esencijalno jedinstvene (neovisno o* $\mathrm{rang}(\mathsf{H})$*).*

Dokaz: Dokaz prvo ilustriramo na 5×5 matrici

$$\mathsf{H} = \begin{pmatrix} x & x & x & x & x \\ * & x & x & x & x \\ \mathbf{0} & * & x & x & x \\ \mathbf{0} & \mathbf{0} & * & x & x \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & * & x \end{pmatrix} \quad \text{u kojoj treba poništiti elemente označene s } *.$$

Za poništavanje pozicije $(2,1)$ koristimo Givensovu rotaciju $\mathsf{G}^{(1)}$

$$\begin{pmatrix} c_1 & s_1 & 0 & 0 & 0 \\ -\overline{s_1} & c_1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x & x & x & x & x \\ \circledast & x & x & x & x \\ 0 & * & x & x & x \\ 0 & 0 & * & x & x \\ 0 & 0 & 0 & * & x \end{pmatrix} = \begin{pmatrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & * & x & x & x \\ 0 & 0 & * & x & x \\ 0 & 0 & 0 & * & x \end{pmatrix} = \mathsf{H}^{(1)},$$

$\mathsf{H}^{(1)} = \mathsf{G}^{(1)}\mathsf{H}$, i odmah prelazimo na poništavanje pozicije $(3,2)$ u $\mathsf{H}^{(1)}$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & c_2 & s_2 & 0 & 0 \\ 0 & -\overline{s}_2 & c_2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & \circledast & x & x & x \\ 0 & 0 & * & x & x \\ 0 & 0 & 0 & * & x \end{pmatrix} = \begin{pmatrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & * & x & x \\ 0 & 0 & 0 & * & x \end{pmatrix} = \mathsf{H}^{(2)},$$

$\mathsf{H}^{(2)} = \mathsf{G}^{(2)}\mathsf{H}^{(1)}$. Dalje, računamo $\mathsf{H}^{(3)} = \mathsf{G}^{(3)}\mathsf{H}^{(2)}$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & c_3 & s_3 & 0 \\ 0 & 0 & -\overline{s}_3 & c_3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & \circledast & x & x \\ 0 & 0 & 0 & * & x \end{pmatrix} = \begin{pmatrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & 0 & x & x \\ 0 & 0 & 0 & * & x \end{pmatrix} = \mathsf{H}^{(3)},$$

i konačno $\mathsf{H}^{(4)} = \mathsf{G}^{(4)}\mathsf{G}^{(3)}\mathsf{G}^{(2)}\mathsf{G}^{(1)}\mathsf{H} \equiv \mathsf{R}$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & c_4 & s_4 \\ 0 & 0 & 0 & -\overline{s}_4 & c_4 \end{pmatrix} \begin{pmatrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & 0 & x & x \\ 0 & 0 & 0 & \circledast & x \end{pmatrix} = \begin{pmatrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & 0 & x & x \\ 0 & 0 & 0 & 0 & x \end{pmatrix} = \mathsf{H}^{(4)}.$$

Općenito trebamo $n-1$ rotaciju i QR faktorizacija je oblika

$$\mathsf{H} = \underbrace{(\mathsf{G}^{(1)})^*(\mathsf{G}^{(2)})^*\cdots(\mathsf{G}^{(n-2)})^*(\mathsf{G}^{(n-1)})^*}_{\mathsf{Q}}\mathsf{R}. \tag{4.1.12}$$

Dalje, u našem 5×5 primjeru je

$$\begin{aligned}\mathsf{Q} &= \begin{pmatrix} c_1 & s_1 & 0 & 0 & 0\\ -\overline{s}_1 & c_1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 1\end{pmatrix}^* \begin{pmatrix} 1 & 0 & 0 & 0 & 0\\ 0 & c_2 & s_2 & 0 & 0\\ 0 & -\overline{s}_2 & c_2 & 0 & 0\\ 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 1\end{pmatrix}^* \begin{pmatrix} 1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & c_3 & s_3 & 0\\ 0 & 0 & -\overline{s}_3 & c_3 & 0\\ 0 & 0 & 0 & 0 & 1\end{pmatrix}^* \begin{pmatrix} 1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & c_4 & s_4\\ 0 & 0 & 0 & -\overline{s}_4 & c_4\end{pmatrix}^*\\ &= \begin{pmatrix} \overline{c}_1 & -s_1\overline{c}_2 & s_1s_2\overline{c}_3 & -s_1s_2s_3\overline{c}_4 & s_1s_2s_3s_4\\ \overline{s}_1 & \overline{c}_1\overline{c}_2 & -s_2\overline{c}_1\overline{c}_3 & s_2s_3\overline{c}_1\overline{c}_4 & -s_2s_3s_4\overline{c}_1\\ 0 & \overline{s}_2 & \overline{c}_2\overline{c}_3 & -s_3\overline{c}_2\overline{c}_4 & s_3s_4\overline{c}_2\\ 0 & 0 & \overline{s}_3 & \overline{c}_3\overline{c}_4 & -s_4\overline{c}_3\\ 0 & 0 & 0 & \overline{s}_4 & \overline{c}_4\end{pmatrix},\end{aligned} \tag{4.1.13}$$

i lako dokažemo da je za svaki $n>2$ matrica Q Hessenbergova.[1] I na kraju, lako se provjeri da je produkt RQ gornje trokutaste i Hessenbergove matrice nužno Hessenbergova matrica.

U slučju strogo Hessenbergove matrice se lako vidi da mora biti $\mathsf{R}_{jj}\neq 0$ za $j=1,\ldots,n-1$. Ako je matrica još i singularna, onda je nužno $\mathsf{R}_{nn}=0$.

Iz prethodno dokazanih tvrdnji znamo da su prvih $n-1$ stupaca u H linearno neovisni, pa teorem o jedinstvenosti QR faktorizacije jedinstveno (do na množenje brojevima modula jedan) određuje prvih $n-1$ stupaca matrice Q. Kako je Q unitarna, onda njen n–ti stupac živi u jednodimenzionalnom potprostoru – ortogonalnom komplementu linearne ljuske prvih $n-1$ stupaca – pa je određen do na množenje skalarom modula jedan. ⊞

Korolar 4.1.7. *Ako QR iteracije* $\mathsf{H}^{(k)}=\mathsf{Q}^{(k)}\mathsf{R}^{(k)}$; $\mathsf{H}^{(k+1)}=\mathsf{R}^{(k)}\mathsf{Q}^{(k)}$ *primijenimo na Hessenbergovu matricu* H*, onda su sve matrice* $\mathsf{H}^{(k)}$, $\mathsf{Q}^{(k)}$ *Hessenbergove.*

Sada ponovo pogledajmo složenost. Inicijalna redukcija matrice A na Hessenbergovu formu H zahtijeva $O(n^3)$ operacija. Kako QR iteracije čuvaju Hessenbergovu formu, svaka QR faktorizacija $\mathsf{H}^{(k)}=\mathsf{Q}^{(k)}\mathsf{R}^{(k)}$ se računa sa $O(n^2)$ operacija i to je bitno ubrzanje jer je QR faktorizacija $\mathsf{A}^{(k)}=\mathsf{Q}^{(k)}\mathsf{R}^{(k)}$ općenite kvadratne matrice proces sa $O(n^3)$ operacija.

[1] Lako se izvedu općenite formule za elemente matrice Q ali nam to sada nije potrebno, dovoljan je zaključak da je Q Hessenbergova.

Nadalje, kako je $\mathsf{Q}^{(k)} = (\mathsf{G}^{(k,1)})^*(\mathsf{G}^{(k,2)})^* \cdots (\mathsf{G}^{(k,n-2)})^*(\mathsf{G}^{(k,n-1)})^*$ produkt od $n-1$ Givensovih rotacija i svaku se može primijeniti s $O(n)$ operacija, onda

$$\mathsf{H}^{(k+1)} = \mathsf{R}^{(k)}(\mathsf{G}^{(k,1)})^*(\mathsf{G}^{(k,2)})^* \cdots (\mathsf{G}^{(k,n-2)})^*(\mathsf{G}^{(k,n-1)})^* \tag{4.1.14}$$

pokazuje da je prijelaz sa $\mathsf{H}^{(k)}$ na $\mathsf{H}^{(k+1)}$ moguć sa samo $O(n^2)$ aritmetičkih operacija.

To nije kraj priče. Uočimo da je jedan korak $\mathsf{H}^{(k+1)} = (\mathsf{Q}^{(k)})^*\mathsf{H}^{(k)}\mathsf{Q}^{(k)}$ transformacija unitarne sličnosti između dvije Hessenbergove matrice. To možemo shvatiti i kao redukciju na Hessenbergovu formu matrice koja je i sama već Hessenbergova, pri čemu unitarna matrica koja realizira tu redukciju ima i drugo svojstvo – računa QR faktorizaciju matrice $\mathsf{H}^{(k)}$. Pitanje je, da li možemo $\mathsf{H}^{(k+1)}$ dobiti direktno kao $(\mathsf{Q}^{(k)})^*\mathsf{H}^{(k)}\mathsf{Q}^{(k)}$, a ne u dva koraka tipa (4.1.12), (4.1.14).

Odgovor je *da*, možemo, koristeći tehniku koja se zove *bulge chasing* ili *naganjanje kvrge*. Radi se o specijalnoj strategiji primjene transformacija sličnosti pomoću ravninskih rotacija, koju ilustriramo na primjeru dimenzije 5: Uzmimo

$$\mathsf{H} = \begin{pmatrix} x & x & x & x & x \\ x & x & x & x & x \\ \mathbf{0} & x & x & x & x \\ \mathbf{0} & \mathbf{0} & x & x & x \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & x & x \end{pmatrix}, \quad \tilde{Q}_1^* = \begin{pmatrix} c_1 & s_1 & 0 & 0 & 0 \\ -\overline{s_1} & c_1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

gdje je $\tilde{Q}_1$ proizvoljna rotacija. Nakon tranformacije sličnosti

$$\tilde{Q}_1^*\mathsf{H}\tilde{Q}_1 = \begin{pmatrix} x & x & x & x & x \\ x & x & x & x & x \\ + & x & x & x & x \\ \mathbf{0} & \mathbf{0} & x & x & x \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & x & x \end{pmatrix} \quad \text{poziciju } (3,1) \text{ poništavamo pomoću}$$

$$\tilde{Q}_2^* = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & c_2 & s_2 & 0 & 0 \\ 0 & -\overline{s}_2 & c_2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \tilde{Q}_2^*\tilde{Q}_1^*\mathsf{H}\tilde{Q}_1\tilde{Q}_2 = \begin{pmatrix} x & x & x & x & x \\ x & x & x & x & x \\ \mathbf{0} & x & x & x & x \\ \mathbf{0} & + & x & x & x \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & x & x \end{pmatrix};$$

zatim novi "poremećaj" na poziciji $(4,2)$ možemo anulirati s

$$\tilde{Q}_3^* = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & c_3 & s_3 & 0 \\ 0 & 0 & -\overline{s}_3 & c_3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \tilde{Q}_3^*\tilde{Q}_2^*\tilde{Q}_1^*\mathsf{H}\tilde{Q}_1\tilde{Q}_2\tilde{Q}_3 = \begin{pmatrix} x & x & x & x & x \\ x & x & x & x & x \\ \mathbf{0} & x & x & x & x \\ \mathbf{0} & \mathbf{0} & x & x & x \\ \mathbf{0} & \mathbf{0} & + & x & x \end{pmatrix}$$

i konačno element na poziciji $(5,3)$ poništavamo s

$$\tilde{Q}_4^* = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & c_4 & s_4 \\ 0 & 0 & 0 & -\overline{s}_4 & c_4 \end{pmatrix}, \quad \tilde{Q}_4^*\tilde{Q}_3^*\tilde{Q}_2^*\tilde{Q}_1^*\mathsf{H}\tilde{Q}_1\tilde{Q}_2\tilde{Q}_3\tilde{Q}_4 = \begin{pmatrix} x & x & x & x & x \\ x & x & x & x & x \\ \mathbf{0} & x & x & x & x \\ \mathbf{0} & \mathbf{0} & x & x & x \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & x & x \end{pmatrix}.$$

Dakle, prva rotacija je u Hessenbergovoj matrici stvorila izbočinu ili kvrgu na poziciji $(3,1)$ (označenu s $+$) i time pokvarila Hessenbergovu formu. Nakon toga smo nizom rotacija tu kvrgu tjerali uzduž dijagonale, sve dok je nismo stjerali u donji desni kut gdje je nestala.

Ako stavimo $\tilde{Q} = \tilde{Q}_1\tilde{Q}_2\tilde{Q}_3\tilde{Q}_4$, onda je $\tilde{Q}^*\mathsf{H}\tilde{Q}$ ponovo Hessenbergova matrica, a matrica $\tilde{Q}$ ima istu strukturu kao i matrica (4.1.13) dobivena QR faktorizacijom od H. Dakle, imamo dvije unitarne matrice koje računaju Hessenbergovu formu od H. Prema Teoremu 3.8.3 o jedinstvenosti Hessenebergove forme, ako se te dvije matrice trivijalno razlikuju u prvom stupcu (do na množenje kompleksnim brojem modula jedan), onda se one na isti način razlikuju u preostalim stupcima. Dakle, jedan korak QR algoritma možemo izvesti tako da $\tilde{Q}_1$ bude ista ona rotacija koja se koristi pri računanju QR faktorizacije matrice $\mathsf{H}^{(k)}$, a ostatak je samo *bulge chasing* tj. korigiranje Hessenbergove forme nizom rotacija. Teorem o jedinstvenosti garantira da smo dobili isti rezultat (do na trivijalnu sličnost dijagonalnom unitarnom matricom) kao u dva koraka tipa (4.1.12), (4.1.14).

Poglavlje 5

Simetrični problem svojstvenih vrijednosti

U ovom poglavlju proučavamo problem svojstvenih vrijednosti $Ax = \lambda x$ u kojem je matrica A realna i simetrična ($A = A^T$) ili, općenitije, kompleksna hermitska matrica ($A = A^*$). Valja odmah reći da u teorijskom dijelu nema bitne razlike između realnih simetričnih i općenito hermitskih matrica. S druge strane, kod razvoja numeričkih algoritama za računala, gdje je nužno provesti efektivan račun, će biti razlika koje su uglavnom tehničke prirode.

5.1 Mini–max karakterizacija

Problem svojstvenih vrijednosti za Hermitske matrice je puno lakši nego opći problem. Prvi razlog je u činjenici da je Schurova forma hermitske (realne simetrične) matrice realna dijagonalna – dakle matricu se može dijagonalizirati unitarnom (ortogonalnom) transformacijom sličnosti. Drugi razlog je u činjenici da su u Hermitskom slučaju svojstvene vrijednosti jako specijalne funkcije matrice.

Neka je $A = U\Lambda U^* = \sum_{i=1}^{n} \lambda_i u_i u_i^*$ Schurova dekompozicija hermitske matrice A u kojoj su svojstvene vrijednosti numerirane tako da je $\lambda_1 \geq \cdots \geq \lambda_n$, a $u_1, \ldots, u_n$ su stupci unitarne matrice U.

Uzmimo proizvoljan jedinični vektor x, $x^*x = 1$, i neka je $y = U^*x$. Vrijedi

$$x^*Ax = x^*U\Lambda \underbrace{U^*x}_{y} = y^*\Lambda y = \sum_{i=1}^{n} \lambda_i |y_i|^2 \in [\lambda_n, \lambda_1], \text{ jer je } \sum_{i=1}^{n} |y_i|^2 = 1. \quad (5.1.1)$$

Odavde odmah vidimo da se λ_1 i λ_n mogu opisati kao

$$\lambda_n = \min_{x^*x=1} x^*Ax = u_n^*Au_n, \quad \lambda_1 = \max_{x^*x=1} x^*Ax = u_1^*Au_1.$$

Nadalje, iz (5.1.1) vidimo da dodatni uvjet npr. $u_1^*x = 0$ (tj. $x \perp u_1$) daje $y_1 = 0$ pa je

$$x^*Ax = \sum_{i=2}^{n} \lambda_i |y_i|^2 \in [\lambda_n, \lambda_2], \quad \text{dakle} \quad \lambda_2 = \max_{\substack{x^*x=1 \\ x \perp u_1}} x^*Ax.$$

Analogno, uzimanjem $x \perp u_n$ dobijemo $\lambda_{n-1} = \min\limits_{\substack{x^*x=1 \\ x \perp u_n}} x^*Ax$. Sada je jasno da dodavanjem novih uvjeta ortogonalnosti na prethodne svojstvene vektore dobivamo sljedeći teorem:

Teorem 5.1.1. *Neka je $A \in \mathbb{C}^{n\times n}$ hermitska matrica i neka su $\lambda_1 \geq \cdots \geq \lambda_n$ njene svojstvene vrijednosti sa pripadnim svojstvenim vektorima $u_1, \ldots, u_n$; $Au_j = \lambda_j u_j$. Vrijedi*

$$\lambda_k = \max_{\substack{x^*x=1 \\ x \perp u_1, \ldots, u_{k-1}}} x^*Ax = \min_{\substack{x^*x=1 \\ x \perp u_{k+1}, \ldots, u_n}} x^*Ax = u_k^*Au_k.$$

Uočimo da iz (5.1.1) slijedi sljedeći jednostavan zaključak: Ako uzmemo proizvoljan jediniči vektor i izračunamo x^*Ax, onda postoji barem jedna svojstvena vrijednost u intervalu $(-\infty, x^*Ax]$ i barem jedna u $[x^*Ax, \infty)$.

Teorem 5.1.2. *Neka je $A \in \mathbb{C}^{n\times n}$ hermitska matrica i neka su $\lambda_1 \geq \cdots \geq \lambda_n$ njene svojstvene vrijednosti sa pripadnim svojstvenim vektorima $u_1, \ldots, u_n$. U proizvoljnom k–dimenzionalnom potprostoru $\mathcal{S}_k$ postoje jedniničnі vektori x, y sa svojstvom*

$$x^*Ax \leq \lambda_k, \quad y^*Ay \geq \lambda_{n-k+1}.$$

Dokaz: Promotrimo presjek potprostora $\mathcal{S}_k$ i linearne ljuske svojstvenih vektora $u_k, u_{k+1}, \ldots, u_n$, $\mathcal{P} = \mathcal{S}_k \bigcap L(u_k, u_{k+1}, \ldots, u_n)$. Očito je $\dim(\mathcal{P}) \geq 1$ pa možemo odabrati jedinični vektor $x \in \mathcal{P}$, i možemo ga prikazati kao $x = \sum_{j=k}^{n} x_j u_j$. Sada je

$$x^*Ax = \sum_{j=k}^{n} \overline{x_j} u_j^* \sum_{j=k}^{n} x_j \lambda_j u_j = \sum_{j=k}^{n} \lambda_j |x_j|^2 \leq \lambda_k.$$

Druga nejednakost se dokaže analogno. $\boxplus$

Jasno je da u prethodnom teoremu pogodnim odabirom potprostora $\mathcal{S}_k$ možemo postići jednakosti. Time smo došli do važnog teorema o spektru hermitske matrice.

Teorem 5.1.3. *(Courant, Fischer) Neka je $A \in \mathbb{C}^{n\times n}$ hermitska matrica i neka su $\lambda_1 \geq \cdots \geq \lambda_n$ njene svojstvene vrijednosti sa pripadnim svojstvenim vektorima $u_1, \ldots, u_n$; $Au_j = \lambda_j u_j$. Tada za $j = 1, \ldots, n$ vrijedi*

$$\lambda_j = \min_{\mathcal{S}_{j-1}} \max_{\substack{x\perp\mathcal{S}_{j-1}\\ x\neq\mathbf{0}}} \frac{x^*Ax}{x^*x} = \min_{\mathcal{S}_{j-1}} \max_{\substack{x\perp\mathcal{S}_{j-1}\\ x^*x=1}} x^*Ax = u_j^*Au_j \tag{5.1.2}$$

$$= \max_{\mathcal{P}_j} \min_{\substack{x\in\mathcal{P}_j\\ x\neq\mathbf{0}}} \frac{x^*Ax}{x^*x} = \max_{\mathcal{P}_j} \min_{\substack{x\in\mathcal{P}_j\\ x^*x=1}} x^*Ax = u_j^*Au_j, \tag{5.1.3}$$

pričemu $\mathcal{S}_{j-1}$, $\mathcal{P}_j$ označavaju proizvoljne $(j-1)$–, j–dimenzionalne potprostore u $\mathbb{C}^n$ ($\mathbb{R}^n$ u slučaju realne simetrične matrice). Specijalno za dvije ekstremne svojstvene vrijednosti imamo

$$\lambda_1 = \max_{x\neq\mathbf{0}} \frac{x^*Ax}{x^*x} = \max_{x^*x=1} x^*Ax = u_1^*Au_1, \tag{5.1.4}$$

$$\lambda_n = \min_{x\neq\mathbf{0}} \frac{x^*Ax}{x^*x} = \min_{x^*x=1} x^*Ax = u_n^*Au_n. \tag{5.1.5}$$

5.2 Sylvesterov teorem

Definicija 5.2.1. Neka hermitska matrica A ima $i_-(A)$ negativnih i $i_+(A)$ pozitivnih svojstvenih vrijednosti, te neka je $i_0(A)$ kratnost nule kao svojstvene vrijednoti (ako je A singularna). Trojku $i(A) = (i_+(A), i_-(A), i_0(A))$ zovemo inercija matrice A.

Definicija 5.2.2. Za hermitske matrice A i B kažemo da su kongruentne ako postoji regularna matrica S takva da je $B = S^*AS$.

Odmah zaključujemo da je A kongruentna matrici

$$I(A) = \begin{pmatrix} \mathsf{I}_{i_+(A)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathsf{I}_{i_-(A)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Teorem 5.2.1. *(Sylvester) Hermitske matrice A i B su kongruentne ako i samo ako je $i(A) = i(B)$.*

Dokaz:
⊞

5.3 Perturbacije spektra

U primjenama često ne možemo apsolutno točno izračunati matricu A koja prirodno opisuje neki aspekt u konkretnom problemu – umjesto toga nam je dostupna matrica $\tilde{A} = A + \delta A$, gdje δA u sebi sadrži sve nesavršenosti i pogreške koje nam onemogućuju pristup matrici A. Ako rješavamo problem svojstvenih vrijednosti, to znači da do spektralne informacije o matrici A moramo doći posredno, radeći s $\tilde{A}$ i uzimajući u obzir veličinu perturbacije δA. Teorija perturbacija spektra matrice proučava kako se mijenjaju svojstvene vrijednosti i vektori matrice A ako se matrica promijeni u $\tilde{A} = A + \delta A$.

U numeričkom rješavanju problema svojstvenih vrijednosti je teorija perturbacija važan alat jer su algoritmi bazirani na različitim aproksimacijama, dakle sa neizbježnim perturbacijama matrica koje mogu uzrokovati netrivijalne promjene svojstvenih vrijednosti.

Ovdje samo navodimo rezultate koji ćemo koristiti u analizi numeričkih algoritama.

Teorem 5.3.1. *(Weyl) Za hermitske matrice A, B i svojstvene vrijednosti numerirane u padajuće nizove vrijedi*

$$\lambda_j(A+B) \leq \lambda_i(A) + \lambda_{j-i+1}(B), \quad za\ i \leq j \tag{5.3.1}$$

$$\lambda_j(A+B) \geq \lambda_i(A) + \lambda_{j-i+n}(B), \quad za\ i \geq j. \tag{5.3.2}$$

Teorem 5.3.2. *Za hermitske matrice A i B vrijedi*

$$\max_{j=1:n} |\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|_2. \tag{5.3.3}$$

U praksi je lakše raditi sa Frobeniusovom normom $\|\cdot\|_F$ umjesto spektralne $\|\cdot\|_2$. Naravno, u Teoremu 5.3.2 možemo trivijalno iskoristiti nejednakost $\|A - B\|_2 \leq \|A - B\|_F$, a pravi pristup perturbacijama u Frobeniusovoj normi je Hoffman–Wielandtov teorem.

Teorem 5.3.3. *(Hoffman, Wielandt) Ako su A i B normalne $n \times n$ matrice sa svojstvenim vrijednostima $\lambda_1(\mathsf{A}), \ldots, \lambda_n(\mathsf{A})$, $\lambda_1(\mathsf{B}), \ldots, \lambda_n(\mathsf{B})$, onda postoji permutacija p tako da je*

$$\sqrt{\sum_{i=1}^{n} |\lambda_i(\mathsf{A}) - \lambda_{p(i)}(\mathsf{B})|^2} \leq \|\mathsf{A} - \mathsf{B}\|_F.$$

Korolar 5.3.4. *Neka je u Teoremu 5.3.3 matrica* A *hermitska, a* B *normalna, te neka su svojstvene vrijednosti indeksirane tako da je* $\lambda_1(\mathsf{A}) \geq \cdots \geq \lambda_n(\mathsf{A})$*, i* $\mathrm{Re}(\lambda_1(\mathsf{B})) \geq \cdots \geq \mathrm{Re}(\lambda_n(\mathsf{B}))$*. Tada je*

$$\sqrt{\sum_{i=1}^{n} |\lambda_i(\mathsf{A}) - \lambda_i(\mathsf{B})|^2} \leq \|\mathsf{A} - \mathsf{B}\|_F.$$

Prethodni korolar nam kaže da, ako simetričnu matricu A napišemo kao $A = D+E$, $D = \mathrm{diag}(a_{ii})_{i=1}^n$, $E = A-D$, onda sortirani dijagonalni elementi $a_{p(1)p(1)} \geq \cdots \geq a_{p(n)p(n)}$ od A (p permutacija) aproksimiraju sortirane svojstvene vrijednosti $\lambda_1 \geq \cdots \geq \lambda_n$ od A s ukupnom pogreškom

$$\sqrt{\sum_{i=1}^{n} (a_{p(i)p(i)} - \lambda_i)^2} \leq \|E\|_F.$$

Dakle, mali izvandijagonalni elementi znače da dijagonalni elementi dobro aproksimiraju svojstvene vrijednosti. Ako $\|E\|_F$ nije mali broj, onda možemo načiniti sljedeće: pokušajmo odrediti ortogonalnu matricu Q tako da matrica $A' = Q^T AQ = D' + E'$ ima normu izvandijagonalnog dijela E' takvu da je $\|E'\|_F < \|E\|_F$. Ako uspijemo, onda dijagonala od A' (koja ima isti spektar kao i A) daje bolju aproksimaciju svojstvenih vrijednosti od A. Time smo zapravo dali osnovnu ideju za razvoj metoda za dijagonalizaciju simetričnih matrica.

5.4 Jacobijeva metoda

Opisat ćemo klasičnu Jacobijevu metodu. Ideja metode je da počevši s $A^{(1)} \equiv A$ generiramo niz ortogonalno sličnih matrica $A^{(2)} = (U^{(1)})^T A^{(1)} U^{(1)}$, $A^{(3)} = (U^{(2)})^T A^{(2)} U^{(2)}$, tj.

$$A^{(k+1)} = (U^{(k)})^T A^{(k)} U^{(k)} = (U^{(1)} U^{(2)} \cdots U^{(k)})^T A (U^{(1)} U^{(2)} \cdots U^{(k)}), k = 1, 2, \ldots \tag{5.4.1}$$

i to tako da matrice $A^{(k)}$ konvergiraju fiksnoj dijagonalnoj matrici kada $k \to \infty$. Pri tome matrice $U^{(k)}$ trebaju biti jednostavne ortogonalne matrice koje u algoritmu možemo jednostavno odrediti i primijeniti u transformaciji sličnosti.

5.4.1 Jacobijeva rotacija

Pogledajmo prvo kako lako možemo dijagonalizirati 2×2 realnu simetričnu matricu

$$M = \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix},\ \gamma \neq 0.$$

Uzet ćemo ravninsku rotaciju

$$J = \begin{pmatrix} \cos\vartheta & \sin\vartheta \\ -\sin\vartheta & \cos\vartheta \end{pmatrix}$$

i odrediti ϑ tako da je $J^T M J$ dijagonalna matrica. Lako izračunamo da je

$$J^T M J = \begin{pmatrix} \alpha\cos^2\vartheta - 2\gamma\sin\vartheta\cos\vartheta & \gamma(\cos^2\vartheta - \sin^2\vartheta) + (\alpha-\beta)\sin\vartheta\cos\vartheta \\ \gamma(\cos^2\vartheta - \sin^2\vartheta) + (\alpha-\beta)\sin\vartheta\cos\vartheta & \alpha\sin^2\vartheta + 2\gamma\sin\vartheta\cos\vartheta + \beta\cos^2\vartheta \end{pmatrix}$$

pa uvjet dijagonalnosti

$$\gamma \underbrace{(\cos^2\vartheta - \sin^2\vartheta)}_{\cos 2\vartheta} + (\alpha-\beta)\underbrace{\sin\vartheta\cos\vartheta}_{\frac{1}{2}\sin 2\vartheta} = 0$$

daje

$$\cot 2\vartheta = \frac{\beta-\alpha}{2\gamma}. \tag{5.4.2}$$

Ako stavimo $\zeta = \cot 2\vartheta$ i $\tau = \tan\vartheta$ onda je $\zeta = (1-\tau^2)/(2\tau)$ pa τ možemo dobiti kao rješenje kvadratne jednadžbe

$$\tau^2 + 2\zeta\tau - 1 = 0.$$

Manje po modulu rješenje je

$$\tau \equiv \tan\vartheta = \frac{\operatorname{sign}(\zeta)}{|\zeta| + \sqrt{1+\zeta^2}}, \tag{5.4.3}$$

pa lako odredimo parametre rotacije

$$\cos\vartheta = \frac{1}{\sqrt{1+\tau^2}},\ \sin\vartheta = \tau\cos\vartheta. \tag{5.4.4}$$

Elementarnom trigonometrijom se lako provjeri da je

$$J^T M J = \begin{pmatrix} \alpha - \gamma\tau & 0 \\ 0 & \beta + \gamma\tau \end{pmatrix}. \tag{5.4.5}$$

Dakle, Jacobijeva rotacija J je matrica svojstvenih vektora matrice M i može se, osim Jacobijevim formulama (5.4.2, 5.4.2, 5.4.4), dobiti i na nekoliko drugih efikasnih načina.

Primijetimo i da je, zbog ortogonalnosti,

$$(\alpha - \gamma\tau)^2 + (\beta + \gamma\tau)^2 = \|J^T M J\|_F^2 = \|M\|_F^2 = \alpha^2 + \beta^2 + 2\gamma^2. \tag{5.4.6}$$

5.4.2 Klasična Jacobijeva metoda

Sada ćemo opisanu operaciju iterativno primjenjivati na 2×2 podmatrice $n \times n$ matrica $A^{(k)}$ u nizu (5.4.1). U k–tom koraku u matrici $A^{(k)}$ uočimo 2×2 podmatricu sa elementima na presjecima redaka i stupaca s indeksima i_k, j_k,

$$\hat{A}^{(k)} = \begin{pmatrix} a^{(k)}_{i_k,i_k} & a^{(k)}_{i_k,j_k} \\ a^{(k)}_{j_k,i_k} & a^{(k)}_{j_k,j_k} \end{pmatrix} = \begin{pmatrix} \boxplus & \boxdot \\ \boxdot & \boxtimes \end{pmatrix}, \; A^{(k)} = \begin{pmatrix} * & * & * & * & * & * & * & * \\ * & * & * & * & * & * & * & * \\ * & * & \boxplus & * & * & \boxdot & * & * \\ * & * & * & * & * & * & * & * \\ * & * & * & * & * & * & * & * \\ * & * & \boxdot & * & * & \boxtimes & * & * \\ * & * & * & * & * & * & * & * \\ * & * & * & * & * & * & * & * \end{pmatrix}.$$

Za $\hat{A}^{(k)}$ odredimo Jacobijevu rotaciju $\hat{U}^{(k)} = \left(\begin{smallmatrix} \cos\vartheta_k & \sin\vartheta_k \\ -\sin\vartheta_k & \cos\vartheta_k \end{smallmatrix}\right)$, koristeći formule (5.4.2, 5.4.3, 5.4.4). (Naravno, ako je $a^{(k)}_{i_k,j_k} = 0$, $\hat{U}^{(k)} = I_2$.) Sada $\hat{U}^{(k)}$ uložimo u $n \times n$ jediničnu matricu tako da u $n \times n$ shemi $\hat{U}^{(k)}$ zauzima iste pozicije kao i $\hat{A}^{(k)}$ u $A^{(k)}$, tj. presjeke redaka i stupaca s indeksima i_k, j_k. Dobijemo ortogonalnu matricu

$$U^{(k)} \equiv U^{(k)}(i_k, j_k, \vartheta_k) = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos\vartheta_k & & & & \sin\vartheta_k & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & -\sin\vartheta_k & & & & \cos\vartheta_k & \\ & & & & & & & & 1 \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix}$$

i računamo $A^{(k+1)} = (U^{(k)})^T A^{(k)} U^{(k)}$. Jasno je da je $a^{(k+1)}_{i_k,j_k} = a^{(k+1)}_{j_k,i_k} = 0$.

Dakle, naučili smo kako u bilo kojoj simetričnoj matrici jednostavnom transformacijom ortogonalne sličnosti (koja mijenja samo dva retka i dva stupca) možemo poništiti dva izvandijagonalna elementa. Sada to sustavno primjenjujemo za $k = 1, 2, 3, \ldots$ Naravno, dvije nule koje smo kreirali u k–tom koraku se mogu pretvoriti u netrivijalne elemente već u sljedećem koraku. Ostaje nam dakle odrediti strategiju kojom ćemo u danom koraku k birati tzv. pivotnu poziciju (i_k, j_k) tako da

u limesu osiguramo konvergenciju niza $A^{(k)}$ dijagonalnoj matrici. U praktičnom računanju se zaustavljamo na indeksu k_* za kojeg je $A^{(k_*)}$ *dovoljno blizu dijagonalnoj matrici.*

Da bismo sve ovo operacionalizirali uvodimo funkciju

$$\Omega(A) = \|A - \mathrm{diag}(A)\|_F = \sqrt{\sum_{i\neq j} |a_{ij}|^2}.$$

Očito, $\Omega(A)$ mjeri koliko je velik izvandijagonalni dio od A, $\Omega(A) = 0$ ako i samo ako je A dijagonalna matrica. Dakle, naš izbor pivotnih elemenata (pivotna strategija) mora osigurati $\lim_{k\to\infty} \Omega(A^{(k)}) = 0$.

Propozicija 5.4.1. *U jednoj primjeni Jacobijeve rotacije* $A^{(k+1)} = (U^{(k)})^T A^{(k)} U^{(k)}$ *sa* $U^{(k)} \equiv U^{(k)}(i_k, j_k, \vartheta_k)$ *je* $\Omega^2(A^{(k+1)}) = \Omega^2(A^{(k)}) - 2(a^{(k)}_{i_k,j_k})^2$.

Dokaz: Prvo uočimo da je $\|A^{(k+1)}\|_F^2 = \Omega^2(A^{(k+1)}) + \sum_{\ell=1}^n (a^{(k+1)}_{\ell\ell})^2$, te da transformacija $A^{(k+1)} = (U^{(k)})^T A^{(k)} U^{(k)}$ mijenja samo dva dijagonalna elementa, (i_k, i_k)–ti i (j_k, j_k)–ti, pri čemu je (vidi (5.4.6))

$$(a^{(k+1)}_{i_k,i_k})^2 + (a^{(k+1)}_{j_k,j_k})^2 = (a^{(k)}_{i_k,i_k})^2 + (a^{(k)}_{j_k,j_k})^2 + 2(a^{(k)}_{i_k,j_k})^2.$$

Odavde je

$$\begin{aligned}\Omega^2(A^{(k+1)}) &= \|A^{(k)}\|_F^2 - \sum_{\ell=1,\ell\neq i_k,j_k}^{n} (a^{(k)}_{\ell\ell})^2 - ((a^{(k)}_{i_k,i_k})^2 + (a^{(k)}_{j_k,j_k})^2 + 2(a^{(k)}_{i_k,j_k})^2) \\ &= \Omega^2(A^{(k)}) - 2(a^{(k)}_{i_k,j_k})^2.\end{aligned}$$

⊞

Propozicija 5.4.2. *Ako je u svakom koraku k pivotna pozicija* (i_k, j_k) *odabrana tako da je*

$$|a^{(k)}_{i_k,j_k}| = \max_{i\neq j} |a^{(k)}_{ij}| \tag{5.4.7}$$

onda je

$$\Omega(A^{(k+1)}) \leq \Omega(A^{(k)}) \sqrt{1 - \frac{2}{n(n-1)}} \tag{5.4.8}$$

pa je $\lim_{k\to\infty} \Omega(A^{(k)}) = 0$.

Dokaz: Iz (5.4.7) je $\Omega(A^{(k)}) \leq n(n-1)(a^{(k)}_{i_k,j_k})^2$ pa je

$$\Omega^2(A^{(k+1)}) \leq \Omega^2(A^{(k)}) - 2\frac{\Omega^2(A^{(k)})}{n(n-1)} = \Omega^2(A^{(k)})(1 - \frac{2}{n(n-1)}).$$

$\boxplus$

Dakle, ako zadamo toleranciju $\epsilon > 0$, postoji indeks k_* sa svojstvom da je za sve $k \geq k_*$ $\Omega(A^{(k)}) < \epsilon$, i algoritam staje na matrici $A^{(k_*)}$. Ako matricu $A^{(k_*)}$ napišemo kao sumu njenog dijagonalnog dijela $D^{(k_*)}$ i izvandijagonalnih elemenata $E^{(k_*)}$, $A^{(k_*)} = D^{(k_*)} + E^{(k_*)}$, te ako stavimo $\tilde{U} = U^{(1)}U^{(2)}\cdots U^{(k_*-1)}$ onda o kvaliteti približne dijagonalizacije $\tilde{U}^T A\tilde{U} \approx D^{(k_*)}$ možemo suditi na nekoliko načina:

• Iz $\tilde{U}^T A\tilde{U} = A^{(k_*)} = D^{(k_*)} + E^{(k_*)}$ je

$$\tilde{U}^T(A - \underbrace{\tilde{U}E^{(k_*)}\tilde{U}^T}_{\delta A})\tilde{U} = D^{(k_*)} \tag{5.4.9}$$

tj. dijagonalni elementi matrice $A^{(k_*)}$ su egzaktne svojstvene vrijednosti matrice $A - \delta A$, a stupci ortogonalne matrice $\tilde{U}$ su pripadni svojstveni vektori. Dakle konačan niz transformacija Jacobijeve metode bi dao egzaktnu dijagonalizaciju matrice $A-\delta A$. Ako je δA dovoljno mala matrica, tako mala da je $A-\delta A \approx A$ onda prihvaćamo $\tilde{U}^T A\tilde{U} \approx D^{(k_*)}$. Primijetimo da je $\|\delta A\|_F = \|E^{(k_*)}\|_F = \Omega(A^{(k_*)}) < \epsilon$, pa je dobar izbor za ϵ reda veličine $\varepsilon\|A\|_F$, gdje je ε strojna točnost.

• Relaciju $\tilde{U}^T A\tilde{U} = A^{(k_*)} = D^{(k_*)} + E^{(k_*)}$ možemo pročitati i ovako: Za svaki indeks i je $A\tilde{U}(:,i) = (D^{(k_*)})_{ii}\tilde{U}(:,i) + \tilde{U}E^{(k_*)}(:,i)$, tj.

$$\|A\tilde{U}(:,i) - (D^{(k_*)})_{ii}\tilde{U}(:,i)\|_2 = \|E^{(k_*)}(:,i)\|_2 < \epsilon, \tag{5.4.10}$$

pa smatramo da je približno $A\tilde{U}(:,i) \approx (D^{(k_*)})_{ii}\tilde{U}(:,i)$.

• U prethodne dvije točke smo pokušali dati smisao izračunatim aproksimacijama. Ponekad je to dovoljno. Ipak, uočimo da u tim diskusijama nismo uopće spominjali svojstvene vrijednosti $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ matrice A. Što možemo reći o nepoznatim vrijednostima λ_i koje su (uz pripadne svojstvene vektore) zapravo ciljevi računanja? Koliko dobro dijagonala od $A^{(k_*)}$ aproksimira spektar od A? Odgovore daje teorija perturbacija. Na primjer, Hoffman–Wielandtov teorem daje da svojstvene vrijednosti matrice A i silazno sortirani dijagonalni elementi $a^{(k_*)}_{p(1)p(1)} \geq a^{(k_*)}_{p(2)p(2)} \geq \cdots \geq a^{(k_*)}_{p(n)p(n)}$ zadovoljavaju

$$\sqrt{\sum_{i=1}^{n}(\lambda_i - a^{(k_*)}_{p(i)p(i)})^2} \leq \Omega(A^{(k_*)}) < \epsilon.$$

Strogo gledano, ostalo je još nekoliko koraka da bi se zaključilo pitanje konvergencije i otvorio prostor za dublju analizu metode. Treba pokazati da je $\lim_{k\to\infty} A^{(k)}$ fiksna dijagonalna matrica i da je $\lim_{k\to\infty} U^{(1)}\cdots U^{(k)}$ ortogonalna matrica svojstvenih vektora matrice A. Pri tome valja naglasiti da je konvergencija svojstvenih vektora netrivijalna u slučaju višestrukih svojstvenih vrijednosti i da se relativno jednostavan dokaz može dobiti kada su sve svojstvene vrijednosti jednostruke.

Teorem 5.4.3. *U klasičnoj Jacobijevoj metodi je limes niza $A^{(k)}$ dijagonalna matrica sa svojstvenim vrijednostima na dijagonali,*

$$\lim_{k\to\infty} A^{(k)} = \begin{pmatrix} \lambda_{\pi(1)} & & \\ & \ddots & \\ & & \lambda_{\pi(n)} \end{pmatrix}$$

pri čemu je $\pi(\cdot)$ neka permutacija. Ako su sve svojstvene vrijednosti jednostruke onda je beskonačni produkt $\lim_{k\to\infty} U^{(1)}\cdots U^{(k)}$ konvergentan i njegov limes je ortogonalna matrica pripadnih svojstvenih vektora.

5.4.2.1 Implementacija

Sada je na redu diskusija o implementaciji u aritmetici konačne preciznosti, gdje osim pogreške koja je nastala uzimanjem konačnog dijela konvergentnog niza u igru ulaze pogreške zaokruživanja u svakoj operaciji koju smo opisali. Jasno je da u aritmetici konačne preciznosti Jacobijevu transformaciju ne možemo odrediti egzaktno, niti je možemo egzaktno primijeniti. U bilo kojem momentu umjesto $A^{(k)}$ imamo $\tilde{A}^{(k)} = A^{(k)}+\delta A^{(k)}$; čak kada bismo krenuli od $\tilde{A}^{(k)}$ kao polazne matrice, niti za nju ne bismo mogli egzaktno odrediti Jacobijevu rotaciju pa bismo umjesto $U^{(k)}$ imali $\tilde{U}^{(k)} = U^{(k)}+\delta U^{(k)}$; na kraju transformaciju sličnosti koristeći $\tilde{U}^{(k)}$ ne možemo izvršiti bez grešaka zaokruživanja $F^{(k)}$. Ukupno, umjesto $A^{(k+1)} = (U^{(k)})^T A^{(k)} U^{(k)}$ imamo

$$\begin{aligned} \tilde{A}^{(k+1)} &= A^{(k+1)} + \delta A^{(k+1)} = (\tilde{U}^{(k)})^T \tilde{A}^{(k)} \tilde{U}^{(k)} + F^{(k)} \\ &= (U^{(k)} + \delta U^{(k)})^T (A^{(k)} + \delta A^{(k)})(U^{(k)} + \delta U^{(k)}) + F^{(k)} \end{aligned} \quad (5.4.11)$$

Odavde odmah slijedi:

- Nema garancije da su nakon transformacije u matrici $\tilde{A}^{(k+1)}$ pivotne pozicije (i_k, j_k), (j_k, i_k) jednake nuli – umjesto toga očekujemo male brojeve. Zapravo nema niti garancije da je $\tilde{A}^{(k+1)}$ simetrična! U programu obično pivotne

pozicije nakon transformacije stavimo eksplicitno na nulu a dva pripadna dijagonalna elementa izračunamo koristeći (5.4.5). (Strogo gledano, zbog pogrešaka (5.4.11) relacija (5.4.5) više ne vrijedi.))

- Uzimajući u obzir prethodne diskusije, razumno je i praktično preskočiti k–ti korak i staviti $\tilde{A}^{(k+1)} = \tilde{A}^{(k)}$ ako je $|\tilde{a}^{(k)}_{i_k,j_k}|$ dovoljno mali broj. Možemo čak pivotne pozicije eksplicitno staviti na nulu.

Vrijednost $\Omega(A^{(k)})$ se u jednom koraku $k \rightsquigarrow k+1$ mijenja jednostavnom formulom iz Propozicije 5.4.1: inicijalno stavimo $\omega_1 = \Omega^2(A)$ i u svakom koraku k je $\omega_{k+1} = \omega_k - 2(a^{(k)}_{i_k,j_k})^2$ suma kvadrata izvandijagonalnih elemenata matrice u novoj iteraciji. Na žalost, ova formula u praktičnom računanju nije pouzdana i to iz barem dva razloga:

- U aritmetici konačne preciznosti Jacobijevu transformaciju ne možemo odrediti egzaktno, niti je možemo egzaktno primijeniti. To znači da u konkretnoj situaciji na računalu formula prije svega nije istinita.

- Uzastopna oduzimanja kojima se dobiva nova vrijednost za ω_k uzrokuju gubitak točnih znamenki tako da se lako može desiti da u nekom momentu izračunamo čak i negativni ω_k.

S obzirom na izbor pivotne pozicije $a^{(k)}_{i_k,j_k}$ prema formuli (5.4.7), te da je prema tome $\Omega(A^{(k)}) \leq n(n-1)|a^{(k)}_{i_k,j_k}|$, dovoljno je pratiti veličine pivotnih elemenata i iteracije zaustaviti onaj čas kada pivot bude ispod $\zeta = \epsilon/(n(n-1))$. Time dobivamo garantiranu gornju ogradu za konkretno izračunatu matricu $\tilde{A}^{(k)}$.

Ovime smo došli do klasične Jacobijeve metode

Algoritam 5.4.1. Algoritam JACOBI_{1846} primjenjuje klasičnu Jacobijevu metodu s pivotnom strategijom (5.4.7). Iteracije se zaustavljaju kada je najveći po modulu izvandijagonalni element manji od zadanog pozitivnog parametra ζ.

$[U, \lambda] = \text{JACOBI}_{1846}(\mathsf{A}, \zeta)$

$U = I_n$;
$\omega = \max_{i \neq j} |a_{ij}|$;
while $\omega \geq \zeta$
  odredi (i', j') tako da je $\omega \equiv |a_{i'j'}| = \max_{i \neq j} |a_{ij}|$;
  if $\omega \geq \zeta$
    odredi Jacobijevu rotaciju $U_{[i',j']}$
    $A = U^T_{[i',j']} A U_{[i',j']}$;
    $U = U U_{[i',j']}$;
  end_if
end
$\lambda = \text{diag}(A)$;

Očita mana klasične metode je da priprema rotacije (traženje pivotne pozicije) zahtijeva $n(n-1)/2$ usporedbi da bi se utvrdila pozicija najvećeg elementa po modulu. Zato su razvijene pivotne strategije koje pivotne elemente uzimaju po unaprijed zadanom poretku i također garantiraju konvergenciju. To su tzv. cikličke strategije.

5.4.3 Cikličke metode

Ideja cikličkih metoda je vrlo jednostavna: umjesto potrage za najvećim elementom, treba jednostavno redom, sustavno, poništavati sve elemente. Konkretno, *ciklička strategija po retcima (stupcima)* bira pivotne pozicije prolazeći gornjim trokutom redak po redak (stupac po stupac):

$$\begin{pmatrix} * & \overset{1}{\rightarrow} & \overset{2}{\rightarrow} & \overset{3}{\rightarrow} & \overset{4}{\rightarrow} & \overset{5}{\rightarrow} & \overset{6}{\rightarrow} & \overset{7}{\rightarrow} \\ & * & \overset{8}{\rightarrow} & \overset{9}{\rightarrow} & \overset{10}{\rightarrow} & \overset{11}{\rightarrow} & \overset{12}{\rightarrow} & \overset{13}{\rightarrow} \\ & & * & \overset{14}{\rightarrow} & \overset{15}{\rightarrow} & \overset{16}{\rightarrow} & \overset{17}{\rightarrow} & \overset{18}{\rightarrow} \\ & & & * & \overset{19}{\rightarrow} & \overset{20}{\rightarrow} & \overset{21}{\rightarrow} & \overset{22}{\rightarrow} \\ & & & & * & \overset{23}{\rightarrow} & \overset{24}{\rightarrow} & \overset{25}{\rightarrow} \\ & & & & & * & \overset{26}{\rightarrow} & \overset{27}{\rightarrow} \\ & & & & & & * & \overset{28}{\rightarrow} \\ & & & & & & & * \end{pmatrix}, \quad \begin{pmatrix} * & 1\downarrow & 2\downarrow & 4\downarrow & 7\downarrow & 11\downarrow & 16\downarrow & 22\downarrow \\ & * & 3\downarrow & 5\downarrow & 8\downarrow & 12\downarrow & 17\downarrow & 23\downarrow \\ & & * & 6\downarrow & 9\downarrow & 13\downarrow & 18\downarrow & 24\downarrow \\ & & & * & 10\downarrow & 14\downarrow & 19\downarrow & 25\downarrow \\ & & & & * & 15\downarrow & 20\downarrow & 26\downarrow \\ & & & & & * & 21\downarrow & 27\downarrow \\ & & & & & & * & 28\downarrow \\ & & & & & & & * \end{pmatrix} \tag{5.4.12}$$

tj. ciklički po retcima biramo

$$(1,2),(1,3),\ldots,(1,n);(2,3),(2,4),\ldots,(2,n);\ldots;(n-2,n-1),(n-2,n);(n-1,n) \tag{5.4.13}$$

a ciklički po stupcima

$$(1,2);(1,3),(2,3);\ldots;(1,n-1),\ldots,(n-2,n-1);(1,n),(2,n),\ldots,(n-1,n) \tag{5.4.14}$$

Kada jednom obiđemo sve izvandijagonalne elemente kažemo da smo napravili jedan *ciklus*. (engl. *sweep*). Cikluse ponavljamo sve dok izvandijagonalni elementi ne postanu dovoljno mali. Formalno, ciklička pivotna strategija

$$\varpi : \mathbb{N} \longrightarrow \{(i,j) : 1 \le i < j \le n\}, \;\; \varpi(k) = (i_k, j_k),$$

je periodička funkcija perioda $n(n-1)/2$, gdje se osnovni period zadaje npr. sa (5.4.12, 5.4.13, 5.4.14).

5.4.3.1 Implementacija

Kada govorimo o implementaciji, vrijede svi komentari is §5.4.2. Naravno, kako je ovdje drugačiji izbor pivotnog elementa, iteracije ćemo zaustaviti kada provjerimo da su svi pivotni elementi manji od zadanog praga tolerancije.

$[U, \lambda] = \text{JACOBI\_CR}(\mathsf{A}, \zeta)$

$U = I_n$;
$\ell = 0$;
for ciklus $= 1:$
  for $i = 1 : n-1$
    for $j = i+1 : n$
      if $|a_{ij}| \ge \zeta$
        odredi Jacobijevu rotaciju $U_{[i',j']}$
        $A = U^T_{[i',j']} A U_{[i',j']}$; $U = U U_{[i',j']}$;
        $\ell = 0$;
      else
        $\ell = \ell + 1$;
        if $\ell = n(n-1)/2$, $\lambda = \text{diag}(A)$; return; *end_if*
      end_if
    end_for(j)
  end_for(i)
end_for(ciklus)
$\lambda = \text{diag}(A)$;

Bibliografija

[1] J. Barlow and J. Demmel, *Computing accurate eigensystems of scaled diagonally dominant matrices*, **27** (1990), no. 3, 762–791.

[2] R. Varga, *Matrix iterative analysis*, Springer Verlag, 2000.

[3] D. Young, *Iterative solution of large linear systems*, Academic Press, 1971.