Sadržaj

Glava 1

Uvod

Postavimo si vrlo jednostavan problem: rješavanje sustava linearnih jednadžbi

$$Ax = b, \tag{1.1}$$

pri čemu je A kvadratna $n \times n$, regularna matrica, a b je n-dimenzionalni vektor. Naoko jednostavan za rješavanje, naročito za maleni broj jednadžbi i nepoznanica, ovaj problem ipak krije mnoge zamke za rješavače, pa čak i danas kada nam na raspolaganju stoje vrlo moćna računala. Zbog toga su se vremenom razvile mnoge različite metode za rješavanje sustava (1.1), a u ovom radu specijalno ćemo se osvrnuti na grupu metoda koje nazivamo *iterativnim metodama*. Kao što samo ime kaže, kod ovih metoda iteriranjem pokušavamo poboljšati neku početnu aproksimaciju, tako da u svakoj iteraciji greška bude što manja. Primjenom odgovarajućeg kriterija zaustavljanja, nakon određenog broja iteracija, dobiveni vektor smatrat ćemo dovoljno dobrom aproksimacijom.

Sada se postavlja pitanje zašto se pored prilično djelotvorne metode Gaussovih eliminacija za rješavanje linearnih sustava razvio i veliki broj iterativnih metoda? Opće je poznato da rješavanje linearnog sustava pomoću Gaussovih eliminacija, u općenitom slučaju, zahtijeva zalihu u memoriji za skladištenje svih n^2 elemenata matrice A, i $\mathcal{O}(n^3)$ operacija za njezino izračunavanje. Matrice koje se pojavljuju u praksi, kao na primjer matrice dobivene iz diskretizacije diferencijabilnih jednadžbi, često imaju posebna svojstva. To se obično svodi na veliki broj elemenata koji su jednaki nuli, pa govorimo o *rijetko popunjenim* matricama, a često je i raspored netrivijalnih elemenata vrlo pravilan, kao na primjer, kod vrpčastih matrica. Gaussove eliminacije obično su u stanju samo djelomično iskoristiti ova svojstva, jer će se u rezultirajućim trokutastim faktorima pojaviti mnogi netrivijalni elementi na mjestima, na kojim je originalna matrica A imala nule. Osim toga vrlo često umjesto skladištenja same matrice, na raspolaganju nam stoji rutina za izračunavanje produkta matrice i vektora, pa su nam sami elementi matrice vrlo teško dostupni. S druge strane rijetka popunjenost i ostala svojstva mogu se vrlo dobro iskorirtiti kod računanja produkta matrice i vektora. Ako matrica ima samo nekoliko netrivijalnih elemenata po retku, tada je broj operacija, potrebnih za računanje produkta takve matrice i danog vektora, jednak $const \cdot n$, za malu konstantu $const$, za razliku od broja $2n^2$ operacija potrebnih za općenitu, gusto popunjenu matricu. Zalihe memorije za skladištenje ovakve matrice mogu biti puno manje od n^2 elemenata, ako spremamo samo netrivijalne elemente. Zato su se razvile iterativne metode za rješavanje sustava (1.1) koje koriste samo množenje matrice i vektora, uz mali broj još dodatnih operacija, i koje mogu prilično nadmašiti Gaussove eliminacije i u memoriji i po broju izvršenih operacija.

U nastavku ovog poglavlja biti će iznesene mnoge definicije i pomoćni rezultati koji će nam koristiti kod razvoja iterativnih metoda i kod teorema koji govore o njihovoj konvergenciji.

1.1 Notacija

U ovoj radnji pretpostavit ćemo da su komponente matrica i vektora elementi skupa kompleksnih brojeva $\mathbb{C}$, osim možda u nekim analizama konvergencije, kada će to biti posebno naglašeno. Oznaku ι koristit ćemo za $\sqrt{-1}$. Ako sa $A = (a_{ij})$ označimo matricu kojoj je element na poziciji (i, j) označen sa a_{ij}, tada sa gornjim indeksom * označavamo *konjugiranje* matrice, pri čemu je $A^* = (\bar{a}_{ji})$. Oznaka $\|\cdot\|$ će uvijek predstavljati proizvoljnu matričnu ili vektorsku normu, ovisno na što je primijenjena.

Za sustav (1.1), $x = A^{-1}b$ biti će rješenje sustava, a ako je x_k aproksimacija rješenja tada sa $r_k = b - Ax_k$ označavamo *rezidual*, a sa $e_k = x - x_k$ *grešku* pridruženu danoj aproksimaciji rješenja. Simbol ξ_j označava j-ti jedinični vektor, odnosno vektor kome je j-ta komponente jednaka 1, a sve ostale komponente su 0, pri čemu će se dimenzija vektora odrediti iz konteksta.

1.2 Vektorske norme i skalarni produkti

Vektorima ćemo smatrati elemente skupa n-dimenzionalnog vektorskog prostora $\mathbb{C}^n$, odnosno vektor je $v \in \mathbb{C}^n$. Sa v_i označavat ćemo i-tu komponentu vektora v. Vektorske norme koje se najčešće upotrebljavaju su

- euklidska ili 2-norma, $\|v\|_2 = (\sum_{i=1}^n |v_i|^2)^{1/2}$,
- 1-norma, $\|v\|_1 = \sum_{i=1}^n |v_i|$,
- ∞-norma, $\|v\|_\infty = \max_{i=1,\ldots,n} |v_i|$.

Ako je $\|\cdot\|$ neka vektorska norma, i ako je B regularna $n \times n$ matrica, tada je $\|v\|_{B^*B} = \|Bv\|$ također vektorska norma.

Euklidska norma je povezana sa standardnim skalarnim produktom

$$\langle v, w\rangle = w^*v = \sum_{i=1}^n \bar{w}_i v_i.$$

Slično B^*B-norma je povezana sa skalarnim produktom

$$\langle v, w\rangle_{B^*B} = \langle B^*Bv, w\rangle = \langle Bv, Bw\rangle.$$

Po definiciji imamo da je $\|v\|_2^2 = \langle v, v\rangle$, i na sličan način je $\|v\|_{B^*B}^2 = \langle Bv, Bv\rangle = \langle v, v\rangle_{B^*B}$.

1.3 Ortogonalnost

Za vektore v i w kažemo da su *ortogonalni* ako je $\langle v, w\rangle = 0$, i da su *ortonormirani* ako je još i $\|v\|_2 = \|w\|_2 = 1$. Za vektore v i w kažemo da su B^*B-*ortogonalni* ako je $\langle B^*Bv, w\rangle = 0$.

$n \times n$ kompleksnu matricu sa ortonormiranim stupcima nazivamo *unitarnom* matricom. Za unitarnu matricu U vrijedi $U^*U = UU^* = I$, gdje je I $n \times n$ matrica identitete. Ako je matrica U realna tada je možemo zvati *ortogonalnom* matricom. Pravokutnu $m \times n$ matricu V, za $m \geq n$, sa ortonormiranim stupcima nazivamo *ortonormalnom matricom*. Važno je napomenuti da za unitarnu matricu U i bilo koji vektor v vrijedi

$$\|Uv\|_2 = \sqrt{v^*U^*Uv} = \sqrt{v^*Iv} = \sqrt{v^*v} = \|v\|_2,$$

odnosno euklidska norma je *unitarno invarijantna*.

Ako imamo linearno nezavisan skup vektora $\{v_1, \ldots, v_n\}$, tada za proizvoljni skalarni produkt možemo konstruirati ortonormirani skup $\{u_1, \ldots, u_n\}$ pomoću *Gram–Schmidtovog* postupka.

Gram–Schmidtov postupak

- $u_1 = v_1/\|v_1\|$,
- $\tilde{u}_k = v_k - \sum_{i=1}^{k-1} \langle v_k, u_i \rangle u_i$, $u_k = \tilde{u}_k/\|\tilde{u}_k\|$, $k = 1, \ldots, n$.

Kod računanja ortogonalnog skupa vektora najčešće se upotrebljava matematički ekvivalentan postupak *modificirane Gram–Schmidtove* metode.

Algoritam 1.3.1. MODIFICIRANI GRAM–SCHMIDTOV ALGORITAM.

Izračunaj $u_1 = v_1/\|v_1\|$.

Za $k = 1, \ldots, n$,

$\tilde{u}_k = v_k$.
Za $i = 1, \ldots, k-1$,
$\tilde{u}_k := \tilde{u}_k - \langle \tilde{u}_k, u_i \rangle u_i$.
$u_k = \tilde{u}_k/\|\tilde{u}_k\|$.

Važno svojstvo ovako konstruiranog ortonormiranog skupa je

$$\text{span}\{u_1, \ldots, u_k\} = \text{span}\{v_1, \ldots, v_k\}$$

koje vrijedi za sve $k = 1, \ldots, n$. Modificirani Gram–Schmidtov postupak predstavlja osnovu mnogih iterativnih metoda za rješavanje linearnih sustava.

1.4 Matrične faktorizacije

Matrice koje ćemo mi promatrati su:

- *kvadratne* matrice koje su elementi skupa M_n (ili $\mathbb{C}^{n \times n}$), odnosno skupa $n \times n$ kompleksinih matrica (izomorfno sa vektorskim prostorom $\mathbb{C}^{n^2}$),
- *pravokutne* matrice koje su elementi skupa $\mathbb{C}^{m \times n}$, odnosno skupa $m \times n$ kompleksnih matrica (izomorfno sa vektorskim prostorom $\mathbb{C}^{mn}$).

Elemente matrice na poziciji (i, j) označavat ćemo sa a_{ij} ili A_{ij}, što će biti jasno iz konteksta u kojem se upotrebljavaju. Matrice možemo podijeliti u mnoge različite klase, što najčešće ovisi o određenim svojstvima matrice, ili rezultatima različitih faktorizacija. Standardne faktorizacije poput Jordanove forme ili Schurove dekompozicije daju obično važne informacije koje se često koriste kod analize numeričkih algoritama. U nastavku ćemo dati nekoliko teorema vezanih uz standardne faktorizacije, ali bez dokaza, budući da se oni mogu naći u bilo kojem udžbeniku linearne algebre.

Teorem 1.4.1 (Jordanova forma). *Neka je A proizvoljna $n\times n$ matrica. Tada postoji regularna matrica S takava da*

$$A = S \begin{bmatrix} J_{n_1}(\lambda_1) & & & \\ & J_{n_2}(\lambda_2) & & \\ & & \ddots & \\ & & & J_{n_m}(\lambda_m) \end{bmatrix} S^{-1} = SJS^{-1},$$

gdje su

$$J_{n_i}(\lambda_i) = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix},$$

$n_i \times n_i$ matrice, i $\sum_{i=1}^{m} n_i = n$.

Matricu J nazivamo *Jordanovom formom* matrice A. Stupci matrice S su *glavni vektori*, a vrijednosti λ_i za $i = 1, \ldots, m$ nazivamo *svojstvenim vrijednostima* matrice A, pri čemu skup svih svojstvenih vrijednosti matrice A čini *spektar* od A, $\sigma(A)$. Za svojstvene vrijednosti općenito vrijedi da za svaki i postoji vektor $v_i \neq 0$ takav da je

$$Av_i = \lambda_i v_i,$$

što se vidi i iz Teorema 1.4.1. Vektor v_i u tom slučaju nazivamo *svojstvenim vektorom* matrice A, koji je pridružen svojstvenoj vrijednosti λ_i, a broj Jordanovih blokova m predstavlja broj nezavisnih svojstvenih vektora matrice A. Budući da za matricu $A-\lambda_i I$, pri čemu je I $n \times n$ matrica identitete, vrijedi $(A - \lambda_i I)v_i = 0$, možemo zaključiti da je ona singularna, pa je shodno tome $\det(A - \lambda_i I) = 0$. Polinom

$$\kappa_A(\lambda) = \det(A - \lambda I)$$

stupnja n nazivamo *karakterističnim polinomom* matrice A, i iz prethodnog razmatranja vidimo da on poništava matricu A, odnosno da je $\kappa_A(A) = 0$. Dakle, postoji polinom stupnja A koji poništava matricu A, pa zaključujemo da postoji i *minimalni polinom* $\mu_A(\lambda)$, najmanjeg mogućeg stupnja, manjeg ili jednakog n, takav da poništava matricu A. Minimalni polinom je djelitelj karakterističnog polinoma, a korijeni oba polinoma su očito svojstvene vrijednosti matrice A.

Matrica A je *dijagonalizabilna* ako i samo ako je $m = n$, i tada svi stupci matrice S predstavljaju svojstvene vektore, a matrica J je *dijagonalana* jer su joj samo dijagonalni elementi J_{ii} eventualno različiti od nule.

Teorem 1.4.2 (Schurova dekompozicija). *Neka je A proizvoljna $n \times n$ matrica sa svojstvenim vrijednostima $\lambda_1, \ldots, \lambda_n$ koje su poredane u bilo kojem poretku. Tada postoji unitarna matrica U takva da je $A = UTU^*$, gdje je T gornje trokutasta matrica sa dijagonalnim elementima $T_{ii} = \lambda_i$.*

Definicija 1.4.3.

- *Matrica A je* normalna *ako vrijedi $A^*A = AA^*$.*
- *Matrica A je* hermitska *ako vrijedi $A^* = A$.*
- *Matrica A je* antihermitska *ako vrijedi $A^* = -A$.*
- *Matrica A je* pozitivno definitna *ako je hermitska, i ako vrijedi $\langle Av, v\rangle > 0$ za sve vektore $v \neq 0$.*
- *Matrica A je* pozitivno semidefinitna *ako je hermitska, i ako vrijedi $\langle Av, v\rangle \geq 0$ za sve vektore v.*

Iz Teorema 1.4.2 vidimo da za normalnu matricu, gornje trokutasta matrica $T = U^*AU$ je također normalna, jer je

$$T^*T = U^*A^*UU^*AU = U^*A^*AU = U^*AA^*U = U^*AUU^*A^*A = TT^*.$$

Odavde slijedi da T mora biti dijagonalna matrica, pa su normalne matrice oblika

$$A = U\Lambda U^*, \tag{1.2}$$

gdje je Λ dijagonalna matrica, a U unitarna. Ovakva dekompozicija se često naziva *spektralnom dekompozicijom*. Normalne matrice, prema tome spadaju pod dijagonalizabilne matrice, samo što su stupci matrice $S = U$ ortonormirani, a na dijagonali matrice Λ su smještene svojstvene vrijednosti λ_i. Hermitske i antihermitske matrice su također normalne matrice, pa se i one mogu faktorizirati u oblik (1.2). Dijagonalna matrica $\Lambda = U^*AU$ je u tom slučaju također hermitska odnosno antihermitska. Za hermitsku matricu tada vrijedi da je $\bar{\lambda}_i = \lambda_i$ za $i = 1, \ldots, n$, odnosno da su joj sve svojstvene vrijednosti realne, a za antihermitsku vrijedi $\bar{\lambda}_i = -\lambda_i$, to jest, svojstvene vrijednosti različite od nule su joj imaginarne. Za svojstvene vrijednosti λ_i pozitivno definitne matrice vrijedi

$$\langle Av_i, v_i\rangle = \langle \lambda_i v_i, v_i\rangle = \lambda_i \|v_i\|_2^2 > 0,$$

gdje je v_i svojstveni vektor različit od nul-vektora. Zbog toga je $\|v_i\|_2 > 0$, pa mora biti $\lambda_i > 0$ za $i = 1, \ldots, n$. Prema tome svojstvene vrijednosti pozitivno definitne matrice su sve pozitivne. Analogno, svojstvene vrijednsti pozitivno semidefinitne matrice su sve veće ili jednake nuli.

Sljedeći teorem govori o dekompoziciji matrice, koja kao rezultat daje informacije o veličinama koje su često vrlo važne kod analize numeričkih algoritama.

Teorem 1.4.4 (Dekompozicija singularnih vrijednosti). *Neka je A proizvoljna $m \times n$ matrica ranga k. Tada postoje unitarna $m \times m$ matrice U i unitarna $n \times n$ matrica V, takve da je*

$$A = U\Sigma V^*,$$

gdje je Σ $m \times n$ *matrica, definirana sa*

$$\begin{bmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix},$$

pri čemu vrijedi $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > 0$.

Veličine σ_i, koje zovemo *singularnim vrijednostima* matrice A, su zapravo kvadratni korijeni netrivijalnih svojstvenih vrijednosti hermitskih matrica $A^*A = V\Sigma^T\Sigma V^*$, gdje je $\Sigma^T\Sigma$ $n \times n$ dijagonalna matrica sa nenegativnim dijagonalnim elementima, i $AA^* = U\Sigma\Sigma^T U^*$, kod koje je $\Sigma\Sigma^T$ $m \times m$ dijagonalna matrica sa nenegativnim dijagonalnim elementima. Stupce matrice V nazivamo *desnim singularnim vektorima* matrice A, i predstavljaju svojstvene vektore matrice A^*A, dok stupce matrice U nazivamo *lijevim singularnim vektorima* od A, koji su uz to još i svojstveni vektori matrice AA^*.

Još jedan važan pojam koji je vezan uz singularnu dekompoziciju je *generalizirani inverz*. Neka je A $m \times n$ matrica, i neka je $A = U\Sigma V^*$, pri čemu su U, V i Σ kao u Teoremu 1.4.4. Tada je generalizirani inverz A^+ matrice A $n \times m$ matrica definirana sa

$$A^+ = V \begin{bmatrix} \sigma_1^{-1} & & & & & \\ & \ddots & & & & \\ & & \sigma_k^{-1} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} U^*.$$

Može se pokazati da za matricu A^+ vrijede *Moore–Penroseovi* uvjeti:

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

Preostale su nam još tri faktorizacije, koje se često koriste kao alat za rješavanje raznih problema

Teorem 1.4.5 (LU faktorizacija). *Naka je* A *regularna* $n \times n$ *matrica. Tada postoji matrica permutacije* P*, takva da se* A *može faktorizirati u oblik*

$$A = PLU,$$

gdje je L *donje trokutasta, a* U *gornje trokutasta matrica.*

LU faktorizacija je osnova standardne direktne metode za rješavanje linearnog sustava $Ax = b$, koja je poznata pod imenom Gaussove eliminacije. Prvo se matrica A faktorizira u oblik PLU, a onda se riješi sustav $Ly = P^*b$, da bi na kraju dobili konačno rješenje rješavnjem sustava $Ux = y$. Za neke matrice A matrica permutacije P za dobivanje LU faktorizacije nije potrebna, odnosno može biti jednaka identiteti, a faktorizacija se tada može izvesti direktno nad matricom A. Takav je slučaj na primjer, sa pozitivno definitnim matricama, kod kojih zbog hermitičnosti još vrijedi da ako L i U^* imaju iste dijagonalne elemente, tada je $U = L^*$. LU faktorizacija pozitivno definitne matrice tada ima oblik $A = LL^* = U^*U$, kojeg nazivamo *faktorizacijom Choleskog*.

Druga direktna metoda za rješavanje linearnih sustava i problema najmanjih kvadrata je QR fakorizacija.

Teorem 1.4.6 (QR faktorizacija). *Neka je A proizvoljna $m \times n$ matrica, sa $m \geq n$. Tada postoji $m \times n$ ortonormalna matrica Q i $n \times n$ gornje trokutasta matrica R, takva da je*

$$A = QR.$$

Druga varijanta iste faktorizacije je kada se matrici Q dodaju dodatni stupci, tako da dobijemo $m \times m$ unitarnu matricu $\hat{Q}$, takvu da je $A = \hat{Q}\hat{R}$, gdje je $\hat{R}$ $m \times n$ matrica kojoj je gornji $n \times n$ blok jednak matrici R, a ostatak jednak nuli.

Jedan način dobivanja QR faktorizacije matrice A je primjena modificiranog Gram–Schmidtovog algoritma na stupce od A. Drugi način je primjena niza unitarnih matrica nad matricom A, kako bi se transformirala u gornje trokutasti oblik. Budući da je produkt unitarnih matrica unitaran, i inverz unitarne matrice je unitaran, ovaj postupak će također dati QR faktorizaciju matrice A. Ako je matrica A punog stupčanog ranga, a dijagonalni elementi matrice R su pozitivni, tada su Q i R faktori jedinstveni, pa će ovakav način dobivanja QR faktorizacije dati isti rezultat kao i QR faktorizacija pomoću modificiranog Gram–Schmidtovog algoritma. Unitarne matrice koje se često koriste u QR faktorizaciji su *Hauseholderovi reflektori* i *Givensove rotacije* koje se primjenjuju kod matrica sa specijalnom strukturom. Kao što ćemo vidjeti, mnoge iterativne metode za rješavanje linearnih sustava koriste Givensove rotacije. Kod rješavanja problema najmanjih kvadrata, jednom kad se $m \times n$ matrica A transformira u oblik $\hat{Q}\hat{R}$ imamo

$$\min_y \|Ay - b\|_2 = \|\hat{Q}\hat{R}y - b\|_2 = \|\hat{R}\hat{y} - \hat{Q}^*b\|_2,$$

što se može riješiti rješavanjem gornje trokutastog sustava $Ry = Q^*b$.

Zadnja faktorizacija odnosi se samo na unitarne matrice.

Teorem 1.4.7 (CS dekompozicija). *Neka je W $n \times n$ unitarna matrica i neka je dana particija*

$$W = \begin{array}{c} \left[\begin{array}{cc} W_{11} & W_{12} \\ W_{21} & W_{22} \end{array}\right] \\ \begin{array}{cc} l & n-l \end{array} \end{array} \begin{array}{c} l \\ n-l \end{array}, \qquad 2l \leq n.$$

Tada postoje unitarne matrice U i V oblika

$$U = \left[\begin{array}{cc} U_{11} & 0 \\ 0 & U_{22} \end{array}\right], \qquad V = \left[\begin{array}{cc} V_{11} & 0 \\ 0 & V_{22} \end{array}\right],$$

gdje su U_{11}, V_{11} $l \times l$ *matrice takve da je*

$$U^*WV = \begin{bmatrix} \Gamma & -\Sigma & 0 \\ \Sigma & \Gamma & 0 \\ 0 & 0 & I \end{bmatrix} \begin{matrix} l \\ l \\ n-2l \end{matrix}$$

$$\begin{matrix} l & l & n-2l \end{matrix}$$

Pri tom vrijedi

$$\begin{aligned} \Gamma &= \operatorname{diag}(\gamma_1, \ldots, \gamma_l) \geq 0, \\ \Sigma &= \operatorname{diag}(\sigma_1, \ldots, \sigma_l) \geq 0, \\ \Gamma^2 + \Sigma^2 &= I, \end{aligned}$$

pri čemu dijagonalni elementi od Γ *mogu biti u bilo kojem poretku.*

1.5 Spektralni radijus i matrične norme

Najprije definirajmo spektralni radijus, koji predstavlja veličinu od velike važnosti kod analize određenih iterativnih metoda.

Definicija 1.5.1. Spektralni radijus $\rho(A)$ $n \times n$ *matrice* A *je*

$$\rho(A) = \max\{|\lambda| : \ \lambda \text{ je svojstvena vrijednost od } A\}.$$

Kao i kod vektorskih prostora $\mathbb{C}^n$, i za matrice se mogu definirati norme, koje mogu dati dosta informacija o samoj matrici. Neka je A $m \times n$ matrica. Najčešće upotrebljavane matrične norme su

- Spektralna norma ili 2-norma, $\|A\|_2 = \sqrt{\rho(A^*A)}$,
- Frobenijusova norma, $\|A\|_F = (\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2)^{1/2}$,
- 1-norma, $\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^m |a_{ij}|$,
- ∞-norma, $\|A\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^n |a_{ij}|$.

Za matričnu normu $\|\cdot\|$ kažemo da je *konzistentna* ako za $m \times n$ matricu A i $n \times p$ matricu B vrijedi $\|AB\| \leq \|A\| \cdot \|B\|$. Lako se može vidjeti da su sve gore navedene norme konzistentne. Još je važno primijetiti da je za spektralnu normu, prema singularnoj dekompoziciji i unitarnoj invarijantnosti $\|A\|_2 = \sigma_{max}(A)$, gdje je $\sigma_{max}(A)$ najveća singularna vrijednost matrice A.

Vrlo važna veličina vezana uz matrične norme, koja se pojavljuje u analizi konvergencije iterativnih metoda, i o kojoj u velikom broji slučajeva ovisi brzina konvergencije je *uvjetovanost matrice*. Za bilo koju matričnu normu $\|\cdot\|$ uvjetovanost invertibilne matrice A definiramo sa $\|A\| \cdot \|A^{-1}\|$. Mi ćemo uglavnom koristiti uvjetovanost definiranu preko spektralne norme

$$\kappa(A) = \kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{max}(A)}{\sigma_{min}(A)},$$

gdje su $\sigma_{max}(A)$ i $\sigma_{min}(A)$ najveća i najmanja singularna vrijednost matrice A.

Definicija 1.5.2. *Neka je $\|\cdot\|$ vektorska norma na $\mathbb{C}^n$.* Operatorska norma, *također označena sa $\|\cdot\|$, je matrična norma definirana na M_n sa*

$$\|A\| = \max_{\|v\|=1} \|Av\|.$$

Kažemo da vektorska norma $\|\cdot\|$ inducira *odgovarajuću matričnu normu.*

Ekvivalentna definicija za operatorsku normu je

$$\|A\| = \max_{v\neq 0} \frac{\|Av\|}{\|v\|}.$$

Primijetimo da za ovako definirane norme vrijedi $\|Av\| \leq \|A\|\|v\|$, odnosno matrična i vektorska norma $\|\cdot\|$ su *konzistente.*

Teorem 1.5.3 ([12]). *Spektralna matrična norma, 1-norma i ∞-norma su operatorske norme, inducirane odgovarajućim vektorskim normama.*

Dokaz: Kao posljedicu singularne dekompozicije imamo da je $A^*A = U\Lambda U^*$, gdje je U unitarna, a $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ dijagonalna matrica nenegativnih svojstvenih vrijednosti od A^*A. Ako uzmemo proizvoljni vektor v i ako definiramo $u = U^*v$, i λ_{max} kao najveću svojstvenu vrijednost matrice A^*A, tada vrijedi

$$\begin{aligned}
\|Av\|_2^2 &= \langle Av, Av\rangle = \langle A^*Av, v\rangle = \langle U\Lambda U^*v, v\rangle = \\
&= \langle \Lambda U^*v, U^*v\rangle = \langle \Lambda u, u\rangle = \sum_{i=1}^{n} \lambda_i |u_i|^2 \leq \\
&\leq \lambda_{max} \sum_{i=1}^{n} |u_i|^2 = \|A\|_2^2 \|u\|_2^2 = \|A\|_2^2 \|v\|_2^2.
\end{aligned}$$

Dakle za proizvoljni vektor v dobili smo nejednakost $\|Av\|_2 \leq \|A\|_2\|v\|_2$, odnosno

$$\max_{\|v\|_2=1} \|Av\|_2 \leq \|A\|_2.$$

Ako v izaberemo tako da on bude svojstveni vekortor matrice A^*A norme jedan, koji odgovara svojstvenoj vrijednosti λ_{max}, tada u gornjoj nejednakosti vrijedi jednakost. U tom je slučaju

$$\|Av\|_2 = \sqrt{\langle Av, Av\rangle} = \sqrt{\langle A^*Av, v\rangle} = \sqrt{\lambda_{max}}\sqrt{\langle v, v\rangle} = \sqrt{\lambda_{max}} = \|A\|_2,$$

odnosno

$$\max_{\|v\|_2=1} \|Av\|_2 \geq \|A\|_2.$$

Time smo pokazali da je spektralna norma inducirana euklidskom normom.

Raspišimo matricu A po stupcima $A = [a_1 \ \ldots \ a_n]$. Tada za proizvoljni vektor v imamo

$$\begin{aligned}
\|Av\|_1 &= \left\| \sum_{i=1}^{n} v_i a_i \right\|_1 \leq \sum_{i=1}^{n} \|v_i a_i\|_1 = \sum_{i=1}^{n} |v_i| \cdot \|a_i\|_1 \leq \\
&\leq \max_{j=1,\ldots n} \|a_j\|_1 \cdot \sum_{i=1}^{n} |v_i| = \|A\|_1 \|v\|_1.
\end{aligned}$$

Ovime smo dobili nejednakost

$$\max_{\|v\|_1=1} \|Av\|_1 \leq \|A\|_1.$$

Ali ako izaberemo v kao jedinični vektor sa jedinicom na poziciji k, pri čemu je k indeks stupca matrice A sa najvećom 1-normom, i nulama na ostalim pozicijama, tada je $\|Av\|_1 = \|A\|_1$, pa moramo imati

$$\max_{\|v\|_1=1} \|Av\|_1 \geq \|A\|_1.$$

Dakle i 1-norma je operatorska norma.

Ponovo uzmimo da je v proizvoljni vektor. Tada je

$$\begin{aligned} \|Av\|_\infty &= \max_{i=1,\ldots,n} \left| \sum_{j=1}^{n} a_{ij} v_j \right| \leq \max_{i=1,\ldots,n} \left(\sum_{j=1}^{n} |a_{ij}| \cdot |v_j| \right) \leq \\ &\leq \max_{j=1,\ldots,n} |v_j| \cdot \max_{i=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| = \|v\|_\infty \cdot \|A\|_\infty, \end{aligned}$$

pa prema tome vrijedi

$$\max_{\|v\|_\infty=1} \|Av\|_\infty \leq \|A\|_\infty.$$

S druge strane, pretpostavimo da k-ti redak od A ima najveću sumu apsolutnih vrijednosti svojih elemenata. Neka je v vektor čije su komponente jednake ± 1, i to tako da predznak j-te komponente od v odgovara predznaku elementa a_{kj}. Tada je

$$|(Av)_k| = \sum_{j=1}^{n} |a_{kj}| = \|A\|_\infty,$$

odnosno

$$\|Av\|_\infty \geq \|A\|_\infty.$$

Pokazali smo da je i ∞-norma operatorska norma. □

Teorem 1.5.4 ([12]). *Ako je $\|\cdot\|$ matrična norma na M_n, i ako je B $n \times n$ regularna matrica, tada je*

$$\|A\|_{B^*B} = \|BAB^{-1}\|$$

matrična norma. Ako je $\|\cdot\|$ inducirana vektorskom normom $\|\cdot\|$, tada je $\|\cdot\|_{B^*B}$ inducirana vektorskom normom $\|\cdot\|_{B^*B}$.

Dokaz: Aksiomi norme se lagano dokažu. Dokažimo da je $\|\cdot\|_{B^*B}$ operatorska norma, ako vrijedi $\|A\| = \max_{v\neq 0} \|Av\|/\|v\|$. Tada za $w = B^{-1}v$ vrijedi

$$\begin{aligned} \|A\|_{B^*B} &= \max_{v\neq 0} \|BAB^{-1}v\|/\|v\| = \\ &= \max_{w\neq 0} \|BAw\|/\|Bw\| = \\ &= \max_{w\neq 0} \|Aw\|_{B^*B}/\|w\|_{B^*B}, \end{aligned}$$

Pa je prema tome $\|\cdot\|_{B^*B}$ matrična norma inducirana vektorskom normom $\|\cdot\|_{B^*B}$. □

Preostali teoremi govore o odnosu spektralnog radijusa i matričnih normi.

Teorem 1.5.5 ([12]). *Ako je $\|\cdot\|$ bilo koja konzistentna matrična norma, i neka je A $n \times n$ matrica, tada je*

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

Dokaz: Neka je λ svojstvena vrijednost od A, za koju je $|\lambda| = \rho(A)$, i neka je $v \neq 0$ odgovarajući svojstveni vektor. Neka je V $n \times n$ matrica kojoj je svaki stupac jednak v. Tada je $AV = \lambda V$, i vrijedi

$$|\lambda| \cdot \|V\| = \|\lambda V\| = \|AV\| \leq \|A\| \cdot \|V\|.$$

Budući da je $V \neq 0$, imamo da je $\|V\| > 0$, odakle slijedi $\rho(A) \leq \|A\|$. □

Teorem 1.5.6 ([18]). *Neka je A $n \times n$ matrica i neka je dan $\epsilon > 0$. Tada postoji operatorska norma $\|\cdot\|_{\epsilon,A}$ takva da je*

$$\|A\|_{\epsilon,A} \leq \rho(A) + \epsilon.$$

Dokaz: Neka je $S^{-1}AS = J$ Jordanova forma matrice A. Definirajmo

$$D_\epsilon = \mathrm{diag}(1, \epsilon, \ldots, \epsilon^{n-1}),$$

i

$$J_\epsilon = D_\epsilon^{-1} J D_\epsilon,$$

tako da vrijedi

$$\begin{aligned} (J_\epsilon)_{ii} &= \epsilon^{-(i-1)} J_{ii} \epsilon^{i-1} = J_{ii} \\ (J_\epsilon)_{i,i+1} &= \epsilon^{-(i-1)} \cdot 1 \cdot \epsilon^{i} = \epsilon \end{aligned}$$

Matričnu normu, za proizvoljnu $n \times n$ matricu B, definirat ćemo na sljedeći način.

$$\|B\|_{\epsilon,A} = \|D_\epsilon^{-1} S^{-1} B S D_\epsilon\|_1,$$

pri čemu je odgovarajuća vektorska norma koja je inducira za proizvoljan vektor v, oblika

$$\|v\|_{\epsilon,A} = \|D_\epsilon^{-1} S^{-1} v\|_1,$$

što se lako može provjeriti iz niza jednakosti

$$\begin{aligned} \max_{v \neq 0} \frac{\|D_\epsilon^{-1} S^{-1} B v\|_1}{\|D_\epsilon^{-1} S^{-1} v\|_1} &= \max_{w \neq 0} \frac{\|D_\epsilon^{-1} S^{-1} B S D_\epsilon w\|_1}{\|w\|_1} = \\ &= \|D_\epsilon^{-1} S^{-1} B S D_\epsilon\|_1 = \|B\|_{\epsilon,A}, \end{aligned}$$

pri čemu je $w = D_\epsilon^{-1} S^{-1} v$. Napokon, vrijedi

$$\|A\|_{\epsilon,A} = \|J_\epsilon\|_1 \leq \rho(A) + \epsilon.$$

□

Kod proučavanja konvergencije jednostavnih iteracija, interesirat će nas uvjeti pod kojima potencije matrice A konvergiraju ka nul-matrici.

Teorem 1.5.7 ([12]). *Neka je A $n \times n$ matrica. Tada je $\lim_{k\to\infty} A^k$ ako i samo ako je $\rho(A) < 1$.*

Dokaz: Prvo pretpostavimo da je $\lim_{k\to\infty} A^k = 0$. Neka je λ svojstvena vrijednost od A sa svojstvenim vektorom $v \neq 0$. Budući da je $A^k v = \lambda^k v$, za proizvoljnu normu $\|\cdot\|$ slijedi

$$0 = \lim_{k\to\infty} \|\lambda^k v\| = \|v\| \lim_{k\to\infty} |\lambda^k|.$$

Kako je $\|v\| > 0$, vrijedi da je $\lim_{k\to\infty} |\lambda^k| = 0$, odakle slijedi da mora biti $|\lambda| < 1$, i to za proizvoljnu svojstvenu vrijednost od A. Dakle vrijedi i $\rho(A) < 1$.

Obrnuto, pretpostavimo da je $\rho(A) < 1$, i tada prema Teoremu 1.5.6 postoji operatorska norma $\|\cdot\|$, takva da je za $0 < \epsilon < 1 - \rho(A)$

$$\|A\| \leq \rho(A) + \epsilon < 1.$$

Odatle slijedi da je

$$\lim_{k\to\infty} \|A^k\| \leq \lim_{k\to\infty} \|A\|^k = 0.$$

Budući da su na M_n sve matrične norme ekvivalentne, znači da je i $\lim_{k\to\infty} \|A^k\|_F = 0$, odakle slijedi da svi elementi od A^k moraju težiti k nuli. □

Korolar 1.5.8 ([12]). *Neka je $\|\cdot\|$ bilo koja matrična norma na M_n. Tada je*

$$\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}$$

za svaku $n \times n$ matricu A.

Dokaz: Budući da je prema definiciji i Teoremu 1.5.5 $\rho(A)^k = \rho(A^k) \leq \|A^k\|$, imamo da je $\rho(A) \leq \|A^k\|^{1/k}$, za sve $k = 1, 2, \ldots$. Za proizvoljni $\epsilon > 0$, matrica $\tilde{A} = [\rho(A) + \epsilon]^{-1} A$ ima spektralni radijus strogo manji od jedan, pa je prema Teoremu 1.5.7 $\lim_{k\to\infty} \|\tilde{A}^k\| = 0$. Dakle, postoji broj K koji ovisi o ϵ, takav da je za svaki $k \geq K$, $\|\tilde{A}^k\| < 1$, što je ekvivalentno tvrdnji $\|A^k\| \leq [\rho(A) + \epsilon]^k$ ili $\|A^k\|^{1/k} \leq \rho(A) + \epsilon$ za sve $k \geq K$. Prema tome imamo

$$\rho(A) \leq \|A^k\|^{1/k} \leq \rho(A) + \epsilon$$

za sve $k \geq K$, a budući da to vrijedi za sve $\epsilon > 0$, slijedi da $\lim_{k\to\infty} \|A^k\|^{1/k}$ postoji i da je jednak $\rho(A)$. □

Preostaje nam još samo jedna lema koja se često koristi u analizi konvergencije iterativnih metoda, a govori o tome kako poznavanje norme matrice A određuje da li je matrica $I - A$ invertibilna, i ako je, na koji način možemo izračunati inverz.

Lema 1.5.9 ([18]). *Neka je A $n \times n$ matrica i neka je $\|\cdot\|$ operatorska norma za koju vrijedi da je $\|A\| < 1$. Tada je $I - A$ regularna matrica i vrijedi*

$$(I - A)^{-1} = \sum_{i=0}^{\infty} A^i.$$

Dokaz: Ako je $\|A\| < 1$ tada su prema Teoremu 1.5.5 i sve svojstvene vrijednosti matrice A po apsolutnoj vrijednosti manje od 1. Prema tome matrica $I - A$ nema niti jednu svojstvenu vrijednost jednaku nuli, pa je stoga regularna.

Definirajmo sada matricu S_k izrazom

$$S_k = \sum_{i=0}^{k} A^i,$$

u tom slučaju imamo da je

$$(I - A)S_k = S_k(I - A) = I - A^{k+1}, \qquad k \geq 1. \tag{1.3}$$

Budući da je $\|A\| < 1$, tada je prema dokazu Teorema 1.5.7

$$\lim_{k\to\infty} A^k = 0.$$

Dakle uzimanjem limesa, od jednakosti (1.3) dobit ćemo

$$(I - A) \lim_{k\to\infty} S_k = \lim_{k\to\infty} S_k(I - A) = I,$$

pa je prema tome

$$(I - A)^{-1} = \lim_{k\to\infty} S_k = \sum_{i=0}^{\infty} A^i.$$

□

1.6 Svojstvene vrijednosti i polje vrijednosti

U kasnijim poglavljima ove radnje vidjet ćemo da svojstvene vrijednosti normalne matrice daju sve važne informacije o matrici, koje su važne za konvergenciju iterativnih metoda za rješavanje sustava. Takav jednostavan skup karaktereristika za ne-normalne matrice, koje bi dale sve potrebne informacije o matrici, na žalost ne postoji. Međutim, polje vrijednosti dati će nam neke važne podatke.

Najprije ćemo navesti neke teoreme koji govore o svojstvenim vrijednostima i na koji način ih možemo locirati.

Teorem 1.6.1 (Gerschgorinov teorem [12]). *Neka je A $n \times n$ matrica, i neka*

$$R_i(A) = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \qquad i = 1, \ldots, n$$

označava sumu apsolutnih vrijednosti svih vandijagonalnih elemenata u retku i. Tada su sve svojstvene vrijednosti od A smještene u uniji krugova, koje nazivamo Gerschgorinovim krugovima

$$\bigcup_{i=1}^{n} \{z \in \mathbb{C} : \ |z - a_{ii}| \leq R_i(A)\}.$$

Dokaz: Neka je λ svojstvena vrijednost matrice A, sa odgovarajućim svojstvenim vektorom $v \neq 0$. Neka je v_p komponenta od v sa najvećom apsolutnom vrijednošću, odnosno $|v_p| \geq \max_{i \neq p} |v_i|$. Budući da je $Av = \lambda v$, imamo

$$\lambda v_p = (Av)_p = \sum_{j=1}^{n} a_{pj} v_j,$$

ili ekvivalentno

$$v_p(\lambda - a_{pp}) = \sum_{\substack{j=1 \\ j \neq p}}^{n} a_{pj} v_j.$$

Iz nejednakosti trokuta slijedi

$$|v_p||\lambda - a_{pp}| = \left| \sum_{\substack{j=1 \\ j \neq p}}^{n} a_{pj} v_j \right| \leq \sum_{\substack{j=1 \\ j \neq p}}^{n} |a_{pj}||v_j| \leq |v_p| R_p(A).$$

Budući da je $|v_p| > 0$, slijedi da je $|\lambda - a_{pp}| \leq R_p(A)$, to jest svojstvena vrijednost leži u Gerschgorinovom krugu, koji odgovara p-tom retku matrice A, pa prema tome sve svojstvene vrijednosti leže u uniji Gerschgorinovih krugova. □

Može se dalje pokazati da ako unija od k Gerschgorinovih krugova predstavlja povezano područje, koje je disjunktno sa ostatkom od $n-k$ krugova, tada to područje sadrži točno k svojstvenih vrijednosti od A.

Sada prelazimo na *polje vrijednosti*, koje je puno značajnije za nehermitske matrice od svojstvenih vrijednosti.

Definicija 1.6.2. Polje vrijednosti $n \times n$ *matrice* A *je*

$$\mathcal{F}(A) = \{\langle Av, v\rangle : \ v \in \mathbb{C}^n, \ \|v\|_2 = 1\}.$$

To se još zove i *numerički rang*. Ekvivalentna definicija je

$$\mathcal{F}(A) = \left\{ \frac{\langle Av, v\rangle}{\langle v, v\rangle} : \ v \in \mathbb{C}^n, \ v \neq 0 \right\}.$$

Polje vrijednosti je kompaktan skup u kompleksnoj ravnini, jer je slika neprekidne funkcije nad kompaktnim skupom euklidske kugle. Također se može pokazati da je ono i konveksan skup, što je poznato kao *Toeplitz–Hausdorffov teorem*,[22, str. 17-24]. *Numerički radijus* $\nu(A)$ je najveća apsolutna vrijednost elementa iz $\mathcal{F}(A)$, odnosno

$$\nu(A) = \max\{|z| : \ z \in \mathcal{F}(A)\}.$$

Lema 1.6.3 ([12]). *Neka je A $n \times n$ matrica, a α je kompleksni skalar. Tada vrijedi*

$$\mathcal{F}(A + \alpha I) = \mathcal{F}(A) + \alpha, \tag{1.4}$$

$$\mathcal{F}(\alpha A) = \alpha \mathcal{F}(A), \tag{1.5}$$

i

$$\mathcal{F}\left(\frac{1}{2}(A + A^*)\right) = Re(\mathcal{F}(A)). \tag{1.6}$$

Dokaz:

$$\begin{aligned}\mathcal{F}(A) &= \{\langle (A+\alpha I)v, v\rangle : \ \|v\|_2 = 1\} = \\ &= \{\langle Av, v\rangle + \alpha\langle v, v\rangle : \ \|v\|_2 = 1\} = \\ &= \{\langle Av, v\rangle : \ \|v\|_2 = 1\} + \alpha = \mathcal{F}(A) + \alpha.\end{aligned}$$

$$\begin{aligned}\mathcal{F}(\alpha A) &= \{\langle \alpha Av, v\rangle : \ \|v\|_2 = 1\} = \\ &= \{\alpha\langle Av, v\rangle : \ \|v\|_2 = 1\} = \alpha\mathcal{F}(A).\end{aligned}$$

$$\left\langle \frac{1}{2}(A+A^*)v, v \right\rangle = \frac{1}{2}(\langle Av, v\rangle + \langle A^*v, v\rangle) = \frac{1}{2}(\langle Av, v\rangle + \overline{\langle Av, v\rangle}) = Re(\langle Av, v\rangle).$$

Prema tome svaka vrijednost iz $\mathcal{F}(\frac{1}{2}(A + A^*))$ je oblika $Re(z)$ za neki $z \in \mathcal{F}(A)$ i obratno. □

Za bilo koju $n \times n$ matricu A, $\mathcal{F}(A)$ sadržava svojstvene vrijednosti od A, budući da je $\langle Av, v\rangle = \langle \lambda v, v\rangle = \lambda\langle v, v\rangle = \lambda$, za svojstvenu vrijednost λ i normalizirani svojstveni vektor v. Također, ako je U unitarna matrica tada je $\mathcal{F}(U^*AU) = \mathcal{F}(A)$, jer vrijednost $\langle U^*AUv, v\rangle$ iz $\mathcal{F}(U^*AU)$ sa $\|v\|_2 = 1$ odgovara vrijednosti $\langle AUv, Uv\rangle = \langle Au, u\rangle$ iz $\mathcal{F}(A)$ za $u = Uv$, $\|u\|_2 = \|v\|_2 = 1$, i obratno.

Za općenitu matricu A, neka $H(A) = \frac{1}{2}(A + A^*)$ označava hermitski dio od A. Tada je je prema (1.6) $\mathcal{F}(H(A)) = Re(\mathcal{F}(A))$. Zato možemo iznijeti teorem koji je analogan Gerschgorinovom, samo što vrijedi za polje vrijednosti, kako bi ga mogli aproksimativno locirati.

Teorem 1.6.4 ([12]). *Neka je A $n \times n$ matrica i neka su*

$$R_i(A) = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \qquad i = 1, \ldots, n,$$

$$C_j(A) = \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}|, \qquad j = 1, \ldots, n.$$

Tada je polje vrijednosti od A sadržano u

$$conv\left(\bigcup_{i=1}^{n}\left\{z \in \mathbb{C} : \ |z - a_{ii}| \leq \frac{1}{2}(R_i(A) + C_i(A))\right\}\right), \tag{1.7}$$

gdje $conv(\cdot)$ označava konveksnu ljusku.

Dokaz: Prvo primijetimo da, budući da je realni dio od $\mathcal{F}(A)$ jednak $\mathcal{F}(H(A))$ i budući da je $\mathcal{F}(H(A))$ konveksna ljuska svojstvenih vrijednosti od $H(A)$, iz Gerschgorinovog teorema primjenjenog na $H(A)$ slijedi

$$Re(\mathcal{F}(A)) \subset \text{conv}\left(\bigcup_{i=1}^{n}\left\{z \in \mathbb{R} : \ |z - Re(a_{ii})| \leq \frac{1}{2}(R_i(A) + C_i(A))\right\}\right). \tag{1.8}$$

Neka je sa $G_F(A)$ označen skup u (1.7). Ako je $G_F(A)$ sadržan u desnoj otvorenoj poluravnini $\{z : \ Re(z) > 0\}$, tada je $Re(a_{ii}) > \frac{1}{2}(R_i(A) + C_i(A))$ za sve i, i zbog toga

je i skup na desnoj strani u (1.8) sadržan u desnoj otvorenoj poluravnini. Budući da je $\mathcal{F}(A)$ konveksan skup, slijedi da i $\mathcal{F}(A)$ leži u desnoj otvorenoj poluravnini.

Sada pretpostavimo da je $G_F(A)$ sadržan u nekoj otvorenoj poluravnini kojoj se rub poklapa sa nekim pravcem kroz ishodište. Budući da je $G_F(A)$ konveksan skup, to je ekvivalentno uvjetu $0 \notin G_F(A)$. Tada postoji neko $\theta \in [0, 2\pi\rangle$, takvo da je $e^{\iota\theta} G_F(A) = G_F(e^{\iota\theta} A)$ sadržano u desnoj otvorenoj poluravnini. Iz prethodne analize slijedi da tada i $\mathcal{F}(e^{\iota\theta} A) = e^{\iota\theta}\mathcal{F}(A)$ leži u desnoj otvorenoj poluravnini, pa je prema tome $0 \notin \mathcal{F}(A)$.

Napokon, za bilo koji kompleksni broj α, ako $\alpha \notin G_F(A)$, tada $0 \notin G_F(A - \alpha I)$, pa prethodna razmatranja pokazuju da i $0 \notin \mathcal{F}(A - \alpha I)$. Prema (1.4) slijedi da $\alpha \notin \mathcal{F}(A)$. Zbog toga je $\mathcal{F}(A) \subset G_F(A)$. □

Za normalne matrice polje vrijednosti je konveksna ljuska spektra. Kako bismo to vidjeli, napišimo svojstvenu dekompoziciju matrice A kao $A = U\Lambda U^*$, gdje U unitarna, i $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$. Zbog svojstva unitarne invarijantnosti polja vrijednosti imamo da je $\mathcal{F}(A) = \mathcal{F}(\Lambda)$. Budući da je za v, takav da je $\|v\|_2 = 1$, $\langle \Lambda v, v\rangle = \sum_{i=1}^n |v_i|^2 \lambda_i$, slijedi da je $\mathcal{F}(\lambda)$ skup konveksnih kombinacija svojstvenih vrijednosti $\lambda_1, \ldots, \lambda_n$. Za hermitske matrice možemo biti još precizniji, budući da je zbog $\overline{\langle Av, v\rangle} = \langle v, Av\rangle = \langle Av, v\rangle$ polje vrijednosti realno.

Numerički radijus $\nu(A)$ također ima mnoga zanimljiva svojstva.

Lema 1.6.5 ([12]). *Neka su A i B proizvoljne $n \times n$ matrice, tada vrijedi*

$$\nu(A + B) \leq \nu(A) + \nu(B), \tag{1.9}$$

$$\frac{1}{2}\|A\|_2 \leq \nu(A) \leq \|A\|_2, \tag{1.10}$$

$$\nu(A^m) \leq [\nu(A)]^m, \qquad m = 1, 2, \ldots. \tag{1.11}$$

Dokaz:

$$\begin{aligned} \nu(A + B) &= \max_{\|v\|_2=1} |\langle (A + B)v, v\rangle| \leq \max_{\|v\|_2=1} (|\langle Av, v\rangle| + |\langle Bv, v\rangle|) \leq \\ &\leq \max_{\|v\|_2=1} |\langle Av, v\rangle| + \max_{\|v\|_2=1} |\langle Bv, v\rangle| = \nu(A) + \nu(B). \end{aligned}$$

Druga nejednakost u (1.10) slijedi iz činjenice da za bilo koji vektor v sa $\|v\|_2 = 1$ slijedi

$$|\langle Av, v\rangle| \leq \|Av\|_2 \|v\|_2 \leq \|A\|_2.$$

Prva nejednakost u (1.10) dobiva se na sljedeći način. Prvo primijetimo da je $\nu(A) = \nu(A^*)$. Ako napišemo A u obliku $A = H(A) + N(A)$, gdje je $N(A) = \frac{1}{2}(A - A^*)$, i ako primijetimo da su obje matrice $H(A)$ i $N(A)$ normalne, tada imamo

$$\|A\|_2 \leq \|H(A)\|_2 + \|N(A)\|_2 = \nu(H(A)) + \nu(N(A)),$$

jer se maksimalne vrijednosti elemenata konveksnog skupa $\mathcal{F}(A)$ mogu postići u njegovim vrhovima, koji su za normalne matrice jednaki svojstvenim vrijednostima. Korištenjem definicije numeričkog radijusa dobivamo

$$\begin{aligned} \|A\|_2 &\leq \frac{1}{2}\left[\max_{\|v\|_2=1} |\langle (A + A^*)v, v\rangle| + \max_{\|v\|_2=1} |\langle (A - A^v)v, v\rangle|\right] \leq \\ &\leq \frac{1}{2}\left[2 \max_{\|v\|_2=1} |\langle Av, v\rangle| + 2 \max_{\|v\|_2=1} |\langle A^* v, v\rangle|\right] = 2\nu(A). \end{aligned}$$

Za dokaz nejednakosti (1.11) vidi [22, Problem 27, str 333–334]. □

Teorem 1.6.6 (Rayleigh–Ritz [22]). *Neka je A $n \times n$ hermitska matrica, i neka su $\lambda_1 \leq \cdots \leq \lambda_n$ njene svojstvene vrijednosti. Tada vrijedi*

$$\lambda_1 = \min_{\|v\|_2=1} \langle Av, v\rangle, \qquad \lambda_n = \max_{\|v\|_2=1} \langle Av, v\rangle.$$

Dokaz: Neka su $u^{(1)}, \ldots, u^{(n)}$ ortonormirani svojstveni vektori matrice A, pridruženi svojstvenim vrijednostima $\lambda_1, \ldots, \lambda_n$. Za proizvoljan vektor v sa $\|v\|_2 = 1$ imamo $v = \sum_{i=1}^n v_i u^{(i)}$ i $\|v\|_2^2 = \sum_{i=1}^n |v_i|^2$. Vrijedi

$$\langle Av, v\rangle = \sum_{i=1}^n \lambda_i |v_i|^2 \geq \lambda_1 \|v\|_2^2 = \lambda_1,$$

odakle je $\min_{\|v\|_2=1}\langle Av, v\rangle \geq \lambda_1$. Ako uzmemo da je v svojstveni vektor $u^{(1)}$ svojstvene vrijednosti λ_1, tada je $\langle Av, v\rangle = \lambda_1$, što znači da je $\min_{\|v\|_2=1}\langle Av, v\rangle \leq \lambda_1$.

S druge strane za proizvoljni vektor v sa $\|v\|_2 = 1$ imamo

$$\langle Av, v\rangle = \sum_{i=1}^n \lambda_i |v_i|^2 \leq \lambda_n \|v\|_2^2 = \lambda_n,$$

odakle je $\max_{\|v\|_2=1}\langle Av, v\rangle \leq \lambda_1$. Ako pak uzmemo da je v svojstveni vektor svojstvene vrijednosti λ_n, tada je $\langle Av, v\rangle = \lambda_n$, što znači da je $\max_{\|v\|_2=1}\langle Av, v\rangle \geq \lambda_n$. Dakle u oba slučaja je tvrdnja dokazana. □

Na kraju ćemo iznijeti još jedan koristan teorem o ogradama svojstvenih vrijednosti hermitske matrice, čiji se dokaz nalazi u [30, str. 202–205].

Teorem 1.6.7 (Cauchyjev teorem ispreplitanja [30]). *Neka je A $n \times n$ hermitska matrica sa svojstvenim vrijednostima $\lambda_1 \leq \cdots \leq \lambda_n$, i neka je H bilo koja $m \times m$ glavna podmatrica (dobivena uzimanjem m stupaca i m redaka iz matrice A, sa indeksima $i_1, \ldots, i_m$), sa svojstvenim vrijednostima $\mu_1 \leq \cdots \leq \mu_m$. Tada za svako $i = 1, \ldots, m$ vrijedi*

$$\lambda_i \leq \mu_i \leq \lambda_{i+n-m}.$$

1.7 Perturbacijska teorija za linearne sustave

Kada linearni sustav $Ax = b$ rješavamo u aritmetici konačne preciznosti, to jest kada nam podaci nisu apsolutno točni, tada se pojavljuju tri važna pitanja:

(1) Koliko će se x promijeniti ako perturbiramo A i b (*greška unaprijed*), to jest, koliko je rješenje osjetljivo na perturbacije podataka?

(2) Koliko moramo perturbirati podatke A i b kako bi aproksimacija rješenja y postala egzaktno rješenje perturbiranog sustava (*povratna greška*)?

(3) Koju ogradu greške unaprijed za zadanu aproksimaciju rješenja, trebamo računati u praksi?

Kako bi odgovorili na ta pitanja potrebne su nam perturbacijske teorije po normi i po komponentama.

Analiza po normi

U nastavku $n \times n$ matrica E i n-dimenzionalni vektor f su proizvoljni i predstavljaju tolerancije u odnosu na koje se uspoređuju perturbacije. To, na primjer, mogu biti $E = A$ i $f = b$.

Prvi rezultat ove analize potvrđuje intuitivnu slutnju da ako je rezidual mali po normi, da tada imamo "dobru" aproksimaciju rješenja.

Teorem 1.7.1 (Rigal i Gaches [21]). *Za bilo koju operatorsku normu $\|\cdot\|$, povratna greška po normi za aproksimaciju rješenja y*

$$\eta_{E,f}(y) = \min\{\epsilon : \ (A + \Delta A)y = b + \Delta b, \ \|\Delta A\| \leq \epsilon\|E\|, \ \|\Delta b\| \leq \epsilon\|f\|\} \tag{1.12}$$

je dana sa

$$\eta_{E,f}(y) = \frac{\|r\|}{\|E\|\|y\| + \|f\|}, \tag{1.13}$$

gdje je $r = b - Ay$.

Dokaz: Uzmimo proizvoljni $\epsilon \geq 0$ i neka su ΔA i Δb takvi da je $\|\Delta A\| \leq \epsilon\|E\|$, $\|\Delta b\| \leq \epsilon\|f\|$ i $(A + \Delta A)y = b + \Delta b$. Tada vrijedi

$$r = b - Ay = \Delta A y - \Delta b,$$

odakle slijedi

$$\|r\| \leq \|\Delta A\|\|y\| + \|\Delta b\| \leq \epsilon(\|E\|\|y\| + \|f\|).$$

Dakle, možemo zaključiti da je desna strana u (1.13) donja ograda od $\eta_{E,f}(y)$, odnosno

$$\frac{\|r\|}{\|E\|\|y\| + \|f\|} \leq \eta_{E,f}(y).$$

Ta donja ograda se može ostvariti za perturbacije

$$\Delta A_{min} = \frac{\|E\|\|y\|}{\|E\|\|y\| + \|f\|} r z^*, \qquad \Delta b_{min} = -\frac{\|f\|}{\|E\|\|y\| + \|f\|} r,$$

gdje je z *dualan* vektoru y, što znači da je $z^*y = \max_{x \neq 0} \frac{|z^*x|}{\|x\|}\|y\| = 1$, vidi [21, str. 119]. Provjerimo da li to zaista vrijedi.

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

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

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

Dakle,

$$\eta_{E,f}(y) \leq \frac{\|r\|}{\|E\|\|y\| + \|f\|}.$$

□

Za izbor $E = A$ i $f = b$, $\eta_{E,f}(y)$ se naziva *relativna povratna greška po normi.* Sljedeći rezultat mjeri osjetljivost sustava, odnosno daje odgovarajuću grešku unaprijed.

Teorem 1.7.2 ([21]). *Neka je $Ax = b$ (A regularna), $(A + \Delta A)y = b + \Delta b$ i $\|\cdot\|$ operatorska norma, pri čemu je $\|\Delta A\| \leq \epsilon\|E\|$ i $\|\Delta b\| \leq \epsilon\|f\|$. Pretpostavimo da je $\epsilon\|A^{-1}\|\|E\| < 1$. Tada*

$$\frac{\|x - y\|}{\|x\|} \leq \frac{\epsilon}{1 - \epsilon\|A^{-1}\|\|E\|}\left(\frac{\|A^{-1}\|\|f\|}{\|x\|} + \|A^{-1}\|\|E\|\right), \tag{1.14}$$

i ova ograda se može postići do na potenciju prvog reda od ϵ.

Dokaz: Ogradu (1.14) možemo dobiti na sljedeći način.

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

odakle slijedi

$$y - x = A^{-1}\Delta b - A^{-1}\Delta Ax - A^{-1}\Delta A(y - x), \tag{1.15}$$

odnosno

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

Ako sada grupiramo sve izraze sa $\|y - x\|$ na lijevu stranu, dobit ćemo

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

odakle direktno slijedi (1.14).

Definirajmo

$$\Delta A = \epsilon\|E\|\|x\|wv^*, \qquad \Delta b = -\epsilon\|f\|w,$$

pri čemu je w vektor sa $\|w\| = 1$, $\|A^{-1}w\| = \|A^{-1}\|$, a v je vektor dualan vektoru x. Budući da je $\|A^{-1}\Delta A\| \leq \|A^{-1}\|\|\Delta A\| \leq \epsilon\|A^{-1}\|\|E\| < 1$, $I + A^{-1}\Delta A$ je regularna prema Lemi 1.5.9, i iz (1.15) slijedi

$$\begin{aligned} (I + A^{-1}\Delta A)(y - x) &= A^{-1}\Delta b - A^{-1}\Delta Ax = \\ &= -\epsilon\|f\|A^{-1}w - \epsilon\|E\|\|x\|A^{-1}wv^*x = \\ &= -\epsilon\|f\|A^{-1}w - \epsilon\|E\|\|x\|A^{-1}w, \end{aligned}$$

iz čega se vidi da je

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

i

$$\frac{\|y - x\|}{\|x\|} = \epsilon\left(\frac{\|f\|}{\|x\|} + \|E\|\right)\|(I + \epsilon\|E\|\|x\|A^{-1}wv^*)^{-1}A^{-1}w\|.$$

Za $u = \epsilon\|E\|\|x\|A^{-1}w$, lako se može provjeriti da vrijedi

$$(I + uv^*)^{-1} = I - \frac{1}{1 + v^*u}uv^*.$$

Zato imamo

$$\begin{aligned}(I+uv^*)^{-1}A^{-1}w &= \left(I-\frac{1}{1+v^*u}uv^*\right)A^{-1}w = \\ &= A^{-1}w-\frac{\epsilon\|E\|\|x\|A^{-1}wv^*A^{-1}w}{1+\epsilon\|E\|\|x\|v^*A^{-1}w} = \\ &= \left(1-\frac{\epsilon\|E\|\|x\|v^*A^{-1}w}{1+\epsilon\|E\|\|x\|v^*A^{-1}w}\right)A^{-1}w = \\ &= \frac{A^{-1}w}{1+\epsilon\|E\|\|x\|v^*A^{-1}w}.\end{aligned}$$

Kako je $\|A^{-1}w\| = \|A^{-1}\|$, imamo

$$\frac{\|y-x\|}{\|x\|} = \epsilon\left(\frac{\|f\|}{\|x\|}+\|E\|\right)\frac{\|A^{-1}\|}{1+\epsilon\|E\|\|x\|v^*A^{-1}w},$$

i taj se izraz zaista razlikuje od izraza na desnoj strani u (1.14) za $\mathcal{O}(\epsilon^2)$. □

U vezi sa načinom mjerenja perturbacija u prethodna dva teorema, definirajmo broj *uvjetovanosti po normi*

$$\kappa_{E,f}(A,x) = \lim_{\epsilon\to 0}\sup\left\{\frac{\|\Delta x\|}{\epsilon\|x\|} \;:\; (A+\Delta A)(x+\Delta x) = b+\Delta b,\right.$$
$$\left.\|\Delta A\| \le \epsilon\|E\|,\ \|\Delta b\| \le \epsilon\|f\|\right\}.$$

Budući da je ograda Teorema 1.7.2 oštra, slijedi

$$\kappa_{E,f}(A,x) = \frac{\|A^{-1}\|\|f\|}{\|x\|}+\|A^{-1}\|\|E\|. \tag{1.16}$$

Za izbor $E = A$ i $f = b$ imamo $\kappa(A) \le \kappa_{A,b}(A,x) \le 2\kappa(A)$, tako da se ograda (1.14) može oslabiti kako bi dobili oblik

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

Analiza po komponentama

Povratna greška po komponentama definira se kao

$$\omega_{E,f}(y) = \min\{\epsilon :\ (A+\Delta A)y = b+\Delta b,\ |\Delta A| \le \epsilon E,\ |\Delta b| \le \epsilon f\}, \tag{1.17}$$

pri čemu se pretpostavlja da E i f imaju nenegativne elemente. Nejednakosti između matrica ili vektora se podrazumijevaju da vrijede po komponentama. U ovoj definiciji svaki element perturbacije je uspoređen sa svojom vlastitom tolerancijom, pa su za razliku od definicije greške po normi, svih n^2+n parametara od E i f iskorišteni.

Kod izbora E i f, najčešći odabir tolerancija je $E = |A|$ i $f = |b|$, koji rezultira *relativnom povratnom greškom po komponentama.* Za taj izbor

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

u (1.17), pa prema tome ako je $\omega_{E,f}(y)$ mali, tada y rješava problem koji je blizak originalnom problemu, u smislu relativnih perturbacija po komponentama, i ima isti raspored nula. Još jedno atraktivno svojstvo relativne povratne greške po komponentama je neosjetljivost na skaliranje sustava: ako se $Ax = b$ skalira u oblik $(S_1AS_2)(S_2^{-1}x) = S_1b$, gdje su S_1 i S_2 dijagonalne matrice, i ako se y skalira kao $S_2^{-1}y$, tada ω ostaje nepromijenjen.

Izbor $E = |A|ee^T$, gdje je $e = [1 \ \dots \ 1]^T$, i $f = |b|$ daje povratnu grešku po retcima. Uvjet $|\Delta A| \leq \epsilon E$ postaje tada $|\Delta a_{ij}| \leq \epsilon\alpha_i$, gdje je α_i jednako 1-normi i-tog retka od A. Prema tome perturbacije $i - tog$ retka od A se uspoređuju sa normom tog retka. Povratna greška po stupcima može se dobiti na sličan način, ako uzmemo $E = ee^T|A|$ i $f = \|b\|_\infty e$.

Treći prirodni odabir toleranicija je $E = \|A\|ee^T$ i $f = \|b\|e$, gdje je $\|\cdot\|$ neka operatorska norma, za koje je $\omega_{E,f}(y)$ jednaka povratnoj grešci po normi $\eta_{E,f}(y)$ do na konstantu.

I za ovaj slučaj postoji jednostavna formula za $\omega_{E,f}(y)$.

Teorem 1.7.3 (Oettli i Prager [21]). *Povratna greška po komponentama je dana sa*

$$\omega_{E,f}(y) = \max_{i=1,\dots n} \frac{|r_i|}{(E|y| + f)_i}, \tag{1.18}$$

gdje je $r = b - Ay$, a $z/0$ se interpretira kao nula ako je $z = 0$ ili kao ∞ ako je $z \neq 0$.

Dokaz: Pokažimo prvo da je desna strana u (1.19) donja ograda od $\omega_{E,f}(y)$. Imamo

$$r = b - Ay = Ay + \Delta Ay - \Delta b - Ay = \Delta Ay - \Delta b,$$

odakle je

$$|r_i| = |(\Delta Ay)_i - (\Delta b)_i| \leq (|\Delta A||y|)_i + |\Delta b_i| \leq \epsilon(E|y| + f)_i,$$

odnosno

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

za svaki i, i $\epsilon \geq 0$, pa vrijedi i za miminimalni ϵ. Prema definiciji (1.17) onda vrijedi

$$\max_{i=1,\dots,n} \frac{|r_i|}{(E|y| + f)_i} \leq \omega_{E,f}(y).$$

Ograda se može postići za perturbacije

$$\Delta A = D_1ED_2, \qquad \Delta b = -D_1f,$$

pri čemu su

$$D_1 = \operatorname{diag}\left(\frac{r_i}{(E|y| + f)_i}\right)_{i=1}^n, \qquad D_2 = \operatorname{diag}(\operatorname{sign}(y_i))_{i=1}^n.$$

Provjerimo valjanost ove tvrdnje. Za $i = 1, \dots, n$ imamo

$$\begin{aligned}
[(A + \Delta A)y]_i &= (Ay)_i + (D_1ED_2y)_i = b_i - r_i + (D_1E|y|)_i = \\
&= b_i - r_i + \frac{r_i}{(E|y| + f)_i}(E|y|)_i = b_i - r_i\left(1 - \frac{(E|y|)_i}{(E|y| + f)_i}\right) = \\
&= b_i - \frac{r_if_i}{(E|y| + f)_i} = (b + \Delta b)_i
\end{aligned}$$

$$|\Delta A| = |D_1|E|D_2| = |D_1|E \leq \max_{i=1,\dots n} \frac{|r_i|}{(E|y| + f)_i} E,$$

$$|\Delta b| = |D_1|f \leq \max_{i=1,\dots n} \frac{|r_i|}{(E|y| + f)_i} f,$$

odakle je

$$\omega_{E,f}(y) \leq \max_{i=1,\dots n} \frac{|r_i|}{(E|y| + f)_i}.$$

□

Sljedeći rezultat daje grešku unaprijed koja odgovara povratnoj grešci po komponentama.

Teorem 1.7.4 ([21]). *Neka je $Ax = b$ (A regularna) i $(A + \Delta A)y = b + \Delta b$, gdje je $|\Delta A| \leq \epsilon E$ i $|\Delta b| \leq \epsilon f$, i pretpostavimo da je $\epsilon \||A^{-1}|E\| < 1$, gdje je $\|\cdot\|$ operatorska norma inducirana apsolutnom vektorskom normom $\|\cdot\|$ ($\||v|\| = \|v\|$). Tada*

$$\frac{\|x - y\|}{\|x\|} \leq \frac{\epsilon}{1 - \epsilon \||A^{-1}|E\|} \frac{\||A^{-1}|E|x| + |A^{-1}|f\|}{\|x\|}, \tag{1.19}$$

a za ∞-normu ta se ograda može dostići do na potenciju prvog reda od ϵ.

Dokaz: Iz jednakosti $(A + \Delta A)y = b + \Delta b$ slijedi

$$A(x - y) = \Delta b - \Delta A x + \Delta A(x - y),$$

odakle je

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

Dalje slijedi da je

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

Budući da je $\epsilon \||A^{-1}|E\| < 1$, prema Lemi 1.5.9 matrica $I - \epsilon|A^{-1}|E$ je invertibilna i vrijedi

$$(I - \epsilon|A^{-1}|E)^{-1} = \sum_{i=0}^{\infty} (\epsilon|A^{-1}|E)^i,$$

odakle je

$$\|(I - \epsilon|A^{-1}|E)^{-1}\| \leq \sum_{i=0}^{\infty} (\epsilon \||A^{-1}|E\|)^i = \frac{1}{1 - \epsilon \||A^{-1}|E\|}.$$

Dakle, imamo

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

Specijalno za ∞-normu definirajmo

$$\Delta A = \epsilon D_1 E D_2, \qquad \Delta b = -\epsilon D_1 f,$$

pri čemu su

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

za k takav da je k-ta komponenta vektora $|A^{-1}|E|x| + |A^{-1}|f$ jednaka ∞-normi tog vektora, odnosno

$$\||A^{-1}|E|x| + |A^{-1}|f\|_\infty = (|A^{-1}|E|x| + |A^{-1}|f)_k.$$

Kako je $A^{-1}\Delta A \leq |A^{-1}||\Delta A| \leq \epsilon|A^{-1}|E$, vrijedi $\|A^{-1}\Delta A\|_\infty \leq \epsilon\||A^{-1}|E\|_\infty < 1$, pa je ponovo prema Lemi 1.5.9 matrica $(I + A^{-1}\Delta A)$ invertibilna. Imamo

$$\begin{aligned} x - y &= (I + A^{-1}\Delta A)^{-1}(A^{-1}\Delta A x - A^{-1}\Delta b) = \\ &= \epsilon(I + \epsilon A^{-1}D_1 E D_2)^{-1}(A^{-1}D_1 E D_2 x + A^{-1}D_1 f) = \\ &= \epsilon\left(\sum_{i=0}^{\infty}(-\epsilon A^{-1}D_1 E D_2)^i\right)(A^{-1}D_1 E|x| + A^{-1}D_1 f) = \\ &= \epsilon(A^{-1}D_1 E|x| + A^{-1}D_1 f) + \mathcal{O}(\epsilon^2), \end{aligned}$$

odnosno, za k-ti jedinični vektor ξ_k je

$$\begin{aligned} |(x-y)_k| &= \epsilon|(\xi_k^T A^{-1}D_1(E|x| + f)| + \mathcal{O}(\epsilon^2) = \\ &= \epsilon\left|\sum_{i=1}^n |A^{-1}_{ki}|(E|x| + f)_i\right| + \mathcal{O}(\epsilon^2) = \\ &= \epsilon[|A^{-1}|(E|x| + f)]_k + \mathcal{O}(\epsilon^2) = \\ &= \epsilon\||A^{-1}|E|x| + |A^{-1}|f\|_\infty + \mathcal{O}(\epsilon^2). \end{aligned}$$

Prema tome je

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

a s druge strane iz (1.19) slijedi

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

odakle se vidi da vrijedi tvrdnja teorema. □

Teorem 1.7.4 upućuje na broj uvjetovanosti

$$\operatorname{cond}_{E,f}(A,x) = \lim_{\epsilon\to 0}\sup\left\{\frac{\|\Delta x\|_\infty}{\|x\|_\infty} \;:\; (A+\Delta A)(x+\Delta x) = b + \Delta b, \; |\Delta A| \leq \epsilon E,\ |\Delta b| \leq \epsilon f\right\},$$

koji je dan sa

$$\operatorname{cond}_{E,f}(A,x) = \frac{\||A^{-1}|E|x| + |A^{-1}|f\|_\infty}{\|x\|_\infty}. \tag{1.20}$$

Za specijalni slučaj kada je $E = |A|$ i $f = |b|$, Skeel je definirao uvjetovanost sa

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

koji se od $\text{cond}_{|A|,|b|}(A,x)$ razlikuje najviše za faktor 2, odnosno $\text{cond}_{|A|,|b|}(A,x) \leq 2\text{cond}(A,x)$. Ako još definiramo $\text{cond}(A) = \text{cond}(A,e)$ za $e = [1 \ \ldots \ 1]^T$, tada imamo

$$\text{cond}(A) = \text{cond}(A,e) = \||A^{-1}||A|\|_\infty \leq \kappa_\infty(A).$$

Numerička stabilnost

Povratne greške koje smo upravo definirali vode do definicije numeričke stabilnosti algoritama za rješavanje linearnih sustava. Kažemo da je numerička metoda za rješavanje kvadratnog, regularnog linearnog sustava $Ax = b$ *povratno stabilna po normi*, ako producira izračunato rješenje $\hat{x}$, takvo da je $\eta_{A,b}(\hat{x})$ reda veličine mašinske točnosti ϵ. Koliko ćemo dozvoliti da $\eta_{A,b}(\hat{x})$ bude velik, a da metodu još uvijek možemo smatrati povratno stabilnom, to ovisi o kontekstu. Obično se podrazumijeva da takva metoda daje $\eta_{A,b}(\hat{x}) = \mathcal{O}(\epsilon)$ za sve A i b.

Značajka povratne stabilnosti po normi je ta da izračunato rješenje $\hat{x}$ rješava malo perturbirani sustav, i ako podaci A i b sadrže netočnosti koje su ograđene samo po normi ($A \to A + \Delta A$ sa $\|\Delta A\| = \mathcal{O}(\epsilon\|A\|)$, i slično za b), tada $\hat{x}$ možemo smatrati egzaktnim rješenjem problema kojeg smo htjeli rješiti, jer ga točnije niti ne možemo riješiti.

Povratna stabilnost po komponentama se definira na sličan način, samo što u ovom slučaju zahtijevamo da je povratna greška po komponentama $\omega_{|A|,|b|}(\hat{x})$ reda veličine ϵ. To je malo stroži zahtjev od povratne greške po normi. Greške zaokruživanja koje su nastale tokom izvođenja metode koja je povratno stabilna po komponentama su po veličini i efektu ekvivalentne greškama koje su se pojavile kod konvertiranja podataka A i b u brojeve pomičnog zareza, kako su prikazani u računalu, prije samog rješavanja problema.

Ako je metoda povratno stabilna po normi tada prema Teoremu 1.7.2 greška unaprijed $\|x - \hat{x}\|/\|x\|$ je ograničena sa $\mathcal{O}(\epsilon\kappa(A))$. Međutim, metoda može dati izračunato rješenje kojemu je greška unaprijed tako ograničena, ali kojoj povratna greška po normi $\eta_{A,b}(\hat{x})$ nije reda veličine $\mathcal{O}(\epsilon)$. Zbog toga je korisno nazvati metodu kojoj je $\|x - \hat{x}\|/\|x\| = \mathcal{O}(\epsilon\kappa(A))$ *stabilnom unaprijed po normi*. Analogno, samo uključujući $\omega_{|A|,|b|}(\hat{x})$, kažemo da je je metoda *stabilna unaprijed po komponentama* ako je $\|x - \hat{x}\|_\infty/\|x\|_\infty = \mathcal{O}(\text{cond}(A,x)\epsilon)$.

Na kraju možemo naglasiti da kod rješavanja linearnog sustava iterativnim metodama, vrlo često se ne isplati računati aproksimaciju rješenja sa velikom točnošću. Naime, u većini slučajeva sam sustav je već aproksimacija nekog polaznog problema kojeg želimo riješiti, a i kod smještavanja elemenata matrice sustava i desne strane sustava u računalo, generiraju se greške. Stoga, i samo egzaktno rješenje sustava u nekoj mjeri odstupa od pravog rješenja polaznog problema. Ako je greška aproksimacije rješenja sustava, dobivene nakon primjene neke iterativne metode, reda velićine tog odstupanja egzaktnog rješenja, moramo biti zadovoljni sa dobivenim rezultatom, jer preciznije računanje nema smisla. Također, i povratna greška za tako dobivenu aproksimaciju će se stopiti sa greškom ulaznih podataka sustava.

Glava 2

Aproksimacije iz Krylovljevih potprostora

2.1 Razvoj iterativnih metoda i prekondicioniranja

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

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

gdje je $\kappa_A(\lambda) = \det(A - \lambda I) = \sum_{i=0}^{n} a_i \lambda^i$ karakteristični polinom matrice A. Slično možemo napisati da je $\mu_A(A) = 0$, gdje je $\mu_A(\lambda)$ minimilni polinom stupnja manjeg ili jednakog n. Linearni sustavi najčešće se rješavaju kada je matrica sutava regularna, jer je tada rješenje jedinstveno. Od sad pa na dalje, u ovoj radnji smatrat ćemo da je matrica sustava uvijek regularna. U tom slučaju nula ne može biti korijen karakterističnog polinoma, pa je $a_0 \neq 0$. Odavde jednostavnim računom možemo dobiti da vrijedi

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

to jest, da je

$$A^{-1} = -\frac{1}{a_0}(a_1 I + \cdots + a_{n-1}A^{n-2} + a_n A^{n-1}). \tag{2.1}$$

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

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

odnosno

$$x \in \ \text{span}\{b, Ab, \ldots, A^{n-1}b\} = \mathcal{K}_n(A, b). \tag{2.2}$$

Prostor koji se pojavljuje na desnoj strani u (2.2) zovemo *Krylovljevim prostorom* matrice A i inicijalnog vektora b. Upravo iz (2.2) dobivamo ideju za metode rješavanja sustava linearnih jednadžbi koje bi se temeljile na aproksimacijama iz Krylovljevih potprostora. Naime kako je vektor b jedini vektor direktno vezan za problem rješavanja

sutava $Ax = b$, čini nam se prirodno uzeti neki multipl od b kao prvu aproksimaciju rješenja, tj.

$$x_1 \in \text{ span}\{b\}.$$

Nakon toga računamo produkt Ab i zahtijevamo da nam je sljedeća aproksimacija jednaka nekoj linearnoj kombinaciji od b i Ab, tj.

$$x_2 \in \text{ span}\{b, Ab\}.$$

Taj se proces nastavlja, tako da nam aproksimacija u k-tom koraku zadovoljava

$$x_k \in \text{ span}\{b, Ab, \ldots, A^{k-1}b\}, \quad k = 1, 2, \ldots.$$

Metoda, naravno, mora odrediti kriterij po kojem biramo vektore iz pojedinih Krylovljevih potprostora u svakom koraku, tako da bi u optimalnom slučaju mogli dobiti rješenje u k-tom koraku, za $k \ll n$. Ako, pak račun vršimo na računalu, u obzir još moramo uzeti i greške zaokruživanja. Konkretne metode bi se zasnivale na aproksimacijama upravo ovako definiranih vektora x_k.

Prije svega promotrimo jednu vrlo jednostavnu činjenicu. Ako je x_k aproksimacija rješenja u k-tom koraku neke iterativne metode za rješavanje linearnih sustava, i ako u $(k+1)$-om koraku uzmemo

$$x_{k+1} = x_k + A^{-1}(b - Ax_k),$$

tada je $x_{k+1} = A^{-1}b$ rješenje sustava. Budući da je izračunavanje vektora $A^{-1}(b - Ax_k)$ ekvivalentno problemu rješavanja polaznog sustava, korekciju vektora x_k radit ćemo pomoću neke njegove aproksimacije, koja se lakše izračunava i koja će nas držati unutar Krylovljevih potprostora. Zato, najprije pretpostavimo da polazimo od iteracije jednostavnog oblika

$$x_{k+1} = x_k + \alpha_k(b - Ax_k), \tag{2.3}$$

gdje je α_k parametar koji određuje točku na pravcu koji prolazi kroz točku x_k i pruža se u smjeru vektora $b - Ax_k$. Promotrimo kako se na taj način možemo približiti vektoru iz (2.2). Lako se vidi, da je na temelju iteracije (2.3) za $k = 0, 1, 2, \ldots$

$$x_k \in x_0 + \text{span } \{r_0, Ar_0, A^2r_0, \ldots, A^{k-1}r_0\}, \tag{2.4}$$

$$r_k \in r_0 + \text{span } \{Ar_0, A^2r_0, A^3r_0, \ldots, A^kr_0\}, \tag{2.5}$$

pri čemu je $r_k = b - Ax_k$ rezidual k-te iteracije, a sa $e_k = A^{-1}r_k = A^{-1}b - x_k$ označavamo grešku. Prostor koji se pojavljuje na desnoj strani (2.4) je također Krylovljev potprostor, ali određen matricom A i vektorom r_0, koji ne mora biti jednak vektoru b. Međutim, ako napišemo da je prema jednakosti (2.1) inverz matrice A jednak $A^{-1} = q_{k-1}(A)$, gdje je $q_{k-1}(\lambda)$ polinom stupnja $(k-1)$ za neki $k \leq n$, tada za bilo koji r_0, u prostoru na desnoj strani izraza (2.4) možemo naći vektor oblika $x_0 + q_{k-1}(A)r_0 = x_0 + A^{-1}r_0 = A^{-1}b$ koji je rješenje linearnog sustava $Ax = b$.

Općenito, iteracija iterativne metode bit će oblika

$$x_k = x_{k-1} + z_{k-1}$$

ili

$$x_k = x_0 + w_k$$

pri čemu će z_{k-1} i w_{k-1} biti birani tako da x_k zadovoljava (2.4) i neki uvjet optimalnosti, najčešće vezan uz normu reziduala ili neku normu greške.

Kao što ćemo kasnije vidjeti, sve metode ovakvog oblika konvergiraju vrlo brzo ukoliko je matrica sustava A blizu identiteti. Naravno, to se događa vrlo rijetko, ali mi možemo cijeli sustav $Ax = b$ zamijeniti sa modificiranim sustavom

$$M^{-1}Ax = M^{-1}b \quad \text{ili} \quad AM^{-1}y = b,\ x = M^{-1}y, \tag{2.6}$$

čija bi matrica bila bliža identiteti od matrice A. Ovdje se radi o *lijevom* odnosno *desnom prekondicioniranju*. Ako je M hermitska i pozitivno definitna matrica, tada možemo izvesti simetrično prekondicioniranje i riješiti modificirani linearni sustav

$$L^{-1}AL^{-*}y = L^{-1}b, \quad x = L^{-*}y, \tag{2.7}$$

pri čemu je $M = LL^*$. Matrica L može biti hermitski drugi korijen od M, ili donje trokutasti faktor Choleskog od M, ili bilo koja druga matrica koja zadovoljava $M = LL^*$. U oba slučaja bitno je samo to da je računanje $M^{-1}z$ za bilo koji vektor z jednostavnije, odnosno da je jednostavnije riješiti sustav sa matricom sustava M. U svakom slučaju, vrlo često mi nećemo računati matrice M^{-1} i L eksplicitno.

Ako matricu prekondicioniranja M možemo izabrati tako da se linearni sustav sa matricom sustava M može jednostavno riješiti i da $M^{-1}A$ ili AM^{-1} ili $L^{-1}AL^{-*}$ aproksimiraju identitetu, tada primjenom iterativne metode na sustave (2.6) ili (2.7) možemo dobiti još bolju tehniku za računanje rješenja sustava. Pravi smisao tvrdnje da "prekondicionirana matrica aproksimira identitetu", ovisi o iterativnoj metodi koju smo koristili, i o njenim svojstvima.

Na kraju još trebamo reći da konvergenciju iterativne metode možemo najbolje kontrolirati kroz normu greške. Odabir norme ovisi o svojstvima metode koju koristimo. Međutim, kako nam je vektor greške $e_k = A^{-1}b - x_k$ nedostupan koliko i samo rješenje problema, najčešće se kontrolira norma reziduala $r_k = b - Ax_k$, koju lako možemo izračunati. U svakom slučaju, u našem interesu je da norma greške ili reziduala teži k nuli, kako se broj iteracija povećava, jer tada možemo reći da iterativna metoda konvergira.

2.2 Jednostavne iteracije

Ako imamo matricu prekondicioniranja M za sustav $Ax = b$, prirodan i najjednostavniji postupak za dobivanje aproksimacije rješenja bilo bi sljedeće. Budući da je matrica prekondicioniranja dizajnirana tako da $M^{-1}A$ u nekom smislu aproksimira identitetu, za $M^{-1}(b - Ax_k)$ može se reći da aproksimira grešku $e_k = A^{-1}b - x_k$ u aproksimaciji rješenja x_k. Bolja aproksimacija rješenja x_{k+1} se tako može postići ako uzmemo da je

$$x_{k+1} = x_k + M^{-1}(b - Ax_k) = x_k + M^{-1}r_k. \tag{2.8}$$

Procedura koja započinje sa početnom aproksimacijom rješenja x_0 i generira uzastopne aproksimacije koristeći (2.8) za $k = 0, 1, \ldots$ ponekad se zove *jednostavna iteracija*, ali često je nalazimo pod različitim imenima ovisno o izboru matrice M. Za M jednak dijagonali od A, procedura se zove *Jacobijeva* metoda; za M jednak donjem trokutu od A, to je *Gauss–Seidelova* metoda; za M oblika $\omega^{-1}D - L$, gdje je D dijagonala od A, L je strogi donji trokut od A, i ω *parametar relaksacije*, to je *SOR* metoda.

Implementacija metode koja se temelji na iteraciji (2.8) svodi se na sljedeći algoritam.

Algoritam 2.2.1. JEDNOSTAVNE ITERACIJE

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

riješi $Mz_0 = r_0$.

Za $k = 1, 2, \ldots$

$x_k = x_{k-1} + z_{k-1}$,
$r_k = b - Ax_k$,
riješi $Mz_k = r_k$.

Analizirajmo sada vektor greške e_k. Iz (2.8) slijedi

$$e_k = (I - M^{-1}A)e_{k-1} = \cdots = (I - M^{-1}A)^k e_0, \tag{2.9}$$

i

$$\|e_k\| \leq \|(I - M^{-1}A)^k\| \cdot \|e_0\|, \tag{2.10}$$

gdje je $\|\cdot\|$ bilo koja vektorska norma, uz uvjet da je matrična norma inducirana tom vektorskom normom $\|B\| = \max_{\|y\|=1} \|By\|$. U ovom slučaju, ograda u (2.10) je stroga, jer za svaki k očito postoji početna greška e_0 za koju vrijedi jednakost.

Lema 2.2.2 ([12]). *Norma greške u iteraciji (2.8) će težiti k nuli, a* x_k *će težiti k* $A^{-1}b$ *za svaku početnu grešku* e_0 *ako i samo ako vrijedi*

$$\lim_{k\to\infty} \|(I - M^{-1}A)^k\| = 0.$$

Dokaz: Iz (2.10) je jasno da ako $\lim_{k\to\infty} \|(I - M^{-1}A)^k\| = 0$ tada $\lim_{k\to\infty} \|e_k\| = 0$. Obratno, pretpostavimo da je $\|(I - M^{-1}A)^k\| \geq \alpha > 0$ za beskonačno mnogo vrijednosti od k. Vektori $e_{0,k}$ norme 1 za koje vrijedi jednakost u (2.10) tvore ograđeni, beskonačni skup u $\mathbf{C}^n$, koji, prema Bolzano–Weierstrassovom teoremu, sadrži konvergentan podniz. Neka je e_0 limes tog podniza. Tada za dovoljno veliki k' imamo $\|e_0 - e_{0,k}\| \leq \varepsilon < 1$ za sve $k \geq k'$, i

$$\begin{aligned} \|(I - M^{-1}A)^k e_0\| &\geq \|(I - M^{-1}A)^k e_{0,k}\| - \|(I - M^{-1}A)^k (e_{0,k} - e_0)\| \\ &\geq \|(I - M^{-1}A)^k\|(1 - \varepsilon) \geq \alpha(1 - \varepsilon). \end{aligned}$$

Budući da ovo vrijedi za beskonačno mnogo vrijednosti od k, slijedi da je $\lim_{k\to\infty} \|(I - M^{-1}A)^k e_0\|$, ako postoji, veći od 0, što je kontradikcija pretpostavci da greška teži k nuli. Znači da za svako $\alpha > 0$ postoji neki $\hat{k}$ takav da za sve $k \geq \hat{k}$, $\|(I - M^{-1}A)^k\| \leq \alpha$, odnosno da je $\lim_{k\to\infty} \|(I - M^{-1}A)^k\| = 0$. □

Koristeći rezultat Korolara 1.5.8 iz uvoda, dobivamo sljedeći rezultat.

Teorem 2.2.3 ([12]). *Iteracija (2.8) konvergira prema* $A^{-1}b$ *za svaku početnu grešku* e_0 *ako i samo ako je* $\rho(I - M^{-1}A) < 1$.

Dokaz: Ako je $\rho(I - M^{-1}A) < 1$, tada

$$\lim_{k\to\infty} \|(I - M^{-1}A)^k\| = \lim_{k\to\infty} \rho(I - M^{-1}A)^k = 0,$$

pa rezultat slijedi iz Leme 2.2.2. Ako pak pretpostavimo da iteracija (2.8) konvergira za svaku početnu grešku, tada prema Lemi 2.2.2 $\lim_{k\to\infty} \|(I - M^{-1}A)^k\| = 0$. S druge strane, uzmimo da je $\rho(I - M^{-1}A) \geq 1$. Tada bi prema gornjoj tvrdnji $\lim_{k\to\infty} \|(I - M^{-1}A)^k\| = \infty$, što je kontradikcija. □

Ako želimo znati koliko iteracija je potrebno izvršiti da bi se dobila aproksimacija čija relativna greška je manja ili jednaka od δ, tada uzimanjem norme iz (2.9) dobivamo

$$\|e_k\| \leq \|I - M^{-1}A\| \cdot \|e_{k-1}\|.$$

Odavde slijedi da ako je $\|I - M^{-1}A\| < 1$, tada se u svakoj iteraciji greška reducira za najmanje taj faktor. Greška će zadovoljavati $\|e_k\|/\|e_0\| \leq \delta$ ako je

$$\|I - M^{-1}A\|^k \leq \delta,$$

odnosno, kao je

$$k \geq \frac{\log \delta}{\log \|I - M^{-1}A\|}.$$

Budući da, prema Teoremu 1.5.6, za svaki $\epsilon > 0$ postoji matrična norma takva da je $\|I - M^{-1}A\| < \rho(I - M^{-1}A) + \epsilon$, tada ako je $\rho(I - M^{-1}A) < 1$ postoji norma za koju je $\|I - M^{-1}A\| < 1$, i za koju se greška reducira monotono, te konvergira najmanje linearno sa faktorom redukcije približno jednakim $\rho(I - M^{-1}A)$. Nažalost, ta norma je često vrlo neprirodna, pa je malo vjerojatno da bi netko zaista želio mjeriti konvergenciju pomoću nje.

2.3 Orthomin(1) i Orthomin(2)

Nadalje, promatrat ćemo metode koje nastoje poboljšati jednostavne iteracije (2.8) uvođenjem dinamički izračunatog parametra u iteraciju. Tako ćemo dobiti

$$x_{k+1} = x_k + \alpha_k(b - Ax_k) = x_k + \alpha_k r_k \tag{2.11}$$

u neprekondicioniranom slučaju, ili

$$x_{k+1} = x_k + \alpha_k M^{-1}(b - Ax_k) = x_k + \alpha_k M^{-1} r_k \tag{2.12}$$

u prekondicioniranom slučaju. Koncentrirat ćemo se najprije na neprekondicionirani slučaj, jer prekondicionirani sustav možemo također pisati kao $Ax = b$, samo što su matrica A i vektor b izmjenjeni.

2.3.1 Orthomin(1)

Jedan način odabira parametra α_k u iteraciji (2.11) je takav da u k-tom koraku iteracije takvog oblika dobijemo rezidual r_{k+1} sa minimalnom euklidskom normom. Najprije ćemo derivirati funkciju $f(\alpha_k) = r_{k+1}^* r_{k+1}$ ($f : \mathbb{R} \longrightarrow \mathbb{R}$) i njenu derivaciju izjednačiti s

nulom, pri čemu dobivamo r_{k+1} sa minimalnom normom. Naime, u tom je slučaju, ako uvrstimo da je $r_{k+1} = r_k - \alpha_k A r_k$, što slijedi iz (2.11),

$$f'(\alpha_k) = 2(\alpha_k r_k^* A^* A r_k - r_k^* A^* r_k),$$

odakle, ćemo dobiti izraz za α_k

$$\alpha_k = \frac{r_k^* A^* r_k}{r_k^* A^* A r_k} = \frac{\langle r_k, A r_k \rangle}{\langle A r_k, A r_k \rangle}. \tag{2.13}$$

Primijetimo da je tada r_{k+1} okomit na Ar_k. Okomitost r_{k+1} i Ar_k rezultira rezidualom koji je jednak rezidualu iz prethodne iteracije r_k minus njegova projekcija na smjer Ar_k. Slijedi da je

$$\|r_k\|_2^2 = \|r_{k+1}\|_2^2 + |\alpha_k|^2 \|A r_k\|_2^2 \geq \|r_{k+1}\|_2^2,$$

odnosno $\|r_{k+1}\|_2 \leq \|r_k\|_2$ pri čemu jednakost vrijedi ako i samo ako je sam r_k okomit na Ar_k. Metodu koja se temelji na takvim iteracijama možemo nazvati *Ortomin(1)*. O konvergenciji ove metode govori sljedeći teorem

Teorem 2.3.1 ([12]). *Euklidska norma reziduala iz iteracije (2.11) sa formulom za koeficijent (2.13) smanjuje se monotono za svaki početni vektor r_0 ako i samo ako $0 \notin \mathcal{F}(A)$, pri čemu $\mathcal{F}(A)$ označava polje vrijednosti matrice A.*

Dokaz: Budući da je polje vrijednosti od A^* jednako konjugiranom polju vrijednosti od A, uvjet teorema je ekvivalentan relaciji $0 \notin \mathcal{F}(A^*)$. Ako je $0 \in \mathcal{F}(A^*)$ i r_0 je vektor različit od nule, koji zadovoljava $\langle r_0, A r_0 \rangle = 0$, tada je $\|r_1\|_2 = \|r_0\|_2$. S druge strane, ako $0 \notin \mathcal{F}(A^*)$, tada $\langle r_k, A r_k \rangle$ ne može biti jednak nuli za bilo koji k i $\|r_{k+1}\|_2 < \|r_k\|_2$. □

Sljedeće ćemo pokazati da u uvjetima prethodnog teorema ne samo što je euklidska norma reziduala reducirana u svakom koraku, već da je ona reducirana najmanje za neki fiksni faktor, neovisan o k.

Teorem 2.3.2 ([12]). *Iteracija (2.11) sa formulom za koeficijent (2.13) konvergira rješenju $A^{-1}b$ za sve početne vektore r_0 ako i samo ako $0 \notin \mathcal{F}(A)$. U tom slučaju, euklidska norma reziduala zadovoljava*

$$\|r_{k+1}\|_2 \leq \sqrt{1 - d^2/\|A\|_2^2}\, \|r_k\|_2 \tag{2.14}$$

za sve k, gdje je d udaljenost od ishodišta do polja vrijednosti od A.

Dokaz: Pretpostavimo da je $0 \notin \mathcal{F}(A)$. Tada analogna tvrdnja vrijedi i za matricu A^*, odnosno vrijedi $0 \notin \mathcal{F}(A^*)$. Kako je $f(y) = \langle A^* y, y \rangle$ neprekidna funkcija, a $\mathcal{F}(A^*)$ je slika od f na kompaktnom skupu $\{y \in \mathbb{C}^n \ : \ \langle y, y \rangle = 1\}$, to znači da je polje vrijednosti također kompaktan skup u $\mathbb{C}$, odnosno ograničen i zatvoren. Zbog zatvorenosti polja vrijednosti i $0 \notin \mathcal{F}(A^*)$ postoji pozitivan broj d kojeg definiramo kao udaljenost $\mathcal{F}(A^*)$ od ishodišta. Slijedi

$$\left| \frac{\langle A^* y, y \rangle}{\langle y, y \rangle} \right| \geq d > 0$$

za sve kompleksne vektore $y \neq 0$. Kako vrijedi $r_{k+1} = r_k - \alpha_k A r_k$, uzimanjem skalarnog produkta od r_{k+1} sa samim sobom dobivamo

$$\langle r_{k+1}, r_{k+1} \rangle = \langle r_k, r_k \rangle - \frac{|\langle r_k, A r_k \rangle|^2}{\langle A r_k, A r_k \rangle},$$

što se može zapisati u obliku

$$\|r_{k+1}\|_2^2 = \|r_k\|_2^2 \left(1 - \left|\frac{\langle A^* r_k, r_k\rangle}{\langle r_k, r_k\rangle}\right|^2 \cdot \left(\frac{\|r_k\|_2}{\|Ar_k\|_2}\right)^2\right). \tag{2.15}$$

Ako nezavisno ogradimo zadnja dva faktora u (2.15), koristeći izraze za d i $\|A\|_2$, tada imamo

$$\|r_{k+1}\|_2 \leq \|r_k\|_2 \sqrt{1 - d^2/\|A\|_2^2}.$$

□

Ograda u (2.14) nije nužno oštra, jer vektor r_k za koje je faktor $|\langle A^* r_k, r_k\rangle / \langle r_k, r_k\rangle|^2$ u (2.15) jednak d^2 nije nužno onaj vektor za koji je faktor $(\|r_k\|_2/\|Ar_k\|_2)^2$ jednak $1/\|A\|_2^2$. Katkada se jača ograda može dobiti ako primijetimo da, zbog odabira parametra α_k, imamo

$$\|r_{k+1}\|_2 = \min_{\alpha \in \mathbb{C}} \|(I - \alpha A) r_k\|_2,$$

pa vrijedi

$$\|r_{k+1}\|_2 \leq \|I - \alpha A\|_2 \cdot \|r_k\|_2 \tag{2.16}$$

za bilo koji koeficijent α. U specijalnom slučaju kada je A hermitska i pozitivno definitna matrica, uzmimo da je $\alpha = 2/(\lambda_{min} + \lambda_{max})$, gdje je λ_{min} najmanja, a λ_{max} najveća svojstvena vrijednost od A. Imamo da je $A = U\Lambda U^*$, gdje su $U^*U = I$ i $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, a λ_i su svojstvene vrijednosti od A. Vrijedi

$$\begin{aligned} \|I - \alpha A\|_2 &= \|I - \alpha U\Lambda U^*\|_2 = \|U(I - \alpha\Lambda)U^*\|_2 = \|I - \alpha\Lambda\|_2 = \\ &= \max_{i=1,\ldots,n} |1 - \alpha\lambda_i| = \max_{i=1,\ldots,n} \left|\frac{\lambda_{max} + \lambda_{min} - 2\lambda_i}{\lambda_{max} + \lambda_{min}}\right|, \end{aligned}$$

za koje se lako pokaže da postiže maksimum za $\lambda_i = \lambda_{min}$ ili $\lambda_i = \lambda_{max}$, koji iznosi $\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}$. Na kraju dobivamo

$$\|r_{k+1}\|_2 \leq \left(\frac{\kappa(A) - 1}{\kappa(A) + 1}\right) \|r_k\|_2, \tag{2.17}$$

gdje je $\kappa(A) = \lambda_{max}/\lambda_{min}$ uvjetovanost matrice A.

Valja napomenuti da $\|r_{k+1}\|_2 \leq \|r_k\|_2$ općenito ne implicira $\|e_{k+1}\|_2 \leq \|e_k\|_2$. Primjer koji ćemo sada navesti demonstrira jedan takav slučaj.

Primjer 2.3.3. *Pretpostavimo da rješavamo sustav sa matricom A i vektorom b, koji su definirani sa*

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix},$$

i da je početna aproksimacija rješenja dana sa

$$x_0 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

Tada odgovarajuća greška i rezidual imaju oblik

$$e_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad r_0 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Nakon prvog koraka Orthomin(1) metode dobivamo da je $\alpha_0 = 1/2$*, odakle slijedi*

$$x_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3/2 \end{bmatrix},$$

pri čemu su

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ -1/2 \end{bmatrix}, \qquad r_1 = \begin{bmatrix} 0 \\ -1/2 \\ 1/2 \end{bmatrix}.$$

Na kraju imamo

$$1 = \|e_0\|_2 < \|e_1\|_2 = \frac{\sqrt{5}}{2} = 1.1180,$$

i

$$1 = \|r_0\|_2 > \|r_1\|_2 = \frac{\sqrt{2}}{2} = 0.7071.$$

U općenitijem nehermitskom slučaju, pretpostavimo da se polje vrijednosti matrice A nalazi u krugu $\mathbf{K} = \{z \in \mathbb{C} : \ |z - c| \leq s\}$ koji ne sadrži ishodište. Razmotrimo izbor $\alpha = 1/c$ u (2.16). Zbog svojstava (1.4) i (1.5) polja vrijednosti vrijedi da je

$$\mathcal{F}(I - (1/c)A) = 1 - (1/c)\mathcal{F}(A) \subseteq \{z \in \mathbb{C} : \ |z| \leq s/|c|\}.$$

Koristeći svojstvo (1.10) koje definira odnos numeričkog radijusa i norme matrice, možemo zaključiti da za ovakav izbor α je

$$\|I - \alpha A\|_2 = \|I - (1/c)A\|_2 \leq 2\frac{s}{|c|},$$

pa je prema tome

$$\|r_{k+1}\|_2 \leq 2\frac{s}{|c|}\|r_k\|_2. \tag{2.18}$$

Ova ocjena može biti i stroža i slabija od (2.14), ovisno o veličini i obliku polja vrijednosti.

Nakon prezentiranja rezultata koji govore o redukciji norme reziduala kod primjene Orthomin(1) metode, pogodan je trenutak da ilustriramo na koji način se ta redukcija može popraviti korištenjem prekondicioniranja. U prekondicioniranom sustavu matrica sustava je $M^{-1}A$, a matricu prekondicioniranja M možemo, na primjer, odabrati tako da je polje vrijednosti matrice $M^{-1}A$ smješteno u krugu radijusa jedan, sa središtem dalekim od ishodišta. U tom slučaju ocjena (2.18) daje vrlo dobar faktor redukcije koji je puno manji od 1.

2.3.2 Orthomin(2)(MINRES)

Metoda Orthomin(1) često radi korake u istom smjeru kojeg je već neki raniji korak prošao. Zato se javlja ideja da za korake unaprijed odaberemo skup ortogonalnih vektora, koje nazivamo *smjerovi traganja* d_0, d_1,..., d_{n-1}. Zapravo u našem slučaju koristit ćemo A^*A-ortogonalnost koja se definira kao

$$\langle d_i, d_j \rangle_{A^*A} = \langle Ad_i, Ad_j \rangle = 0,$$

pri čemu je, za pozitivno definitnu matricu B, dobro definiran skalarni produkt $\langle x, y \rangle_B = \langle Bx, y \rangle$. Lako se pokaže da su takvi vektori linearno nezavisni. U svakom smjeru d_k napravit ćemo točno jedan korak, i taj korak će biti takve dužine da ćemo poništiti komponentu vektora reziduala r_k u smjeru Ad_k. Novi rezidual r_{k+1} će u $(k+1)$-om koraku biti jednak početnom rezidualu r_0, kojem su odstranjene sve komponente u smjerovima Ad_0,...,Ad_k. Nakon najviše n koraka bit ćemo gotovi.

Općenito, u svakom koraku biramo novu aproksimaciju rješenja oblika

$$x_{k+1} = x_k + \alpha_k d_k, \tag{2.19}$$

pri čemu za rezidual vrijedi

$$r_{k+1} = r_k - \alpha_k A d_k. \tag{2.20}$$

Da bismo našli vrijednost od α_k, koristit ćemo činjenicu da je r_{k+1} ortogonalan na Ad_0,...,Ad_k, zbog prethodno navedenoga, pa više nikada nećemo morati ići u tim smjerovima. Na analogan način kao i kod Orthomin(1) okomitost r_{k+1} na Ad_k je ekvivalentna traženju r_{k+1} oblika (2.20) sa najmanjom euklidskom normom. I u tom slučaju dobivamo α_k istim postupkom, samo što je on sada oblika

$$\alpha_k = \frac{\langle r_k, Ad_k \rangle}{\langle Ad_k, Ad_k \rangle}. \tag{2.21}$$

Ovu metodu nazvat ćemo *Orthomin(2a).*

Orthomin(2a)

- Dana je početna iteracija x_0, i skup A^*A-ortogonalnih vektora $\{d_0, d_1, \ldots, d_{n-1}\}$.
- $r_0 = b - Ax_0$.
- Za $k = 1, 2, \ldots$
  - $\alpha_{k-1} = \langle r_{k-1}, Ad_{k-1} \rangle / \langle Ad_{k-1}, Ad_{k-1} \rangle$,
  - $x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$,
  - $r_k = r_{k-1} - \alpha_{k-1} A d_{k-1}$.

Da zaista vrijedi ortogonalnost r_{k+1} i $Ad_0, \ldots, Ad_{k-1}$ pokazat ćemo u sljedećem teoremu.

Teorem 2.3.4. *Za metodu Orthomin(2a) vrijede sljedeća svojstva:*

$$\langle Ad_i, Ad_j \rangle = 0 \quad (i \neq j) \tag{2.22}$$

$$\langle r_i, Ad_j \rangle = 0 \quad (j < i) \tag{2.23}$$

$$\langle r_0, Ad_i\rangle = \langle r_1, Ad_i\rangle = \cdots = \langle r_i, Ad_i\rangle. \tag{2.24}$$

Skalar α_k može se zato napisati kao

$$\alpha_k = \frac{\langle r_0, Ad_k\rangle}{\langle Ad_k, Ad_k\rangle}. \tag{2.25}$$

Dokaz: Prva jednakost je očita jer smo tako birali smjerove traganja. Koristeći činjenicu da je $r_{i+1} = r_i - \alpha_i Ad_i$ imamo

$$\langle r_{i+1}, Ad_j\rangle = \langle r_i, Ad_j\rangle - \alpha_i\langle Ad_i, Ad_j\rangle.$$

Ako je $j = i$ tada, zbog odabira α_i (2.21), vrijedi $\langle r_{i+1}, Ad_i\rangle = 0$. Što više, zbog (2.22) je $\langle r_{i+1}, Ad_j\rangle = \langle r_i, Ad_j\rangle$, za $j \neq i$, pa jednakosti (2.23) i (2.24) slijede iz tih tih relacija indukcijom. Na kraju, formula (2.25) slijedi iz (2.24) i (2.21). □

Teorem 2.3.5. *Metoda Orthomin(2a) je m-koračna metoda ($m \leq n$) u smislu da je u m-tom koraku aproksimacija x_m jednaka rješenju $A^{-1}b$.*

Dokaz: Neka je m najmanji cijeli broj takav da se r_0 nalazi u prostoru razapetom sa $Ad_0,\ldots,Ad_{m-1}$. Očito je $m \leq n$, budući da su vektori Ad_0,Ad_1,... linearno nezavisni, pa ih maksimalno može biti n. Zatim, izaberimo skalare $a_0,\ldots,a_{m-1}$ takve da je

$$r_0 = a_0 Ad_0 + \cdots + a_{m-1}Ad_{m-1}.$$

Odavde slijedi iz $r_0 = Ae_0$ da je

$$e_0 = a_0 d_0 + \cdots + a_{m-1}d_{m-1},$$

odnosno iz $e_0 = x - x_0$, za $x = A^{-1}b$, slijedi da je

$$x = x_0 + a_0 d_0 + \cdots + a_{m-1}d_{m-1}.$$

Koristeći se činjenicom da su smjerovi traženja d_i medjusobno A^*A-ortogonalni i jednakošću (2.25) iz Teorema 2.3.4, računanjem skalarnog produkta $\langle r_0, Ad_i\rangle$ dobivamo

$$\langle r_0, Ad_i\rangle = a_i\langle Ad_i, Ad_i\rangle$$

odnosno

$$a_i = \frac{\langle r_0, Ad_i\rangle}{\langle Ad_i, Ad_i\rangle} = \alpha_i.$$

Budući da je, primjenom indukcije na (2.19)

$$x_m = x_0 + \alpha_0 d_0 + \cdots + \alpha_{m-1}d_{m-1},$$

možemo zaključiti da tvrdnja $x = x_m$ vrijedi. □

Sada još treba definirati skup A^*A-ortogonalnih smjerova traganja $\{d_i : i = 0, 1, \ldots, n-1\}$. To možemo napraviti primjenom *Gram–Schmidtove metode A^*A-ortogonalizacije* na niz linearno nezavisnih vektora $u_0,\ldots,u_{n-1}$, koja je zapravo klasična Gramm–Schmidtova metoda uz skalarni produkt $\langle Ax, Ay\rangle = \langle x, y\rangle_{A^*A}$. Prema tome metoda glasi:

$$d_k = u_k + \sum_{i=0}^{k-1}\beta_{ki}d_i, \tag{2.26}$$

pri čemu su koeficijenti

$$\beta_{ki} = -\frac{\langle Au_k, Ad_i\rangle}{\langle Ad_i, Ad_i\rangle}, \tag{2.27}$$

i nakon svakog koraka vrijedi da je

$$\text{span}\{d_0, d_1, \ldots, d_i\} = \text{span}\{u_0, u_1, \ldots u_i\}.$$

Za Orthomin(2a) metodu uzet ćemo da je $u_i = r_i$. Uz pomoć Teorema 2.3.4 može se provjeriti da su vektori r_i linearno nezavisni. Naime, ako pretpostavimo da do k-tog koraka nismo došli do rješenja, tada su svi d_i, $i = 0, \ldots, k$ različiti od nul-vektora. U suprotnom, ako je $d_k = 0$, tada bi prema Teoremu 2.3.7, koji ćemo kasnije pokazati, moralo biti $r_k = 0$, odnosno rješenje je dostignuto. Dalje, pretpostavimo da je $\sum_{i=0}^{k} \gamma_i r_i = 0$ za neke konstante γ_i, $i = 0, \ldots, k$. Iz (2.22) i (2.26) slijedi da je $\langle Ar_i, Ad_i\rangle = \langle Ad_i, Ad_i\rangle$ i da je $\langle Ar_i, Ad_j\rangle = 0$ za $j > i$. Tada imamo

$$0 = \left\langle A\left(\sum_{i=0}^{k} \gamma_i r_i\right), Ad_k \right\rangle = \sum_{i=0}^{k} \gamma_i \langle Ar_i, Ad_k\rangle = \gamma_k \langle Ad_k, Ad_k\rangle,$$

pa budući da je $\langle Ad_k, Ad_k\rangle$ različito od nule, slijedi da je $\gamma_k = 0$. Nakon toga, skalarnim množenjem $A(\sum_{i=0}^{k-1} \gamma_i r_i)$ sa Ad_{k-1} dobit ćemo da je $\gamma_{k-1} = 0$, i tako redom sve do $\gamma_0 = 0$. Dakle, slijedi da su vektori $r_0, \ldots, r_k$ linearno nezavisni. Iz (2.23) i (2.26) može se vidjeti da vrijedi

$$\langle r_i, Ar_j\rangle = 0 \qquad (j < i). \tag{2.28}$$

Sljedeći zadatak je odrediti koeficijente β_{ki} za $i = 0, \ldots, k-1$. Problem se pojednostavljuje ako promatramo matrice A koje su hermitske. U tom slučaju jednakost (2.28) vrijedi za sve $j \neq i$, pa nadalje promatramo sljedeće

$$\langle Ar_k, r_{i+1}\rangle = \langle Ar_k, r_i\rangle - \bar{\alpha}_i \langle Ar_k, Ad_i\rangle,$$

odnosno vrijedi

$$\langle Ar_k, Ad_i\rangle = \frac{1}{\bar{\alpha}_i}(\langle Ar_k, r_i\rangle - \langle Ar_k, r_{i+1}\rangle). \tag{2.29}$$

Za $i < k-1$ lijeva strana u (2.29) je jednaka 0, pa su $\beta_{ki} = 0$ za $i = 0, 1, \ldots, k-2$, a $\beta_k = \beta_{k,k-1} = -\frac{\langle Ar_k, Ad_{k-1}\rangle}{\langle Ad_{k-1}, Ad_{k-1}\rangle}$. Sada smo u potpunosti definirali metodu *Orthomin(2)* za hermitske matrice, čiji algoritam onda izgleda ovako

Algoritam 2.3.6. ORTHOMIN(2)

Dana je početna iteracija x_0,

$d_0 = r_0 = b - Ax_0$.

Za $k = 1, 2, \ldots$

izračunaj Ad_{k-1},

$$\alpha_{k-1} = \frac{\langle r_{k-1}, Ad_{k-1}\rangle}{\langle Ad_{k-1}, Ad_{k-1}\rangle},$$

$$x_k = x_{k-1} + \alpha_{k-1} d_{k-1},$$

$$r_k = r_{k-1} - \alpha_{k-1} Ad_{k-1},$$

izračunaj Ar_k,

$$\beta_k = \frac{\langle Ar_k, Ad_{k-1}\rangle}{\langle Ad_{k-1}, Ad_{k-1}\rangle},$$

$$d_k = r_k - \beta_k d_{k-1}.$$

Navedimo još nekoliko svojstva Orthomin(2) metode.

Teorem 2.3.7. *Rezidual* r_k *dobiven u* k*-tom koraku metode Orthomin(2) primijenjene na hermitski sustav ima najmanju euklidsku normu na prostoru*

$$r_0 + span\{Ar_0, A^2 r_0, \ldots, A^k r_0\}. \tag{2.30}$$

Dokaz: Ako gledamo kako su definirani vektori r_k i d_k u ovoj metodi, tada možemo zaključiti sljedeće: $d_0 = r_0$, pa je $r_1 = r_0 - \alpha_0 Ar_0$ i $d_1 = (1 - \beta_1) r_0 - \alpha_0 Ar_0$, odnosno $r_1, d_1 \in \text{span}\{r_0, Ar_0\}$. Pretpostavimo da vrijedi da je $r_k, d_k \in \text{span}\{r_0, Ar_0, \ldots, A^k r_0\}$, tada za r_{k+1} imamo

$$r_{k+1} = r_k - \alpha_k Ad_k \in \text{span}\{r_0, \ldots, A^{k+1} r_0\},$$

pri čemu se iz indukcije vidi da je koeficijent uz r_0 uvijek jednak 1, odnosno

$$r_{k+1} \in r_0 + \text{span}\{Ar_0, A^2 r_0, \ldots, A^{k+1} r_0\} \tag{2.31}$$

i zbog $d_{k+1} = r_{k+1} - \beta_k d_k$

$$d_{k+1} \in \text{span}\{r_0, Ar_0, A^2 r_0, \ldots, A^{k+1} r_0\}. \tag{2.32}$$

Iz (2.32) slijedi da je $\text{span}\{Ad_0, Ad_1, \ldots, Ad_k\} = \text{span}\{Ar_0, \ldots, A^{k+1} r_0\}$, a kako je zbog (2.23) r_{k+1} okomit na $\text{span}\{Ad_0, Ad_1, \ldots, Ad_k\}$, onda slijedi da je r_{k+1} vektor u prostoru (2.31) sa najmanjom euklidskom normom. Za $k = n - 1$ zbog linearne nezavisnosti vektora $\{Ad_0, \ldots, Ad_{n-1}\}$ slijedi da je $r_n = 0$. □

Napomena. Ako pretpostavimo da je $d_k = 0$, tada prema (2.32) iz dokaza prethodnog teorema možemo zaključiti da postoje konstante a_i, $i = 0, 1, \ldots, k$ takve da je

$$\sum_{i=0}^{k} a_i A^i r_0 = 0. \tag{2.33}$$

Bez smanjenja općenitosti možemo pretpostaviti da je $a_0 \neq 0$, a u slučaju da to ne vrijedi tada prethodnu jednakost množimo s A^{-1} tako dugo dok ne dobijemo konstantu uz član r_0 koja je različita od nule. Jednakost (2.33) sada možemo pomnožiti sa a_0^{-1}, pri čemu dobivamo da se nula nalazi u prostoru (2.30). Budući da je r_k vektor iz prostora (2.30) sa najmanjom euklidskom normom, mora biti $r_k = 0$. Dakle, ako do k-tog koraka nismo došli do rješenja, vektori d_i, $i = 0, \ldots, k$ moraju biti različiti od nule.

Algoritam 2.3.6 za metodu Orthomin(2) koristi se i za nehermitske matrice no tada se gubi ortogonalnost Ad_i vektora. Ovaj algoritam može se zaustaviti i neobavljena posla, ukoliko je $\langle r_0, Ar_0 \rangle = 0$. Kao i kod $Orthomin(1)$, može se pokazati da Orthomin(2) konvergira ako je $0 \notin \mathcal{F}(A)$. Ako je korak iteracije Orthomin(2) metode izvediv, tada se euklidska norma reziduala reducira za najmanje onoliko koliko bi se reducirala u Orthomin(1) koraku iz iste točke. To je zbog toga što se norma reziduala minimizira po većem prostoru $r_k + \text{span}\{Ar_k, Ad_{k-1}\}$ umjesto po prostoru $r_k + \text{span}\{Ar_k\}$. Ograda (2.14) također vrijedi i za Orthomin(2) kada $0 \notin \mathcal{F}(A)$.

Algoritam koji aproksimira rješenje hermitskog linearnog sustava $Ax = b$ tako da minimizira rezidual po prostoru iz (2.31) je poznat pod imenom MINRES (Minimal residual iteration) algoritam.

2.3.3 Analiza greške i konvergencija Orthomin(2) metode

Kao poslijedica relacije (2.31), rezidual u k-tom koraku metode ima oblik

$$r_k = r_0 + \sum_{i=1}^{k} \psi_i A^i r_0 = \left(I + \sum_{i=1}^{k} \psi_i A^i \right) r_0.$$

Koeficijenti ψ_i su u vezi sa koeficijentima α_i i β_i, ali točna veza nam sada nije bitna. Bitno je to da Orthomin(2) metoda bira ψ_j takve da oni minimiziraju $\|r_k\|_2$. Izraz u zagradama može se shvatiti kao polinom koji za argument ima matricu A, pa se izraz za rezidual možemo izraziti kao

$$r_k = p_k(A) r_0, \tag{2.34}$$

gdje je p_k polinom k-tog stupnja, kod kojeg zahitjevamo da je $p_k(0) = 1$. Orthomin(2) metoda odabire taj polinom kada bira koeficijente ψ_j.

U slučaju da je matrica A hermitska, tada matricu možemo zapisati kao $A = U\Lambda U^*$, pri čemu su $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, $\lambda_1, \ldots, \lambda_n$ svojstvene vrijednosti od A i $U^*U = UU^* = I$. Zbog toga slijedi

$$\begin{aligned} \|r_k\|_2 &= \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \|p_k(A) r_0\|_2 \leq \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \|p_k(A)\|_2 \|r_0\|_2 = \\ &= \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \|U p_k(\Lambda) U^*\|_2 \|r_0\|_2 = \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \|p_k(\Lambda)\|_2 \|r_0\|_2, \end{aligned}$$

odnosno vrijedi

$$\|r_k\|_2 \leq \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \max_{i=1,\ldots,n} |p_k(\lambda_i)| \|r_0\|_2, \tag{2.35}$$

gdje je sa $\mathbb{P}_k$ definiran skup polinoma stupnja k. Općenito, polinom n-tog stupnja može biti jednoznačno određen ako je zadan na $n+1$ točaka. Ako uzmemo da je $p_n(0) = 1$ i da u svim svojstvenim vrijednostima od A (njih n) p_n poprima vrijednost 0, tada je to polinom koji minimizira (2.35), pri čemu je onda r_n jednak nuli. U egzaktnoj aritmetici

broj iteracija k koje su potrebne da se izračuna egzaktno rješenje je manji ili jednak broju različitih svojstvenih vrijednosti, jer možemo naći polinom stupnja k koji je u 0 jednak 1, i u svim različitim svojstvenim vrijednostima jednak 0. Za takav polinom je $\max_{i=1,\ldots,n} |p_k(\lambda_i)| = 0$, pa iz ocjene (2.35) slijedi da je $r_k = 0$. Time se vidi da je Orthomin(2) metoda brža ako A ima višestrukih svojstvenih vrijednosti.

Mi ćemo sada prikazati analizu za indefinitni problem, i to u specijalnom slučaju kada su svojstvene vrijednosti od A sadržane u dva intervala $[a, b] \bigcup [c, d]$, gdje je $a < b < 0 < c < d$ i $b - a = d - c$. Budući da je desna strana nejednakosti (2.35) manja ili jednaka minimizaciji pa maksimizaciji po uniji tih skupova, jednostavniji pristup je minimizacija po tim segmentima nego po konačnom broju točaka. Polinomi s kojima se to postiže bazirani su na Čebiševljevim polinomima.

Čebiševljev polinom stupnja k je

$$T_k(\omega) = \frac{1}{2}\left[(\omega + \sqrt{\omega^2 - 1})^k + (\omega - \sqrt{\omega^2 - 1})^k\right].$$

Čebiševljevi polinomi imaju sljedeća svojstva:

$$|T_k(\omega)| \leq 1, \qquad \omega \in [-1, 1],$$

rapidno osciliraju između -1 i 1, odnosno

$$T_k\left(\cos\left(\frac{i\pi}{k}\right)\right) = (-1)^i, \qquad i = 0, 1, \ldots, i,$$

i k nultočaka polinoma T_k moraju se nalaziti između $k + 1$ ekstrema od T_k u segmentu $[-1, 1]$. Najbitnije mu je svojstvo to da između svih normiranih polinoma stupnja k polinom $2^{1-k}T^k$ ima najmanju ∞-normu na intervalu $[-1, 1]$, koja iznosi 2^{1-k}.

U ovom slučaju, polinom k-tog stupnja koji ima vrijednost 1 u ishodištu, i ima minimalnu ∞-normu na $[a, b] \bigcup [c, d]$ je dan sa

$$p_k(z) = \frac{T_l(q(z))}{T_l(q(0))}, \quad q(z) = 1 + \frac{2(z - b)(z - c)}{ad - bc}, \tag{2.36}$$

gdje je $l = [\frac{k}{2}]$, $[\cdot]$ označava cjelobrojni dio, i T_l je l-ti Čebiševljev polinom. Primijetimo da funkcija $q(z)$ preslikava svaki od intervala $[a, b]$ i $[c, d]$ na interval $[-1, 1]$. Slijedi da za $z \in [a, b] \bigcup [c, d]$ apsolutna vrijednost brojnika u izrazu za $p_k(z)$ iz (2.36), je ograničena sa 1. Nazivnik je određen tako da zadovolji svojstvo $p_k(0) = 1$. Znači, imamo sljedeće

$$\|r_k\|_2 \leq T_k\left(\frac{ad + bc}{ad - bc}\right)^{-1} \|r_0\|_2,$$

i nakon što izračunamo vrijednost Čebiševljevog polinoma, dobivamo

$$\|r_k\|_2 \leq 2\left[\left(\frac{\sqrt{|ad|} + \sqrt{|bc|}}{\sqrt{|ad|} - \sqrt{|bc|}}\right)^{[\frac{k}{2}]} + \left(\frac{\sqrt{|ad|} - \sqrt{|bc|}}{\sqrt{|ad|} + \sqrt{|bc|}}\right)^{[\frac{k}{2}]}\right]^{-1} \|r_0\|_2. \tag{2.37}$$

Drugi sumand u (2.37) konvergira k nuli kada k raste, pa se ocjena greške Orthomin(2) metode za indefinitne hermitske matrice često izražava sa slabijom formulacijom

$$\|r_k\|_2 \leq 2\left(\frac{\sqrt{|ad|} - \sqrt{|bc|}}{\sqrt{|ad|} + \sqrt{|bc|}}\right)^{[\frac{k}{2}]} \|r_0\|_2. \tag{2.38}$$

U slučaju pozitivno definitne hermitske matrice A ocjena greške se izvodi analogno kao i u slučaju metode konjugiranih gradijenata, koja će biti prezentirana u sljedećem poglavlju, a slična je prethodnoj analizi. U tom slučaju dobivamo ocjenu

$$\|r_k\|_2 \leq 2 \left[\left(\frac{\sqrt{\kappa(A)}+1}{\sqrt{\kappa(A)}-1} \right)^k + \left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1} \right)^k \right]^{-1} \|r_0\|_2, \tag{2.39}$$

odnosno

$$\|r_k\|_2 \leq 2 \left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1} \right)^k \|r_0\|_2. \tag{2.40}$$

Na žalost, nikakve jače a priori ograde na normu reziduala nisu poznate za slučaj kada se Orthomin(2) primijeni na općenitu matricu, čije polje vrijednosti ne sadrži ishodište, iako u praksi može biti značajnije bolja od Orthomin(1). Naime, kod primjene Orthomin(2) metode na nehermitsku matricu, gubi se svojstvo opisano u Teoremu 2.3.7. Međutim, u jednom od daljnjih odjeljaka biti će opisana GMRES metoda, za koju tvrdnja Teorema 2.3.7 uvijek vrijedi, pa i za nehermitske matrice. Također će biti pokazano da se za tu metodu može konstruirati linearni sustav sa nehermitskom matricom sustava A, koja može imati proizvoljne svojstvene vrijednosti, i sa vektorom desne strane b, takvim da u k-tom ($k < n$) koraku GMRES metode producira rezidual r_k^{GMRES} sa proizvoljnom normom. Za $k = n$ vrijedit će $r_n^{GMRES} = 0$. Kako za svaki k vrijedi

$$\|r_k^{GMRES}\|_2 \leq \|r_k^{Orthomin(2)}\|_2,$$

to znači da u svakom koraku k metode Orthomin(2) primijenjene na nehermitski sustav možemo dobiti rezidual $r_k^{Orthomin(2)}$, čija je donja ograda norme proizvoljno velika. Dakle, možemo zaključiti da ne možemo dobiti gornju ocjenu norme reziduala u svakom koraku Orthomin(2) metode, koja bi ovisila o spektru matrice sustava A.

2.3.4 Prekondicionirana Othomin(2) metoda

Ovom metodom možemo riješiti i prekondicionirani sustav $M^{-1}Ax = M^{-1}b$, pri čemu M biramo tako da matrica $M^{-1}A$ ima manji broj uvjetovanosti, kako bi smanjili broj iteracija za postizanje zadane točnosti. Dakle, u slučaju prekondicioniranog sustava imamo sljedeći algoritam

Algoritam 2.3.8. PREKONDICIONIRANI ORTHOMIN(2)

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

riješi $Mp_0 = r_0$,

$d_0 = p_0$.

Za $k = 1, 2, \ldots$

izračunaj Ad_{k-1},
riješi $Mg_{k-1} = Ad_{k-1}$,
$\alpha_{k-1} = \dfrac{\langle p_{k-1}, g_{k-1}\rangle}{\langle g_{k-1}, g_{k-1}\rangle}$,
$x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$,
$r_k = r_{k-1} - \alpha_{k-1} Ad_{k-1}$,
riješi $Mp_k = r_k$,
izračunaj Ap_k,
riješi $Mq_k = Ap_k$,
$\beta_k = \dfrac{\langle q_k, g_{k-1}\rangle}{\langle g_{k-1}, g_{k-1}\rangle}$,
$d_k = p_k - \beta_k d_{k-1}$.

Ako je matrica A hermitska, tada bismo željeli da se i nakon prekondicioniranja to svojstvo zadrži. Zato ćemo uzeti da je i matrica prekondicioniranja M hermitska, pa ako je možemo faktorizirati kao $M = LL^*$, tada algoritam možemo primijeniti na sustav

$$L^{-1}AL^{-*}\hat{x} = L^{-1}b, \tag{2.41}$$

koji i dalje ostaje hermitski, a svojstvene vrijednosti matrice $L^{-1}AL^{-*}$ su iste kao i kod $M^{-1}A$. Ako sada označimo sve veličine vezane uz prekondicionirani sustav (2.41) sa $\hat{}$, a veličine vezane uz početni sustav $Ax = b$ sa standardnim oznakama, tada uz pomoć relacija

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

$$d_k = L^{-*}\hat{d}_k,$$

Orthomin(2) metodu primijenjenu na sustav (2.41) možemo transformirati u algoritam, koji ne ovisi o faktorizaciji matrice prekondicioniranja M.

Algoritam 2.3.9. HERMITSKI PREKONDICIONIRANI ORTHOMIN(2)

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

riješi $Mp_0 = r_0$,

$d_0 = p_0$.

Za $k = 1, 2, \ldots$

izračunaj Ad_{k-1},
riješi $Mg_{k-1} = Ad_{k-1}$,
$\alpha_{k-1} = \dfrac{\langle r_{k-1}, g_{k-1}\rangle}{\langle Ad_{k-1}, g_{k-1}\rangle}$,
$x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$,
$r_k = r_{k-1} - \alpha_{k-1} Ad_{k-1}$,
riješi $Mp_k = r_k$,
izračunaj Ap_k,
$\beta_k = \dfrac{\langle Ap_k, g_{k-1}\rangle}{\langle Ad_{k-1}, g_{k-1}\rangle}$,
$d_k = p_k - \beta_k d_{k-1}$.

2.4 Metoda najbržeg silaska, konjugirani smjerovi i konjugirani gradijenti(CG)

2.4.1 Metoda najbržeg silaska

Drugi način odabira parametra α_k u iteraciji (2.11) je takav da u k-tom koraku iteracije dobijemo grešku e_{k+1} sa minimalnom A-normom, pri čemu je A, matrica sustava $Ax = b$, pozitivno definitna. A-norma definira se kao $\|a\|_A = \sqrt{\langle a, a\rangle_A} = \sqrt{\langle Aa, a\rangle}$. Od sada pa na dalje, u ovom odjeljku, smatrat ćemo da je matrica sustava $Ax = b$ pozitivno definitna hermitska matrica.

Ponovo ćemo derivirati jednu funkciju po α_k, kao i kod Orthomin(1) metode, samo što se ovaj puta radi o funkciji $f(\alpha_k) = e_{k+1}^* A e_{k+1}$ ($f : \mathbb{R} \longrightarrow \mathbb{R}$). Njenu derivaciju ćemo izjednačiti je s nulom, pri čemu dobivamo e_{k+1} sa minimalnom A-normom. Razvoj ove metode, koju nazivamo *metodom najbržeg silaska*, puno detaljnije je opisan u [33], dok će u ovoj radnji biti izneseni samo osnovni elementi važni za razumijevanje. Ako uvrstimo da je $e_{k+1} = e_k - \alpha_k r_k$ i $r_{k+1} = r_k - \alpha_k A r_k$, što slijedi iz (2.11), dobit ćemo da je

$$f'(\alpha_k) = 2(\alpha_k r_k^* A r_k - r_k^* r_k),$$

odakle slijedi izraz za α_k

$$\alpha_k = \frac{r_k^* r_k}{r_k^* A r_k} = \frac{\langle r_k, r_k\rangle}{\langle Ar_k, r_k\rangle}. \tag{2.42}$$

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

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

odakle se vidi da se A-norma greške smanjuje u svakom koraku.

Analiza konvergencije za metodu najbržeg silaska je slična analizi greške za Orthomin(2) metodu. Iz jednakosti (2.11) možemo dobiti rekurziju za e_k, koja glasi

$$e_{k+1} = e_k - \alpha_k r_k = e_k - \alpha_k A e_k = (I - \alpha_k A) e_k,$$

koju na drugačiji način možemo zapisati kao

$$e_{k+1} = p_1(A) e_k,$$

pri čemu je p_1 polinom prvog stupnja, takav da je $p_1(\lambda) = 1 - \alpha_k \lambda$. Prema gore navedenome, za grešku e_{k+1} onda vrijedi

$$\|e_{k+1}\|_A = \min_{p_1 \in \mathbb{P}_1, p_1(0)=1} \|p(A) e_k\|_A, \tag{2.43}$$

gdje se minimum uzima po svim polinomima p prvog stupnja, za koje je $p(0) = 1$, odnosno koji su oblika $p(\lambda) = 1 - \alpha\lambda$. Za nastavak analize trebat će nam još sljedeće svojstvo A-norme za proizvoljni vektor u:

$$\|u\|_A = (u^* A u)^{1/2} = (u^* A^{1/2} A^{1/2} u)^{1/2} = ((A^{1/2}u)^* A^{1/2} u)^{1/2} = \|A^{1/2} u\|_2. \tag{2.44}$$

Ako simetričnu, pozitivno definitnu matricu A prikažemo kao $A = U\Lambda U^T$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, $U^*U = UU^* = I$, tada je $A^{1/2}$ simetrični kvadratni korijen matrice A, oblika $A^{1/2} = U\Lambda^{1/2}U^*$ koji komutira sa A i bilo kojim polinomom od A. $\lambda_1, \ldots, \lambda_n$ su svojstvene vrijednosti matrice A. Zato vrijedi

$$\begin{aligned} \|e_{k+1}\|_A &= \min_{p_1 \in \mathbb{P}_1, p_1(0)=1} \|A^{1/2} p_1(A) e_k\|_2 = \min_{p_1 \in \mathbb{P}_1, p_1(0)=1} \|U p_1(\Lambda) U^* A^{1/2} e_k\|_2 \leq \\ &\leq \min_{p_1 \in \mathbb{P}_1, p_1(0)=1} \|p_1(\Lambda)\|_2 \|e_k\|_A = \min_{p_1 \in \mathbb{P}_1, p_1(0)=1} \max_{j=1,\ldots,n} |p_1(\lambda_j)| \|e_k\|_A. \end{aligned} \tag{2.45}$$

Može se pokazati da se taj minimum postiže za polinom

$$p_1(\lambda) = 1 - \frac{2}{\lambda_{min} + \lambda_{max}} \lambda$$

i iznosi

$$|p_1(\lambda_{min})| = |p_1(\lambda_{max})| = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = \frac{\kappa(A) - 1}{\kappa(A) + 1} < 1,$$

pri čemu je λ_{min} minimalna, a λ_{max} maksimalna svojstvena vrijednost matrice A, te $\kappa(A) = \lambda_{max}/\lambda_{min}$ uvjetovanost matrice A. Dakle dobivamo rezultat

$$\|e_{k+1}\|_A \leq \left(\frac{\kappa(A) - 1}{\kappa(A) + 1} \right) \|e_k\|_A,$$

i kao konačni oblik

$$\|e_k\|_A \leq \left(\frac{\kappa(A) - 1}{\kappa(A) + 1} \right)^k \|e_0\|_A. \tag{2.46}$$

2.4.2 Metoda konjugiranih smjerova

Kao i kod izvoda metode Orthomin(2a), a u svrhu da metoda najbržeg silaska ne bi radila korake u smjeru kojim je neki raniji korak prošao, unaprijed odabiremo skup A-ortogonalnih vektora, odnosno smjerove traganja d_0, d_1,..., d_{n-1}. Dva vektora d_i i d_j su A-ortogonalna ili *konjugirana* ako vrijedi da je $\langle d_i, d_j\rangle_A = \langle Ad_i, d_j\rangle = 0$. Lagano se može provjeriti da su A-ortogonalni vektori linearno nezavisni. Znači u svakom koraku biramo točku

$$x_{k+1} = x_k + \alpha_k d_k \tag{2.47}$$

s minimalnom A-normom greške.

Dakle, u svakom smjeru d_k napravit ćemo točno jedan korak, i taj korak će biti takve dužine da ćemo poništiti komponentu vektora greške e_k u smjeru Ad_k. Nakon n koraka bit ćemo gotovi. U $(k+1)$-om koraku onda biramo e_{k+1} takav da bude jednak početnoj grešci, kojoj su odstranjene sve komponente u smjerovima Ad_0,...,Ad^k, odnosno on je A-ortogonalan na d_0,...,d_k. A-ortogonalnost između e_{k+1} i d_k je ekvivalentna nalaženju točke minimuma duž smjera traganja d_k, kao i u metodi najbržeg silaska. Da bi to vidjeli, ponovo ćemo derivirati po α_k funkciju $f(e_{k+1}) = e_{k+1}^* A e_{k+1}$ i izjednačiti je s nulom, samo što je u tom slučaju r_{k+1} okomit na d_k. Ako opet uvrstimo da je $r_{k+1} = r_k - \alpha_k A d_k$ i $e_{k+1} = e_k - \alpha_k d_k$, dobit ćemo izraz za α_k

$$\alpha_k = \frac{d_k^* r_k}{d_k^* A d_k} = \frac{\langle r_k, d_k\rangle}{\langle Ad_k, d_k\rangle}. \tag{2.48}$$

Ovako dobivena metoda naziva se metoda *konjugiranih smjerova*.

Metoda konjugiranih smjerova

- Dana je početna iteracija x_0, i skup A-ortogonalnih vektora $\{d_0, d_1, \ldots, d_{n-1}\}$.
- $r_0 = b - Ax_0$.
- Za $k = 1, 2, \ldots$
  - $\alpha_{k-1} = \langle r_{k-1}, d_{k-1}\rangle / \langle Ad_{k-1}, d_{k-1}\rangle$,
  - $x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$,
  - $r_k = r_{k-1} - \alpha_{k-1} A d_{k-1}$.

A-ortogonalnost e_{k+1} i d_0,...,d_k pokazat ćemo u sljedećem teoremu.

Teorem 2.4.1 ([20]). *Za metodu konjugiranih smjerova vrijede sljedeća svojstva:*

$$\langle Ad_i, d_j\rangle = 0 \quad (i \neq j) \tag{2.49}$$

$$\langle r_i, d_j\rangle = \langle Ae_i, d_j\rangle = 0 \quad (j < i) \tag{2.50}$$

$$\langle r_0, d_i\rangle = \langle r_1, d_i\rangle = \cdots = \langle r_i, d_i\rangle. \tag{2.51}$$

Skalar α_i može se zato napisati kao

$$\alpha_k = \frac{\langle r_0, d_k\rangle}{\langle Ad_k, d_k\rangle}. \tag{2.52}$$

Dokaz: Prva jednakost je očita jer smo tako birali smjerove traganja. Koristeći činjenicu da je zbog (2.47) $r_{i+1} = r_i - \alpha_i A d_i$ imamo

$$\langle r_{i+1}, d_j \rangle = \langle r_i, d_j \rangle - \alpha_i \langle A d_i, d_j \rangle.$$

Ako je $j = i$ tada, zbog (2.48), vrijedi $\langle r_{i+1}, d_i \rangle = 0$. Što više, zbog (2.49) je $\langle r_{i+1}, d_j \rangle = \langle r_i, d_j \rangle$, za $j \neq i$, pa jednakosti (2.50) i (2.51) slijede iz tih tih relacija matematičkom indukcijom. Na kraju, formula (2.52) slijedi iz (2.51) i (2.48). □

Poslijedica formule (2.52) je ta da se aproksimacije x_1,x_2,... od x mogu izračunati bez računanja reziduala r_1,r_2,..., odnosno osigurava da je izbor smjerova traganja d_1,d_2,... neovisan o tim rezidualima. Ovisan je samo o početnom rezidualu r_0. Takodjer zbog (2.50) vrijedi da je $\langle e_k, d_j \rangle_A = \langle A e_k, d_j \rangle = 0$ za $j < k$ čime smo dokazali napomenu koja je spomenuta prije teorema.

I kod ove metode nas interesira konvergencija, o čemu govori sljedeći teorem.

Teorem 2.4.2 ([20]). *Metoda konjugiranih smjerova je m-koračna metoda ($m \leq n$), u smislu da je u m-tom koraku aproksimacija x_m jednaka rješenju $x = A^{-1}b$.*

Dokaz: Neka je m najmanji cijeli broj takav da se $e_0 = x - x_0$ nalazi u prostoru razapetom sa d_0,...,d_{m-1}. Očito je $m \leq n$, budući da su vektori d_0,d_1,... linearno nezavisni, pa ih maksimalno može biti n. Zatim, izaberimo skalare a_0,...,a_{m-1} takve da je

$$e_0 = a_0 d_0 + \cdots + a_{m-1} d_{m-1}.$$

Odavde slijedi

$$x = x_0 + a_0 d_0 + \cdots + a_{m-1} d_{m-1}.$$

Nadalje,

$$r_0 = b - A x_0 = A(x - x_0) = a_0 A d_0 + \cdots + a_{m-1} A d_{m-1}.$$

Koristeći se činjenicom da su smjerovi traženja d_i medjusobno konjugirani i jednakošću (2.52) iz Teorema 2.4.1, računanjem skalarnog produkta $\langle r_0, d_i \rangle$ dobivamo

$$\langle r_0, d_i \rangle = a_i \langle A d_i, d_i \rangle$$

odnosno

$$a_i = \frac{\langle r_0, d_i \rangle}{\langle A d_i, d_i \rangle} = \alpha_i.$$

Budući da je, primjenom indukcije na (2.47)

$$x_m = x_0 + \alpha_0 d_0 + \cdots + \alpha_{m-1} d_{m-1},$$

možemo zaključiti da tvrdnja $x = x_m$ vrijedi. □

Kao i kod Orthomin(2a) skup A-ortogonalnih smjerova $\{d_i\}$ možemo dobiti uz pomoć Gramm–Schmidtove metode A-ortogonalizacije na niz linearno nezavisnih vektora u_0,...,u_{n-1} sa skalrnim produktom $\langle \cdot, \cdot \rangle_A$. Dakle u (2.26) koeficijenti su oblika

$$\beta_{ki} = -\frac{\langle A u_k, d_i \rangle}{\langle A d_i, d_i \rangle}. \tag{2.53}$$

2.4.3 Metoda konjugiranih gradijenata (CG)

Metoda konjugiranih smjerova (CG) je, zapravo, metoda konjugiranih smjerova kod koje se smjerovi traganja konstruiraju primjenom Gram–Schmidtove metode A-ortogonalnosti na reziduale, tj. uzima se da je $u_i = r_i$. Činjenica da su vektori r_i dobiveni metodom konjugiranih smjerova linearno nezavisni, može se provjeriti uz pomoć (2.50) i (2.51). Ponovo vrijedi

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

i budući da je r_k ortogonalan na prethodne smjerove traganja zbog (2.50), on je onda zbog prethodne tvrdnje, ortogonalan i na prethodne reziduale, odnosno vrijedi

$$\langle r_i, r_j \rangle = 0, \qquad i \neq j. \tag{2.54}$$

Promatramo sljedeći skalarni produkt

$$\langle r_k, r_{i+1} \rangle = \langle r_k, r_i \rangle - \bar{\alpha}_i \langle r_k, Ad_i \rangle,$$

pa odavde vrijedi

$$\langle Ar_k, d_i \rangle = \frac{1}{\bar{\alpha}_i}(\langle r_k, r_i \rangle - \langle r_k, r_{i+1} \rangle). \tag{2.55}$$

Za $i < k-1$ lijeva strana u (2.55) je jednaka 0, pa su $\beta_{ki} = 0$ za $i = 0, 1, \ldots, k-2$, a za $\beta_k = \beta_{k,k-1}$ vrijedi

$$\begin{aligned} \beta_k &= \frac{\langle r_k, r_k \rangle}{\langle d_{k-1}, r_{k-1} \rangle} = \qquad \text{zbog (2.48)}, \\ &= \frac{\langle r_k, r_k \rangle}{\langle r_{k-1}, r_{k-1} \rangle}, \qquad \text{zbog (2.50) i (2.26)}. \end{aligned} \tag{2.56}$$

Zbog (2.48), (2.50) i (2.26) možemo i α_k napisati u ljepšem obliku

$$\alpha_k = \frac{\langle r_k, r_k \rangle}{\langle Ad_k, d_k \rangle}, \tag{2.57}$$

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

Sada smo u potpunosti definirali metodu konjugiranih gradijenata čiji algoritam onda izgleda ovako

Algoritam 2.4.3. KONJUGIRANI GRADIJENTI

Dana je početna iteracija x_0,

$d_0 = r_0 = b - Ax_0$.

Za $k = 1, 2, \ldots$

izračunaj Ad_{k-1},

$$\alpha_{k-1} = \frac{\langle r_{k-1}, r_{k-1}\rangle}{\langle Ad_{k-1}, d_{k-1}\rangle},$$
$$x_k = x_{k-1} + \alpha_{k-1} d_{k-1},$$
$$r_k = r_{k-1} - \alpha_{k-1} Ad_{k-1},$$
$$\beta_k = \frac{\langle r_k, r_k\rangle}{\langle r_{k-1}, r_{k-1}\rangle},$$
$$d_k = r_k + \beta_k d_{k-1}.$$

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

Teorem 2.4.4 ([12]). *Greška* e_k *dobivena u* k*-tom koraku metode konjugiranih gradijenata ima najmanju* A*-normu na prostoru*

$$e_0 + span\{Ae_0, A^2e_0, \ldots, A^k e_0\}. \tag{2.58}$$

Dokaz: Ako gledamo kako su definirani vektori e_k i d_k u ovoj metodi, tada možemo zaključiti sljedeće: $d_0 = r_0 = Ae_0$, pa je $e_1 = e_0 - \alpha_0 r_0 = e_0 - \alpha_0 Ae_0$. Pretpostavimo da vrijedi da je $e_k \in \text{span}\{e_0, Ae_0, \ldots, A^k e_0\}$ i $d_{k-1} \in \text{span}\{Ae_0, A^2e_0, \ldots, A^k e_0\}$, tada za e_{k+1} imamo

$$e_{k+1} = e_k - \alpha_k d_k = e_k - \alpha_k \beta_k d_{k-1} - \alpha_k Ae_k \in \text{span}\{e_0, \ldots, A^{k+1}e_0\}.$$

Matematičkom indukcijom možemo zaključiti da je koeficijent uz e_0 uvijek jednak 1, odnosno

$$e_{k+1} \in e_0 + \text{span}\{Ae_0, A^2e_0, \ldots, A^{k+1}e_0\} \tag{2.59}$$

i zbog $d_k = Ae_k - \beta_k d_{k-1}$

$$d_k \in \text{span}\{Ae_0, A^2e_0, \ldots, A^{k+1}e_0\}. \tag{2.60}$$

Iz (2.60) slijedi da je $\text{span}\{d_0, d_1, \ldots, d_k\} = \text{span}\{Ae_0, \ldots, A^{k+1}e_0\}$, a kako je zbog (2.50) e_{k+1} A-ortogonalan na $\text{span}\{d_0, d_1, \ldots, d_k\}$, onda slijedi da je e_{k+1} vektor u prostoru (2.59) sa najmanjom A-normom. Za $k = n - 1$ zbog linearne nezavisnosti vektora $\{d_0, \ldots, d_{n-1}\}$ slijedi da je $e_n = 0$. □

U sljedećem teoremu promatramo promjenu euklidske norme greške, koja za metodu konjugiranih gradijenata ima dobra svojstva, u smislu da niz normi vektora greški predstavlja nerastući niz.

Teorem 2.4.5 ([20]). *U svakom koraku CG algoritma, duljina vektora greške* $e_k = x - x_k$ *se reducira, pri čemu je* $A^{-1}b = x = x_m$*, za neki* $m \leq n$.

Dokaz: Najprije dokažimo nekoliko činjenica koje će nam trebati u dokazu. Vrijedi:

$$\langle d_i, d_j\rangle = \frac{\langle d_i, d_i\rangle}{\langle r_i, r_i\rangle}\langle r_j, r_j\rangle \qquad (i \le j). \tag{2.61}$$

Da bismo to pokazali, dokažimo prvo jednu jednakost, koja glasi:

$$d_k = \langle r_k, r_k\rangle \sum_{j=0}^{k} \frac{r_j}{\langle r_j, r_j\rangle} \qquad k = 0, 1, 2, \ldots \tag{2.62}$$

Indukcijom iz jednakosti $d_k = r_k + \beta_k d_{k-1}$ dobijemo da je

$$d_k = r_k + \beta_k r_{k-1} + \beta_k\beta_{k-1} r_{k-2} + \cdots + \beta_k\beta_{k-1}\cdots\beta_1 r_0.$$

Uvrštavanjem vrijednosti za β_i iz (2.56) u prethodnu jednakost, i uz neophodna kračenja imamo

$$d_k = r_k + \frac{\langle r_k, r_k\rangle}{\langle r_{k-1}, r_{k-1}\rangle} r_{k-1} + \frac{\langle r_k, r_k\rangle}{\langle r_{k-2}, r_{k-2}\rangle} r_{k-2} + \cdots + \frac{\langle r_k, r_k\rangle}{\langle r_0, r_0\rangle} r_0.$$

Ako d_i i d_j, za $i \le j$, raspišemo kao u (2.62) i uz korištenje (2.54) dobivamo

$$\langle d_i, d_j\rangle = \langle r_i, r_i\rangle\langle r_j, r_j\rangle \sum_{l=0}^{i} \frac{1}{\langle r_l, r_l\rangle}$$

pa za $i = j$ imamo

$$\langle d_i, d_i\rangle = \langle r_i, r_i\rangle^2 \sum_{l=0}^{i} \frac{1}{\langle r_l, r_l\rangle}.$$

Prema tome je

$$\langle d_i, d_j\rangle = \frac{\langle r_j, r_j\rangle}{\langle r_i, r_i\rangle}\langle r_i, r_i\rangle^2 \sum_{k=0}^{i} \frac{1}{\langle r_k, r_k\rangle} = \frac{\langle r_j, r_j\rangle}{\langle r_i, r_i\rangle}\langle d_i, d_i\rangle.$$

Time je jednakost (2.61) dokazana.
Induktivnom primjenom jednakosti $x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$ imamo i

$$x_k = x_0 + \sum_{j=0}^{k-1} \alpha_j d_j.$$

Promotrimo sada sljedeće:

$$\begin{aligned}
&\langle e_{k-1}, e_{k-1}\rangle - \langle e_k, e_k\rangle = \\
&= \langle e_{k-1} - e_k, e_{k-1}\rangle + \langle e_k, e_{k-1} - e_k\rangle = \\
&= \langle x_k - x_{k-1}, e_{k-1}\rangle + \langle e_k, x_k - x_{k-1}\rangle = \\
&= \alpha_{k-1}\langle d_{k-1}, x_n - x_{k-1}\rangle + \alpha_{k-1}\langle x_n - x_k, d_{k-1}\rangle = \\
&= \alpha_{k-1}(\alpha_{k-1}\langle d_{k-1}, d_{k-1}\rangle + 2\alpha_k\langle d_k, d_{k-1}\rangle + \cdots + 2\alpha_{m-1}\langle d_{m-1}d_{k-1}\rangle) = \\
&= \frac{\|d_{k-1}\|_2^2}{\|r_{k-1}\|_2^2}(\alpha_{k-1}\|r_{k-1}\|_2^2 + 2\alpha_k\|r_k\|_2^2 + \cdots + 2\alpha_{m-1}\|r_{m-1}\|_2^2)
\end{aligned}$$

pri čemu je ovaj zadnji izraz veći od nule ukoliko do k-tog koraka nismo došli do rješenja.
□

2.4.4 Analiza greške i konvergencija metode konjugiranih gradijenata

Isto kao i kod Orthomin(2) metode, zbog relacije (2.59), greška u k-tom koraku metode ima oblik

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

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

$$e_k = p_k(A)e_0, \tag{2.63}$$

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

Kako je matrica A hermitska i pozitivno definitna, tada matricu možemo zapisati kao produkt matrica $A = U\Lambda U^*$, pri čemu su za $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, $\lambda_1, \ldots, \lambda_n$ svojstvene vrijednosti od A i $U^*U = UU^* = I$. $A^{1/2}$ je hermitski drugi korijen od A i vrijedi $A^{1/2} = U\Lambda^{1/2}U^*$, pa komutira sa A. Zbog toga slijedi

$$\begin{aligned}
\|e_k\|_A &= \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \|p_k(A)e_0\|_A = \\
&= \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \|A^{1/2}p_k(A)e_0\|_2 = \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \|Up_k(\Lambda)U^T A^{1/2}e_0\|_2 \leq \\
&\leq \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \|p_k(\Lambda)\|_2 \|e_0\|_A = \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \max_{i=1,\ldots,n} |p_k(\lambda_i)| \|e_0\|_A.
\end{aligned} \tag{2.64}$$

Daljnje opservacije vezane uz polinome i svojstvene vrijednosti su analogne onima kod Orthomin(2) metode. Sve svojstvene vrijednosti od A smještene su u segmentu $[\lambda_{min}, \lambda_{max}]$, pri čemu je $\lambda_{min} > 0$ najmanja, a $\lambda_{max} > 0$ najveća svojstvena vrijednost.

Budući da je desna strana nejednakosti (2.64) manja ili jednaka minimizaciji pa maksimizaciji po cijelom intervalu $[\lambda_{min}, \lambda_{max}]$, jednostavniji pristup je minimizacija po tom segmentu nego po konačnom broju točaka. Polinomi s kojima se to postiže ponovo su bazirani na Čebiševljevim polinomima. U ovom slučaju, polinom k-tog stupnja koji ima vrijednost 1 u ishodištu, i ima minimalnu ∞-normu na $[\lambda_{min}, \lambda_{max}]$ je dan sa

$$p_k(\lambda) = \frac{T_k\left(\frac{\lambda_{max}+\lambda_{min}-2\lambda}{\lambda_{max}-\lambda_{min}}\right)}{T_k\left(\frac{\lambda_{max}+\lambda_{min}}{\lambda_{max}-\lambda_{min}}\right)}. \tag{2.65}$$

Argument brojnika je takav da segment $[\lambda_{min}, \lambda_{max}]$ prebacuje u segment $[-1, 1]$. Slijedi da za $z \in [\lambda_{min}, \lambda_{max}]$ apsolutna vrijednost brojnika iz (2.65) je ograničena sa 1. Nazivnik je određen tako da zadovolji svojstvo $p_k(0) = 1$. Znači, imamo sljedeće

$$\|e_k\|_A \leq T_k\left(\frac{\lambda_{max}+\lambda_{min}}{\lambda_{max}-\lambda_{min}}\right)^{-1} \|e_0\|_A = T_k\left(\frac{\kappa(A)+1}{\kappa(A)-1}\right)^{-1} \|e_0\|_A,$$

i nakon što izračunamo vrijednost Čebiševljevog polinoma, dobivamo

$$\|e_k\|_A \leq 2\left[\left(\frac{\sqrt{\kappa(A)}+1}{\sqrt{\kappa(A)}-1}\right)^k + \left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^k\right]^{-1} \|e_0\|_A \tag{2.66}$$

odnosno

$$\|e_k\|_A \leq 2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^k \|e_0\|_A. \tag{2.67}$$

pri čemu je $\kappa(A) = \lambda_{max}/\lambda_{min}$ broj uvjetovanosti matrice A.

2.4.5 Prekondicionirana CG metoda

Kao i kod Orthomin(2) metode kada rješavamo hermitski sustav, želimo da se to svojstvio zadrži i nakon prekondicioniranja. Zato ponovo biramo matricu prekondicioniranja M koja je hermitska. Ako matricu M faktoriziramo kao $M = LL^*$, tada algoritam možemo primijeniti na sustav

$$L^{-1}AL^{-*}\hat{x} = L^{-1}b, \tag{2.68}$$

koji i dalje ostaje hermitski, a svojstvene vrijednosti matrice $L^{-1}AL^{-*}$ su iste kao i kod $M^{-1}A$. Matricu M također biramo i uz kriterij da $L^{-1}AL^{-*}$ ima što manju uvjetovanost. Ako sada označimo sve veličine vezane uz prekondicionirani sustav (2.68) sa $\hat{}$, a veličine vezane uz početni sustav $Ax = b$ sa standardnim oznakama, tada uz pomoć relacija

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

CG metodu primijenjenu na sustav (2.68) možemo transformirati u algoritam, koji ne ovisi o faktorizaciji matrice prekondicioniranja M.

Algoritam 2.4.6. HERMITSKI PREKONDICIONIRANI KONJUGIRANI GRADIJENTI

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

riješi $Mp_0 = r_0$,

$d_0 = p_0$.

Za $k = 1, 2, \ldots$

izračunaj Ad_{k-1},

$$\alpha_{k-1} = \frac{\langle r_{k-1}, p_{k-1}\rangle}{\langle Ad_{k-1}, d_{k-1}\rangle},$$

$x_k = x_{k-1} + \alpha_{k-1}d_{k-1}$,

$r_k = r_{k-1} - \alpha_{k-1}Ad_{k-1}$,

riješi $Mp_k = r_k$,

$$\beta_k = \frac{\langle r_k, p_k\rangle}{\langle r_{k-1}, p_{k-1}\rangle},$$

$d_k = p_k + \beta_k d_{k-1}$.

2.5 GMRES

2.5.1 Razvoj i implementacija GMRES metode

GMRES metoda (Generalized minimal residual algorithm) ili generalizirani algoritam minimalnog reziduala koristi modificirani Gram–Schmidtov postupak kako bi konstru-

irao ortonormiranu bazu za niz Krylovljevih potprostora $\text{span}\{r_0, Ar_0, \ldots, A^k r_0\}$. Kada se modificirani Gram–Schmidtov postupak primijeni na ovakav prostor naziva se *Arnoldijeva metoda*.

Algoritam 2.5.1. ARNOLDIJEV ALGORITAM

Dan je vektor q_1 *sa* $\|q_1\|_2 = 1$.

Za $j = 1, 2, \ldots, n-1$

$\tilde{q}_{j+1} = Aq_j$.

Za $i = 1, \ldots, j$

$h_{i,j} = \langle \tilde{q}_{j+1}, q_i \rangle$,

$\tilde{q}_{j+1} := \tilde{q}_{j+1} - h_{i,j} q_i$.

$h_{j+1,j} = \|\tilde{q}_{j+1}\|_2$,

$q_{j+1} = \dfrac{\tilde{q}_{j+1}}{h_{j+1,j}}$.

Ako je Q_k $n \times k$ matrica čije stupce čine vektori ortonormirane baze q_1,...,q_k, tada se Arnoldijeva iteracija može napisati u matričnom obliku

$$AQ_k = Q_k H_k + h_{k+1,k} q_{k+1} \xi_k^T = Q_{k+1} H_{k+1,k}. \tag{2.69}$$

Ovdje je H_k $k \times k$ gornja Hessenbergova matrica kojoj je (i,j)-ti element jednak $h_{i,j}$ za $j = 1$,...,k, $i = 1$,...,$\min\{j+1, k\}$, a svi ostali elementi su jednaki nuli. Vektor ξ_k je k-ti jedinični vektor $[0 \ \ldots \ 0 \ 1]^T$. $(k+1) \times k$ matrica $H_{k+1,k}$ je matrica kojoj je gornji $k \times k$ blok jednak H_k, a zadnji redak jednak nuli, osim na poziciji $(k+1, k)$, na kojoj je element jednak $h_{k+1,k}$. Rekurzivna definicija za matricu vektora ortonormirane baze $(k+1)$-dimenzionalnog Krylovljevog potprostora Q_{k+1} može se u matričnom obliku napisati kao

$$[q_1 \quad AQ_k] = Q_{k+1} R_{k+1},$$

pri čemu je R_{k+1} $(k+1) \times (k+1)$ gornje trokutasta matrica

$$R_{k+1} = [\xi_1 \quad H_{k+1,k}],$$

a $\xi_1 = [1, 0, \ldots, 0]^T$. Prema tome Arnoldijev algoritam od prvog do $(k+1)$-og koraka možemo smatrati rekurzivnom QR faktorizacijom matrice $[q_1 \quad AQ_k]$.

U GMRES metodi, aproksimacija rješenja u k-tom koraku tada ima oblik $x_k = x_0 + Q_k y_k$ za neki k-dimenzionalni vektor y_k, tj. x_k se dobiva tako da se početnoj aproksimaciji x_0 doda neka linearna kombinacija vektora ortonormirane baze Krylovljevog potprostora. Da bi dobili aproksimaciju za koju $r_k = r_0 - AQ_k y_k$ ima minimalnu euklidsku normu, vektor y_k mora biti rješenje problema najmanjih kvadrata

$$\min_{y \in \mathbb{C}^k} \|r_0 - AQ_k y\|_2 = \min_{y \in \mathbb{C}^k} \|r_0 - Q_{k+1} H_{k+1,k} y\|_2 =$$

$$= \min_{y \in \mathbb{C}^k} \|Q_{k+1}(\beta \xi_1 - H_{k+1,k} y)\|_2 = \min_{y \in \mathbb{C}^k} \|\beta \xi_1 - H_{k+1,k} y\|_2, \tag{2.70}$$

gdje je $\beta = \|r_0\|_2$, ξ_1 prvi jedinični $(k+1)$-dimenzionalni vektor $[1\ 0\ \ldots\ 0]^T$. Druga jednakost je ostvarena uz korištenje činjenice da je $Q_{k+1}\xi_1$ prvi vektor ortonormirane baze koji je jednak r_0/β.

Glavni koraci GMRES algoritma su sljedeći:

- *Dana je početna iteracija x_0, izračunaj $r_0 = b - Ax_0$ i $q_1 = r_0/\|r_0\|_2$.*
- *Za k=1,2,...*
  - *Izračunaj q_{k+1} i $h_{i,k}$, $i = 1, \ldots, k+1$ pomoću Arnoldijevog algoritma.*
  - *Nađi $x_k = x_0 + Q_k y_k$, gdje je y_k rješenje problema najmanjih kvadrata $\min_y \|\beta\xi_1 - H_{k+1,k}y\|_2$.*

Standardna metoda za rješavanje problema najmanjih kvadrata (2.70) je faktoriziranje $(k+1) \times k$ matrice $H_{k+1,k}$ na produkt $(k+1) \times (k+1)$ unitarne matrice $F^{(k)*}$ i $(k+1) \times k$ gornje trokutaste matrice $R^{(k)}$. Gornji $k \times k$ blok matrice $R^{(k)}$ je gornje trokutasti, a zadnji redak je jednak nul-vektoru. Ta faktorizacija, a ponovo se radi o QR faktorizaciji, može se ostvariti upotrebom Givensovih rotacija ili Hauseholderovih refleksija. Budući da je

$$\|\beta\xi_1 - H_{k+1,k}y\|_2 = \|F^{(k)*}(\beta F^{(k)}\xi_1 - R^{(k)}y)\|_2 = \|\beta F^{(k)}\xi_1 - R^{(k)}y\|_2,$$

i da je $(k+1)$-a koordinata vektora $R^{(k)}y$ jednaka nuli, rješenje y_k se tada može dobiti rješavanjem gornje trokutastog sustava

$$R^{(k)}_{k\times k}y = \beta(F^{(k)}\xi_1)_{k\times 1}, \tag{2.71}$$

gdje je $R^{(k)}_{k\times k}$ gornji $k \times k$ blok od $R^{(k)}$, a $(F^{(k)}\xi_1)_{k\times 1}$ je dio od k gornjih elemenata prvog stupca od $F^{(k)}$.

Važno je primijetiti da je u tom slučaju apsolutna vrijednost zadnjeg elementa $(k+1)$-dimenzionalnog vektora $\beta F^{(k)}\xi_1$ jednaka euklidskoj normi reziduala u k-tom koraku GMRES metode jer

$$\|b - Ax_k\|_2 = \|\beta F^{(k)}\xi_1 - R^{(k)}y_k\|_2,$$

a $\beta F^{(k)}\xi_1 - R^{(k)}y_k$ je jednak nuli u svim komponentama osim u zadnjoj, $(k+1)$-oj, koja je jednaka zadnjem elementu od $\beta F^{(k)}\xi_1$.

Promatrat ćemo sada slučaj kada se QR faktorizacija ostvaruje pomoću Givensovih rotacija. Ako nam je dana QR faktorizacija matrice $H_{k+1,k}$, tada nam je namjera da izračunamo QR faktorizaciju sljedeće matrice $H_{k+2,k+1}$ sa što manje utrošenog posla. Da bi to postigli najprije označimo sa F_i matricu rotacije koja rotira ravninu razapetu sa jediničnim vektorima ξ_i i ξ_{i+1} za kut θ_i:

$$F_i = \begin{bmatrix} I & & & \\ & c_i & s_i & \\ & -\bar{s}_i & c_i & \\ & & & I \end{bmatrix},$$

gdje je $c_i = \cos(\theta_i)$ i $s_i = \sin(\theta_i)$. Dimenzija matrice F_i kao i dimenzija drugog identičnog bloka ovisi o kontekstu u kojem se koristi. Pretpostavimo da su rotacije F_i, $i = 1, \ldots, k$

prethodno bile upotrebljene na matrici $H_{k+1,k}$ tako da je

$$(F_kF_{k-1}\cdots F_1)H_{k+1,k} = R^{(k)} = \begin{bmatrix} x & x & \dots & x \\ & x & \dots & x \\ & & \ddots & \vdots \\ & & & x \\ 0 & 0 & \dots & 0 \end{bmatrix},$$

gdje x-evi označavaju netrivijalne elemente. Tada je $F^{(k)} = F_kF_{k-1}\cdots F_1$. Da bismo dobili matricu $R^{(k+1)}$, odnosno gornje trokutasti faktor od $H_{k+2,k+1}$, prvo trebamo izmnožiti zadnji stupac od $H_{k+2,k+1}$ sa prethodnim rotacijama jer, kao što smo već prije napomenuli, Arnoldijev algoritam možemo smatrati rekurzivnom QR faktorizacijom matrice $[q_1 \quad AQ_k]$ kojoj broj stupaca raste u svakom koraku algoritma. Ako smo izračunali QR faktorizaciju za takvu matricu u k-tom koraku, tada u $(k+1)$-om koraku moramo izračunati QR faktorizaciju matrice koja je jednaka prethodnoj matrici ali ima još jedan dodatan stupac. Trokutasti faktor matrice u $(k+1)$-om koraku je zbog toga jednak trokutastom faktoru matrice iz k-tog koraka, samo kojemu je također dodan još jedan stupac, pa zato je dovoljno obraditi samo novi, $(k+1)$-i stupac matrice $H_{k+2,k+1}$. Time dobivamo

$$(F_kF_{k-1}\cdots F_1)H_{k+2,k+1} = \begin{bmatrix} x & x & \dots & x & x \\ & x & \dots & x & x \\ & & \ddots & \vdots & \vdots \\ & & & x & x \\ 0 & 0 & \dots & 0 & d \\ 0 & 0 & \dots & 0 & h \end{bmatrix},$$

gdje je $(k+2, k+1)$-i element h upravo $h_{k+2,k+1}$ budući da na taj element ne utječu prethodne rotacije. Sljedeća rotacija F_{k+1} se bira tako da eliminira taj element, odnosno da vrijedi $-\bar{s}_{k+1}d + c_{k+1}h = 0$, a to možemo postići ako stavimo da je

$$c_{k+1} = \frac{|d|}{\sqrt{|d|^2+|h|^2}}, \quad \bar{s}_{k+1} = \frac{c_{k+1}h}{d}, \qquad \text{za } d \neq 0,\ h \neq 0,$$

$$c_{k+1} = 0, \quad s_{k+1} = 1, \qquad \text{za } d = 0,$$

$$(c_{k+1} = 1, \quad s_{k+1} = 0, \qquad \text{za } h = 0).$$

Primijetimo da ako je $h = 0$, tada je egzaktno rješenje linearnog sustava dostignuto u $(k+1)$-om koraku. Naime, u tom slučaju je $F_{k+1} = I$, pa su prvih $k+1$ komponenti $(k+2)$-dimenzionalnog vektora $\beta F^{(k+1)}\xi_1$ jednake $(k+1)$-dimenzionalnom vektoru $\beta F^{(k)}\xi_1$ iz prethodnog koraka, a zadnja, $(k+2)$-a komponenta je ostala nepromijenjena, odnosno jednaka nuli. Budući da je apsolutna vrijednost te zadnje komponente jednaka euklidskoj normi reziduala u tom koraku GMRES metode, zaključujemo da je $r_{k+1} = 0$, i rješenje je dostignuto u $(k+1)$-om koraku. Prema tome, ako egzaktno rješenje još nije izračunato, tada je $h \neq 0$, i $(k+1)$-ti dijagonalni element od $R^{(k+1)}$ je različit od nule. Za $d \neq 0$ taj dijagonalni element je jednak

$$c_{k+1}d + s_{k+1}h = \frac{d}{|d|}\sqrt{|d|^2+|h|^2},$$

dok za $d = 0$ on je jednak h.

GMRES algoritam može se napisati u sljedećem obliku.

Algoritam 2.5.2. GMRES

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

$\beta = \|r_0\|_2$,

$q_1 = \dfrac{r_0}{\beta}$,

$l = (1, 0, \ldots, 0)^T$.

Za $k = 1, 2, \ldots$

Izračunaj q_{k+1} *i* $h_{i,k} = H(i,k)$ *za* $i = 1, \ldots, k+1$, *koristeći Arnoldijev algoritam.*

Primijeni $F_1, \ldots, F_{k-1}$ *na zadnji stupac od* H, *odnosno:*

Za $i = 1, \ldots, k-1$

$$\begin{bmatrix} H(i,k) \\ H(i+1,k) \end{bmatrix} := \begin{bmatrix} c_i & s_i \\ -\bar{s}_i & c_i \end{bmatrix} \begin{bmatrix} H(i,k) \\ H(i+1,k) \end{bmatrix}.$$

Izračunaj k*-tu Givensovu rotaciju* F_k *kako bi se poništio* $(k+1,k)$ *element od* H*:*

$$c_k = \frac{|H(k,k)|}{\sqrt{|H(k,k)|^2 + |H(k+1,k)|^2}},$$

ako je $c_k \neq 0$ *tada* $s_k = c_k \dfrac{\overline{H(k+1,k)}}{\overline{H(k,k)}}$, *ako je* $c_k = 0$ *tada* $s_k = 1$.

Primijeni k*-tu rotaciju na* l *i na zadnji stupac od* H*:*

$$\begin{bmatrix} l(k) \\ l(k+1) \end{bmatrix} := \begin{bmatrix} c_k & s_k \\ -\bar{s}_k & c_k \end{bmatrix} \begin{bmatrix} l(k) \\ 0 \end{bmatrix},$$

$H(k,k) := c_k H(k,k) + s_k H(k+1,k)$,

$H(k+1,k) = 0$.

Ako je ocjena norme reziduala $\beta|l(k+1)|$ *dovoljno mala, tada:*

Riješi gornje trokutasti sustav $H_{k\times k} y_k = \beta l_{k\times 1}$.

Izračunaj $x_k = x_0 + Q_k y_k$.

Potpun GMRES algoritam može biti nepraktičan zbog velikog zahtjeva za memorijom i za brojem operacija, ako je broj iteracija, koje su potrebne za rješavanje sustava, velik. GMRES(j) algoritam se definira tako da restartamo GMRES svakih j koraka, koristeći zadnju iteraciju kao početnu za sljedeći GMRES ciklus. Poznati problem sa GMRES-om sa restartom je taj što on može stagnirati. Naime može se dogoditi da pri restartanju ponovno pretražujemo smjerove koje smo već u prethodnom ciklusu pretraživali.

2.5.2 Svojstva GMRES metode

Iz same implementacije metode vidljiva su sljedeća svojstva GMRES metode.

Teorem 2.5.3 ([32]). *Neka su F_i, $i = 1, \ldots, k$ matrice rotacija koje su korištene za svođenje matrice $H_{k+1,k}$ na trokutasti oblik, te $R^{(k)} = F^{(k)} H_{k+1,k}$ i $g^{(k)} = \beta F^{(k)} \xi_1$ matrica i desna strana sustava dobivenog tokom rješavanja GMRES metodom. Označimo sa R_k $k \times k$ gornje trokutastu matricu dobivenu iz $R^{(k)}$ brisanjem zadnjeg retka, i sa g_k k-dimenzionalan vektor dobiven iz $g^{(k)}$ brisanjem zadnje komponente. Tada,*

(i) Rang od AQ_k je jednak rangu od R_k. Posebno, ako je $(R_k)_{k,k} = 0$ tada A mora biti singularna.

(ii) Vektor y_k koji minimizira $\|\beta\xi_1 - H_{k+1,k} y\|_2$ dan je sa

$$y_k = R_k^{-1} g_k.$$

(iii) Rezidual u k-tom koraku zadovoljava

$$r_k = b - Ax_k = Q_{k+1}(\beta\xi_1 - H_{k+1,k} y_k) = Q_{k+1} F^{(k)*}((g^{(k)})_{k+1} \xi_{k+1}) \tag{2.72}$$

i, kao rezultat

$$\|r_k\|_2 = \|b - Ax_k\|_2 = |(g^{(k)})_{k+1}|. \tag{2.73}$$

Dokaz: (i) Iz (2.69) dobivamo jednakost

$$AQ_k = Q_{k+1} H_{k+1,k} = Q_{k+1} F^{(k)*} F^{(k)} H_{k+1,k} = Q_{k+1} F^{(k)*} R^{(k)}.$$

Budući da je $Q_{k+1} F^{(k)*}$ unitarna, rang od AQ_k je jednak rangu od $R^{(k)}$, koji je opet jednak rangu od R_k, budući da se te dvije matrice razlikuju samo za zadnji nul–redak od $R^{(k)}$. Ako je pak $(R_k)_{k,k} = 0$ tada je rang od R_k manji ili jednak $k-1$, a kao rezultat je i rang od AQ_k manji ili jednak $k-1$. Kako je Q_k punog ranga, to znači da je A singularna.

(ii) To je već uglavnom pokazano kod implementacije. Za svaki vektor y imamo

$$\begin{aligned} \|\beta\xi_1 - H_{k+1,k} y\|_2^2 &= \|F^{(k)}(\beta\xi_1 - H_{k+1,k} y)\|_2^2 = \|g^{(k)} - R^{(k)} y\|_2^2 = \\ &= |(g^{(k)})_{k+1}|^2 + \|g_k - R_k y\|_2^2 \end{aligned} \tag{2.74}$$

Minimum lijeve strane se postiže kad je drugi izraz desne strane od (2.74) jednak nuli. Ako smo krenuli od regularne matrice A tada je i R_k regularna, pa se to postiže kada je $y = R_k^{-1} g_k$.

(iii) Za bilo koji $x = x_0 + Q_k y$ vrijedi

$$\begin{aligned} b - Ax &= Q_{k+1}(\beta\xi_1 - H_{k+1,k} y) = Q_{k+1} F^{(k)*} F^{(k)}(\beta\xi_1 - H_{k+1,k} y) = \\ &= Q_{k+1} F^{(k)*}(g^{(k)} - R^{(k)} y). \end{aligned}$$

Kao što smo vidjeli u dokazu *(ii)*, euklidska norma od $g^{(k)} - R^{(k)} y$ postiže svoj minimum kada y poništi sve komponente od $g^{(k)}$ osim zadnje, koja je jednaka

$(g^{(k)})_{k+1}$. Kako je zadnji redak od $R^{(k)}$ jednak nul-retku, to znači da je i zadnja komponenta od $R^{(k)}y$ jednaka nuli, pa kao rezultat dobivamo

$$b - Ax_k = Q_{k+1}F^{(k)*}((g^{(k)})_{k+1}\xi_{k+1})$$

što dokazuje (2.72). Jednakost (2.73) slijedi iz ortonormiranosti stupaca matrice $Q_{k+1}F^{(k)*}$.

□

Primijetimo da, kada vektor $\beta\xi_1$ izmnožimo redom sa $F_1, F_2, \ldots, F_{k-1}$, zadnja komponenta tako dobivenog vektora $g^{(k,k-1)}$ ostaje nepromijenjena, odnosno jednaka nuli. To znači da je on oblika $g^{(k,k-1)} = [\gamma_1\ \gamma_2\ \ldots\ \gamma_k\ 0]^T$. U zadnjem koraku, kada taj vektor na kraju još pomnožimo i sa F_k, dobivamo da je $g^{(k)} = [\gamma_1\ \gamma_2\ \ldots\ c_k\gamma_k\ -s_k\gamma_k]$, odnosno dobivamo korisnu jednakost za $(g^{(k)})_{k+1}$ koji se pojavljuje u izrazu za rezidual u k-tom koraku:

$$(g^{(k)})_{k+1} = -s_k\gamma_k. \tag{2.75}$$

Posebno, ako je $s_k = 0$ tada norma reziduala mora biti jednaka nuli, što znači da je egzaktno rješenje dostignuto u k-tom koraku.

Ako proučimo Algoritam 2.5.2, tada vidimo da je jedini mogući slučaj kada k-ti korak neće biti izvediv, i kada će doći do prekida GMRES metode, je u Arnoldijevom algoritmu za $\tilde{q}_{k+1} = 0$, odnosno $h_{k+1,k} = 0$. U tom slučaju se algoritam zaustavlje jer se sljedeći Arnoldijev vektor neće moći generirati zbog dijeljenja s nulom. Ali, u tom slučaju, kao što smo već vidjeli, rezidual je jednak nuli, pa smo dostigli egzaktno rješenje u tom koraku. Obrat je također istinit: ako se algoritam zaustavi u k-tom koraku sa $r_k = b - Ax_k = 0$, tada je $h_{k+1,k} = 0$.

Teorem 2.5.4 ([32]). *Neka je A regularna matrica. Tada se GMRES algoritam prekida u k-tom koraku ($h_{k+1,k} = 0$) ako i samo ako je aproksimacija x_k jednaka egzaktnom rješenju.*

Dokaz: Nužnost smo već pokazali, ali promotrimo tu situaciju iz drugog ugla. Pretpostavimo da je $h_{k+1,k} = 0$, tada je $s_k = 0$. Zaista, ako definiramo $h_{k,k}^{(k-1)} = (F_{k-1}F_{k-2}\cdots F_1H_{k+1,k})_{k,k}$, tada je zapravo množenje sa F_k suvišno, pa je već u $(k-1)$-om koraku dobivena trokutasta forma za $H_{k+1,k}$, odnosno $(R_k)_{k,k} = h_{k,k}^{(k-1)}$. Ako ipak krenemo računati F_k tada, budući da je A regularna, $(R_k)_{k,k}$ je različito od nule zbog Teorema 2.5.3 (i), pa iz formule $s_k = \text{sign}(h_{k,k}^{(k-1)})h_{k+1,k}/\sqrt{|h_{k,k}^{(k-1)}|^2 + |h_{k+1,k}|^2}$ dobijemo da je s_k definirano i jednako nuli. Tada iz jednakosti (2.73) i (2.75) dobivamo da je $r_k = 0$.

Da bi dokazali dovoljnost tvrdnje, pretpostavljamo da je $r_k = 0$ i ponovo koristimo (2.75). Budući da smo u k-tom koraku dostigli egzaktno rješenje, a ne u $(k-1)$-tom, znači da je $(g^{(k,k-1)})_k = \gamma_k$ različito od nule, pa onda mora biti $s_k = 0$. Ponovo iz gornje formule za s_k dobivamo da je $h_{k+1,k} = 0$. □

Pogledajmo još samo kako zaista izgleda $(k+1)$-dimenzionalan vektor $g^{(k)} = \beta F^{(k)}\xi_1$. Jednostavno, indukcijom može se pokazati da vrijedi

$$g^{(k)} = \begin{bmatrix} c_1 \\ -\bar{s}_1c_2 \\ \bar{s}_1\bar{s}_2c_3 \\ \vdots \\ (-1)^{k-1}\bar{s}_1\cdots\bar{s}_{k-1}c_k \\ (-1)^k\bar{s}_1\cdots\bar{s}_{k-1}\bar{s}_k \end{bmatrix},$$

pa se prema (2.73) norma reziduala egzaktno može dobiti sa

$$\|r_k\|_2 = \beta|s_1 s_2 \cdots s_k|. \tag{2.76}$$

Prije nego što počnemo promatrati konvergenciju GMRES metode, zgodno bi bilo promotriti situacije kada ona stagnira. Naime iz minimizacijskog svojstva metode vidljivo je da norma reziduala neće rasti kako povećavamo broj iteracija, jer promatramo minimum na sve većem i većem skupu. Sljedeći teorem govori o tome kada se ta norma neće smanjivati.

Teorem 2.5.5 ([4]). *Pretpostavimo da su izvršena k koraka Arnoldijevog algoritma, i da je matrica H_k, definirana u (2.69) kao $AQ_k = Q_k H_k + h_{k+1,k} q_{k+1} \xi_1^T$, singularna. Tada*

$$\min_{y \in \mathbb{C}^k} \|\beta\xi_1 - H_{k+1,k} y\|_2 = \min_{y \in \mathbb{C}^{k-1}} \|\beta\xi_1 - H_{k,k-1} y\|_2. \tag{2.77}$$

Ako sa y_k označimo rješenje lijeve strane od (2.77), a sa y_{k-1} rješenje desne strane, tada je $y_k = [y_{k-1}^T \ 0]^T$, odakle slijedi da je $x_k = x_{k-1}$. Obrnuto, pretpostavimo da je izvršeno k koraka Arnoldijevog algoritma i da vrijedi (2.77). Tada je H_k singularna.

Dokaz: Prvo primijetimo, da ako postoji mogućnost za izvođenjem k-tog koraka Arnoldijevog algoritma, tada mora biti $h_{j+1,j} \neq 0$ za $j = 1, \ldots, k-1$. Inače bi se, prema Teoremu 2.5.4 Arnoldijev algoritam zaustavio u nekom ranijem koraku, dostižući egzaktno rješenje. U tom slučaju bi bilo $h_{j+1,j} = 0$, i $AQ_j = Q_j H_j$, što povlači da zbog toga što je AQ_j punog ranga, H_j mora biti regularna matrica. Dakle, ako pretpostavimo da je H_k singularna, tada mora biti $h_{k+1,k} \neq 0$. Također, kada je matrica H_k singularna, dimenzija nul–potprostora (jezgre) je jednaka 1, zbog toga što je ona gornje Hessenbergova matrica koja na donjoj sporednoj dijagonali ima sve elemente različite od nule ($h_{j+1,j} \neq 0$ za $j = 1, \ldots, k-1$), što znači da H_k u slici ima $k-1$ linearno nezavisnih vektora.

Dalje promatramo QR faktorizaciju matrica $H_{k+1,k}$ i $H_{k,k-1}$. Neka su

$$H_{k+1,k} = F^{(k)*} R^{(k)} \quad \text{i} \quad H_{k,k-1} = F^{(k-1)*} R^{(k-1)},$$

QR faktorizacije, kod kojih su

$$R^{(k)} = \begin{bmatrix} R_k \\ 0 \end{bmatrix} \quad \text{i} \quad R^{(k-1)} = \begin{bmatrix} R_{k-1} \\ 0 \end{bmatrix}.$$

Primijetimo da je

$$H_{k+1,k} = \left[\begin{array}{c|c} & h_{1,k} \\ H_{k,k-1} & \vdots \\ & h_{k,k} \\ \hline 0 \cdots 0 & h_{k+1,k} \end{array}\right] = \begin{bmatrix} H_{k,k-1} & h_k \\ 0 & h_{k+1,k} \end{bmatrix}$$

i $H_k = [H_{k,k-1} \ h_k]$, za $h_k = [h_{1,k} \ \ldots \ h_{k,k}]^T$. Budući da je H_k singularna, a da $H_{k,k-1}$ zbog prije iznešene opservacije ima puni rang, slijedi da je zadnji stupac matrice H_k, kojeg smo označili sa h_k, linearna kombinacija prvih $k-1$. Znači, postoji neki $z \in \mathbb{C}^{k-1}$ takav da je $h_k = H_{k,k-1} z$. Zbog toga imamo

$$h_k = F^{(k-1)*} R^{(k-1)} z = F^{(k-1)*} \begin{bmatrix} R_{k-1} z \\ 0 \end{bmatrix}. \tag{2.78}$$

Nadalje, možemo napisati da je $F^{(k)} = F_k \widetilde{F}^{(k-1)}$, gdje su

$$F_k = \begin{bmatrix} I_{k-1} & & \\ & c_k & s_k \\ & -\bar{s}_k & c_k \end{bmatrix} \quad \text{i} \quad \widetilde{F}^{(k-1)} = \begin{bmatrix} F^{(k-1)} & 0 \\ 0 & 1 \end{bmatrix}.$$

Zbog toga je

$$F_k \widetilde{F}^{(k-1)} H_{k+1,k} = R^{(k)}$$

ili

$$F_k \begin{bmatrix} F^{(k-1)} H_{k,k-1} & F^{(k-1)} h_k \\ 0 & h_{k+1,k} \end{bmatrix} = R^{(k)}. \tag{2.79}$$

Iz (2.78) imamo da je $F^{(k-1)} h_k = [(R_{k-1} z)^T \ 0]^T$. Korištenjem ove jednakosti, uz činjenicu da je $F^{(k-1)} H_{k,k-1} = R^{(k-1)}$, dobivamo jednakost

$$F_k \begin{bmatrix} R^{(k-1)} & \begin{bmatrix} R_{k-1} z \\ 0 \end{bmatrix} \\ 0 \cdots 0 & h_{k+1,k} \end{bmatrix} = R^{(k)}.$$

Zbog toga su c_k i s_k izabrani tako da

$$\begin{bmatrix} c_k & s_k \\ -\bar{s}_k & c_k \end{bmatrix} \begin{bmatrix} 0 \\ h_{k+1,k} \end{bmatrix} = \begin{bmatrix} \pm h_{k+1,k} \\ 0 \end{bmatrix}.$$

Vrijednosti c_k i s_k tada moraju zadovoljavati $c_k = 0$ i $s_k = \pm 1$. Sada iz (2.76) slijedi

$$\begin{aligned} \|r_k\|_2 = \|r_0 - AQ_k y_k\|_2 &= \beta |s_1 \cdots s_{k-1} s_k| = \beta |s_1 \cdots s_{k-1}| = \\ &= \|r_0 - AQ_{k-1} y_{k-1}\|_2 = \|r_{k-1}\|_2. \end{aligned}$$

Zbog jedinstvenosti rješenja problema najmanjih kvadrata, (jer je R_k regularna) y_k je prema Teoremu 2.5.3 jedinstven, pa mora biti $y_k = [y_{k-1}^T \ 0]^T$. Time, također, ispada da je $x_k = x_0 + Q_{k-1} y_{k-1} = x_{k-1}$.

Za obrat tvrdnje, pretpostavimo da je izvršeno k koraka Arnoldijevog algoritma, i pretpostavimo da vrijedi (2.77). Iz (2.76) slijedi da mora biti $|s_k| = 1$ i $c_k = 0$. Upotrebom jednakosti (2.79) možemo napisati

$$F_k \begin{bmatrix} R^{(k-1)} & F^{(k-1)} h_k \\ 0 & h_{k+1,k} \end{bmatrix} = R^{(k)}.$$

F_k je izabrana tako da $R^{(k)}$ bude gornje trokutasta, sa zadnjim nul-retkom. Ali mi znamo da F_k ima oblik

$$F_k = \begin{bmatrix} I_{k-1} & & \\ & 0 & s_k \\ & -\bar{s}_k & 0 \end{bmatrix}.$$

Odavde ispada da zadnja komponenta od k-dimenzionalnog vektora $F^{(k-1)} h_k$ je jednaka 0, odnosno $F^{(k-1)} h_k = [w^T \ 0]^T$ za neki $(k-1)$-dimenzionalni vektor w. Ako je $w = 0$ tada je i $h_k = 0$ jer je $F^{(k-1)}$ unitarna matrica, pa je $H_k = [H_{k,k-1} \ h_k]$ singularna. Sada pretpostavimo da je $w \neq 0$. Neka $v = [s^T \ t]$, sa $s \in \mathbb{C}^{k-1}$ i $t \in \mathbb{C}$. Tada imamo

$$\begin{aligned} H_k v &= [H_{k,k-1} \ h_k] v = [F^{(k-1)*} R^{(k-1)} \ h_k] v = \\ &= F^{(k-1)*} [R^{(k-1)} \ F^{(k-1)} h_k] v = F^{(k-1)*} (R^{(k-1)} s + t F^{(k-1)} h_k). \end{aligned}$$

Sada, zbog toga što je $R^{(k-1)} = [R_k^T \ 0]^T$ imamo

$$H_k v = F^{(k-1)*} \begin{bmatrix} R_{k-1}s + tw \\ 0 \end{bmatrix}. \tag{2.80}$$

Kako je R_k regularna, za bilo koji t postoji s takav da je desna strana u (2.80) jednaka nuli. Prema tome postoji vektor v za koji je $H_k v = 0$, pa je prema tome H_k singularna. □

2.5.3 Analiza greške i konvergencija GMRES metode

GMRES algoritam primijenjen na općeniti linearni susutav daje u koraku k rezidual koji ponovo zadovoljava

$$\|r_k\|_2 = \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \|p_k(A)r_0\|_2, \tag{2.81}$$

gdje se minimum uzima po svim polinomima p_k stupnja k ili manje uz svojstvo da je $p_k(0) = 1$. To se vidi iz sljedeće leme.

Lema 2.5.6 ([32]). *Neka je x_k aproksimacija rješenja ostvarena u k-tom koraku GMRES algoritma, i neka je $r_k = b - Ax_k$. Tada postoji $q_{k-1} \in \mathbb{P}_{k-1}$ takav da je x_k oblika*

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

i

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

Dokaz: Iz definicije metode u k-tom koraku bira se aproksimacija rješenja x_k takva da za $\mathcal{K}_k(A, r_0) = \text{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$ vrijedi

$$\|b - Ax_k\|_2 = \min_{x \in x_0 + \mathcal{K}_k(A,r_0)} \|b - Ax\|_2,$$

jer je x_k oblika $x_k = x_0 + Q_k y_k$, pri čemu stupci od Q_k čine bazu Krylovljevog potprostora $\mathcal{K}_k(A, r_0)$. Kako je svaki $x \in x_0 + \mathcal{K}_k(A, r_0)$ oblika $x = x_0 + \sum_{j=0}^{k-1} \alpha_j A^j r_0$, tada možemo pisati

$$x = x_0 + q_{k-1}(A)r_0, \qquad \text{za } q_{k-1} \in \mathbb{P}_{k-1}.$$

U tom slučaju je

$$r = b - Ax = b - Ax_0 - Aq_{k-1}(A)r_0 = (I - Aq_{k-1}(A))r_0 = p_k(A)r_0,$$

za $p_k \in \mathbb{P}_k$, sa $p_k(0) = 1$, što dokazuje tvrdnju leme. □

Sljedeći teorem i korolar govore o konvergenciji GMRES metode, primijenjene na dijagonalizabilne matrice, pri čemu pod konvergencijom smatramo brzinu kojom se norma reziduala približava nuli. Isto tako nas interesiraju uvjeti pod kojima bi GMRES metoda mogla doći do rješenja i prije n-tog koraka.

Teorem 2.5.7 ([32]). *Pretpostavimo da je A dijagonalizabilna matrica i neka je $A = V\Lambda V^{-1}$ spektralna dekompozicija od A, gdje je $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ dijagonalna matrica svojstvenih vrijednosti, a stupci regularne matrice V su svojstveni vektori od A. Definirajmo,*

$$\epsilon_k = \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \max_{i=1,\ldots,n} |p_k(\lambda_i)|.$$

Tada, norma reziduala postignutog u k-tom koraku GMRES metode zadovoljava nejednakost

$$\|r_k\|_2 \leq \kappa(V)\epsilon_k\|r_0\|_2,$$

gdje je $\kappa(V) = \|V\|_2\|V^{-1}\|_2$.

Dokaz: Neka je p_k bilo koji polinom stupnja manjeg ili jednakog k, koji zadovoljava uvjet $p_k(0) = 1$, i neka je $x \in x_0 + \mathcal{K}_k(A, r_0)$ vektor za koji je $r = b - Ax = p_k(A)r_0$ iz $\mathcal{K}_{k+1}(A, r_0)$. Tada

$$\|b - Ax\|_2 = \|Vp_k(\Lambda)V^{-1}r_0\|_2 \leq \|V\|_2\|V^{-1}\|_2\|p_k(\Lambda)\|_2\|r_0\|_2.$$

Budući da je Λ dijagonalna matrica, vrijedi

$$\|p_k(\Lambda)\|_2 = \max_{i=1,\ldots,n} \|p_k(\lambda_i)\|_2.$$

Kako x_k minimizira normu reziduala na $x_0 + \mathcal{K}_k(A, r_0)$, tada za bilo koji polinom p_k sa gornjim svojstvima vrijedi

$$\|b - Ax_k\|_2 \leq \|b - Ax\|_2 \leq \|V\|_2\|V^{-1}\|_2\|r_0\|_2 \max_{i=1,\ldots,n} |p_k(\lambda_i)|.$$

Sada možemo u gornjoj nejednakosti uzeti polinom p_k koji minimizira desnu stranu nejednakosti, to odgovara odabiru odgovarajućeg $x \in x_0 + \mathcal{K}_k(A, r_0)$. Time dobivamo željeni rezultat

$$\|b - Ax_k\|_2 \leq \|b - Ax\|_2 \leq \|V\|_2\|V^{-1}\|_2\|r_0\|_2\epsilon_k.$$

□

Pretpostavit ćemo da su stupci od V skalirani tako da uvjetovanost $\kappa(V)$ bude što manja. Međutim, polinom koji minimizira $\|Vp_k(\Lambda)V^{-1}r_0\|_2$ ne mora biti polinom koji minimizira $\|p_k(\Lambda)\|_2$, pa na prvi pogled nije jasno da li je ocjena Teorema 2.5.7 stroga.

Zadržimo se i dalje na dijagonalizabilnim matricama, ali ćemo promatrati one matrice kojima spektar možemo smjestiti u neki pravilni geometrijski lik, koji je po mogućnosti udaljen od ishodišta. Najjednostavniji slučaj je kada spektar matrice A možemo smjestititi unutar elipse $\mathbf{E}(c, d, a)$ sa centrom u c, gdje je udaljenost između fokusa jednaka $2d$, a veća poluos je a. Još se zahtijeva da se ishodište nalazi izvan te elipse.

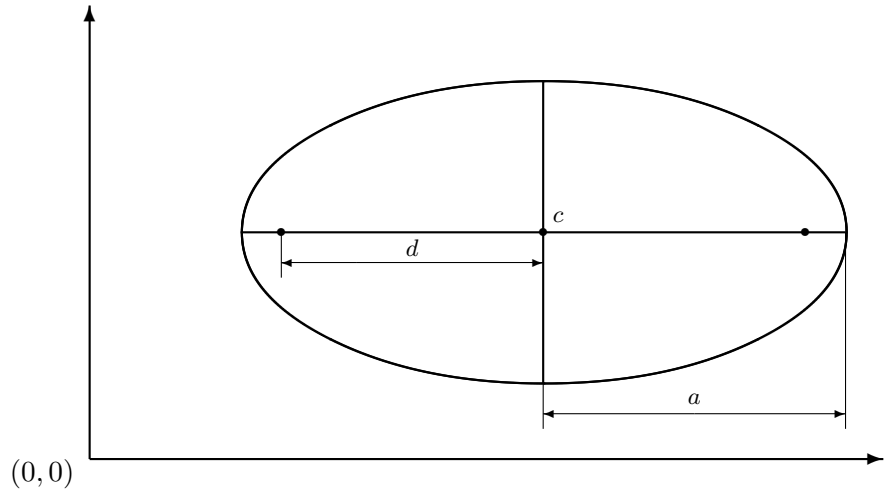
Korolar 2.5.8 ([32]). *Neka je A dijagonalizabilna matrica, to jest neka je* $A = V\Lambda V^{-1}$ *gdje je* $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ *dijagonalna matrica svojstvenih vrijednosti. Pretpostavimo da su sve svojstvene vrijednosti od A smještene unutar elipse* $\mathbf{E}(c, d, a)$ *u kojoj se ne nalazi ishodište. Tada norma reziduala, postignutog u k-tom koraku GMRES metode, zadovoljava nejednakost*

$$\|r_k\|_2 \leq \kappa(V)\frac{T_k\left(\frac{a}{d}\right)}{\left|T_k\left(\frac{c}{d}\right)\right|}\|r_0\|_2,$$

gdje je T_k *Čebiševljev polinom stupnja k*

Dokaz: Ono što trebano pronaći je gornja ograda za skalar ϵ_k iz Teorema 2.5.7 pod zadanim pretpostavkama. Po definiciji je

$$\begin{aligned}\epsilon_k &= \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \max_{i=1,\ldots,n} |p_k(\lambda_i)| \\ &\leq \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \max_{\lambda\in\mathbf{E}(c,d,a)} |p_k(\lambda)|.\end{aligned}$$



Slika 2.1: Elipsa $\mathbf{E}(c, d, a)$, koja ne sadrži ishodište.

Nejednakost vrijedi zbog toga što se maksimum modula analitičke funkcije poprima na rubu domene. Za danu analizu koristit ćemo Čebiševljeve polinome.

Čebiševljeve polinome kompleksne varijable možemo definirati na sljedeći način

$$T_k(z) = \cosh(k\zeta), \qquad \text{gdje je } z = \cosh(\zeta).$$

Ako definiramo varijablu $w = e^{\zeta}$, gornja formula ekvialentna je

$$T_k(z) = \frac{1}{2}[w^k + w^{-k}], \qquad \text{gdje je } z = \frac{1}{2}[w + w^{-1}].$$

Ova definicija Čebiševljevih polinoma koristiti se u skupu $\mathbb{C}$. Primijetimo da jednadžba $\frac{1}{2}[w + w^{-1}] = z$ ima dva rješenja $w_{1,2} = z \pm \sqrt{z^2 - 1}$ koja su međusobno inverzna, a iz definicije polinoma T_k vidimo da on ne ovisi o izboru tih rješenja. Direktnom provjerom može se provjeriti da su T_k-ovi zaista polinomi po varijabli z i da zadovoljavaju rekurziju

$$T_{k+1}(z) = 2zT_k(z) - T_{k-1}(z),$$

$$T_0 = 1, \qquad T_1 = z.$$

Neka je $\mathbf{C}_\rho$ kružnica radijusa ρ sa centrom u ishodištu. Tada, takozvano Joukowskovo preslikavanje

$$J(w) = \frac{1}{2}[w + w^{-1}]$$

transformira $\mathbf{C}_\rho$ u elipsu sa centrom u ishodištu, fokusima -1, 1, velikom poluosi $\frac{1}{2}[\rho + \rho^{-1}]$ i malom poluosi $\frac{1}{2}[\rho - \rho^{-1}]$. Postoje dvije kružnice koje se mogu preslikati pomoću preslikavanja $J(w)$ u istu elipsu, jedna sa radijusom ρ, a druga sa radijusom ρ^{-1}. Prema tome dovoljno je promatrati samo kružnice sa radijusom $\rho \geq 1$ (za $\rho = 1$ dobivamo degenerični slučaj u kojem se elipsa svodi na interval $[-1, 1]$). Ako sada promatramo elipsu $\mathbf{E}_\rho$ dobivenu iz $\mathbf{C}_\rho$ preslikavanjem J, uz pretpostavku da je γ bilo koja točka kompleksne ravnine izvan elipse, tada vrijedi

$$\min_{p_k \in \mathbb{P}_k, p_k(0)=1} \max_{\lambda \in \mathbf{E}_\rho} |p_k(\lambda)| \leq \frac{\rho^k + \rho^{-k}}{|w_\gamma^k + w_\gamma^{-k}|}, \tag{2.82}$$

gdje je w_γ dominanatni korijen jednadžbe $J(w) = \gamma$. Da bi to pokazali, prvo primijetimo da se bilo koji polinom p_k stupnja k, koji zadovoljava uvjet $p_k(\gamma) = 1$, može napisati kao

$$p_k(z) = \frac{\sum_{j=0}^{k} a_j z^j}{\sum_{j=0}^{k} a_j \gamma^j}.$$

Točka z na elipsi $\mathbf{E}_\rho$ je dobivena iz neke točke w iz $\mathbf{C}_\rho$ djelovanjem preslikavanja J, a također i γ je, na isti način dobivena iz w_γ. Tada, se polinom p_k, koji djeluje na točkama elipse, može na drugačiji način napisati kao

$$p_k(z) = \frac{\sum_{j=0}^{k} b_j (w^j + w^{-j})}{\sum_{j=0}^{k} b_j (w_\gamma^j + w_\gamma^{-j})},$$

gdje se koeficijenti b_j mogu dobiti kao neke linearne kombinacije koeficijenata a_j. Promotrimo jedan posebni polinom, koji se dobiva ako uzmemo da je $b_k = 1$ i $b_j = 0$ za $j \neq k$,

$$p_k^*(z) = \frac{w^k + w^{-k}}{w_\gamma^k + w_\gamma^{-k}} = \frac{T_k(z)}{T_k(\gamma)}$$

što je zapravo skalirani Čebiševljev polinom stupnja k po varijabli z. Kako točke oblika $w^k + w^{-k}$ ponovo pripadaju nekoj elipsi sa poluosima $\rho^k + \rho^{-k}$ i $\rho^k - \rho^{-k}$ i centrom u ishodištu, tada je očito da se maksimum modula polinoma p_k^* postiže u tjemenima velikih poluosi te elipse, to jest kada je $w = \rho e^{i\phi}$ realan i jednak $\pm\rho$. Ta se vrijednost postiže za $z = \frac{1}{2}(\rho + \rho^{-1})$, što je tjeme velike poluosi elipse $\mathbf{E}_\rho$. Prema tome je

$$\max_{z \in \mathbf{E}_\rho} |p_k^*(z)| = \frac{\rho^k + \rho^{-k}}{|w_\gamma^k + w_\gamma^{-k}|}$$

što dokazuje traženu nejednakost (2.82).

Za općenitiju elipsu $\mathbf{E}(c, d, a)$, jednostavnom zamjenom varijabli koja c prebacuje u 0, $c + d$ u -1, i $c - d$ u 1, dobivamo da se polinom p_k^* transformirao u

$$\hat{T}_k(z) = \frac{T_k\left(\frac{c-z}{d}\right)}{T_k\left(\frac{c-\gamma}{d}\right)}. \tag{2.83}$$

Analogno, kao u slučaju kada je centar elipse bio u ishodištu, lako se vidi da se maksimum modula polinoma $\hat{T}_k(z)$ postiže u točki $c - a$, tjemenu velike poluosi elipse $\mathbf{E}(c, d, a)$. Dakle, imamo

$$\max_{z \in \mathbf{E}(c,d,a)} |\hat{T}_k(z)| = \frac{T_k\left(\frac{a}{d}\right)}{\left|T_k\left(\frac{c-\gamma}{d}\right)\right|}.$$

Ovdje treba napomenuti da, d i a mogu biti i imaginarni. U tom slučaju su velike poluosi vertikalne, ali je a/d ponovo realan, pa je brojnik gornjeg izraza uvijek realan. Sada možemo iskoristiti polinom $\hat{T}_k$ sa $\gamma = 0$

$$\begin{aligned} \epsilon_k &\leq \min_{p_k \in \mathbb{P}_k, p_k(0)=1} \max_{\lambda \in \mathbf{E}(c,d,a)} |p_k(\lambda)| \\ &\leq \max_{\lambda \in \mathbf{E}(c,d,a)} |\hat{T}_k(\lambda)| = \frac{T_k\left(\frac{a}{d}\right)}{|T_k\left(\frac{c}{d}\right)|}. \end{aligned}$$

Time je dokaz dovršen. □

Ekspilcitan izraz za $T_k(\frac{a}{d})/T_k(\frac{c}{d})$ može se dobiti iz definicije Čebiševljevih polinoma:

$$\begin{aligned}\frac{T_k\left(\frac{a}{d}\right)}{T_k\left(\frac{c}{d}\right)} &= \frac{\left(\frac{a}{d}+\sqrt{\left(\frac{a}{d}\right)^2-1}\right)^k+\left(\frac{a}{d}+\sqrt{\left(\frac{a}{d}\right)^2-1}\right)^{-k}}{\left(\frac{c}{d}+\sqrt{\left(\frac{c}{d}\right)^2-1}\right)^k+\left(\frac{c}{d}+\sqrt{\left(\frac{c}{d}\right)^2-1}\right)^{-k}} \\ &\approx \left(\frac{a+\sqrt{a^2-d^2}}{c+\sqrt{c^2-d^2}}\right)^k.\end{aligned}$$

Budući da uvjetovanost $\kappa(V)$ matrice svojstvenih vektora V obično nije poznata i može biti vrlo velika, ocjena Korolara 2.5.8 ponovo ne mora biti precizna. Ona može biti korisna jedino ako znamo da je matrica A normalna, pa je tada $\kappa(V) = 1$. U tom slučaju dobili smo vrlo primjenljivu ocjenu konvergencije samo pomoću analize distribucije svojstvenih vrijednosti. U normalnom slučaju, kao i kod hermitskih matrica, problem konvergencije GMRES metode općenito svodi se na problem teorije aproksimacija: kako se dobro može aproksimirati nula na skupu kompleksnih svojstvenih vrijednosti, koristeći polinome stupnja k koji u ishodištu poprimaju vrijednost 1. U ovom problemu može se dobro primijeniti informacija o tome da li su svojstvene vrijednosti dobro ili loše distribuirane u kompleksnoj ravnini. Svojstvene vrijednosti nakupljene u uskom području oko jedne točke c, daleke od ishodišta daju dobru distribuciju, budući da je vrijednost polinoma $(1 - z/c)^k$ mala u svim točkama kompleksne ravnine koje su bliske točki c . Dobar primjer iskorištavanje ovakve distribucije je i Korolar 2.5.8. Svojstvene vrijednosti koje su raspoređene svuda oko ishodišta su loše distribuirane, jer je nemoguće naći polinom (po principu maksimuma modula analitičke funkcije) čija je vrijednost u ishodištu jednaka 1, a u svakoj točki na nekoj zatvorenoj krivulji oko ishodišta, manja od 1. Također polinom niskog stupnja ne može biti 1 u ishodištu, i malen po apsolutnoj vrijednosti u mnogim točkama distribuiranim svuda oko ishodišta.

Općenito se, međutim, ponašanje GMRES metode ne može odrediti samo iz svojstava svojstvenih vrijednosti. Zapravo, bilo koja nerastuća krivulja može predstavljati graf normi reziduala po broju iteracija za GMRES metodu primijenjenu na nekom nenormalnom problemu, štoviše, matrica problema može imati bilo koje svojstvene vrijednosti. To znači da, kada radimo sa krajnje ne-normalnim matricama, informacija o svojstvenim vrijednostima sama ne može garantirati brzu konvergenciju GMRES metode. Pa tako, na primjer, svojstvene vrijednosti koje su nakupljene usko oko 1 ne moraju nužno biti dobre za ne-normalne matrice, kao što jesu za normalne. Gornje tvrdnje potkrijepit ćemo sljedećim primjerom.

Primjer 2.5.9 ([14]). *Norme reziduala dobivene u nizu koraka GMRES metode su nerastući niz, budući da se reziduali minimiziraju po skupu ekspandirajućih podprostora. Postavlja se sljedeće pitanje: ako je dan nerastući niz pozitivnih realnih brojeva $f(0) \geq f(1) \geq \cdots \geq f(n-1) > 0$, i skup kompleksnih brojeva $\{\lambda_1 \ldots, \lambda_n\}$, različitih od nule, da li tada postoje $n \times n$ matrica A sa svojstvenim vrijednostima $\lambda_1,\ldots,\lambda_n$ i početni vektor r_0 sa $\|r_0\|_2 = f(0)$, takvi da kod primjene GMRES algoritma na linearni sustav $Ax = b$, sa početnim rezidualom r_0, dobijemo aproksimacije x_k takve da je $\|r_k\|_2 = f(k)$, $k = 1,\ldots,n-1$? Odgovor na ovo pitanje je potvrdan. No, prije samog odgovora primijetimo da pretpostavka $f(n-1) > 0$ znači da GMRES algoritam ne konvergira k egzaktnom rješenju sve do zadnjeg, n-tog, koraka, kada su dimenizije potprostora $\mathcal{K}_n(A, r_0)$ i $A\mathcal{K}_n(A, r_0)$ jednake n.*

Započet ćemo najprije sa jednostavnom analizom nekih svojstava traženog rješenja. Budući da su reziduali generirani primjenom GMRES algoritma na linearni sustav $Ax = b$, sa početnom aproksimacijom x_0, potpuno određeni matricom A i početnim rezidualom r_0, bez smanjenja općenitosti možemo pretpostaviti da je početna aproksimacija x_0 jednaka nuli, i da je vektor desne strane sustava b zapravo početni rezidual. Pretpostavimo da A i b predstavljaju nepoznatu matricu i desnu stranu sustava. Neka je $\mathcal{W} = \{w_1, \ldots, w_n\}$ ortonormirana baza Krylovljevog prostora reziduala $A\mathcal{K}_n(A, b)$, takva da je span$\{w_1, \ldots, w_j\} = A\mathcal{K}_j(A, b)$, $j = 1, \ldots, n$, i neka je W matrica sa ortonormiranim stupcima $[w_1 \ \ldots \ w_n]$. Prema Lemi 2.5.6 minimizacijsko svojstvo reziduala možemo preformulirati u

$$\|r_k\|_2 = \min_{u \in A\mathcal{K}_k(A,b)} \|b - u\|_2,$$

odakle se vidi da je $r_k \perp A\mathcal{K}_k(A, b)$ za $k = 0, 1, \ldots, n-1$, pa je on oblika

$$r_k = \sum_{j=k+1}^{n} \langle b, w_j \rangle w_j,$$

r_n je nula, a $b = r_0$ se može raspisati kao

$$b = \sum_{j=1}^{n} \langle b, w_j \rangle w_j.$$

Iz prethodnoga se vidi da je $|\langle b, w_j \rangle| = \sqrt{\|r_{j-1}\|_2^2 - \|r_j\|_2^2}$. Ako je zadan nerastući pozitivan niz $f(0) \geq f(1) \geq \cdots \geq f(n-1) > 0$, definirajmo $f(n) = 0$ i i nove vrijednosti $g(k)$ sa

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

Uvjeti $\|b\|_2 = f(0)$, $\|r_k\|_2 = f(k)$, $k = 1, 2, \ldots, n-1$ bit će tada zadovoljeni ako su koordinate od b u bazi $\mathcal{W}$ određene sa

$$W^* b = [g(1) \ \ldots \ g(n)]^T.$$

Zaista, u tom slučaju je $r_k = \sum_{j=k+1}^{n} \langle b, w_j \rangle w_j = \sum_{j=k+1}^{n} g(j) w_j$, pa slijedi

$$\|r_k\|_2^2 = \sum_{j=k+1}^{n} |g(j)|^2 = \sum_{j=k+1}^{n} [(f(j-1))^2 - (f(j))^2] = (f(k))^2.$$

Neka je $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, $\lambda_j \neq 0$, $j = 1, 2, \ldots, n$ skup točaka različitih od nule u kompleksnoj ravnini. Promotrimo normirani polinom

$$a(z) = z^n - \sum_{j=0}^{n-1} \alpha_j z^j = (z - \lambda_1)(z - \lambda_2) \cdots (z - \lambda_n).$$

Zbog toga što su svi λ_j različiti od nule, slijedi $\alpha_0 \neq 0$.

Konstrukcija matrice A i desne strane susutava b sada je jednostavna. Matricu A možemo smatrati linearnim operatorom na n-dimenzionalnim Hilbertovim prostorom $\mathbb{C}^n$. Taj operator označit ćemo sa $\mathcal{A}$, a njegova reprezantacija u standardnoj bazi $\mathcal{E} = \{\xi_1, \ldots, \xi_n\}$, pri čemu je ξ_i i-ti jedinični vektor, daje traženu matricu A:

$$\mathcal{A}^{\mathcal{E}} = A.$$

$\mathcal{A}$ je jedinstveno određen djelovanjem na bilo koji skup vektora baze.

Neka je $\mathcal{V} = \{v_1, \ldots, v_n\}$ bilo koja ortonormirana baza u $\mathbb{C}^n$, i neka je V matrica sa ortonormiranim stupcima $[v_1 \ \ldots \ v_n]$. Neka b zadovoljava

$$V^*b = [g(1) \ \ldots \ g(n)]^T$$

pri čemu ukoliko je dan b sa $\|b\|_2 = f(0)$, V može biti izabran, ili ako je V zadan, b može biti izabran. Budući da je $g(n) = f(n-1)$ različit od nule, skup vektora $\mathcal{B} = \{b, v_1, \ldots v_{n-1}\}$ je linearno nezavisan i također tvori bazu za $\mathbb{C}^n$. Neka je B matrica sa stupcima $[b \ v_1 \ \ldots \ v_{n-1}]$. Tada je operator $\mathcal{A}$ jednostavno određen jednadžbama

$$\begin{aligned} \mathcal{A}b &= v_1, \\ \mathcal{A}v_1 &= v_2, \\ &\vdots \\ \mathcal{A}v_{n-2} &= v_{n-1}, \\ \mathcal{A}v_{n-1} &= \alpha_0 b + \alpha_1 v_1 + \cdots + \alpha_{n-1} v_{n-1}. \end{aligned}$$

U matričnoj reprezentaciji u bazi $\mathcal{B}$ matrica

$$\mathcal{A}^{\mathcal{B}} = \begin{bmatrix} 0 & \ldots & 0 & \alpha_0 \\ 1 & \ldots & 0 & \alpha_1 \\ & \ddots & \vdots & \vdots \\ & & 1 & \alpha_{n-1} \end{bmatrix},$$

tada ima svojstvene vrijednosti koje odgovaraju skupu Λ. Radi se o matrici koju još nazivamo pratilac polinoma*, i čija svojstva su dana u [27]. Na kraju, matrica A je dana sa*

$$A = \mathcal{A}^{\mathcal{E}} = B\mathcal{A}^{\mathcal{B}}B^{-1}.$$

Sve ovo dokazuje tvrdnju sljedećeg teorema.

Teorem 2.5.10. *Neka je dan nerastući niz $f(0) \geq f(1) \geq \cdots \geq f(n-1) > 0$ pozitivnih brojeva i skup kompleksnih brojeva različitih od nule $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, tada postoji matrica A sa svojstvenim vrijednostima $\lambda_1, \lambda_2, \ldots, \lambda_n$ i desna strana sustava b sa $\|b\|_2 = f(0)$ takvi da reziduali r_k u svakom koraku GMRES algoritma primijenjenog na sustav $Ax = b$ sa $x_0 = 0$, zadovoljavaju $\|r_k\|_2 = f(k)$, $k = 1, 2, \ldots, n-1$.*

Dakle, ne možemo ništa više reći o svojstvima konvergencije GMRES metode, primijenjene na ne-normalne sustave.

Da rezimiramo slučaj dijagonalizabilnih matrica. Dakle, kada je matrica svojstvenih vektora ekstremno loše uvjetovana, ocjena Teorema 2.5.7, je manje korisna. Ona može biti veća od $\|r_0\|_2$ za sve $k < n$, ali mi znamo, zbog svojstava normi reziduala objašnjenih u prethodnom primjeru, da je $\|r_k\|_2 \leq \|r_0\|_2$ za sve k. U takvim slučajevima nije jasno da li GMRES metoda slabo konvergira ili je ocjena Teorem 2.5.7 precijenila stvarnu normu reziduala.

Drugačije ograde norme reziduala mogu se ostvariti preko polja vrijednosti $\mathcal{F}(A)$ matrice A, uz uvjet $0 \notin \mathcal{F}(A)$. Na primjer, pretpostavimo da je $\mathcal{F}(A)$ sadržan u krugu

$\mathbf{K} = \{z \in \mathbb{C} : \ |z - c| \leq s\}$ koji ne sadrži ishodište. Razmotrimo polinom $p_k(z) = (1 - z/c)^k$. Iz pravila (1.4) i (1.5), koja vrijede za polje vrijednosti slijedi

$$\mathcal{F}(I - \frac{1}{c}A) = 1 - \frac{1}{c}\mathcal{F}(A) \subseteq \{z \in \mathbb{C} : \ |z| \leq \frac{s}{|c|}\}$$

pa je tada, $\nu((I-(1/c)A) \leq s/|c|$. Nejednakost potencija (1.11) daje $\nu((I-(1/c)A)^k) \leq (s/|c|)^k$, odakle je, zbog svojstva (1.10), koje određuje odnos euklidske norme i numeričkog radijusa,

$$\|p_k(A)\|_2 \leq 2\left(\frac{s}{|c|}\right)^k.$$

Slijedi da norma GMRES-ovog reziduala zadovoljava nejednakost

$$\|r_k\|_2 \leq 2\left(\frac{s}{|c|}\right)^k \|r_0\|_2, \tag{2.84}$$

Ocjena (2.84) ponekad dosta precijenjuje pravu normu reziduala GMRES metode. U mnogim slučajevima, krug $\mathbf{K}$ mora biti puno veći nego sam $\mathcal{F}(A)$ kako bi obuhvatio cijeli $\mathcal{F}(A)$, a osim toga ne smije sadržavati ishodište.

Još jedan drugačiji pristup za ocjenjivanje $\|p(A)\|_2$ je primjena *pseudospektra*. Za bilo koji polinom p, ako ga gledamo kao funkciju, on je analitički na svakom otvorenom skupu koji sadrži kompaktan skup $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$, odnosno spektar od A, pa se, prema [28, str. 475], matrica $p(A)$ može napisati kao Cauchyjev integral

$$p(A) = \frac{1}{2\pi i}\int_\Gamma p(z)(zI - A)^{-1}dz,$$

gdje je Γ bilo koja jednostavna zatvorena krivulja ili unija jednostavnih zatvorenih krivulja (pozitivno orijentiranih) koje sadrže spektar od A. Ako u prethodnoj jednakosti uzmemo norme na svakoj strani, i ako zamijenimo normu integrala sa duljinom krivulje $\mathcal{L}(\Gamma)$ puta maksimum norma od integranda dobivamo

$$\|p(A)\|_2 \leq \frac{\mathcal{L}(\Gamma)}{2\pi}\max_{z\in\Gamma}\|p(z)(zI - A)^{-1}\|_2.$$

Nadalje, ako promotrimo krivulju Γ_ϵ na kojoj je norma rezolvente $\|(zI - A)^{-1}\|_2$ konstantna, recimo $\|(zI - A)^{-1}\|_2 = \epsilon^{-1}$, tada imamo

$$\|p(A)\|_2 \leq \frac{\mathcal{L}(\Gamma_\epsilon)}{2\pi\epsilon}\max_{z\in\Gamma_\epsilon}|p(z)|.$$

Krivulju za koju vrijedi $\|(zI - A)^{-1}\|_2 = \epsilon^{-1}$ zovemo granicom ϵ-*pseudospektra* matrice A:

$$\Lambda_\epsilon = \{z : \ \|(zI - A)^{-1}\|_2 \geq \epsilon^{-1}\}.$$

Iz optimalnosti GMRES aproksimacije slijedi da rezidual r_k GMRES metode zadovoljava

$$\|r_k\|_2 \leq \frac{\mathcal{L}(\Gamma_\epsilon)}{2\pi\epsilon}\min_{p_k\in\mathbb{P}_k, p_k(0)=1}\max_{z\in\Gamma_\epsilon}|p(z)|\|r_0\|_2 \tag{2.85}$$

za bilo koji izbor parametra ϵ. Za neke probleme, i sa pažljivo odabranim vrijednosti ϵ ograda (2.85) može biti puno manja od one u Teoremu 2.5.7. Ali ni ograda (2.85)

nije stroga, i za neke probleme ne može se naći povoljan ϵ koji bi dao realističnu ocjenu pravog GMRES reziduala, vidi [16]. Lako se može vidjeti što dovodi do precijenjivanja ocjene. Norma integrala funkcije može biti puno manja od produkta duljine krivulje i maksimuma norme integranda.

Svaka od ocjena danih u Teoremu 2.5.7, (2.84) i (2.85) daju ograde reziduala GMRES metode tako što ograđuju veličinu $\min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|p_k(A)\|_2$. Najgori slučaj ponašanja GMRES metode dan je sa

$$\|r_k\|_2 = \max_{\|r_0\|_2=1} \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|p_k(A)r_0\|_2. \tag{2.86}$$

Sad se postavlja pitanje kada je desna strana u (2.86) jednaka sljedećoj veličini

$$\min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|p_k(A)\|_2 = \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \max_{\|r_0\|_2=1} \|p_k(A)r_0\|_2? \tag{2.87}$$

Dakle, svaki korak GMRES metode je matematički ekvivalentan minimizaciji $\|p_k(A)r_0\|_2$ po skupu polinoma, kojeg možemo označiti sa $\mathbf{P}_k$, pri čemu je

$$\mathbf{P}_k = \{p_k \in \mathbb{P}_k : \ p_k(0) = 1\}.$$

Pokazat ćemo kasnije da, za svaki r_0, takav GMRES-ov polinom, označen sa p_{r_0} postoji i jedinstven je ako je $\|p_{r_0}(A)\|_2 > 0$. Znači, brzina konvergencije GMRES iteracije ovisi o tome kako brzo $\|p_{r_0}(A)r_0\|_2$ konvergira k nuli kada se k povećava, a to ponovo ovisi o matrici sustava A i početnom rezidualu r_0. Ipak se čini da dominantnu ulogu u konvergeniciji ima matrica A, pa stoga se javlja potreba za proučavanjem problema minimiziranja $\|p_k(A)\|_2$ za $p_k \in \mathbb{P}_k$, $p_k(0) = 1$, kojeg možemo nazvati "problemom idealne GMRES metode". Također ćemo pokazati da polinom idealnog GMRES-a, kojeg ćemo označiti sa p_*, postoji i jedinstven je u slučaju kada je $\|p_*(A)\|_2 > 0$.

Prema tome, GMRES metoda pronalazi $p_{r_0} \in \mathbf{P}_k$ takav da je

$$\|p_{r_0}(A)r_0\|_2 \text{ minimalno.} \tag{2.88}$$

Lema 2.5.6 daje ekvivalentnu tvrdnju

$$r_0 \approx \text{span}\{Ar_0, A^2r_0, \ldots, A^kr_0\} = A\mathcal{K}_k(A, r_0), \tag{2.89}$$

pri čemu "$y \approx V$" označava problem pronalaženja najbolje aproksimacije obzirom na normu $\|\cdot\|_2$ točke y u potprostoru V.

S druge strane, idealni GMRES metoda pronalazi $p_* \in \mathbf{P}_k$ takav da je

$$\|p_*(A)\|_2 \text{ minimalno.} \tag{2.90}$$

Ekvivalentno

$$I \approx \text{span}\{A, A^2, \ldots, A^k\}. \tag{2.91}$$

Nadalje, za svaki b, zbog svojstva minimalnosti, vrijedi

$$\|p_{r_0}(A)r_0\|_2 \le \|p_*(A)r_0\|_2 \le \|p_*(A)\|_2\|r_0\|_2,$$

što daje rezultat i za najgori slučaj, za svaki k

$$\max_{r_0\in\mathbb{C}^n, \|r_0\|_2=1} \|p_{r_0}(A)r_0\|_2 \le \|p_*(A)\|_2. \tag{2.92}$$

Time smo zadali odnos između (2.86) i (2.87), odnosno p_{r_0} i p_*. Odgovor na pitanje egzistencije i jedinstvenosti ova dva polinoma daje sljedeći teorem.

Teorem 2.5.11 ([17],[26]). *Optimalni polinomi p_{r_0} i p_* postoje. U slučaju kada su minimumi u (2.88) i (2.90) za GMRES i idealni GMRES različiti od nule, i kada je matrica A regularna, oni su jedinstveni.*

Dokaz: Promatrat ćemo alternativne formulacije problema (2.89) i (2.91). U oba slučaja imamo problem oblika $w \approx V$, gdje je V konačnodimenzionalan potprostor normiranog vektorskog prostora W i $w \in W$. Za GMRES je $W = \mathbb{C}^n$, a za idealni GMRES je $W = \mathbb{C}^{n^2}$. Dakle, tražimo $v_{min} \in V$ takav da je

$$\|w - v_{min}\| = e(w, V) = \inf_{v \in V} \|w - v\|.$$

Trebamo pokazati da je skup $E(w, V) = \{v \in V : \ \|w - v\| = e(w, V)\}$ neprazan. Budući da je $0 \in V$ imamo

$$e(w, V) \leq \|w - 0\| = \|w\|.$$

Promotrimo sada proizvoljni element $v \in V$ koji zadovoljava $\|v\| > 2\|w\|$. Tada vrijedi

$$\|w - v\| \geq \|v\| - \|w\| > \|w\| \geq e(w, V),$$

odakle slijedi da $v \notin E(w, V)$ i $E(w, V) \subset B = \{v \in \text{V:} \ \|v\| \leq 2\|w\|\}$. Rezultat toga je $e(w, V) = \inf_{v \in B} \|w - v\|$. Kako je svaki zatvoren i ograničen podskup konačno dimenzionalnog prostora kompaktan, i kako svaka realna neprekidna funkcija na kompaktnom skupu poprima svoj infimum na tom skupu, to znači da, budući da je B kompaktan i funkcija $V \ni v \mapsto \|w - v\|$ je neprekidna, dobivamo egzistenciju $v_{min} \in V$ takvog da je $\|w - v_{min}\| = e(w, V)$. Prema tome $E(w, V)$ je neprazan, pa polinomi p_{r_0} i p_* postoje.

Dokaz jedinstvenosti polinoma p_{r_0} i p_* može se podijeliti na dva dijela:

(a) Da li je za w najbliža točka $v \in V$ jedinstvena?

(b) Da li se taj vektor v može na jedinstveni način napisati kao linearna kombinacija k vektora koji razapinju prostore u (2.89) i (2.91)?

Odgovor na pitanje iz dijela (b) može se izvesti na sljedeći način. Ono što mi zapravo trebamo pokazati je da su k vektora, koje razmatramo, linearno nezavisni. Za problem idealne GMRES metode (2.91), pretpostavimo da su $A, A^2, \ldots, A^k$ linearno zavisni. To znači da postoji polinom $p \in \mathbb{P}_k$ za koji je $p(A) = 0$. Budući da je u ovom slučaju konstantni koeficijent, koeficijent uz nultu potenciju varijable polinoma p, jednak nuli, tada zbog regularnosti matrice A, $p(z)$ možemo pomnožiti sa z^{-1} jednom ili više puta, tako da konstantni koeficijent postane različit od nule. Ako polinom sada pomnožimo sa odgovarajućom konstantom $p(0)$, dobivamo $p(A) = 0$ za $p \in \mathbf{P}_k$, što je kontradikcija sa pretpostavkom da je minimum u (2.90) različit od nule. Analogan dokaz može se primijeniti i u slučaju pravog GMRES-a (2.89).

Sada još moramo pokazati da vrijedi i potvrdan odgovor na pitanje iz (a). Dokaz jedinstvenosti vektora v za problem (2.89) slijedi iz činjenice da je vektorska norma $\|\cdot\|_2$ strogo konveksna, odnosno da vrijedi da za svaka dva različita elementa $w_1, w_2 \in W$, takvih da je $\|w_1\|_2 = \|w_2\|_2 = 1$, imamo $\|(w_1 + w_2)/2\|_2 < 1$. Naime, pretpostavimo da neki element $w \in W$ ima dvije različite najbolje aproksimacije $v_1, v_2 \in V$. Tada je

$$v_0 = \frac{v_1 + v_2}{2} \in V \quad \text{i} \quad w - v_0 = e(w, V)\frac{u_1 + u_2}{2},$$

gdje je, zbog činjenice da je minimum $e(w,V)$ različit od nule

$$u_1 = \frac{w - v_1}{e(w,V)} \quad \text{i} \quad u_2 = \frac{w - v_2}{e(w,V)}.$$

Budući da je $u_1 \neq u_2$ i $\|u_1\|_2 = \|u_2\|_2 = 1$ uvjet stroge konveksnosti daje

$$\|w - v_0\|_2 = e(w,V) \left\| \frac{u_1 + u_2}{2} \right\|_2 < e(w,V),$$

što je kontradikcija sa definicijom od $e(w,V)$ kao minimuma. Dakle, minimum za GMRES problem mora biti jedinstven.

S druge strane, matrična norma $\|\cdot\|_2$ nije strogo konveksna, pa ne možemo iskoristiti gornju argumentaciju kod dokaza jedinstvenosti za problem idealnog GMRES-a. Ponovo pretpostavimo da postoje dva različita rješenja p_1 i p_2 problema (2.90), i neka je minimalna norma koji oni dostižu

$$\|p_1(A)\|_2 = \|p_2(A)\|_2 = C. \tag{2.93}$$

Ako definiramo $p(z) = \frac{1}{2}(p_1(z) + p_2(z))$ tada $\|p(A)\|_2 \leq C$, pa mora biti $\|p(A)\|_2 = C$ zbog minimalnosti. Neka je $\{w_1, \ldots, w_J\}$ skup maksimalnih desnih singularnih vektora za $p(A)$, to jest skup ortonormiranih vektora sa svojstvom

$$\|p(A)w_j\|_2 = C, \qquad 1 \leq j \leq J,$$

sa najvećim mogućim brojem J. Tada za svaki w_j, zbog

$$C = \left\| \frac{1}{2}p_1(A)w_j + \frac{1}{2}p_2(A)w_j \right\|_2 \leq \frac{1}{2}\|p_1(A)w_j\|_2 + \frac{1}{2}\|p_2(A)w_j\|_2 \leq C,$$

i zbog (2.93) imamo

$$\|p_1(A)w_j\|_2 = \|p_2(A)w_j\|_2 = C$$

i

$$p_1(A)w_j = p_2(A)w_j,$$

jer bi inače, zbog stroge konveksnosti vektorske norme $\|\cdot\|_2$, imali da je $\|p(A)w_j\|_2 < C$. Zbog toga

$$(p_1 - p_2)(A)w_j = 0, \qquad 1 \leq j \leq J.$$

Budući da $(p_1 - p_2)(z)$ nije identički jednak nulpolinomu, mi ga možemo pomnožiti sa odabranim skalarom i pogodnom potencijom od z^{-1}, zbog regularnosti matrice A, tako da dobijemo polinom $\Delta p \in \mathbf{P}_k$ takav da je

$$\Delta p(A)w_j = 0, \qquad 1 \leq j \leq J.$$

Za $\epsilon \in \langle 0, 1\rangle$, promotrimo polinom $p_\epsilon \in \mathbf{P}_k$ definiran konveksnom linearnom kombinacijom

$$p_\epsilon(z) = (1 - \epsilon)p(z) + \epsilon\Delta p(z).$$

Ako sa $\{w_{J+1}, \ldots, w_n\}$ označimo ostatak od n singularnih vektora matrice $p(A)$, sa odgovarajućim singularnim vrijednostima $C > \sigma_{J+1} \geq \cdots \geq \sigma_n \geq 0$, tada imamo

$$\|p_\epsilon(A)w_j\|_2 \leq \begin{cases} (1-\epsilon)C & 1 \leq j \leq J, \\ (1-\epsilon)\sigma_{J+1} + \epsilon\|\Delta p(A)\|_2 & J+1 \leq j \leq n. \end{cases}$$

Za $1 \leq j \leq J$ dana norma je manja od C za proizvoljni ϵ, a za $J+1 \leq j \leq n$ ta norma je manja C za dovoljno mali ϵ, budući da je $\sigma_{J+1} < C$. Kako singularni vektori $w_1 \ldots, w_n$ čine ortonormiranu bazu za $\mathbb{C}^n$, to znači da je svaki vektor v norme 1 linearna kombinacija tih vektora $v = \sum_{j=1}^{n} \gamma_j w_j$ sa $\sum_{j=1}^{n} \gamma_j^2 = 1$, pa slijedi $\|p_\epsilon(A)\|_2 < C$ za dovoljno mali ϵ, što je kontradikcija sa pretpostavkom da su p_1 i p_2 minimalni. □

Dakle sada znamo da polinomi p_{r_0} i p_* zaista postoje i jedinstveni su u većini slučajeva, pa ćemo se stoga vratiti na nejednakost (2.92). Ona pokazuje da je $\|p_*(A)\|_2$ ili izraz u (2.87) gornja ograda za $\|p_{r_0}(A)r_0\|_2$ odnosno za izraz iz (2.86) koji je jednak normi reziduala r_k GMRES metode u k-tom koraku, za najgori slučaj. Pitanje je sada koliko je to dobra ograda i da li (2.87) može poslužiti kao ograda za normu od r_k. Od velike važnosti je slučaj kada zapravo nejednakost (2.92) prelazi u jednakost. Ova nejednakost je jednakost za mnoge matrice uključujući i normalne matrice, kada je $\kappa(V) = 1$ pa je ocjena Teorem 2.5.7 stroga. To vrijedi i za općenite matrice za korak k=1 i za matrice čija je dimenzija manja ili jednaka 3, vidi [25] i [13]. Mnogi numerički eksperimenti su pokazali da se radi o jednakosti za mnoge matrice i veličine k, međutim to općenito ne vrijedi što se može pokazati kontraprimjerom, kojeg je izradio K. C. Toh u [35].

Primjer 2.5.12 ([35]). *Radi se o 4×4 matrici*

$$A = \begin{bmatrix} 1 & \epsilon & 0 & 0 \\ 0 & -1 & c/\epsilon & 0 \\ 0 & 0 & 1 & \epsilon \\ 0 & 0 & 0 & -1 \end{bmatrix}, \quad \epsilon > 0,\ 0 < c < 2. \tag{2.94}$$

Za takvu matricu vrijede sljedeći teoremi.

Teorem 2.5.13 ([35]). *Za matricu A iz (2.94), polinom 3. stupnja p_* idealog GMRES-a je*

$$p_*(z) = 1 + (\alpha - 1)z^2$$

sa

$$\alpha = \frac{2c^2}{4 + c^2}.$$

Odgovarajuća matrica je

$$p_*(A) = \begin{bmatrix} \alpha & 0 & \gamma & 0 \\ 0 & \alpha & 0 & \gamma \\ 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & \alpha \end{bmatrix},$$

gdje je

$$\gamma = (\alpha - 1)c,$$

sa normom

$$\|p_*(A)\|_2 = \frac{4c}{4 + c^2}.$$

Teorem 2.5.14 ([35]). *Neka je matrica A dana sa (2.94). Tada za bilo koji vektor $r_0 \in \mathbb{C}^4$ odgovarajući polinom 3. stupnja p_{r_0} pravog GMRES-a za A zadovoljava*

$$\|p_{r_0}(A)r_0\|_2 < \|p_*(A)\|_2 \|r_0\|_2.$$

Ovaj teorem pokazuje da se za najgori slučaj pravog GMRES-a i idealnog GMRES-a norme razlikuju. Što više te dvije norme mogu se razlikovati za proizvoljno mali faktor, o čemu govori sljedeći teorem.

Teorem 2.5.15 ([35]). *Neka je matrica A dana sa (2.94), za $0 < \epsilon \leq 1$. Tada*

$$\max_{\|r_0\|_2=1} \|p_{r_0}(A)r_0\|_2 \leq 2(1+c)\sqrt{\epsilon} + (2+3c)\epsilon,$$

i za svaki $0 < c < 2$,

$$\frac{\max_{\|r_0\|_2=1} \|p_{r_0}(A)r_0\|_2}{\|p_*(A)\|_2} \longrightarrow 0 \quad kada \quad \epsilon \to 0.$$

Dakle niti jedan od pristupa koji vodi do rezultata danog u Teoremu 2.5.7 ili rezultata (2.84) i (2.85), u općenitom slučaju, ne moraju dati preciznu ogradu za normu reziduala GMRES metode, i za sada problem opisa konvergencije GMRES metode uz pomoć nekog jednostavnog svojstva matrice sustava ostaje otvoren.

2.5.4 Prekondicionirana GMRES metoda

Prekondicioniranje se koristi, kao i kod svih drugih metoda, kako bi transformiralo početni linearni sustav u sustav koji se lakše i bolje rješava. U slučaju GMRES metode, ono se može upotrijebiti u svrhu transformacije polja vrijednosti u neki povoljniji oblik ili poboljšanja distribucije svojstvenih vrijednosti (iako one uvijek ne određuju konvergenciju). Postoje tri načina, kao i kod CG metode, na koje možemo primijeniti prekondicioniranje: lijevo, desno i dvostrano.

Lijevo–prekondicionirana GMRES metoda

Lijevo–prekondicionirana GMRES metoda, je GMRES algoritam primijenjen na sustav

$$M^{-1}Ax = M^{-1}b. \tag{2.95}$$

Algoritam je sličan Algoritmu 2.5.2, samo što se umjesto reziduala početne iteracije koristi $z_0 = M^{-1}(b - Ax_0) = M^{-1}r_0$, i svugdje gdje se upotrebljava matrica sustava stavlja se $M^{-1}A$. Arnoldijev algoritam tada konstruira ortogonalnu bazu lijevo–prekondicioniranog Krylovljevog potprostora

$$\mathcal{K}_k(M^{-1}A, z_0) = \text{span}\{z_0, M^{-1}Az_0, \ldots, (M^{-1}A)^{k-1}z_0\}.$$

Svi reziduali i njihove norme koji se računaju u toku izvršavanja algoritma u vezi su sa neprekondicioniranim rezidualima preko jednakosti $z_k = M^{-1}(b - Ax_k) = M^{-1}r_k$, pa se stoga ne radi o istim vektorima kao kod neprekondicioniranog sustava. Dosta često nema jednostavnog načina da dođemo do neprekondicioniranih reziduala, osim da ih izračunamo egzaktno, množenjem prekondicioniranih reziduala sa M. To može prouzročiti neke probleme kod kriterija zaustavljanja koji su bazirani na pravim rezidualima, umjesto na onim prekondicioniranim.

Desno–prekondicionirana GMRES metoda

Desno–prekondicionirana GMRES metoda se zasniva na rješavanju sustava

$$AM^{-1}u = b, \qquad u = Mx. \tag{2.96}$$

Naravno, nova varijabla u zapravo se nikada ne treba računati. Naime, jednom kada se izračuna početni rezidual $r_0 = b - Ax_0 = b - AM^{-1}u_0$, svi daljnji vektori Krylovljevog potprostora mogu se dobiti bez referenciranja na u varijable. Rješenje sustava u (2.96) dano je sa

$$u_k = u_0 + \sum_{i=1}^{k} v_i q_i$$

za $u_0 = Mx_0$, pri čemu su v_i, $i = 1, \ldots, k$ komponente nekog k-dimenzionalnog vektora v. Ako sve to pomnožimo sa M^{-1} dobivamo željenu aproksimaciju izraženu preko x varijabli

$$x_k = x_0 + M^{-1}\left(\sum_{i=1}^{k} v_i q_i\right).$$

Dakle algoritam za desno–prekondicioniranu GMRES metodu je ponovo sličan Algoritmu 2.5.2, samo što se na kraju x_k računa kao $x_0 = x_0 + M^{-1}Q_k v$ i gdje god se primijenjuje matrica sustava koristi se matrica AM^{-1}. U ovom slučaju Arnoldijev algoritam vraća ortogonalnu bazu desno–prekondicioniranog Krylovljevog potprostora

$$\mathcal{K}_k(AM^{-1}, r_0) = \text{span}\{r_0, AM^{-1}r_0, \ldots, (AM^{-1})^{k-1}r_0\}.$$

Primijetimo samo da je sada rezidul analogan rezidualu neprekondicioniranog sustava jer je $b - AM^{-1}u_k = b - Ax_k$.

Dvostrano prekondicioniranje pomoću faktora

U mnogim slučajevima M se može faktorizirati u oblik

$$M = LU.$$

Tada postoji mogućnost primijene GMRES metode na dvostrano prekondicioniranom sustavu

$$L^{-1}AU^{-1}u = L^{-1}b, \qquad x = U^{-1}u.$$

U ovakvoj situaciji, kao početni rezidual uzimamo neprekondicionirani početni rezidual pomnožen sa L^{-1}, a na kraju za formiranje aproksimacije x_k, vektor $Q_k y$ množimo sa U^{-1}. Rezidual u k-tom koraku je tada oblika $L^{-1}(b - Ax_k)$. Dvostrano prekondicioniranje često se koristi kada je matrica A skoro simetrična, kako se svojstvo simetričnosti nebi pokvarilo (kao kod CG).

Usporedba lijevog i desnog prekondicioniranja

Postavlja se pitanje da li postoji razlika između lijevog, desnog i dvostranog prekondicioniranja? Činjenica, da su kod raznih vrsta prekondicioniranja dostupne različite verzije reziduala, može utjecati na kriterij zaustavljanja, i može izazvati prerano ili prekasno zaustavljanje algoritma. To može biti prilično štetno ukoliko je M vrlo loše uvjetovana

matrica. Općenito postoje male razlike između ova tri odabira prekondicioniranja. Kada uspoređujemo lijevo, desno i dvostrano prekondicioniranje, prvo što možemo primijetiti je da su spektri triju matrica sustava $M^{-1}A$, AM^{-1} i $L^{-1}AU^{-1}$ identični. Zbog toga bi, u glavnom, mogli očekivati sličnu konvergenciju, iako znamo da svojstvene vrijednosti ne utječu uvijek na konvergenciju.

Kod lijevog prekondicioniranja, GMRES minimizira normu reziduala

$$\|M^{-1}b - M^{-1}Ax\|_2,$$

među svim vektorima iz afinog potprostora

$$x_0 + \mathcal{K}_k(M^{-1}A, z_0) = x_0 + \text{span}\{z_0, M^{-1}Az_0, \ldots, (M^{-1}A)^{k-1}z_0\} \tag{2.97}$$

u kojem je z_0 prekondicionirani početni rezidual $z_0 = M^{-1}r_0$. Tada se aproksimacija x_k može izraziti kao

$$x_k = x_0 + s_{k-1}(M^{-1}A)z_0$$

gdje je s_{k-1} polinom stupnja $k-1$ koji, prema Lemi 2.5.6, minimizira normu

$$\|z_0 - M^{-1}As(M^{-1}A)z_0\|_2$$

među svim polinomima s stupnja manjeg ili jednakog $k-1$. Ovaj uvjet minimizacije moguće je izraziti i preko originalnog vektora reziduala r_0.

$$z_0 - M^{-1}As(M^{-1}A)z_0 = M^{-1}[r_0 - As(M^{-1}A)M^{-1}r_0].$$

Jednostavna algebarska manipulacija pokazuje da za bilo koji polinom s vrijedi

$$s(M^{-1}A)M^{-1}r = M^{-1}s(AM^{-1})r, \tag{2.98}$$

odakle dobivamo relaciju

$$z_0 - M^{-1}As(M^{-1}A)z_0 = M^{-1}[r_0 - AM^{-1}s(AM^{-1})r_0]. \tag{2.99}$$

Razmotrimo sada slučaj desno prekondicionirane GMRES metode. Ovdje moramo obratiti pažnju na razliku između originalne varijable x i uvedene varijable u, koja je sa x vezana preko jednakosti $x = M^{-1}u$. Za varijablu u, desno prekondicionirani GMRES minimizira normu od $r = b - AM^{-1}u$, gdje u pripada afinom potprostoru

$$u_0 + \mathcal{K}_k(AM^{-1}, r_0) = u_0 + \text{span}\{r_0, AM^{-1}r_0, \ldots, (AM^{-1})^{k-1}r_0\} \tag{2.100}$$

u kojem je r_0 rezidual $r_0 = b - AM^{-1}u_0$. Ovaj rezidual je identičan rezidualu koji je izražen preko varijable x jer je $M^{-1}u_0 = x_0$. Množenjem (2.100) sa M^{-1} i ponovnim korištenjem identitete (2.98), dobivamo da varijabla x odgovarajuće varijable u iz potprostora (2.100), pripada afinom potprostoru

$$M^{-1}u_0 + M^{-1}\mathcal{K}_k(AM^{-1}, r_0) = x_0 + \text{span}\{z_0, M^{-1}Az_0, \ldots, (M^{-1}A)^{k-1}z_0\}.$$

To je identično sa afinim prostorom (2.97) koji se pojavljuje kod lijevog prekondicioniranja. Dakle, kod desno prekondicioniranog GMRES-a, aproksimacija x može se izraziti kao

$$x_k = x_0 + M^{-1}t_{k-1}(AM^{-1})r_0,$$

odnosno, također kao

$$x_k = x_0 + t_{k-1}(M^{-1}A)z_0.$$

Međutim, sada je t_{k-1} polinom $(k-1)$-og stupnja koji minimizira normu

$$\|r_0 - AM^{-1}t(AM^{-1})r_0\|_2 \tag{2.101}$$

među svim polinomima t stupnja manjeg ili jednakog $k-1$. Dvije veličine koje se minimiziraju u (2.99) i (2.101) razlikuju se samo za množenje sa matricom M^{-1}. Preciznije, lijevo prekondicionirani GMRES minimizira $M^{-1}r$, dok desno prekondicionirani algoritam minimizira r, gdje se r uzima iz istog potprostora u oba slučaja. Time smo pokazali sljedeći teorem.

Teorem 2.5.16 ([32]). *Aproksimacija rješenja ostvarena preko lijevo ili desno prekondicionirane GMRES metode je oblika*

$$x_k = x_0 + s_{k-1}(M^{-1}A)z_0 = x_0 + M^{-1}s_{k-1}(AM^{-1})r_0$$

gdje je $z_0 = M^{-1}r_0$, i s_{k-1} je polimom stupnja $k-1$. Polinom s_{k-1} minimizira normu reziduala $\|b - Ax_k\|_2$ u slučaju desnog prekondicioniranja, i normu prekondicioniranog reziduala $\|M^{-1}(b - Ax_k)\|_2$ u slučaju lijevog prekondicioniranja.

U mnogim praktičnim situacijama razlika u konvergenciji u ova dva slučaja nije velika. Jedina iznimka je slučaj kada je M loše uvjetovana, tada razlike mogu biti značajne, što se vidi i iz Teorem 2.5.16.

Napomena. Kod lijevog i desnog prekondicioniranja GMRES algoritam se jednostavno primijeni na prekondicionirani sustav, jedino se još na kraju, kod desno prekondicioniranog sustava treba posebno izračunati aproksimacija u x varijabli. U oba slučaja algoritam se ne može transformirati u jednostavniji oblik. Kod dvostranog prekondicioniranja, algoritam također primjenjujemo direktno na prekondicionirani sustav u općenitom slučaju, jer za faktorizaciju $M = LU$, za koju je $U \neq L^*$ Arnoldijev algoritam se ne može transformirati tako da izbjegnemo upotrebu faktora L i U. Ako uzmemo da je matrica prekondicioniranja M hermitska, i da je $M = LL^*$, tada se algoritam može transformirati tako da se upotrijebi množenje samo sa matricom M. To ima smisla samo kod hermitskih sustava, pa se izvod takvog algoritma nalazi u sljedećem odjeljku kod prekondicioniranja MINRES metode.

2.6 MINRES i CG preko Lanczosovog algoritma

2.6.1 Lanczosov algoritam

U ovom odjeljku promatrat ćemo isključivo hermitske sustave. Kada je matrica sustava A hermitska, Arnoldijev algoritam može se pojednostaviti do trokoračne rekurzije poznate pod imenom *Lanczosov algoritam*. Naime, $T_k = Q_k^* A Q_k$ je hermitska ali i Hessenbergova matrica, pa prema tome mora biti tridijagonalna. Matrica $Q_k = [q_1\ q_2\ \cdots\ q_k]$ je $n \times k$ matrica sa ortonormiranim stupcima q_i, $i = 1, \ldots, k$, za koju vrijedi $AQ_k = Q_k T_k + t_{k+1,k} q_{k+1} \xi_{k+1}$. Tridijagonalnost matrice T_k dosta pojednostavljuje algoritam.

Algoritam 2.6.1. LANCZOSOV ALGORITAM

Dan je vektor q_1 *sa* $\|q_1\|_2 = 1$, *i* $\beta_0 = 0$.

Za $j = 1, 2, \ldots, n-1$

$$\begin{aligned}
&\tilde{q}_{j+1} = Aq_j - \beta_{j-1}q_{j-1},\\
&\alpha_j = \langle \tilde{q}_{j+1}, q_j \rangle,\\
&\tilde{q}_{j+1} := \tilde{q}_{j+1} - \alpha_j q_j.\\
&\beta_j = \|\tilde{q}_{j+1}\|_2,\\
&q_{j+1} = \frac{\tilde{q}_{j+1}}{\beta_j}.
\end{aligned}$$

Pokažimo sada da vektori konstruirani ovim algoritmom čine ortonormiranu bazu Krylovljevog potprostora određenog sa matricom A i vektorom q_1. Prema samom Lanczosovom algoritmu vidi se da ti vektori zaista pripadaju Krylovljevom potprostoru i da su norme 1, zbog definicije skalara β_j. Iz definicije skalara α_j slijedi da je $\langle q_{j+1}, q_j \rangle = 0$. Zbog dokaza, kojeg ćemo izvesti matematičkom indukcijom, pretpostavimo da vrijedi $\langle q_k, q_i \rangle = 0$ za $i \neq j$ kada je $k, i \leq j$. Tada

$$\begin{aligned}
\langle \tilde{q}_{j+1}, q_{j-1} \rangle &= \langle Aq_j - \alpha_j q_j - \beta_{j-1}q_{j-1}, q_{j-1} \rangle = \langle Aq_j, q_{j-1} \rangle - \beta_{j-1} =\\
&= \langle q_j, Aq_{j-1} \rangle - \beta_{j-1} = \langle q_j, \tilde{q}_j + \alpha_{j-1}q_{j-1} + \beta_{j-2}q_{j-2} \rangle - \beta_{j-1} =\\
&= \langle q_j, \tilde{q}_j \rangle - \beta_{j-1} = \langle q_j, \beta_{j-1}q_j \rangle - \beta_{j-1} = 0.
\end{aligned}$$

Kako je $\tilde{q}_{j+1} = \beta_j q_{j+1}$, to znači da smo pokazali da je $\langle q_{j+1}, q_{j-1} \rangle = 0$. Za $i < j-1$, imamo

$$\begin{aligned}
\langle \tilde{q}_{j+1}, q_i \rangle &= \langle Aq_j - \alpha_j q_j - \beta_{j-1}q_{j-1}, q_i \rangle =\\
&= \langle Aq_j, q_i \rangle = \langle q_j, Aq_i \rangle =\\
&= \langle q_j, \tilde{q}_{i+1} + \alpha_i q_i + \beta_{i-1}q_{i-1} \rangle = 0.
\end{aligned}$$

Prema tome, vektori $q_1, \ldots, q_{j+1}$ čine ortonormiranu bazu Krylovljevog potprostora $\text{span}\{q_1, Aq_1, \ldots, A^j q_1\}$.

Kao i kod Arnoldijevog algoritma, Lanczosov algoritam može se u matričnom obliku napisati kao

$$AQ_k = Q_k T_k + \beta_k q_{k+1} \xi_k^T = Q_{k+1} T_{k+1,k}, \tag{2.102}$$

gdje je Q_k $n \times k$ matrica sa ortonormiranim stupcima $q_1, \ldots, q_k$, ξ_k je k-ti jedinični vektor, a T_k je $k \times k$ hermitska tridijagonalna matrica koeficijenata rekurzije:

$$T_k = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{k-1} \\ & & \beta_{k-1} & \alpha_k \end{bmatrix}. \tag{2.103}$$

Gornji $k \times k$ blok $(k+1) \times k$ matrice $T_{k+1,k}$ je T_k, a zadnji redak joj je $\beta_k \xi_k^T$.

Kao što je već prije pokazano MINRES i CG algoritmi u svakom koraku generiraju aproksimaciju rješenja x_k iz Krylovljevog potprostora za koju je euklidska norma reziduala, odnosno A-norma greške minimalna. Lanczosov algoritam, kao i Arnoldijev, generira ortonormiranu bazu $\{q_1, \ldots, q_k\}$ za Krylovljev potprostor definiran sa A i r_0, pri čemu je $q_1 = r_0/\beta$, $\beta = \|r_0\|_2$, to znači da gore spomenuti algoritmi generiraju aproksimacije rješenja oblika

$$x_k = x_0 + Q_k y_k, \tag{2.104}$$

gdje je y_k izabran tako da minimizira odgovarajuću normu greške.

2.6.2 MINRES

Za MINRES algoritam, y_k u (2.104) rješava problem najmanjih kvadrata

$$\begin{aligned} \min_{y\in\mathbb{C}^k} \|r_0 - AQ_k y\|_2 &= \min_{y\in\mathbb{C}^k} \|r_0 - Q_{k+1}T_{k+1,k}y\|_2 = \\ &= \min_{y\in\mathbb{C}^k} \|Q_{k+1}(\beta\xi_1 - T_{k+1,k}y)\|_2 = \\ &= \min_{y\in\mathbb{C}^k} \|\beta\xi_1 - T_{k+1,k}y\|_2, \end{aligned} \tag{2.105}$$

slično GMRES algoritmu, samo što je ovdje situacija jednostavnija zbog hermitičnosti matrice A, odnosno T_k. Naime kod MINRES algoritma nema potrebe za spremanjem svih ortonormiranih vektora dobivenih iz Lanczosovog algoritma, jer u svakom koraku algoritma baratamo samo sa tri poslijednja vektora. Nadalje, razmotrimo način na koji možemo pojednostavniti GMRES algoritam za hermitsku matricu. Neka je R_k gornji $k \times k$ blok gornje trokutastog faktora QR faktorizacije matrice $T_{k+1,k} = F^{(k)*}R^{(k)}$. Budući da je $T_{k+1,k}$ tridijagonalna matrica, i da koristimo uzastopno množenje Givensovim rotacijama za dobivanje QR faktorizacije, kao i kod GMRES algoritma, R_k ima samo tri netrivijalne dijagonale. Definirajmo $P_k = [p_0\ p_1\ \ldots p_{k-1}] = Q_k R_k^{-1}$, to možemo jer je R_k regularna ukoliko još nismo dostigli egzaktno rješnje. (vidi GMRES: $(R_k)_{k,k} = |(F^{(k-1)}T_k)_{k,k}|^2 + \beta_k^2$, za $F^{(k-1)} = F_{k-1}\cdots F_1$, i da bi to bilo jednako 0, mora biti i $\beta_k = (T_{k+1,k})_{k+1,k} = 0$, a u tom slučaju bi bilo $AQ_k = Q_kT_k$, pa bi zbog regularnosti matrice A matrica T_k i gornje trokutasta matrica $F^{(k-1)}T_k$ bile regularne. To povlači $(F^{(k-1)}T_k)_{k,k} \neq 0$, odnosno $(R_k)_{k,k} \neq 0$, što je kontradikcija. Dakle, $(R_k)_{k,k} \neq 0$, a matematičkom indukcijom možemo zaključiti da su i svi ostali dijagonalni elementi matrice R_k različiti od nule, pa je R_k regularna. Osim toga za $\beta_k = 0$ bi i norma reziduala bila jednaka nuli pa bismo u tom koraku već našli rješenje.) Ako sa $[0\ \ldots\ 0\ r_{j-3}^{(j-1)}\ r_{j-2}^{(j-1)}\ r_{j-1}^{(j-1)}\ 0\ \ldots\ 0]^T$ označimo elemente j-tog stupca matrice R_k, tada iz $P_kR_k = Q_k$ dobivamo da je $p_0 = (1/r_0^{(0)})q_1$, a ostali stupci mogu se dobiti iz sljedeće formule

$$p_{k-1} = \frac{1}{r_{k-1}^{(k-1)}}\left(q_k - r_{k-2}^{(k-1)}p_{k-2} - r_{k-3}^{(k-1)}p_{k-3}\right).$$

Zbog gornje opservacije, ako još nismo došli do rješenja, $r_{k-1}^{(k-1)} \neq 0$. Aproksimacija rješenja x_k tada se može dobiti iz x_{k-1} na sljedeći način

$$\begin{aligned} x_k &= x_0 + P_k\beta(F^{(k)}\xi_1)_{k\times 1} = \\ &= x_0 + \beta[P_{k-1} \quad p_{k-1}][(F^{(k-1)}\xi_1)_{(k-1)\times 1} \quad (F^{(k)}\xi_1)_k]^T = \\ &= x_0 + \beta P_{k-1}(F^{(k-1)}\xi_1)_{(k-1)\times 1} + \beta(F^{(k)}\xi_1)_k p_{k-1} = \\ &= x_{k-1} + \beta(F^{(k)}\xi_1)_k p_{k-1}, \end{aligned}$$

jer je $y_k = \beta R_k^{-1}(F^{(k)}\xi_1)_{k\times 1}$ kao kod GMRES-a, a F_k kad djeluje na vektor utječe samo na k-tu i $(k+1)$-u komponentu vektora. To rezultira sljedećom implementacijom MINRES algoritma.

Algoritam 2.6.2. MINRES (za hermitske matrice)

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

$\beta = \|r_0\|_2$,

$q_1 = \dfrac{r_0}{\beta}$,

$l = (1, 0, \ldots, 0)^T$.

Za $k = 1, 2, \ldots$

Izračunaj q_{k+1}, $\alpha_k = T(k,k)$ *i* $\beta_k = T(k+1,k) = T(k,k+1)$, *koristeći Lanczosov algoritam.*

Primijeni F_{k-2} *i* F_{k-1} *na zadnji stupac od* T, *odnosno:*

ako je $k > 2$ *tada* $\begin{bmatrix} T(k-2,k) \\ T(k-1,k) \end{bmatrix} := \begin{bmatrix} c_{k-2} & s_{k-2} \\ -\bar{s}_{k-2} & c_{k-2} \end{bmatrix} \begin{bmatrix} 0 \\ T(k-1,k) \end{bmatrix}$,

ako je $k > 1$ *tada* $\begin{bmatrix} T(k-1,k) \\ T(k,k) \end{bmatrix} := \begin{bmatrix} c_{k-1} & s_{k-1} \\ -\bar{s}_{k-1} & c_{k-1} \end{bmatrix} \begin{bmatrix} T(k-1,k) \\ T(k,k) \end{bmatrix}$.

Izračunaj k*-tu Givensovu rotaciju* F_k *kako bi se poništio* $(k+1,k)$ *element od* T*:*

$c_k = \dfrac{|T(k,k)|}{\sqrt{|T(k,k)|^2 + |T(k+1,k)|^2}}$,

ako je $c_k \neq 0$ *tada* $s_k = c_k \dfrac{\overline{T(k+1,k)}}{\overline{T(k,k)}}$, *ako je* $c_k = 0$ *tada* $s_k = 1$.

Primijeni k*-tu rotaciju na* l *i na zadnji stupac od* T*:*

$\begin{bmatrix} l(k) \\ l(k+1) \end{bmatrix} := \begin{bmatrix} c_k & s_k \\ -\bar{s}_k & c_k \end{bmatrix} \begin{bmatrix} l(k) \\ 0 \end{bmatrix}$,

$T(k,k) := c_k T(k,k) + s_k T(k+1,k)$,

$T(k+1,k) = 0 \qquad (*)$.

Izračunaj $p_{k-1} = (1/T(k,k))[q_k - T(k-1,k)p_{k-2} - T(k-2,k)p_{k-3}]$, *gdje su* p_{-1}, p_{-2}, *jednaki nuli.*

$x_k = x_{k-1} + \beta l(k) p_{k-1}$.

$(*)$ $T(k+1,k) = 0$ treba zapravo izvesti u sljedećem, $(k+1)$-om koraku, jer je originalna vrijednost $T(k+1,k)$ potrebna za izvođenje $(k+1)$-og koraka Lanczosovog algoritma.

2.6.3 Konjugirani gradijenti

Prema analizi razvoja metode konjugiranih gradijenata (CG), znamo da je y_k izabran tako da rezidual r_k bude ortogonalan na Krylovljev potprostor, odnosno na stupce matrice Q_k. Naime, za pozitivno definitnu matricu A biramo y_k koji minimizira A-normu greške $e_k = e_0 - Q_k y_k$, pa tada, zbog toga što tražimo najbolju aproksimaciju vektora e_0 u potprostoru unitarnog prostora sa $\|\cdot\|_A$ normom, kojeg razapinju stupci od Q_k, vrijedi da je e_k A-ortogonalan na stupce od Q_k, to jest $Q_k^* A e_k = Q_k^* r_k = 0$. Tada y_k za CG metodu zadovoljava

$$Q_k^*(r_0 - AQ_k y_k) = \beta\xi_1 - T_k y_k = 0, \tag{2.106}$$

odnosno, y_k je rješenje $k \times k$ linearnog sustava $T_k y = \beta\xi_1$. Dok problem najmanjih kvadrata (2.105) uvijek ima rješenje, linearan sustav (2.106) ima jedinstveno rješenje ako i samo ako je T_k regularna. Kada je A pozitvno definitna hermitska matrica, tada je tridijagonalna matrica $T_k = Q_k^* A Q_k$, također hermitska i pozitivno definitna, pa time i regularna, naravno ukoliko nismo došli do rješenja (ili kada je $\beta_k = 0$). Zbog toga je za matricu T_k izvediva faktorizacija Choleskog, odnosno T_k možemo svesti na oblik

$$T_k = L_k D_k L_k^*, \tag{2.107}$$

gdje su L_k donje trokutasta bidijagonalna matrica sa jediničnom dijagonalom, i D_k dijagonalna matrica. Ponovno ćemo iskoristiti hermitičnost matrice T_k kako ne bismo morali pamtiti sve vektore $q_1, \ldots, q_k$. Faktorizacija (2.107) u svakom koraku lako se može dobiti iz prethodnog, budući da su L_k i D_k glavne $k \times k$ podmatrice matrica L_{k+1} i D_{k+1}. Ako definiramo $P_k = [p_0\ p_1 \ldots p_{k-1}] = Q_k L_k^{-*}$, tada su stupci ovako definirane matrice P_k A-ortogonalni jer je

$$P_k^* A P_k = L_k^{-1} Q_k^* A Q_k L_k^{-*} = L_k^{-1} T_k L_k^{-*} = D_k,$$

a budući da P_k zadovoljava jednadžbu $P_k L_k^* = Q_k$, slijedi da je $p_0 = q_1$, i da se svi ostali stupci matrice P_k mogu dobiti preko rekurzije

$$p_{k-1} = q_k - \bar{b}_{k-1} p_{k-2},$$

gdje je b_{k-1} $(k, k-1)$ element sporedne dijagonale matrice L_k. Uzimajući u obzir ovu faktorizaciju, $x_k = x_0 + Q_k y_k$ za $y_k = \beta T_k^{-1}\xi_1$ je tada dan sa

$$\begin{aligned}
x_k &= x_0 + \beta Q_k T_k^{-1}\xi_1 = x_0 + \beta P_k D_k^{-1} L_k^{-1}\xi_1 = \\
&= x_0 + \beta[P_{k-1} \quad p_{k-1}][D_{k-1}^{-1} L_{k-1}^{-1}(\xi_1)_{(k-1)\times 1} \quad (D_k^{-1} L_k^{-1})_{k,1}]^T = \\
&= x_0 + \beta P_{k-1} D_{k-1}^{-1} L_{k-1}^{-1}(\xi_1)_{(k-1)\times 1} + \beta(D_k^{-1} L_k^{-1})_{k,1} p_{k-1} = \\
&= x_{k-1} + \beta(D_k^{-1} L_k^{-1})_{k,1} p_{k-1} =
\end{aligned}$$

Koeficijent $\beta(D_k^{-1} L_k^{-1})_{k,1}$ je definiran, ukoliko je L_k invertibilna matrica, i $(D_k)_{k,k} \neq 0$, što vrijedi ako je T_k regularna. Preostali su samo tehnički računi za dobivanje skalara b_k, $d_k = (D_k)_{k,k}$, i $(D_k^{-1} L_k^{-1})_{k,1} = d_k^{-1}(L_k^{-1})_{k,1}$.

Prema tome CG algoritam izveden preko Lanczosovog algoritma ima oblik

Algoritam 2.6.3. CG

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

$\beta = \|r_0\|_2$,

$q_1 = \dfrac{r_0}{\beta}$,

$c = 1, \qquad d = 1, \qquad f = 0.$

Za $k = 1, 2, \ldots$

Izračunaj q_{k+1}, $\alpha_k = T(k,k)$ *i* $\beta_k = T(k+1,k) = T(k,k+1)$, *koristeći Lanczosov algoritam.*

Izračunaj $p_{k-1} = q_k - \bar{c}p_{k-2}$, *gdje je* $p_{-1} = 0$.

$f = T(k,k) - f \cdot |c|^2$,

$x_k = x_{k-1} + \beta f^{-1} d p_{k-1}$.

$c = \dfrac{T(k+1,k)}{f}$,

$d = -c \cdot d$,

Promotrimo sada neka svojstva CG algoritma dobivenog preko Lanczosovog algoritma, i pokušajmo pronaći vezu između ovog algoritma (LCG) i CG algoritma dobivenog preko minimizacije kvadratne forme (kfCG). Najprije obratimo pažnju na izraz za rezidual u k-tom koraku LCG algoritma koji je sljedećeg oblika:

$$\begin{aligned} r_k &= r_0 - AQ_k y_k = \beta Q_k \xi_1 - Q_k T_k y_k - \beta_{k+1}(q_{k+1}\xi_k^T) y_k = \\ &= Q_k(\beta\xi_1 - T_k y_k) - \beta_{k+1}\xi_k^T y_k q_{k+1} = \\ &= -\beta_{k+1}\xi_k^T y_k q_{k+1}. \end{aligned} \tag{2.108}$$

Dakle, rezidual r_k je vektor istog smjera kao i q_{k+1}, što rezultira zaključkom kako su vektori reziduala međusobno ortogonalni. Uz to, već smo prije pokazali da su vektori p_k međusobno A-ortogonalni i da je $p_0 = q_1 = \beta r_0$. Iz (2.108) slijedi

$$-\beta_{k+1}\xi_k^T y_k p_k = r_k + \bar{b}_k \beta_{k+1}\xi_k^T y_k p_{k-1},$$

pa ako sa $\tilde{p}_k$ definiramo

$$\tilde{p}_k = -\beta_{k+1}\xi_k^T y_k p_k,$$

imamo

$$\tilde{p}_k = r_k + \tilde{\beta}_k \tilde{p}_{k-1},$$

pri čemu je

$$\tilde{\beta}_k = -\bar{b}_k \frac{\beta_{k+1}\xi_k^T y_k}{\beta_k \xi_{k-1}^T y_{k-1}}.$$

Za $k = 1$ $\tilde{p}_0 = \beta r_0$ i $\tilde{\beta}_1 = -(\beta_2 \xi_1^T y_1)/\beta$. Nadalje je

$$x_k = x_{k-1} + \tilde{\alpha}_{k-1}\tilde{p}_{k-1},$$

pri čemu je

$$\tilde{\alpha}_{k-1} = -\frac{\beta(D_k^{-1}L_k^{-1})_{k,1}}{\beta_k \xi_{k-1}^T y_{k-1}},$$

a za $k = 1$ je $\tilde{\alpha}_0 = -\beta(D_1 L_1)_{1,1}/\beta$. Time smo dobili analogne rekurzivne jednadžbe za x_k i $\tilde{p}_k$ kao i kod kfCG, kod kojih su koeficijenti $\tilde{\alpha}_k = \langle r_k, r_k\rangle/\langle A\tilde{p}_k, \tilde{p}_k\rangle$ i $\tilde{\beta}_k = \langle r_k, r_k\rangle/\langle r_{k-1}, r_{k-1}\rangle$ analogno definirani zbog ortogonalnosti vektora r_k i A-ortogonalnosti vektora $\tilde{p}_k$.

2.6.4 Prekondicionirani Lanczosov algoritam

Budući da se Lanczosov algoritam primjenjuje za rješavanje sustava sa hermitskom matricom, prekondicionirani sustav bi treba zadržati to svojstvo. Dakle, prije svega prekondicioniranje treba biti hermitsko, s time da zbog zadržavanja hermitičnosti matrica prekondicioniranja morala bi imati faktorizaciju oblika $M = LL^*$. Za to bi najpogodnija bila pozitivno definitna matrica prekondicioniranja.

Sada promatramo primjenu Lanczosovog agoritma na prekondicionirani sustav $L^{-1}A\cdot$ $\cdot L^{-*}u = L^{-1}b$, gdje je $x = L^{-*}u$. Rezidual prekondicioniranog sustava, kojeg ćemo označiti sa $\hat{r}_k$ u k-tom koraku, sa rezidualom početnog sustava $Ax = b$, kojeg označavamo sa r_k, je u vezi preko relacije

$$\hat{r}_k = L^{-1}r_k,$$

čije norma je oblika

$$\|\hat{r}_k\|_2 = \sqrt{\langle L^{-1}r_k, L^{-1}r_k\rangle} = \sqrt{\langle r_k, M^{-1}r_k\rangle}.$$

Nadalje, ako sa q_k označimo vektore koji čine ortonormiranu bazu Krylovljevog potprostora prekondicioniranog sustava tada prvi korak j-te iteracije Lanczosovog algoritma daje

$$\tilde{q}_{j+1} = L^{-1}AL^{-*}q_j - \beta_{j-1}q_{j-1}.$$

Ako sada označimo sa $v_i = Lq_i$ i $\tilde{v}_i = L\tilde{q}_i$ za svako i, tada imamo

$$\tilde{v}_{j+1} = AM^{-1}v_j - \beta_j v_{j-1}.$$

sljedeći korak je

$$\alpha_j = \langle \tilde{q}_{j+1}, q_j\rangle = \langle L^{-1}\tilde{v}_{j+1}, L^{-1}v_j\rangle = \langle \tilde{v}_{j+1}, M^{-1}v_j\rangle,$$

pa je zbog

$$\tilde{q}_{j+1} = \tilde{q}_{j+1} - \alpha_j q_j,$$

$$\tilde{v}_{j+1} = \tilde{v}_{j+1} - \alpha_j v_j.$$

I na kraju,

$$\beta_j = \|\tilde{q}_{j+1}\|_2 = \sqrt{\langle L^{-1}\tilde{v}_{j+1}, L^{-1}\tilde{v}_{j+1}\rangle} = \sqrt{\langle \tilde{v}_{j+1}, M^{-1}\tilde{v}_{j+1}\rangle},$$

i

$$v_{j+1} = \frac{\tilde{v}_{j+1}}{\beta_j}.$$

Ako još uvedemo dvije nove varijable $w_j = M^{-1}v_j$ i $\tilde{w}_{j+1} = M^{-1}\tilde{v}_{j+1}$, tada prekondicionirani Lanczosov algoritam možemo napisati na sljedeći način:

Algoritam 2.6.4. PREKONDICIONIRANI LANCZOSOV ALGORITAM (ZA HERMITSKE MATRICE A SA POZITIVNO DEFINITNOM MATRICOM PREKONDICIONIRANJA M).

Dan je vektor r_0,

riješi $M\tilde{w}_1 = r_0$,

$\beta_0 = \sqrt{\langle r_0, \tilde{w}_1 \rangle}$,

$v_1 = \dfrac{r_0}{\beta_0}$,

$w_1 = \dfrac{\tilde{w}_1}{\beta_0}$,

$v_0 = 0$.

Za $j = 1, 2, \ldots, n-1$

$\tilde{v}_{j+1} = Aw_j - \beta_{j-1}v_{j-1}$,
$\alpha_j = \langle \tilde{v}_{j+1}, w_j \rangle$,
$\tilde{v}_{j+1} := \tilde{v}_{j+1} - \alpha_j v_j$,
riješi $M\tilde{w}_{j+1} = \tilde{v}_{j+1}$,
$\beta_j = \sqrt{\langle \tilde{v}_{j+1}, \tilde{w}_{j+1} \rangle}$,
$v_{j+1} = \dfrac{\tilde{v}_{j+1}}{\beta_j}$,
$w_{j+1} = \dfrac{\tilde{w}_{j+1}}{\beta_j}$.

Ako se Algoritam 2.6.2 i Algoritam 2.6.3 primijene direktno na prekondicionirani sustav i ako sa $\hat{x}_k$ i $\hat{p}_k$ označimo iteraciju i vektor smjera generirani pomoću danih algoritama, i ako tada uzmemo da je $x_k = L^{-*}\hat{x}_k$ i $p_k = L^{-*}\hat{p}_k$, tada prekondicionirani algoritmi imaju oblik :

Algoritam 2.6.5. PREKONDICIONIRANI MINRES (ZA HERMITSKE MATRICE SA POZITIVNO DEFINITNOM MATRICOM PREKONDICIONIRANJA)

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

riješi $Mz_0 = r_0$,

$\beta = \sqrt{\langle r_0, z_0 \rangle}$,

$v_1 = \dfrac{r_0}{\beta}$,

$w_1 = \dfrac{z_0}{\beta}$,

$l = (1, 0, \ldots, 0)^T$.

Za $k = 1, 2, \ldots$

Izračunaj v_{k+1}, w_{k+1}, $\alpha_k = T(k,k)$ *i* $\beta_k = T(k+1,k) = T(k,k+1)$, *koristeći prekondicionirani Lanczosov algoritam.*

Primijeni F_{k-2} *i* F_{k-1} *na zadnji stupac od* T, *odnosno:*

ako je $k > 2$ *tada* $\begin{bmatrix} T(k-2,k) \\ T(k-1,k) \end{bmatrix} := \begin{bmatrix} c_{k-2} & s_{k-2} \\ -\bar{s}_{k-2} & c_{k-2} \end{bmatrix} \begin{bmatrix} 0 \\ T(k-1,k) \end{bmatrix}$,

ako je $k > 1$ *tada* $\begin{bmatrix} T(k-1,k) \\ T(k,k) \end{bmatrix} := \begin{bmatrix} c_{k-1} & s_{k-1} \\ -\bar{s}_{k-1} & c_{k-1} \end{bmatrix} \begin{bmatrix} T(k-1,k) \\ T(k,k) \end{bmatrix}$.

Izračunaj k*-tu Givensovu rotaciju* F_k *kako bi se poništio* $(k+1,k)$ *element od* T*:*

$c_k = \dfrac{|T(k,k)|}{\sqrt{|T(k,k)|^2 + |T(k+1,k)|^2}}$,

ako je $c_k \neq 0$ *tada* $s_k = c_k \dfrac{\overline{T(k+1,k)}}{\overline{T(k,k)}}$, *ako je* $c_k = 0$ *tada* $s_k = 1$.

Primijeni k*-tu rotaciju na* l *i na zadnji stupac od* T*:*

$\begin{bmatrix} l(k) \\ l(k+1) \end{bmatrix} := \begin{bmatrix} c_k & s_k \\ -\bar{s}_k & c_k \end{bmatrix} \begin{bmatrix} l(k) \\ 0 \end{bmatrix}$,

$T(k,k) := c_k T(k,k) + s_k T(k+1,k)$,

$T(k+1,k) = 0 \qquad (*)$.

Izračunaj $p_{k-1} = (1/T(k,k))[w_k - T(k-1,k)p_{k-2} - T(k-2,k)p_{k-3}]$, *gdje su* p_{-1}, p_{-2}, *jednaki nuli.*

$x_k = x_{k-1} + \beta l(k) p_{k-1}$.

$(*)$ $T(k+1,k) = 0$ treba zapravo izvesti u sljedećem, $(k+1)$-om koraku, jer je originalna vrijednost $T(k+1,k)$ potrebna za izvođenje $(k+1)$-og koraka Lanczosovog algoritma.

Algoritam 2.6.6. Prekondicionirani CG (sa pozitivno definitnom matricom prekondicioniranja)

Dana je početna iteracija x_0,

riješi $Mz_0 = r_0$,

$r_0 = b - Ax_0$,

$\beta = \sqrt{\langle r_0, z_0 \rangle}$,

$v_1 = \dfrac{r_0}{\beta}$,

$w_1 = \dfrac{z_0}{\beta}$,

$c = 1, \qquad d = 1, \qquad f = 0.$

Za $k = 1, 2, \ldots$

Izračunaj v_{k+1}, w_{k+1}, $\alpha_k = T(k,k)$ *i* $\beta_k = T(k+1,k) = T(k,k+1)$, *koristeći prekondicionirani Lanczosov algoritam.*

Izračunaj $p_{k-1} = w_k - \bar{c}p_{k-2}$, *gdje je* $p_{-1} = 0$.

$f = T(k,k) - f \cdot |c|^2$,

$x_k = x_{k-1} + \beta f^{-1} d p_{k-1}$.

$c = \dfrac{T(k+1,k)}{f}$,

$d = -c \cdot d$,

2.7 BCG i srodne metode

Budući da GMRES metoda za nehermitske probleme svakom iteracijom povećava količinu posla kojeg se treba obaviti i memorije koju treba zauzeti, pojavila se potreba za razvojem drugačijih metoda koje bi zahtijevale fiksnu količinu posla i memorije po iteraciji. Takve metode, kao što će se pokazati, ipak imaju neke nedostatke. Na primjer, one mogu reducirati euklidsku normu reziduala na određenu veličinu sa više iteracija nego standardne metode. Osim toga, ovi algoritmi mogu zakazati, iako se to može izbjeći tzv. *provjerama unaprijed*. Međutim, provjerama unaprijed ponovo gubimo svojstvo fiksnog posla i memorije, pa zahtjev za njima raste iz iteracije u iteraciju kao kod GMRES-a.

2.7.1 Dvostrani Lanczosov algoritam

Dvostrani Lanczosov algoritam je proširenje simetričnog Lanczosovog algoritma na nesimetrične matrice, tako da se, umjesto jedne trokoračne rekurzije, dobiva par trokoračnih rekurzija, jedna vezana uz A, a druga uz A^*. Za razliku od Arnoldijevog algoritma, ova metoda konstruira *biortogonalne* baze za dva Krylovljeva potprostora, jednog definira-

nog za matricu A

$$\mathcal{K}_k(A, v_1) = \text{span}\{v_1, Av_1, \ldots, A^{k-1}v_1\},$$

i drugog definiranog za matricu A^*

$$\mathcal{K}_k(A^*, w_1) = \text{span}\{w_1, A^*w_1, \ldots, (A^*)^{k-1}w_1\}.$$

Biortogonalnost se očituje u svojstvu

$$\langle v_i, w_j \rangle = 0, \qquad \text{za } i \neq j.$$

Algoritam je sljedeći.

Algoritam 2.7.1. DVOSTRANI LANCZOSOV ALGORITAM

Dani su vektori v_1 *i* w_1 *sa* $\|v_1\|_2 = 1$ *i* $\langle v_1, w_1 \rangle = 1$.

Neka su $\beta_0 = \gamma_0 = 0$ *i* $v_0 = w_0 = 0$.

Za $j = 1, 2, \ldots$

Izračunaj Av_j i A^*w_j,
$\alpha_j = \langle Av_j, w_j \rangle$,
$\tilde{v}_{j+1} = Av_j - \alpha_j v_j - \beta_{j-1} v_{j-1}$,
$\tilde{w}_{j+1} = A^*w_j - \bar{\alpha}_j w_j - \gamma_{j-1} w_{j-1}$,
$\gamma_j = \|\tilde{v}_{j+1}\|_2$,
$v_{j+1} = \dfrac{\tilde{v}_{j+1}}{\gamma_j}$,
$\beta_j = \langle v_{j+1}, \tilde{w}_{j+1} \rangle$,
ako je $\beta_j = 0$ *stani, inače*
$w_{j+1} = \dfrac{\tilde{w}_{j+1}}{\bar{\beta}_j}$.

U ovom algoritmu vektori baze su skalirani tako da vektori v_j imaju normu 1, i da je $\langle v_j, w_j \rangle = 1$.

Neka je V_k matrica sa stupcima $v_1, \ldots, v_k$ i W_k matrica sa stupcima $w_1, \ldots, w_k$, tada se par rekurzija iz gornjeg algoritma može napisati u matričnom obliku kao

$$AV_k = V_k T_k + \gamma_k v_{k+1} \xi_k^T = V_{k+1} T_{k+1,k}, \tag{2.109}$$

$$A^*W_k = W_k T_k^* + \bar{\beta}_k w_{k+1} \xi_k^T = W_{k+1} \hat{T}_{k+1,k}, \tag{2.110}$$

gdje je T_k $k \times k$ tridijagonalna matrica koeficijenata rekurzije

$$T_k = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \gamma_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{k-1} \\ & & \gamma_{k-1} & \alpha_k \end{bmatrix}.$$

$(k+1)\times k$ matrice $T_{k+1,k}$ i $\hat{T}_{k+1,k}$ imaju T_k odnosno T_k^* kao gornji $k\times k$ blok, i zadnji red popunjen s nulama osim $(k+1,k)$ elementa koji je jednak γ_k odnosno $\bar{\beta}_k$. Svojstvo biortogonalnosti matrično se može napisati kao

$$V_k^* W_k = I.$$

Primijetimo da ako je $A = A^*$ i $w_1 = v_1$, tada se dvostrani Lanczosov algoritam svodi na obični Lanczosov algoritam za hermitske matrice. Također primijetimo, da prema algoritmu vektori v_j pripadaju potprostoru $\mathcal{K}_k(A, v_1)$, a vektori w_j su u $\mathcal{K}_k(A^*, w_1)$. Zapravo, može se pokazati sljedeći teorem.

Teorem 2.7.2 ([12]). *Ako dvostrani Lanczosov algoritam ne zakaže do k-tog koraka, a to znači da su definirani svi vektori v_j i w_j iz dvostranog Lanczosovog algoritma, odnosno da je $\langle v_j, w_j\rangle \neq 0$ za $j = 1,\ldots,k+1$, tada je*

$$\langle v_i, w_j\rangle = 0 \quad \textit{za svake } i,j \leq k+1, \quad i \neq j.$$

Dokaz: Dokaz se provodi indukcijom. Pretpostavimo da tvrdnja teorema vrijedi za $i,j \leq k$. Izbor koeficijenata β_j i γ_j u algoritmu osigurava da je za sve j $\langle v_j, w_j\rangle = 1$ i da je $\|v_j\|_2 = 1$. Prema definiciji koeficijenta α_k, zbog pretpostavke indukcije, imamo

$$\begin{aligned}
\langle \tilde{v}_{k+1}, w_k\rangle &= \langle Av_k, w_k\rangle - \alpha_k\langle v_k, w_k\rangle - \beta_{k-1}\langle v_{k-1}, w_k\rangle = \langle Av_k, w_k\rangle - \alpha_k = 0\\
\langle \tilde{w}_{k+1}, v_k\rangle &= \langle A^* w_k, v_k\rangle - \bar{\alpha}_k\langle w_k, v_k\rangle - \gamma_{k-1}\langle w_{k-1}, v_k\rangle = \overline{\langle Av_k, w_k\rangle} - \bar{\alpha}_k = 0
\end{aligned}$$

Iz rekurzivnih jednakosti za $\tilde{v}_{k+1}$ i $\tilde{w}_k$ zajedno sa pretpostavkom indukcije imamo

$$\begin{aligned}
\langle \tilde{v}_{k+1}, w_{k-1}\rangle &= \langle Av_k, w_{k-1}\rangle - \alpha_k\langle v_k, w_{k-1}\rangle - \beta_{k-1}\langle v_{k-1}, w_{k-1}\rangle =\\
&= \langle Av_k, w_{k-1}\rangle - \beta_{k-1} = \langle v_k, A^* w_{k-1}\rangle - \beta_{k-1} =\\
&= \langle v_k, \tilde{w}_k\rangle + \alpha_{k-1}\langle v_k, w_{k-1}\rangle + \gamma_{k-2}\langle v_k, w_{k-2}\rangle - \beta_{k-1} =\\
&= \langle v_k, \tilde{w}_k\rangle - \beta_{k-1} = 0,
\end{aligned}$$

a slično slijedi i da je $\langle \tilde{w}_{k+1}, v_{k-1}\rangle = 0$. I na kraju, za $j < k-1$, imamo

$$\begin{aligned}
\langle \tilde{v}_{k+1}, w_j\rangle &= \langle Av_k, w_j\rangle - \alpha_k\langle v_k, w_j\rangle - \beta_{k-1}\langle v_{k-1}, w_j\rangle =\\
&= \langle Av_k, w_j\rangle = \langle v_k, A^* w_j\rangle =\\
&= \langle v_k, \tilde{w}_{j+1}\rangle + \alpha_j\langle v_k, w_j\rangle + \gamma_{j-1}\langle v_k, w_{j-1}\rangle = \beta_j\langle v_k, w_{j+1}\rangle = 0,
\end{aligned}$$

slično slijedi i $\langle \tilde{w}_{k+1}, v_j\rangle = 0$. Budući da su v_{k+1} i w_{k+1} samo multipli od $\tilde{v}_{k+1}$ i $\tilde{w}_{k+1}$ tvrdnja teorema je dokazana za svako k. □

U ovom algoritmu matrice A i A^* imaju dualne uloge, jer su nad njima izvršene slične operacije. Zapravo, indirektno se rješavaju dva linearna sustava, jedan sa A i drugi sa A^*. Ako to zaista trebamo, onda je ovaj algoritam pogodan za takve zahtjeve, međutim ukoliko nam ne treba rješenje sustava sa matricom A^*, operacije s njom su zapravo potrošene uzalud.

Sa praktičnog stanovišta dvostrani Lanczosov algoritam ima značajnu prednost nad Arnoldijevim algoritmom jer zahtijeva pamćenje samo nekoliko vektora, konkretnije šest vektora duljine n, plus okupacija memorija za smještaj tridijagonalne matrice.

S druge strane, u ovom algoritmu postoje potencijalne mogućnosti zakazivanja, a to je slučaj kada Lanczosovi vektori, iz nekog razloga nisu definirani, to jest kada je

$\beta_j = 0$. U praksi se problemi često javljaju kada se dogodi "skoro" zakazivanje, to jest kada se Lanczosovi vektori moraju skalirati sa malim koeficijentima. Nakon nekoliko koraka akumulirani efekt tih skaliranja može dovesti do pojave prekomjernih grešaka zaokruživanja. Postoje dvije različite situacije kada može doći do zakazivanja. Prvo, ako je $\tilde{v}_{j+1} = 0$ ili $\tilde{w}_{j+1} = 0$, tada je dvostrani Lanczosov algoritam pronašao invarijantni potprostor. Ako je $\tilde{v}_{j+1} = 0$, tada Lanczosovi vektori $v_1, \ldots, v_j$ razapinju A-invarijantni potprostor, a kao što ćemo kasnije vidjeti, BICG algoritam dostići će egzaktno rješenje. Ako je $\tilde{w}_{j+1} = 0$, tada Lanczosovi vektori $w_1, \ldots, w_j$ razapinju A^*-invarijantni potprostor. U toj situaciji ništa ne možemo reći o aproksimaciji rješenja susutava sa matricom A. S druge strane, ako se algoritam koristi za rješavanje para linearnih susutava, jednog sa A, i drugog dualnog sustava sa A^*, tada u tom slučaju aproksimacija dualnog sustava je egzaktno rješenje. Ovakav slučaj označavamo kao *regularno zaustavljanje*.

Drugi slučaj, kojeg označavamo kao *ozbiljni slom algoritma*, nastupa kada je $\langle \tilde{v}_{j+1}, \tilde{w}_{j+1} \rangle = 0$, ali niti jedan od vektora $\tilde{v}_{j+1}$ i $\tilde{w}_{j+1}$ nije jednak nuli. U tom slučaju, netrivijalni vektori $v_{j+1} \in \mathcal{K}_{j+1}(A, v_1)$ i $w_{j+1} \in \mathcal{K}_{j+1}(A^*, w_1)$ sa svojstvom da je $\langle v_{j+1}, w_i \rangle = \langle w_{j+1}, v_i \rangle = 0$ za $i \leq j$ jednostavno ne postoje. Primijetimo da iako takvi vektori ne moraju postojati u $(j+1)$-om koraku, u nekom kasnijem $(j+k)$-tom koraku oni mogu postojati. Procedure koje jednostavno preskaču korake u kojima su Lanczosovi vektori nedefinirani, i koji omogučavaju da se algoritam nastavi izvršavati u večini slučajeva, označavamo kao *Lanczosove algoritme sa provjerama unaprijed*. U nastavku biti će prikazan Lanczosov algoritam sa provjerom unaprijed kako su ga opisali Parlett, Taylor i Liu u [31].

Glavna ideja koju koristi Lanczosov algoritam sa provjerama unaprijed je da par v_{j+2}, w_{j+2} često može biti definiran iako v_{j+1}, w_{j+1} nisu definirani. Ako pak niti par v_{j+2}, w_{j+2} nije definiran, onda se može pokušati sa parom v_{j+3}, w_{j+3}, itd. Da bi bolje opisali ideju provjera unaprijed, potrebno je vratiti se na vezu Lanczosovih vektora sa polinomima. Prema definiciji dvostranog Lanczosovog algoritma, ukoliko ne dođe do zakazivanja, u k tom koraku imamo definirane vektore $v_1, \ldots, v_{k+1}$ i $w_1, \ldots, w_{k+1}$ koje možemo napisati na sljedeći način

$$v_{j+1} = \frac{1}{\prod_{i=1}^{j} \gamma_i} p_j(A) v_1,$$

$$w_{j+1} = \frac{1}{\prod_{i=1}^{j} \bar{\beta}_i} \bar{p}_j(A^*) w_1,$$

pri čemu je $p_j \in \mathbb{P}_j$ polinom koji je rekurzivno zadan sa

$$p_0(t) = 1,$$

$$p_j(t) = (t - \alpha_j) p_{j-1}(t) - \beta_{j-1}\gamma_{j-1} p_{j-2}(t),$$

a lako se može pokazati da je p_j karakteristični polinom trodijagonalne matrice T_j. Dakle isti se polinom pojavljuje u definicijama za v_{j+1} i w_{j+1}, samo što se kod w_{j+1} radi o konjugiranom i skaliranom polinomu p_j iz definicije vektora v_{j+1}. Ako definiramo Krylovljeve matrice

$$K_{k+1} = [v_1 \ Av_1 \ \ldots \ A^k] \quad \text{i} \quad N_{k+1} = [w_1 \ A^* w_1 \ \ldots \ (A^*)^k w_1],$$

tada prethodne rekurzije za polinome možemo napisati u sljedećem matričnom obliku

$$V_{k+1} = K_{k+1} Z_{k+1}^{-1} G_{k+1}^{-1},$$

$$W_{k+1} = N_{k+1}\bar{Z}_{k+1}^{-1}\bar{H}_{k+1}^{-1},$$

gdje su

$$G_{k+1} = \mathrm{diag}(1, \gamma_1, \gamma_1\gamma_2, \ldots, \prod_{i=1}^{k}\gamma_k), \qquad H_{k+1} = \mathrm{diag}(1, \beta_1, \beta_1\beta_2, \ldots, \prod_{i=1}^{k}\beta_i),$$

a Z_{k+1} je gornje trokutasta matrica sa jediničnom dijagonalom, takva da matrica Z_{k+1}^{-1} ima smještene koeficijente polinoma p_j u j-tom stupcu, iznad dijagonale. Sada definirajmo matricu momenta kao

$$M_{k+1} = N_{k+1}^* K_{k+1},$$

gdje je $(M_{k+1})_{i+1,j+1} = w_1^* A^{i+j} v_1$. Budući da vrijedi $W_{k+1}^* V_{k+1} = I$, imamo

$$I = H_{k+1}^{-1} Z_{k+1}^{-T} N_{k+1}^* K_{k+1} Z_{k+1}^{-1} G_{k+1}^{-1},$$

odakle slijedi

$$M_{k+1} = Z_{k+1}^T D_{k+1} Z_{k+1}, \tag{2.111}$$

za $D_{k+1} = H_{k+1}G_{k+1}$. Ako sada označimo matrice $L_{k+1} = Z_{k+1}^T$ i $U_{k+1} = D_{k+1}Z_{k+1}$ tada vidimo da je dvostrani Lanczosov algoritam ekvivalentan računanju LU faktorizacije matrice momenta M_{k+1}, odnosno računanju $M_{k+1} = L_{k+1}U_{k+1}$, gdje je produkt $d_j = \prod_{i=1}^{j-1}(\beta_i\gamma_i)$ j-ti pivotni element u izvođenju Gaussovih eliminacija na matrici M_{k+1}.

Dakle, kod pojave ozbiljnog sloma, kada je $\beta_k = 0$ ili vrlo mali, pivotni element je tada isto jednak ili približno jednak nuli. Da bi se izbjeglo dijeljenje sa nulom kao pivotni element se tada uzima 2×2 matrica umjesto 1×1, pa se onda istovremeno računaju v_{k+1} i v_{k+2}, kao i w_{k+1} i w_{k+2}. Znači, nakon k-tog koraka standardnog algoritma imamo

$$\tilde{v}_{k+1} = Av_k - \alpha_k v_k - \beta_{k-1}v_{k-1} = \frac{1}{\prod_{i=1}^{k-1}\gamma_k} p_k(A)v_1,$$

$$\tilde{w}_{k+1} = A^* w_k - \bar{\alpha}_k w_k - \gamma_{k-1}w_{k-1} = \frac{1}{\prod_{i=1}^{k-1}\bar{\beta}_k} p_k(A^*)w_1.$$

Umjesto normaliziranja $\tilde{v}_{k+1}$ i $\tilde{w}_{k+1}$ za dobivanje vektora v_{k+1} i w_{k+1} mi ćemo potražiti bilo koja dva vektora $\tilde{v}_{k+2}$ i $\tilde{w}_{k+2}$ takve da je

$$\mathrm{span}\{\tilde{v}_{k+1}, \tilde{v}_{k+2}\} = \mathrm{span}\{v_{k+1}, v_{k+2}\},$$

i

$$\mathrm{span}\{\tilde{w}_{k+1}, \tilde{w}_{k+2}\} = \mathrm{span}\{w_{k+1}, w_{k+2}\}.$$

Najjednostavniji izbor je

$$\tilde{v}_{k+2} = A\tilde{v}_{k+1} - \omega_{k+1}v_k,$$

i

$$\tilde{w}_{k+2} = A^*\tilde{w}_{k+1} - \bar{\omega}_{k+1}w_k.$$

Koeficijent ω_k osigurava ortogonalnost $\tilde{v}_{k+2}$ sa $w_1, \ldots, w_k$, i $\tilde{w}_{k+2}$ sa $v_1, \ldots, v_k$. Naime, za $j < k$ to već vrijedi jer

$$\langle \tilde{v}_{k+2}, w_j\rangle = \langle A\tilde{v}_{k+1}, w_j\rangle - \omega_{k+1}\langle v_k, w_j\rangle = \langle \tilde{v}_{k+1}, A^* w_j\rangle = 0,$$

zbog toga što je $A^*w_j \in \mathcal{K}_{j+1}(A^*, w_1) \subseteq \mathcal{K}_k(A^*, w_1)$, a $\tilde{v}_{k+1}$ je okomit na $\mathcal{K}_k(A^*, w_1)$ prema definiciji. Analogno vrijedi i za $\tilde{w}_{k+2}$ i v_j. Koeficijent ω_{k+1} dobije se iz uvjeta ortogonalnosti $\tilde{v}_{k+2}$ i w_k ili $\tilde{w}_{k+2}$ sa v_k, gdje u oba slučaja ispada da je $\omega_{k+1} = \langle \tilde{v}_{k+1}, \tilde{w}_{k+1} \rangle$. Na primjer,

$$0 = \langle \tilde{v}_{k+2}, w_k \rangle = \langle A\tilde{v}_{k+1}, w_k \rangle - \omega_{k+1} \langle v_k, w_k \rangle,$$

pa slijedi, zbog prethodno provjerenih ortogonalnosti, da je

$$\omega_{k+1} = \langle \tilde{v}_{k+1}, A^* w_k \rangle = \langle \tilde{v}_{k+1}, \tilde{w}_{k+1} \rangle.$$

Iz prethodnih definicija također se može provjeriti da je

$$\theta_{k+1} = \langle \tilde{v}_{k+2}, \tilde{w}_{k+1} \rangle = \langle \tilde{v}_{k+1}, \tilde{w}_{k+2} \rangle = \langle A\tilde{v}_{k+1}, \tilde{w}_{k+1} \rangle.$$

Sada definirajmo sljedeće matrice

$$\widetilde{V}_{k+1,k+2} = [\tilde{v}_{k+1} \quad \tilde{v}_{k+2}], \qquad \widetilde{W}_{k+1,k+2} = [\tilde{w}_{k+1} \quad \tilde{w}_{k+2}],$$

i matricu

$$C_{k+1} = \widetilde{W}^*_{k+1,k+2} \widetilde{V}_{k+1,k+2} = \begin{bmatrix} \omega_{k+1} & \theta_{k+1} \\ \theta_{k+1} & \omega_{k+2} \end{bmatrix}.$$

Ideja je sljedeća: ukoliko je C_{k+1} regularna, trebamo pronaći neku njenu faktorizaciju

$$C_{k+1} = \widetilde{W}^*_{k+1,k+2} \widetilde{V}_{k+1,k+2} = R_{k+1} S_{k+1},$$

koju možemo preformulirati u

$$I = R^{-1}_{k+1} \widetilde{W}^*_{k+1,k+2} \widetilde{V}_{k+1,k+2} S^{-1}_{k+1} = (\widetilde{W}_{k+1,k+2} R^{-*}_{k+1})(\widetilde{V}_{k+1,k+2} S^{-1}_{k+1}).$$

Ako sada definiramo $V_{k+1,k+2} = [v_{k+1} \quad v_{k+2}]$ i $W_{k+1,k+2} = [w_{k+1} \quad w_{k+2}]$ sa

$$V_{k+1,k+2} = \widetilde{V}_{k+1,k+2} S^{-1}_{k+1}, \qquad W_{k+1,k+2} = \widetilde{W}_{k+1,k+2} R^{-*}_{k+1},$$

Dobili smo tražene vektore v_{k+1}, v_{k+2}, w_{k+1} i w_{k+2} za koje vrijedi

$$W^*_{k+1,k+2} V_{k+1,k+2} = I.$$

Na kraju, dvostrani Lanczosov algoritam sa provjerom u naprijed kreira blok tridijagonalnu matricu T_j

$$T_j = \begin{bmatrix} A_1 & B_1 & & \\ \Gamma_1 & \ddots & \ddots & \\ & \ddots & \ddots & B_{j-1} \\ & & \Gamma_{j-1} & A_j \end{bmatrix}.$$

Dijagonalni blokovi A_k su ili 1×1 ili 2×2 matrice, a matrice B_k i Γ_k su oblikovane adekvatno tome. Matrice Γ_k imaju jedan od sljedećih oblika

$$[x], \qquad [0 \quad x], \qquad \begin{bmatrix} x \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} 0 & x \\ 0 & 0 \end{bmatrix},$$

gdje je x neki pozitivan broj. Također ispada da i matrice B_k imaju rang 1. Lijevi i desni Lancosovi vektori grupiraju se u svakom koraku, što označavamo isto sa $v_1, \ldots, v_j$

i $w_1, \ldots, w_j$ samo što su v_i i w_i nekad $n \times 1$, a nekad $n \times 2$ matrice. Oni zajedno tvore matrice $V_j = [v_1, \ldots, v_j]$ i $W_j = [w_1, \ldots, w_j]$. Na kraju vrijedi ista jednakost kao i za standardni algoritam a to je

$$W_j^* A V_j = T_j.$$

Dakle ako je $\omega_{k+1} = 0$, tada dolazi do zakazivanja Lanczosovog algoritma, a ako je i B_{k+1} singularna tada trebamo promatrati slučajeve veće dimenzije od 2. Također izbor faktorizacije uvjetuje oblik metode, vidi [31]. Loša strana metode sa provjerom unaprijed je da njena implementacija značajno povećava složenost algoritma. Osim problema identificiranja situacija kod koje je došlo do skorog sloma, matrica T_k još gubi i svoj tridijagonalni oblik. Naime na mjestu gdje smo pokušali savladati slom pojavljuju se elementi i izvan dosadašnjih triju dijagonala. Kada se Lanzcosova metoda koristi za rješavanje linearnih sustava, zakazivanje metode i nije tako katastrofalno. Naime, jednostavnije i jeftinije bi bilo, restartati metodu nego implementirati metodu sa pogledom unaprijed.

2.7.2 Bikonjugirani gradijenti (BCG)

Ako pretpostavimo da dvostrani Lanczosov algoritam neće zakazati, tada vektori baza Krylovljevih prostora $\mathcal{K}_k(A, r_0)$ i $\mathcal{K}_k(A^*, \hat{r}_0)$ generirani tim algoritmom mogu biti iskotišteni za računanje aproksimacije rješenja susutava $Ax = b$, sa početnom iteracijom x_0 i rezidualom $r_0 = b - Ax_0$, na isti način kao i kod hermitskog slučaja. Dakle, uzet ćemo x_k oblika

$$x_k = x_0 + V_k y_k,$$

i onda nam još preostaje izbor y_k, vektora dimenzije k. Ako uzmemo da je $v_1 = r_0/\|r_0\|_2$ i kao vektor w_1 uzmemo proizvoljni vektor za kojeg je $\langle v_1, w_1 \rangle = 1$, time su jedinstveno definirani Krylovljevi prostori $\mathcal{K}_k(A, v_1)$ i $\mathcal{K}_k(A^*, w_1)$, kao i Lanczosovi vektori $v_1, \ldots, v_k$ i $w_1, \ldots, w_k$ za svako k. Jedan od prirodnih izbora vektora y_k u k-toj iteraciji metode bio bi takav da odgovarajući $r_k = r_0 - AV_k y_k$ bude ortogonalan na $\mathcal{K}_k(A^*, w_1)$, odnosno na vektore $w_1, \ldots, w_k$. To nas vodi do jednakosti

$$W_k^* r_k = W_k^* r_0 - W_k^* A V_k y_k = 0.$$

Zbog biortogonalnosti Lanczosovih vektora i zbog (2.109) slijedi da je $W_k^* r_0 = \beta \xi_1$, sa $\beta = \|r_0\|_2$, i $W_k^* A V_k = T_k$, stoga jednadžba za y_k izgleda ovako,

$$T_k y_k = \beta \xi_1. \tag{2.112}$$

Kao i kod hermitskog slučaja, i za dvostrani Lanczosov algoritam, uz izbor y_k koji zadovoljava (2.112), zbog (2.109) rezidual ima oblik

$$\begin{aligned} r_k &= r_0 - AV_k y_k = r_0 - V_k T_k y_k - \gamma_k \xi_k^T y_k v_{k+1} = \\ &= V_k(\beta \xi_1 - T_k y_k) - \gamma_k \xi_k^T y_k v_{k+1} = -\gamma_k \xi_k^T y_k v_{k+1}, \end{aligned}$$

odnosno on je u smjeru v_{k+1} i zaista je ortogonalan na vektore $w_1, \ldots, w_k$.

Ako pak, paralelno rješavamo i dualni sustav $A^* \hat{x} = \hat{b}$, tada bi w_1 bio definiran kao skalirani početni rezidual $\hat{r}_0 = \hat{b} - A^* \hat{x}_0$, odnosno bilo bi $w_1 = \hat{r}_0 / \langle \hat{r}_0, v_1 \rangle$. Pa, čak ako nam rješenje dualnog problema i ne treba, vektor w_1 uvijek možemo smatrati kao početni rezidual za neki dualni sustav. Po analogiji stvari, u svakom koraku bi se tražilo aproksimativno rješenje dualnog sustava $\hat{x}_k = \hat{x}_0 + W_k \hat{y}_k$, uz izbor vektora $\hat{y}_k$, takvog

da je $\hat{r}_k = \hat{r}_0 - A^* W_k \hat{y}_k$ okomit na vektore $v_1, \ldots, v_k$. Istim postupkom kao i za r_k može se pokazati da je $\hat{r}_k = -\bar{\beta}_k \xi_k^T \hat{y}_k w_{k+1}$. Dakle, ovakvim postupkom dobili smo reziduale našeg polaznog i njemu dualnog sustava koji su međusobno biortogonalni.

Dalje nastavljamo kao kod dobivanja algoritma za konjugirane gradijente preko hermitskog Lanczosovog algoritma, i zapisujemo LU dekompoziciju od T_k kao

$$T_k = L_k U_k$$

i definiramo

$$P_k = V_k U_k^{-1}.$$

Aproksimacija rješenja u k-tom koraku tada se može izraziti kao

$$\begin{aligned} x_k &= x_0 + V_k T_k^{-1}(\beta\xi_1) = x_0 + V_k U_k^{-1} L_k^{-1}(\beta\xi_1) = \\ &= x_0 + P_k L_k^{-1}(\beta\xi_1). \end{aligned}$$

Također aproksimacija x_k može se dobiti iz x_{k-1} na sličan način kao kod konjugiranih gradijenata. Sve to možemo definirati i za dualni problem, pa definirajmo matricu

$$\widehat{P}_k = W_k L_k^{-*}.$$

Vektori $\hat{p}_k$ koji čine stupce od $\widehat{P}_k$ i vektori p_k koji čine stupce od P_k su A-konjugirani jer

$$\widehat{P}_k^* A P_k = L_k^{-1} W_k^* A V_k U_k^{-1} = L_k^{-1} T_k U_k^{-1} = I.$$

Dakle, dobili smo niz reziduala r_k i $\hat{r}_k$ koji su biortogonalni, i niz vektora smjera p_k i $\hat{p}_k$ koji su A-ortogonalni, i kao kod konjugiranih gradijenata možemo dobiti rekurzivne relacije koje ih definiraju. Sa ovim napomenama, lako možemo dobiti algoritam koji liči na konjugirane gradijente iz dvostranog Lanczosevog algoritma i kojeg nazivamo algoritmom *bikonjugiranih gradijenata*.

Algoritam 2.7.3. BIKONJUGIRANI GRADIJENTI (BCG)

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$.

Izaberi $\hat{r}_0$ *takav da je* $\langle r_0, \hat{r}_0 \rangle \neq 0$.

$p_0 = r_0$,

$\hat{p}_0 = \hat{r}_0$.

Za $k = 1, 2, \ldots$

izračunaj Ap_{k-1},
izračunaj $A^*\hat{p}_{k-1}$,
$\alpha_{k-1} = \dfrac{\langle r_{k-1}, \hat{r}_{k-1} \rangle}{\langle Ap_{k-1}, \hat{p}_{k-1} \rangle}$,
$x_k = x_{k-1} + \alpha_{k-1} p_{k-1}$,
$r_k = r_{k-1} - \alpha_{k-1} Ap_{k-1}$,
$\hat{r}_k = \hat{r}_{k-1} - \bar{\alpha}_{k-1} A^* \hat{p}_{k-1}$,
$\beta_{k-1} = \dfrac{\langle r_k, \hat{r}_k \rangle}{\langle r_{k-1}, \hat{r}_{k-1} \rangle}$,
$p_k = r_k + \beta_{k-1} p_{k-1}$,
$\hat{p}_k = \hat{r}_k + \bar{\beta}_{k-1} \hat{p}_{k-1}$.

Kada je A hermitska i $\hat{r}_0 = r_0$, ovaj se algoritam reducira na konjugirane gradijente. Ako je u (2.112) T_k singularna, tada ta jednakost ne mora imati rješenje. U tom slučaju ne postoji aproksimacija $x_k = x_0 + V_k y_k$ za koje je $W_k^* r_k = 0$. Tada će algoritam zakazati, ali to je drugačiji tip sloma od onoga koji se može dogoditi u dvostranom Lanczosovom algoritmu. Standardni algoritam ne može izaći na kraj sa takvim slomom, iako matrica T_j za neki $j > k$ može biti regularna.

Općenito gledajući, od metode koja rješava linearni sustav sa nehermitskom matricom zahtijevamo dvije stvari

(i) da zadovoljava neko svojstvo minimalnosti na Krylovljevim potprostorima koje generira matrica A,

(ii) da se može izračunati sa što manje operacija i da zahtijeva malo memorije po iteraciji.

Na žalost, metode obično zadovoljavaju samo jedno od ta dva svojstva. Na primjer, GMRES metoda zadovoljava (i) ali ne i (ii), budući da u svakoj iteraciji potreba za memorijom sve više raste. S druge strane BCG metoda zadovoljava (ii), jer zbog trokoračne rekurzije broj operacija i količina memorije su konstantni u svakoj iteraciji. Međutim ona ne zadovoljava (i), već određena svojstva biortogonalnosti, zbog čega algoritam često ispoljava nepravilno ponašanje u konvergenciji sa velikim oscilacijama norme reziduala.

Uz to još, ova metoda može zakazati na dva načina, prvi je slom dvostranog Lanczosovog algoritma, kada bi došlo do dijeljenja sa nulom zbog koeficijenta $\beta_k = 0$ u Algoritmu 2.7.3, a drugi je slučaj singularne matrice T_k, što je ekvivalentno dijeljenju sa nulom kod računanja koeficijenta α_k u istom algoritmu ($\widetilde{P}_k A P_k = L_k^{-1} T_k U_k^{-1} = I$ u regularnom slučaju).

2.7.3 Metoda kvazi–minimalnog reziduala (QMR)

Metoda kvazi–minimalnog reziduala (QMR) je metoda slična BCG metodi, bazirana na dvostranom Lanczosovom algoritmu, koja se primijenjuje za rješavanje nehermitskih linearnih sustava, ali koja je u stanju prevladati probleme BCG-a. Ona prije svega neće zakazati kada je T_k singularna, iako će taj slučaj izazvati stagnaciju kod konvergencije što ćemo kasnije vidjeti, a problem ozbiljnog sloma dvostranog Lanczosovog algoritma može se riješiti metodom s provjerom unaprijed. QMR metoda može se također implementirati pomoću kratkih rekurzivnih izraza, i tako ona zadovoljava uvjet (ii) iz prošlog odjeljka. S druge strane, ova metoda se približila i uvjetu (i), jer kvazi–minimalno svojstvo osigurava glatku konvergenciju, za razliku od BCG. Štoviše iteracije BCG metode mogu se lako dobiti iz QMR iteracije, pa čak kad BCG iteracija ne može biti definirana zbog singularnosti matrice T_k.

U QMR metodi ponovo se uzima da je aproksimacija x_k oblika

$$x_k = x_0 + V_k y_k,$$

samo što se y_k bira tako da minimizira kvazi–rezidual, vrijednost usko povezanu sa euklidskom normom reziduala. Budući da je $r_k = r_0 - A V_k y_k$, kao i kod BCG metode dobiva se da je

$$r_k = V_{k+1}(\beta\xi_1 - T_{k+1,k} y_k), \tag{2.113}$$

tako da norma reziduala r_k zadovoljava

$$\|r_k\|_2 \leq \|V_{k+1}\|_2 \|\beta\xi_1 - T_{k+1,k} y_k\|_2. \tag{2.114}$$

Budući da stupci od V_{k+1} nisu ortogonalni, rješavanje problema najmanjih kvadrata $\min_{y\in\mathbb{C}^k} \|r_0 - V_{k+1} T_{k+1,k} y\|_2$ zahtijevalo bi previše operacija. Stoga ćemo umjesto toga minimizirati drugi faktor u (2.114). Svi stupci matrice V_{k+1} imaju normu jedan, zato prvi faktor u (2.114) zadovoljava

$$\|V_{k+1}\|_2 \leq \|V_{k+1}\|_F \leq \sqrt{k+1}.$$

Dakle, y_k u QMR metodi rješava problem najmanjih kvadrata

$$\min_{y\in\mathbb{C}^k} \|\beta\xi_1 - T_{k+1,k} y\|_2, \tag{2.115}$$

koji uvijek ima rješenje, čak i kad je matrica T_k singularna. Naime, matrica $T_{k+1,k}$ je $(k+1)\times k$ tridijagonalna matrica koja na donjoj sporednoj dijagonali, ukoliko se dvostrani Lanczosov algoritam nije zbog nekog razloga zaustavio, ima elemente $\gamma_j > 0$ za $j = 1, \ldots, k$, pa je ona punog ranga. Što više rješnje je jedinstveno. Znači, QMR iteracije su definirane ako dvostrani Lanczososov algoritam ne zakaže, a kod njegovog zakazivanja upotrebljava se algoritam sa provjerom unaprijed, čime matrica $T_{k+1,k}$ postaje blok-tridijagonalna.

Norma QMR reziduala može se povezati sa normom optimalnog GMRES reziduala na sljedeći način.

Teorem 2.7.4 ([12]). *Ako sa r_k^G označimo GMRES rezidual u k-tom koraku, a sa r_k^Q QMR rezidual u k-tom koraku, tada*

$$\|r_k^Q\|_2 \leq \kappa(V_{k+1})\|r_k^G\|_2,$$

gdje je V_{k+1} matrica vektora baze prostora $\mathcal{K}_{k+1}(A, r_0)$ koji su konstruirani dvostranim Lanczosovim algoritmom, $\kappa(X) = \|X\|_2\|X^+\|_2 = \sigma_{max}(X)/\sigma_{min}(X)$ označava broj uvjetovanosti, a $\sigma_{max}(X)$ i $\sigma_{min}(X)$ najveću i najmanju singularnu vrijednost od X

Dokaz: GMRES rezidual je također oblika (2.113), ali y_k^G je izabran tako da minimizira euklidsku normu GMRES reziduala. Neka je V_{k+1}^+ generalizirani inverz matrice V_{k+1}, te zbog toga što je ona punog ranga vrijedi $V_{k+1}^+ = (V_{k+1}^* V_{k+1})^{-1} V_{k+1}^*$, i $V_{k+1}^+ V_{k+1} = I$. Zato slijedi

$$\begin{aligned}\|\beta\xi_1 - T_{k+1,k}y_k^Q\|_2 &= \min_{y\in\mathbb{C}^k}\|\beta\xi_1 - T_{k+1,k}y\|_2 = \min_{y\in\mathbb{C}^k}\|V_{k+1}^+V_{k+1}(\beta\xi_1 - T_{k+1,k}y)\|_2 \leq \\ &\leq \|V_{k+1}^+\|_2 \min_{y\in\mathbb{C}^k}\|V_{k+1}(\beta\xi_1 - T_{k+1,k}y)\|_2 = \frac{1}{\sigma_{min}(V_{k+1})}\|r_k^G\|_2,\end{aligned}$$

a zajedno sa (2.114) daje

$$\|r_k^Q\|_2 \leq \|V_{k+1}\|_2\|\beta\xi_1 - T_{k+1,k}y_k\|_2 \leq \frac{\sigma_{max}(V_{k+1})}{\sigma_{min}(V_{k+1})}\|r_k^G\|_2.$$

□

Na žalost, uvjetovanost matrice V_{k+1} ne može se apriori ograditi, čak ta matrica može biti vrlo loše uvjetovana.

Implementacija QMR algoritma je vrlo slična onoj od MINRES-a. Problem najmanjih kvadrata (2.115) rješava se QR faktorizacijom $(k+1)\times k$ matrice $T_{k+1,k}$ na produkt $(k+1)\times(k+1)$ unitarne matrice $F^{(k)*}$ i $(k+1)\times k$ gornje trokutaste matrice $R^{(k)}$. To se ponovo postiže uz pomoć k Givensovih rotacija $F_1,\ldots,F_k$, gdje F_k rotira jedinične vektore ξ_k i ξ_{k+1} za kut θ_k. Budući da je $T_{k+1,k}$ tridijagonalna, $R^{(k)}$ ima oblik

$$R = \begin{bmatrix} \rho_1 & \sigma_1 & \tau_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \tau_{k-2} \\ & & & \ddots & \sigma_{k-1} \\ & & & & \rho_k \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix}.$$

QR faktorizacija matrice $T_{k+1,k}$ se lagano dobije iz QR dekompozicije od $T_{k,k-1}$. Da bi dobili $R^{(k)}$ prvo treba pomnožiti zadnji stupac od $T_{k+1,k}$ sa rotacijama iz koraka $k-2$ i $k-1$, jer ostale rotacije ne utječu na njega. Tada dobivamo matricu oblika

$$F_{k-1}F_{k-2}F_{k-3}\cdots F_1T_{k+1,k} = \begin{bmatrix} x & x & x & & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & x \\ & & & \ddots & x \\ & & & & d \\ 0 & \cdots & \cdots & \cdots & h \end{bmatrix},$$

gdje x-ovi označavaju elemente različite od nule, a $(k+1,k)$-ti element je zapravo γ_k, budući da na njega nisu utjecale prethodne rotacije. Rotacija F_k se bira tako da poništi taj element tako što postavljamo da je $c_k = |d|/\sqrt{|d|^2+|h|^2}$ i $\bar{s}_k = c_k h/d$ ako je $d \neq 0$, te $c_k = 0$ i $s_k = 1$ ako je $d = 0$. Taj isti niz od k rotacija primjenjuje se i na vektor $\beta\xi_1$ kako bismo dobili $g^{(k)} = \beta F_k \cdots F_1 \xi_1$. $g^{(k)}$ se od $g^{(k-1)}$ razlikuje samo po k-toj i $(k+1)$-toj komponenti. Ako sa R_k označimo gornji $k \times k$ blok matrice $R^{(k)}$ i sa $g^{(k)}_{k\times 1}$ prvih k komponenti od $g^{(k)}$, tada rješenje problema najmanjih kvadrata je rješenje trokutastog linearnog sustava

$$R_k y_k = g^{(k)}_{k\times 1}.$$

Za dobivanje aproksimacije x_k, definirat ćemo pomoćne vektore

$$P_k = [p_0\ p_1\ \ldots\ p_{k-1}] = V_k R_k^{-1}.$$

Tada je

$$x_k + V_k y_k = x_0 + P_k g^{(k)}_{k\times 1}$$

a budući da se $g^{(k)}$ i $g^{(k-1)}$ poklapaju u prvih $k-1$ komponenti, vrijedi

$$x_{k-1} = x_0 + P_{k-1} g^{(k)}_{(k-1)\times 1}.$$

Sada možemo napisati

$$x_k = x_{k-1} + g^{(k)}_k p_{k-1}, \tag{2.116}$$

gdje je $g^{(k)}_k$ k-ta komponenta od $g^{(k)}$. Na kraju, iz jednadžbe $P_k R_k = V_k$, možemo dobiti jednostavnu rekurzijua za pomoćne vektore p_k

$$p_{k-1} = \frac{1}{\rho_k}(v_k - \sigma_{k-1}p_{k-2} - \tau_{k-2}p_{k-3}). \tag{2.117}$$

Napokon dobivamo implementaciju QMR algoritma.

Algoritam 2.7.5. METODA KVAZI–MINIMALNOG REZIDUALA (QMR)

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

$\beta = \|r_0\|_2$,

$v_1 = \dfrac{r_0}{\beta}$,

Dan je $\hat{r}_0$,

$w_1 = \dfrac{\hat{r}_0}{\|\hat{r}_0\|_2}$,

$l = (1, 0, \ldots, 0)^T$.

Za $k = 1, 2, \ldots$

Izračunaj v_{k+1}, w_{k+1}, $\alpha_k = T(k,k)$, $\beta_k = T(k,k+1)$, *i* $\gamma_k = T(k+1,k)$, *koristeći dvostrani Lanczosov algoritam.*

Primijeni F_{k-2} *i* F_{k-1} *na zadnji stupac od* T, *odnosno:*

ako je $k > 2$ *tada* $\begin{bmatrix} T(k-2,k) \\ T(k-1,k) \end{bmatrix} := \begin{bmatrix} c_{k-2} & s_{k-2} \\ -\bar{s}_{k-2} & c_{k-2} \end{bmatrix} \begin{bmatrix} 0 \\ T(k-1,k) \end{bmatrix}$,

ako je $k > 1$ *tada* $\begin{bmatrix} T(k-1,k) \\ T(k,k) \end{bmatrix} := \begin{bmatrix} c_{k-1} & s_{k-1} \\ -\bar{s}_{k-1} & c_{k-1} \end{bmatrix} \begin{bmatrix} T(k-1,k) \\ T(k,k) \end{bmatrix}$.

Izračunaj k*-tu Givensovu rotaciju* F_k *kako bi se poništio* $(k+1,k)$ *element od* T*:*

$c_k = \dfrac{|T(k,k)|}{\sqrt{|T(k,k)|^2 + |T(k+1,k)|^2}}$,

ako je $c_k \neq 0$ *tada* $s_k = c_k \dfrac{\overline{T(k+1,k)}}{\overline{T(k,k)}}$, *ako je* $c_k = 0$ *tada* $s_k = 1$.

Primijeni k*-tu rotaciju na* l *i na zadnji stupac od* T*:*

$\begin{bmatrix} l(k) \\ l(k+1) \end{bmatrix} := \begin{bmatrix} c_k & s_k \\ -\bar{s}_k & c_k \end{bmatrix} \begin{bmatrix} l(k) \\ 0 \end{bmatrix}$,

$T(k,k) := c_k T(k,k) + s_k T(k+1,k)$,

$T(k+1,k) = 0 \qquad (*)$.

Izračunaj $p_{k-1} = (1/T(k,k))[v_k - T(k-1,k)p_{k-2} - T(k-2,k)p_{k-3}]$, *gdje su* p_{-1}, p_{-2}, *jednaki nuli.*

$x_k = x_{k-1} + \beta l(k) p_{k-1}$.

(*) $T(k+1,k) = 0$ treba zapravo izvesti u sljedećem, $(k+1)$-om koraku, jer je originalna vrijednost $T(k+1,k)$ potrebna za izvođenje $(k+1) - og$ koraka dvostranog Lanczosovog algoritma.

2.7.4 Konvergencija metoda BCG i QMR

Iz prethodnih razmatranja možemo zaključiti da su rješenja linearnog sustava (2.112) i problema najmanjih kvadrata (2.115) u uskoj vezi. Zato možemo očekivati istu takvu vezu između normi reziduala u BCG i QMR algoritmima.

Obje metode završit će za $m \leq n$ iteracija, pri čemu je n dimenzija sustava. Kod BCG metode rezidual se bira tako da bude ortogonalan na određeni potprostor, kojemu dimenzija raste u svakoj iteraciji. U najgorem slučaju on mora biti okomit na cijeli prostor, a to vrijedi samo za nul-vektor. Kod QMR metode, situacija je slična kao kod GMRES-a. Rezidual se bira tako da rješavamo problem najmanjih kvadrata na sve većem i većem potprostoru. U jednom trenutku rješenje će generirati vektor $T_{m+1,m}y_m$ koji će se nalaziti u istom potprostoru kao i $\beta\xi_1$, a budući da kvazi-rezidual $T_{m+1,m}y_m - \beta\xi_1$ ima najmanju normu, taj vektor bit će upravo jednak vektoru $\beta\xi_1$, pa će kvazi-rezidual, a s njime i rezidual, biti jednak nuli.

Najprije promotrimo jednu lemu. Neka $\mathcal{H}_k$, $k = 1, 2, \ldots$ predstavlja klasu gornje Hessenbergovih $k \times k$ matrica H_k, pri čemu sa H_{k-1} označavamo $(k-1)\times(k-1)$ glavnu podmatricu od H_k. Za svaki k definirajmo $(k+1) \times k$ matricu $H_{k+1,k}$ sa

$$H_{k+1,k} = \begin{bmatrix} H_k \\ h_{k+1,k}\xi_k^T \end{bmatrix}. \tag{2.118}$$

Matrica $H_{k+1,k}$ može se faktorizirati u oblik $F^{(k)*}R^{(k)}$, gdje je $F^{(k)}$ $(k+1) \times (k+1)$ unitarna matrica, a $R^{(k)}$ $(k+1)\times k$ gornje trokutasta matrica. Ta se faktorizacija može ostvariti primjenom Givensovih rotacija na način kako je opisano u GMRES algoritmu:

$$(F_k \cdots F_1)H_{k+1,k} = R^{(k)}, \qquad \text{gdje je } F_k = \begin{bmatrix} I_{i-1} & & & \\ & c_k & s_k & \\ & -\bar{s}_k & c_k & \\ & & & I_{k-i} \end{bmatrix}. \tag{2.119}$$

Bitno je samo napomenuti to da prvih $k-1$ rotacija F_i, $i = 1, \ldots, k-1$ isto tako sudjeluje u faktorizaciji matrice $H_{k,k-1}$.

Neka je $\beta > 0$ dan, i pretpostavimo da je H_k regularna. Neka je sa $\tilde{y}_k$ označeno rješenje linearnog sustava $H_k y = \beta\xi_1$, i neka je sa y_k označeno rješenje problema najmanjih kvadrata $\min_{y\in\mathbb{C}^k} \|H_{k+1,k}y - \beta\xi_1\|_2$. Na kraju, neka su

$$\tilde{\nu}_k = H_{k+1,k}\tilde{y}_k - \beta\xi_1, \qquad \nu_k = H_{k+1,k}y_k - \beta\xi_1.$$

Lema 2.7.6 ([5]). *Koristeći gornju notaciju, norme od ν_k i $\tilde{\nu}_k$ su povezane sa sinusima i kosinusima kuteva Givensovih rotacija na sljedeći način*

$$\|\nu_k\|_2 = \beta|s_1 s_2 \cdots s_k| \qquad i \qquad \|\tilde{\nu}_k\|_2 = \beta\frac{1}{|c_k|}|s_1 s_2 \cdots s_k|. \tag{2.120}$$

Slijedi da je

$$\|\tilde{\nu}_k\|_2 = \frac{\|\nu_k\|_2}{\sqrt{1 - \left(\frac{\|\nu_k\|_2}{\|\nu_{k-1}\|_2}\right)^2}}, \tag{2.121}$$

ili ekvivalentno

$$\left(\frac{\|\nu_k\|_2}{\|\tilde{\nu}_k\|_2}\right)^2 + \left(\frac{\|\nu_k\|_2}{\|\nu_{k-1}\|_2}\right)^2 = 1.$$

Dokaz: Neka je $F^{(k)} = F_k \cdots F_1$. Problem najmanjih kvadrata može se napisati kao

$$\min_{y\in\mathbb{C}^k} \|H_{k+1,k}y - \beta\xi_1\|_2 = \min_{y\in\mathbb{C}^k} \|F^{(k)}(H_{k+1,k}y - \beta\xi_1)\|_2 = \min_{y\in\mathbb{C}^k} \|R^{(k)}y - \beta F^{(k)}\xi_1\|_2.$$

Rješenje y_k se dobiva rješavanjem trokutastog linearnog sustava sa matricom R_k koja je jednaka gornjem $k \times k$ bloku od $R^{(k)}$ i sa desnom stranom koja je jednaka prvim k komponentama vektora $\beta F^{(k)}\xi_1$. Razlika $R^{(k)}y_k - \beta F^{(k)}\xi_1$ je zato jednaka nuli, osim u zadnjoj komponenti, koja je jednaka zadnjoj komponenti od $-\beta F^{(k)}\xi_1$, koja je opet, kao i kod GMRES-a, jednaka $(-1)^{k+1}\beta\bar{s}_1 \cdots \bar{s}_k$. Dakle

$$\|\nu_k\|_2 = \min_{y\in\mathbb{C}^k} \|H_{k+1,k}y - \beta\xi_1\|_2 = \beta|s_1 \cdots s_k|,$$

pa smo time dokazali prvu jednakost u (2.120).

Za rješenje linearnog sustava $\tilde{y}_k = H_k^{-1}\beta\xi_1$, imamo

$$\tilde{\nu}_k = H_{k+1,k}H_k^{-1}\beta\xi_1 - \beta\xi_1,$$

ali zbog definicije matrice $H_{k+1,k}$ imamo da je

$$H_{k+1,k}H_k^{-1} = \begin{bmatrix} H_kH_k^{-1} \\ h_{k+1,k}\xi_k^T H_k^{-1} \end{bmatrix} = \begin{bmatrix} I_k \\ h_{k+1,k}\xi_k^T H_k^{-1} \end{bmatrix}.$$

Slijedi

$$\tilde{\nu}_k = \begin{bmatrix} 0_{k\times k} \\ \beta h_{k+1,k}(H_k^{-1})_{k,1} \end{bmatrix}$$

gdje je $(H_k^{-1})_{k,1} = \xi_k^T H_k^{-1}\xi_1$ $(k,1)$-ta komponenta matrice H_k^{-1}. Matrica H_k može se faktorizirati kao $F^{(k-1)*}\tilde{R}_k$, gdje je $F^{(k-1)} = \tilde{F}_{k-1} \cdots \tilde{F}_1$, a $\tilde{F}_k$ je glavna $k\times k$ podmatrica od F_k. Matrica $H_{k+1,k}$, nakom primjene prvih $k-1$ rotacija, ima oblik

$$F_{k-1} \cdots F_1 H_{k+1,k} = \begin{bmatrix} x & x & x & & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & x \\ & & & \ddots & x \\ & & & & r \\ 0 & \cdots & \cdots & \cdots & h \end{bmatrix},$$

gdje je r (k,k)-ti element matrice $\tilde{R}^{(k)}$, a $h = h_{k+1,k}$. k-ta rotacija je izabrana tako da poništi netrivijalni element u zadnjem retku:

$$c_k = \frac{|r|}{\sqrt{|r|^2 + |h|^2}}, \qquad \bar{s}_k = \frac{\operatorname{sign}(r)h}{\sqrt{|r|^2 + |h|^2}}.$$

Budući da je po pretpostavci H_k regularna matrica tada je $r \neq 0$ pa shodno tome i $c_k \neq 0$. Sada imamo $H_k^{-1} = \tilde{R}_k^{-1}\tilde{F}^{(k)}$, a $(k,1)$-ti element te matrice je $1/r$ puta $(k,1)$-ti element od $\tilde{F}^{(k)}$. To je ekvivalentno tome da gledamo k-tu komponentu vektora $\tilde{F}^{(k)}\xi_1 = \tilde{F}_{k-1} \cdots \tilde{F}_1\xi_1$, što je ekvivalentno dobivanju reziduala prethodnog problema najmanjih kvadrata (kao kod GMRES), i jednako je $(-1)^{k-1}\bar{s}_1 \cdots \bar{s}_{k-1}$. Slijedi da je

jedina netrivijalna komponenta od $\tilde{\nu}_k$ jednaka $(-1)^{k-1}\beta(h_{k+1,k}/r)\bar{s}_1\cdots\bar{s}_{k-1}$. Napokon korištenjem činjenice da je $|s_k/c_k| = |h/r| = |h_{k+1,k}/r|$ dobivamo

$$\|\tilde{\nu}_k\|_2 = \beta\frac{|s_k|}{|c_k|}|s_1\cdots s_{k-1}|.$$

Time smo dokazali i drugu jednakost u (2.120).

Iz rezultata (2.120) je jasno da je

$$\frac{\|\nu_k\|_2}{\|\nu_{k-1}\|_2} = |s_k|, \qquad \frac{\|\tilde{\nu}_k\|_2}{\|\nu_k\|_2} = \frac{1}{|c_k|}.$$

Rezultat (2.121) slijedi iz jednakosti $|c_k| = \sqrt{1-|s_k|^2}$. □

Označimo sada reziduale i iteracije u BCG metodi sa gornjim indeksom B, a u QMR metodi sa gornjim indeksom Q. Najprije ćemo promatrati normu reziduala BCG metode, odnosno njenu konvergenciju.

Teorem 2.7.7 ([9]). *Ako u k-tom koraku BCG metode aproksimacija x_k^B postoji tada je*

$$x_k^B = x_k^Q + \frac{g_k^{(k)}|s_k|^2}{c_k^2}p_{k-1},$$

i

$$\|r_k^B\|_2 = \frac{1}{|c_k|}|s_1\cdots s_k|\|r_0\|_2,$$

gdje je $g_k^{(k)}$ k-ta komponenta vektora $g^{(k)} = \beta F^{(k)}\xi_1$ a p_{k-1} je k-ti stupac matrice $P_k = V_k R_k^{-1}$.

Dokaz: Ako stavimo da je $H_k = T_k$ tada Lemu 2.7.6 možemo primijeniti i na BCG metodu. Iz njenog dokaza vidljiva je ekvivalentnost između postajanja aproksimacije x_k^B, regularnosti matrice T_k i nejednakosti $c_k \neq 0$. Nadalje koristimo iste oznake kao u dokazu prethodne leme. Iz (2.118), (2.119) i definicije $F^{(k-1)}$ imamo

$$T_k = F^{(k-1)*}\begin{bmatrix} I_{k-1} & 0 \\ 0 & c_k \end{bmatrix}R_k.$$

Odavde, zbog činjenice da je $y_k^B = \beta T_k^{-1}\xi_1$, dobivamo

$$y_k^B = \beta R_k^{-1}\begin{bmatrix} I_{k-1} & 0 \\ 0 & \frac{1}{c_k} \end{bmatrix}F^{(k-1)}\xi_1.$$

Kako je prema definiciji $\beta = \|r_0\|_2$, i $y_k^Q = R_k^{-1}g_{k\times 1}^{(k)}$, izraz za y_k^B dalje možemo raspisati

$$\begin{aligned} y_k^B &= y_k^Q + R_k^{-1}\left(\begin{bmatrix} I_{k-1} & 0 \\ 0 & \frac{1}{c_k} \end{bmatrix}g^{(k-1)} - g_{k\times 1}^{(k)}\right) = \\ &= y_k^Q + R_k^{-1}\begin{bmatrix} 0_{(k-1)\times 1} \\ (-1)^{k-1}\beta\bar{s}_1\cdots\bar{s}_{k-1}\left(\frac{1}{c_k} - c_k\right) \end{bmatrix} = \\ &= y_k^Q + (-1)^{k-1}\beta\bar{s}_1\cdots\bar{s}_{k-1}c_k\frac{|s_k|^2}{c_k^2}R_k^{-1}\begin{bmatrix} 0_{(k-1)\times 1} \\ 1 \end{bmatrix} \\ &= y_k^Q + g_k^{(k)}\frac{|s_k|^k}{c_k^2}R_k^{-1}\xi_k. \end{aligned}$$

Nadalje, iz izraza za y_k^B možemo dobiti izraz za x_k^B

$$\begin{aligned} x_k^B &= x_0 + V_k y_k^B = x_0 + V_k y_k^Q + g_k^{(k)} \frac{|s_k|^2}{c_k^2} V_k R_k^{-1} \xi_k = \\ &= x_k^Q + g_k^{(k)} \frac{|s_k|^2}{c_k^c} p_{k-1}. \end{aligned}$$

Izraz za normu reziduala $\|r_k^B\|_2$ direktno slijedi iz Leme 2.7.6. □

Teorem 2.7.7 demonstrira kako se egzistirajuća BCG iteracija može dobiti iz QMR algoritma, pa čak iako u prethodnom koraku BCG iteracija nije postajala zbog singularnosti tridijagonalne matrice T_{k-1}. U tom slučaju BCG algoritam bi stao, međutim za neki $j \geq k$ može postajati T_j koja je regularna, pa se na ovaj način proces može nastaviti. Nadalje, $\|r_k^B\|_2$ može se, bez ikakvih naknadnih troškova, izračunati iz veličina koje su generirane QMR algoritmom. Osim toga, iz izraza za normu reziduala u Teoremu 2.7.7 vidimo razlog njenog osciliranja. Naime, kako je niz $|s_1 \cdots s_k|$ nerastući, globalno možemo očekivati da će se i norma reziduala BCG metode tako ponašati. Međutim, c_k koji se nalazi u nazivniku, u slučaju da je blizu nule, izazivat će skokove i šiljke na krivulji koja ocrtava normu reziduala u odnosu na broj iteracija.

Vratimo se sada QMR metodi.

Teorem 2.7.8 ([9]). *Norma reziduala aproksimacije x_k^Q QMR metode zadovoljava relaciju*

$$\|r_k^Q\|_2 \leq \|V_{k+1}\|_2 |s_1 \cdots s_k| \|r_0\|_2.$$

Dokaz: Definirajmo izraz *kvazi-reziduala* sa

$$z_k^Q = \beta \xi_1 - T_{k+1,k} y_k^Q.$$

Prema (2.113) je

$$r_k^Q = V_{k+1}(\beta \xi_1 - T_{k+1,k} y_k^Q) = V_{k+1} z_k^Q,$$

pri čemu y_k^Q minimizira normu izraza z_k^Q. Prema Lemi 2.7.6 vrijedi

$$\|z_k^Q\|_2 = \beta |s_1 \cdots s_k|,$$

pa za normu reziduala prema (2.114) vrijedi

$$r_k^Q \leq \|V_{k+1}\|_2 \|z_k^Q\|_2,$$

odakle slijedi tvrdnja teorema. □

Kako je norma svakog stupca od V_{k+1} jednaka 1, onda vrijedi $\|V_{k+1}\|_2 \leq \sqrt{k+1}$ pa se tvrdnja Teorema 2.7.8 može napisati i kao

$$\|r_k^Q\|_2 \leq \sqrt{k+1} |s_1 \cdots s_k| \|r_0\|_2.$$

Pogledajmo još jedan slučaj konvergencije QMR metode, koji je vrlo sličan analognom slučaju kod GMRES metode.

Teorem 2.7.9 ([9]). *Neka je k iteracija u kojoj se QMR regularno zaustavio (u slučaju ozbiljnog sloma koristi se provjera unaprijed). Pretpostavimo da je $k \times k$ tridijagonalna matrica T_k, generirana u k-tom koraku dvostranog Lanczosovog algoritma, dijagonalizabilna, odnosno da vrijedi $T_k = X\Lambda X^{-1}$ za dijagonalnu matricu Λ i regularnu matricu X. Tada za $j = 1, 2, \ldots, k-1$, reziduali QMR algoritma zadovoljavaju*

$$\|r_j^Q\|_2 \leq \|r_0\|_2 \kappa(X)\sqrt{j+1}\epsilon_j,$$

gdje je

$$\epsilon_j = \min_{p_j \in \mathbb{P}_j, p_j(0)=1} \max_{\lambda \in \sigma(A)} |p_j(\lambda)|.$$

Nadalje, ako se dvostrani Lanczosov algoritam zaustavio sa $\gamma_k = \|\tilde{v}_{k+1}\|_2 = 0$, tada je $x_k^Q = A^{-1}b$ egzaktno rješenje sustava $Ax = b$.

Dokaz: Imamo $r_j^Q = \beta V_{j+1}(\xi_1 - T_{j+1,j}u_j^Q)$, gdje je $\beta u_j^Q = y_j^Q$. Odavde slijedi

$$\|r_j^Q\|_2 \leq \|r_0\|_2\sqrt{j+1}\theta_j,$$

gdje je θ_j dan sa

$$\theta_j = \min_{u \in \mathbb{C}^j} \|\xi_1 - T_{j+1,j}u\|_2, \qquad \xi_1 = [1\ 0\ \ldots\ 0] \in \mathbb{R}^{j+1}.$$

Zbog toga, potrebno je još dokazati da je

$$\theta_j \leq \kappa(X)\epsilon_j.$$

Neka je $j \in \{1, 2, \ldots, k-1\}$ proizvoljan, ali fiksiran. Kako vrijedi

$$T_k = \begin{bmatrix} T_{j+1,j} & * \\ 0 & * \end{bmatrix}$$

imamo

$$T_k \begin{bmatrix} u \\ 0 \end{bmatrix} = \begin{bmatrix} T_{j+1,j}u \\ 0 \end{bmatrix} \quad \text{za sve } u \in \mathbb{C}^j,$$

pri čemu postoji $j+1$ netrivijalnih komponenata vektora na desnoj strani jednakosti. Iz toga slijedi da za ξ_1 vrijedi

$$T_k\xi_1 = \begin{bmatrix} f_2 \\ 0_{(k-2)\times 1} \end{bmatrix}, \quad T_k^2\xi_1 = \begin{bmatrix} f_3 \\ 0_{(k-3)\times 1} \end{bmatrix} \ldots \quad T_k^j\xi_1 = \begin{bmatrix} f_{j+1} \\ 0_{(k-j-1)\times 1} \end{bmatrix},$$

gdje su f_i vektori dimenzije i. Na kraju možemo zaključiti da je

$$\left\{ \begin{bmatrix} t \\ 0 \end{bmatrix} :\ t \in \mathbb{C}^{j+1} \right\} = \{p_j(T_k)\xi_1 :\ p_j \in \mathbb{P}_j\}.$$

Sada možemo napisati da je

$$\theta_j = \min_{u \in \mathbb{C}^j} \left\| \xi_1 - T_k \begin{bmatrix} u \\ 0 \end{bmatrix} \right\|_2 = \min_{p_j \in \mathbb{P}_j, p_j(0)=1} \|p_j(T_k)\xi_1\|_2.$$

Budući da vrijedi

$$\|p_j(T_k)\xi_1\|_2 \leq \|p_j(T_k)\|_2 \leq \|X\|_2\|p_j(\Lambda)\|_2\|X^{-1}\|_2,$$

za θ_j možemo dobiti relaciju

$$\theta_j \leq \kappa(X) \min_{p_j \in \mathbb{P}_j, p_j(0)=1} \max_{\lambda \in \sigma(T_k)} |p_j(\lambda)|.$$

Po pretpostavci u k-tom koraku je došlo do regularnog zaustvljanja to znači da je ili $\|\tilde{v}_{k+1}\|_2 = 0$, pa je prema (2.109) V_k A-invarijantni potprostor ili je $\|\tilde{w}_{k+1}\|_2 = 0$ pa je prema (2.110) W_k A^*-invarijantni potprostor. Pogledajmo prvi slučaj. Neka je

$$AV_k = V_k T_k, \qquad T_k \in \mathbb{C}^{k\times k},$$

tada za svako $\lambda \in \sigma(T_k)$, $T_k y = \lambda y$ za neki $y \neq 0$, vrijedi

$$A(V_k y) = V_k T_k y = V_k(\lambda y) = \lambda(V_k y),$$

pa je $\lambda \in \sigma(A)$, čime smo dobili da je $\sigma(T_k) \subseteq \sigma(A)$. Drugi slučaj se dokazuje analogno. Time smo dokazali prvu tvrdnju teorema.

Druga tvrdnja se dokazuje analogno kao i kod GMRES metode. Kad bi htjeli izračunati k-ti korak do kraja, za $\gamma_k = 0$, tada bi $F^{(k-1)}T_{k+1,k}$ već bila gornje trokutasta pa bi $F_k = I$, odnosno $s_k = 0$. Iz Teorema 2.7.8 slijedi da bi $r_k = 0$. □

Na kraju pogledajmo odnos između konvergencija BCG i QMR metoda.

Teorem 2.7.10 ([5]). *Pretpostavimo da su do k-tog koraka Lanczosovi vektori definirani, i da je tridijagonalna matrica T_k generirana dvostranim Lanczosovim algoritmom u k-tom koraku regularna. Tada su BCG rezidual r_k^B i QMR kvazi-rezidual z_k^Q povezani relacijom*

$$\|r_k^B\|_2 = \frac{\|z_k\|_2}{\sqrt{1 - \left(\frac{\|z_k^Q\|_2}{\|z_{k-1}^Q\|_2}\right)^2}}. \tag{2.122}$$

Dokaz: Iz (2.109), činjenice da je $x_k^B = x_0 + V_k y_k^B$, i (2.112), BCG rezidual ima oblik

$$\begin{aligned} r_k^B &= r_0 - AV_k y_k^B = \\ &= r_0 - V_{k+1} T_{k+1,k} y_k^B = \\ &= V_{k+1}(\beta\xi_1 - \beta T_{k+1,k} T_k^{-1} \xi_1). \end{aligned}$$

Vektor, zadan izrazom u zagradama, ima samo jednu netrivijalnu komponentu, onu $(k+1)-u$, a budući da je $\|v_{k+1}\|_2 = 1$, imamo

$$\|r_k^B\|_2 = \|\beta\xi_1 - T_{k+1,k} T_k^{-1} \beta\xi_1\|_2.$$

Tvrdnja teorema sada slijedi iz Leme 2.7.6 i definicije kvazireziduala z_k^Q. □

Teorem 2.7.10 pokazuje da ako se norma QMR kvazi-reziduala reducira sa značajnim faktorom u k-tom koraku, tada će norma BCG reziduala biti približno jednaka normi QMR kvazi-reziduala u k-tom koraku, jer će u tom slučaju djelitelj u desnoj strani od (2.122) biti blizu 1. Ako norma QMR kvazi-reziduala ostane skoro konstantna, tada će djelitelj u desnoj strani od (2.122) biti blizu 0, a norma BCG reziduala bit će puno veća. U tom slučaju, dok će krivulja norme QMR kvazi-reziduala biti skoro horizontalna, krivulja norme BCG reziduala imat će oscilacije, odnosno šiljak okrenut prema gore. Stupanj "horizontalnosti" jedne krivulje uvjetuje amplitudu oscilacije druge. Relacija (2.122) uglavnom demonstrira povezanost BCG i QMR metoda: ili će obje metode dobro konvergirati, ili će obje imati slabu konvergenciju za dani problem.

2.7.5 Metoda kvadriranih konjugiranih gradijenata (CGS)

Metode BCG i QMR zahtijevaju množenje vektora sa matricom A ali i sa matricom A^*. To zahtijeva dodatni posao, a naročito kada je kompliciranije množiti sa A^* nego sa A. To se događa, na primjer, kada A nije eksplicitno smještena u memoriji, već kada postoji samo procedura koja definira djelovanje matrice na zadani vektor, a posebno je problematično kada se sustav rješava na paralelnim računalima. Zbog toga je poželjno imati iterativnu metodu koja će zahtijevati samo množenje sa matricom A, a neće biti puno zahtjevnija od BCG metode. Takva metoda je *metoda kvadriranih konjugiranih gradijenata* (Conjugate gradient squared method) ili CGS metoda.

Krenut ćemo od BCG algoritma, odakle možemo primijetiti da vrijedi

$$r_k = \phi_k(A)r_0, \qquad \hat{r}_k = \bar{\phi}_k(A^*)\hat{r}_0,$$

$$p_k = \psi_k(A)r_0, \qquad \hat{p}_k = \bar{\psi}_k(A^*)\hat{r}_0$$

za određene polinome k-tog stupnja ϕ_k i ψ_k. Ako algoritam dobro konvergira, tada je $\|\phi_k(A)r_0\|_2$ mala, pa bi mogli očekivati da je $\|\phi_k^2(A)r_0\|_2$ još manja. Ako još pokušamo izračunati $\phi_k^2(A)r_0$ sa otprilike jednako mnogo operacija kao i $\phi_k(A)r_0$ tada bi to najvjerojatnije bio brzo konvergirajući algoritam. To je bila osnovna ideja razvoja CGS metode.

Sada ćemo raspisati BCG rekurzije preko izraza sa polinomima ϕ_k i ψ_k. Imamo

$$\phi_k(A)r_0 = \phi_{k-1}(A)r_0 - \alpha_{k-1}A\psi_{k-1}(A)r_0, \tag{2.123}$$

$$\psi_k(A)r_0 = \phi_k(A)r_0 + \beta_{k-1}\psi_{k-1}(A)r_0, \tag{2.124}$$

gdje su

$$\alpha_{k-1} = \frac{\langle \phi_{k-1}(A)r_0, \bar{\phi}_{k-1}(A^*)\hat{r}_0\rangle}{\langle A\psi_{k-1}(A)r_0, \bar{\psi}_{k-1}(A^*)\hat{r}_0\rangle} = \frac{\langle \phi_{k-1}^2(A)r_0, \hat{r}_0\rangle}{\langle A\psi_{k-1}^2(A)r_0, \hat{r}_0\rangle}, \tag{2.125}$$

$$\beta_{k-1} = \frac{\langle \phi_k(A)r_0, \bar{\phi}_k(A^*)\hat{r}_0\rangle}{\langle \phi_{k-1}(A)r_0, \bar{\phi}_{k-1}(A^*)\hat{r}_0\rangle} = \frac{\langle \phi_k^2(A)r_0, \hat{r}_0\rangle}{\langle \phi_{k-1}^2(A)r_0, \hat{r}_0\rangle}. \tag{2.126}$$

Dakle, koeficijenti rekurzija mogu se izračunati ako znamo $\hat{r}_0$, $\phi_j^2(A)r_0$, i $\psi_j^2(A)r_0$ za $j = 1, 2, \ldots$.

Iz (2.123) i (2.124), vidimo da polinomi $\phi_k(z)$ i $\psi_k(z)$ zadovoljavaju rekurzije

$$\phi_k(z) = \phi_{k-1}(z) - \alpha_{k-1}z\psi_{k-1}(z),$$

$$\psi_k(z) = \phi_k(z) + \beta_{k-1}\psi_{k-1}(z),$$

pa kvadrirajući obje jednakosti imamo

$$\phi_k^2(z) = \phi_{k-1}^2(z) - 2\alpha_{k-1}z\phi_{k-1}(z)\psi_{k-1}(z) + \alpha_{k-1}^2z^2\psi_{k-1}^2(z),$$

$$\psi_k^2(z) = \phi_k^2(z) + 2\beta_{k-1}\phi_k(z)\psi_{k-1}(z) + \beta_{k-1}^2\psi_{k-1}^2(z).$$

Da nema mješovitih izraza $\phi_{k-1}(z)\psi_{k-1}(z)$ i $\phi_k(z)\psi_{k-1}(z)$, ove jednakosti bi tvorile jednostavne iteracije za ϕ_k^2 i ψ_k^2. Izlaz iz ove situacije je uvođenje jednog od tih mješovitih izraza, $\phi_k(z)\psi_{k-1}(z)$ kao trećeg člana rekurzije. Drugi izraz možemo izraziti preko ova tri člana rekurzije. Množenjem ϕ_{k-1} sa rekurzijom za ψ_{k-1} dobivamo

$$\phi_{k-1}(z)\psi_{k-1}(z) = \phi_{k-1}^2(z) + \beta_{k-2}\phi_{k-1}(z)\psi_{k-2}(z),$$

a množenje rekurzije za ϕ_k sa ψ_{k-1} daje

$$\begin{aligned}\phi_k(z)\psi_{k-1}(z) &= \phi_{k-1}(z)\psi_{k-1}(z) - \alpha_{k-1}z\psi_{k-1}^2(z) = \\ &= \phi_{k-1}^2(z) + \beta_{k-2}\phi_{k-1}(z)\psi_{k-2}(z) - \alpha_{k-1}z\psi_{k-1}^2(z).\end{aligned}$$

Ako sakupimo sve ove relacije na jedno mjesto, dobit ćemo rekurzije za polinome, čija je definicija

$$\Phi_k(z) = \phi_k^2(z), \quad \Theta_k(z) = \phi_k(z)\psi_{k-1}(z), \quad \Psi_k(z) = \psi_k^2(z),$$

koje glase

$$\begin{aligned}\Phi_k(z) &= \Phi_k(z) - 2\alpha_{k-1}z(\Phi_{k-1}(z) + \beta_{k-2}\Theta_{k-1}(z)) + \alpha_{k-1}^2 z^2\Psi_{k-1}(z), \\ \Theta_k(z) &= \Phi_{k-1}(z) + \beta_{k-2}\Theta_{k-1}(z) - \alpha_{k-1}z\Psi_{k-1}(z), \\ \Psi_k(z) &= \Phi_k(z) + 2\beta_{k-1}\Theta_k(z) + \beta_{k-1}^2\Psi_{k-1}(z).\end{aligned}$$

Ove rekurzije su osnova algoritma. Ako sada definiramo nove vrijednosti

$$\begin{aligned}r_k &= \Phi_k(A)r_0, \\ q_k &= \Theta_k(A)r_0, \\ p_k &= \Psi_k(A)r_0,\end{aligned}$$

tada gornje rekurzije za polinome prelaze u

$$\begin{aligned}r_k &= r_{k-1} - \alpha_{k-1}A(2r_{k-1} + 2\beta_{k-2}q_{k-1} - \alpha_{k-1}Ap_{k-1}), \\ q_k &= r_{k-1} + \beta_{k-2}q_{k-1} - \alpha_{k-1}Ap_{k-1}, \\ p_k &= r_k + 2\beta_{k-1}q_k + \beta_{k-1}^2 p_{k-1}.\end{aligned}$$

Pogodno bi bilo uvesti najprije jedan pomoćni vektor

$$s_{k-1} = 2r_{k-1} + 2\beta_{k-2}q_{k-1} - \alpha_{k-1}Ap_{k-1}.$$

Sa ovime dobivamo sljedeći niz operacija za računanje aproksimacije rješnja u svakoj iteraciji x_k, koje su generirane tako da je r_k zadanog oblika $r_k = \phi_k^2(A)r_0$, pri čemu započinjemo sa $r_0 = b - Ax_0$, $p_0 = r_0$, $q_0 = 0$ i $\beta_0 = 0$.

$$\alpha_{k-1} = \frac{\langle r_{k-1}, \hat{r}_0\rangle}{\langle Ap_{k-1}, \hat{r}_0\rangle},$$

$$s_{k-1} = 2r_{k-1} + 2\beta_{k-2}q_{k-1} - \alpha_{k-1}Ap_{k-1},$$

$$q_k = r_{k-1} + \beta_{k-2}q_{k-1} - \alpha_{k-1}Ap_{k-1},$$

$$x_k = x_{k-1} + \alpha_{k-1}s_{k-1},$$

$$r_k = r_{k-1} - \alpha_{k-1}As_{k-1},$$

$$\beta_{k-1} = \frac{\langle r_k, \hat{r}_0\rangle}{\langle r_{k-1}, \hat{r}_0\rangle},$$

$$p_k = r_k + 2\beta_{k-1}q_k + \beta_{k-1}^2 p_{k-1}.$$

Malo pojednostavljenje algoritma može se napraviti ako uvedemo još jedan pomoćni vektor

$$u_{k-1} = r_{k-1} + \beta_{k-2} q_{k-1}.$$

Ta definicija vodi do relacija

$$q_k = u_{k-1} - \alpha_{k-1} A p_{k-1},$$

$$s_{k-1} = u_{k-1} + q_k,$$

$$p_k = u_k + \beta_{k-1}(q_k + \beta_{k-1} p_{k-1}),$$

pa kao rezultat ovoga vektor s_{k-1} više nije potreban. Sljedeći algoritam generira aproksimaciju rješenja x_k sa traženim rezidualom r_k.

Algoritam 2.7.11. KVADRIRANI KONJUGIRANI GRADIJENTI (CGS)

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

Izaberi $\hat{r}_0$ *takav da je* $\langle r_0, \hat{r}_0 \rangle \neq 0$.

$p_0 = r_0$,

$u_0 = r_0$.

Za $k = 1, 2, \ldots$

izračunaj Ap_{k-1},
$\alpha_{k-1} = \dfrac{\langle r_{k-1}, \hat{r}_0 \rangle}{\langle Ap_{k-1}, \hat{r}_0 \rangle}$,
$q_k = u_{k-1} - \alpha_{k-1} A p_{k-1}$,
$x_k = x_{k-1} + \alpha_{k-1}(u_{k-1} + q_k)$,
izračunaj $A(u_{k-1} + q_k)$,
$r_k = r_{k-1} - \alpha_{k-1} A(u_{k-1} + q_k)$,
$\beta_{k-1} = \dfrac{\langle r_k, \hat{r}_0 \rangle}{\langle r_{k-1}, \hat{r}_0 \rangle}$,
$u_k = r_k + \beta_{k-1} q_k$,
$p_k = u_k + \beta_{k-1}(q_k + \beta_{k-1} p_{k-1})$.

Primijetimo da u ovom algoritmu zaista nama množenja vektora sa matricom A^*, ali umjesto toga u svakom koraku se izvode dva množenja vektora sa matricom A.

2.7.6 Metoda stabiliziranih bikonjugiranih gradijenata (BICGSTAB)

CGS metoda je bazirana na kvadriranju rezidualnog polinoma, i u slučaju neregularne konvergencije, ona može dovesti do značajnih grešaka zaokruživanja, pa čak može proizvesti i vrijednosti koje su izvan raspoloživog raspona u skupu strojnih brojeva. Naime,

kada je norma BCG reziduala mala, norma CGS reziduala je obično još puno manja, ali ako je norma BCG reziduala velika, norma CGS reziduala je još veća, pa su oscilacije kod CGS puno veće nego kod BCG metode. *Metoda stabiliziranih bikonjugiranih gradijenata* (BICGSTAB) je varijacija CGS metode koja je bila razvijena s ciljem da popravi ovaj problem.

U CGS metodi definirali smo rekurzije tako da rezidual r_k zadovoljava jednakost $r_k = \phi_k(A)^2 r_0$, kod kojeg je $\phi_k(A)r_0$ rezidual BCG metode. Prema konstrukciji imali smo $\langle \phi_k(A)r_0, \bar{\phi}_j(A)r_0\rangle = 0$ za $j < k$, zbog biortogonalnosti nizova BCG reziduala r_k i $\hat{r}_k$, što je zapravo ekvivalentno sa činjenicom da je $\phi_k(A)r_0$ okomit na $\mathcal{K}_k(A^*, \hat{r}_0) = \text{span}\{\hat{r}_0, A^*\hat{r}_0, \ldots, (A^*)^{k-1}\hat{r}_0\}$. Iz toga dalje slijedi da mi možemo dobiti BCG parametre zahtijevajući da je, na primijer, r_k okomit na $\bar{\chi}_j(A^*)\hat{r}_0$, odnosno da je $\langle \chi_j(A)\phi_k(A)r_0, \hat{r}_0\rangle = 0$ za neki drugi pogodni polinom χ_j stupnja $j < k$. U BCG metodi je $\chi_j = \phi_j$, jer je $\hat{r}_j = \bar{\phi}_j(A^*)\hat{r}_0$, što je iskorišteno u CGS metodi, budući da se rekurzije za vektore $\phi_j^2(A)r_0$ mogu dobiti iz onih za $\phi_j(A)r_0$.

Naravno, mi možemo konstruirati druge iterativne metode, za koje su aproksimacije x_k generirane tako da je

$$r_k = \chi_k(A)\phi_k(A)r_0,$$

gdje je ϕ_k ponovo BCG rezidualni polinom, a χ_k je izabran tako da pokuša držati normu reziduala malom u svakom koraku, ali s druge strane, da zadrži globalnu brzu konvergenciju. Jedna mogućnost izbora polinoma χ_k je polinom oblika

$$\chi_k = (1 - \omega_k z)(1 - \omega_{k-1} z)\cdots(1 - \omega_1 z), \tag{2.127}$$

pri čemu koeficijenti ω_j mogu biti izabrani tako da u svakom koraku minimiziraju

$$\|r_k\|_2 = \|(I - \omega_k A)\chi_{k-1}(A)\phi_k(A)r_0\|_2.$$

Ovo je glavna ideja BICGSTAB metode, koja se može smatrati kombinacijom metoda BCG i Orthomin(1).

Ponovo, neka $\phi(A)r_0$ označva BCG rezidual u k-tom koraku, a $\psi_k(A)r_0$ neka označava BCG vektor smjera p_k, i neka oba polinoma zadovoljavaju rekurzije (2.123) i (2.124). U BICGSTAB algoritmu želimo naći rekurzije za

$$r_k = \chi_k(A)\phi_k(A)r_0$$

i

$$p_k = \chi_k(A)\psi_k(A)r_0.$$

Iz (2.127), (2.123) i (2.124) slijedi

$$\begin{aligned} r_k &= (1 - \omega_k A)\chi_{k-1}(A)(\phi_{k-1}(A) - \alpha_{k-1}A\psi_{k-1}(A))r_0 = \\ &= (I - \omega_k A)(r_{k-1} - \alpha_{k-1}Ap_{k-1}), \end{aligned}$$

$$\begin{aligned} p_k &= \chi_k(A)(\phi_k(A) + \beta_{k-1}\psi_{k-1}(A))r_0 = \\ &= (\chi_k(A)\phi_k(A) + \beta_{k-1}(I - \omega_k A)\chi_{k-1}(A)\psi_{k-1}(A))r_0 = \\ &= r_k + \beta_{k-1}(I - \omega_k A)p_{k-1}. \end{aligned} \tag{2.128}$$

Još nam je preostalo izraziti BCG koeficijente α_{k-1} i β_{k-1} preko skalarnih produkata upravo definiranih vektora. Ono što ćemo koristiti je svojstvo biortogonalnosti koje se

pojavljuje u BCG metodi, naime, znamo da u toj metodi vrijedi $\langle r_k, \hat{r}_j\rangle = 0$ i $\langle Ap_k, \hat{p}_j\rangle = 0$ za $j \neq k$, pri čemu su $r_k, p_k \in \mathcal{K}_k(A, r_0)$, i $\hat{r}_j, \hat{p}_j \in \mathcal{K}_j(A^*, \hat{r}_0)$. To znači da je

$$\langle \phi_k(A)r_0, (A^*)^j\hat{r}_0\rangle = 0, \qquad \text{za } j = 0, 1, \ldots, k-1,$$

i

$$\langle A\psi_k(A)r_0, (A^*)^j\hat{r}_0\rangle = 0, \qquad \text{za } j = 0, 1, \ldots, k-1.$$

Krenimo od definicije koeficijenta β_{k-1} (2.126). Iz rekurzija (2.123) i (2.124) imamo

$$\phi_k(z) = -\alpha_{k-1}z\phi_{k-1}(z) + (\phi_{k-1}(z) - \alpha_{k-1}\beta_{k-2}z\psi_{k-2}(z)),$$

$$\psi_k(z) = \phi_k(z) + \beta_{k-1}\psi_{k-1}(z),$$

odakle se vidi da je vodeći koeficijent polinoma ϕ_k jednak $-\alpha_{k-1}$ puta vodeći koeficijent od ϕ_{k-1}, dok je vodeći koeficijent od ψ_k jednak onom od ϕ_k. Matematičkom indukcijom možemo zaključiti da je vodeći koeficijent od ϕ_k jednak $(-1)^k\alpha_0\cdots\alpha_{k-1}$. S druge strane vodeći koeficijent polinoma χ_k je očigledno dan sa $(-1)^k\omega_1\cdots\omega_k$. Dakle, skalarni produkti koji se pojavljuju u β_{k-1}, definirani su preko $\phi_k^2(A)r_0$ i $\phi_{k-1}^2 r_0$, međutim, niti jedan od ta dva vektora nam nije dostupan, već na raspolaganju imamo vektore oblika $\chi_k(A)\phi_k(A)r_0$. Sada ćemo iskoristiti prethodna razmatranja kako bi dobili sljedeće relacije.

$$\begin{aligned}
\langle \phi_k(A)r_0, \bar{\phi}_k(A^*)\hat{r}_0\rangle &= (-1)^k\alpha_0\cdots\alpha_{k-1}\langle \phi_k(A)r_0, (A^*)^k\hat{r}_0\rangle = \\
&= \frac{(-1)^k\alpha_0\cdots\alpha_{k-1}}{(-1)^k\omega_1\cdots\omega_k}\langle \phi_k(A)r_0, \bar{\chi}_k(A^*)\hat{r}_0\rangle = \\
&= \frac{\alpha_0\cdots\alpha_{k-1}}{\omega_1\cdots\omega_k}\langle r_k, \hat{r}_0\rangle,
\end{aligned}$$

odakle slijedi da je

$$\beta_{k-1} = \frac{\alpha_{k-1}}{\omega_k}\frac{\langle r_k, \hat{r}_0\rangle}{\langle r_{k-1}, \hat{r}_0\rangle}.$$

Sličan postupak napravimo i za α_{k-1} iz definicije (2.125). Za to nam još treba na pogodan način transformirati skalarni produkt $\langle A\psi_{k-1}(A)r_0, \bar{\psi}_{k-1}(A^*)\hat{r}_0\rangle$.

$$\begin{aligned}
\langle A\psi_{k-1}(A)r_0, \bar{\psi}_{k-1}(A^*)\hat{r}_0\rangle &= (-1)^{k-1}\alpha_0\cdots\alpha_{k-2}\langle A\psi_{k-1}(A)r_0, (A^*)^{k-1}\hat{r}_0\rangle = \\
&= \frac{(-1)^{k-1}\alpha_0\cdots\alpha_{k-2}}{(-1)^{k-1}\omega_1\cdots\omega_{k-1}}\langle A\psi_{k-1}(A)r_0, \bar{\chi}_{k-1}(A^*)\hat{r}_0\rangle \\
&= \frac{\alpha_0\cdots\alpha_{k-2}}{\omega_1\cdots\omega_{k-1}}\langle Ap_{k-1}r_0, \hat{r}_0\rangle,
\end{aligned}$$

odakle imamo da je

$$\alpha_{k-1} = \frac{\langle r_{k-1}, \hat{r}_0\rangle}{\langle Ap_{k-1}, \hat{r}_0\rangle}.$$

Nadalje, trebamo još odrediti parametar ω_k. Najjednostavniji izbor bi bio da ω_k minimizira euklidsku normu vektora $(I - \omega_k A)\chi_{k-1}(A)\phi_k(A)r_0$. Rekurziju za r_k tada možemo drugačije napisati kao

$$r_k = (I - \omega_k A)r_{k-1/2},$$

pri čemu je

$$r_{k-1/2} = \chi_{k-1}(A)(\phi_{k-1}(A) - \alpha_{k-1}A\psi_{k-1}(A))r_0 = r_{k-1} - \alpha_{k-1}Ap_{k-1}.$$

Nakon minimiziranja funkcije $f(\omega_k) = \|(I - \omega_k A)r_{k-1/2}\|_2^2$, kao kod Orthomin(1), dobivamo optimalni parametar ω_k dan sa

$$\omega_k = \frac{\langle r_{k-1/2}, Ar_{k-1/2}\rangle}{\langle Ar_{k-1/2}, Ar_{k-1/2}\rangle}.$$

Na kraju trebamo još dobiti jednakost iz koje se aproksimacija x_k dobiva iz aproksimacije x_{k-1}. Imamo

$$r_k = r_{k-1/2} - \omega_k Ar_{k-1/2} = r_{k-1} - \alpha_{k-1}Ap_{k-1} - \omega_k Ar_{k-1/2},$$

što daje

$$x_k = x_{k-1} + \alpha_{k-1}p_{k-1} + \omega_{k-1}r_{k-1/2}.$$

Sada smo izveli sve potrebne relacije, čime napokon možemo zaokružiti kompletan BICGSTAB algoritam.

Algoritam 2.7.12. Stabilizirani bikonjugirani gradijenti (BICGSTAB)

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

Izaberi $\hat{r}_0$ *takav da je* $\langle r_0, \hat{r}_0\rangle \neq 0$.

$p_0 = r_0$.

Za $k = 1, 2, \ldots$

izračunaj Ap_{k-1},
$\alpha_{k-1} = \dfrac{\langle r_{k-1}, \hat{r}_0\rangle}{\langle Ap_{k-1}, \hat{r}_0\rangle}$,
$x_{k-1/2} = x_{k-1} + \alpha_{k-1}p_{k-1}$,
$r_{k-1/2} = r_{k-1} - \alpha_{k-1}Ap_{k-1}$,
izračunaj $Ar_{k-1/2}$,
$\omega_k = \dfrac{\langle r_{k-1/2}, Ar_{k-1/2}\rangle}{\langle Ar_{k-1/2}, Ar_{k-1/2}\rangle}$,
$x_k = x_{k-1/2} + \omega_k r_{k-1/2}$,
$r_k = r_{k-1/2} - \omega_k Ar_{k-1/2}$,
$\beta_{k-1} = \dfrac{\alpha_{k-1}}{\omega_k}\dfrac{\langle r_k, \hat{r}_0\rangle}{\langle r_{k-1}, \hat{r}_0\rangle}$,
$p_k = r_k + \beta_{k-1}(p_{k-1} - \omega_k Ap_{k-1})$.

Zbog ortogonalnog svojstva $\langle \phi_i(A)r_0, \chi_j(A^*)\hat{r}_0\rangle = 0$, za $j < i$, slijedi da je BICGSTAB, kao i CGS, je konačna motoda, jer u egzaktnoj aritmetici ona će završiti nakon $m \leq n$ iteracija. U tom slučaju je $r_{m-1/2} = 0$, pa ω_m nije definiran, što predstavlja regularno zaustavljanje. Tada se može preskočiti množenje sa $(I - \omega_m A)$, r_m je tako i onako jednak 0, a aproksimacije se može izračunati samo sa $x_m = x_{m-1/2} = x_{m-1} + \alpha_{m-1}p_{m-1}$. Po broju operacija BICGSTAB obavlja nešto malo više posla po iteraciji nego CGS, ali se to nadoknađuje gotovo uvijek kroz manji broj iteracija. Jedini problem koji se javlja i u CGS, i u BICGSTAB metodi je mogućnost zakazivanja algotitma. Naime, obje metode zakazat će u istim iteracijama kada bi zakazala i BCG metoda, jer se u pozadini obaju metoda nalazi dvostrani Lanczosov algoritam. Za rješavanje ovakvog sloma algoritma ponovo treba uvesti tehnike provjere unaprijed.

2.7.7 Konvergencija metoda CGS i BICGSTAB

Probleme konvergencije BCG metode, kao i metoda koje su bazirane na njoj, opisao je van der Vorst u [36]. Oni su bili glavni razlog koji su ga ponukali na razvoj nove metode: BICGSTAB.

Kod problema kod kojih BCG metoda dobro konvergira, CGS metoda obično zahtijeva duplo manje iteracija za postizanje željene točnosti od BCG metode. Dakle, u mnogim slučajevima kada BCG konvergira CGS konvergira duplo brže, ali, slično tome, kada BCG divergira, tada CGS divergira otprilike dva puta brže. To ponašanje se može očekivati jer, se BCG operator $\phi_k(A)$ primijenjuje dva puta na r_0 u CGS metodi. Kada se norma BCG reziduala smanji ili poveća u nekom koraku, tada se norma CGS reziduala otprilike smanji ili poveća za kvadrat smanjenja ili povećanja norme BCG reziduala. Zbog toga krivulja konvergencije CGS metode može imati divlje oscilacije, veće od onih kod BCG metode, što može dovesti i do numeričke nestabilnosti.

Međutim, postoji jedan problem kod promatranja CGS reziduala na ovakav način. Slaba točka je u tome što je redukcijski BCG operator $\phi_k(A)$ jako ovisan o početnom rezidualu r_0, i što vjerojatno on neće biti reduktivni operator za bilo koji drugi vektor, pa čak ni za vektor $\phi_k(A)r_0$ na koji se primijenjuje. Zaista se mogu konstruirati primjeri za koje je $\phi_k(A)r_0$ mali po normi, a $\phi_k^2(A)r_0$ mnogo veći po normi. Tako nešto se može dogoditi ako pretpostavimo da r_0 ima male koordinate u smjeru nekih svojstvenih vektora matrice A. Tada $|\phi_k(\lambda)|$ može biti velik za odgovarajuće svojstvene vrijednosti, naročito kada su one više ili manje izolirane od ostalih, ali smjerovi tih svojstvenih vektora BCG reziduala ne moraju davati veliki doprinos njegovoj normi. Međutim, vrijednosti $|\phi_k^2(\lambda)|$ mogu biti tako velike da postanu dominante, i odnesu prevagu u odnosu na ostale kod utjecaja na iznos norme CGS reziduala.

Takve situacije se često događaju, pa u jednoj krivulji konvergencije CGS metode možemo primijetiti mnoge lokalne šiljke. S druge strane oni izgleda ne usporavaju konvergenciju CGS metode. Ali, ipak oni mogu biti vrlo štetni, jer mogu biti tako veliki, da odgovarajuća lokalna korekcija (padajuća strana šiljka) može izazvati fatalno kračenje. Kao rezultat, konačno rješenje može imati veliki gubitak u točnosti, što se može provjeriti računanjem pravog reziduala $r = b - Ax$.

U mnogim primjerima krivulja konvergencije BICGSTAB metode je puno glađa od CGS metode, što je jasno zbog odabira oblika reziduala BICGSTAB metode. Također, u mnogim slučajevima BICGSTAB je puno efektivnija metoda od CGS, u smislu da je potrebno obaviti manje posla za postizanje iste točnosti.

2.7.8 Prekondicionirani algoritmi

Za sve algoritme u ovom poglavlju promatrat ćemo dvostruko prekondicioniranje, jer se lijevo i desno prekondicioniranje lako mogu dobiti iz njega. Dakle, promatramo matricu prekondicioniranja M koju na neki način možemo faktorizirati na

$$M = LU.$$

Tada zapravo rješavamo prekondicionirani sustav

$$L^{-1}AU^{-1}\check{x} = L^{-1}b, \qquad x = U^{-1}\check{x}. \tag{2.129}$$

Označimo sve veličine vezane uz prekondicionirani sustav (2.129) sa $\check{}$, a sve veličine vezane uz početni sustav neka ostanu označene standardnim oznakama.

Za prekondicioniranu BCG metodu vrijede sljedeće relacije:

$$r_k = L\check{r}_k, \qquad \hat{r}_k = U^*\check{\hat{r}}_k,$$

$$p_k = U^{-1}\check{p}_k, \qquad \hat{p}_k = L^{-*}\check{\hat{p}}_k,$$

tada se BCG algoritam primijenjen na sustav (2.129) može transformirati u sljedeći algoritam.

Algoritam 2.7.13. Prekondicionirani BCG

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

riješi $Mq_0 = r_0$.

Izaberi $\hat{r}_0$ *takav da je* $\langle q_0, \hat{r}_0 \rangle \neq 0$.

Riješi $M^*\hat{q}_0 = \hat{r}_0$,

$p_0 = q_0$,

$\hat{p}_0 = \hat{q}_0$.

Za $k = 1, 2, \ldots$

izračunaj Ap_{k-1},
izračunaj $A^*\hat{p}_{k-1}$,
$\alpha_{k-1} = \dfrac{\langle q_{k-1}, \hat{r}_{k-1} \rangle}{\langle Ap_{k-1}, \hat{p}_{k-1} \rangle}$,
$x_k = x_{k-1} + \alpha_{k-1} p_{k-1}$,
$r_k = r_{k-1} - \alpha_{k-1} Ap_{k-1}$,
$\hat{r}_k = \hat{r}_{k-1} - \bar{\alpha}_{k-1} A^*\hat{p}_{k-1}$,
riješi $Mq_k = r_k$,
riješi $M^*\hat{q}_k = \hat{r}_k$,
$\beta_{k-1} = \dfrac{\langle q_k, \hat{r}_k \rangle}{\langle q_{k-1}, \hat{r}_{k-1} \rangle}$,
$p_k = q_k + \beta_{k-1} p_{k-1}$,
$\hat{p}_k = \hat{q}_k + \bar{\beta}_{k-1} \hat{p}_{k-1}$.

Na žalost kod QMR metode rješavanje prekondicioniranog sustava (2.129) ne može se transformirati u oblik koji koristi samo matricu prekondicioniranja M, već se pojavljuju relacije u kojima ne možemo izbjeći pojavu faktora L i U matrice M. Zato se QMR metoda jednostavno primijeni na sustav (2.129), kao rezultat dobije se rješenje z, koje se na kraju transformira u rješenje $x = U^{-1}z$.

Algoritam 2.7.14. Prekondicionirani QMR

Dana je početna iteracija x_0,

riješi $Lr_0 = b - Ax_0$,

$\beta = \|r_0\|_2$,

$v_1 = \dfrac{r_0}{\beta}$,

Dan je $\hat{r}_0$,

$w_1 = \dfrac{\hat{r}_0}{\|\hat{r}_0\|_2}$,

$d = (1, 0, \ldots, 0)^T$.

Za $k = 1, 2, \ldots$

Izračunaj v_{k+1}, w_{k+1}, $\alpha_k = T(k,k)$, $\beta_k = T(k,k+1)$, *i* $\gamma_k = T(k+1,k)$, *koristeći dvostrani Lanczosov algoritam primijenjen na matricu* $L^{-1}AU^{-1}$.

Primijeni F_{k-2} *i* F_{k-1} *na zadnji stupac od* T, *odnosno:*

ako je $k > 2$ *tada* $\begin{bmatrix} T(k-2,k) \\ T(k-1,k) \end{bmatrix} := \begin{bmatrix} c_{k-2} & s_{k-2} \\ -\bar{s}_{k-2} & c_{k-2} \end{bmatrix} \begin{bmatrix} 0 \\ T(k-1,k) \end{bmatrix}$,

ako je $k > 1$ *tada* $\begin{bmatrix} T(k-1,k) \\ T(k,k) \end{bmatrix} := \begin{bmatrix} c_{k-1} & s_{k-1} \\ -\bar{s}_{k-1} & c_{k-1} \end{bmatrix} \begin{bmatrix} T(k-1,k) \\ T(k,k) \end{bmatrix}$.

Izračunaj k*-tu Givensovu rotaciju* F_k *kako bi se poništio* $(k+1,k)$ *element od* T*:*

$c_k = \dfrac{|T(k,k)|}{\sqrt{|T(k,k)|^2 + |T(k+1,k)|^2}}$,

ako je $c_k \neq 0$ *tada* $s_k = c_k \dfrac{\overline{T(k+1,k)}}{\overline{T(k,k)}}$, *ako je* $c_k = 0$ *tada* $s_k = 1$.

Primijeni k*-tu rotaciju na* d *i na zadnji stupac od* T*:*

$\begin{bmatrix} d(k) \\ d(k+1) \end{bmatrix} := \begin{bmatrix} c_k & s_k \\ -\bar{s}_k & c_k \end{bmatrix} \begin{bmatrix} d(k) \\ 0 \end{bmatrix}$,

$T(k,k) := c_k T(k,k) + s_k T(k+1,k)$,

$T(k+1,k) = 0 \qquad (*)$.

Izračunaj $p_{k-1} = (1/T(k,k))[v_k - T(k-1,k)p_{k-2} - T(k-2,k)p_{k-3}]$, *gdje su* p_{-1}, p_{-2}, *jednaki nuli.*

$z_k = z_{k-1} + \beta d(k) p_{k-1}$.

Ako je ocjena norme reziduala $\beta\sqrt{k+1}|s_1 \cdots s_k|$ *dovoljno mala, tada :*

riješi $Ux_k = z_k$.

(*) $T(k+1,k)=0$ treba zapravo izvesti u sljedećem, $(k+1)$-om koraku, jer je originalna vrijednost $T(k+1,k)$ potrebna za izvođenje $(k+1)-og$ koraka dvostranog Lanczosovog algoritma.

Za prekondicioniranu CGS metodu vrijede sljedeće relacije:

$$r_k = L\check{r}_k, \qquad q_k = U^{-1}\check{q}_k,$$

$$u_k = U^{-1}\check{u}_k, \qquad p_k = U^{-1}\check{p}_k,$$

$$\hat{r}_0 = U^*\check{\hat{r}}_0,$$

pa se CGS algoritam primijenjen na sustav (2.129) može transformirati u sljedeći algoritam.

Algoritam 2.7.15. PREKONDICIONIRANI CGS

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

riješi $Ms_0 = r_0$.

Izaberi $\hat{r}_0$ *takav da je* $\langle s_0, \hat{r}_0\rangle \neq 0$.

$p_0 = s_0$,

$u_0 = s_0$.

Za $k = 1, 2, \ldots$

izračunaj Ap_{k-1},
riješi $Mt_{k-1} = Ap_{k-1}$,
$\alpha_{k-1} = \dfrac{\langle s_{k-1}, \hat{r}_0\rangle}{\langle t_{k-1}, \hat{r}_0\rangle}$,
$q_k = u_{k-1} - \alpha_{k-1}t_{k-1}$,
$x_k = x_{k-1} + \alpha_{k-1}(u_{k-1} + q_k)$,
izračunaj $A(u_{k-1} + q_k)$,
$r_k = r_{k-1} - \alpha_{k-1}A(u_{k-1} + q_k)$,
riješi $Ms_k = r_k$,
$\beta_{k-1} = \dfrac{\langle s_k, \hat{r}_0\rangle}{\langle s_{k-1}, \hat{r}_0\rangle}$,
$u_k = s_k + \beta_{k-1}q_k$,
$p_k = u_k + \beta_{k-1}(q_k + \beta_{k-1}p_{k-1})$.

U prekondicioniranoj CGS metodi, kao i kod BCG metode, faktori L i U ne igraju nikakvu ulogu u algoritmu, pa bilo koji izbor faktora matrice M odgovara nekom izboru vektora $\check{\hat{r}}_0$ u prekondicioniranom algoritmu. To znači da umjesto traženja pogodnog oblika prekondicioniranja, možemo tražiti pogodan izbor vektora $\hat{r}_0$.

Za prekondicioniranu BICGSTAB metodu vrijede sljedeće relacije:

$$r_k = L\check{r}_k, \qquad r_{k-1/2} = L\check{r}_{k-1/2},$$

$$p_k = L\check{p}_k, \qquad \hat{r}_0 = L^{-*}\check{\hat{r}}_0,$$

pa se BICGSTAB algoritam primijenjen na sustav (2.129) može transformirati u sljedeći algoritam.

Algoritam 2.7.16. PREKONDICIONIRANI BICGSTAB

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

Izaberi $\hat{r}_0$ *takav da je* $\langle r_0, \hat{r}_0 \rangle \neq 0$.

$p_0 = r_0$.

Za $k = 1, 2, \ldots$

riješi $Ms_{k-1} = p_{k-1}$,
izračunaj As_{k-1},
$\alpha_{k-1} = \dfrac{\langle r_{k-1}, \hat{r}_0 \rangle}{\langle As_{k-1}, \hat{r}_0 \rangle}$,
$x_{k-1/2} = x_{k-1} + \alpha_{k-1} s_{k-1}$,
$r_{k-1/2} = r_{k-1} - \alpha_{k-1} As_{k-1}$,
riješi $Mt_{k-1/2} = r_{k-1/2}$,
izračunaj $At_{k-1/2}$,
riješi $Lu_{k-1/2} = r_{k-1/2}$,
riješi $Lv_{k-1/2} = At_{k-1/2}$,
$\omega_k = \dfrac{\langle u_{k-1/2}, v_{k-1/2} \rangle}{\langle v_{k-1/2}, v_{k-1/2} \rangle}$,
$x_k = x_{k-1/2} + \omega_k t_{k-1/2}$,
$r_k = r_{k-1/2} - \omega_k At_{k-1/2}$,
$\beta_{k-1} = \dfrac{\alpha_{k-1}}{\omega_k} \dfrac{\langle r_k, \hat{r}_0 \rangle}{\langle r_{k-1}, \hat{r}_0 \rangle}$,
$p_k = r_k + \beta_{k-1}(p_{k-1} - \omega_k As_{k-1})$.

Kod BICGSTAB metode, međutim postoji razlike između izbora različitih oblika prekondicioniranja, ne mogu se prikazati kao pogodan izbor vektora $\hat{r}_0$, i to zbog izraza za ω_k. Naime, kod rješavanja prekondicioniranog sustava, algoritam se ne može transformirati da u potpunosti ne bude ovisan o faktorima L i U. Uvijek će se pojaviti inverz prvog faktora kada pokušamo izračunati ω_k za prekondicionirani sustav.

2.8 Simetrizacija problema

2.8.1 CGNR i CGNE metode

Postoji još jedan način na koji možemo pristupiti rješavanju nehermitskog sustava, a to je simetrizacija problema. On se svodi na transformaciju nehermitskog problema na hermitski, rješavanjem jedne od *normalnih jednadžbi*

$$A^*Ax = A^*b \quad \text{ili} \quad AA^*\hat{x} = b,\ x = A^*\hat{x}. \tag{2.130}$$

Rješavanje se može ostvariti bez eksplicitnog računanja hermitskih pozitivno definitnih matrica A^*A ili AA^*, a u drugom slučaju nije niti potrebno generitati aproksimacije za $\hat{x}$, već se direktno računaju aproksimacije za x. Ako upotrijebimo CG metodu za rješavanje bilo kojeg sustava u (2.130) tada odgovarajuće algoritme nazivamo CGNR za prvi, i CGNE za drugi sustav. Ako sa $\hat{}$ označimo sve veličine u normalnim jednadžbama, a sa standardnim oznakama označimo veličine vezane uz polazni sustav $Ax = b$, tada uz pomoć relacija

$$r_k = A^{-*}\hat{r}_k \qquad p_k = \hat{p}_k,$$

za CGNR algoritam, i

$$x_k = A^*\hat{x}_k, \qquad , r_k = \hat{r}_k, \qquad p_k = A^*\hat{p}_k,$$

za CGNE algoritam, algoritme možemo implementirati na sljedeći način

Algoritam 2.8.1. CG za normalne jednadžbe (CGNR i CGNE)

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

izračunaj A^*r_0,

$p_0 = A^*r_0$.

Za $k = 1, 2, \ldots$

izračunaj Ap_{k-1},

$\alpha_{k-1} = \dfrac{\langle A^*r_{k-1}, A^*r_{k-1}\rangle}{\langle Ap_{k-1}, Ap_{k-1}\rangle}$ *za CGNR,* $\alpha_{k-1} = \dfrac{\langle r_{k-1}, r_{k-1}\rangle}{\langle p_{k-1}, p_{k-1}\rangle}$ *za CGNE,*

$x_k = x_{k-1} + \alpha_{k-1}p_{k-1}$,

$r_k = r_{k-1} - \alpha_{k-1}Ap_{k-1}$,

izračunaj A^*r_k,

$\beta_k = \dfrac{\langle A^*r_k, A^*r_k\rangle}{\langle A^*r_{k-1}, A^*r_{k-1}\rangle}$ *za CGNR,* $\beta_k = \dfrac{\langle r_k, r_k\rangle}{\langle r_{k-1}, r_{k-1}\rangle}$ *za CGNE,*

$p_k = A^*r_k + \beta_k p_{k-1}$.

CGNR algoritam minimizira A^*A-normu greške, što je ekvivalentno minimiziranju euklidske norme reziduala $r_k = b - Ax_k$, po afinom potprostoru

$$x_k \in x_0 + \text{span}\{A^*r_0, (A^*A)A^*r_0, \ldots, (A^*A)^{k-1}A^*r_0\}.$$

CGNE algoritam minimizira AA^*-normu greške po $\hat{x}_k = A^{-*}x_k$, što je opet ekvivalentno minimiziranju euklidske norme greške $e_k = x - x_k$ po afinom potprostoru

$$x_k \in x_0 + \text{span}\{A^*r_0, A^*(AA^*)r_0, \ldots, A^*(AA^*)^{k-1}r_0\}.$$

Možemo primijetiti da se u obje metode radi o istom prostoru, ali o različitim normama koje se minimiziraju.

2.8.2 Analiza greške i konvergencija CGNR i CGNE metode

Najprije promatramo konvergenciju metode CGNR. Kao što je već napomenuto ova metoda minimizira euklidsku normu reziduala aproksimacija oblika

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

pri čemu je q_{k-1} polinom $(k-1)$-og stupnja. Prema tome, rezidual je tada oblika

$$r_k = r_0 - Aq_{k-1}(A^*A)A^*r_0 = (I - AA^*q_{k-1}(AA^*))r_0.$$

Neka je sada $A = U\Sigma V^*$ singularna dekompozicija od A, pri čemu su U i V unitarne matrice, a $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$ je dijagonalna matrica singularnih vrijednosti. Tada rezidual CGNR aproksimacije zadovoljava sljedeće

$$\begin{aligned}\|r_k\|_2 &= \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|p_k(AA^*)r_0\|_2 = \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|Up_k(\Sigma^2)U^*r_0\|_2 \leq \\ &\leq \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|p_k(\Sigma^2)\|_2\|r_0\|_2 = \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \max_{i=1,\ldots,n} |p_k(\sigma_i^2)|\|r_0\|_2.\end{aligned}$$

Na analogan način kao i kod CG metode, korištenjem Čebiševljevih polinoma, možemo doći do zaključka da je

$$\|r_k\|_2 \leq T_k\left(\frac{\sigma_{max}^2 + \sigma_{min}^2}{\sigma_{max}^2 - \sigma_{min}^2}\right)^{-1}\|r_0\|_2,$$

pri čemu su σ_{max} i σ_{min} najveća, odnosno najmanja singularna vrijednost. Raspisivanjem vrijednosti Čebiševljevog polinoma dobivamo

$$\|r_k\|_2 \leq 2\left[\left(\frac{\kappa(A)+1}{\kappa(A)-1}\right)^k + \left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^k\right]^{-1}\|r_0\|_2,$$

odnosno

$$\|r_k\|_2 \leq 2\left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^k\|r_0\|_2,$$

gdje je $\kappa(A) = \|A\|_2\|A^{-1}\|_2 = \sigma_{max}/\sigma_{min}$ uvjetovanost matrice A. Time smo pokazali da konvergencija CGNR metode ovisi o $\kappa(A)$, odnosno o singularnim vrijednostima matrice A, a ne o njenom spektru, što je karakteristično za CG metodu.

Sličan zaključak možemo izvesti i za CGNE metodu. Ona minimizira euklidsku normu greške aproksimacije oblika

$$x_k = x_0 + A^*s_{k-1}(AA^*)r_0 = x_0 + A^*s_{k-1}(AA^*)Ae_0,$$

pri čemu je s_{k-1} polinom $(k-1)$-og stupnja. Greška je tada oblika

$$e_k = e_0 - A^* s_{k-1}(AA^*)Ae_0 = (I - A^* A s_{k-1}(A^*A))e_0.$$

Greška CGNE metode tada zadovoljava

$$\begin{aligned} \|e_k\|_2 &= \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|p_k(A^*A)e_0\|_2 = \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|V p_k(\Sigma^2)V^* e_0\|_2 \leq \\ &\leq \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \|p_k(\Sigma^2)\|_2 \|e_0\|_2 = \min_{p_k\in\mathbb{P}_k, p_k(0)=1} \max_{i=1,\ldots,n} |p_k(\sigma_i^2)| \|e_0\|_2. \end{aligned}$$

Istom analizom dolazimo do rezultata

$$\|e_k\|_2 \leq 2\left[\left(\frac{\kappa(A)+1}{\kappa(A)-1}\right)^k + \left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^k\right]^{-1} \|e_0\|_2,$$

odnosno

$$\|e_k\|_2 \leq 2\left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^k \|e_0\|_2.$$

Još treba samo napomenuti, da prema analizi CG metode, primijenjenoj na normalne jednadžbe, za CGNR metodu euklidske norme reziduala čine nerastući niz, a kod CGNE to vrijedi za euklidske norme greške.

2.8.3 Prekondicionirane CGNR i CGNE metode

Kako konvergencija CGNR i CGNE metode ovisi o singularnim vrijednostima, odnosno uvjetovanosti matrice A, ove metode možemo primijeniti i na prekondicioniranim normalnim jednadžbama, s namjerom da prekondicionirana matrica sustava bude volje uvjetovana, i da iteracije imaju bolju konvergenciju. Promatrat ćemo posebno svaku metodu.

Kod CGNR metode, ona se primijenjuje na prekondicionirane normalne jednadžbe

$$M^{-*}A^*AM^{-1}\hat{x} = M^{-*}A^*b, \qquad x = M^{-1}\hat{x}. \tag{2.131}$$

sa dvostranim prekondicioniranajem, i matricom prekondicioniranja M^*M. Ako sa $\hat{}$ označimo sve veličine vezane uz sustav (2.131), a sa standardnim oznakama označimo veličine vezane uz polazne normalne jednadžbe u CGNR algoritmu, tada su one vezane relacijama

$$r_k = A^{-*}M^*\hat{r}_k, \qquad p_k = M^{-1}\hat{p}_k,$$

pa uz adekvatne transformacije prekondicionirani CGNR algoritam izgleda ovako

Algoritam 2.8.2. PREKONDICIONIRANI CGNR

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

izračunaj A^*r_0,

riješi $M^*q_0 = A^*r_0$,

riješi $Ms_0 = q_0$,

$p_0 = s_0$.

Za $k = 1, 2, \ldots$

izračunaj Ap_{k-1},

$$\alpha_{k-1} = \frac{\langle q_{k-1}, q_{k-1}\rangle}{\langle Ap_{k-1}, Ap_{k-1}\rangle},$$

$x_k = x_{k-1} + \alpha_{k-1}p_{k-1}$,

$r_k = r_{k-1} - \alpha_{k-1}Ap_{k-1}$,

izračunaj A^*r_k,

riješi $M^*q_k = A^*r_k$,

riješi $Ms_k = q_k$,

$$\beta_k = \frac{\langle q_k, q_k\rangle}{\langle q_{k-1}, q_{k-1}\rangle}$$

$p_k = s_k + \beta_k p_{k-1}$.

Kod CGNE metode, ona se primijenjuje na prekondicionirane normalne jednadžbe

$$M^{-1}AA^*M^{-*}\hat{x} = M^{-1}b, \qquad x = A^*M^{-*}\hat{x}. \tag{2.132}$$

sa dvostranim prekondicioniranajem, i matricom prekondicioniranja MM^*. Ako ponovo sa$\hat{}$označimo sve veličine vezane uz sustav (2.132), a sa standardnim oznakama označimo veličine vezane uz polazne normalne jednadžbe u CGNE algoritmu, tada su one vezane relacijama

$$r_k = M\hat{r}_k, \qquad p_k = A^*M^{-*}\hat{p}_k,$$

pa uz adekvatne transformacije prekondicionirani CGNE algoritam izgleda ovako

Algoritam 2.8.3. PREKONDICIONIRANI CGNE

Dana je početna iteracija x_0,

$r_0 = b - Ax_0$,

riješi $Mq_0 = r_0$,

riješi $M^*s_0 = q_0$,

izračunaj A^*s_0,

$p_0 = A^*s_0$.

Za $k = 1, 2, \ldots$

izračunaj Ap_{k-1},

$$\alpha_{k-1} = \frac{\langle q_{k-1}, q_{k-1}\rangle}{\langle p_{k-1}, p_{k-1}\rangle},$$

$x_k = x_{k-1} + \alpha_{k-1}p_{k-1}$,

$r_k = r_{k-1} - \alpha_{k-1}Ap_{k-1}$,

riješi $Mq_k = r_k$,

riješi $M^*s_k = q_k$,

izračunaj A^*s_k,

$$\beta_k = \frac{\langle q_k, q_k\rangle}{\langle q_{k-1}, q_{k-1}\rangle}$$

$p_k = A^*s_k + \beta_k p_{k-1}$.

2.9 Primjena iterativnih metoda

2.9.1 Izbor metode

Nakon prikaza raznih iterativnih metoda za rješavanje linearnih sustava, postavlja se pitanje koju metodu primijeniti za konkretan sustav. Prilikom odabira metode, potrebno je razmotriti dva osnovna kriterija, a to su:

- Ukoliko matrica sustava A ima neko posebno svojstvo (hermitičnost, pozitivnu definitnost, ...) izbor metode se sužava na one metode koje su specijalizirane za rješavanje sustava sa upravo takvim svojstvom.
- Među pogodnim metodama bira se ona metoda kojoj je potrebno izvršiti najmanji broj operacija za postizanje tražene točnosti i koja ima što manje zahtjeva za prostorom memorije.

Potrebno je napomenuti da uglavnom ne postoji najbolja metoda za sve sustave općenito, tako da su za neke klase matrica pogodne jedne metode, a za druge klase, druge metode. Isto tako, uspješnost metode ovisi i o dobrom izboru matrice prekondicioniranja, pa se uvijek preporuča rješavati prekondicionirani sustav.

Pogledajmo najčešće slučajeve. Za hermitske sustave izbor iterativne metode je vrlo jednostavan. Za pozitivno definitne probleme treba upotrijebiti CG ili MINRES, a za indefinitne samo MINRES. Odnos između normi reziduala CG i MINRES metoda je jednak kao i odnos između norme reziduala BCG metode i kvazireziduala QMR metode. Ova tvrdnja se lako pokaže iz Leme 2.7.6, jer se norma kvazireziduala poklapa sa normom reziduala MINRES metode. Zbog tog svojstva je krivulja konvergencije MINRES metode, određena normama reziduala, doljnja ograda krivulje konvergencije CG metode, pa se zato češće preferira MINRES.

Izbor metode za nehermitske probleme nije tako lagan. Ako je produkt matrice i vektora jako skup, na primjer ako je matrica A gusto popunjena i nema specijalnih svojstava koje bi omogućile brzo računanje tog produkta, tada bi najčešći izbor bila GMRES metoda, zato što zahtijeva najmanje množenja matrice i vektora za reduciranje norme reziduala na zadanu veličinu. Ako računanje produkta matrice i vektora nije tako skupo, ili ako zahtjev za memorijom kod GMRES metode postane prevelik, tada bi dobar izbor bila jedna od metoda BCG, QMR, CGS ili BICGSTAB. Zbog (2.122), obično se preporuča QMR metoda kao bolja od BCG. Izbor između QMR, CGS, i BICGSTAB ovisi o problemu. Za svaku od ovih metoda može se naći primjer za koji je ona najbolja, i s druge strane može se naći primjer za kojeg je ona najgora. Zbog toga se ne može reći niti za jednu metodu da je najbolji izbor za nehermitske probleme. Drugi pristup je simetrizacija problema i rješavanje metodama CGNR i CGNE. Kao što smo vidjeli njihova konvergencija ovisi o distribuciji singularnih vrijednosti. Zato postoje problemi u kojima su CGNR i CGNE metode najbolji izbor, ali postoje problemi u kojima jedna od ostalih metoda za rješavanje nehermitskih sustava daleko nadmašuje ove dvije metode, vidi primjere u [29]. U praksi prevladava ova druga situacija.

Najbolji pregled metoda dan je u [1], a možemo ga dobiti ako sažmemo sve najbitnije karakteristike pojedinih metoda za rješavanje sustava $Ax = b$, i smjestimo ih u jednu listu.

1. **Minimalni reziduali (MINRES)**
   - Primjenljiva na hermitskim matricama.
   - Odabir aproksimacije u k-toj iteraciji je takav da minimizira euklidsku normu reziduala na Krylovljevom potprostoru $r_0 + A\mathcal{K}_k(A, r_0)$.
   - Brzina konvergencije ovisi o korijenu uvjetovanosti matrice sustava.
2. **Konjugirani gradijenti (CG)**
   - Primjenljiva na pozitvno definitnim hermitskim matricama.
   - Odabir aproksimacije u k-toj iteraciji je takav da minimizira A-normu greške na Krylovljevom potprostoru $e_0 + A\mathcal{K}_k(A, e_0)$.
   - Brzina konvergencije ovisi o korijenu uvjetovanosti matrice sustava.
3. **Generalizirani minimalni reziduali (GMRES)**
   - Primjenljiva na nehermitskim matricama.
   - Odabir aproksimacije u k-toj iteraciji je takav da minimizira euklidsku normu reziduala na Krylovljevom potprostoru $r_0 + A\mathcal{K}_k(A, r_0)$.
   - Rastući zahtjev za prostorom u memoriji, zato je potrebno restartanje.

- Brzina konvergencije kod ne-normalnog slučaja ne ovisi o spektru, ali je vezana uz pseudospektar. Ona se općenito ne može ograničiti samo na svojstva matrice, već na nju utjeće i desna strana sustava, kao i izbor početne iteracije.

4. **Bikonjugirani gradijenti (BCG)**

   - Primjenljiva na nehermitskim matricama.
   - Odabir aproksimacije u k-toj iteraciji je takav da zadovoljava uvjet biortogonalnosti; rezidual je iz Krylovljevog potprostora $r_0 + A\mathcal{K}_k(A, r_0)$ i ortogonalan je na Krylovljev potprostor dualnog reziduala $\mathcal{K}_k(A^*, \hat{r}_0)$, a dualni rezidual je ortogonalan na Krylovljev potprostor $\mathcal{K}_k(A, r_0)$.
   - Osim produkta vektora sa matricom sustava, zahtijeva se produkt i sa transponiranom matricom sustava.
   - Može doći do ozbiljnog zakazivanja metode.
   - Krivulja konvergencije može biti jako oscilirajuća, ali brzina konvergencije se poklapa sa brzinom QMR metode, što više završit će u koraku u kojem je završio i QMR.

5. **Kvazi–minimalni reziduali (QMR)**

   - Primjenljiva na nehermitskim matricama.
   - Odabir aproksimacije u k-toj iteraciji je takav da minimizira euklidsku normu kvazi–reziduala.
   - Metoda je dizajnirana da izbjegne oscilirajuću konvergenciju BCG metode, i izbjegava jednu od dvije situacije mogućeg sloma BCG-a.
   - Ako BCG metoda u jednoj iteraciji ostvari značajan napredak u konvergenciji, tada QMR ostvaruje od prilike isti rezultat u tom istom koraku. Ako BCG stagnira ili divergira, QMR i dalje može reducirati rezidual ali napredak je vrlo polagan.
   - Brzina konvergencije ovisi o uvjetovanosti matrice Lanczosovih vektora.

6. **Kvadrirani konjugirani gradijenti (CGS)**

   - Primjenljiva na nehermitskim matricama.
   - Odabir aproksimacije u k-toj iteraciji je takav da rezidual bude iz Krylovljevog potprostora $r_0 + A\mathcal{K}_{2k}(A, r_0)$, i da je oblika $r_k = p_k^2(A)r_0$, pri čemu je $p_k(A)r_0$ rezidual k-tog koraka BCG metode.
   - Metoda je dizajnirana da eliminira računanje produkta vektora sa transponiranom matricom sustava.
   - Obično konvergira (ili divergira) dvostruko brže od BCG metode. Krivulja konvergencije ima još jače oscilacije od BCG metode, što može izazvati gubitak točnosti izračunatog reziduala.

7. **Stabilizirani bikonjugirani gradijenti (BICGSTAB)**

   - Primjenljiva na nehermitskim matricama.

- Odabir aproksimacije u k-toj iteraciji je takav da rezidual bude iz Krylovljevog potprostora $r_0 + A\mathcal{K}_{2k}(A, r_0)$, i da je oblika $r_k = \chi_k(A)p_k(A)r_0$, pri čemu je $p_k(A)r_0$ rezidual k-tog koraka BCG metode, a $\chi_k(z) = (1 - \omega_k z)\cdots(1 - \omega_1 z)$. r_k ima najmanju euklidsku normu od vektora oblika $(I - \omega A)\chi_{k-1}(A)\cdot$ $\cdot p_k(A)r_0$.
- Metoda ne zahtijeva računanje produkta vektora sa transponiranom matricom sustava, i dizajnirana je tako da izbjegne oscilacije krivulje konvergencije CGS metode.
- Brzina konvergencije je otprilike ista kao i kod CGS metode, ali dolazi do manje gubitaka u točnosti izračunatog reziduala nego kod CGS-a.

8. **Metode za rješavanje normalnih jednadžbi (CGNR i CGNE)**

- Primjenljiva na nehermitskim matricama.
- Odabir aproksimacije u k-toj iteraciji je takav da za CGNR metodu minimizira euklidsku normu reziduala na Krylovljevom potprostoru $r_0 + AA^*\mathcal{K}_k(AA^*, r_0)$, a za CGNE metodu minimizira euklidsku normu greške na Krylovljevom potprostoru $e_0 + A^*A\mathcal{K}_k(A^*A, e_0)$.
- Brzina konvergencije ovisi o uvjetovanosti matrice sustava.

Jedan od važnih kriterija za odabir metode je broj operacija sa vektorima i matricama po iteraciji, budući da one daju daleko značajniji doprinos složenosti algoritma od skalarnih operacija. Drugi takav kriterij bi bio potrošnja memorije po iteraciji. Takav pregled vidljiv je u sljedećim tabelama.

Tablica 2.1 prikazuje broj operacija potrebnih za izvršavanje k-te iteraciji pojedine me-

Metode	$\langle v, w\rangle$	αv	$v + w$	Av	A^*v	$v = M^{-1}w$	$v = M^{-*}w$
MINRES	2	7	5	1	0	1	0
CG	2	3	3	1	0	1	0
GMRES	$k+1$	$k+1$	k	1	0	1	0
BCG	2	5	5	1	1	1	1
QMR	3	10	7	1	1	1	1
CGS	2	6	7	2	0	2	0
BICGSTAB	4	6	6	2	0	2	0
CGNR/CGNE	2	3	3	1	1	1	1

Tablica 2.1: Broj operacija po iteraciji pojedinih iterativnih metoda

tode, pri čemu je α skalar, v, w su n-dimenzionalni vektori, A je $n \times n$ matrica sustava, i M je $n \times n$ matrica prekondicioniranja. Zadnja dva stupca odnose se na rješavanja sustava sa matricom prekondicioniranja ili njenom transponiranom matricom.

U Tablici 2.2 prikazana je potrošnja memorije pojedine metode, ponovo u k-toj iteraciji. n je dimenzija sustava.

2.9.2 Odabir početne iteracije

Kada odaberemo iterativnu metodu kojem ćemo riješiti zadani linearni sustava, sljedeće pitanje koje se postavlja je kako započeti iteriranje. Mnogi linearni sustavi dolaze kao

Metoda	Potrošnja memorije
MINRES	matrica+$5n$
CG	matrica+$5n$
GMRES	matrica+$(k+3)n$
BCG	matrica+$9n$
QMR	matrica+$11n$
CGS	matrica+$10n$
BICGSTAB	matrica+$7n$
CGNR/CGNE	matrica+$6n$

Tablica 2.2: Potrošnja memorije po iteraciji pojedinih iterativnih metoda

konkretni problemi iz mnogih egzaktnih znanosti. Zato iz samih svojstava tih problema, kao i procedura koje su početni problem svele na linearni sustav, često možemo naslutiti grubu aproksimaciju traženog rješenja, koja se vrlo dobro može iskotistiti kao početna iteracija za iterativnu metodu. Ukoliko nemamo takvih informacija, onda se vrlo često uzima da je početna iteracija $x_0 = 0$. U tom slučaju iterativne metode koje se koriste aproksimacijama iz Krylovljevih potprostora, generirat će aproksimacije iz vrlo poželjnog potprostora

$$x_k \in \mathcal{K}_k(A, b) = \text{span}\{b, Ab, \ldots, A^{k-1}b\},$$

u kojem se, sa dovoljno velikom dimenzijom $m \leq n$, nalazi rješenje.

U svakom slučaju konvergencija metode, naročito kod nehermitskih sustava, ovisi o početnoj iteraciji, pa se zato njenom povoljnom izboru treba posvetiti malo pažnje. Kao što ćemo kasnije pokazati, za točnost konačne aproksimacije rješenja, kod većine metoda, važno je izabrati početnu aproksimaciju sa ne prevelikom normom.

2.10 Kriterij zaustavljanja i točnost iterativnih metoda

2.10.1 Kriterij zaustavljanja

Kada koristimo iterativne metode, navedene u ovom poglavlju, možemo očekivati da će one doći do traženog rješenja za manje od n koraka, pri čemu je n dimenzija sustava, ali samo ako računamo u egzaktnoj aritmetici. Budući da se iterativne metode primjenjuju uglavnom na računalima, egzaktna aritmetika nam neće biti dostupna, pa nećemo biti u stanju dobiti točno rješenje. Zato, moramo dati neki kriterij, po kojemu ćemo aproksimaciju u nekom koraku smatrati dovoljno dobrom, što znači da bi, u slučaju da je taj kriterij ispunjen, danu aproksimaciju mogli smatrati dovoljno točnom. Iteriranje se tada može zaustaviti u tom koraku. Jedino što nam je dostupno, a što daje neke naznake o odstupanja aproksimacije x_k u k-tom koraku od rješenja $x = A^{-1}b$, je rezidual $r_k = b - Ax_k$, kojeg možemo dobiti u svakoj iteraciji spomenutih iterativnih metoda direktno ili preko rekurzije. Iako bi najindikativnija mjera tog odstupanja bila neka standardna norma greške $e_k = x - x_k$, najčešće euklidska norma ili ∞-norma, to nam nije dostupno, jer da znamo vektor greške znali bismo i točno rješenje. Zato se, uglavnom, kao kriterij zaustavljanja provjerava da li je norma reziduala pala ispod nekog praga tolerancije. Sada se postavlja pitanje kakav je odnos između norme reziduala

i norme greške, odnosno ako je norma reziduala mala da li to isto vrijedi i za normu greške? Potvrdan odgovor bi dakako bio vrlo poželjan, ali to nije uvijek slučaj.

Norma relativne greške i norma relativnog reziduala su povezane sljedećom nejednakošću

$$\frac{1}{\kappa(A)}\frac{\|b-Ax_k\|}{\|b\|} \leq \frac{\|A^{-1}b-x_k\|}{\|A^{-1}b\|} \leq \kappa(A)\frac{\|b-Ax_k\|}{\|b\|}, \tag{2.133}$$

gdje je $\kappa(A) = \|A\| \cdot \|A^{-1}\|$ uvjetovanost matrice A, a $\|\cdot\|$ je bilo koja vektorska norma i njezina inducirana matrična norma. Da bi pokazali da (2.133) vrijedi primijetimo da, budući da je $b - Ax_k = A(A^{-1}b - x_k)$, $\|A^{-1}b\| \leq \|A^{-1}\| \cdot \|b\|$ i $\|b\| = \|AA^{-1}b\| \leq \|A\| \cdot \|A^{-1}b\|$, imamo

$$\begin{aligned}\frac{\|b-Ax_k\|}{\|b\|} &\leq \frac{\|A\| \cdot \|A^{-1}b-x_k\|}{\|b\|} = \frac{\|A\| \cdot \|A^{-1}\| \cdot \|A^{-1}b-x_k\|}{\|A^{-1}\| \cdot \|b\|} \leq \\ &\leq \frac{\kappa(A)\|A^{-1}b-x_k\|}{\|A^{-1}b\|},\end{aligned}$$

$$\begin{aligned}\frac{\|A^{-1}b-x_k\|}{\|A^{-1}b\|} &\leq \frac{\|A^{-1}\| \cdot \|b-Ax_k\|}{\|A^{-1}b\|} = \frac{\|A\| \cdot \|A^{-1}\| \cdot \|b-Ax_k\|}{\|A\| \cdot \|A^{-1}b\|} \leq \\ &\leq \frac{\kappa(A)\|b-Ax_k\|}{\|b\|}.\end{aligned}$$

Da bismo ostvarili gornju i donju ogradu norme greške potrebno je moći ocijeniti uvjetovanost matrice A. U svakom slučaju kada je norma relativnog reziduala mala, ako je matrica dobro uvjetovana, to jest kada je $\kappa(A)$ malo, tada će i norma relativne greške biti mala. U suprotnoj situaciji, kada je matrica loše uvjetovana, raspon između donje i gornje ograde je velik, pa ocjena za normu relativne greške ne mora biti stroga, ali pokazuje da norma greške može biti i vrlo velika. Zato kod loše uvjetovanih matrica mala norma relativnog reziduala ne mora značiti da je aproksimacija zaista i blizu rješenja.

2.10.2 Točnost iterativnih metoda

Točnost koju zahtijevamo od iterativne metode ne bi smjela biti veća od točnosti ulaznih parametara, koja često nije jako velika, zbog toga što su i sama matrica A i vektor desne strane b rezultat nekih prethodnih transformacija, pa u svakom slučaju sadrže neku grešku. Zato zahtijevana točnost metoda, obično isto tako nije jako velika, štoviše manja je od točnosti koja bi se mogla postići. Ponekad se iterativne metode primjenju na vrlo loše uvjetovanim problemima, i tada je važno znati koliko mašinska preciznost i broj uvjetovanosti ograničavaju točnost metode. Kod svih navedenih iterativnih metoda nigdje se rezidual ne računa direktno, već kao rekurzija kod metoda baziranih na CG metodi, ili se njegova norma računa preko određenih relacija kao kod GMRES. U egzaktnoj aritmetici te dvije stvari su ekvivalentne, međutim u aritmetici konačne preciznosti to nije tako. Dakle, nakon što smo na neki način odredili kolika nam treba biti norma pravog reziduala aproksimacije, važno nam je znati koliko norma izračunatog reziduala odstupa od nje. Takva analiza će nam dati potpuniju sliku dostignute točnosti. Traba još samo napomenuti da se u analizi točnosti algoritma promatraju ralni sustavi.

Iterativne metode sa rekurzijom za računanje reziduala

Najprije ćemo promotriti skupinu iterativnih metoda u kojima se rezidual r_k računa pomoću rekurzije. Većina rezultata u ovom odjeljku su prezentirani u [11]. Rekurzije koje ćemo koristiti su zadane preko formula

$$x_k = x_{k-1} + \alpha_{k-1} p_{k-1}, \qquad r_k = r_{k-1} - \alpha_{k-1} A p_{k-1}. \tag{2.134}$$

Ovdje je p_{k-1} neki vektor smjera, a α_{k-1} koeficijent, oboje određeni nekom konkretnom metodom iz te skupine. Vekotor r_k se računa prema rekurziji umjesto da se računa direktno kao $b - Ax_k$. Ako promatramo prekondicionirane sustave, tada se u takvim algoritmima računa z_k kao rješenje sustava $Mz_k = r_k$ koji se puno lakše rješava od sustava sa matricom A. Vektor z_k se tada koristi za određivanje novog koeficijenta i vektora smjera, ali on ne mijenja oblik formula (2.134). Upravo te formule ćemo analizirati u aritmetici konačne preciznosti.

Još jedna napomena prije same analize. Veličina $\|b - Ax_k\|_2/\|b\|_2$, čija vrijednost se obično prati, može biti puno veća od $\|b - Ax_k\|_2/(\|A\|_2\|x\|_2)$ za $Ax = b$ i $\|b\|_2 \ll \|A\|_2\|x\|_2$. Rezultati koji će biti prezentirani koriste samo ovaj drugi izraz, kojeg označavamo kao norma relativnog reziduala. Vrijedi sljedeća relacija

$$\frac{\|b - Ax_k\|_2}{\|A\|_2\|x\|_2} \leq \frac{\|b\|_2 + \|A\|_2\|x_k\|_2}{\|A\|_2\|x\|_2} \leq 1 + \frac{\|x_k\|_2}{\|x\|_2},$$

odakle se nazire važnost veličine $\max_k \frac{\|x_k\|_2}{\|x\|_2}$ u daljnjoj analizi.

Implementacija formula (2.134) u aritmetici konačne preciznosti

Pretpostavit ćemo sljedeći model aritmetike pomičnog zareza na računalu sa mašinskom preciznošću ϵ:

$$\mathrm{fl}(a \pm b) = a(1 + \epsilon_1) \pm b(1 + \epsilon_2), \quad |\epsilon_1|, |\epsilon_2| \leq \epsilon, \tag{2.135}$$

$$\mathrm{fl}(a \text{ op } b) = (a \text{ op } b)(1 + \epsilon_3), \quad |\epsilon_3| \leq \epsilon, \quad \text{op} = *, /. \tag{2.136}$$

Ovaj model vrijedi i za ona računala koja ne koriste sigurnosnu znamenku kod zbrajanja i oduzimanja.

Sa ovakvim modelom, vrijede sljedeći rezultati za operacije sa n-dimenzionalnim vektorima v i w, $n \times n$ matricom A, i skalarom α, koji su dokazani u [21].

$$\|\alpha v - \mathrm{fl}(\alpha v)\|_2 \leq \epsilon \|\alpha v\|_2, \tag{2.137}$$

$$\|v + w - \mathrm{fl}(v + w)\|_2 \leq \epsilon(\|v\|_2 + \|w\|_2), \tag{2.138}$$

$$|\langle v, w\rangle - \mathrm{fl}(\langle v, w\rangle)| \leq n(\epsilon + \mathcal{O}(\epsilon^2))\|v\|_2\|w\|_2, \tag{2.139}$$

$$\|Av - \mathrm{fl}(Av)\|_2 \leq c(\epsilon + \mathcal{O}(\epsilon^2))\|A\|_2\|v\|_2, \tag{2.140}$$

pri čemu konstanta c ovisi o rutini koja računa produkt matrice i vektora. Rezultat za takav produkt koji se računa na standardni način, gdje je A $n \times n$ matrica sa najviše m netrivijalnih elemenata po retku je

$$|Av - \mathrm{fl}(Av)| \leq m(\epsilon + \mathcal{O}(\epsilon^2))|A| \cdot |v|,$$

gdje je za matricu $A = (a_{ij})$ $|A| = (|a_{ij}|)$, a za vektor $v = (v_i)$ $|v| = (|v_i|)$. Tada vrijedi

$$\begin{aligned}\|Av - \mathrm{fl}(Av)\|_2 &\leq m(\epsilon + \mathcal{O}(\epsilon^2))\||A|\|_2\|v\|_2 \leq m(\epsilon + \mathcal{O}(\epsilon^2))\|A\|_F\|v\|_2 \leq \\ &\leq \sqrt{n}m(\epsilon + \mathcal{O}(\epsilon^2))\|A\|_2\|v\|_2.\end{aligned}$$

U ovom slučaju je tada $c = m\sqrt{n}$. U općenitom slučaju to ne vrijedi jer ponekad matrice nisu eksplicitno spremljene u memoriji, već postoje samo rutine koje računaju produkt matrice s vektorom, pa se za takve rutine treba napraviti posebna analiza.

Uz pomoć upravo navedenih pravila izvest ćemo analizu implementacije formula (2.134) u aritmetici konačne preciznosti. Sa x_k, r_k, p_{k-1} i α_{k-1} označit ćemo veličine koje su izračunate u konačnoj artimetici. Zbog jednostavnosti, izraze koji uključuju ϵ^2 ili više potencije od ϵ označavat ćemo sa $\mathcal{O}(\epsilon^2)$, budući da su takvi izrazi vrlo mali, i ne utječu na ocjenu. Kada su formule (2.134) implementirane u aritmetici konačne preciznosti, tada izračunate iteracije zadovoljavaju

$$x_k = x_{k-1} + \alpha_{k-1}p_{k-1} + \zeta_k, \tag{2.141}$$

$$r_k = r_{k-1} - \alpha_{k-1}Ap_{k-1} + \eta_k, \tag{2.142}$$

gdje je

$$\|\zeta_k\|_2 \leq \epsilon\|x_{k-1}\|_2 + (2\epsilon + \epsilon^2)\|\alpha_{k-1}p_{k-1}\|_2 \tag{2.143}$$

i

$$\|\eta_k\|_2 \leq \epsilon\|r_{k-1}\|_2 + (2\epsilon+\epsilon^2)\|\alpha_{k-1}Ap_{k-1}\|_2 + (1+\epsilon)^2\|\alpha_{k-1}(\mathrm{fl}(Ap_{k-1}) - Ap_{k-1})\|_2. \tag{2.144}$$

Nejednakosti (2.143) i (2.144) su direktna primjena ocjena (2.137) i (2.138). Za zadnji izraz u (2.144) znamo da vrijedi

$$\|\mathrm{fl}(Ap_{k-1}) - Ap_{k-1}\|_2 \leq c(\epsilon + \mathcal{O}(\epsilon^2))\|A\|_2\|p_{k-1}\|_2.$$

U nastavku promatramo pravi rezidual izračunate aproksimacije x_k. Ako jednakost (2.141) pomnožimo sa A, i oduzmemo od b, dobivamo rekurziju za pravi rezidual $b - Ax_k$. Ako sada od toga oduzmemo još i rekurziju (2.142) za r_k, imamo

$$\begin{aligned}b - Ax_k - r_k &= (b - Ax_{k-1} - r_{k-1}) - A\zeta_k - \eta_k = \\ &= (b - Ax_0 - r_0) - \sum_{j=1}^{k}(A\zeta_j + \eta_j).\end{aligned}$$

Nadalje, uzmimo norme na obje strane, i podijelimo sa $\|A\|_2\|x\|_2$ za $x = A^{-1}b$, čime dobivamo

$$\frac{\|b - Ax_k - r_k\|_2}{\|A\|_2\|x\|_2} \leq \frac{\|b - Ax_0 - r_0\|_2}{\|A\|_2\|x\|_2} + \sum_{j=1}^{k}\left(\frac{\|\zeta_j\|_2}{\|x\|_2} + \frac{\|\eta_j\|_2}{\|A\|_2\|x\|_2}\right). \tag{2.145}$$

Vektor r_0 se jedini računa direktno, tako da se prvi izraz na lijevoj strani (2.145) lako može ocijeniti pomoću (2.138) i (2.140), što rezultira sa

$$\|b - Ax_0 - r_0\|_2 \leq \epsilon((1 + c)\|A\|_2\|x_0\|_2 + \|b\|) + \mathcal{O}(\epsilon^2)\|A\|_2\|x_0\|_2,$$

a, kako vrijedi $\|b\|_2 \leq \|A\|_2\|x\|_2$, možemo napisati

$$\frac{\|b - Ax_0 - r_0\|_2}{\|A\|_2\|x\|_2} \leq \epsilon(1 + c)\frac{\|x_0\|_2}{\|x\|_2} + \epsilon + \mathcal{O}(\epsilon^2)\frac{\|x_0\|_2}{\|x\|_2}. \tag{2.146}$$

sljedeća lema daje ocjene za ostale izraze u (2.145).

Lema 2.10.1 ([11]). *Definirajmo*

$$\Theta_k = \max_{j\le k} \frac{\|x_j\|_2}{\|x\|_2}. \tag{2.147}$$

Pretpostavimo da je $1-2\epsilon-\epsilon^2 > 0$*, što je razumljiva pretpostavka. Tada izrazi na desnoj strani u (2.145) zadovoljavaju*

$$\sum_{j=1}^{k} \frac{\|\zeta_j\|_2}{\|x\|_2} \le (5\epsilon + \mathcal{O}(\epsilon^2))k\Theta_k, \tag{2.148}$$

$$\sum_{j=1}^{k} \frac{\|\eta_j\|_2}{\|A\|_2\|x\|_2} \le (\epsilon + \mathcal{O}(\epsilon^2))k(1 + (5+2c)\Theta_k). \tag{2.149}$$

Dokaz: Prema (2.141) možemo pisati

$$\alpha_{j-1}p_{j-1} = x_j - x_{j-1} - \zeta_j, \tag{2.150}$$

pa uvrštavanjem ovog izraza u (2.143) dobivamo

$$\begin{aligned} \|\zeta_j\|_2 &\le \epsilon\|x_{j-1}\|_2 + (2\epsilon+\epsilon^2)(\|x_j\|_2 + \|x_{j-1}\|_2 + \|\zeta_j\|_2) = \\ &\le 3\epsilon\|x_{j-1}\|_2 + 2\epsilon\|x_j\|_2 + \epsilon^2(\|x_{j-1}\|_2 + \|x_j\|_2) + (2\epsilon+\epsilon^2)\|\zeta_j\|_2. \end{aligned}$$

Koristeći pretpostavku $1-2\epsilon-\epsilon^2 > 0$, ovo se može napisati u obliku

$$\begin{aligned} \|\zeta_j\|_2 &\le \frac{3\epsilon\|x_{j-1}\|_2 + 2\epsilon\|x_j\|_2 + \epsilon^2(\|x_{j-1}\|_2 + \|x_j\|_2)}{1-2\epsilon-\epsilon^2} \le \\ &\le \epsilon(3\|x_{j-1}\|_2 + 2\|x_j\|_2) + \mathcal{O}(\epsilon^2)(\|x_{j-1}\|_2 + \|x_j\|_2) \le \\ &\le \epsilon\cdot 5\Theta_k\|x\|_2 + \mathcal{O}(\epsilon^2)\Theta_k\|x\|_2. \end{aligned} \tag{2.151}$$

Odavde slijedi (2.148), sumiranjem ovog izraza k puta.

Nadalje, iz (2.140) treći izraz u (2.144) može se ocijeniti sa

$$(1+\epsilon)^2\|\alpha_{j-1}(\mathrm{fl}(Ap_{k-1}) - Ap_{k-1})\|_2 \le c(\epsilon + \mathcal{O}(\epsilon^2))(1+\epsilon)^2\|A\|_2\|\alpha_{j-1}p_{j-1}\|_2,$$

i raspisivanjem $\alpha_{j-1}p_{j-1}$ kao u (2.150) i korištenjem ocjene (2.151) za $\|\zeta_j\|_2$, on dobiva konačan oblik

$$\begin{aligned} &(1+\epsilon)^2\|\alpha_{j-1}(\mathrm{fl}(Ap_{k-1}) - Ap_{k-1})\|_2 \le \\ &\le c(\epsilon+\mathcal{O}(\epsilon^2))(1+\epsilon)^2\|A\|_2(\|x_{j-1}\|_2 + \|x_j\|_2 + \\ &\quad +\epsilon(3\|x_{j-1}\|_2 + 2\|x_j\|_2) + \mathcal{O}(\epsilon^2)(\|x_{j-1}\|_2 + \|x_j\|_2)) = \\ &\le c(\epsilon+\mathcal{O}(\epsilon^2))\|A\|_2(\|x_{j-1}\|_2 + \|x_j\|_2). \end{aligned} \tag{2.152}$$

Iz (2.150) slijedi

$$\alpha_{j-1}Ap_{j-1} = A(x_j - x_{j-1} - \zeta_j),$$

pa umetanjem ove jednakosti u drugi izraz u (2.144), uz korištenje ocjene (2.151) za $\|\zeta_j\|_2$, imamo

$$\begin{aligned} &(2\epsilon+\epsilon^2)\|\alpha_{j-1}Ap_{j-1}\|_2 \le \\ &\le (2\epsilon+\epsilon^2)\|A\|_2(\|x_{j-1}\|_2 + \|x_j\|_2 + \epsilon(3\|x_{j-1}\|_2 + 2\|x_j\|_2) + \\ &\quad +\mathcal{O}(\epsilon^2)(\|x_{j-1}\|_2 + \|x_j\|_2)) = \\ &\le (2\epsilon+\mathcal{O}(\epsilon^2))\|A\|_2(\|x_{j-1}\|_2 + \|x_j\|_2). \end{aligned} \tag{2.153}$$

Na poslijetku, kod dokaza indukcijom, pretpostavimo da je svaki izraz $\|\eta_i\|_2$, $i = 1,\ldots,j-1$ ocijenjen sa $\mathcal{O}(\epsilon)\|A\|_2(\|x\|_2 + \max_{l\le i}\|x_l\|_2)$. Iz (2.144), (2.152),(2.153), te iz ocjene za $\|b - Ax_0 - r_0\|_2$ jasno se vidi da ta ocjena vrijedi za η_1, jer

$$\begin{aligned}\|\eta_1\|2 &\le \epsilon\|b - Ax_0 - r_0\|_2 + \epsilon\|b\|_2 + \epsilon\|A\|_2\|x\|_2 + (2\epsilon+\epsilon^2)\|\alpha_0 Ap_0\|_2 + \\ &\quad +(1+\epsilon)^2\|\alpha_0(\mathrm{fl}(Ap_0) - Ap_0)\|_2 \le \\ &\le \epsilon\|A\|_2\|x\|_2 + \epsilon\|A\|_2(5+2c)\max_{l\le 1}\|x_l\|_2 + \mathcal{O}(\epsilon^2)\|A\|_2(\|x\|_2 + \max_{l\le 1}\|x_l\|_2).\end{aligned}$$

Budući da r_{j-1} zadovoljava

$$r_{j-1} = b - Ax_{j-1} - (b - Ax_0 - r_0) + \sum_{i=1}^{j-1}(A\zeta_i + \eta_i),$$

koristeći (2.146),(2.148), kojeg smo već dokazali, i pretpostavku indukcije imamo

$$\begin{aligned}\|r_{j-1}\|_2 &\le \|A\|_2\|x - x_{j-1}\|_2 + \epsilon(1+c)\|A\|_2(\max_{i\le j-1}\|x_i\|_2 + \|x\|_2) + \\ &\quad +\|A\|_2 5(j-1)\epsilon\max_{i\le j-1}\|x_i\|_2 + (j-1)\mathcal{O}(\epsilon)\|A\|_2(\|x\|_2 + \max_{i\le j-1}\|x_i\|_2) + \\ &\quad \mathcal{O}(\epsilon^2)\|A\|_2(\|x\|_2 + \max_{i\le j-1}\|x_i\|_2) = \\ &\le \|A\|_2\|x - x_{j-1}\|_2 + \mathcal{O}(\epsilon^2)\|A\|_2(\|x\|_2 + \max_{i\le j-1}\|x_i\|_2). \qquad (2.154)\end{aligned}$$

Uvrštavajući (2.152), (2.153) i (2.154) u ocjenu (2.144) za $\|\eta_j\|_2$ imamo

$$\begin{aligned}\|\eta_j\|_2 &\le \epsilon\|A\|_2\|x - x_{j-1}\|_2 + (2+c)\epsilon\|A\|_2(\|x_{j-1}\|_2 + \|x_j\|_2) + \\ &\quad +\mathcal{O}(\epsilon^2)\|A\|_2(\|x\|_2 + \max_{i\le j}\|x_i\|_2) \le \\ &\le (\epsilon + \mathcal{O}(\epsilon^2))\|A\|_2(\|x\|_2 + (5+2c)\max_{i\le j}\|x_i\|_2) \le \qquad (2.155)\\ &\le (\epsilon + \mathcal{O}(\epsilon^2))\|A\|_2(\|x\|_2 + (5+2c)\Theta_k\|x\|_2) \qquad (2.156)\end{aligned}$$

To nam pokazuje da je $\|\eta_j\|_2$ također ograđen izrazom $\mathcal{O}(\epsilon)\|A\|_2(\|x\|_2 + \max_{i\le j}\|x_i\|_2)$, čime je dokaz indukcijom završen. Uvrštavanjem (2.155) u (2.149), sumiranjem k puta dobivamo traženu ocjenu. □

Uvrštavanjem ocjena (2.146)–(2.149) u (2.145) daje rezultat sljedećeg teorema.

Teorem 2.10.2 ([11]). *Razlika između pravog reziduala $b - Ax_k$ i izračunatog vektora r_k zadovoljava nejednakost*

$$\frac{\|b - Ax_k - r_k\|_2}{\|A\|_2\|x\|_2} \le (\epsilon + \mathcal{O}(\epsilon^2))[k + 1 + (1 + c + k(10+2c))\Theta_k], \qquad (2.157)$$

gdje je c definiran sa (2.140), a Θ_k sa (2.147).

Dokaz: Prema (2.145) imamo

$$\begin{aligned}&\frac{\|b - Ax_k - r_k\|_2}{\|A\|_2\|x\|_2} \le \\ &\le \epsilon(1+c)\Theta_k + \epsilon + \epsilon 5k\Theta_k + \epsilon k(1 + (5+2c)\Theta_k) + \mathcal{O}(\epsilon^2)(1+\Theta_k) \le \\ &\le \epsilon[k + 1 + (1 + c + k(10+2c))\Theta_k] + \mathcal{O}(\epsilon^2)(1+\Theta_k),\end{aligned}$$

što je bila i tvrdnja teorema. □

Primijetimo da Teorem 2.10.2 slijedi iz analize grešaka zaokruživanja formula (2.134). Nikakve pretpostavke nisu postavljene u vezi sa koeficijentima α_{k-1} ili o vektorima smjera p_{k-1}, ili o tome da li algoritam uopće konvergira. To je stvar analize konvergencije za pojedine konkretne metode.

Ocjene u Lemi 2.10.1 nisu stroge. Naročito u slučaju kada je algoritam blizu konvergencije, možemo očekivati da je norma od r_{j-1} puno manja od $\|A\|_2(\|x\|_2 + \|x_{j-1}\|_2)$, pa ocjena za $\|\eta_j\|_2$ u Lemi 2.10.1 može jako precijenjivati pravu vrijednost. Na osnovi jednakosti (2.141) i pravila za računanje u aritmetici pomičnog zareza (ocjena (2.143)), možemo očekivati da je

$$\|\zeta_j\|_2 \approx \epsilon\|x_{j-1}\|_2,$$

tako da će se najznačajnije greške zaokruživanja dogoditi u koraku u kojem je $\|x_{j-1}\|_2$ najveći. Uglavnom, dok su konstantni izraz i ovisnost o k u (2.157) najvjerojatnije precijenjeni, možemo očekivati da faktor Θ_k igra važnu ulogu u veličini razlike između pravog i izračunatog reziduala.

Promotrimo sada što se događa kada je algoritam blizu konvergencije, budući da u većini slučajeva možemo očekivati da r_k teži k 0 kada $k \to \infty$. Jednom kada se r_{k-1} reducira ispod određenog stupnja, kada je vrlo mali, aporoksimacija rješenja x_k ostaje približno nepromijenjna u odnosu na prethodnu iteraciju. To je zbog toga što je norma izraza, koji mijenja vrijednost aproxsimacije x_{k-1} u x_k, u bliskoj vezi sa veličinom vektora r_{k-1}, kao što se može vidjeti iz (2.142)

$$\alpha_{k-1}p_{k-1} = A^{-1}(r_{k-1} - r_k + \eta_k).$$

Iz ovoga slijedi da kada vektori r_k, a prema tome i $\alpha_{k-1}p_{k-1}$, konvergiraju ka nuli, i kada su iteracije algoritma prošle točku u kojoj $\|A^{-1}r_{k-1}\|_2$ dostiže $\mathcal{O}(\epsilon)\|x_{k-1}\|_2$, pravi rezidual $b - Ax_k$ neće ovisti o k, kao što pretpostavlja ocjena (2.157), već će ostati skoro konstantan. Ako sa S označimo broj koraka koji su potrebni da dostignemo ovo nepromijenjivo stanje, ocjena (2.157) može se zamijeniti sa

$$\frac{\|b - Ax_k - r_k\|_2}{\|A\|_2\|x\|_2} \leq (\epsilon + \mathcal{O}(\epsilon^2))[S + 1 + (1 + c + S(10 + 2c))\Theta], \tag{2.158}$$

gdje je

$$\Theta = \max_j \frac{\|x_j\|_2}{\|x\|_2}. \tag{2.159}$$

Primijetimo da ako iteracije započinju sa ekstremno velikom početnom aproksimacijom x_0, dok je $\|x\|_2$ razumne veličine, faktor Θ_k u (2.157) biti će veliki za sve k, i nijedna metoda oblika (2.134) vjerojatno neće biti u stanju naći dobru aproksimaciju rješenja. To je zbog toga, što se čak niti početni rezidual neće moći izračunati sa pristojnom točnošću, a svi koeficijenti i vektori smjera su definirani pomoću izraza sa izračunatim rezidualom r_k. Mi ćemo pretpostaviti da je, od sada pa na dalje, početna aproksimacija razumne veličine u odnosu na pravo rješenje, i da faktor Θ_k postaje velik samo ako algoritam generira aproksimaciju koja je puno veća i od početne aproksimacije, i od rješenja, u nekoj iteraciji.

Specifični algoritmi

Za konkretne algoritme, koji zadovoljavaju (2.134), interesira nas odnos norme pravog relativnog reziduala $\|b - Ax_k\|_2/\|A\|_2\|x\|_2$ i norme izračunatog relativnog reziduala

$\|r_k\|_2/\|A\|_2\|x\|_2$. Sada ćemo u egzaktnoj aritmetici pokušati ocijeniti Θ u (2.159), što će više–manje vrijediti i u aritmetici konačne preciznosti za algoritme sa određenim svojstvima. Dok je rezultat Teorema 2.10.2 neovisan o prekondicioniranju sustava, ocjena za Θ može biti različita za različito prekondicioniranje.

Ako algoritam reducira euklidsku normu greške u svakom koraku, ili općenitije, ako aproksimacije rješenja u svakoj iteraciji zadovoljavaju,

$$\|x - x_k\|_2 \leq \|x - x_0\|_2, \qquad \forall k,$$

tada imamo

$$\|x_k\|_2 - \|x\|_2 \leq \|x - x_k\|_2 \leq \|x - x_0\|_2 \leq \|x\|_2 + \|x_0\|_2,$$

odakle slijedi

$$\|x_k\|_2 \leq 2\|x\|_2 + \|x_0\|_2 \implies \Theta \leq 2 + \Theta_0. \tag{2.160}$$

Ako euklidska norma greške može rasti, ali se u svakom koraku reducira B-norma greške, $\|x-x_k\|_B = \|B^{1/2}(x-x_k)\|_2$ za neku pozitivno definitnu matricu B, odnosno općenitije, ako vrijedi

$$\|x - x_k\|_B \leq \|x - x_0\|_B,$$

tada imamo

$$\begin{aligned} \|x - x_k\|_2 &= \|B^{-1/2}B^{1/2}(x - x_k)\|_2 \leq \|B^{-1/2}\|_2\|x - x_k\|_B \leq \\ &\leq \|B^{-1/2}\|_2\|x - x_0\|_B \leq \|B^{-1/2}\|_2\|B^{1/2}\|_2\|x - x_0\|_2 \\ &\leq \kappa^{1/2}(B)\|x - x_0\|_2, \end{aligned}$$

odakle slijedi

$$\|x_k\|_2 \leq \|x\|_2 + \kappa^{1/2}(B)(\|x\|_2 + \|x_0\|_2) \implies \Theta \leq 1 + \kappa^{1/2}(B)(1 + \Theta_0). \tag{2.161}$$

Kako bi ostvarili najbolju apriornu ocjenu za pravi rezidual, trebamo odrediti neku normu greške koja se uvijek reducira u odnosu na svoju početnu vrijednost kod izvođenja algoritma, i koja je po mogučnosti što bliža euklidskoj normi.

CG metoda

Pokazali smo da u egzaktnoj aritmetici, svaki korak CG algoritma reducira A normu greške. Tada iz (2.161) slijedi da je

$$\Theta \leq 1 + \kappa^{1/2}(A)(1 + \Theta_0). \tag{2.162}$$

Međutim, znamo da postoji i jača tvrdnja, koja tvrdi da euklidska norma greške također pada monotono. Prema tome iz (2.160) slijedi da je

$$\Theta \leq 2 + \Theta_0. \tag{2.163}$$

Na žalost, svojstvo ortogonalnosti reziduala, koje se koristi za dobivanje redukcije euklidske norme greške u CG algoritmu može se u potpunosti izgubiti u aritmetici konačne preciznosti. Međutim, može se napraviti analogna analiza i u tom slučaju, što nam omogučava sličan zaključak i za aritmetiku konačne preciznosti. U [10] i [15] je pokazano da je greška aproksimacije rješenja x_k generiranog CG algoritmom u aritmetici konačne preciznosti za $Ax = b$ aproksimativno jednako grešci aproksimacije rješenja $\tilde{x}_k$

generiranog algoritmom u egzaktnoj aritmetici, ali koji je primijenjen na većem problemu $\tilde{A}\tilde{x} = \tilde{b}$, sa početnom aproksimacijom $\tilde{x}_0$, koja zadovoljava $\|\tilde{x} - \tilde{x}_0\|_2 \approx \|x - x_0\|_2$. Argumenti koje smo iskoristili da bi dobili monotonu konvergenciju euklidske norme greške mogu se sada primijeniti na egzaktnu CG iteraciju $\tilde{x}_k$ kako bi dobili

$$\|x - x_k\|_2 \approx \|\tilde{x} - \tilde{x}_k\|_2 \leq \|\tilde{x} - \tilde{x}_0\|_2 \approx \|x - x_0\|_2.$$

Slijedi da, u aritmetici konačne preciznosti, pod pogodnim pretpostavkama vezanih uz $\kappa(A)$, možemo očekivati da će (2.163) također vrijediti.

Za broj iteracije u kojem je već r_k dovoljno mali možemo tvrditi da $\|b - Ax_k\|_2/(\|A\|_2\|x\|_2)$ zadovoljava ocjenu (2.158). Prema tome, za CG algoritam procinjujemo da je

$$\min_k \frac{\|b - Ax_k\|_2}{\|A\|_2\|x\|_2} \leq \mathcal{O}(\epsilon)S, \tag{2.164}$$

i taj izraz je neovisan o $\kappa(A)$. Broj koraka S koji su potrebni do se dostigne nepromijenjivo stanje za loše uvjetovane probleme može biti prilično velik, ali on je često i precijenjen.

Kada se hermitska pozitivno definitna matrica prekondicioniranja M upotrijebi u CG algoritmu, to je ekvivalentno primjeni neprekondicioniranog algoritma na problem $M^{-1/2}AM^{-1/2}\hat{x} = M^{-1/2}b$, $x = M^{-1/2}\hat{x}$. U tom slučaju u svakom koraku minimizira se $M^{-1/2}AM^{-1/2}$-norma od $\hat{x} - \hat{x}_k$, što je A-norma greške $x - x_k$, pa ocjena (2.162) i dalje vrijedi. Međutim, sada postoji mogućnost da euklidska norma od $x - x_k$ ipak može rasti. S druge strane, mi znamo da euklidska norma od $\hat{x} - \hat{x}_k$, što je M-norma od $x - x_k$, monotono pada. Za prekondicionirani problem ocjena (2.164) mora se zamijeniti sa

$$\min_k \frac{\|b - Ax_k\|_2}{\|A\|_2\|x\|_2} \leq \mathcal{O}(\epsilon) \min\{\kappa^{1/2}(A), \kappa^{1/2}(M)\}S. \tag{2.165}$$

MINRES(Orthomin(2)) metoda

Analizu za MINRES(Orthomin(2)) metodu možemo izvesti na analogan način kao i za CG metodu. Znamo da u svakom koraku MINRES algoritam minimizira euklidsku normu reziduala, što je A^2-norma greške, pa stoga vrijedi ocjena

$$\Theta \leq 1 + \kappa(A)(1 + \Theta_0). \tag{2.166}$$

Iz toga sada ocjenjujemo normu relatvnog reziduala za MINRES metodu sa

$$\min_k \frac{\|b - Ax_k\|_2}{\|A\|_2\|x\|_2} \leq \mathcal{O}(\epsilon)\kappa(A)S. \tag{2.167}$$

Ako sada ponovo gledamo hermitski prekondicionirani sustav, kao kod CG metode, tada prekondicionirana MINRES metoda je zapravo primjena neprekondicionirane MINRES metode na taj prekondicionirani sustav. Tada se u svakom koraku minimizira euklidska norma reziduala prekondicioniranog sustava, što je $M^{-1/2}AM^{-1}AM^{-1/2}$-norma od $\hat{x} - \hat{x}_k$, odnosno $AM^{-1}A$-norma greške $x - x_k$. Zato u tom slučaju imamo ocjenu

$$\min_k \frac{\|b - Ax_k\|_2}{\|A\|_2\|x\|_2} \leq \mathcal{O}(\epsilon)\kappa(A)\kappa^{1/2}(M)S. \tag{2.168}$$

BCG i CGS metode

Kao što smo pokazali BCG metoda u svakom koraku zadovoljava određena svojstva

biortogonalnosti, a ne zadovoljava nikakva minimizacijska svojstva, tako da norma reziduala i greške prilično oscilira. Zato norma aproksimacije rješenja x_k može postati proizvoljno velika, pa stoga ne možemo dati nikakvu apriorinu ocjenu za Θ u (2.159). Može se dogoditi da algoritam nema uspjeha u postizanju malog pravog reziduala, čak i kad $\|r_k\|_2 \to 0$.

Za CGS metodu znamo da vrijede slični zaključci kao za BCG. Oscilacije norme reziduala mogu biti još jače nego kod BCG metode, pa norma aproksimacije rješenja x_k može postati proizvoljno velika, i zato ponovo ne možemo dati apriornu ocjenu za Θ.

U slučaju kada je Θ jako velik za obje metode, onda možemo očekivati da je

$$\min_k \frac{\|b - Ax_k\|_2}{\|A\|_2\|x\|_2} \leq \mathcal{O}(\epsilon)\Theta.$$

<u>*CGNR i CGNE metode*</u>

CGNR je CG algoritam primijenjen na $A^*Ax = A^*b$ koji u svakom koraku minimizira A^*A-normu greške, što je ekvivalentno minimizaciji euklidske norme reziduala, pa iz (2.161) slijedi da je $\Theta \leq 1+\kappa(A)(1+\Theta_0)$. Budući da se euklidska norma greške također reducira u svakom koraku CG algoritma primijenjenog na sustav $A^*Ax = A^*b$, to isto vrijedi i za CGNR metodu. Kao što je prije napomenuto, može se očekivati da CG metoda reducira grešku i u aritmetici konačne preciznosti, pa tada iz (2.160) slijedi $\Theta \leq 2 + \Theta_0$.

CGNE je ekvivalentan CG algoritmu primijenjenom na $AA^*\hat{x} = b$, $x = A^*\hat{x}$, pa prema tome on u svakom koraku minimizira AA^*-normu od $\hat{x} - \hat{x}_k$, što je ekvivalentno minimizaciji euklidske norme greške. Iz toga slijedi da u egzaktnoj aritmetici vrijedi $\Theta \leq 2 + \Theta_0$, a budući da ova ocjena ne zahtijeva ortogonalnost reziduala, može se očekivati da ona vrijedi i u aritmetici konačne preciznosti.

Za oba algoritma možemo dati konačnu ocjenu

$$\min_k \frac{\|b - Ax_k\|_2}{\|A\|_2\|x\|_2} \leq \mathcal{O}(\epsilon)S \tag{2.169}$$

koja je neovisna o $\kappa(A)$.

Kada se CGNR metoda prekondicionira matricom M, to je ekvivalentno primjeni CG metode na sustav $M^{-*}A^*AM^{-1}\hat{x} = M^{-*}A^*b$, $x = M^{-1}\hat{x}$. U svakom koraku se tada minimizira $M^{-*}A^*AM^{-1}$-norma od $\hat{x} - \hat{x}_k$, što je A^*A-norma greške, pa prema (2.161) slijedi da je $\Theta \leq 1+\kappa(A)(1+\Theta_0)$. Budući da se u svakom koraku još reducira i euklidska norma od $\hat{x} - \hat{x}_k$, što je M^*M-norma greške, to znači da ponovo prema (2.161) slijedi da je i $\Theta \leq 1 + \kappa(M)(1 + \Theta_0)$. Konačna ocjena za prekondicionirani CGNR algoritam glasi

$$\min_k \frac{\|b - Ax\|_2}{\|A\|_2\|x\|_2} \leq \mathcal{O}(\epsilon)\min\{\kappa(A), \kappa(M)\}S. \tag{2.170}$$

Prekondicionirana CGNE metoda je ekvivalentna primjeni CG metode na sustav $M^{-1}AA^*M^{-*}\hat{x} = M^{-1}b$, $x = A^*M^{-*}\hat{x}$. U svakom koraku ona minimizira $M^{-1}AA^*M^{-*}$-normu od $\hat{x} - \hat{x}_k$, što je ponovo ekvivalentno minimizaciji euklidske norme greške. Zbog toga ocjena (2.169) vrijedi i za prekondicionirani CGNE algoritam.

BICGSTAB metoda

Za razliku od prethodno navedenih metoda, BICGSTAB metoda u svakoj iteraciji rekurzivno dobiva x_k iz x_{k-1}, ali ovaj puta pribrajanjem linearne kombinacije dvaju vektora.

Aproksimacija rješenja x_k i rezidual r_k, $k = 1, 2, \ldots$ u BICGSTAB algoritmu računaju se rekurzijama

$$x_k = x_{k-1} + \alpha_{k-1}p_{k-1} + \omega_k q_{k-1}, \qquad r_k = r_{k-1} - \alpha_{k-1}Ap_{k-1} - \omega_k Aq_{k-1}, \tag{2.171}$$

pri čemu su p_{k-1}, α_{k-1} i ω_k definirani u BICGSTAB algoritmu, a $q_{k-1} = r_{k-1/2}$. Kada se ove formule implementiraju u aritmetici konačne preciznosti, dobivaju oblik

$$x_k = x_{k-1} + \alpha_{k-1}p_{k-1} + \omega_k q_{k-1} + \zeta_k, \tag{2.172}$$

$$r_k = r_{k-1} - \alpha_{k-1}Ap_{k-1} - \omega_k Aq_{k-1} + \eta_k, \tag{2.173}$$

kod kojeg se direktnom primjenom pravila (2.137)–(2.140) dobiva

$$\|\zeta_k\|_2 \leq (2\epsilon + \epsilon^2)\|x_{k-1}\|_2 + (3\epsilon + 3\epsilon^2 + \epsilon^3)\|\alpha_{k-1}p_{k-1}\|_2 + (2\epsilon + \epsilon^2)\|\omega_k q_{k-1}\|_2, \tag{2.174}$$

i

$$\begin{aligned}\|\eta_k\|_2 \quad \leq \quad & (2\epsilon + \epsilon^2)\|r_{k-1}\|_2 + (3\epsilon + 3\epsilon^2 + \epsilon^3)\|\alpha_{k-1}Ap_{k-1}\|_2 + \\ & (1+\epsilon)^3\|\alpha_{k-1}(\mathrm{fl}(Ap_{k-1}) - Ap_{k-1})\|_2 + (2\epsilon + \epsilon^2)\|\omega_k Aq_{k-1}\|_2 + \\ & +(1+\epsilon)^2\|\omega_k(\mathrm{fl}(Aq_{k-1}) - Aq_{k-1})\|_2.\end{aligned} \tag{2.175}$$

I u ovom slučaju za odnos između izračunatog reziduala i pravog reziduala vrijede ocjene (2.145) i (2.146), a za ocjene ostalih izraza u (2.145) potreban nam je analogon Leme 2.10.1.

Lema 2.10.3. *Definirajmo Θ_k kao u (2.147) i*

$$\Phi_{k-1} = \max_{j \leq k} \frac{\|\alpha_{j-1}p_{j-1}\|_2 + \|\omega_j q_{j-1}\|_2}{\|\alpha_{j-1}p_{j-1} + \omega_j q_{j-1}\|_2}. \tag{2.176}$$

Pretpostavimo da je $1 - (3\epsilon + 3\epsilon^2 + \epsilon^3)\Phi_{k-1} > 0$. Tada izrazi na desnoj strani (2.145) kod BICGSTAB algoritma zadovoljavaju

$$\sum_{j=1}^{k} \frac{\|\zeta_j\|_2}{\|x\|_2} \leq (\epsilon + \Phi_{k-1}\mathcal{O}(\epsilon^2))k(2 + 6\Phi_{k-1})\Theta_k, \tag{2.177}$$

$$\sum_{j=1}^{k} \frac{\|\eta_j\|_2}{\|A\|_2\|x\|_2} \leq (\epsilon + \Phi_{k-1}\mathcal{O}(\epsilon^2))k[2 + (2 + (6 + 2c)\Phi_{k-1})\Theta_k]. \tag{2.178}$$

Dokaz: Dokaz ove leme sličan je dokazu Leme 2.10.1. Koristit ćemo činjenicu da je

$$\|\alpha_{j-1}p_{j-1}\|_2 + \|\omega_j q_{j-1}\|_2 \leq \Phi_{j-1}\|\alpha_{j-1}p_{j-1} + \omega_j q_{j-1}\|_2. \tag{2.179}$$

Iz (2.172) slijedi

$$\alpha_{j-1}p_{j-1} + \omega_j q_{j-1} = x_j - x_{j-1} - \zeta_j, \tag{2.180}$$

i uvrštavanjem ovog izraza i (2.179) u (2.174) dobivamo

$$\begin{aligned}\|\zeta_j\|_2 \quad \leq \quad & ((2 + 3\Phi_{j-1})\epsilon + (1 + 3\Phi_{j-1})\epsilon^2 + \Phi_{j-1}\epsilon^3)\|x_{j-1}\|_2 + \\ & +(3\epsilon + 3\epsilon^2 + \epsilon^3)\Phi_{j-1}\|x_j\|_2 + (3\epsilon + 3\epsilon^2 + \epsilon^3)\Phi_{j-1}\|\zeta_j\|_2\end{aligned}$$

Koristeći pretpostavku da je $1-(3\epsilon+3\epsilon^2+\epsilon^3)\Phi_{k-1}>0$, što će vrijediti onda i za Φ_{j-1} sa $j\leq k$, to se može napisati kao

$$\|\zeta_j\|_2 \leq \epsilon((2+3\Phi_{j-1})\|x_{j-1}\|_2+3\Phi_{j-1}\|x_j\|_2)+\mathcal{O}(\epsilon^2)\Phi_{j-1}^2(\|x_{j-1}\|_2+\|x_j\|_2). \quad (2.181)$$

Odatle slijedi da je

$$\frac{\|\zeta_j\|_2}{\|x\|_2} \leq (\epsilon+\Phi_{j-1}\mathcal{O}(\epsilon^2))(2+6\Phi_{j-1})\Theta_j,$$

pa sumiranjem po j, i uzimanjem maksimuma za Θ_j i Φ_{j-1} dobivamo ocjenu (2.177).

Nakon uvrštavanja (2.140), (2.179) i (2.181) u (2.175) imamo sljedeći oblik za normu od η_j

$$\begin{aligned}\|\eta_j\|_2 &\leq (2\epsilon+\epsilon^2)\|r_{j-1}\|_2+(\epsilon+\mathcal{O}(\epsilon^2))(3+c)\Phi_{j-1}\|A\|_2(\|x_{j-1}\|_2+\|x_j\|_2)+\\ &\quad +\mathcal{O}(\epsilon^2)\Phi_{j-1}^2\|A\|_2(\|x_{j-1}\|_2+\|x_j\|_2). \quad (2.182)\end{aligned}$$

Ponovo ćemo matematičkom indukcijom provjeriti da $\|\eta_j\|_2$ ima određeni oblik. Pretpostavimo da je svaki izraz $\|\eta_i\|_2$, $i=1,\ldots,j-1$ ocijenjen sa

$$\mathcal{O}(\epsilon)\|A\|_2(\|x\|_2+\Phi_{i-1}\max_{l\leq i}\|x_l\|_2).$$

Za η_1 to vrijedi, jer je, nakon uvrštavanja ocjene za $\|r_0\|_2\leq\|b-Ax_0-r_0\|_2+\|b-Ax_0\|_2$ u (2.182) za $j=1$

$$\begin{aligned}\|\eta_1\|_2 &\leq (2\epsilon+\epsilon^2)\|A\|_2\|x\|_2+(\epsilon+\mathcal{O}(\epsilon^2))(6+2c)\Phi_0\|A\|_2\max_{l\leq 1}\|x_l\|_2+\\ &\quad +\mathcal{O}(\epsilon^2)\|A\|_2(\|x\|_2+\Phi_0^2\max_{l\leq 1}\|x_l\|_2).\end{aligned}$$

Uvrštavanjem pretpostavke indukcije i (2.146) u (2.145) imamo

$$\|r_{j-1}\|_2\leq\|A\|_2\|x-x_{j-1}\|_2+\mathcal{O}(\epsilon)\|A\|_2(\|x\|_2+\Phi_{j-2}\max_{l\leq j-1}\|x_l\|_2). \quad (2.183)$$

Sada kad napokon i (2.183) uvrstimo u (2.182) imamo

$$\begin{aligned}\|\eta_j\|_2 &\leq \epsilon\|A\|_2[2\|x\|_2+(2+(6+2c)\Phi_{j-1})\max_{l\leq j}\|x_l\|_2]+\\ &\quad +\mathcal{O}(\epsilon^2)\|A\|_2(\|x\|_2+\Phi_{j-1}^2\max_{l\leq j}\|x_l\|_2), \quad (2.184)\end{aligned}$$

čime je indukcija dokazana. Odavde je

$$\frac{\|\eta_j\|_2}{\|A\|_2\|x\|_2}\leq(\epsilon+\Phi_{j-1}\mathcal{O}(\epsilon^2))[2+(2+(6+2c)\Phi_{j-1})\Theta_j],$$

pa nakon sumiranja i maksimiziranja dobivamo traženu ocjenu (2.178). □

Uvrštavanje (2.146), (2.177) i (2.178) u (2.145) daje tvrdnju sljedećeg teorema.

Teorem 2.10.4. *Razlika između pravog reziduala $b-Ax_k$ i izračunatog vektora r_k kod BICGSTAB algoritma zadovoljava nejednakost*

$$\frac{\|b-Ax_k-r_k\|_2}{\|A\|_2\|x\|_2}\leq(\epsilon+\Phi_{k-1}\mathcal{O}(\epsilon^2))[2k+1+(1+c+k(4+(12+2c)\Phi_{k-1}))\Theta_k], \quad (2.185)$$

gdje je c definiran sa (2.140), Θ_k sa (2.147) a Φ_{k-1} sa (2.176).

Primijetimo da niti ocjene Leme (2.10.3) nisu stroge. Ovdje isto vrijedi da, kako vektori r_k teže k nuli, tada i vektori $\alpha_{k-1}p_{k-1} + \omega_k q_{k-1}$ teže k nuli. U slučaju da smo prošli točku u kojoj je $\|A^{-1}r_{k-1}\|_2$ dostigao $\mathcal{O}(\epsilon)\|x_{k-1}\|_2$, pravi rezidual $b - Ax_k$ će ostati skoro konstantan. Ako sa S označimo broj koraka potrebnih da postignemo ovo nepromijenjivo stanje tada se ograda (2.185) može zamijeniti sa

$$\frac{\|b - Ax_k - r_k\|_2}{\|A\|_2\|x\|_2} \leq (\epsilon + \Phi_{k-1}\mathcal{O}(\epsilon^2))[2S + 1 + (1 + c + S(4 + (12 + 2c)\Phi))\Theta], \quad (2.186)$$

gdje je

$$\Phi = \max_j \frac{\|\alpha_{j-1}p_{j-1}\|_2 + \|\omega_j q_{j-1}\|_2}{\|\alpha_{j-1}p_{j-1} + \omega_j q_{j-1}\|_2}, \quad (2.187)$$

a Θ je definirano sa (2.159). Ono što je još potrebno je, pronaći nekakvu ocjenu za Φ. Φ ovisi o kutu između vektora p_j i q_j za sve j-ove, pa u slučaju da su u nekom koraku ta dva vektora skoro kolinearna ali suprotnih orijentacija, ili da je $\alpha_{j-1}p_{j-1} + \omega_j q_{j-1}$ mali a sami vektori $\alpha_{j-1}p_{j-1}$ i $\omega_j q_{j-1}$ veliki, Φ može biti prilično velika konstanta, a ocjena (2.186) precijenjena.

Iterativne metode koje se svode na problem najmanjih kvadrata

Druga grupa metoda, koje ćemo promatrati, su one koje se zasnivaju na Arnoldijevom ili Lanczosevom algoritmu, i koje minimiziranje norme reziduala ili kvazireziduala svode na rješavanje problema najmanjih kvadrata dimenzije manje od dimenzije prostora. U ovakvim metodama numerika potrebna za dobivanje ocjene pravog reziduala je puno kompliciranija nego kod metoda sa rekurzijom za dobivanje reziduala, jer se računanje aproksimacije rješenja u svakom koraku odvija u nekoliko faza, a aproksimacija norme reziduala dobiva se obično preko određene komponente vektora koji je nastao kao međurezultat u računu aproksimacije rješenja. U analizi greške metoda koje se svode na problem najmanjih kvadrata, koriste se analize greške raznih tipova QR faktorizacije i rješavanja trokutastih sustava. Najbolji primjeri takvih metoda su GMRES i QMR metode. Prije same analize važno je napomenuti da će sve veličine izražene u algoritmima biti izračunate vrijednosti u aritmetici konačne preciznosti, a sa $\mathrm{fl}(\,\cdot\,)$ ćemo označavati rezultat operacije $(\,\cdot\,)$ u istoj aritmetici. Dimenzija sustava je n.

GMRES metoda

Numerička stabilnost GMRES metode prezentirana je u [8], a analiza u ovoj radnji temelji se na tom članku. Prva faza GMRES algoritma je Arnoldijev algoritma, kojeg možemo smatrati QR faktorizacijom matrice $[q_1\ AQ_k]$, jer za gornjetrokutastu matricu $\bar{R}_{k+1} = [\xi_1\ H_{k+1,k}]$, u egzaktnoj aritmetici vrijedi da je

$$[q_1 \quad AQ_k] = Q_{k+1}\bar{R}_{k+1}.$$

U aritmetici konačne preciznosti ovu dekompoziciju možemo prikazati na sljedeći način

$$\mathrm{fl}(AQ_k) = AQ_k + F_{A,k},$$

i

$$[q_1 \quad \mathrm{fl}(AQ_k)] = Q_{k+1}\bar{R}_{k+1} + F_{0,k},$$

prema redoslijedu vršenja operacija, tako da kao konačni rezultat imamo

$$[q_1 \quad AQ_k + F_{A,k}] = [q_1 \quad Q_{k+1}H_{k+1,k}] + F_{0,k},$$

odakle je

$$AQ_k + F_{A,k} = Q_{k+1}H_{k+1,k} + \widetilde{F}_{0,k},$$

pri čemu je $\widetilde{F}_{0,k}$ $n \times k$ matrica zadnjih k stupaca matrice $F_{0,k}$. Dakle, grešku Arnoldijevog algoritma možemo izraziti kao

$$AQ_k = Q_{k+1}H_{k+1,k} + F_k, \tag{2.188}$$

gdje je

$$F_k = \widetilde{F}_{0,k} - F_{A,k},$$

pa je

$$\|F_k\|_2 \leq \|\widetilde{F}_{0,k}\|_2 + \|F_{A,k}\|_2 \leq \|F_{0,k}\|_2 + \|F_{A,k}\|_2.$$

Potrebno je sada naći ocjene za norme matrica $F_{0,k}$ i $F_{A,k}$. Prema (2.140) za $c = m\sqrt{n}$, gdje je m najveći broj netrivijalnih elemenata u retku matrice A, vrijedi

$$\begin{aligned} \|F_{A,k}\|_2 &\leq \|F_{A,k}\|_F \leq m(\epsilon + \mathcal{O}(\epsilon^2))\|A\|_F\|Q_k\|_F \leq \\ &\leq m\sqrt{n}(\epsilon + \mathcal{O}(\epsilon^2))\|A\|_2\sqrt{\sum_{i=1}^{k}\|q_i\|_2^2} \end{aligned}$$

Izračunati vektori q_i, $i = 1, \ldots, k$ su vjerojatno izgubili svojstvo ortogonalnosti, ali im je norma približno jednaka 1. Naime, oni su dobiveni normalizacijom, odnosno

$$q_i = \mathrm{fl}\left(\frac{\tilde{q}_i}{\mathrm{fl}(\|\tilde{q}_i\|_2)}\right).$$

Prema (2.139) i (2.137) vrijedi

$$\begin{aligned} \mathrm{fl}(\langle\tilde{q}_i, \tilde{q}_i\rangle) &\leq [1 + n(\epsilon + \mathcal{O}(\epsilon^2))]\|\tilde{q}_i\|_2^2, \\ \mathrm{fl}(\|\tilde{q}_i\|_2) &\leq (1+\epsilon)\sqrt{\mathrm{fl}(\langle\tilde{q}_i, \tilde{q}_i\rangle)} \leq (1+\epsilon)\sqrt{[1 + n(\epsilon + \mathcal{O}(\epsilon^2))]\|\tilde{q}_i\|_2^2} \leq \\ &\leq (1+\epsilon)\left[1 + \frac{n}{2}(\epsilon + \mathcal{O}(\epsilon^2))\right]\|\tilde{q}_i\|_2^2 \leq \\ &\leq \left(1 + \frac{n+2}{2}\epsilon + \mathcal{O}(\epsilon^2)\right)\|\tilde{q}_i\|_2, \\ \mathrm{fl}(\|\tilde{q}_i\|_2) &\geq (1-\epsilon)\sqrt{\mathrm{fl}(\langle\tilde{q}_i, \tilde{q}_i\rangle)} \geq \\ &\geq \left(1 - \frac{n+2}{2}\epsilon + \mathcal{O}(\epsilon^2)\right)\|\tilde{q}_i\|_2, \\ \|q_i\|_2 &\leq (1+\epsilon)\frac{\|\tilde{q}_i\|_2}{\mathrm{fl}(\|\tilde{q}_i\|_2)}, \end{aligned}$$

što daje rezultat

$$\begin{aligned} \|q_i\|_2^2 &\leq (1+\epsilon)^2\frac{\|\tilde{q}_i\|_2^2}{\mathrm{fl}(\|\tilde{q}_i\|_2)^2} \leq (1 + 2\epsilon + \epsilon^2)\frac{\|\tilde{q}_i\|_2^2}{\left(1 - \frac{n+2}{2}\epsilon + \mathcal{O}(\epsilon^2)\right)^2\|\tilde{q}_i\|_2^2} \leq \\ &\leq \frac{1 + 2\epsilon + \epsilon^2}{1 + (n+2)\epsilon + n\mathcal{O}(\epsilon^2)} \leq 1 + (n+4)\epsilon + n\mathcal{O}(\epsilon^2). \end{aligned}$$

Prema tome je

$$\begin{aligned}\|F_{A,k}\|_2 &\leq m\sqrt{n}(\epsilon+\mathcal{O}(\epsilon^2))\sqrt{k}\sqrt{1+(n+4)\epsilon+n\mathcal{O}(\epsilon^2)}\|A\|_2 \leq \\ &\leq m\sqrt{n}\sqrt{k}(\epsilon+\mathcal{O}(\epsilon^2))\left(1+\frac{n+4}{2}\epsilon+n\mathcal{O}(\epsilon^2)\right)\|A\|_2 \leq \\ &\leq m\sqrt{n}\sqrt{k}(\epsilon+n\mathcal{O}(\epsilon^2))\|A\|_2,\end{aligned}$$

odnosno

$$\|Q_k\|_F \leq \sqrt{k}\left(1+\frac{n+4}{2}\epsilon+n\mathcal{O}(\epsilon^2)\right).$$

Za daljnje ocjene bit će nam potreban teorem o povratnoj grešci modificiranog Gram–Schmidtovog postupka (MGS) kao metode za dobivanje QR faktorizacije.

Teorem 2.10.5 ([21]). *Pretpostavimo da je MGS postupak primijenjen na $B \in \mathbb{R}^{p\times q}$, $p \geq q$ ranga q, tako da su kao rezultat u aritmetici konačne preciznosti dobivene izračunate matrice $\widetilde{Q} \in \mathbb{R}^{p\times q}$ i $\widetilde{R} \in \mathbb{R}^{q\times q}$. Tada postoje konstante c_i, $i = 1,\ldots,4$ takve da je*

(i) $B+\Delta B_1 = \widetilde{Q}\widetilde{R}, \qquad \|\Delta B_1\|_F \leq c_1 q^2(\epsilon+\mathcal{O}(\epsilon^2))\|B\|_F,$

(ii) *i postoji ortonormalna matrica $Q \in \mathbb{R}^{p\times q}$ takva da je*
$B+\Delta B_2 = Q\widetilde{R}, \qquad \|\Delta B_F\|_2 \leq c_2 pq(\epsilon+\mathcal{O}(\epsilon^2))\|B\|_F,$

(iii) $\|\widetilde{Q}-Q\|_2 \leq (c_1 q^{5/2}+c_2 q^{3/2}p)(\kappa(B)\epsilon+\mathcal{O}((\kappa(B)\epsilon)^2)),$

(iv) $\|\widetilde{Q}^T\widetilde{Q}-I\|_2 \leq (c_3 q^3+c_4 q^2 p)(\kappa(B)\epsilon+\mathcal{O}((\kappa(B)\epsilon)^2)).$

Dokaz: (i) MGS postupak u egzaktnoj aritmetici može se izraziti u matričnom obliku pomoću matrica $B_k = [q_1 \ldots q_{k-1}\ b_k^{(k)} \ldots b_q^{(k)}]$, pri čemu je $B = B_1 = [b_1^{(1)} \ldots b_q^{(1)}]$, i $b_j^{(k+1)} = b_j^{(k)} - \langle b_j^{(k)}, q_k\rangle q_k$ za $j = k+1,\ldots q$, te $q_{k+1} = b_k^{(k)}/\|b_k^{(k)}\|_2$. Dakle, lako se vidi da MGS transformira matricu B u ortonormalnu matricu $B_{q+1} = Q$ preko niza transformacija $B_k = B_{k+1}R_k$, gdje je R_k ekvivalentna jediničnoj matrici svugdje osim u k-tom retku, u kojem se podudara sa konačnim R. Imamo

$$[q_1 \ldots q_{k-1}\ b_k^{(k)} \ldots b_q^{(k)}] =$$

$$= [q_1 \ldots q_k\ b_{k+1}^{(k+1)} \ldots b_q^{(k+1)}] \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & r_{kk} & \cdots & \cdots & r_{kq} & \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & \end{bmatrix}.$$

Za izračunate matrice tada vrijedi

$$\widetilde{B}_k = \widetilde{B}_{k+1}\widetilde{R}_k + \Delta B_k, \qquad |\Delta B_k| \leq (2\epsilon+\mathcal{O}(\epsilon^2))|\widetilde{B}_{k+1}||\widetilde{R}_k|,$$

jer se u svakom stupcu od $\widetilde{R}_k$ nalaze najviše dva netrivijalna elementa. Uzastopnim izvođenjem svih koraka ove rekurzije dobivamo

$$B = \widetilde{Q}\widetilde{R} + \Delta B_q \widetilde{R}_{q-1}\cdots\widetilde{R}_1 + \cdots + \Delta B_2\widetilde{R}_1 + \Delta B_1.$$

Odatle je

$$|B-\widetilde{Q}\widetilde{R}| \leq (2\epsilon+\mathcal{O}(\epsilon^2))(|\widetilde{B}_{q+1}||\widetilde{R}_q|\cdots|\widetilde{R}_1|+\cdots+|\widetilde{B}_3||\widetilde{R}_2||\widetilde{R}_1|+|\widetilde{B}_2||\widetilde{R}_1|), \quad (2.189)$$

a tipiči član ove sume ima oblik

$$|\widetilde{B}_k||\widetilde{R}_{k-1}|\cdots|\widetilde{R}_1| = |[\tilde{q}_1 \ldots \tilde{q}_{k-1}\ \tilde{b}_k^{(k)} \ldots \tilde{b}_q^{(k)}]|S_{k-1},$$

gdje se S_{k-1} podudara sa $|\widetilde{R}|$ u prvih $k-1$ redaka, a sa identitetom na zasnjih $q-k+1$ redaka. Za izračunate $\tilde{q}_i$ također vrijedi da je

$$\|\tilde{q}_i\|_2^2 = 1 + (p+4)\epsilon + p\mathcal{O}(\epsilon^2),$$

zato za $\tilde{b}_j^{(k+1)} = \mathrm{fl}(\tilde{b}_j^{(k)} - \langle\tilde{q}_k, \tilde{b}_j^{(k)}\rangle\tilde{q}_k)$ direktnim računom iz (2.137), (2.138) i (2.139) imamo

$$\begin{aligned}\|\tilde{b}_j^{(k+1)} - (\tilde{b}_j^{(k)} - \langle\tilde{q}_k, \tilde{b}_j^{(k)}\rangle\tilde{q}_k)\|_2 &\leq [\epsilon + ((p+2)\epsilon + p\mathcal{O}(\epsilon^2))\|\tilde{q}_k\|_2^2]\|\tilde{b}_j^{(k)}\|_2 \leq \\ &\leq [(p+3)\epsilon + p\mathcal{O}(\epsilon^2)]\|\tilde{b}_j^{(k)}\|_2\end{aligned}$$

a s druge strane je

$$\begin{aligned}&\|\tilde{b}_j^{(k)} - \langle\tilde{q}_k, \tilde{b}_j^{(k)}\rangle\tilde{q}_k\|_2 \\ &\leq \sqrt{\|\tilde{b}_j^{(k)}\|_2^2 - 2\langle\tilde{q}_k, \tilde{b}_j^{(k)}\rangle^2 + \langle\tilde{q}_k, \tilde{b}_j^{(k)}\rangle^2\|\tilde{q}_k\|_2^2} \leq \\ &\leq \sqrt{\|\tilde{b}_j^{(k)}\|_2^2 - \langle\tilde{q}_k, \tilde{b}_j^{(k)}\rangle^2 + \langle\tilde{q}_k, \tilde{b}_j^{(k)}\rangle^2[(p+4)\epsilon + p\mathcal{O}(\epsilon^2)]} \leq \\ &\leq \sqrt{\|\tilde{b}_j^{(k)}\|_2^2 + \|\tilde{b}_j^{(k)}\|_2^2[1 + (p+4)\epsilon + p\mathcal{O}(\epsilon^2)][(p+4)\epsilon + p\mathcal{O}(\epsilon^2)]} \leq \\ &\leq \|\tilde{b}_j^{(k)}\|_2\sqrt{1 + (p+4)\epsilon + p\mathcal{O}(\epsilon^2)} \leq \\ &\leq \|\tilde{b}_j^{(k)}\|_2\left(1 + \frac{p+4}{2}\epsilon + p\mathcal{O}(\epsilon^2)\right)\end{aligned}$$

tako da na kraju dobivamo

$$\|\tilde{b}_j^{(k+1)}\|_2 \leq [1 + (3/2p+5)\epsilon + p\mathcal{O}(\epsilon^2)]\|\tilde{b}_j^{(k)}\|_2.$$

Iz $\tilde{r}_{kj} = \mathrm{fl}(\langle\tilde{q}_k, \tilde{b}_j^{(k)}\rangle)$ slijedi

$$\begin{aligned}|\tilde{r}_{kj}| &\leq [1 + p(\epsilon + \mathcal{O}(\epsilon^2)]\|\tilde{q}_k\|_2\|\tilde{b}_j^{(k)}\|_2 \leq \\ &\leq [1 + p(\epsilon + \mathcal{O}(\epsilon^2)]\left(1 + \frac{p+4}{2}\epsilon + \mathcal{O}(\epsilon^2)\right)\cdot \\ &\quad \cdot\left[1 + \left(\frac{3}{2}p + 5\right)\epsilon + p\mathcal{O}(\epsilon^2)\right]^{k-1}\|b_j\|_2 \leq \\ &\leq (1 + p\mathcal{O}(\epsilon))\|b_j\|_2,\end{aligned}$$

odnosno

$$\|\tilde{R}\|_F \leq \sqrt{q}(1 + p\mathcal{O}(\epsilon))\|B\|_F.$$

Dalje, imamo

$$S_{k-1} = \begin{bmatrix} |\widetilde{R}_{(k-1)\times q}| \\ [0\ I_{(q-k+1)\times(q-k+1)}] \end{bmatrix},$$

pri čemu je $\widetilde{R}_{(k-1)\times q}$ gornji $(k-1)\times q$ blok matrice $\widetilde{R}$, pa vrijedi

$$|\widetilde{B}_k||\widetilde{R}_{k-1}|\cdots|\widetilde{R}_1| = |[\tilde{q}_1\ldots\tilde{q}_{k-1}]||\widetilde{R}_{(k-1)\times q}| + |[0\ldots0\ \tilde{b}_k^{(k)}\ldots\tilde{b}_q^{(k)}]|.$$

Slijedi

$$\begin{aligned}
&\||\widetilde{B}_k||\widetilde{R}_{k-1}|\cdots|\widetilde{R}_1|\|_F \\
&\leq \|[\tilde{q}_1\ldots\tilde{q}_{k-1}]\|_F\|\widetilde{R}\|_F + \|[\tilde{b}_k^{(k)}\ldots\tilde{b}_q^{(k)}]\|_F \leq \\
&\leq \sqrt{k-1}\left(1+\frac{p+4}{2}\epsilon+p\mathcal{O}(\epsilon^2)\right)\sqrt{q}(1+p\mathcal{O}(\epsilon))\|B\|_F + \\
&\quad +\sqrt{q-k+1}\left[1+\left(\frac{3}{2}p+5\right)\epsilon+p\mathcal{O}(\epsilon^2)\right]^{k-1}\|B\|_F \leq \\
&\leq [q(1+p\mathcal{O}(\epsilon))+\sqrt{q}(1+pq\mathcal{O}(\epsilon))]\|B\|_F \leq \\
&\leq (q+\sqrt{q}+pq^{3/2}\mathcal{O}(\epsilon))\|B\|_F \leq q(2+pq^{1/2}\mathcal{O}(\epsilon))\|B\|_F,
\end{aligned}$$

pa koristeći (2.189) imamo

$$\begin{aligned}
\|B-\widetilde{Q}\widetilde{R}\|_F &\leq (2\epsilon+\mathcal{O}(\epsilon^2))q^2(2+pq^{1/2}\mathcal{O}(\epsilon))\|B\|_F \leq \\
&\leq 4q^2(\epsilon+\mathcal{O}(\epsilon^2)\|B\|_F.
\end{aligned}$$

(ii) Analizu greške MGS metode pojednostavljuje činjenica da je MGS metoda ekvivalentna matematički i numerički Householderovoj QR faktorizaciji, izvedenoj na proširenoj matrici $\begin{bmatrix} 0_q \\ B \end{bmatrix} \in \mathbb{R}^{(p+q)\times q}$, za koju već postoje gotove ocjene greške, vidi [2]. Može se pokazati da je u egzaktnoj aritmetici

$$P^T\begin{bmatrix} 0_q \\ B \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix}, \qquad P^T = P_q\cdots P_1, \tag{2.190}$$

pri čemu je R ekvivalentna gornje trokutastoj matrici koja se dobije MGS metodom, a

$$P_k = I - v_k v_k^T, \quad v_k = \begin{bmatrix} -\xi_k \\ q_k \end{bmatrix},\ k=1,\ldots,q,$$

ξ_k je jednak k-tom jediničnom vektoru, a q_k je k-ti ortonormirani vektor dobiven MGS metodom. Primjenjujući Teorem 18.4 iz [21, str. 368] na (2.190) dobivamo da postoji ortogonalna matrica $\widetilde{P}$ takva da vrijedi

$$\begin{bmatrix} \Delta B_3 \\ B+\Delta B_4 \end{bmatrix} = \widetilde{P}\begin{bmatrix} \widetilde{R} \\ 0 \end{bmatrix} = \begin{bmatrix} \widetilde{P}_{11} \\ \widetilde{P}_{21} \end{bmatrix}\widetilde{R}, \tag{2.191}$$

sa

$$\left\|\begin{bmatrix} \Delta B_3 \\ \Delta B_4 \end{bmatrix}\right\|_F \leq dq(p+q)(\epsilon+\mathcal{O}(\epsilon^2))\|B\|_F,$$

za neku malu konstantu d.

Promotrimo sada danu particiju matrice $\widetilde{P}$

$$\widetilde{P} = \begin{array}{c} \\ q \\ p \end{array}\begin{array}{c} \begin{array}{cc} q & p \end{array} \\ \begin{bmatrix} \widetilde{P}_{11} & \widetilde{P}_{12} \\ \widetilde{P}_{21} & \widetilde{P}_{22} \end{bmatrix} \end{array}$$

dimenzije $(q+p)\times(q+p)$, za koju, zbog $q\le p$, vrijedi da je $2q\le(q+p)$. Iz tog razloga možemo na $\widetilde{P}$ primijeniti teorem o CS dekompoziciji iz [19], koji kaže da, u ovom slučaju, postoje unitarne matrice $\widetilde{U}=\mathrm{diag}(U_{11},U_{22})$ i $\widetilde{V}=\mathrm{diag}(V_{11},V_{22})$, gdje su $U_{11},V_{11}\in\mathbb{R}^{q\times q}$, i $U_{22},V_{22}\in\mathbb{R}^{p\times p}$ ortogonalne matrice, takve da je

$$\begin{bmatrix} U_{11}^T & 0\\ 0 & U_{22}^T\end{bmatrix}\begin{bmatrix}\widetilde{P}_{11} & \widetilde{P}_{12}\\ \widetilde{P}_{21} & \widetilde{P}_{22}\end{bmatrix}\begin{bmatrix} V_{11} & 0\\ 0 & V_{22}\end{bmatrix}=\begin{array}{c} \\ q\\ q\\ p-q\end{array}\begin{array}{c}\begin{array}{ccc} q & q & p-q\end{array}\\ \begin{bmatrix}\Gamma & -\Sigma & 0\\ \Sigma & \Gamma & 0\\ 0 & 0 & I\end{bmatrix}\end{array},$$

pri čemu su

$$\begin{aligned}\Gamma &= \mathrm{diag}(\gamma_1,\ldots,\gamma_q)\ge 0\\ \Sigma &= \mathrm{diag}(\sigma_1,\ldots,\sigma_q)\ge 0\end{aligned}$$

i vrijedi

$$\Gamma^2+\Sigma^2=I.$$

Odavde slijedi

$$\begin{aligned}U_{11}^T\widetilde{P}_{11}V_{11} &= \Gamma\\ U_{22}^T\widetilde{P}_{21}V_{11} &= \begin{bmatrix}\Sigma\\ 0\end{bmatrix},\end{aligned}\tag{2.192}$$

odnosno

$$\begin{aligned}\widetilde{P}_{11} &= U_{11}\Gamma V_{11}^T\\ \widetilde{P}_{21} &= (U_{22})_{p\times q}\Sigma V_{11}^T,\end{aligned}$$

gdje je $(U_{22})_{p\times q}$ matrica koju čine prvih q stupaca matrice U_{22}. Zato možemo tvrditi da postoje ortonormalne matrice $U=U_{11}$ i $V=(U_{22})_{p\times q}$, ortogonalna matrica $W=V_{11}$, te kvadratne pozitivne dijagonalne matrice Γ i Σ, sa svojstvom $\Gamma^2+\Sigma^2=I$, takve da vrijedi $\widetilde{P}_{11}=U\Gamma W^T$ i $\widetilde{P}_{21}=V\Sigma W^T$. Neka je sada $Q=VW^T$, tada je

$$\begin{aligned}\Delta B_3 &= U\Gamma W^T\widetilde{R}\\ B+\Delta B_4 &= V\Sigma W^T\widetilde{R}\end{aligned}$$

i, zbog toga što je $(I-\Sigma)(I+\Sigma)=I-\Sigma^2$, odnosno $I+\Sigma>0$, pa je $I-\Sigma=(I+\Sigma)^{-1}(I-\Sigma^2)=(I+\Sigma)^{-1}\Gamma^2$, vrijedi sljedeće

$$\begin{aligned}Q\widetilde{R} &= \widetilde{P}_{21}\widetilde{R}+(Q-\widetilde{P}_{21})\widetilde{R}=B+\Delta B_4+V(I-\Sigma)W^T\widetilde{R}=\\ &= B+\Delta B_4+V(I+\Sigma)^{-1}\Gamma U^TU\Gamma W^T\widetilde{R}\\ &= B+\Delta B_4+V(I+\Sigma)^{-1}\Gamma U^T\Delta B_3.\end{aligned}$$

Ako sada definiramo

$$F=V(I+\Sigma)^{-1}\Gamma U^T,$$

i

$$\Delta B_2 = \Delta B_4 + F\Delta B_3,$$

tada je

$$\|F\|_2 = \|(I+\Sigma)^{-1}\Gamma\|_2 = \max_i \left|\frac{\gamma_i}{1+\sigma_i}\right| \leq 1.$$

Na kraju imamo

$$\begin{aligned}\|\Delta B_2\|_F &\leq \|\Delta B_3\|_F + \|\Delta B_4\|_F \leq \\ &\leq 2\left\|\begin{bmatrix}\Delta B_3\\ \Delta B_4\end{bmatrix}\right\|_F \leq 2dq(p+q)(\epsilon+\mathcal{O}(\epsilon^2))\|B\|_F \\ &\leq 4dqp(\epsilon+\mathcal{O}(\epsilon^2))\|B\|_F\end{aligned}$$

(iii) Iz (i) i (ii) slijedi da je $\widetilde{Q}-Q = (\Delta B_1 - \Delta B_2)\widetilde{R}^{-1}$, pa vrijedi

$$\widetilde{R} = Q^T B + Q^T \Delta B_2,$$

i

$$\begin{aligned}\widetilde{R}^{-1} &= \widetilde{R}^{-1}Q^T B(Q^T B)^{-1} = \\ &= \widetilde{R}^{-1}(Q^T B + Q^T\Delta B_2 - Q^T\Delta B_2)(Q^T B)^{-1} = \\ &= [I - \widetilde{R}^{-1}Q^T\Delta B_2](Q^T B)^{-1},\end{aligned}$$

odnosno

$$\begin{aligned}\|\widetilde{R}^{-1}\|_2 &\leq \|(Q^T B)^{-1}\|_2(1+\|\Delta B_2\|_2\|\widetilde{R}^{-1}\|_2) \leq \\ &\leq \|(Q^T B)^{-1}\|_2(1+c_2pq^{3/2}(\epsilon+\mathcal{O}(\epsilon^2))\|B\|_2\|\widetilde{R}^{-1}\|_2).\end{aligned}$$

Sada nam još ostaje ocijeniti normu $\|(Q^T B)^{-1}\|_2$, a za što će nam biti potrebni neki rezultati vezani uz generalizirani inverz. Prema točki 9. u Teoremu 2.13 iz [19] vrijedi da je

$$(Q^T B)^{-1} = (Q^T B)^+ = B^+(Q^+)^T = B^+Q,$$

pa iz toga slijedi

$$\|(Q^T B)^{-1}\|_2 \leq \|B^+\|_2.$$

Sada konačno možemo pisati

$$\|\widetilde{R}^{-1}\|_2 \leq \|B^+\|_2(1+c_2pq^{3/2}(\epsilon+\mathcal{O}(\epsilon^2))\|B\|_2\|\widetilde{R}^{-1}\|_2).$$

Uz uvjet da je $c_2pq^{3/2}(\epsilon+\mathcal{O}(\epsilon^2))\kappa(B) < 1$, za $\kappa(B) = \|B\|_2\|B^+\|_2$, imamo

$$\|\widetilde{R}^{-1}\|_2 \leq \frac{\|B^+\|_2}{1-c_2pq^{3/2}(\epsilon+\mathcal{O}(\epsilon^2))\kappa(B)}.$$

Sada napokon možemo dati ocjenu za $\|\widetilde{Q}-Q\|_2$.

$$\begin{aligned}\|\widetilde{Q}-Q\|_2 &\leq (\|\Delta B_1\|_F + \|\Delta B_2\|_F)\|\widetilde{R}^{-1}\|_2 \leq \\ &\leq (c_1q^2 + c_2pq)(\epsilon+\mathcal{O}(\epsilon^2))\|B\|_F\|\widetilde{R}^{-1}\|_2 \leq \\ &\leq (c_1q^{5/2} + c_2pq^{3/2})(\epsilon+\mathcal{O}(\epsilon^2))\frac{\kappa(B)}{1-c_2pq^{3/2}(\epsilon+\mathcal{O}(\epsilon^2))\kappa(B)} \leq \\ &\leq (c_1q^{5/2} + c_2pq^{3/2})(\kappa(B)\epsilon + \mathcal{O}((\kappa(B)\epsilon)^2)).\end{aligned}$$

(iv) Vrijedi sljedeće

$$\widetilde{Q}^T\widetilde{Q} - I = (\widetilde{Q} - Q)^T(\widetilde{Q} + Q) + Q^T\widetilde{Q} - \widetilde{Q}^T Q,$$

a s druge strane, ako definiramo $\Delta Q = \widetilde{Q} - Q$, je

$$\begin{aligned} Q^T\widetilde{Q} - \widetilde{Q}^T Q &= Q^T(Q + \Delta Q) - (Q^T + \Delta Q^T)Q = \\ &= I + Q^T\Delta Q - I - \Delta Q^T Q = Q^T\Delta Q - \Delta Q^T Q. \end{aligned}$$

Jer je $\|\widetilde{Q}\|_2 \leq \|\widetilde{Q}\|_F \leq \sqrt{q}(1 + (p+4)/2\epsilon + p\mathcal{O}(\epsilon^2))$, vrijedi

$$\begin{aligned} \|\widetilde{Q}^T\widetilde{Q} - I\|_2 &\leq \|\widetilde{Q} - Q\|_2(\|\widetilde{Q}\|_2 + 1) + 2\|\widetilde{Q} - Q\|_2 \leq \\ &\leq [3 + \sqrt{q}(1 + (p+4)/2\epsilon + p\mathcal{O}(\epsilon^2))]\|\widetilde{Q} - Q\|_2 \leq \\ &\leq [3 + \sqrt{q}(1 + (p+4)/2\epsilon + p\mathcal{O}(\epsilon^2))] \cdot \\ &\quad \cdot (c_1 q^{5/2} + c_2 pq^{3/2})(\kappa(B)\epsilon + \mathcal{O}((\kappa(B)\epsilon)^2)) \leq \\ &\leq (3 + \sqrt{q})(c_1 q^{5/2} + c_2 pq^{3/2})(\kappa(B)\epsilon + \mathcal{O}((\kappa(B)\epsilon)^2)). \end{aligned}$$

□

Sada ovaj teorem primjenjujemo na naš problem QR faktorizacije. Iz *(i)* dijela teorema slijedi da postoji neka konstanata c, i postoji konstanta d za koju vrijedi $1 + nk\|A\|_2^2 \leq dnk\|A\|_2^2$, takve da je

$$\begin{aligned} \|F_{0,k}\|_2 &\leq c(k+1)^2(\epsilon + \mathcal{O}(\epsilon^2))\|[q_1 \quad AQ_k + F_{A,k}]\|_F \leq \\ &\leq c(k+1)^2(\epsilon + \mathcal{O}(\epsilon^2))\sqrt{\|q_1\|_2^2 + (\|AQ_k\|_F + \|F_{A,k}\|_F)^2} \leq \\ &\leq c(k+1)^2(\epsilon + \mathcal{O}(\epsilon^2))\sqrt{\|q_1\|_2^2 + [1 + m(\epsilon + \mathcal{O}(\epsilon^2))]^2\|A\|_F^2\|Q_k\|_F^2} \leq \\ &\leq c(k+1)^2(\epsilon + \mathcal{O}(\epsilon^2))\sqrt{1 + nk\|A\|_2^2 + n^2k\|A\|_2^2\mathcal{O}(\epsilon)} \leq \\ &\leq c(k+1)^2(\epsilon + \mathcal{O}(\epsilon^2))\sqrt{dnk\|A\|_2^2(1 + n\mathcal{O}(\epsilon))} \leq \\ &\leq c(k+1)^2(\epsilon + \mathcal{O}(\epsilon^2))\sqrt{d}\sqrt{k}\sqrt{n}(1 + n\mathcal{O}(\epsilon))\|A\|_2 \leq \\ &\leq c\sqrt{d}(k+1)^2\sqrt{k}\sqrt{n}(\epsilon + n\mathcal{O}(\epsilon^2))\|A\|_2 \leq \\ &\leq c_5 k^{5/2} n^{1/2}(\epsilon + \mathcal{O}(\epsilon^2))\|A\|_2, \end{aligned}$$

za neku konstantu c_5. Dakle, vrijedi konačna ocjena

$$\begin{aligned} \|F_k\|_2 &\leq \|F_{0,k}\|_2 + \|F_{A,k}\|_2 \leq \\ &\leq (c_5 k^{5/2} n^{1/2} + mk^{1/2} n^{1/2})(\epsilon + n\mathcal{O}(\epsilon^2))\|A\|_2 = \\ &= p(m,k,n)(\epsilon + n\mathcal{O}(\epsilon^2))\|A\|_2. \end{aligned}$$

Dalje, prema *(ii)* dijelu teorema postoji ortonormalna matrica $\widehat{Q}_{k+1}$ takva da je

$$[q_1 \quad AQ_k + F_{A,k}] - \overline{\overline{F}}_k = \widehat{Q}_{k+1}[\xi_1 \quad H_{k+1,k}],$$

takva da postoji konstanta c za koju vrijedi

$$\begin{aligned} \|\overline{\overline{F}}_k\|_2 &\leq c(k+1)n(\epsilon + \mathcal{O}(\epsilon^2))\|[q_1 \quad AQ_k + F_{A,k}]\|_F \leq \\ &\leq c(k+1)n(\epsilon + n\mathcal{O}(\epsilon^2))\sqrt{d}\sqrt{k}\sqrt{n}(1 + n\mathcal{O}(\epsilon))\|A\|_2 \leq \\ &\leq c\sqrt{d}(k+1)\sqrt{k}n^{3/2}(\epsilon + n\mathcal{O}(\epsilon^2))\|A\|_2 \leq \\ &\leq c_6 k^{3/2} n^{3/2}(\epsilon + n\mathcal{O}(\epsilon^2))\|A\|_2, \end{aligned}$$

za neku konstantu c_6. Sada, ako definiramo $\overline{F}_k$ kao matricu koju čine zadnjih k stupaca matrice $\overline{\overline{F}}_k$ minus $F_{A,k}$, tada vrijedi da je

$$AQ_k = \widehat{Q}_{k+1}H_{k+1,k} + \overline{F}_k,$$

i

$$\begin{aligned}\|\overline{F}_k\|_2 &\leq \|\overline{\overline{F}}_k\|_2 + \|F_{A,k}\|_2 \leq \\ &\leq (c_6k^{3/2}n^{3/2} + mk^{1/2}n^{1/2})(\epsilon + n\mathcal{O}(\epsilon^2))\|A\|_2 \leq \\ &\leq c_7k^{3/2}n^{3/2}(\epsilon + n\mathcal{O}(\epsilon^2))\|A\|_2,\end{aligned}$$

za neku konstantu c_7.

Preostalo nam je još provjeriti koliko je izračunati Q_{k+1} udaljen od ortonormalne matrice. U *(iii)* i *(iv)* dijelovima teorema $\|Q_{k+1} - \widehat{Q}_{k+1}\|_2$ i $\|Q_{k+1}^TQ_{k+1} - I\|_2$ su ograđeni sa $\kappa([q_1\ AQ_k + F_{A,k}])$, što je zapravo $\kappa([q_1\ \mathrm{fl}(AQ_k)])$, odnosno broj uvjetovanosti za izračunatu matricu nad kojom izvršavamo QR faktorizaciju, koja nam tako i onako stoji na raspolaganju u svakoj iteraciji GMRES algoritma. Imamo

$$\begin{aligned}\|Q_{k+1} - \widehat{Q}_{k+1}\|_2 &\leq (c_1(k+1)^{5/2} + c_2(k+1)^{3/2}n)[\kappa([q_1\ \mathrm{fl}(AQ_k)])\epsilon + \\ &\quad + \mathcal{O}((\kappa([q_1\ \mathrm{fl}(AQ_k)])\epsilon)^2)] \leq \\ &\leq c_8k^{3/2}n[\kappa([q_1\ \mathrm{fl}(AQ_k)])\epsilon + \mathcal{O}((\kappa([q_1\ \mathrm{fl}(AQ_k)])\epsilon)^2)]\end{aligned}$$

i

$$\begin{aligned}\|Q_{k+1}^TQ_{k+1} - I\|_2 &\leq (c_3(k+1)^3 + c_2(k+1)^2n)[\kappa([q_1\ \mathrm{fl}(AQ_k)])\epsilon + \\ &\quad + \mathcal{O}((\kappa([q_1\ \mathrm{fl}(AQ_k)])\epsilon)^2)] \leq \\ &\leq c_9k^2n[\kappa([q_1\ \mathrm{fl}(AQ_k)])\epsilon + \mathcal{O}((\kappa([q_1\ \mathrm{fl}(AQ_k)])\epsilon)^2)]\end{aligned}$$

za neke konstante c_8 i c_9. Ovime smo završili analizu Arnoldijevog algoritma u aritmetici konačne preciznosti.

Dalje nas interesira koliko izračunata norma Arnoldijevog reziduala $\mathrm{fl}(\|\beta\xi_1 - H_{k+1,k}\cdot \cdot y_k\|_2)$ odstupa od norme pravog reziduala $\|b - Ax_k\|_2$, u k-toj iteraciji GMRES algoritma izvođenog u aritmetici konačne preciznosti. Aproksimacija rješenja x_k, u k-tom koraku, je u toj aritmetici tada izražena sa

$$x_k = x_0 + Q_ky_k + \Delta x_k, \tag{2.193}$$

gdje za $\Delta x_k = \mathrm{fl}(x_0 + Q_ky_k) - (x_0 + Q_ky_k)$, prema (2.138) i (2.140), te ogradi za normu od Q_k, vrijedi ocjena

$$\begin{aligned}\|\Delta x_k\|_2 &\leq \|\mathrm{fl}(x_0 + Q_ky_k) - (x_0 + \mathrm{fl}(Q_ky_k))\|_2 + \|\mathrm{fl}(Q_ky_k) - Q_ky_k\|_2 \leq \\ &\leq \epsilon(\|x_0\|_2 + \|\mathrm{fl}(Q_ky_k)\|_2) + \|\mathrm{fl}(Q_ky_k) - Q_ky_k\|_2 \leq \\ &\leq \epsilon(\|x_0\|_2 + \|Q_ky_k\|_2) + (1+\epsilon)\|\mathrm{fl}(Q_ky_k) - Q_ky_k\|_2 \leq \\ &\leq \epsilon(\|x_0\|_2 + \|Q_k\|_F\|y_k\|_2) + (1+\epsilon)k(\epsilon + \mathcal{O}(\epsilon^2))\|Q_k\|_F\|y_k\|_2 \leq \\ &\leq \epsilon\|x_0\|_2 + [(1+k)\epsilon + k\mathcal{O}(\epsilon^2)]\|Q_k\|_F\|y_k\|_2 \leq \\ &\leq \epsilon\|x_0\|_2 + [(1+k)\epsilon + k\mathcal{O}(\epsilon^2)]\sqrt{k}\left(1 + \frac{n+4}{2}\epsilon + n\mathcal{O}(\epsilon^2)\right)\|y_k\|_2 \leq \\ &\leq \epsilon\|x_0\|_2 + \sqrt{k}[(k+1)\epsilon + nk\mathcal{O}(\epsilon^2)]\|y_k\|_2 \leq \\ &\leq \epsilon\|x_0\|_2 + 2k^{3/2}(\epsilon + n\mathcal{O}(\epsilon^2))\|y_k\|_2\end{aligned}$$

Koristeći (2.188) i (2.193) imamo

$$\begin{aligned} b - Ax_k &= b - Ax_0 - AQ_k - A\Delta x_x = \\ &= b - Ax_0 - Q_{k+1}H_{k+1,k}y_k - F_k y_k - A\Delta x_k = \\ &= (b - Ax_0 - \beta q_1) + (\beta q_1 - Q_{k+1}H_{k+1,k}y_k) - F_k y_k - A\Delta x_k = \\ &= (b - Ax_0 - \beta q_1) + Q_{k+1}(\beta\xi_1 - H_{k+1,k}y_k) - F_k y_k - A\Delta x_k. \end{aligned}$$

Dakle, vrijedi

$$\|(b-Ax_k)-Q_{k+1}(\beta\xi_1-H_{k+1,k}y_k)\|_2 \leq \|b-Ax_0-\beta q_1\|_2+\|F_k y_k\|_2+\|A\Delta x_k\|_2. \quad (2.194)$$

U sljedećem koraku trebamo odrediti ocjene normi na desnoj strani nejednakosti. Za δ_i, takve da je $|\delta_i| \leq \epsilon$, $i = 1, \ldots, n$, iz (2.136) imamo

$$\begin{aligned} \|b - Ax_0 - \beta q_1\|_2 &= \left\| b - Ax_0 - \beta \left(I + \begin{bmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_n \end{bmatrix} \right) \frac{\mathrm{fl}(b - Ax_0)}{\beta} \right\|_2 \leq \\ &\leq \|b - Ax_0 - \mathrm{fl}(b - Ax_0)\|_2 + \epsilon\|\mathrm{fl}(b - Ax_0)\|_2 \leq \\ &\leq (1+\epsilon)\|b - Ax_0 - \mathrm{fl}(b - Ax_0)\|_2 + \epsilon\|b - Ax_0\|_2 \leq \\ &\leq (1+\epsilon)[\|\mathrm{fl}(b - Ax_0) - (b - \mathrm{fl}(Ax_0))\|_2 + \\ &\quad +\|\mathrm{fl}(Ax_0) - Ax_0\|_2] + \epsilon\|b - Ax_0\|_2 \leq \\ &\leq (1+\epsilon)[\epsilon(\|b\|_2 + \|\mathrm{fl}(Ax_0)\|_2) + \|\mathrm{fl}(Ax_0) - Ax_0\|_2] + \\ &\quad +\epsilon\|b - Ax_0\|_2 \leq \\ &\leq (1+\epsilon)[\epsilon\|b\|_2 + (1+\epsilon)\|\mathrm{fl}(Ax_0) - Ax_0\|_2 + \epsilon\|Ax_0\|_2] + \\ &\quad +\|b - Ax_0\|_2 \leq \\ &\leq (1+\epsilon)[\epsilon\|b\|_2 + (1+\epsilon)m\sqrt{n}(\epsilon + \mathcal{O}(\epsilon^2))\|A\|_2\|x_0\|_2 + \\ &\quad +\epsilon\|A\|_2\|x_0\|_2] + \epsilon\|b\|_2 + \epsilon\|A\|_2\|x_0\|_2 \leq \\ &\leq (mn^{1/2} + 2)\epsilon\|A\|_2\|x_0\|_2 + 2\epsilon\|b\|_2 + \\ &\quad +\mathcal{O}(\epsilon^2)(mn^{1/2}\|A\|_2\|x_0\|_2 + \|b\|_2), \end{aligned} \quad (2.195)$$

$$\begin{aligned} \|A\Delta x_k\|_2 &\leq \epsilon\|A\|_2\|x_0\|_2 + 2k^{3/2}(\epsilon + n\mathcal{O}(\epsilon^2))\|A\|_2\|y_k\|_2, \\ \|F_k y_k\|_2 &\leq (c_5 k^{5/2}n^{1/2} + mk^{1/2}n^{1/2})(\epsilon + n\mathcal{O}(\epsilon^2))\|A\|_2\|y_k\|_2, \end{aligned}$$

što daje, prema (2.194) ocjenu

$$\begin{aligned} &\|(b - Ax_k) - Q_{k+1}(\beta\xi_1 - H_{k+1,k}y_k)\|_2 \leq \\ &\leq (2k^{3/2} + c_5 k^{5/2}n^{1/2} + mk^{1/2}n^{1/2})\epsilon\|A\|_2\|y_k\|_2 + \\ &\quad +(mn^{1/2} + 3)\epsilon\|A\|_2\|x_0\|_2 + 2\epsilon\|b\|_2 + \\ &\quad +\mathcal{O}(\epsilon^2)(\|A\|_2\|y_k\|_2 + \|A\|_2\|x_0\|_2 + \|b\|_2). \end{aligned} \quad (2.196)$$

Uz pretpostavku da $\|x_0\|_2$ nije ekstremno velik, izraz $(mn^{1/2} + 3)\epsilon\|A\|_2\|x_0\|_2 + 2\epsilon\|b\|_2$ nije od presudnog značaja u prethodnoj ocjeni. On može postati velik jedino ako $\|y_k\|_2$ poćne nekontrolirano rasti. Iz (2.196) i ocjene za normu od Q_{k+1} imamo konačnu ocjenu

$$\|b - Ax_k\|_2 \leq (k+1)^{1/2}\left(1 + \frac{n+4}{2}\epsilon\right)\|\beta\xi_1 - H_{k+1,k}y_k\|_2 +$$

$$+(2k^{3/2}+c_5k^{5/2}n^{1/2}+mk^{1/2}n^{1/2})\epsilon\|A\|_2\|y_k\|_2+$$
$$+(mn^{1/2}+3)\epsilon\|A\|_2\|x_0\|_2+2\epsilon\|b\|_2+$$
$$+\mathcal{O}(\epsilon^2)(\|A\|_2\|y_k\|_2+\|A\|_2\|x_0\|_2+\|b\|_2). \quad (2.197)$$

U praksi, mi naravno nemamo $\|\beta\xi_1 - H_{k+1,k}y_k\|_2$ već izračunatu normu Arnoldijevog reziduala, pa sada još moramo ustanoviti odnos između te dvije veličine. Izračunata norma Arnoldijevog reziduala je zapravo apsolutna vrijednost $(k+1)$-e komponente vektora $g^{(k)} = \mathrm{fl}(F^{(k)}(\beta\xi_1))$ dobivenog primjenom Givensovih rotacija, koje Hessenbergovu matricu $H_{k+1,k}$ svode na gornje trokutasti oblik, u aritmetici konačne preciznosti. Zbog toga bi trebali ocjenu norme pravog reziduala izraziti preko $|g^{(k)}_{k+1}|$, a ne $\|\beta\xi_1 - H_{k+1,k}y_k\|_2$. Za nastavak analize trebat će nam dva teorema i jedna lema iz [21, str. 154, 373, 375]. Lema 18.7 daje ocjenu povratne greške djelovanja jedne Givensove rotacije na vektor. Budući da je $F^{(k)}$ produkt od k Givensovih rotacija F_i, ovu lemu primjenjujemo k puta da bi dobili ocjenu za $g^{(k)}$. Imamo

$$\begin{aligned} g^{(k)} &= (F_k^r + \Delta F_k^r)\cdots(F_1^r + \Delta F_1^r)(\beta\xi_1) = \\ &= (F_k^r\cdots F_1^r + \Delta F_k^r F_{k-1}^r\cdots F_1^r + \cdots + F_k^r\cdots F_2^r\Delta F_1^r + \Delta F^{(k)})(\beta\xi_1) \end{aligned}$$

pri čemu su F_i^r, $i = 1,\ldots,k$ egzaktne $(k+1)\times(k+1)$ Givensove rotacije, a Frobeniusova norma od $\Delta F^{(k)}$ je reda $k\mathcal{O}(\epsilon^2)$. Prema Lemi 18.7 vrijedi

$$\begin{aligned} \|g^{(k)}\|_2 &\le \beta(1+\|\Delta F_k\|_2+\cdots+\|\Delta F_1\|_2+k\mathcal{O}(\epsilon^2)) \le \\ &\le \beta[1+6\sqrt{2}k(\epsilon+\mathcal{O}(\epsilon^2))]. \end{aligned}$$

Teorem 8.5 daje ocjenu povratne greške rješavanja trokutastog sustava. Njega primjenjujemo u procesu dobivanja vektora y_k, koji se dobiva kao rješenje gornje trokutastog sustava $R_k y_k = g^{(k)}_{k\times 1}$ u aritmetici konačne preciznosti, gdje je $g^{(k)}_{k\times 1}$ vektor sastavljen od prvih k komponenti vektora $g^{(k)}$. Vrijedi

$$(R_k + \Delta R_k)y_k = g^{(k)}_{k\times 1},$$

gdje je

$$\|\Delta R_k\|_2 \le k^{3/2}(\epsilon+\mathcal{O}(\epsilon^2))\|R_k\|_2.$$

I naposlijetku nam treba Teorem 18.9 koji kaže da postoji ortogonalna $(k+1)\times(k+1)$ matrica $\widehat{F}^{(k)}$ takva da je

$$H_{k+1,k} + \Delta H_{k+1,k} = \widehat{F}^{(k)T}\begin{bmatrix} R_k \\ 0 \end{bmatrix},$$

gdje je

$$\|\Delta H_{k+1,k}\|_2 \le c_{10}k^{3/2}(\epsilon+\mathcal{O}(\epsilon^2))\|H_{k+1,k}\|_2,$$

za neku konstantu $c_{10} \ge 1$, a $\widehat{F}^{(k)}$ je produkt egzaktnih Givensovih rotacija F_i^r. Ovo možemo na drugačiji način napisati kao

$$\widehat{F}^{(k)}H_{k+1,k} = \begin{bmatrix} R_k \\ 0 \end{bmatrix} + \Delta\widehat{R}_k,$$

sa

$$\|\Delta\widehat{R}_k\|_2 = \|\widehat{F}^{(k)}\Delta H_{k+1,k}\|_2 \le c_{10}k^{3/2}(\epsilon+\mathcal{O}(\epsilon^2))\|H_{k+1,k}\|_2.$$

Prema Lemi 18.7 još vrijedi da je

$$\|g^{(k)} - \widehat{F}^{(k)}(\beta\xi_1)\|_2 \leq 6\sqrt{2}k\beta(\epsilon + \mathcal{O}(\epsilon^2)).$$

Sada se konačno vračamo na Arnoldijev rezidual. Imamo

$$\begin{aligned}
\|\beta\xi_1 - H_{k+1,k}y_k\|_2 &= \|\widehat{F}^{(k)}(H_{k+1,k}y_k - \beta\xi_1)\|_2 \leq \\
&\leq \left\| \begin{bmatrix} R_k \\ 0 \end{bmatrix} y_k + \Delta\widehat{R}_k y_k - g^{(k)} \right\|_2 + \|g^{(k)} - \widehat{F}^{(k)}(\beta\xi_1)\|_2 \leq \\
&\leq \|R_k y_k - g^{(k)}_{k\times 1}\|_2 + |g^{(k)}_{k+1}| + \|\Delta\widehat{R}_k\|_2\|y_k\|_2 + \\
&\quad +6\sqrt{2}k\beta(\epsilon + \mathcal{O}(\epsilon^2)) \leq \\
&\leq \|\Delta R_k\|_2\|y_k\|_2 + |g^{(k)}_{k+1}| + \|\Delta\widehat{R}_k\|_2\|y_k\|_2 + \\
&\quad +6\sqrt{2}k\beta(\epsilon + \mathcal{O}(\epsilon^2)) \leq \\
&\leq |g^{(k)}_{k+1}| + k^{3/2}(\epsilon + \mathcal{O}(\epsilon^2))\|R_k\|_2\|y_k\|_2 + \\
&\quad +c_{10}k^{3/2}(\epsilon + \mathcal{O}(\epsilon^2))\|H_{k+1,k}\|_2\|y_k\|_2 + 6\sqrt{2}k\beta(\epsilon + \mathcal{O}(\epsilon^2)).
\end{aligned}$$

Na isti način kao što smo dobili ocjenu za Δx_k, možemo dobiti i ocjenu za $\beta = \mathrm{fl}(b - Ax_0)$,

$$\beta \leq \|A\|_2\|x_0\|_2 + \|b\|_2 + \mathcal{O}(\epsilon)(n\|A\|_2\|x_0 + \|b\|_2).$$

Tako, dobivamo da je

$$\begin{aligned}
\|\beta\xi_1 - H_{k+1,k}y_k\|_2 &\leq |g^{(k)}_{k+1}| + k^{3/2}\epsilon(\|R_k\|_2 + c_{10}\|H_{k+1,k}\|_2)\|y_k\|_2 + \\
&\quad +6\sqrt{2}k\epsilon(\|A\|_2\|x_0\|_2 + \|b\|_2) + \mathcal{O}(\epsilon^2)(\|R_k\|_2\|y_k\|_2 + \\
&\quad +\|H_{k+1,k}\|_2\|y_k\|_2 + \|A\|_2\|x_0\|_2 + \|b\|_2).
\end{aligned}$$

Na kraju analize točnosti GMRES metode napokon možemo dati odnos između norme pravog reziduala i izračunate norme Arnoldijevog reziduala u k-toj iteracije.

$$\begin{aligned}
\|b - Ax_k\|_2 &\leq (k+1)^{1/2}\left(1 + \frac{n+4}{2}\epsilon\right)|g^{(k)}_{k+1}| + \\
&\quad +(2k^{3/2} + c_5k^{5/2}n^{1/2} + mk^{1/2}n^{1/2})\epsilon\|A\|_2\|y_k\|_2 + \\
&\quad +c_{11}k^2\epsilon(\|R_k\|_2 + c_{10}\|H_{k+1,k}\|_2)\|y_k\|_2 + \\
&\quad +(mn^{1/2} + c_{12}k^{3/2} + 3)\epsilon\|A\|_2\|x_0\|_2 + (c_{12}k^{3/2} + 2)\epsilon\|b\|_2 + \\
&\quad +\mathcal{O}(\epsilon^2)[(\|A\|_2 + \|R_k\|_2 + \|H_{k+1,k}\|_2)\|y_k\|_2 + \|A\|_2\|x_0\|_2 + \|b\|_2],
\end{aligned}$$

za neke konstante c_{11} i c_{12}.

QMR metoda

Analiza greške QMR metode slična je analizi GMRES metode, osim što se aproksimacija x_k u k-toj iteraciji računa rekurzivno, preko aproksimacije x_{k-1}, a ne direktno iz x_0. Prva faza QMR algoritma je dvostrani Lanczosov algoritam, kojeg najprije promatramo. U njemu se vektori v_i dobivaju normalizacijom vektora $\tilde{v}_i$, pa i za njih vrijedi da je

$$\|v_i\|_2^2 \leq 1 + (n+4)\epsilon + n\mathcal{O}(\epsilon^2).$$

Nadalje, promatramo postupak za dobivanje vektora $\tilde{v}_i$, koji je jednak za sve vektore osim za slučaj $i = 2$, kojeg ćemo najprije analizirati. U aritmetici konačne preciznosti definirajmo

$$\tilde{v}_2 = Av_1 - \alpha_1 v_1 + \Delta\tilde{v}_2,$$

pa prema (2.138), (2.140) za $c = m\sqrt{n}$, gdje kao i kod analize GMRES metode, m najveći broj netrivijalnih elemenata u retku matrice A, te (2.137), vrijedi

$$\begin{aligned}
\|\Delta\tilde{v}_2\|_2 &= \|\mathrm{fl}(Av_1 - \alpha_1 v_1) - (Av_1 - \alpha_1 v_1)\|_2 \leq \\
&\leq \|\mathrm{fl}(Av_1 - \alpha_1 v_1) - (\mathrm{fl}(Av_1) - \mathrm{fl}(\alpha_1 v_1))\|_2 + \|\mathrm{fl}(Av_1) - Av_1\|_2 + \\
&\quad + \|\mathrm{fl}(\alpha_1 v_1) - \alpha_1 v_1\|_2 \leq \\
&\leq \epsilon(\|\mathrm{fl}(Av_1)\|_2 + \|\mathrm{fl}(\alpha_1 v_1)\|_2) + \|\mathrm{fl}(Av_1) - Av_1\|_2 + \|\mathrm{fl}(\alpha_1 v_1) - \alpha_1 v_1\|_2 \leq \\
&\leq \epsilon(\|Av_1\|_2 + \|\alpha_1 v_1\|_2) + (1+\epsilon)(\|\mathrm{fl}(Av_1) - Av_1)\|_2 + \|\mathrm{fl}(\alpha_1 v_1) - \alpha_1 v_1\|_2) \leq \\
&\leq \epsilon(\|A\|_2\|v_1\|_2 + |\alpha_1|\|v_1\|_2) + \\
&\quad + (1+\epsilon)[m\sqrt{n}(\epsilon + \mathcal{O}(\epsilon^2))\|A\|_2\|v_1\|_2 + \epsilon|\alpha_1|\|v_1\|_2] \leq \\
&\leq [(1+m\sqrt{n})\epsilon + m\sqrt{n}\mathcal{O}(\epsilon^2)]\|A\|_2\|v_1\|_2 + (2\epsilon + \mathcal{O}(\epsilon^2))|\alpha_1|\|v_1\|_2 \leq \\
&\leq [(1+m\sqrt{n})\epsilon\|A\|_2 + 2\epsilon|\alpha_1| + \mathcal{O}(\epsilon^2)(m\sqrt{n}\|A\|_2 + \|T_{k+1,k}\|_1)] \cdot \\
&\quad \cdot \left(1 + \frac{n+4}{2}\epsilon + n\mathcal{O}(\epsilon^2)\right) \leq \\
&\leq (1+m\sqrt{n})\epsilon\|A\|_2 + 2\epsilon|\alpha_1| + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1).
\end{aligned}$$

Kako v_2 dobivamo dijeljenjem $\tilde{v}_2$ sa γ_1, prema (2.136), nadalje vrijedi

$$v_2 = \mathrm{fl}\left(\frac{\tilde{v}_2}{\gamma_1}\right) = \left(I + \begin{bmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_n \end{bmatrix}\right)\frac{\tilde{v}_2}{\gamma_1},$$

pri čemu je $|\delta_i| \leq \epsilon$ za $i = 1, \ldots, n$. Iz prethodnog slijedi

$$\begin{aligned}
\gamma_1 v_2 &= \left(I + \begin{bmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_n \end{bmatrix}\right)\tilde{v}_2 = Av_1 - \alpha_1 v_1 + \\
&\quad + \begin{bmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_n \end{bmatrix}(Av_1 - \alpha_1 v_1) + \left(I + \begin{bmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_n \end{bmatrix}\right)\Delta\tilde{v}_2 = \\
&= Av_1 - \alpha_1 v_1 - f_1,
\end{aligned}$$

gdje je norma od f_1 ograđena sa

$$\begin{aligned}
\|f_1\|_2 &\leq \epsilon(\|A\|_2 + |\alpha_1|)\left(1 + \frac{n+4}{2}\epsilon + n\mathcal{O}(\epsilon^2)\right) + \\
&\quad + (1+\epsilon)[(1+m\sqrt{n})\epsilon\|A\|_2 + 2\epsilon|\alpha_1| + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1)] \leq \\
&\leq \epsilon[(2+m\sqrt{n})\|A\|_2 + 3|\alpha_1|] + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1). \qquad (2.198)
\end{aligned}$$

Dakle za $i = 2$ vrijedi

$$Av_1 = \alpha_1 v_1 + \gamma_1 v_2 + f_1, \qquad (2.199)$$

uz ocjenu (2.198). Sličnu analizu radimo i za $i > 2$, jedina razlika je ta što rekurzija za $\tilde{v}_i$ sadrži tri člana. U aritmetici konačne preciznosti imamo

$$\tilde{v}_{i+1} = Av_i - \alpha_i v_i - \beta_{i-1} v_{i-1} + \Delta\tilde{v}_{i+1},$$

za koje ponovo prema (2.138), (2.140), (2.137) i ocjeni za $\|\Delta\tilde{v}_2\|_2$, vrijedi

$$\begin{aligned}
\|\Delta\tilde{v}_{i+1}\|_2 &= \|\mathrm{fl}(Av_i - \alpha_i v_i - \beta_{i-1}v_{i-1}) - (Av_i - \alpha_i v_i - \beta_{i-1}v_{i-1})\|_2 \leq \\
&\leq \|\mathrm{fl}(Av_i - \alpha_i v_i - \beta_{i-1}v_{i-1}) - [\mathrm{fl}(Av_i - \alpha_i v_i) - \mathrm{fl}(\beta_{i-1}v_{i-1})]\|_2 + \\
&\quad + \|\mathrm{fl}(Av_i - \alpha_i v_i) - (Av_i - \alpha_i v_i)\|_2 + \|\mathrm{fl}(\beta_{i-1}v_{i-1}) - \beta_{i-1}v_{i-1}\|_2 \leq \\
&\leq \epsilon(\|\mathrm{fl}(Av_i - \alpha_i v_i)\|_2 + \|\mathrm{fl}(\beta_{i-1}v_{i-1})\|_2\|_2) + \\
&\quad + \|\mathrm{fl}(Av_i - \alpha_i v_i) - (Av_i - \alpha_i v_i)\|_2 + \|\mathrm{fl}(\beta_{i-1}v_{i-1}) - \beta_{i-1}v_{i-1}\|_2 \leq \\
&\leq \epsilon(\|Av_i - \alpha_i v_i\|_2 + \|\beta_{i-1}v_{i-1}\|_2) + (1+\epsilon)\cdot \\
&\quad \cdot[\|\mathrm{fl}(Av_i - \alpha_i v_i) - (Av_i - \alpha_i v_i)\|_2 + \|\mathrm{fl}(\beta_{i-1}v_{i-1}) - \beta_{i-1}v_{i-1}\|_2] \leq \\
&\leq \epsilon(\|A\|_2 + |\alpha_i| + |\beta_{i-1}|)\left(1 + \frac{n+4}{2}\epsilon + n\mathcal{O}(\epsilon^2)\right) + \\
&\quad + (1+\epsilon)[(1+m\sqrt{n})\epsilon\|A\|_2 + 2\epsilon|\alpha_i| + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1) + \\
&\quad + \epsilon|\beta_{i-1}|\left(1 + \frac{n+4}{2}\epsilon + n\mathcal{O}(\epsilon^2)\right)] \leq \\
&\leq (2 + m\sqrt{n})\epsilon\|A\|_2 + 3\epsilon|\alpha_i| + 2\epsilon|\beta_{i-1}| + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1).
\end{aligned}$$

Nastavljamo kao i za v_2,

$$v_{i+1} = \mathrm{fl}\left(\frac{\tilde{v}_{i+1}}{\gamma_i}\right) = \left(I + \begin{bmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_n \end{bmatrix}\right)\frac{\tilde{v}_{i+1}}{\gamma_i},$$

za $\delta_i \leq \epsilon$, $i = 1, \ldots, n$, pa je zato

$$\begin{aligned}
\gamma_i v_{i+1} &= \left(I + \begin{bmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_n \end{bmatrix}\right)\tilde{v}_{i+1} = Av_i - \alpha_i v_i - \beta_{i-1}v_{i-1} + \\
&\quad + \begin{bmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_n \end{bmatrix}(Av_i - \alpha_i v_i - \beta_{i-1}v_{i-1}) + \left(I + \begin{bmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_n \end{bmatrix}\right)\Delta\tilde{v}_{i+1} = \\
&= Av_i - \alpha_i v_i - \beta_{i-1}v_{i-1} - f_i,
\end{aligned}$$

pri čemu vrijedi ocjena

$$\begin{aligned}
\|f_i\|_2 &\leq \epsilon(\|A\|_2 + |\alpha_i| + |\beta_{i-1}|)\left(1 + \frac{n+4}{2}\epsilon + n\mathcal{O}(\epsilon^2)\right) + \\
&\quad + (1+\epsilon)[(2+m\sqrt{n})\epsilon\|A\|_2 + 3\epsilon|\alpha_i| + 2\epsilon|\beta_{i-1}| + \\
&\quad + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1)] \leq \\
&\leq \epsilon[(3+m\sqrt{n})\|A\|_2 + 4|\alpha_i| + 3|\beta_{i-1}|] + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1) \quad (2.200)
\end{aligned}$$

Dakle, i za $i > 1$ vrijedi

$$Av_i = \beta_{i-1}v_{i-1} + \alpha_i v_i + \gamma_i v_{i+1} + f_i, \quad (2.201)$$

uz ocjenu (2.200). Sada ako definiramo matricu F_k kao

$$F_k = [f_1 \ f_2 \ \dots \ f_k],$$

tada dvostrani Lanczosov algoritam, izveden u aritmetici konačne preciznosti, možemo napisati u matričnom obliku, na sljedeći način

$$AV_k = V_{k+1}T_{k+1,k} + F_k, \tag{2.202}$$

pri čemu je $T_{k+1,k}$ tridijagonalna matrica, kod koje je i-ti stupac označen sa $(T_{k+1,k})_i$, a ocjena za normu matrice F_k dana je sa

$$\begin{aligned}
\|F_k\|_2 &\leq \|F_k\|_F = \left(\sum_{i=1}^{k}\|f_i\|_2^2\right)^{1/2} \leq \\
&\leq \Big(\{\epsilon[(2+m\sqrt{n})\|A\|_2 + 3|\alpha_1|] + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1)\}^2 + \\
&\quad + \sum_{i=2}^{k}\{\epsilon[(3+m\sqrt{n})\|A\|_2 + 4|\alpha_i| + 3|\beta_{i-1}|] + \\
&\quad + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1)\}^2\Big)^{1/2} \leq \\
&\leq \Big(\{\epsilon[(3+m\sqrt{n})\|A\|_2 + 4(|\alpha_1| + |\gamma_1|)] + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1)\}^2 + \\
&\quad + \sum_{i=2}^{k}\{\epsilon[(3+m\sqrt{n})\|A\|_2 + 4(|\beta_{i-1}| + |\alpha_i| + |\gamma_i|)] + \\
&\quad + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1)\}^2\Big)^{1/2} \leq \\
&\leq \Big(\{\epsilon[(3+m\sqrt{n})\|A\|_2 + 4\|(T_{k+1,k})_1\|_1] + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1)\}^2 + \\
&\quad + \sum_{i=2}^{k}\{\epsilon[(3+m\sqrt{n})\|A\|_2 + 4\|(T_{k+1,k})_i\|_1] + \\
&\quad + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1)\}^2\Big)^{1/2} \leq \\
&\leq \Big(k\{\epsilon[(3+m\sqrt{n})\|A\|_2 + 4\|T_{k+1,k}\|_1] + \\
&\quad + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1)\}^2\Big)^{1/2} \leq \\
&\leq \epsilon\sqrt{k}[(3+m\sqrt{n})\|A\|_2 + 4\|T_{k+1,k}\|_1] + \mathcal{O}(\epsilon^2)(\|A\|_2 + \|T_{k+1,k}\|_1).
\end{aligned} \tag{2.203}$$

Budući da nakon koraka dvostrukog Lanczosovog algoritma slijedi Givensova QR faktorizacija matrice $T_{k+1,k}$, za nastavak analize greške QMR metode ponovo nam je potreban Teorem 18.9 iz [21, str. 375] koji tvrdi da postoji ortogonalna $(k+1)\times(k+1)$ matrica $\widehat{F}^{(k)}$ takva da je

$$\widehat{F}^{(k)}T_{k+1,k} = \begin{bmatrix} R_k \\ 0 \end{bmatrix} + \Delta\widehat{R}_k, \tag{2.204}$$

pri čemu je

$$\begin{aligned}\|\Delta\widehat{R}_k\|_F &\leq 6\sqrt{2}k(\epsilon+\mathcal{O}(\epsilon^2))\|T_{k+1,k}\|_F \leq \\ &\leq 9k(\epsilon+\mathcal{O}(\epsilon^2))\|T_{k+1,k}\|_F. \end{aligned} \tag{2.205}$$

Sljedeći korak je računanje matrice $P_k = \mathrm{fl}(V_k R_k^{-1})$, koja se zapravo računa povratnim supstitucijama, kao rješenja trokutastog sustava. Naime, neka su p^i i v^i stupci matrica P_k^T i V_k^T, tada prema Teoremu 8.5 iz [21, str. 154] za svako i vrijedi da je

$$(R_k^T + \Delta R_i^T)p^i = v^i,$$

i, budući da R_k^T u svakom retku ima najviše 3 netrivijalna elementa, vrijedi ocjena

$$|\Delta R_i^T| \leq (3\epsilon+\mathcal{O}(\epsilon^2))|R_k^T|.$$

Prema tome, imamo

$$R_k^T p^i = v^i + \Delta v^i,$$

sa

$$|\Delta v^i| \leq (3\epsilon+\mathcal{O}(\epsilon^2))|R_k^T||p^i|,$$

odnosno

$$P_k R_k = V_k + \Delta V_k,$$

sa

$$|\Delta V_k| \leq (3\epsilon+\mathcal{O}(\epsilon^2))|P_k||R_k|.$$

Napokon, promatramo grešku koja se javlja koda same matrice P_k,

$$\begin{aligned} P_k &= V_k R_k^{-1} + \Delta V_k R_k^{-1} = \\ &= V_k R_k^{-1} + \Delta P_k, \end{aligned} \tag{2.206}$$

sa ocjenom

$$\begin{aligned}\|\Delta P_k\|_2 &\leq \|\Delta V_k\|_F \|R_k^{-1}\|_2 \leq \\ &\leq 3\sqrt{k}(\epsilon+\mathcal{O}(\epsilon^2))\|P_k\|_F\kappa(R_k). \end{aligned} \tag{2.207}$$

Na kraju iteracije QMR algoritma provjerava se da li je kvazirezidual dostigao određenu granicu, što je u egzaktnoj aritmetici ekvivalentno provjeravanju apsolutne vrijednosti $(k+1)$-ve komponente vektora $F^{(k)}(\beta\xi_1)$. U aritmetici konačne preciznosti, na kraju svake iteracije provjerava se $(k+1)$-va komponenta vektora $g^{(k)} = \mathrm{fl}(F^{(k)}(\beta\xi_1))$, za kojeg, ekvivalentno kao i kod GMRES metode vrijedi,

$$\|g^{(k)}\|_2 \leq \beta[1+6\sqrt{2}k(\epsilon+\mathcal{O}(\epsilon^2))].$$

Preostaje nam još naći gornju ogradu za $\beta = \mathrm{fl}(\|b-Ax_0\|_2)$. Neka je $z = \mathrm{fl}(b-Ax_0)$, tada prema (2.138) i (2.140) imamo

$$\begin{aligned}\|z-(b-Ax_0)\|_2 &\leq \|\mathrm{fl}(b-Ax_0)-[b-\mathrm{fl}(Ax_0)]\|_2 + \|Ax_0-\mathrm{fl}(Ax_0)\|_2 \leq \\ &\leq \epsilon(\|b\|_2+\|\mathrm{fl}(Ax_0)\|_2)+\|Ax_0-\mathrm{fl}(Ax_0)\|_2 \leq \\ &\leq \epsilon(\|b\|_2+\|Ax_0\|_2)+(1+\epsilon)\|Ax_0-\mathrm{fl}(Ax_0)\|_2 \leq \\ &\leq \epsilon(\|b\|_2+\|A\|_2\|x_0\|_2)+ \\ &\quad +(1+\epsilon)m\sqrt{n}(\epsilon+\mathcal{O}(\epsilon^2))\|A\|_2\|x_0\|_2 \leq \\ &\leq \epsilon\|b\|_2+(1+m\sqrt{n})\epsilon\|A\|_2\|x_0\|_2+\mathcal{O}(\epsilon^2)\|A\|_2\|x_0\|_2. \end{aligned}$$

U daljnjem postupku za dobivanje konstante β vektor z sudjeluje u skalarnom produktu, koji se nakon toga korjenuje, stoga imamo

$$\begin{aligned} |\mathrm{fl}(z^Tz) - z^Tz| &\leq n(\epsilon + \mathcal{O}(\epsilon^2))\|z\|_2^2, \\ |\mathrm{fl}(\sqrt{\mathrm{fl}(z^Tz)}) - \sqrt{\mathrm{fl}(z^Tz)}| &\leq \epsilon\sqrt{\mathrm{fl}(z^Tz)} \leq \epsilon\sqrt{[1+n(\epsilon+\mathcal{O}(\epsilon^2)]\|z\|_2^2} \leq \\ &\leq \epsilon\left[1+\frac{n}{2}(\epsilon+\mathcal{O}(\epsilon^2))\right]\|z\|_2 \leq (\epsilon + n\mathcal{O}(\epsilon^2))\|z\|_2. \end{aligned}$$

Sada možemo analizirati odnos između β i $\|b - Ax_0\|_2$.

$$\begin{aligned} |\beta - \|b-Ax_0\|_2| &\leq |\beta - \sqrt{\mathrm{fl}(z^Tz)}| + |\sqrt{\mathrm{fl}(z^Tz)} - \|z\|_2| + |\|z\|_2 - \|b-Ax_0\|_2| \leq \\ &\leq (\epsilon + n\mathcal{O}(\epsilon^2))\|z\|_2 + \frac{n}{2}(\epsilon+\mathcal{O}(\epsilon^2))\|z\|_2 + \|z-(b-Ax_0)\|_2 \leq \\ &\leq \left[\left(1+\frac{n}{2}\right)\epsilon + n\mathcal{O}(\epsilon^2)\right]\|z\|_2 + \|z-(b-Ax_0)\|_2 \leq \\ &\leq \frac{n+2}{2}(\epsilon+\mathcal{O}(\epsilon^2))\|b-Ax_0\|_2 + \left(1+\frac{n+2}{2}(\epsilon+\mathcal{O}(\epsilon^2)\right)\cdot \\ &\quad \cdot\|z-(b-Ax_0)\|_2 \leq \\ &\leq \frac{n+2}{2}(\epsilon+\mathcal{O}(\epsilon^2))(\|b\|_2 + \|A\|_2\|x_0\|_2) + \\ &\quad + \left[1+\frac{n+2}{2}(\epsilon+\mathcal{O}(\epsilon^2))\right]\cdot \\ &\quad \cdot[\epsilon\|b\|_2 + (1+m\sqrt{n})\epsilon\|A\|_2\|x_0\|_2 + \mathcal{O}(\epsilon^2)\|A\|_2\|x_0\|_2] \leq \\ &\leq \frac{n+4}{2}\epsilon\|b\|_2 + \frac{n+2m\sqrt{n}+4}{2}\epsilon\|A\|_2\|x_0\|_2 + \\ &\quad + \mathcal{O}(\epsilon^2)(\|A\|_2\|x_0\|_2 + \|b\|_2), \end{aligned}$$

odakle je

$$\begin{aligned} \beta &\leq \|b-Ax_0\|_2 + |\beta - \|b-Ax_0\|_2| \leq \\ &\leq \left(1+\frac{n+4}{2}\epsilon\right)\|b\|_2 + \left(1+\frac{n+2m\sqrt{n}+4}{2}\epsilon\right)\|A\|_2\|x_0\|_2 + \\ &\quad + \mathcal{O}(\epsilon^2)(\|A\|_2\|x_0\|_2 + \|b\|_2). \end{aligned} \tag{2.208}$$

Ograda za β bila nam je potrebna da dovršimo ocjenu za $\|g^{(k)}\|_2$, pa zato slijedi

$$\begin{aligned} \|g^{(k)}\|_2 &\leq \left[\left(1+\frac{n+4}{2}\epsilon\right)\|b\|_2 + \left(1+\frac{n+2m\sqrt{n}+4}{2}\epsilon\right)\|A\|_2\|x_0\|_2 + \right. \\ &\quad \left. + \mathcal{O}(\epsilon^2)(\|A\|_2\|x_0\|_2 + \|b\|_2)\right][1+6\sqrt{2}k(\epsilon+\mathcal{O}(\epsilon^2))] \leq \\ &\leq \left(1+\frac{n+12\sqrt{2}k+4}{2}\epsilon\right)\|b\|_2 + \left(1+\frac{n+2m\sqrt{n}+12\sqrt{2}k+4}{2}\epsilon\right)\cdot \\ &\quad \cdot\|A\|_2\|x_0\|_2 + \mathcal{O}(\epsilon^2)(\|A\|_2\|x_0\|_2 + \|b\|_2) \end{aligned} \tag{2.209}$$

I na samom kraju svake iteracije algoritma izračunava se aproksimacija rješenja, tekuće iteracije, čije je oblik u aritmetici konačne preciznosti

$$x_k = x_{k-1} + g_k^{(k)}p_{k-1} + \Delta x_k, \tag{2.210}$$

i koji je istovjetan obliku (2.141), pa zato možemo koristiti rezultat (2.151) iz Leme 2.10.1 uz uvjet da je $1 - 2\epsilon - \epsilon^2 > 0$, koji daje ocjenu

$$\|\Delta x_k\|_2 \leq \epsilon(3\|x_{k-1}\|_2 + 2\|x_k\|_2) + \mathcal{O}(\epsilon^2)(\|x_{k-1}\|_2 + \|x_k\|_2).$$

Ako definiramo

$$\chi_k = \max_{i \leq k} \|x_i\|_2, \tag{2.211}$$

tada ocjena norme vektora Δx_k glasi

$$\|\Delta x_k\|_2 \leq (5\epsilon + \mathcal{O}(\epsilon^2))\chi_k. \tag{2.212}$$

Izvršavanjem rekurzije (2.210) do kraja, dobivamo konačni oblik za aproksimaciju rješenja

$$x_k = x_0 + P_k g_{k\times 1}^{(k)} + \sum_{i=1}^{k} \Delta x_i, \tag{2.213}$$

gdje je $g_{k\times 1}^{(k)}$ k-dimenzionalni vektor sačinjen od prvih k komponenti vektora $g^{(k)}$. U tom slučaju je

$$\sum_{i=1}^{k} \|\Delta x_i\|_2 \leq (5\epsilon + \mathcal{O}(\epsilon^2))k\chi_k. \tag{2.214}$$

Ovime smo napravili ocjene svih elemenata koji će se pojaviti u analizi odnosa pravog reziduala i kvazireziduala, računatih u aritmetici konačne preciznosti. Ta analiza je vrlo slična onoj za GMRES metodu, osim u činjenici da kod QMR metode mi niti ne očekujemo da matrica V_k bude blizu ortonormalnoj matrici. Imamo

$$\begin{aligned}
&b - Ax_k - V_{k+1}(\beta\xi_1 - T_{k+1,k}R_k^{-1}g_{k\times 1}^{(k)}) = \\
&\quad = \quad b - Ax_0 - AP_k g_{k\times 1}^{(k)} - \sum_{i=1}^{k} A\Delta x_i - \beta v_1 + V_{k+1}T_{k+1,k}R_k^{-1}g_{k\times 1}^{(k)} = \\
&\quad = \quad (b - Ax_0 - \beta v_1) - AP_k g_{k\times 1}^{(k)} + AV_k R_k^{-1} g_{k\times 1}^{(k)} - F_k R_k^{-1} g_{k\times 1}^{(k)} - \sum_{i=1}^{k} A\Delta x_i = \\
&\quad = \quad (b - Ax_0 - \beta v_1) - A\Delta P_k g_{k\times 1}^{(k)} - F_k R_k^{-1} g_{k\times 1}^{(k)} - \sum_{i=1}^{k} A\Delta x_i,
\end{aligned}$$

odakle je

$$\begin{aligned}
\|b - Ax_k\|_2 \quad \leq \quad & \|V_{k+1}\|_2 \|\beta\xi_1 - T_{k+1,k}R_k^{-1}g_{k\times 1}^{(k)}\|_2 + \|b - Ax_0 - \beta v_1\|_2 + \\
& + \|A\|_2 \|\Delta P_k\|_2 \|g^{(k)}\|_2 + \|F_k\|_2 \|R_k^{-1}\|_2 \|g^{(k)}\|_2 + \\
& + \|A\|_2 \left(\sum_{i=1}^{k} \|\Delta x_i\|_2 \right).
\end{aligned} \tag{2.215}$$

Prema (2.195), (2.207), (2.203), (2.209) i (2.214) dalje slijedi

$$\|b - Ax_k\|_2 \leq$$

$$
\begin{aligned}
\leq\ & \sqrt{k+1}\left(1+\frac{n+4}{2}\epsilon+n\mathcal{O}(\epsilon^2)\right)\|\beta\xi_1-T_{k+1,k}R_k^{-1}g_{k\times 1}^{(k)}\|_2+\\
&+(m\sqrt{n}+2)\epsilon\|A\|_2\|x_0\|_2+2\epsilon\|b\|_2+\mathcal{O}(\epsilon^2)(m\sqrt{n}\|A\|_2\|x_0\|_2+\|b\|_2)\\
&+\|A\|_2[3\sqrt{k}(\epsilon+\mathcal{O}(\epsilon^2))\|P_k\|_F\kappa(R_k)]\cdot\\
&\cdot\left[\left(1+\frac{n+12\sqrt{2}k+4}{2}\epsilon\right)\|b\|_2+\left(1+\frac{n+2m\sqrt{n}+12\sqrt{2}k+4}{2}\epsilon\right)\cdot\right.\\
&\left.\cdot\|A\|_2\|x_0\|_2+\mathcal{O}(\epsilon^2)(\|A\|_2\|x_0\|_2+\|b\|_2)\right]+\\
&+\{\epsilon\sqrt{k}[(3+m\sqrt{n})\|A\|_2+4\|T_{k+1,k}\|_1]+\mathcal{O}(\epsilon^2)(\|A\|_2+\|T_{k+1,k}\|_1)\}\|R_k^{-1}\|_2\cdot\\
&\cdot\left[\left(1+\frac{n+12\sqrt{2}k+4}{2}\epsilon\right)\|b\|_2+\left(1+\frac{n+2m\sqrt{n}+12\sqrt{2}k+4}{2}\epsilon\right)\cdot\right.\\
&\left.\cdot\|A\|_2\|x_0\|_2+\mathcal{O}(\epsilon^2)(\|A\|_2\|x_0\|_2+\|b\|_2)\right]+\|A\|_2(5\epsilon+\mathcal{O}(\epsilon^2))k\chi_k \qquad (2.216)
\end{aligned}
$$

Kao i kod GMRES metode, umjesto kvazireziduala, nama je na raspolaganju skalar $|g_{k+1}^{(k)}|$, koji se za razliku od egzaktne aritmetike ne mora poklapati za normom kvazireziduala u aritemtici konačne preciznosti. Zato ćemo sada promatrati njihov odnos. Prema Lemi 18.7 iz [21, str. 373] (2.204), (2.205), (2.209) i (2.208) vrijedi

$$
\begin{aligned}
&\|\beta\xi_1-T_{k+1,k}R_k^{-1}g_{k\times 1}^{(k)}\|_2=\|\widehat{F}^{(k)}(T_{k+1,k}R_k^{-1}g_{k\times 1}^{(k)}-\beta\xi_1)\|_2\leq\\
&\leq\ \left\|\begin{bmatrix}R_k\\0\end{bmatrix}R_k^{-1}g_{k\times 1}^{(k)}+\Delta\widehat{R}_kR_k^{-1}g_{k\times 1}^{(k)}-g^{(k)}\right\|_2+\|g^{(k)}-\widehat{F}^{(k)}(\beta\xi_1)\|_2\leq\\
&\leq\ \left\|\begin{bmatrix}g_{k\times 1}^{(k)}\\0\end{bmatrix}-g^{(k)}\right\|_2+\|\Delta\widehat{R}_k\|_2\|R_k^{-1}\|_2\|g^{(k)}\|_2+6\sqrt{2}k\beta(\epsilon+\mathcal{O}(\epsilon^2))\leq\\
&\leq\ |g_{k+1}^{(k)}|+6\sqrt{2}k(\epsilon+\mathcal{O}(\epsilon^2))\|T_{k+1,k}\|_F\|R_k^{-1}\|_2\cdot\\
&\quad\cdot\left[\left(1+\frac{n+12\sqrt{2}k+4}{2}\epsilon\right)\|b\|_2+\left(1+\frac{n+2m\sqrt{n}+12\sqrt{2}k+4}{2}\epsilon\right)\cdot\right.\\
&\quad\left.\cdot\|A\|_2\|x_0\|_2+\mathcal{O}(\epsilon^2)(\|A\|_2\|x_0\|_2+\|b\|_2)\right]+6\sqrt{2}k(\epsilon+\mathcal{O}(\epsilon^2))\cdot\\
&\quad\cdot\left[\left(1+\frac{n+4}{2}\epsilon\right)\|b\|_2+\left(1+\frac{n+2m\sqrt{n}+4}{2}\epsilon\right)\|A\|_2\|x_0\|_2+\right.\\
&\quad\left.+\mathcal{O}(\epsilon^2)(\|A\|_2\|x_0\|_2+\|b\|_2)\right]. \qquad (2.217)
\end{aligned}
$$

Uvrštavanjem (2.217) u (2.216), i sređivanjem nejednakosti, dobivamo konačan odnos između norme pravog reziduala i izračunatog kvazireaziduala.

$$
\begin{aligned}
&\|b-Ax_k\|_2\leq\sqrt{k+1}\left(1+\frac{n+4}{2}\epsilon\right)|g_{k+1}^{(k)}|+\\
&\quad+\epsilon\{(m\sqrt{n}+9k\sqrt{k+1}+2)\|A\|_2\|x_0\|_2+(9k\sqrt{k+1}+2)\|b\|_2+\\
&\quad+[3\sqrt{k}\|A\|_2\|P_k\|_F\kappa(R_k)+(3+m\sqrt{n})\sqrt{k}\|A\|_2\|R_k^{-1}\|_2+\\
&\quad+(9k\sqrt{k+1}+4\sqrt{k}\sqrt{k+1})\|T_{k+1,k}\|_F\|R_k^{-1}\|_2](\|A\|_2\|x_0\|_2+\|b\|_2)+5k\chi_k\|A\|_2\}+
\end{aligned}
$$

$$+\mathcal{O}(\epsilon^2)\{\|A\|_2 + (\|A\|_2\|x_0\|_2 + \|b\|_2)[1 + (\|A\|_2 + \|T_{k+1,k}\|_F)\|R_k^{-1}\|_2 + \|A\|_2\|P_k\|_F\kappa(R_k)]\}.$$

Glava 3

Prekondicioniranje

3.1 Osnove prekondicioniranja

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

$$M^{-1}Ax = M^{-1}b \tag{3.1}$$

ima isto rješenje kao i originalni sustav $Ax = b$, ali svojstva matrice $M^{-1}A$ mogu biti bolja. Matrica prekondicioniranja M bira se uglavnom tako, da je rješavanje prekondicioniranog sustava iterativnom metodom brže, u smislu da će metoda zahtijevati manje iteracija do postizanja konvergencije od rješavanja originalnog sustava. To se najbolje može postići ako je matrica prekondicioniranog sustava blizu identiteti. U većini slučaja matrica M^{-1} se ne računa eksplicitno, već se radi o rješavanju sustava sa matricom M, pri čemu odabir matrice M treba biti takav da se taj sustav puno jednostavnije rješava od onog sa matricom A.

Budući da upotreba prekondicioniranja u iterativnoj metodi zahtijeva neke dodatne troškove u vremenu i memoriji, trebamo voditi računa o odnosu troškova konstrukcije i primjene prekondicioniranja, i dobitka u brzini konvergencije. Ukoliko su troškovi prekondicioniranja veći tada se to mora amortizirati ili drastično manjim brojem iteracija koje metoda treba izvesti, ili kroz višestruku upotrebu prekondicioniranja na više linearnih sustava. U svakom slučaju, konačni rezultat mora biti povoljan, na način da ušteda u količina rada kod korištenja prekondicioniranja mora biti veća od troškova same upotrebe prekondicioniranja.

Transformacija $A \to M^{-1}A$ nije jedina moguća. Kod sustava (3.1) radi se o *lijevom* prekondicioniranju, dok kod sustava

$$AM^{-1}y = b, \qquad x = M^{-1}y \tag{3.2}$$

radi se o *desnom* prekondicioniranju. Ako, s druge strane, postoji neka faktorizacija matrice prekondicioniranja $M = L_1L_2$ tada možemo definirati i *dvostrano* prekondici-

oniranje pomoću faktora, pri čemu se tada rješava sustav

$$L_1^{-1}AL_2^{-1}y = L_1^{-1}b, \qquad x = L_2^{-1}y. \tag{3.3}$$

U ovom slučaju treba napomenuti da, kod onih iterativnih metoda kod kojih konvergencija ovisi o spektru matrice sustava, matrice $M^{-1}A$ i $L_1^{-1}AL_2^{-1}$ imaju isti spektar, jer su slične. I kod primjene ovakvog načina prekondicioniranja, kao što smo već vidjeli, kod velike većine iterativnih metoda biti će potrebno samo rješavanje sustava sa matricom M, dok se faktori L_1 i L_2 neće trebati eksplicitno izračunavati. Matrice L_1 i L_2 zovu se *lijeva* i *desna* matrica prekondicioniranja, a kod rješavanja sustava, jednostavno se primijeni iterativna metoda na tako prekondicionirani sustav. Lijevo i desno prekondicioniranje je očigledno varijanta dvostranog prekondicioniranja, kada je jedan od faktora matrice M jednak identiteti. Kod prekondicioniranih algoritama, potrebne su u glavnom dvije osnovne modifikacije: $r_0 \leftarrow L_1^{-1}r_0$ prije iterativnog procesa, i $x_k \leftarrow L_2^{-1}x_k$ nakon iteriranja. Dvostrano prekondicioniranje ima posebnu ulogu kod prekondicioniranja hermitskog sustava sa hermitskom pozitivno definitnom matricom prekondicioniranja M. Tada se se matrica M može faktorizirati npr. metodom Choleskog na $M = LL^*$, pa matrica prekondicioniranog sustava $L^{-1}AL^{-*}$ ostaje hermitska.

Način na koji prekondicionirana matrica treba aproksimirati identitetu ovisi o iterativnoj metodi koju koristimo. Za jednostavne iteracije željeli bismo postići $\rho(I - M^{-1}A) \ll 1$ kako bismo imali brzu asimptotsku konvergenciju, ili $\|I - M^{-1}A\| \ll 1$ kako bismo imali veliku redukciju norme greške u svakom koraku.

Za CG ili MINRES metodu primijenjenu na hermitski pozitivno definitni problem, težnja nam je imati hermitsku prekondicioniranu matricu $L^{-1}AL^{-*}$ sa brojem uvjetovanosti blizu jedan, tako da ograda greške bazirana na Čebiševljevim polinomima bude mala. S druge strane, znamo da za konvergenciju ovih metoda značajnu ulogu igra i distribucija svojstvenih vrijednosti. U tom smislu mi možemo zahtijevati da prekondicionirana matrica ima, na primjer, samo nekoliko velikih svojstvenih vrijednosti, a ostatak gusto nakupljnih oko jedne točke, ili da prekondicionirana matrica ima samo nekoliko različitih svojstvenih vrijednosti. Za MINRES metodu primijenjenu na hermitski indefinitan linearan sustav, ali sa pozitvno definitnom matricom prekondicioniranja, ponovno je važna distribucija svojstvenih vrijednosti. U svakom slučaju svojstvene vrijednosti bi trebale biti distribuirane tako da polinom malog stupnja, koji je jednak jedinici u ishodištu, bude mali u svim svojstvenim vrijednostima.

Za GMRES metodu bila bi dobra prekondicionirana matrica koja je bliska normalnoj matrici i čije su svojstvene vrijednosti tijesno nakupljene oko neke točke daleko od ishodišta, ali također i neka druga svojstva bila bi dovoljna da se definira dobra matrica prekondicioniranja. Nadalje, nije baš sasvim jasno koja bi se svojstva trebala promatrati za ostale nehermitske iterativne metode (kao npr. BCG, QMR, CGS ili BICGSTAB), ali kako bi sve ove metode trebale konvergirati u jednoj iteraciji ukoliko je matrica sustava jednaka identiteti, intuitivno nam je jasno da bi prekondicionirana matrica trebala na neki način aproksimirati identitetu.

Prekondicioniranje se u grubo može podijeliti u tri kategorije:

1. Prekondicioniranja definirana za općenite klase matrica, npr. matrice sa netrivijalnim dijagonalnim elementima, pozitivno definitne matrice, M-matrice. Primjeri takvih prekondicioniranja su Jacobi, Gauss–Seidel, i SOR prekondicioniranja, nekompletna LU i Cholesky faktorizacija, i modificirana Cholesky faktorizacija.

2. Prekondicioniranja dizajnirana za široke klase raznih polaznih problema, npr. eliptičke parcijalne diferencijalne jednadžbe. Primjeri su prekondicioniranje multigrid metodom i dekompozicijom domene.

3. Prekondicioniranja dizajnirana za specifične matrice ili polazne probleme, npr. transportna jednadžba. Primjer je difuzijska sintetička akceleracija.

Prednost 1. kategorije prekondicionioniranja je ta, da se takva prekondicioniranja mogu upotrebljavati u situacijama kada točno porijeklo problema nije poznato. Većina prekondicioniranja u toj kategoriji zahtijevaju poznavanje najmanje nekih elemenata matrice sustava A. Za klasu problema koji dolaze od parcijalnih diferencijalnih jednadžbi, ponekad je moguće pokazati da prekondicioniranje mijenja ovisnost broja uvjetovanosti prekondicionirane matrice o koraku mreže, koja se koristi u aproksimacijama konačnim diferencijama ili konačnim elementima. To ne pomaže puno ako nam je cilj riješiti jedan sustav $Ax = b$, nego kad želimo riješiti polaznu parcijalno diferencijalnu jednadžbu, kod koje se težina rješavanja linearnog sustava određuje u usporedbi sa točnosti konačnih diferencija ili konačnih elemenata.

Usprkos velikim naporima u razvoju prekondicioniranja za općenite linearne sustave ili za široke klase polaznih problema, još uvijek je moguće u mnogim situacijama koristiti fizikalnu intuiciju u vezi sa specifičnim problemom, kako bi se razvilo još efektivnije prekondicioniranje. Zbog tog razloga je problem traženja pravog načina prekondicioniranja linearnog sustava vrlo širok, i obuhvaća razne grane znanosti.

3.2 Klasične iterativne metode

Najprije ćemo razmotriti klasične iterativne metode, zajedno sa nekim njihovim svojstvima i konvergencijom. Ekvivalentan način na koji možemo opisati Algoritam 2.2.1 za jednostavne iteracije, kod rješavanja sustava $Ax = b$, je sljedeći. Rastavimo matricu A na oblik $A = M - N$ tako da linearni sustav $Ax = b$ transformiramo u

$$Mx = Nx + b,$$

pri čemu je matrica M regularna. Ako nam je dana aproksimacija x_{k-1}, novu aproksimaciju x_k možemo dobiti kao rješenje jednadžbe

$$Mx_k = Nx_{k-1} + b. \tag{3.4}$$

U tom slučaju imamo iteracije oblika

$$x_k = M^{-1}Nx_{k-1} + M^{-1}b,$$

kod kojih se traži fiksna točka. S druge strane, da bismo vidjeli da je ovo ekvivalentno jednostavnim iteracijama zamijenimo $M^{-1}N$ sa $I - M^{-1}A$ kako bismo dobili

$$x_k = (I - M^{-1}A)x_{k-1} + M^{-1}b = x_{k-1} + M^{-1}r_{k-1} = x_{k-1} + z_{k-1}.$$

Dakle radi se o iteracijama oblika

$$x_k = Gx_{k-1} + f, \tag{3.5}$$

sa $G = I - M^{-1}A$, i $f = M^{-1}b$. U ovom smislu, rastav matrice $A = M - N$ i prekondicioniranje sa matricom M su sinonimi. Ovisno o izboru matrice M dobit ćemo različite

metode. Također, ovako dobivenu matricu M možemo koristiti i kao matricu prekondicioniranja za prekondicionirani oblik bilo koje od metoda aproksimacije iz Krylovljevih potprostora.

Pogledajmo sada neka osnovna svojstva konvergencije iterativnih metoda oblika (3.5). Ako iteracije započnemo sa proizvoljnim x_0, tada se postavlja pitanje u kojem slučaju će x_k konvergirati ka rješenju $A^{-1}b$, kad k teži prema beskonačnosti. Promotrimo sljedeće.

$$\begin{aligned} x_k &= Gx_{k-1} + f = G(Gx_{k-2} + f) + f = G^2x_{k-2} + Gf + f = \cdots \\ &= G_kx_0 + (G^{k-1}f + G^{k-2}f + \cdots + Gf + f) = \\ &= G^kx_0 + \left(\sum_{i=0}^{k-1} G^i\right) f. \end{aligned}$$

Ako koristimo operatorsku normu $\|\cdot\|$, i ako je $\|G\| < 1$, tada prema Lemi 1.5.9 vrijedi da je

$$\lim_{k\to\infty} \sum_{i=0}^{k} G^i = (I - G)^{-1},$$

i

$$\lim_{k\to\infty} G^k = 0.$$

Uzimanjem limesa imamo

$$\begin{aligned} x &= \lim_{k\to\infty} x_k = \lim_{k\to\infty} G^kx_0 + \lim_{k\to\infty} \left(\sum_{i=0}^{k-1} G^i\right) f = \\ &= 0 \cdot x_0 + (I - G)^{-1}f = \\ &= (I - G)^{-1}f, \end{aligned} \tag{3.6}$$

odakle je

$$(I - G)x = f,$$

odnosno

$$x = Gx + f$$

što je ekvivalentno sa $Ax = b$. Jedino što nam je preostalo, je pronaći odgovarajuću operatorsku normu.

Na kraju možemo izraziti rezultat koji govori o konvergenciji iteracija (3.5), koji je analogan onome za jednostavne iteracije. U ovom slučaju je dokaz malo drugačiji.

Teorem 3.2.1 ([18]). *Niz x_k za $k \geq 0$, dobiven pomoću rekurzije (3.5), konvergira za proizvoljnu početnu iteraciju x_0 i proizvoljnu desnu stranu f ako i samo ako je $\rho(G) < 1$.*

Dokaz: Najprije pretpostavimo da je $\rho(G) < 1$. Tada postoji $\epsilon > 0$ takav de je $\rho(G)+\epsilon < 1$. Prema Lemi 1.5.6 postoji operatorska norma $\|\cdot\|$ takva da je

$$\|B\| < \rho(G) + \epsilon < 1.$$

Prije iskaza ovog teorema već smo pokazali da ako za neku normu vrijedi da je $\|G\| < 1$ tada niz x_k konvergira, odnosno

$$\lim_{k\to\infty} x_k = x, \quad \text{i} \quad x = Gx + f.$$

Pretpostavimo sada obrat, to jest da niz x_k konvergira za svako x_0 i f. Sljedeće, pretpostavimo da je u tom slučaju $\rho(G) \geq 1$. To znači da postoji svojstvena vrijednost λ od G, takva da je $|\lambda| \geq 1$. Neka je $f \neq 0$ svojstveni vektor od G, koji pripada λ, i $x_0 = 0$. Tada imamo

$$\begin{aligned}
x_1 &= Gx_0 + f = G \cdot 0 + f = f \\
x_2 &= Gx_1 + f = Gf + f = (\lambda + 1)f \\
x_3 &= Gx_2 + f = (\lambda + 1)Gf + f = (\lambda^2 + \lambda + 1)f \\
&\vdots \\
x_k &= Gx_{k-1} + f = (\lambda^{k-1} + \cdots + \lambda + 1)f
\end{aligned}$$

Prema tome za iteraciju u k-tom koraku vrijedi

$$x_k = \begin{cases} \dfrac{\lambda^k - 1}{\lambda - 1} f, & \lambda \neq 1 \\ \\ kf, & \lambda = 1 \end{cases}$$

što u oba slučaja divergira kada k teži ka beskonačnosti. Dakle našli smo niz koji za određene x_0 i f divergira, što je u kontradikciji sa pretpostavkom. Znači, da mora biti $\rho(B) < 1$. □

Teorem 3.2.2 ([18]). *Ako za matricu G iteracije (3.5) konvergiraju za bilo koji x_0 i f, tada za svaku operatorsku normu, za koju je $\|G\| < 1$ vrijedi*

$$\|x - x_k\| \leq \frac{\|G\|}{1 - \|G\|}\|x_k - x_{k-1}\| \leq \frac{\|G\|^k}{1 - \|G\|}\|x_1 - x_0\|,$$

gdje je $x = \lim_{k\to\infty} x_k$.

Dokaz: Neka su $k \geq 1$ i $i \geq 0$. Tada imamo

$$\begin{aligned}
x_{k+i+1} - x_{k+i} &= G(x_{k+i} - x_{k+i-1}) = G^2(x_{k+i-1} - x_{k+i-2}) = \cdots \\
&= G^i(x_{k+1} - x_k).
\end{aligned}$$

S druge strane prema prethodnom, za $j \geq 1$ imamo

$$\begin{aligned}
x_{k+j} - x_k &= (x_{k+j} - x_{k+j-1}) + (x_{k+j-1} - x_{k+j-2}) + \cdots + (x_{k+1} - x_k) = \\
&= (G^{j-1} + G^{j-2} + \cdots + G + I)(x_{k+1} - x_k)
\end{aligned}$$

odakle je

$$\begin{aligned}
\|x_{k+j} - x_k\| &\leq (\|G\|^{j-1} + \|G\|^{j-2} + \cdots + \|G\| + 1)\|x_{k+1} - x_k\| = \\
&= \frac{1 - \|G\|^j}{1 - \|G\|}\|x_{k+1} - x_k\|.
\end{aligned}$$

Kako je $\lim_{j\to\infty} x_{k+j} = x$ i $\lim_{j\to\infty} \|G\|^j = 0$ vrijedi

$$\lim_{j\to\infty} \|x_{k+j} - x_k\| \leq \lim_{j\to\infty} \frac{1-\|G\|^j}{1-\|G\|}\|x_{k+1} - x_k\|,$$

odnosno

$$\|x - x_k\| \leq \frac{1}{1-\|G\|}\|x_{k+1} - x_k\|.$$

Ponovo je iz prethodno pokazanog, za $l \leq k$

$$\|x_{k+1} - x_k\| = G^l(x_{k+1-l} - x_{k-l}),$$

odnosno

$$\|x_{k+1} - x_k\| \leq \|G^l\|\|x_{k+1-l} - x_{k-l}\|,$$

pa vrijedi

$$\|x - x_k\| \leq \frac{\|G\|^l}{1-\|G\|}\|x_{k+1-l} - x_{k-l}\|.$$

Za $l = 1$ dobije se prva jednakost tvrdnje teorema, a za $l = k$ dobije se druga. □

Nakon što smo uspostavili nužne i dovoljne uvjete konvergencije, potrebno je još razmotriti i *stopu konvergencije*, pomoću koje ćemo moći uspoređivati različite rastave matrice A, odnosno prekondicioniranja. Promotrimo sada rastav $A = M - N$. Kao i kod jednostavnih iteracija imamo

$$\begin{aligned} e_k &= x - x_k = x - (I - M^{-1}A)x_{k-1} - M^{-1}b = \\ &= (I - M^{-1}A)x - (I - M^{-1}A)x_{k-1} = \\ &= (I - M^{-1}A)e_{k-1} = Ge_{k-1}. \end{aligned}$$

Prema tome, vrijedi

$$\|x - x_k\| \leq \|G\|\|x - x_{k-1}\|,$$

za bilo koju operatorsku normu. Prema Lemi 1.5.6 za svako $\epsilon > 0$ postoji neka norma za koju je $\|G\| < \rho(G) + \epsilon$, tako da spektralni radijus možemo smatrati ocjenom stope konvergencije, odnosno njene brzine.

3.2.1 Jacobijeva metoda

Ako je matrica M, dobivena iz rastava matrice A, dijagonalna, tada se jednostavne iteracije sa takvim rastavom matrice nazivaju *Jacobijevom metodom*. Ovdje pretpostavljamo da su dijagonalni elementi matrice A različiti od nule, tako da postoji M^{-1}. To je jedan od najlakših načina prekondicioniranja. Ako matricu zapišemo u obliku $A = D - L - U$, gdje je D dijagonalna, L strogo donje trokutasta i U strogo gornje trokutasta matrica tada za Jacobijevu metodu vrijedi $x_k = G_J x_{k-1} + f_J$, gdje je

$$G_J = I - D^{-1}A = D^{-1}(L + U), \qquad f_J = D^{-1}b. \tag{3.7}$$

U matričnom obliku iteracije Jacobijeve metode izgledaju kao

$$x_k = D^{-1}(L + U)x_{k-1} + D^{-1}b. \tag{3.8}$$

Raspisivanjem jednadžbe (3.8) po komponentama, možemo vidjeti kako se odvija korekcija vektora aproksimacije rješenja. Ako sa zagradama označimo komponente vektora, tada se Jacobijeva metoda može napisati u obliku

$$x_k(i) = \frac{1}{a_{ii}} \left(-\sum_{j \neq i} a_{ij} x_{k-1}(j) + b(i) \right), \quad i = 1, \ldots, n. \tag{3.9}$$

Primijetimo da kod skladištenja vektora u memoriji, novi vektor x_k ne može se prepisati preko starog vektora x_{k-1} sve dok sve komponenete vektora x_k nisu izračunate. Dakle, prije kraja jedne iteracije, oba vektora moramo posebno skladištiti.

U nastavku, interesirat će nas konvergencija Jacobijeve metode, međutim, samo za matrice sa posebnim svojstvima postoje rezultati koji govore o tome. Definirajmo zato pojam *strogo dijagonalne dominantnosti*.

Definicija 3.2.3. *Kažemo da je matrica A strogo* dijagonalno dominantna *ako vrijedi*

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \qquad i = 1, \ldots, n. \tag{3.10}$$

Za takve matrice vrijedi sljedeći teorem.

Teorem 3.2.4 ([18]). *Ako je matrica A strogo dijagonalno dominanatna, tada Jacobijeva metoda konvergira za svaku početnu iteraciju x_0.*

Dokaz: Prema (3.7) i (3.10) vrijedi

$$\|G_J\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |(G_J)_{ij}| = \max_{i=1,\ldots,n} \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| < 1,$$

pa prema Teoremu 3.2.1, Jacobijeve iteracije konvergiraju za svaku početnu iteraciju. □

3.2.2 Gauss–Seidelova metoda

Ako je matrica M, iz rastava matrice A sa svojstvom $a_{ii} \neq 0$ za $i = 1, \ldots, n$, jednaka donjem trokutastom dijelu od A, odnosno $M = D - L$, tada se procedura jednostavnih iteracija naziva *Gauss–Seidelova metoda*. Matrica i vektor iteracije su tada jednaki

$$G_{GS} = (D - L)^{-1} U, \qquad f_{GS} = (D - L)^{-1} b, \tag{3.11}$$

u matričnom obliku Gauss–Seidelove iteracije su oblika

$$x_k = (D - L)^{-1} U x_{k-1} + (D - L)^{-1} b, \tag{3.12}$$

a po komponentama je možemo raspisati kao

$$x_k(i) = \frac{1}{a_{ii}} \left(-\sum_{j=1}^{i-1} a_{ij} x_k(j) - \sum_{j=i+1}^{n} a_{ij} x_{k-1}(j) + b(i) \right), \quad i = 1, \ldots, n. \tag{3.13}$$

Kod Gauss–Seidelove metode, tekuće aproksimacije prethodnih komponenata od x koriste se da bi se izračunala njegova sljedeća komponenta. Time se omogućava da komponente od x_k prebrišu stare vrijednosti komponenata od x_{k-1}, čim se izračunaju. Zato nam nije potreban dodatni pomoćni vektor za pamćenje komponenti prethodne iteracije.

O konvergenciji Gauss–Seidelove metode govore sljedeći teoremi.

Teorem 3.2.5 ([34]). *Ako je matrica A strogo dijagonalno dominanatna, tada Gauss–Seidelova metoda konvergira za svaku početnu iteraciju x_0.*

Dokaz: Izaberimo proizvoljnu svojstvenu vrijednost λ matrice G_{GS} sa svojstvenim vektorom $y \neq 0$. Tada imamo

$$(D-L)^{-1}Uy = \lambda y,$$

što, množenjem sa $D-L$, daje

$$Uy = \lambda(D-L)y.$$

Zbog daljnje analize, promatrajmo gornju jednakost po komponenetama,

$$\lambda a_{ii}y(i) = -\sum_{j=i+1}^{n} a_{ij}y(j) - \lambda\sum_{j=1}^{i-1} a_{ij}y(j), \qquad i = 1,\ldots,n,$$

odakle, prema nejednakosti trokuta imamo

$$\begin{aligned} |\lambda||a_{ii}||y(i)| &\leq \sum_{j=i+1}^{n} |a_{ij}||y(j)| + |\lambda|\sum_{j=1}^{i-1}|a_{ij}||y(j)| \leq \\ &\leq \left(\sum_{j=i+1}^{n}|a_{ij}| + |\lambda|\sum_{j=1}^{i-1}|a_{ij}|\right)\max_{j=1,\ldots,n}|y(j)|, \qquad (3.14) \\ &\quad i = 1,\ldots,n. \end{aligned}$$

Neka je $j_0 \in \{1,\ldots,n\}$ takav da

$$|y(j_0)| = \max_{j=1,\ldots,n}|y(j)| > 0.$$

Ubacujući to u (3.14), za $i = j_0$ imamo

$$|\lambda||a_{j_0j_0}| \leq \sum_{j=j_0+1}^{n}|a_{j_0j}| + |\lambda|\sum_{j=1}^{j_0-1}|a_{j_0j}|.$$

Ako sada pretpostavimo da je $|\lambda| \geq 1$, tada iz prethodnog imamo

$$|a_{j_0j_0}| \leq \frac{1}{|\lambda|}\sum_{j=j_0+1}^{n}|a_{j_0j}| + \sum_{j=1}^{j_0-1}|a_{j_0j}| \leq \sum_{j\neq j_0}|a_{j_0j}|,$$

što je kontradiktorno sa svojstvom stroge dijagonalne dominantnosti matrice A. Stoga zaključujemo da za svaku svojstvenu vrijednost λ matrice G_{GS} vrijedi $|\lambda| < 1$, što povlači da je i $\rho(G_{GS}) < 1$, pa prema Teoremu 3.2.1, Gauss–Seidelova metoda konvergira za svaku početnu iteraciju. □

Teorem 3.2.6 ([18]). *Ako je matrica A hermitska i pozitivno definitna, tada Gauss–Seidelova metoda konvergira za svaku početnu iteraciju x_0.*

Dokaz. Neka je λ proizvoljna svojstvena vrijednost matrice G_{GS} i neka je $y \neq 0$ odgovarajući svojstveni vektor. Tada ponovo imamo

$$Uy = \lambda(D - L)y, \tag{3.15}$$

pa ako toj jednakosti dodamo $-\lambda Uy$ s lijeve i desne strane, vrijedi

$$(1 - \lambda)Uy = \lambda(D - L - U)y = \lambda Ay. \tag{3.16}$$

S druge strane je, zbog (3.15)

$$Ay = (D - L)y - Uy = (1 - \lambda)(D - L)y. \tag{3.17}$$

Pomnožimo sada (3.16) skalarno sa λy, a (3.17) sa y, i oduzmimo prvu jednadžbu od druge, tada vrijedi

$$(1 - |\lambda|^2)\langle Ay, y\rangle = (1 - \lambda)(\langle Dy, y\rangle - \langle Ly, y\rangle - \langle Uy, \lambda y\rangle). \tag{3.18}$$

Promatramo sada samo zadnji izraz u prethodnoj jednakosti. Zbog hermitičnosti matrice A vrijedi da je $U^* = L$, zbog pozitivne definitnosti $D > 0$, a zbog (3.15) vrijedi da je $\lambda Ly = \lambda Dy - Uy$, pa imao

$$\begin{aligned} \langle Uy, \lambda y\rangle &= \langle y, \lambda Ly\rangle = \langle y, \lambda Dy\rangle - \langle y, Uy\rangle = \\ &= \langle Dy, \lambda y\rangle - \langle Ly, y\rangle. \end{aligned}$$

Uz ovu jednakost, (3.18) ima oblik

$$\begin{aligned} (1 - |\lambda|^2)\langle Ay, y\rangle &= (1 - \lambda)(\langle Dy, y\rangle - \langle Ly, y\rangle - \langle Dy, \lambda y\rangle + \langle Ly, y\rangle) = \\ &= (1 - \lambda)(1 - \bar{\lambda})\langle Dy, y\rangle, \end{aligned}$$

odakle zbog $\langle Ay, y\rangle > 0$ i zbog $\lambda \neq 1$ (budući da bi iz (3.16) za $\lambda = 1$ slijedilo da je $Ay = 0$, što je kontradiktorno sa svojstvom pozitivne definitnosti matrice A), slijedi

$$1 - |\lambda|^2 = \frac{\langle Dy, y\rangle}{\langle Ay, y\rangle}|1 - \lambda|^2 > 0.$$

Znači za svaku svojstvenu vrijednost λ od G_{GS} vrijedi da je $|\lambda| < 1$, pa to vrijedi i za svojstvenu vrijednost koja ima maksimalnu apsolutnu vrijednost, odnosno $\rho(G_{GS}) < 1$, iz čega prema Teoremu 3.2.1 slijedi da Gauss–Seidelova metoda konvergira za svaku početnu iteraciju. □

3.2.3 JOR metoda

Jacobi overrelaxation, ili skračeno JOR metoda je modifikacija Jacobijeve metode sa svrhom poboljšanja konvergencije, odnosno svođenja spektralnog radijusa matrice iteracije na minimum. Ako je matrica A matrica sustava kojeg želimo riješiti, tada kao i kod Jacobijeve metode, pretpostavljamo da je $a_{ii} \neq 0$ za $i = 1, \ldots, n$, i da matricu možemo rastaviti na $A = D - L - U$, pri čemu je D dijagonala matrice A, a L i U strogi donji i strogi gornji trokut od A. Promatrajmo sada novi rastav $A = \omega^{-1}D - (1 - \omega)\omega^{-1}D - L - U$, za $\omega \in \mathbb{R}$, tako da vrijedi

$$M = \frac{1}{\omega}D, \qquad N = \frac{1 - \omega}{\omega}D + L + U.$$

Sada možemo definirati matricu iteracija $G_{JOR,\omega} = M^{-1}N$ i vektor kao

$$G_{JOR,\omega} = (1-\omega)I + \omega D^{-1}(L+U) = (1-\omega)I + \omega G_J, \quad f_{JOR,\omega} = \omega D^{-1}b. \tag{3.19}$$

Iteracije tada izgledaju kao

$$x_k = [(1-\omega)I + \omega D^{-1}(L+U)]x_{k-1} + \omega D^{-1}b. \tag{3.20}$$

Time smo zapravo dobili težinski prosjek između prijašnje iteracije i izračunate Jacobijeve iteracije, jer se iteracije mogu zapisati kao

$$x_k = (1-\omega)x_{k-1} + \omega(G_J x_{k-1} + f_J). \tag{3.21}$$

Po komponentama, iteracije imaju oblik

$$x_k(i) = (1-\omega)x_{k-1}(i) + \omega\frac{1}{a_{ii}}\left(-\sum_{j\neq i} a_{ij}x_{k-1}(j) + b_i\right),\ i = 1,\ldots,n. \tag{3.22}$$

Osnovna ideja je izabrati koeficijent ω tako da ubrza konvergenciju iteracija, i to tako da konvergencija bude najbrža moguća.

Najčešće se pod "overrelaxation" podrazumijeva JOR postupak za svako $\omega \neq 0$, ali ima i podjela na "overrelaxation" za $\omega > 1$ i "underrelaxation" za $\omega < 1$. Primijetimo da za $\omega = 1$ dobivamo prije definiranu Jacobijevu metodu.

O konvergenciji JOR metode govore nam sljedeći teoremi.

Teorem 3.2.7 ([34]). *Ako Jacobijeva metoda konvergira za svaku početnu iteraciju x_0, onda za proizvoljno $\omega \in \langle 0, 1]$ konvergira i JOR metoda i to za svako x_0.*

Dokaz: Neka je λ proizvoljna svojstvena vrijednost Jacobijeve matrice G_J. Tada zbog konvergencije Jacobijeve metode za svako x_0, na osnovu Teorema 3.2.1, slijedi

$$|\lambda| < 1.$$

Matrica iteracije JOR metode $G_{JOR,\omega}$ je oblika (3.19), pa je njezina svojstvena vrijednost oblika

$$\mu = 1 - \omega + \omega\lambda.$$

Neka je $\lambda = \alpha + \iota\beta$, za $\alpha, \beta \in \mathbb{R}$, i $\iota^2 = -1$. Tada zbog $|\lambda| < 1$ slijedi da je $\alpha^2 + \beta^2 < 1$ i $\alpha^2 < 1$. Zbog prethodno rečenog i zbog činjenice da je $1 - \omega \geq 0$, vrijedi

$$\begin{aligned} |\mu|^2 &= (1-\omega+\omega\alpha)^2 + \omega^2\beta^2 \leq \\ &\leq (1-\omega)^2 + 2\omega(1-\omega)|\alpha| + \omega^2(\alpha^2+\beta^2) < \\ &< (1-\omega)^2 + 2\omega(1-\omega) + \omega^2 = (1-\omega+\omega)^2 = 1. \end{aligned}$$

Odavde slijedi da je $\rho(G_{JOR,\omega}) < 1$, pa prema Teoremu 3.2.1, JOR metoda konvergira za svaku početnu iteraciju. □

Teorem 3.2.8 ([34]). *Neka je A hermitska pozitivno definitna matrica i neka Jacobijeva metoda konvergira. Tada konvergira i JOR metoda za*

$$0 < \omega < \frac{2}{1-\lambda} \leq 2,$$

gdje je $\lambda \leq 0$ najmanja svojstvena vrijednost Jacobijeve matrice G_J.

Dokaz: Neka su λ_i, $i = 1, \ldots, n$ svojstvene vrijednosti matrice G_J. Zbog konvergencije Jacobijeve metode iz Teorema 3.2.1 slijedi da je $|\lambda_i| < 1$ za sve $i = 1, \ldots, n$. Kako su, zbog hermitičnosti matrice $G_J = D^{-1}(L + L^*)$, što pak slijedi iz hermitičnosti matrice A, sve λ_i realni brojevi i

$$0 = \mathrm{tr}(G_J) = \sum_{i=1}^{n} \lambda_i$$

slijedi da nisu sve λ_i pozitivne, te vrijedi

$$\lambda = \min_{i=1,\ldots,n} \lambda_i \leq 0.$$

Ponovo vrijedi da su svojstvene vrijednosti matrice $G_{JOR,\omega}$ oblika

$$\mu_i = 1 - \omega(1 - \lambda_i), \qquad i = 1, \ldots, n.$$

Kako je prema uvjetima teorema

$$0 < \omega(1 - \lambda) < 2, \quad \text{i} \quad \lambda_i < 1, \quad \text{za } i = 1, \ldots, n$$

slijedi

$$0 < \omega(1 - \lambda_i) < \omega(1 - \lambda) < 2,$$

odnosno

$$-1 < \mu_i = 1 - \omega(1 - \lambda_i) < 1,$$

odakle je $\rho(G_{JOR,\omega}) < 1$, pa prema Teoremu 3.2.1 JOR metoda konvergira za svaku početnu iteraciju. □

Teorem 3.2.9 ([34]). *JOR metoda ne konvergira za*

$$\omega < 0, \quad i \quad \omega \geq 2.$$

Dokaz: Iz definicije (3.19) matrice $G_{JOR,\omega}$ vidimo da je

$$\mathrm{tr}(G_{JOR,\omega}) = n(1 - \omega) = \sum_{i=1}^{n} \mu_i,$$

pri čemu su μ_i svojstvene vrijednosti matrice $G_{JOR,\omega}$, imamo

$$n|1 - \omega| \leq \sum_{i=1}^{n} |\mu_i| \leq n\rho(G_{JOR,\omega}).$$

Dakle,

$$\rho(G_{JOR,\omega}) \geq |1 - \omega|,$$

odakle se vidi da je za $\omega < 0$ i $\omega \geq 2$ $\rho(G_{JOR,\omega}) \geq 1$, iz čega, prema Teoremu 3.2.1, slijedi tvrdnja teorema. □

3.2.4 SOR i SSOR metode

Succesive overrelaxation, ili skraćeno SOR metoda je, u ovom slučaju modifikacija Gauss–Seidelove metode sa svrhom poboljšanja konvergencije. Ako je matrica A matrica sustava kojeg želimo riješiti, pretpostavljamo da je $a_{ii} \neq 0$ za $i = 1, \ldots, n$, i da je $A = D - L - U$, pri čemu je D dijagonala matrice A, a L i U strogi donji i strogi gornji trokut od A. Pononovo, promatramo novi rastav $A = \omega^{-1}D - (1-\omega)\omega^{-1}D - L - U$, za $\omega \in \mathbb{R}$, ali ovaj puta uzimamo da su

$$M = \frac{1}{\omega}D - L, \qquad N = \frac{1-\omega}{\omega}D + U.$$

Sada možemo definirati matricu iteracija $G_{SOR,\omega} = M^{-1}N$ i vektor, kao

$$G_{SOR,\omega} = (D - \omega L)^{-1}[(1-\omega)D + \omega U], \quad f_{SOR,\omega} = \omega(D - \omega L)^{-1}b. \tag{3.23}$$

Iteracije tada izgledaju kao

$$x_k = (D - \omega L)^{-1}[(1-\omega)D + \omega U]x_{k-1} + \omega(D - \omega L)^{-1}b. \tag{3.24}$$

Ako iteracije rastavimo po komponentama, tada za $i = 1, \ldots, n$ one imaju oblik

$$x_k(i) = (1-\omega)x_{k-1}(i) + \omega\frac{1}{a_{ii}}\left(-\sum_{j=1}^{i-1} a_{ij}x_k(j) - \sum_{j=i+1}^{n} a_{ij}x_{k-1}(j) + b(i)\right). \tag{3.25}$$

Time smo, kod SOR metode, zapravo dobili da je komponenta novog vektora jednaka težinskom prosjeku između komponente prijašnje iteracije i izračunate komponente Gauss-Seidelove iteracije, jer se iteracije po komponentama mogu zapisati kao

$$x_k(i) = (1-\omega)x_{k-1}(i) + \omega(x_k^{(GS)}(i)), \tag{3.26}$$

gdje je $x_k^{(GS)}(i)$ i-ta komponenta vektora dobivenog primjenom jedne Gauss-Seidelove iteracije nad vektorom x_{k-1}.

Ponovo je osnovna ideja izabrati koeficijent ω takav da ubrza konvergenciju iteracija, i vrijede iste podjele na "overrelaxation" i "underrelaxation". Primijetimo da za $\omega = 1$ dobivamo prije definiranu Gauss–Seidelovu metodu.

Ako pretpostavimo da je matrica sustava A realna simetrična ili kompleksna hermitska, tada se mogu definirati simetrične ili hermitske verzije Gauss–Seidelove i SOR metode. Simetričnu verziju SOR metode nazivamo *Symmetric succesive overrelaxation* metodom ili skraćeno SSOR, a ona izvodi dva SOR koraka zajedno tako da rezultirajuća matrica iteracije bude nalik hermitskoj matrici. Preciznije, prvi SOR korak se izvodi kao kod (3.25), ali u drugom SOR koraku nepoznanice se računaju u obrnutom poretku. To znači, da je jedan SSOR korak ekvivalentan jednom SOR koraku u naprijed, kojeg slijedi jedan SOR korak unazad. Sličnost SSOR matrice sa hermitskom, omogućava upotrebu SSOR metode za prekondicioniranje hermitskih sustava. To je zaista i glavna motivacija za korištenje SSOR metode, budući da njena stopa konvergencije sa optimalnim izborom vrijednosti za ω je obično manja nego stopa konvergencije SOR metode sa optimalnim ω. Ako definiramo $M_1 = \omega^{-1}D - L$ i $M_2 = \omega^{-1}D - U$, tada rješavamo sljedeća dva uzastopna problema

$$\begin{aligned} M_1 x_{k-1/2} &= N_1 x_{k-1} + b, \quad N_1 = M_1 - A, \\ M_2 x_k &= N_2 x_{k-1/2} + b, \quad N_2 = M_2 - A, \end{aligned}$$

odakle je

$$x_k = M_2^{-1}N_2M_1^{-1}N_1x_{k-1} + (M_2^{-1}N_2M_1^{-1} + M_2^{-1})b.$$

Ako u prethodnu iteraciju uvrstimo prave vrijednosti za M_1, N_1, M_2 i N_2 tada dobivamo konačan oblik SSOR iteracije

$$\begin{aligned} x_k &= (D-\omega U)^{-1}[(1-\omega)D+\omega L](D-\omega L)^{-1}[(1-\omega)D+\omega U]x_{k-1} + \\ &\quad +\omega(2-\omega)(D-\omega U)^{-1}D(D-\omega L)^{-1}b, \end{aligned} \tag{3.27}$$

iz čega se vidi da je

$$G_{SSOR,\omega} = (D-\omega U)^{-1}[(1-\omega)D+\omega L](D-\omega L)^{-1}[(1-\omega)D+\omega U], \tag{3.28}$$

i

$$f_{SSOR,\omega} = \omega(2-\omega)(D-\omega U)^{-1}D(D-\omega L)^{-1}b. \tag{3.29}$$

Ako želimo dobiti točan oblik matrice prekondicioniranja koju inducira SSOR metoda, tada najprije (3.27) trebamo svesti na oblik

$$M_{SSOR,\omega}x_k = N_{SSOR,\omega}x_{k-1} + b,$$

pri čemu je onda $M_{SSOR,\omega}$ tražena matrica. Množenjem (3.27) sa $1/\omega(2-\omega)(D-\omega L)D^{-1}(D-\omega U)$, kako bi matricu uz b sveli na identitetu, dobit ćemo da je matrica prekondicioniranja upravo oblika

$$M_{SSOR,\omega} = \frac{1}{\omega(2-\omega)}(D-\omega L)D^{-1}(D-\omega U). \tag{3.30}$$

Ovakvo prekondicioniranje ponekad se koristi sa CG metodom za hermitski pozitivno definitni problem. Za $\omega = 1$ dobivamo simetričnu Gauss-Seidelovu metodu, kod koje je

$$M_{SGS} = (D-L)D^{-1}(D-U).$$

SSOR matrica prekondicionioniranja dana je u svom faktoriziranom obliku, pa tako ovakav način prekondicioniranja dijeli mnoga svojstva drugih metoda baziranih na faktorizaciji, koje će biti prezentirane u sljedećem odjeljku. Primijetimo da je $D-\omega L$ donje trokutasta matrica, a $D-\omega U$ gornje trokutasta matrica, pa zbog toga na SSOR postupak možemo gledati kao na LU faktorizaciju matrice prekondicioniranja, odnosno imamo

$$M_{SSOR,\omega} = L_{SSOR,\omega}U_{SSOR,\omega},$$

gdje su

$$L_{SSOR,\omega} = (D-\omega L)D^{-1}, \qquad U_{SSOR,\omega} = \frac{1}{\omega(2-\omega)}(D-\omega U).$$

Matrica $L_{SSOR,\omega}$ je jedinična donje trokutasta matrica, a $U_{SSOR,\omega}$ je gornje trokutasta. Znači za izračunavanje $w = M_{SSOR,\omega}^{-1}v$ treba riješiti dva uzastopna trokutasta problema

$$\begin{aligned} L_{SSOR,\omega}z &= v \\ U_{SSOR,\omega}w &= z. \end{aligned}$$

Analiza SOR metode

Analizi konvergencije SOR metode posvetit ćemo malo više pažnje, budući da se ona i najčešće koristi, od svih metoda spomenutih u ovom odjeljku. Konvergenciju ćemo promatrati za nekoliko standardnih tipova problema. Najprije, pogledajmo kada SOR metoda, u općenitom slučaju, ne konvergira.

Teorem 3.2.10 ([34]). *SOR metoda ne konvergira za*

$$\omega \leq 0, \quad i \quad \omega \geq 2.$$

Dokaz: Očito je da za SOR matricu (3.23) vrijedi

$$\begin{aligned} \det(G_{SOR,\omega}) &= (\det(D-\omega L))^{-1} \det((1-\omega)D+\omega U) = \\ &= (\det(D))^{-1}(1-\omega)^n \det(D) = (1-\omega)^n, \end{aligned}$$

jer su $D-\omega L$ i $(1-\omega)D+\omega U$ donja i gornja trokutasta matrica. Kako je

$$(\rho(G_{SOR,\omega}))^n \geq \prod_{i=1}^{n} |\lambda_i| = |\det(G_{SOR,\omega})| = |1-\omega|^n,$$

gdje su λ_i svojstvene vrijednosti matrice $G_{SOR,\omega}$, onda je

$$\rho(G_{SOR,\omega}) \geq |1-\omega|.$$

Prema Teoremu 3.2.1 SOR metoda konvergirat će ako i samo ako je $\rho(G_{SOR,\omega}) < 1$, pa ako imamo da je $|1-\omega| \geq 1$, tada SOR metoda neće konvergirati. Ovaj uvjet bit će ispunjen ako i samo ako je $\omega \leq 0$ i $\omega \geq 2$. □

Iz ovog teorema se vidi da za općenite matrice vrijedi, ako SOR metoda konvergira, tada za ω mora vrijediti $\omega \in \langle 0, 2\rangle$. Slijedi analogon Teorema 3.2.6, samo za SOR metodu.

Teorem 3.2.11 ([34]). *Neka je A hermitska pozitivno definitna matrica. Tada SOR metoda konvergira za $\omega \in \langle 0, 2\rangle$.*

Dokaz: Neka je λ proizvoljna svojstvena vrijednost SOR matrice $G_{SOR,\omega}$ i neka je $y \neq 0$ odgovarajući svojstveni vektor. Tada vrijedi

$$G_{SOR,\omega} y = \lambda y,$$

odnosno

$$[(1-\omega)D+\omega U]y = \lambda(D-\omega L)y. \tag{3.31}$$

Za lijevu stranu jednakosti (3.31), direktnim računom dobivamo

$$2[(1-\omega)D+\omega U] = (2-\omega)D - \omega A + \omega(U-L),$$

a za desnu stranu vrijedi

$$2(D-\omega L) = (2-\omega)D + \omega A + \omega(U-L).$$

Koristeći se time, uz činjenicu da je zbog hermitičnosti matrice A $L = U^*$, poslije skalarnog množenja sa y, iz (3.31) dobivamo

$$\begin{aligned} &(2-\omega)\langle Dy, y\rangle - \omega\langle Ay, y\rangle + \omega\langle (U-U^*)y, y\rangle = \\ &= \lambda[(2-\omega)\langle Dy, y\rangle + \omega\langle Ay, y\rangle + \omega\langle (U-U^*)y, y\rangle]. \end{aligned} \tag{3.32}$$

Budući da je A pozitivno definitna matrica, vrijedi da je $a_{ii} > 0$ za svako $i = 1, \ldots, n$, a kako je $y \neq 0$ imamo

$$\langle Ay, y\rangle > 0, \qquad \langle Dy, y\rangle > 0.$$

Matrica $U - U^*$ je antihermitska pa je $\langle (U - U^*)y, y\rangle$ čisto imaginarni broj ili nula. Definirajmo

$$d = \langle Dy, y\rangle, \quad a = \langle Ay, y\rangle, \quad \iota u = \langle (U - U^*)y, y\rangle,$$

gdje je $\iota = \sqrt{-1}$. Sada (3.32) glasi

$$(2 - \omega)d - \omega a + \iota\omega u = \lambda[(2 - \omega)d + \omega a + \iota\omega u],$$

odakle je

$$\lambda = \frac{(2 - \omega)d - \omega a + \iota\omega u}{(2 - \omega)d + \omega a + \iota\omega u}.$$

Imaginarni dijelovi brojnika i nazivnika u izrazu za λ su jednaki, pa je $|\lambda| < 1$ ako i samo ako je

$$|(2 - \omega)d + \omega a| > |(2 - \omega)d - \omega a|,$$

što je, nakon kvadriranja, ekvivalentno sa

$$4\omega(2 - \omega)da > 0.$$

Nadalje, budući da su $d > 0$ i $a > 0$, to će vrijediti ako i samo ako je

$$\omega(2 - \omega) > 0,$$

što se postiže ako i samo ako je $\omega \in \langle 0, 2\rangle$. Za takav izbor omega, za proizvoljni λ vrijedi $|\lambda| < 1$, pa će vrijediti i $\rho(G_{SOR,\omega}) < 1$, odakle prema Teoremu 3.2.1 SOR metoda konvergira za svaku početnu iteraciju. □

Matrice sa posebnom strukturom

U praksi se pokazuje, da mnoge matrice koje dolaze iz konkretnih parcijalnih diferencijalnih jednadžbi imaju posebnu strukturu. Zato ćemo se usredotočiti na matrice koje imaju određeni raspored nula među svojim elementima. Za takvo nešto koristit ćemo terminologiju iz teorije grafova, što se pokazalo kao korisno sredstvo za rad sa takvim matricama. Najprije ćemo definirati nekoliko pojmova.

Definicija 3.2.12. *Kažemo da je* $\mathbb{G} = \{\mathbb{V}, \mathbb{E}\}$ orijentirani graf *$n \times n$ matrice A, ako je* $\mathbb{V} = \{1, \ldots, n\}$ *i* $(i, j) \in \mathbb{E}$ *za* $i, j \in \{1, \ldots, n\}$, $i \neq j$, *ako i samo ako je* $a_{ij} \neq 0$.

Definicija 3.2.13. *Za $n \times n$ matricu A sa usmjerenim grafom* $\mathbb{G} = \{\mathbb{V}, \mathbb{E}\}$ *kažemo da zadovoljava* svojstvo $\mathcal{A}$ *ako postoji particija* $\mathbb{V} = \mathbb{S}_1 \cup \mathbb{S}_2$, *kod koje je* $\mathbb{S}_1 \cap \mathbb{S}_2 = \emptyset$, *takva da za svako* $(i, j) \in \mathbb{E}$, *ili* $i \in \mathbb{S}_1$, $j \in \mathbb{S}_2$, *ili* $j \in \mathbb{S}_1$, $i \in \mathbb{S}_2$.

Alternativnu definiciju svojstva $\mathcal{A}$ daje sljedeći teorem.

Teorem 3.2.14 ([12]). *Matrica A zadovoljava svojstvo $\mathcal{A}$ ako i samo ako je A dijagonalna matrica, ili postoji matrica permutacije P takva da $P^{-1}AP$ ima oblik*

$$\begin{bmatrix} D_1 & B \\ C & D_2 \end{bmatrix}, \tag{3.33}$$

gdje su D_1 i D_2 kvadratne dijagonalne matrice.

Dokaz: Ako matrica A zadovoljava svojstvo $\mathcal{A}$, tada ako je jedan od skupova $\mathbb{S}_1$ i $\mathbb{S}_2$ prazan, A je dijagonalna. U suprotnom slučaju treba poredati retke i stupce od A tako, da se najprije izredaju indeksi iz $\mathbb{S}_1$, iza kojih slijede indeksi iz $\mathbb{S}_2$. Iz Definicije 3.2.13 gdje se definiraju skupovi $\mathbb{S}_1$ i $\mathbb{S}_2$, slijedi da će tada dva dijagonalna bloka reda $\text{card}(\mathbb{S}_1)$ i $\text{card}(\mathbb{S}_2)$ biti dijagonalne matrice.

Obrnuto, ako se A može ispermutirati u oblik (3.33), tada uzmimo da je $\mathbb{S}_1$ skup indeksa koji odgovaraju prvom dijagonalnom bloku, a da je $\mathbb{S}_2$ skup indeksa koji odgovaraju drugom dijagonalnom bloku. Tada $\mathbb{S}_1$ i $\mathbb{S}_2$ zadovoljavaju svojstva koja se zahtijevaju u Definiciji 3.2.13. □

Mnoge matrice dobivene diskretizacijom parcijalnih diferencijalnih jednadžbi zadovoljavaju svojstvo $\mathcal{A}$. Na primjer, takva matrica je matrica koja dolazi iz aproksimacije s 5 točaka Poissonove jednadžbe na kvadratu, pomoću konačnih diferencija. Ako čvorove na mreži numeriramo na crveno–crni način, poput polja na šahovskoj ploči, tada dobivamo matricu oblika (3.33). Dakle, isplati se posebno promatrati matrice sa svojstvom $\mathcal{A}$ i njihovo ponašanje u SOR procesu, što nam je sljedeći korak.

Najprije ćamo iznijeti nekoliko pomoćnih definicija i tvrdnji, koje će nam biti potrebne u teoremima o konvergenciji. Započnimo prvo sa definicijom još jednog pojma, vezanog uz orijentirani graf matrice.

Definicija 3.2.15. *Za danu $n \times n$ matricu A, kažemo da je $u \in \mathbb{Z}^n$* vektor uređaja, *ako za njegove komponente vrijedi $|u(i) - u(j)| = 1$ za svako $(i,j) \in \mathbb{E}$. Štoviše, u je* kompatibilan vektor uređaja *ako, uz ostalo zadovoljava i*

$$\begin{aligned} i \geq j+1 \quad &\Longrightarrow \quad u(i) - u(j) = +1, \\ i \leq j-1 \quad &\Longrightarrow \quad u(i) - u(j) = -1. \end{aligned}$$

Lema 3.2.16 ([24]). *Ako za matricu A postoji vektor uređaja, tada postoji matrica permutacije P takva da za matricu $\widetilde{A} = PAP^{-1}$ postoji kompatibilan vektor uređaja.*

Dokaz: Svaka transformacija sličnosti sa matricom permutacije je zapravo preindeksiranje varijabli. Dakle, orijentirani graf matrice $\widetilde{A}$ je graf $\widetilde{\mathbb{G}} = \{\pi(\mathbb{V}), \pi(\mathbb{E})\}$, gdje je π odgovarajuća permutacija skupa $\{1, 2, \ldots, n\}$. Drugim riječima, $\pi(\mathbb{V}) = \{\pi(1), \pi(2), \ldots, \pi(n)\}$, i $(\pi(i), \pi(j)) \in \pi(\mathbb{E})$ ako i samo je $(i,j) \in \mathbb{E}$.

Neka je u vektor uređaja, koji postoji prema pretpostavci leme, i neka je $v(i) = u(\pi^{-1}(i))$, $i = 1, \ldots, n$ gdje π^{-1} inverz permutacije π. Budući da iz $(i,j) \in \pi(\mathbb{E})$ slijedi $(\pi^{-1}(i), \pi^{-1}(j)) \in \mathbb{E}$, tada prema definiciji vektora uređaja vrijedi da je $|u(\pi^{-1}(i)) - u(\pi^{-1}(j))| = 1$. Prema konstrukciji vektora v, slijedi da je $|v(i) - v(j)| = 1$, odakle je v vektor uređaja matrice $\widetilde{A}$.

Izaberimo permutaciju π takvu da je $v(1) \leq v(2) \leq \cdots \leq v(n)$, što odgovara situaciji $u(\pi^{-1}(1)) \leq u(\pi^{-1}(2)) \leq \cdots \leq u(\pi^{-1}(n))$ koja se uvijek može izvesti. Neka je dan uređen par $(i,j) \in \pi(\mathbb{E})$, $i \geq j+1$, čime dobivamo da je $v(i) - v(j) \geq 0$, a kako je v vektor uređaja, tada mora biti $v(i) - v(j) = 1$. Na isti način, iz $(i,j) \in \pi(\mathbb{E})$, $i \leq j-1$, slijedi da je $v(i) - v(j) = -1$. Prema tome možemo zaključiti da je v kompatibilan vektor uređaja. □

Važnost definicije pojma vektora uređaja izražena je u sljedećem rezultatu.

Lema 3.2.17 ([24]). *Matrica A zadovoljava svojstvo $\mathcal{A}$ ako i samo ako za nju postoji vektor uređaja.*

Dokaz: Pretpostavimo prvo da matrica A zadovoljava svojstvo $\mathcal{A}$ i uzmimo vektor u čije su komponente definirane sa

$$u(i) = \begin{cases} 1, & i \in \mathbb{S}_1, \\ 2, & i \in \mathbb{S}_2, \end{cases} \qquad i = 1, \ldots, n.$$

Za svako $(i, j) \in \mathbb{E}$ indeksi i i j pripadaju različitim skupovima, pa je zbog toga $u(i) - u(j) = \pm 1$, odakle možemo zaključiti da je u vektor uređaja.

Kako bi dokazali obrat, pretpostavimo da za matricu A postoji vektor uređaja u i neka su

$$\mathbb{S}_1 = \{i \in \mathbb{V} : \ u(i) \text{ je neparan}\}, \qquad \mathbb{S}_2 = \{i \in \mathbb{V} : \ u(i) \text{ je paran}\}.$$

Jasno je, da je u tom slučaju $\mathbb{S}_1 \cup \mathbb{S}_2 = \mathbb{V}$ i $\mathbb{S}_1 \cap \mathbb{S}_2 = \emptyset$, pa je zbog toga $\{\mathbb{S}_1, \mathbb{S}_2\}$ particija skupa $\mathbb{V}$. Za bilo koje $i, j \in \mathbb{V}$ takve da je $(i, j) \in \mathbb{E}$ iz Definicije 3.2.15 vektora uređaja slijedi da je $u(i) - u(j) = \pm 1$. Drugim riječima, brojevi $u(i)$ i $u(j)$ moraju biti različitog pariteta, pa prema gornjoj konstrukciji slijedi da i i j pripadaju različitim skupovima particije. Dakle, matrica zadovoljava svojstvo $\mathcal{A}$. □

Napomenimo, da ako matrica A zadovoljava svojstvo $\mathcal{A}$, da tada matrica $\widetilde{A} = P^{-1}AP$, koja ima oblik (3.33), ima kompatibilni vektor uređaja.

Postojanje kompatibilnog vektora uređaja je uzrok mnogih zanimljivih svojstava martrica, koja su od velike važnosti za ponašanje SOR metode.

Lema 3.2.18 ([24]). *Ako za matricu A postoji kompatibilni vektor uređaja, tada je funkcija*

$$g(s, t) = \det\left(tL + \frac{1}{t}U - sD\right), \qquad s \in \mathbb{R}, \ t \in \mathbb{R} \setminus \{0\}$$

neovisna o t.

Dokaz: Budući da A i $H(s, t) = tL + (1/t)U - sD$ dijele isti raspored nula za sve $t \neq 0$, tada prema definiciji slijedi da je svaki kompatibilni vektor uređaja matrice A također i kompatibilni vektor uređaja matrice $H(s, t)$. Specijalno, možemo zaključiti da za matricu $H(s, t)$ postoji kompatibilni vektor uređaja.

Prema definiciji determinante je

$$g(s, t) = \sum_{\pi \in \Pi_n} (-1)^{|\pi|} \prod_{i=1}^{n} h_{i,\pi(i)}(s, t),$$

gdje su $h_{ij}(s, t)$ elementi matrice $H(s, t)$, Π_n skup svih permutacija na skupu $\{1, \ldots, n\}$ i $|\pi|$ predznak permutacije $\pi \in \Pi_n$. Kako je

$$h_{ij}(s, t) = -t^{\sigma_{i-j}} s^{1-|\sigma_{i-j}|} a_{ij}, \qquad i, j = 1, \ldots, n,$$

pri čemu je

$$\sigma_l = \begin{cases} +1, & l > 0, \\ -1, & l < 0, \\ 0, & l = 0, \end{cases}$$

možemo zaključiti da je

$$g(s, t) = (-1)^n \sum_{\pi \in \Pi_n} (-1)^{|\pi|} t^{n_L(\pi) - n_U(\pi)} s^{n - n_L(\pi) - n_U(\pi)} \prod_{i=1}^{n} a_{i,\pi(i)}, \tag{3.34}$$

gdje $n_L(\pi)$ i $n_U(\pi)$ označavaju broj elemenata $i \in \{1, \ldots, n\}$ takvih da su redom $i > \pi(i)$, odnosno $i < \pi(i)$.

Neka je u kompatibilan vektor uređaja matrice $H(s,t)$, za kojeg znamo da postoji, i izaberimo proizvoljnu permutaciju $\pi \in \Pi_n$ takvu da je (jer radimo sa regularnim matricama, takva permutacija mora postojati)

$$a_{1,\pi(1)}, a_{2,\pi(2)}, \ldots, a_{n,\pi(n)} \neq 0.$$

Tada su svi parovi $(i, \pi(i))$, $i = 1, \ldots, n$ u $\mathbb{E}$. Iz defincije slijedi

$$n_L(\pi) = \sum_{\pi(i)<i} [u(i) - u(\pi(i))], \qquad n_U(\pi) = \sum_{\pi(i)>i} [u(\pi(i)) - u(i)],$$

odakle je

$$n_L(\pi) - n_U(\pi) = \sum_{\pi(i)\neq i} [u(i) - u(\pi(i))] = \sum_{i=1}^{n} [u(i) - u(\pi(i))] = \sum_{i=1}^{n} u(i) - \sum_{i=1}^{n} u(\pi(i)).$$

Kako je π permutacija skupa $\{1, \ldots, n\}$, onda je

$$\sum_{i=1}^{n} u(i) = \sum_{i=1}^{n} u(\pi(i))$$

i $n_L(\pi) - n_U(\pi) = 0$ za sve $\pi \in \Pi_n$ takve da je $a_{i,\pi(i)} \neq 0$ za sve $i = 1, \ldots, n$. Zato

$$t^{n_L(\pi)-n_U(\pi)} \prod_{i=1}^{n} a_{i,\pi(i)} = \prod_{i=1}^{n} a_{i,\pi(i)}, \qquad \pi \in \Pi_n,$$

i iz (3.34) slijedi da je $g(s,t)$ neovisan o $t \in \mathbb{R} \setminus \{0\}$. □

Napomena: Prema Lemi 3.2.17, Lemi 3.2.16 i Lemi 3.2.18, slijedi da ako matrica A zadovoljava svojstvo $\mathcal{A}$, tada postoji matrica permutacije P takva da matrica $P^{-1}AP$ zadovoljava

$$\det(L + U - sD) = \det(tL + t^{-1}U - sD), \qquad t \in \mathbb{R} \setminus \{0\}. \tag{3.35}$$

Sada ćemo napokon iznijeti niz rezultata o svojstvenim vrijednostima i konvergenciji SOR metode, vezanih uz matrice koje zadovoljavaju svojstvo $\mathcal{A}$, odnosno za koje vrijedi tvrdnja Leme 3.2.18.

Teorem 3.2.19 ([12]). *Pretpostavimo da matrica $A = D - L - U$ zadovoljava svojstvo (3.35). Tada vrijede sljedeća svojstva:*

(i) Ako je μ svojstvena vrijednost Jacobijeve matrice G_J, tada je i $-\mu$ svojstvena vrijednost od G_J sa istom kratnosti.

(ii) Ako je $\lambda = 0$ svojstvena vrijednost SOR matrice $G_{SOR,\omega}$, tada je $\omega = 1$.

(iii) Ako je $\lambda \neq 0$ svojstvena vrijednost od $G_{SOR,\omega}$ za neko $\omega \in \langle 0, 2\rangle$, tada je

$$\mu = \frac{\lambda + \omega - 1}{\omega \lambda^{1/2}} \tag{3.36}$$

svojstvena vrijednost od G_J.

(iv) Ako je μ svojstvena vrijednost od G_J i λ zadovoljava (3.36) za neko $\omega \in \langle 0, 2\rangle$, tada je λ svojstvena vrijednost od $G_{SOR,\omega}$.

Dokaz: Iz svojstva (3.35) sa $t = -1$, imamo da je za bilo koji broj μ,

$$\begin{aligned}\det(G_J - \mu I) &= \det(D^{-1}(L+U) - \mu I) = \det[D^{-1}(L+U-\mu D)] = \\ &= \frac{1}{\det D}\det(L+U-\mu D) = \frac{1}{\det D}\det(-L-U-\mu D) = \\ &= \frac{(-1)^n}{\det D}(L+U+\mu D) = (-1)^n \det(D^{-1}(L+U)+\mu I) = \\ &= (-1)^n \det(G_J + \mu I).\end{aligned}$$

Budući da su svojstvene vrijednosti od G_J brojevi za koje je $\det(G_J - \mu I) = 0$, a njihove kratnosti su također određene karakterističnim polinomom, dobivamo tvrdnju *(i)*.

Primijetimo da matricu $G_{SOR,\omega}$ možemo napisati na drugačiji način,

$$\begin{aligned}G_{SOR,\omega} &= (D-\omega L)^{-1}[(1-\omega)D + \omega U] = \\ &= (I - \omega D^{-1}L)^{-1}D^{-1}D[(1-\omega)I + \omega D^{-1}U] = \\ &= (I - \omega D^{-1}L)^{-1}[(1-\omega)I + D^{-1}U]. \qquad (3.37)\end{aligned}$$

Budući da je matrica $I - \omega D^{-1}L$ donje trokutasta sa jedinicama na dijagonali, njezina detrminanta je jednaka 1, tako da za bilo koji broj λ imamo

$$\begin{aligned}\det(G_{SOR,\omega} - \lambda I) &= \det[(I-\omega D^{-1}L)^{-1}[(1-\omega)I + \omega D^{-1}U] - \lambda I] = \\ &= \det[(1-\omega)I + \omega D^{-1}U - \lambda(I - \omega D^{-1}L)]. \qquad (3.38)\end{aligned}$$

Ako je $\lambda = 0$ svojstvena vrijednost od $G_{SOR,\omega}$, tada slijedi da je

$$\det[(1-\omega)I + \omega D^{-1}U] = 0.$$

Kako je ta matrica gornje trokutasta sa $(1-\omega)$ na dijagonali, zaključujemo da je

$$\det[(1-\omega)I + \omega D^{-1}U] = (1-\omega)^n = 0,$$

i da je $\omega = 1$, čime smo pokazali tvrdnju *(ii)*.

Za $\lambda \neq 0$, iz jednakosti (3.38) slijedi

$$\begin{aligned}\det(G_{SOR,\omega} - \lambda I) &= \frac{1}{\det D}\det\left[(1-\omega-\lambda)D + \omega U + \lambda\omega L\right] = \\ &= \frac{\lambda^{n/2}\omega^n}{\det D}\det\left[\frac{1-\omega-\lambda}{\lambda^{1/2}\omega}D + \lambda^{-1/2}U + \lambda^{1/2}L\right].\end{aligned}$$

Koristeći svojstvo (3.35) za $t = \lambda^{1/2}$ imamo

$$\begin{aligned}\det(G_{SOR,\omega} - \lambda I) &= \frac{\lambda^{n/2}\omega^n}{\det D}\det\left[\frac{1-\omega-\lambda}{\lambda^{1/2}\omega}D + U + L\right] = \\ &= \lambda^{n/2}\omega^n \det\left(D^{-1}(L+U) - \frac{\lambda+\omega-1}{\lambda^{1/2}\omega}I\right) = \\ &= \lambda^{n/2}\omega^n \det\left(G_J - \frac{\lambda+\omega-1}{\lambda^{1/2}\omega}I\right). \qquad (3.39)\end{aligned}$$

Iz ovoga slijedi da ako je $\lambda \neq 0$ svojstvena vrijednost od $G_{SOR,\omega}$ i ako μ zadovoljava (3.36), tada je μ svojstvena vrijednost od G_J. Obrnuto, ako je μ svojstvena vrijednost od G_J i ako λ zadovoljava (3.36), tada je λ svojstvena vrijednost od $G_{SOR,\omega}$. Time smopokazali tvrdnje *(iii)* i *(iv)*. □

Korolar 3.2.20 ([12]). *Kada matrica sustava A zadovoljava svojstvo (3.35), tada je $\rho(G_{GS}) = [\rho(G_J)]^2$.*

Dokaz: Kako je $G_{GS} = G_{SOR,1}$, kad u (3.36) uvrstimo $\omega = 1$ onda se ona svodi na

$$\mu = \lambda^{1/2}.$$

Ako su sve svojstvene vrijednosti od G_{GS} jednake nuli, tada je $G_{GS} = (D-L)^{-1}U = 0$, odakle slijedi da je $U = 0$. Kao posljedicu toga, imamo da je $G_J = D^{-1}L$ strogo donje trokutasta matrica, sa nulama na dijagonali. Prema tome, u toj situaciji i sve svojstvene vrijednosti od G_J su jednake nuli pa je $\rho(G_{GS}) = 0 = [\rho(G_J]^2$. Ako postoji svojstvena vrijednost λ od G_{GS}, različita od nule, tada prema tvrdnji *(iii)* Teorema 3.2.19 slijedi da postoji svojstvena vrijednost μ od G_J, takva da je $\mu = \lambda^{1/2}$. To vrijedi i za λ za koji je $|\lambda| = \rho(G_{GS})$. Odavde slijedi da je $[\rho(G_J)]^2 \geq \rho(G_{GS})$. Iz tvrdnje *(iv)* Teorema 3.2.19 slijedi da ne postoji svojstvena vrijednost μ od G_J, takva da je $|\mu|^2 > \rho(G_{GS})$, jer kad bi postojala takva svojstvena vrijednost μ, tada bi $\lambda = \mu^2$ bila svojstvena vrijednost od G_{GS} sa $|\lambda| > \rho(G_{GS})$, što je kontradiktorno sa definicijom spektralnog radijusa. Dakle $[\rho(G_J)]^2 = \rho(G_{GS})$. □

Prema ovom korolaru Gauss–Seidelova metoda dvostruko brže konvergira nego Jacobijeva.

Sljedeći teorem govori o optimalnoj vrijednosti parametra ω kod SOR metode, i odgovarajućoj optimalnoj stopi konvergencije.

Teorem 3.2.21 ([12]). *Pretpostavimo da matrica A zadovoljava svojstvo (3.35), da G_J ima samo realne svojstvene vrijednosti, i da je $\rho_\mu = \rho(G_J) < 1$. Tada SOR metoda konvergira za svako $\omega \in \langle 0, 2\rangle$, i spektralni radijus SOR matrice je*

$$\rho(G_{SOR,\omega}) = \begin{cases} \frac{1}{4}\left[\omega\rho_\mu + \sqrt{(\omega\rho_\mu)^2 - 4(\omega-1)}\right]^2 & za\ 0 < \omega \leq \omega_{opt}, \\ \omega - 1 & za\ \omega_{opt} \leq \omega < 2, \end{cases} \tag{3.40}$$

gdje je ω_{opt}, optimalna vrijednost od ω, jednak

$$\omega_{opt} = \frac{2}{1 + \sqrt{1-\rho_\mu^2}}. \tag{3.41}$$

Za bilo koju drugu vrijednost od ω, imamo

$$\rho(G_{SOR,\omega_{opt}}) < \rho(G_{SOR,\omega}), \qquad \omega \in \langle 0, 2\rangle \setminus \{\omega_{opt}\}. \tag{3.42}$$

Dokaz: Napomenimo najprije da matrica G_J može imati sve svojstvene vrijednosti realne u slučaju kad je, na primjer, matrica A hermitska.

Rješavanje jednadžbe (3.36) po λ, za proizvoljni ω, rezultira sa

$$\lambda = \frac{1}{4}\left(\omega\mu \pm \sqrt{(\omega\mu)^2 - 4(\omega-1)}\right)^2. \tag{3.43}$$

Iz Teorema 3.2.19 slijedi da ako je μ svojstvena vrijednost od G_J, tada su oba korijena λ svojstvene vrijednosti od $G_{SOR,\omega}$.

Budući da je μ realan, izraz unutar kvadratnog korijena u (3.43), za $\mu \neq 0$, je negativan ako

$$\frac{2(1-\sqrt{1-\mu^2})}{\mu^2} < \omega < \frac{2(1+\sqrt{1-\mu^2})}{\mu^2}.$$

Kako je $|\mu| < 1$, to znači da je

$$\frac{2(1+\sqrt{1-\mu^2})}{\mu^2} > 2,$$

pa kako za $\omega \geq 2$ znamo sa SOR metoda ne konvergira, nas interesiraju samo $\omega < 2$. Dakle imamo da je taj izraz negativan ako

$$\omega_\mu = \frac{2(1-\sqrt{1-\mu^2})}{\mu^2} < \omega < 2,$$

i u tom slučaju je

$$\begin{aligned} |\lambda| &= \frac{1}{4}\left|\omega\mu \pm \sqrt{(\omega\mu)^2 - 4(\omega-1)}\right|^2 = \\ &= \frac{1}{4}[(\omega\mu)^2 + |(\omega\mu)^2 - 4(\omega-1)|] = \\ &= \frac{1}{4}[(\omega\mu)^2 + 4(\omega-1) - (\omega\mu)^2] = \\ &= \omega - 1, \qquad \omega \in \langle \omega_\mu, 2\rangle. \end{aligned}$$

Za $\mu = 0$ je izraz pod korijenom u (3.43) jednak $-4(\omega-1)$, a $\omega_\mu = 1$. U oba slučaja, za $\omega > \omega_\mu$, i $\omega \leq \omega_\mu$ je $|\lambda| = |\omega - 1|$. Zato, u slučaju da je $G_J = 0$, kada su joj sve svojstvene vrijednosti jednake nuli, $\rho(G_{SOR,\omega}) = |\omega - 1| < 1$ za $\omega \in \langle 0, 2\rangle$, $\omega_{opt} = 1$ odakle slijedi da je $\rho(G_{SOR,\omega_{opt}}) = 0 < \rho(G_{SOR,\omega})$ za $\omega \in \langle 0, 2\rangle \setminus \{0\}$, čime je za $G_J = 0$ tvrdnja teorema dokazana. Od sada pa na dalje ćemo pretpostaviti da postoji svojstvena vrijednost μ od G_J različita od nule, i upravo nju ćemo promatrati.

U preostalom intervalu $\langle 0, \omega_\mu]$, za $\omega \in \langle 0, \omega_\mu]$, oba korijena λ su pozitivna, a veći od njih je oblika

$$\frac{1}{4}\left[\omega|\mu| + \sqrt{(\omega|\mu|)^2 - 4(\omega-1)}\right]^2, \qquad \omega \in \langle 0, \omega_\mu]. \tag{3.44}$$

Ova vrijednost je veća ili jednaka od $\omega - 1$ za $\omega \in \langle 0, \omega_\mu]$, jer za taj interval, budući da je izraz pod korijenom u (3.43) veći od nule, imamo

$$\frac{1}{4}\left[\omega|\mu| + \sqrt{(\omega|\mu|)^2 - 4(\omega-1)}\right]^2 \geq \frac{1}{4}(\omega|\mu|)^2 \geq \omega - 1.$$

Definirajmo sada funkciju $f(|\mu|) = \omega_\mu$, i nju derivirajmo po varijabli $|\mu|$. Dobivamo

$$f'(|\mu|) = \frac{6 - |\mu|^2 - 4\sqrt{1-|\mu|^2}}{|\mu|^3\sqrt{1-|\mu|^2}},$$

te zatim pretpostavimo da je $f'(|\mu|) \leq 0$. To vrijedi ako i samo ako je

$$6 - |\mu|^2 - 4\sqrt{1-|\mu|^2} \leq 0,$$

što je ekvivalentno sa $|\mu| \geq \sqrt{5}/2 > 1$, čime smo dobili kontradikciju sa pretpostavkom teorema da je $\rho(G_J) < 1$. Dakle, ω_μ je strogo rastuća funkcija od $|\mu|$, i imamo

$$\omega_\mu \leq \frac{2(1-\sqrt{1-\rho_\mu^2})}{\rho_\mu^2} = \frac{2}{1+\sqrt{1-\rho_\mu^2}} = \omega_{opt}.$$

Označimo sada sa $g(|\mu|)$, za $\mu \neq 0$ izraz u (3.44) za fiksni $\omega \in \langle 0, \omega_\mu]$, i uzmimo neku svojstvenu vrijednost ν od G_J, takvu da je $|\nu| < |\mu|$, te označimo sa $\lambda_\mu = g(|\mu|)$ i λ_ν odgovarajuće svojstvene vrijednosti od $G_{SOR,\omega}$. Imamo dvije situacije. Mi znamo da je $\omega_\nu < \omega_\mu$, ali ne znamo u kom se odnosu nalaze ω i ω_ν. U prvom slučaju, ako je $\omega > \omega_\nu$ tada vrijedi

$$|\lambda_\nu| = \omega - 1 \leq g(|\mu|) = \lambda_\mu,$$

što smo već prije pokazali. U drugom slučaju, ako je $\omega \leq \omega_\nu$, tada je $\lambda_\nu = g(|\nu|)$, i pretpostavimo da je $\lambda_\nu \geq \lambda_\mu$. To je ekvivalentno sa sljedećim nizom ekvivalentnih nejednakosti:

$$\begin{gathered}
\omega|\nu| + \sqrt{\omega^2|\nu|^2 - 4(\omega-1)} \geq \omega|\mu| + \sqrt{\omega^2|\mu|^2 - 4(\omega-1)} \quad \Longrightarrow \\
0 < \omega(|\mu| - |\nu|) \leq \sqrt{\omega|\nu|^2 - 4(\omega-1)} - \sqrt{\omega|\mu|^2 - 4(\omega-1)} \quad \Longrightarrow \\
\sqrt{\omega|\mu|^2 - 4(\omega-1)} < \sqrt{\omega|\nu|^2 - 4(\omega-1)} \quad \Longrightarrow \\
|\mu| < |\nu|,
\end{gathered}$$

što je kontradikcija sa izborom ν. Dakle, i u tom slučaju ispada da je $\lambda_\nu < \lambda_\mu$. Možemo zaključiti da, za $\omega \in \langle 0, \omega_\mu]$ i za $|\nu| < |\mu|$ vrijedi $|\lambda_\nu| \leq |\lambda_\mu|$.

Odavde slijedi da svojstvena vrijednost λ od $G_{SOR,\omega}$ za koju je $|\lambda| = \rho(G_{SOR,\omega})$ odgovara svojstvenoj vrijednosti μ od G_J za koju je $|\mu| = \rho_\mu$, jer je takva svojstvena vrijednost veća od onih koje odgovaraju svojstvenim vrijednostima matrice G_J sa manjom apsolutnom vrijednosti $|\mu|$, ako je $\omega \in \langle 0, \omega_{opt}]$, ili je po apsolutnoj vrijednosti jednaka svim ostalim za $\omega \in \langle \omega_{opt}, 2\rangle$. Prema tome, možemo zaključiti da (3.40) vrijedi za ω_{opt} danog sa (3.41).

Nadalje, ako opet definiramo funkciju $h(\omega)$ koja je jednaka (3.44) za fiksni $|\mu| = \rho_\mu$, tada ta funkcija poprima vrijednosti $h(0) = 1$ i $h(\omega_{opt}) = \omega_{opt} - 1 < 1$. Derivacija te funkcije po ω za $\omega \in \langle 0, \omega_{opt}\rangle$ je oblika

$$h'(\omega) = \frac{1}{2}\left[\omega\rho_\mu + \sqrt{(\omega\rho_\mu)^2 - 4(\omega-1)}\right]\left[\rho_\mu + \frac{\rho_\mu^2\omega - 2}{\sqrt{(\omega\rho_\mu)^2 - 4(\omega-1)}}\right].$$

Ako pretpostavimo da je $h'(\omega) \geq 0$, tada je

$$\rho_\mu + \frac{\rho_\mu^2\omega - 2}{\sqrt{(\omega\rho_\mu)^2 - 4(\omega-1)}} \geq 0,$$

što je opet ekvivalentno sa $\rho_\mu \geq 1$. To je kontradikcija sa pretpostavkom teorema, pa zaključujemo da je $h(\omega)$, odnosno $\rho(G_{SOR,\omega})$ strogo padajuća funkcija od ω na $\langle 0, \omega_{opt}]$. Prema tome vrijedi

$$\rho(G_{SOR,\omega}) < h(0) = 1, \qquad \text{za } \omega \in \langle 0, \omega_{opt}],$$

i

$$\rho(G_{SOR,\omega}) = \omega - 1 < 1, \qquad \text{za } \omega \in \langle \omega_{opt}, 2\rangle,$$

odakle možemo zaključiti da SOR metoda konvergira za $\omega \in \langle 0, 2\rangle$. Dalje, vrijedi

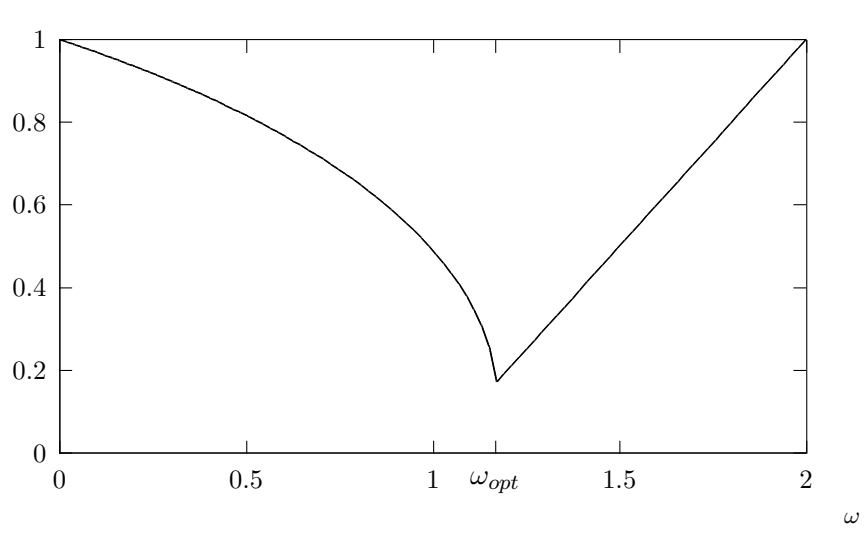
$$\rho(G_{SOR,\omega_{opt}}) < \rho(G_{SOR,\omega}), \qquad \text{za } \omega \in \langle 0, \omega_{opt}],$$

i

$$\rho(G_{SOR,\omega_{opt}}) = 1 - \omega_{opt} < 1 - \omega, \qquad \text{za } \omega \in \langle \omega_{opt}, 2\rangle,$$

čime smo pokazali i nejednakost (3.42). □

Valja napomenuti da, budući da izraz za ω_{opt} sadrži spektralni radijus matrice G_J, on često nije poznat, i tada je bolje da se on precijeni, nego da se podcijeni. Razlog tome se vidi iz izraza za h' i Slike 3.1, jer je $h'(\omega_{opt}^-) = \infty$, pa lijevo od ω_{opt} $\rho(G_{SOR,\omega})$ puno brže pada nego što desno od ω_{opt} raste (tu je brzina jednaka 1).



Slika 3.1: Spektralni radijus SOR matrice za različite vrijednosti $\omega \in \langle 0, 2\rangle$.

3.2.5 Blok metode

Ponekad se javlja potreba za korištenjem blok matrica prekondicioniranja, obično zbog blok strukture same matrice sustava kojeg rješavamo. Prirodni odabir za takav pristup je:

- U problemima sa više fizičkih varijabli po čvoru diskretizacijske mreže, blok se može oformiti grupiranjem jednadžbi za jedan čvor.

- Kod strukturiranih matrica, kao kod matrica dobivenih iz parcijalnih diferencijalnih jednadžbi na regularnim mrežama, gdje se particioniranje u blokove može zasnivati na fizičkoj domeni. Primjeri ovakvog slučaja su particioniranje duž linija u 2D slučaju, ili duž ravnina u 3D slučaju.

- Kod paralelnih računala prirodno je da se particioniranje varijabli podudara sa raspodjelom varijabli po procesorima.

Najčešće se koristi blok-dijagonalna matrica prekondicioniranja, što je osnova za blok verziju Jacobijeve metode. Ona se može dobiti iz particioniranja varijabli. Ako se skup indeksa $S = \{1, \ldots, n\}$ particionira kao $S = \bigcup_i S_i$, pri čemu su skupovi S_i medjusobno disjunktni, tada definiramo matricu M_{BJ}, čiji su elementi definirani sa

$$m_{ij} = \begin{cases} a_{ij}, & \text{ako su } i \text{ i } j \text{ u istom indeksnom podskupu,} \\ 0, & \text{inače.} \end{cases}$$

Druga mogućnost je definiranje blok-rastava matrice $A = D_B - L_B - U_B$ za istu particiju varijabli, pri čemu je za $D_B = [d_{ij}]$, $L_B = [l_{ij}]$ i $U_B = [u_{ij}]$

$$d_{ij} = \begin{cases} a_{ij}, & \text{ako su } i \text{ i } j \text{ u istom indeksnom podskupu,} \\ 0, & \text{inače,} \end{cases}$$

$$l_{ij} = \begin{cases} a_{ij}, & \text{ako su } i \in S_{p(i)} \text{ i } j \in S_{p(j)} \text{ za koje je } p(i) > p(j), \\ 0, & \text{inače,} \end{cases}$$

$$u_{ij} = \begin{cases} a_{ij}, & \text{ako su } i \in S_{p(i)} \text{ i } j \in S_{p(j)} \text{ za koje je } p(i) < p(j), \\ 0, & \text{inače.} \end{cases}$$

Tada blok verziju Gauss–Seidelove metode možemo definirati kao $M_{BGS} = D_B - L_B$ i $N_{BGS} = U_B$, kod koje kad tražimo inverz od M_{BGS} umjesto inverza dijagonalnih elemenata, tražimo inverz male blok dijagonalne matrice koja se često može lagano direktno riješiti.

Us pomoć prethodnog rastava, potpuno analogno kao i kod standardnih metoda, možemo definirati blok verzije JOR i SOR metoda.

3.3 Uspoređivanje prekondicioniranja

3.3.1 Perron–Frobeniusova teorija

Sada ćemo dati mali uvod u teoriju koja se bavi asimptotskom stopom konvergencije jednostavnih iteracija, kada su upotrebljene sa određenom klasom rastava matrice, poznatim pod imenom *regularni rastavi*. Ta teorija se zasniva na radu Perrona i Frobeniusa na nenegativnim matricama, i dat će nam neke pomoćne rezultate za usporedbu regularnih rastava.

Definicija 3.3.1. *Koristit ćemo notaciju $A \geq B$ ($A > B$) sa značenjem da svaki element realne matrice A je veći ili jednak od (strogo veći od) odgovarajućeg elementa od B. Matrica sa (i,j)-tim elementom $|a_{ij}|$ bit će označena sa $|A|$. Za matricu A kažemo da je* pozitivna *(*nenegativna*) ako je $A > 0$ ($A \geq 0$).*

Lema 3.3.2. *Neka su A i B $n \times n$ matrice, i neka je v n-vektor. Tada vrijedi:*

(i) $|A^k| \leq |A|^k$ *za sve* $k = 1, 2, \ldots$.

(ii) Ako je $0 \leq A \leq B$, *tada* $0 \leq A^k \leq B^k$ *za sve* $k = 1, 2, \ldots$.

(iii) Ako je $A > 0$, *tada* $A^k > 0$ *za sve* $k = 1, 2, \ldots$.

(iv) Ako je $A > 0$, $v \geq 0$ *i* v *nije nul-vektor, tada* $Av > 0$.

(v) Ako je $A \geq 0$, $v \geq 0$ *i* $Av \geq \alpha v$ *za neko* $\alpha > 0$, *tada* $A^k v \geq \alpha^k v$ *za sve* $k = 1, 2, \ldots$.

Dokaz: (i) Dokaz se izvodi matematičkom indukcijom. Prvo, promatramo slučaj za $k = 2$. Imamo

$$|A^2|_{ij} = \left| \sum_{l=1}^{n} a_{il} a_{lj} \right| \leq \sum_{l=1}^{n} |a_{il}||a_{lj}| = (|A|^2)_{ij},$$

za proizvoljne $i, j = 1, \ldots, n$, pa možemo zaključiti da je $|A^2| \leq |A|^2$. Pretpostavimo da tvrdnja *(i)* vrijedi za $k = 2, 3, \ldots$ i provjerimo da li ona vrijedi i za $k+1$.

$$|A^{k+1}|_{ij} = \left| \sum_{l=1}^{n} (A^k)_{il} a_{lj} \right| \leq \sum_{l=1}^{n} |A^k|_{il} |a_{lj}|,$$

za proizvoljne $i, j = 1 \ldots, n$, odakle možemo zaključiti da je $|A^{k+1}| \leq |A^k||A|$. Iz ove nejednakosti i pretpostavke indukcije, slijedi

$$|A^{k+1}| \leq |A|^k |A| = |A|^{k+1},$$

pa možemo zaključiti da tvrdnja *(i)* vrijedi za svako k.

(ii) Za $k = 2$, prema pretpostavci tvrdnje, vrijedi

$$(A^2)_{ij} = \sum_{l=1}^{n} a_{il} a_{lj} \leq \sum_{l=1}^{n} b_{il} b_{lj} = (B^2)_{ij},$$

za proizvoljne $i, j = 1, \ldots, n$, pa prema tome je $0 \leq A^2 \leq B^2$. Pretpostavimo da tvrdnja *(ii)* vrijedi za $k = 2, 3, \ldots$, pa imamo

$$(A^{k+1})_{ij} = \sum_{l=1}^{n} (A^k)_{il} a_{lj} \leq \sum_{l=1}^{n} (B^k)_{il} b_{lj},$$

za proizvoljne $i, j = 1, \ldots, n$, odakle je

$$0 \leq A^{k+1} \leq B^k B = B^{k+1}.$$

Tvrdnja *(ii)* vrijedi za svako k.

(iii) Za $k = 2$, prema pretpostavci tvrdnje, vrijedi

$$(A^2)_{ij} = \sum_{l=1}^{2} a_{il} a_{lj} > 0,$$

za proizvoljne $i, j = 1, \ldots, n$, odnosno $A^2 > 0$. Pretpostavimo da tvrdnja vrijedi za $k = 2, 3 \ldots$, tada je

$$(A^{k+1}) = \sum_{l=1}^{n} (A^k)_{il} a_{lj} > 0,$$

za proizvoljne $i, j = 1, \ldots, n$, odnosno $A^{k+1} > 0$, pa pretpostavka *(iii)* vrijedi za svako k.

(iv) Za $i = 1, \ldots, n$ imamo

$$(Av)_i = \sum_{j=1}^{n} a_{ij} v_j.$$

Kako v nije nul-vektor, tada postoji $j_0 \in \{1, \ldots, n\}$ takav da je $v_{j_0} > 0$ tada je

$$(Av)_i \geq a_{ij_0} v_{j_0} > 0.$$

Budući da ova nejednakost vrijedi za svaku komponentu vektora Av, tada imamo da je $Av > 0$.

(v) Za $k = 2$, prema pretpostavci tvrdnje, za $i = 1, \ldots, n$ vrijedi

$$(A^2 v)_i = \sum_{r=1}^{n} a_{ir} \left(\sum_{s=1}^{n} a_{rs} v_s \right) \leq \alpha \sum_{r=1}^{n} a_{ir} v_r \leq \alpha^2 v_i,$$

odakle je $A^2 v \leq \alpha^2 v$. Pretpostavimo da tvrdnja vrijedi za $k = 2, 3, \ldots$, tada, za $i = 1, \ldots, n$, i uz tvrdnju *(iii)* samo sa "$\leq$", imamo

$$(A^{k+1} v)_i = \sum_{r=1}^{n} (A^k)_{ir} \left(\sum_{s=1}^{n} a_{rs} v_s \right) \leq \alpha \sum_{r=1}^{n} (A^k)_{ir} v_r \leq \alpha^{k+1} v_i,$$

odakle je $A^{k+1} v \geq \alpha^{k+1} v$, čime smo pokazali da tvrdnja *(v)* vrijedi za svako k.

□

Slijede nekoliko rezultata koji omogućavaju usporedbu matrica.

Teorem 3.3.3 ([12]). *Neka su A i B $n \times n$ matrice. Ako je $|A| \leq B$, tada $\rho(A) \leq \rho(|A|) \leq \rho(B)$.*

Dokaz: Iz tvrdnji *(i)* i *(ii)* Leme 3.3.2 slijedi da za svako $k = 1, 2, \ldots$ imamo da je $|A^k| \leq |A|^k \leq B^k$, tako da Frobeniusove norme tih matrica zadovoljavaju

$$\|A^k\|_F^{1/k} \leq \||A|^k\|_F^{1/k} \leq \|B^k\|_F^{1/k}. \tag{3.45}$$

Budući da je prema Korolaru 1.5.8 spektralni radijus matrice C jednak $\lim_{k\to\infty} ||C^k||^{1/k}$, gdje je $\|\cdot\|$ bilo koja matrična norma, uzimanje limesa u (3.45) daje $\rho(A) \leq \rho(|A|) \leq \rho(B)$. □

Korolar 3.3.4 ([12]). *Naka su A i B $n \times n$ matrice. Ako je $0 \leq A \leq B$, tada $\rho(A) \leq \rho(B)$.*

Korolar 3.3.5 ([12]). *Neka su A i B $n \times n$ matrice. Ako je $0 \leq A < B$, tada $\rho(A) < \rho(B)$.*

Dokaz: Postoji broj

$$\alpha = \min_{i,j=1,\ldots,n, a_{ij} \neq 0} \frac{a_{ij} + b_{ij}}{2a_{ij}} > 1,$$

takav da je za svake $i, j = 1, \ldots, n$, kada je $a_{ij} = 0$, $a_{ij} = \alpha a_{ij} = 0 < b_{ij}$, odnosno kada je $a_{ij} \neq 0$ $a_{ij} < \alpha a_{ij} \leq \frac{a_{ij}+b_{ij}}{2} < b_{ij}$. Dakle, postoji $\alpha > 1$ takav da je $0 \leq A \leq \alpha A < B$. Iz Korolara 3.3.4 slijedi da je $\rho(B) \geq \alpha \rho(A)$, tako da ako je $\rho(A) \neq 0$, tada $\rho(B) > \rho(A)$. Ako je $\rho(A) = 0$, definirajmo matricu C koja na $(1,1)$ poziciji ima element jednak $b_{1,1} > 0$, a svi ostali elementi su jednaki nuli. Spektralni radijus te matrice je b_{11}, pa imamo $C = |C| \leq B$, pa je ponovo prema Korolaru 3.3.4 $\rho(B) \geq b_{11} > 0 = \rho(A)$. □

Peron je došao do važnih rezultata za pozitivne matrice, od kojih ćemo iznijeti neke od njih.

Teorem 3.3.6 ([12]). *Neka je A $n \times n$ matrica i pretpostavimo da je $A > 0$. Tada je $\rho(A) > 0$, $\rho(A)$ je svojstvena vrijednost od A, i postoji pozitivan vektor $v > 0$ takav da je $Av = \rho(A)v$.*

Dokaz: Iz Korolara 3.3.5 slijedi da je $\rho(A) > 0$. Prema definiciji spektralnog radijusa, postoji svojstvena vrijednost λ od A, sa svojstvom $|\lambda| = \rho(A)$. Neka je v odgovarajući svojstveni vektor, različit od nule. Imamo

$$\rho(A)|v| = |\lambda| \cdot |v| = |\lambda v| = |Av| \leq |A| \cdot |v| = A|v|,$$

tada je $y = A|v| - \rho(A)|v| \geq 0$. Ako je y nul-vektor, tada iz toga slijedi da je $\rho(A)$ svojstvena vrijednost od A sa nenegativnim svojstvenim vektorom $|v|$. Kada bi $|v|$ imao komponentu jednaku nuli, tada bi ta ista komponenta od $A|v|$ morala biti jednaka nuli, a budući da su svi elementi od A pozitivni, slijedi prema tvrdnji *(iv)* Leme 3.3.2, da je v nul-vektor, što je kontradikcija. Zbog toga, ako je $y = 0$ tvrdnja teorema je dokazana. Preostalo je još pokazati da je y uvijek jednak nuli.

Ako y nije nul-vektor, tada je ponovo prema tvrdnji *(iv)* Leme 3.3.2 $Ay > 0$. Označimo sa $z = A|v| > 0$, pa imamo $0 < Ay = Az - \rho(A)z$ ili $Az > \rho(A)z$. Slično kao kod dokaza Korolara 3.3.5, slijedi da postoji broj $\alpha > \rho(A)$ takav da je $Az \geq \alpha z$. Iz tvrdnje *(v)* Leme 3.3.2 slijedi da za svako $k \geq 1$, $A^k z \geq \alpha^k z$. Odavde zaključujemo da je $\|A^k\|_2^{1/k} \geq \alpha > \rho(A)$ za sve k. Budući da je $\lim_{k\to\infty} \|A^k\|_2^{1/k} = \rho(A)$, dobivamo kontradiktornu nejednakost $\rho(A) \geq \alpha > \rho(A)$. □

Vektor v iz ovog teorema se često naziva *Perronovim vektorom*, a $\rho(A)$ *Perronovim korijenom* od A.

U mnogim primjerima pojavljuju se nenegativne matrice, koje ne moraju nužno biti pozitivne, pa je poželjno proširiti tvrdnju prethodnog teorema i na ovakve matrice.

Teorem 3.3.7 ([12]). *Ako je A $n \times n$ matrica, i $A \geq 0$, tada je $\rho(A)$ svojstvena vrijednost od A, i postoji nenegativan vektor $v \geq 0$, sa $\|v\|_2 = 1$, takav da je $Av = \rho(A)v$.*

Dokaz: Za bilo koji $\epsilon > 0$, definirajmo $A(\epsilon) = [a_{ij} + \epsilon] > 0$. Neka je $v(\epsilon) > 0$ sa $\|v(\epsilon)\|_2 = 1$ Perronov vektor od $A(\epsilon)$, a $\rho(\epsilon)$ Perronov korijen. Budući da je skup vektora $v(\epsilon)$ sadržan u kompaktnom skupu $\{w : \|w\|_2 = 1\}$, postoji monotono padajući niz $\epsilon_1 > \epsilon_2 > \ldots$ sa $\lim_{k\to\infty} = 0$, takav da postoji $\lim_{k\to\infty} v(\epsilon_k) = v$ i da zadovoljava $\|v\|_2 = 1$ (To znamo, budući da ako definiramo niz $\tilde{\epsilon}_k = 1/k$, koji teži ka nuli, i za koji niz $v(\tilde{\epsilon}_k)$ leži u kompaktnom skupu, tada postoji podniz $\{\epsilon_k\}$ od $\tilde{\epsilon}_k$, koji teži k nuli i za koje je $v(\epsilon_k)$ konvergentan, jer niz u kompaktnom skupu ima konvergentan podniz.) Budući da je $v(\epsilon_k) > 0$, slijedi da je $v \geq 0$.

Prema Teoremu 3.3.3 niz brojeva $\{\rho(\epsilon_k)\}_{k=1,2,\ldots}$ je monotono padajući niz, ograničen odozdo s nulom. Prema tome, postoji $\rho = \lim_{k\to\infty} \rho(\epsilon_k)$, i kako je ponovo prema istom teoremu $\rho(\epsilon_k) \geq \rho(A)$, vrijedi da je i $\rho \geq \rho(A)$. Ali iz činjenice da je

$$Av = \lim_{k\to\infty} [A(\epsilon_k)v(\epsilon_k)] = \lim_{k\to\infty} [\rho(\epsilon_k)v(\epsilon_k)] = \rho v$$

i činjenice da v nije nul-vektor (inače ne bi imao normu jednaku 1), slijedi da je ρ svojstvena vrijednost od A, i $\rho \leq \rho(A)$. Dakle, mora biti $\rho = \rho(A)$. □

3.3.2 Uspoređivanje regularnih rastava

Prethodno iznjetu Perron–Frobeniusovu teoriju iskoristiti ćemo sada za uspoređivanje regularnih rastava kada je matrica sustava A "inverzno-pozitivna". Glavne rezultate ovog odjeljka dao je Varga u [37].

Definicija 3.3.8. *Za $n \times n$ realne matrice A, M i N, rastav $A = M - N$ je* regularni rastav *ako je M regularna sa $M^{-1} \geq 0$ i $M \geq A$ ($N \geq 0$).*

Teorem 3.3.9 ([12]). *Neka je $A = M - N$ regularni rastav matrice A, za koju je $A^{-1} \geq 0$. Tada*

$$\rho(M^{-1}N) = \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)} < 1.$$

Dokaz: Budući da je $M^{-1}A = I - M^{-1}N$ regularna matrica, slijedi da $M^{-1}N$ ne može imati svojstvenu vrijednost jednaku 1, jer bi u suprotnom, za svojstveni vektor $y \neq 0$ pridružen jedinici, bilo $0 = y - M^{-1}Ny = (I - M^{-1}N)y = M^{-1}Ay$, što je kontradikcija. Kako je $M^{-1}N \geq 0$, prema Teoremu 3.3.7 $\rho(M^{-1}N)$ je svojstvena vrijednost, pa zbog prethodne primjedbe, niti spektralni radijus od $\rho(M^{-1}N)$ ne može biti jednak 1. Iz istog teorema slijedi da postoji vektor $v \geq 0$, takav da je $M^{-1}Nv = \rho(M^{-1}N)v$. Nadalje, vrijedi

$$A^{-1}N = (I - M^{-1}N)^{-1}M^{-1}N,$$

tako da je

$$A^{-1}Nv = \frac{\rho(M^{-1}N)}{1 - \rho(M^{-1}N)}v. \tag{3.46}$$

Da je $\rho(M^{-1}N) > 1$, tada bi slijedilo da $A^{-1}Nv$ ima negativne komponente, što je nemoguće prema Lemi 3.3.2, jer je $A^{-1} \geq 0$, $N \geq 0$ i $v \geq 0$. Time smo dokazali da je $\rho(M^{-1}N) < 1$. Iz (3.46) slijedi da je $\rho(M^{-1}N)/(1 - \rho(M^{-1}N))$ svojstvena vrijednost od $A^{-1}N$, pa prema tome imamo

$$\rho(A^{-1}N) \geq \frac{\rho(M^{-1}N)}{1 - \rho(M^{-1}N)},$$

ili, ekvivalentno, budući da je $1 - \rho(M^{-1}N) > 0$,

$$\rho(M^{-1}N) \leq \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)}. \tag{3.47}$$

Sada imamo da je $A^{-1}N \geq 0$, pa ponovo prema Teoremu 3.3.7 slijedi da postoji vektor $w \geq 0$, takav da je $A^{-1}Nw = \rho(A^{-1}N)w$. Uz pomoć relacije

$$M^{-1}N = (A + N)^{-1}N = (I + A^{-1}N)^{-1}A^{-1}N,$$

imamo

$$M^{-1}Nw = \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)}w,$$

pa je $\rho(A^{-1}N)/(1 + \rho(A^{-1}N))$ svojstvena vrijednost od $M^{-1}N$. Prema tome slijedi da je

$$\rho(M^{-1}N) \geq \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)},$$

što zajedno sa (3.47) rezultira tvrdnjom teorema. □

Prema ovom teoremu, svaka metoda jednostavnih iteracija koja je dobivena iz regularnog rastava $A = M - N$, sa $G = M^{-1}N$ konvergira za svaku početnu iteraciju.

Korolar 3.3.10 ([12]). *Neka su $A = M_1 - N_1 = M_2 - N_2$ dva regularna rastava od A, pri čemu je $A^{-1} \geq 0$. Ako je $N_1 \leq N_2$, tada*

$$\rho(M_1^{-1}N_1) \leq \rho(M_2^{-1}N_2).$$

Dokaz: Definirajmo funkciju $f(x) = x/(1+x)$, tada je $\rho(M_i^{-1}N_i) = f(\rho(A^{-1}N_i))$, za $i = 1, 2$. Ako funkciju f deriviramo, tada vidimo da je $f'(x) = 1/(1+x)^2 > 0$ za $x \neq 0$, odnosno da je funkcija f strogo rastuća. Kako je $0 \leq N_1 \leq N_2$ i $A^{-1} \geq 0$, onda slijedi da je $0 \leq A^{-1}N_1 \leq A^{-1}N_2$, pa je, prema Korolaru 3.3.4 $\rho(A^{-1}N_1) \leq \rho(A^{-1}N_2)$. Zbog strogo rastuće momotonosti funkcije f, slijedi da je onda i $\rho(M_1^{-1}N_1) \leq \rho(M_2^{-1}N_2)$, čime smo pokazali tvrdnju korolara. □

Dosta često nije lako odrediti da li je inverz matrice sustava A nenegativan ili pozitivan, a što se zahtijeva kao uvjet u Korolaru 3.3.10. Zato ćemo iznijeti nekoliko ekvivalentnih uvjeta, koji garantiraju nenegativnost inverza matrice.

Definicija 3.3.11. *Za $n \times n$ matricu A kažemo da je M-matrica ako*

(i) $a_{ii} > 0$, $i = 1, \ldots, n$,

(ii) $a_{ij} \leq 0$, $i, j = 1, \ldots, n$, $j \neq i$, i

(iii) *A je regularna i $A^{-1} \geq 0$.*

Mnoge matrice, dobivene iz parcijalnih diferencijalnih jednadžbi, kao na primjer iz transportne jednadžbe, jesu M-matrice, pa zato ova definicija ima smisla. U sljedećem teoremu iznijet ćemo ostale ekvivalentne tvrdnje, prema kojima možemo prepoznati M-matricu, pa prema tome i garantirati da takve matrice imaju nenegativan inverz. Njegov dokaz nalazi se u [23, str. 114–116].

Teorem 3.3.12 ([23]). *Neka je A realna $n \times n$ matrica sa nepozitivnim vandijagonalnim elementima. Sljedeće tvrdnje su ekvivalentne:*

(i) A je M-matrica.

(ii) A je regularna i $A^{-1} \geq 0$. (Primijetimo da je uvjet (i) u Definiciji 3.3.11 suvišan.)

(iii) Sve svojstvene vrijednosti od A imaju pozitivan realni dio.

(iv) Svaka realna svojstvena vrijednost od A je pozitivna.

(v) Sve glavne minore od A su M-matrice.

(vi) A se može faktorizirati u oblik $A = LU$, gdje je L donje trokutasta, U gornje trokutasta, a svi dijagonalni elementi obaju matrica su pozitivni.

(vii) Dijagonalni elementi od A su pozitivni, i AD je strogo dijagonalno dominanatna za proizvoljnu pozitivnu, dijagonalnu matricu D.

Za realan simetričan slučaj imamo sljedeći rezultat.

Definicija 3.3.13. *Realna matrica A je* Stieltjesova matrica *ako je A simetrična, pozitivno definitna i svi vandijagonalni elementi od A su nepozitivni.*

Teorem 3.3.14 ([12]). *Svaka Stieltjesova matrica je M-matrica.*

Dokaz. Neka je A Stieltjesova matrica. Dijagonalni elementi od A su pozitivni jer je A pozitivno definitna, zato još trebamo samo provjeriti da je $A^{-1} \geq 0$. Rastavimo matricu A na $A = D - C$, pri čemu je $D = \text{diag}(A)$ sa pozitivnim elementima na dijagonali, a C je nenegativna. Budući da je A pozitivno definitna, ona je i regularna, i $A^{-1} = [D(I - B)]^{-1} = (I - B)^{-1}D^{-1}$, gdje je $B = D^{-1}C$. Ako je $\rho(B) < 1$, tada je inverz od $I - B$ dan Neumanovim redom

$$(I - B)^{-1} = I + B + B^2 + \cdots,$$

i budući da je $B \geq 0$, tada bi slijedilo da je $(I - B)^{-1} \geq 0$, pa zbog toga i $A^{-1} \geq 0$. Zbog toga, jedina stvar koju još moramo pokazati je da $\rho(B) < 1$.

Pretpostavimo da je $\rho(B) \geq 1$. Kako je $B \geq 0$, prema Teoremu 3.3.7 slijedi da je $\rho(B)$ svojstvena vrijednost od B. Ali tada $D^{-1}A = I - B$ mora imati nepozitivnu svojstvenu vrijednost $1 - \rho(B)$. Ta matrica je slična simetričnoj pozitivno definitnoj matrici $D^{-1/2}AD^{-1/2}$, pa smo dobili kontradikciju. Zbog toga mora biti $\rho(B) < 1$. □

Matrica dobivena iz difuzijske jednadžbe je Stieltjesova matrica, pa prema tome i M-matrica.

Na temelju prethodno danih rezultata, sada možemo dati nekoliko zaključaka u vezi stopa konvergencije klasičnih iterativnih metoda, baziranih na prekondicioniranim jednostavnim iteracijama.

Korolar 3.3.15. *Ako je matrica sustava A M-matrica, tada stopa konvergencije Gauss-Seidelove metode je najmanje tako dobra kao stopa Jacobijeve metode.*

Dokaz. Prema definicijama metoda, za rastav $A = D - L - U$ pri čemu je D dijagonalna, L strogo donje trokutasta i U strogo gornje trokutasta matrica, za Jacobijevu metodu imamo

$$M_J = D, \qquad N_J = L + U \geq 0,$$

gdje je $M_J^{-1} = D^{-1} > 0$ jer je $D > 0$, i za Gauss–Seidelovu metodu

$$M_{GS} = D - L, \qquad N_{GS} = U \geq 0,$$

gdje je M_{GS} donje trokutasta matrica, sa nepozitivnim vandijagonalnim elementima i pozitivnim dijagonalnim elementima, što znači sa pozitivnim realnim svojstvenim vrijednostima. Prema tvrdnji *(iv)* Teorema 3.3.12 ona je M-matrica pa stoga je i njen inverz nenegativan odnosno $M_{GS}^{-1} \geq 0$. Kako je $A = M_J - N_J = M_{GS} - N_{GS}$, obje metode definiraju regularne rastave. Prema pretpostavci korolara je $A^{-1} \geq 0$ i očito je $N_{GS} \leq N_J$, pa iz Korolara 3.3.10 slijedi da je

$$\rho(M_{GS}^{-1}N_{GS}) \leq \rho(M_J^{-1}N_J).$$

□

Ovdje možemo primijetiti i činjenicu da je M_{GS} bliži matrici A od M_J, odnosno Gauss–Seidelova matrica prekondicioniranja je u ovom smislu bolja od Jacobijeve.

Ako sada promatramo samo dijagonalne matrice prekondicioniranja M, čiji su elementi veći ili jednaki onima na dijagonali matrice A, zbog tvrdnje Korolara 3.3.10 možemo zaključiti da je Jacobijev rastav $M = \text{diag}(A)$ najbolji. Slično, ako razmatramo regularne rastave, u kojima se zahtijeva da matrica M ima određeni raspored

nula (na primjer, vrpčaste matrice), tada isti korolar nudi najbolji izbor za M. Sa stanovišta stope konvergencije jednostavnih iteracija najbolje je elemente matrice M, različite od nule uzeti takve, da budu jednaki onima iz A, na istoj poziciji.

Korolar 3.3.10 potvrđuje našu intuitivnu spoznaju, a to je da u specijalnom slučaju regularnih rastava inverzno-nenegativnih ili inverzno-pozitivnih matrica, što je matrica prekondicioniranja M bliža matrici sustava A, po komponentama, to je stopa konvergencije prekondicioniranih jednostavnih iteracija bolja.

3.3.3 Regularni rastavi i CG metoda

Za hermitske pozitivno definitne sustave, A norma greške kod CG metode, i euklidska norma reziduala kod MINRES metode, mogu se ograditi izrazom koji sadrži korijen broja uvjetovanosti matrice sustava A. Kod prekondicioniranih sustava, radi se o $L^{-1}AL^{-*}$-normi greške za CG metodu, pri čemu je $M = LL^*$, i o M^{-1} normi reziduala za MINRES metodu. Prema tome, kod mjerenja uspješnosti prekondicioniranja, umjesto spektralnog radijusa matrice $I - M^{-1}A$, kao kod jednostavnih iteracija, nas će sada zanimati broj uvjetovanosti od $L^{-1}AL^{-*}$, odnosno kvocijent najveće i najmanje svojstvene vrijednosti od $M^{-1}A$. Što je broj uvjetovanosti manji, brzina konvergencije je veća.

Teorem 3.3.16 ([12]). *Neka su A, M_1, i M_2 simetrične, pozitivno definitne, realne matrice, takve da zadovoljavaju uvjete Korolara 3.3.10, i pretpostavimo da je najveća svojstvena vrijednost od $M_2^{-1}A$ veća ili jednaka od 1. Tada kvocijenti najveće i najmanje svojstvene vrijednosti od $M_1^{-1}A$ i $M_2^{-1}A$ zadovoljavaju*

$$\frac{\lambda_{max}(M_1^{-1}A)}{\lambda_{min}(M_1^{-1}A)} < 2\frac{\lambda_{max}(M_2^{-1}A)}{\lambda_{min}(M_2^{-1}A)}. \tag{3.48}$$

Dokaz: Budući da su elementi od $M_2^{-1}N_2$ nenegativni, iz Teorema 3.3.7 slijedi da je spektralni radijus od $M_2^{-1}N_2$ jednak najvećoj svojstvenoj vrijednosti:

$$\rho(M_2^{-1}N_2) = \rho(I - M_2^{-1}A) = 1 - \lambda_{min}(M_2^{-1}A). \tag{3.49}$$

Isto vrijedi i za M_1. Iz rezultata Korolara 3.3.10 $\rho(M_1^{-1}N_1) \leq \rho(M_2^{-1}N_2)$ slijedi sljedeći niz nejednakosti. Kako je $1 - \lambda_{max}(M_1A)$ u spektru od $M_1^{-1}N_1$, vrijedi

$$\lambda_{max}(M_1^{-1}A) - 1 \leq |1 - \lambda_{max}(M_1^{-1}A)| \leq \rho(M_1^{-1}N_1) = 1 - \lambda_{min}(M_1^{-1}A),$$

pa zato imamo

$$1 - \lambda_{min}(M_1^{-1}A) \leq 1 - \lambda_{min}(M_2^{-1}A) \quad \text{i} \quad \lambda_{max}(M_1^{-1}A) - 1 \leq 1 - \lambda_{min}(M_2^{-1}A),$$

ili ekvivalentno,

$$\lambda_{min}(M_1^{-1}A) \geq \lambda_{min}(M_2^{-1}A) \quad \text{i} \quad \lambda_{max}(M_1^{-1}A) \leq 2 - \lambda_{min}(M_2^{-1}A).$$

Kako je $L^{-1}AL^{-*}$ pozitivno definitna i slična $M^{-1}A$, onda su sve svojstvene vrijednosti od $M_1^{-1}A$ i $M_2^{-1}A$ pozitivne. Zato, dijeljenjem druge nejednakosti s prvom dobivamo

$$\frac{\lambda_{max}(M_1^{-1}A)}{\lambda_{min}(M_1^{-1}A)} \leq \frac{\lambda_{max}(M_2^{-1}A)}{\lambda_{min}(M_2^{-1}A)} \left(\frac{2 - \lambda_{min}(M_2^{-1}A)}{\lambda_{max}(M_2^{-1}A)} \right).$$

Kako je, prema pretpostavci, $\lambda_{max}(M_2^{-1}A) \geq 1$, i kako je $\lambda_{min}(M_2^{-1}A) > 0$, drugi faktor desne strane nejednakosti je manji od 2, čime je teorem dokazan. □

Teorem 3.3.17 ([12]). *Pretpostavka u Teoremu 3.3.16, da je najveća svojstvena vrijednosto od $M_2^{-1}A$ veća ili jednaka od jedan, je zadovoljena ako i samo ako A i M_2 imaju najmanje jedan zajednički dijagonalni element.*

Dokaz: Ako A i M_2 imaju zajednički dijagonalni element, onda simetrična matrica N_2 mora imati nulu na dijagonali. Iz toga dalje slijedi da $M_2^{-1}N_2$ ima nepozitivnu svojstvenu vrijednost, jer prema Teoremu 1.6.6 najmanja svojstvena vrijednost te matrice zadovoljava

$$\lambda_{min}(M_2^{-1}N_2) = \min_{v\neq 0} \frac{v^*N_2v}{v^*M_2v} \leq \frac{\xi_j^*N_2\xi_j}{\xi_jM_2\xi_j} = \frac{(N_2)_{jj}}{(M_2)_{jj}} = 0,$$

pri čemu je ξ_j vektor sa 1 na j-toj poziciji, koja je jednaka poziciji nule na dijagonali matrice N_2, i sa nulama na ostalim pozicijama. Zbog toga, $M_2^{-1}A = I - M_2^{-1}N^{-1}$ ima svojstvenu vrijednost veću ili jednaku od 1. □

Teoremi 3.3.16 i 3.3.17 pokazuju da, kad je par regularnih rastava pravilno skaliran, što znači da je M_2 bio pomnožen s konstantom, ako je bilo potrebno, tako da A i M_2 imaju najmanje jedan zajednički dijagonalni element, tada onaj rastav koji je bio bliži matrici A po komponentama daje manji broj uvjetovanosti za prekondicioniranu matricu sustava, do na faktor 2. To znači da će Čebiševljeve ograde greške za svaki korak CG i MINRES metode biti manje, ili u gorem slučaju malo veće, za bližu matricu prekondicioniranja.

3.3.4 Optimalno dijagonalno i blok-dijagonalno prekondicioniranje

Uz regularne rastave, skoro jedina preostala klasa prekondicioniranja, za koje se može odrediti optimalna ili skoro optimalna matrica prekondicioniranja, je klasa dijagonalnih i blok-dijagonalnih prekondicioniranja. Ako se "optimalnost" definira kao posjedovanje malog broja uvjetovanosti simetrično prekondicionirane matrice, tada je (blok) dijagonala hermitske pozitivno definitne matrice A blizu najbolje (blok) dijagonalne matrice prekondicioniranja.

Definirat ćemo još jedno svojstvo matrice, slično svojstvu A.

Definicija 3.3.18. *Matrica A je* blok 2-ciklička *ako se ona može ispermutirati u oblik*

$$A = \begin{bmatrix} D_1 & B \\ C & D_2 \end{bmatrix},$$

gdje su D_1 i D_2 blok-dijagonalne matrice

$$D_i = \begin{bmatrix} D_{i,1} & & \\ & \ddots & \\ & & D_{i,m_i} \end{bmatrix}, \quad i = 1,2, \quad D_{i,j} \in \mathbb{C}^{n_{i,j}\times n_{i,j}}.$$

Sljedeći rezultat o optimalnosti blok-dijagonalne matrice prekondicioniranaj dao je Elsner.

Teorem 3.3.19 (Elsner [12]). *Ako hermitska pozitivno definitna matrica A ima oblik*

$$A = \begin{bmatrix} I_{n_1} & B \\ B^* & I_{n_2} \end{bmatrix}, \tag{3.50}$$

tada je $\kappa(A) \leq \kappa(D^*AD)$ *za bilo koju regularnu matricu* D *oblika*

$$\begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}. \qquad D_1 \in \mathbb{C}^{n_1 \times n_1}, \quad D_2 \in \mathbb{C}^{n_2 \times n_2}. \tag{3.51}$$

Dokaz: Ako je λ svojstvena vrijednost od A, sa svojstvenim vektorom $[v,\ w]^T$, čiji je prvi blok označen sa v, a drugi sa w, tada iz (3.50) slijedi

$$\begin{bmatrix} I & B \\ B^* & I \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \lambda \begin{bmatrix} v \\ w \end{bmatrix},$$

odnosno,

$$Bw = (\lambda - 1)v, \qquad B^*v = (\lambda - 1)w.$$

Odavde možemo zaključiti da je i $[v,\ -w]^T$ svojstveni vektor od A, sa svojstvenom vrijednosti $2 - \lambda$, jer je

$$A \begin{bmatrix} v \\ -w \end{bmatrix} = \begin{bmatrix} v - Bw \\ B^*v - w \end{bmatrix} = (2 - \lambda) \begin{bmatrix} v \\ -w \end{bmatrix}.$$

Iz pozitivne definitnosti matrice A slijedi, ako je λ_{max} najveća svojstvena vrijednost od A, tada je $\lambda_{min} = 2 - \lambda_{max}$ najmanja, a $\kappa(A) = \lambda_{max}/(2 - \lambda_{max})$.

Neka je

$$S = \begin{bmatrix} I_{n_1} & 0 \\ 0 & -I_{n_2} \end{bmatrix},$$

tako da je $S[v,\ w]^T = [v,\ -w]^T$. Ako je $Az = \lambda_{max} z$, tada je $SAz = \lambda_{max} Sz$ i

$$A^{-1}SAz = \frac{\lambda_{max}}{2 - \lambda_{max}} Sz, \qquad SA^{-1}SAz = \frac{\lambda_{max}}{2 - \lambda_{max}} z.$$

Prema tome, imamo

$$\rho(SA^{-1}SA) \geq \frac{\lambda_{max}}{2 - \lambda_{max}} = \kappa(A),$$

i za bilo koju regularnu matricu D, zbog toga što sličnost matrica čuva spektar, vrijedi

$$\kappa(A) \leq \rho(SA^{-1}SA) = \rho(D^{-1}SA^{-1}SAD) \leq \|D^{-1}SA^{-1}SAD\|_2. \tag{3.52}$$

Sada, ako je D oblika (3.51), tada S i D komutiraju. Također $\|S\|_2 = 1$, pa imamo

$$\begin{aligned} \|D^{-1}SA^{-1}SAD\|_2 &= \|S(D^{-1}A^{-1}D^{-*})S(D^*AD)\|_2 \leq \\ &\leq \|D^{-1}A^{-1}D^{-*}\|_2 \cdot \|D^*AD\|_2 = \kappa(D^*AD). \end{aligned} \tag{3.53}$$

Kombiniranjem (3.52) i (3.53) dobivamo traženi rezultat. □

Dakle, prema ovom teoremu, optimalna matrica prekondicioniranja oblika (3.51), u smislu najmanjeg broja uvjetovanosti, je matrica kojoj su dijagonalni blokovi jednaki odgovarajućim blokovima matrice A.

Pretpostavimo da A nije oblika (3.50), ali da se može ispermutirati u taj oblik. Neka je $A = P^T\tilde{A}P$, gdje je P matrica permutacije, i $\tilde{A}$ je oblika (3.50). Tada za bilo koju blok-dijagonalnu matricu D oblika (3.51), vrijedi

$$\kappa(D^*AD) = \kappa(PD^*P^T\tilde{A}PDP^T).$$

Ako je permutacija takva da je PDP^T ponovo blok-dijagonalna matrica oblika (3.51), tada je A, kao i $\tilde{A}$, optimalno skalirana matrica među takvim blok-dijagonalnim matricama. Naime, kad bi $\kappa(D^*AD)$ bio manji od $\kappa(A)$ za neku matricu D oblika (3.51), tada bi $\kappa(\tilde{D}^*\tilde{A}\tilde{D})$ bio manji od $\kappa(\tilde{A})$, za $\tilde{D} = PDP^T$, što je kontradikcija. Naročito, ako matrica A zadovoljava svojstvo A, tada je optimalna dijagonalna matrica prekondicioniranja jednaka $M = \text{diag}(A)$, sa $D = M^{1/2}$, jer budući da je D dijagonalna, onda je i PDP^T dijagonalna za bilo koju matricu permutacije. Ako je A zapisana u blok obliku, gdje se blokovi mogu ispermutirati u 2-blok ciklički oblik, ali bez permutiranja elemenata iz jednog bloka u drugi, tada je optimalna blok-dijagonalna matrica prekondicioniranja jednaka $M = \text{blok\_diag}(A)$, jer će tada PDP^T ostati blok dijagonalna matrica.

Dakle, iz Teorema 3.3.19 slijedi da u određenim slučajevima, kada je matrica sustava hermitska i pozitivno definitna, blok dijagonala matrice, s dva bloka, je najbolja blok-dijagonalna matrica prekondicioniranja, u smislu minimiziranja broja uvjetovanosti prekondicionirane matrice. Za proizvoljne hermitske pozitivno definitne matrice imamo još i sljedeće rezultate: od van der Sluisa, koji se odnosi na dijagonalne, i od Demmela, koji se odnosi na blok-dijagonalne matrice prekondicioniranja.

Teorem 3.3.20 (van der Sluis [12]). *Ako hermitska pozitivno definitna matrica A ima sve dijagonalne elemente jednake 1, tada*

$$\kappa(A) \leq m \cdot \min_{D \in \mathcal{D}} \kappa(DAD), \tag{3.54}$$

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

Dokaz: Izračunajmo faktorizaciju Choleskog matrice A, tako da je $A = U^*U$, pri čemu je U gornje trokutasta matrica. Budući da A ima sve dijagonalne elemente jednake 1, svaki stupac od U ima normu 1. Također, svaki vandijagonalni element od A ima apsolutnu vrijednost manju ili jednaku od 1, budući da je

$$|a_{ij}| = |u_i^* u_j| \leq \|u_i\|_2 \cdot \|u_j\|_2 = 1,$$

gdje su u_i i u_j i-ti i j-ti stupac od U. Nadalje, iz Gerschgorinovog teorema slijedi da najveća svojstvena vrijednost $\lambda_{max}(A)$ od A mora biti u nekom od Gercshgorinovih krugova, odnosno da postoji indeks i takav da je

$$\lambda_{max}(A) - 1 \leq |\lambda_{max}(A) - 1| \leq \sum_{j \neq i} |a_{ij}|.$$

Odavde imamo

$$\|A\|_2 = \lambda_{max}(A) \leq \max_i \sum_{j=1}^{n} |a_{ij}| \leq m.$$

Za bilo koju regularnu matricu D možemo napisati

$$\begin{aligned} \kappa(D^*AD) &= \|D^*AD\|_2 \cdot \|D^{-1}A^{-1}D^{-*}\|_2 \\ &= \|D^*U^*UD\|_2 \cdot \|D^{-1}U^{-1}U^{-*}D^{-*}\|_2 = \\ &= \|UD\|_2^2 \cdot \|D^{-1}U^{-1}\|_2^2. \end{aligned}$$

Sada je $\|D^{-1}U^{-1}\|_2^2 \geq \|U^{-1}\|_2^2/\|D\|_2^2 = \|A^{-1}\|_2/\|D\|_2^2$, tako da imamo

$$\kappa(D^*AD) \geq \|UD\|_2^2 \frac{\|A^{-1}\|_2}{\|D\|_2^2} = \kappa(A) \cdot \frac{\|UD\|_2^2}{\|A\|_2 \cdot \|D\|_2^2} \geq \frac{\kappa(A)}{m} \left(\frac{\|UD\|_2}{\|D\|_2} \right)^2. \tag{3.55}$$

Sada pretpostavimo da je D pozitivno definitna dijagonalna matrica, sa najvećim dijagonalnim elementom jednakim d_{jj}. Neka je ξ_j j-ti jediniči vektor. Tada je

$$\|UD\|_2 = \max_{\|v\|_2=1} \|UDv\|_2 \geq \|UD\xi_j\|_2 = \|d_{jj}u_j\|_2 = d_{jj} = \|D\|_2. \tag{3.56}$$

Kombiniranjem (3.55) i (3.56) dobivamo traženi rezultat. □

Teorem 3.3.21 (Demmel [6]). *Ako hermitska pozitivno definitna matrica A ima sve dijagonalne blokove jednake identiteti, odnosno, ako je oblika*

$$A = \begin{bmatrix} I_{n_1} & A_{12} & \dots & A_{1m} \\ A_{12}^* & I_{n_2} & \dots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1m}^* & A_{2m}^* & \dots & I_{n_m} \end{bmatrix},$$

tada

$$\kappa(A) \leq m \cdot \min_{D \in \mathcal{D}_B} \kappa(D^*AD),$$

gdje je $\mathcal{D}_B = \{regularne\ blok-dijagonalne\ matrice,\ sa\ blokovima\ reda\ n_1, \dots, n_m\}$, a m je broj dijagonalnih blokova u A.

Dokaz: Dokaz je vrlo sličan dokazu prethodnog teorema. Ponovo faktoriziramo matricu na $A = U^*U$, i matricu U particioniramo na

$$U = [U_1,\ U_2, \dots, U_m],$$

pri čemu su matrice U_i dimenzije $n \times n_i$, za $i = 1, \dots, m$. Tada vrijedi

$$U_i^*U_i = A_{ii} = I_{n_1}, \qquad i = 1, \dots, m,$$

pa je svaki od blokova U_i ortonormalan. Nadalje je,

$$\|A_{ij}\|_2 = \|U_i^*U_j\|_2 \leq \|U_i\|_2\|U_j\|_2 = 1,$$

za $i \neq j$. Prema blok verziji Gerschgorinovog teorema, vrijedi da za najveću svojstvenu vrijednost $\lambda_{max}(A) = \|A\|_2$ postoji indeks i, takav da je

$$\|(\lambda_{max}(A)I - A_{ii})^{-1}\|_2^{-1} \leq \sum_{j \neq i} \|A_{ij}\|_2.$$

Kako je $A_{ii} = I$ imamo

$$\|(\lambda_{max}(A)I - A_{ii})^{-1}\|_2^{-1} = [|\lambda_{max}(A) - 1|^{-1}\|I\|_2]^{-1} = |\lambda_{max}(A) - 1|,$$

pa na kraju dobivamo

$$\|A\|_2 \leq \sum_{j=1}^{m} \|A_{ij}\|_2 \leq m.$$

Kao i u prethodnom teoremu, za svaku regularnu matricu D vrijedi

$$\kappa(D^*AD) \geq \frac{\kappa(A)}{m} \left(\frac{\|UD\|_2}{\|D\|_2} \right)^2. \tag{3.57}$$

Sada pretpostavimo da je D regularna blok-dijagonalna matrica. Tada su singularne vrijednosti svakog od dijagonalnih blokova ujedno i singularne vrijednosti cijele matrice D. Uzmimo da je $\sigma_{max}(D)$ singularna vrijednost i-tog dijagonalnog bloka D_i, takva da je $\sigma_{max}(D) = \|D\|_2$. Tada postoje singularni vektori u i v, zapisani u blok obliku

$$u = [0 \ \ldots \ u_i \ 0 \ \ldots \ 0]^T, \qquad v = [0 \ \ldots \ 0 \ v_i \ 0 \ \ldots \ 0]^T,$$

kod kojih je i-ti blok dimenzije n_i netrivijalan, i kojima je norma jednaka 1, takvi da je $Dv = \sigma_{max}(D)u$. Tada je

$$\begin{aligned} \|UD\|_2 &= \max_{\|w\|_2=1} \|UDw\|_2 \geq \|UDv\|_2 = \\ &= \sigma_{max}(D)\|Uu\|_2 = \|D\|_2\|U_i u_i\|_2 = \|D\|_2. \end{aligned} \tag{3.58}$$

(3.57) zajedno sa (3.58) daje traženi rezultat. □

Tvrdnje ovih teorema mogu se na drugačiji način iskazati kao: za proizvoljnu hermitsku pozitivno definitnu matricu njena (blok) dijagonala je skoro optimalna (blok) dijagonalna matrica prekondicioniranja.

Na osnovi ovih rezultata, razumno je reći da bi se kod pozitivno definitnih matrica skoro uvijek trebalo koristiti barem dijagonalno prekondicioniranje zajedno sa primjenom CG ili MINRES metoda. Ponekad matrice koje dolaze u praksi iz konkretnih problema imaju dijagonalne elemente koji jako variraju u redu veličine. Tada i svojstvene vrijednosti također variraju po veličini, pa je uvjetovanost velika. Jednostavnim skaliranjem, što u biti i je dijagonalno prekondicioniranje, možemo prilično reducirati uvjetovanost matrice, uz minimalan trošak.

3.4 Nekompletne faktorizacije

Kao što smo vidjeli u prethodnom odjeljku, SSOR ili SGS matrice prekondicioniranja su oblika $M = L_M U_M$ gdje L_M i U_m imaju isti raspored nula kao donje trokutasti L i gornje trokutasti U dijelovi matrice sustava A. Ako izračunamo matricu greške $A - L_M U_M$, tada bi, na primjer za SGS, imali

$$A - L_M U_M = D - L - U - (I - LD^{-1})(D - U) = -LD^{-1}U.$$

Ako L_M ima isti raspored nula kao i L dio od A, i U_M ima isti raspored kao i U dio od A, postavlja se pitanje, da li postoje L_M i U_M takvi da rezultiraju manjom greškom, u nekom smislu, od gore navedene. Mi možemo, na primjer, pokušati pronaći nekompletnu faktorizaciju u kojoj matrica reziduala $A - L_M U_M$ ima nule na mjestima gdje A ima netrivijalne elemente. Pokazat ćemo da je to općenito moguće za ILU(0) faktorizaciju. Naime, mnoge direktne metode za rješavanje sustava koriste se raznim faktorizacijama, međutim za velike i rijetko popunjene matrice, koje se često pojavljuju kao rezultat diskretizacije parcijalne diferencijalne jednadžbe, te metode su neprektične i preskupe. Umjesto toga, tražit ćemo aproksimativne, nekompletne faktorizacije, $A \approx L_M U_M$, gdje su L_M i U_M rijetko popunjene donja i gornja trokutasta matrica. Produkt $M = L_M U_M$ tada možemo koristiti kao matricu prekondicioniranja, u očekivanju da će M^{-1} dobro aproksimirati inverz matrice A. Općenito, razni rasporedi se mogu specificirati, a matrice L_M i U_M se mogu definirati tako da zadovoljavaju određena svojstva. To nas vodi do općenitije klase nekompletnih faktorizacija, koje ćemo u kratko razložiti.

Nekompletnima smatramo one faktorizacije, kojima se tokom samog procesa faktorizacije određeni elementi zanemaruju. To na primjer mogu biti netrivijalni elementi u faktorizaciji na pozicijama u kojima originalna matrica sustava ima nulu. Takve se matrice prekondicioniranja tada uvijek ostavljaju u faktoriziranom obliku. Njihova efikasnost onda ovisi o tome kako dobro njihov inverz aproksimira A^{-1}. Medjutim, kod nekompletnih faktorizacija se može pojaviti problem, a to je da one mogu zakazati, zbog pokušaja djeljenja s nulom, ili mogu rezultirati indefinitnom matricom, zbog negativnog pivotnog elementa, iako potpuna faktorizacija iste matrice postoji i treba biti pozitivno definitna, kao na primjer kod faktorizacije Choleskog. Isto tako je važno razmotriti njihov trošak izvođenja. Čak i kad nekompletna faktorizacija postoji, broj operacija koji je sudjelovao u njenom dobivanju je najmanje jednak broju operacija koji bi sudjelovao u rješavanju sustava s takvom matricom, tako da trošak može biti jednak trošku jedne ili više iteracija iterativne metode. Takvi troškovi se mogu nadoknaditi ako znamo da će iterativna metoda zahtijevati puno iteracija, ili ako će ista matrica prekondicioniranja biti upotrijebljenja na nekoliko linearnih sustava.

Dakle, nekompletne faktorizacije mogu biti dane u raznim oblicima. Ako je $M = L_M U_M$ sa L_M i U_M regularnim trokutastim matricama, rješavanje sustava sa matricom M se svodi na standardni način, prvo supstitucijama u naprijed, a zatim povratnim supstitucijama. Međutim, često su nekompletne faktorizacije dane sa $M = (D_M + L_M)D_M^{-1}(D_M + U_M)$, pri čemu je D_M dijagonalna, a L_M i U_M su strogo trokutaste matrice. U tom slučaju možemo koristiti jedan od sljedećih ekvivalentnih oblika za rješavanje sustava $Mx = y$

$$(D_M + L_M)z = y, \qquad (I + D_M^{-1}U_M)x = z,$$

ili

$$(I + L_M D_M^{-1})z = z, \qquad (D_M + U_M)x = z.$$

3.4.1 Nekompletna LU faktorizacija (ILU)

Pretpostavimo da je A proizvoljna, rijetko popunjena matrica, čije elemente smo označili sa a_{ij}, $i, j = 1, \ldots, n$. Općenita nekompletna LU faktorizacija (ILU) računa rijetko popunjenu donje trokutastu matricu L_M i rijetko popunjenu gornje trokutastu matricu U_M, tako da rezidual $R = L_M U_M - A$ zadovoljava određena svojstva, kao na primjer, da ima nule na određenim pozicijama. Algoritam za dobivanje nekompletne LU faktorizacije može se dobiti iz postupka Gaussovih eliminacija uz izbacivanje određenih elemenata sa unaprijed određenih vandijagonalnih pozicija. Postavlja se sada pitanja egzistencije takve faktorizacije. Ona postoji u mnogim slučajevima, a mi ćemo prvo iznijeti Ky-Fanov rezultat koji govori o egzistenciji nekompletne LU faktorizacije za M-matrice.

Lema 3.4.1 ([32]). *Neka je $A = [a_{ij}]$ M-matrica, i neka je $A^{(1)} = [a_{ij}^{(1)}]$ matrica dobivena iz prvog koraka Gaussovih eliminacija. Tada je i $A^{(1)}$ M-matrica.*

Dokaz: Prvo, razmotrimo vandijagonalne elemente od $A^{(1)}$:

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}}.$$

Budući da su a_{ij}, a_{i1}, a_{1j} nepozitivni, a a_{11} je pozitivan, slijedi da je i $a_{ij}^{(1)} \leq 0$ za $i \neq j$.

Drugo, činjenica, da je $A^{(1)}$ regularna matrica, posljedica je standardne relacije Gaussovih eliminacija

$$A = L^{(1)}A^{(1)}, \quad \text{gdje je} \quad L^{(1)} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & 0 & \cdots & 0 \\ -\frac{a_{31}}{a_{11}} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{11}} & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{3.59}$$

Pa kako je A regularna, a prema upravo pokazanom je i $L^{(1)}$ regularna, tada slijedi da je i $A^{(1)}$ regularna.

Na kraju, pokazat ćemo da je $[A^{(1)}]^{-1}$ nenegativno tako da što ćemo ispitati predznake svakog stupca $[A^{(1)}]^{-1}e_j$ za $j = 1, \ldots, n$. Za $j = 1$ jasno je da je $[A^{(1)}]^{-1}e_1 = \frac{1}{a_{11}}$, zbog toga što prvi stupac od $A^{(1)}$ ima u prvom retku element jednak a_{11}, a u ostalima je nula. Za slučaj kada je $j \neq 1$, iz (3.59) slijedi

$$[A^{(1)}]^{-1}e_j = A^{-1}L^{(1)}e_j = A^{-1}e_j \geq 0.$$

Prema tome, svi stupci od $[A^{(1)}]^{-1}$ su nenegativni, i time smo dovršili dokaz, jer je prema tvrdnji *(ii)* Teorema 3.3.12 matrica $A^{(1)}$ M-matrica. □

Također, $(n-1) \times (n-1)$ matrica $A_{22}^{(1)}$ dobivena iz $A^{(1)}$ odstranjujući prvi redak i prvi stupac je isto M-matrica. Prvo, budući da dijagonalne kao i vandijagonalne elemente dijeli s matricom $A^{(1)}$, koja je prema upravo dokazanom teoremu M-matrica, i matrica $A_{22}^{(1)}$ ima dijagonalne elemente pozitivne, a vandijagonalne nepozitivne. Predznaci elemenata inverza te matrice su ponovo vidljivi iz strukture matrice $A^{(1)}$, koja u prvom stupcu ima svuda nule osim u prvom retku. Naime, inverz matrice $A_{22}^{(1)}$ je jednak $(n-1) \times (n-1)$ donjem dijagonalnom bloku inverza matrice $A^{(1)}$, čiji su elementi svi nenegativni.

Sljedeći rezultat dao je Varga.

Lema 3.4.2 (Varga [12]). *Ako je $A = [a_{ij}]$ M-matrica i ako elementi matrice $B = [b_{ij}]$ zadovoljavaju*

$$0 < a_{ii} \leq b_{ii}, \qquad a_{ij} \leq b_{ij} \leq 0 \text{ za } i \neq j,$$

tada je B također M-matrica.

Dokaz: Rastavimo matricu B na dijagonalni dio D i vandijagonalni dio $-C$, tada je $B = D - C = D(I - G)$, gdje je $D \geq 0$, $C, \geq 0$ i $G = D^{-1}C \geq 0$. Imamo $B^{-1} = (I - G)^{-1}D^{-1}$, i ako je $\rho(G) < 1$, tada

$$(I - G)^{-1} = I + G + G^2 + \cdots \geq 0,$$

odakle slijedi da je $B^{-1} \geq 0$, i da je zbog toga B M-matrica. Da bi vidjeli da je $\rho(G) < 1$, primijetimo prvo da ako rastavimo A kao $A = M - N$, gdje je $M = \text{diag}(A)$, tada je to regularni rastav, pa je prema Teoremu 3.3.9 $\rho(M^{-1}N) < 1$. Prema pretpostavci leme o elementima od B, slijedi da je $0 \leq G \leq M^{-1}N$, odakle je prema Korolaru 3.3.4

$$\rho(G) \leq \rho(M^{-1}N) < 1.$$

□

Neka je $\mathcal{P}$ podskup indeksa $\{(i,j) : \ j \neq i, \ i,j = 1,\ldots,n\}$. Indeksi u skupu $\mathcal{P}$ biti će oni indeksi, na čijim će pozicijama u nekompletnoj LU faktorizaciji elementi biti postavljeni na 0. Sljedeći teorem ne samo što dokazuje egzistenciju nekompletne LU faktorizacije, nego i demonstrira način na koji je možemo izračunati. Dokazali su ga Meijerink i van der Vorst, zajedno s njegovim korolarom.

Teorem 3.4.3 (Meijerink, van der Vorst [12]). *Ako je $A = [a_{ij}]$ $n \times n$ M-matrica, tada za bilo koji podskup $\mathcal{P}$ vandijagonalnih indeksa postoje donje trokutasta matrica $L = [l_{ij}]$ sa jediničnom dijagonalom i donje trokutasta matrica $U = [u_{ij}]$, takve da je $A = LU - R$, gdje*

$$\begin{aligned} l_{ij} &= 0, && \text{za } (i,j) \in \mathcal{P}, \ i > j, \\ u_{ij} &= 0, && \text{za } (i,j) \in \mathcal{P}, \ i < j, \\ r_{ij} &= 0, && \text{za } (i,j) \notin \mathcal{P}. \end{aligned}$$

Faktori L i U su jedinstveni, a rastav $A = LU - R$ je regularan rastav.

Dokaz: Dokaz se provodi kroz konstrukciju $n-1$ koraka analognih onima iz Gaussovih eliminacija. U k-tom koraku, prvo zamijenimo elemente tekuće matrice, s indeksima (k,j) i $(i,k) \in \mathcal{P}$, sa 0. Tada izvršimo korak Gaussovih eliminacija, tako da se eliminiraju elementi u redovima od $(k+1)$-og do n-tog, k-tog stupca, pribrajajući odgovarajuće multiple k-tog retka recima $k+1$ do n. Preciznije, definirajmo matrice

$$A^{(k)} = \left[a_{ij}^{(k)}\right], \quad \tilde{A}^{(k)} = \left[\tilde{a}_{ij}^{(k)}\right], \quad L^{(k)} = \left[l_{ij}^{(k)}\right], \quad R^{(k)} = \left[r_{ij}^{(k)}\right]$$

relacijama

$$A^{(0)} = A, \quad \tilde{A}^{(k)} = A^{(k-1)} + R^{(k)}, \quad A^{(k)} = L^{(k)}\tilde{A}^{(k)}, \quad k = 1,\ldots,n-1,$$

gdje je $R^{(k)}$ svugdje jednaka 0 osim na pozicijama $(k,j) \in \mathcal{P}$ i na pozicijama $(i,k) \in \mathcal{P}$, gdje je $r_{kj}^{(k)} = -a_{kj}^{(k-1)}$ i $r_{ik}^{(k-1)} = -a_{ik}^{(k-1)}$. Donje trokutasta matrica $L^{(k)}$ je svuda jednaka identiteti, osim k-tog stupca, koji je jednak

$$\left[0 \ \ldots \ 0 \ 1 \ -\frac{\tilde{a}_{k+1,k}^{(k)}}{\tilde{a}_{kk}^{(k)}} \ \ldots \ -\frac{\tilde{a}_{nk}^{(k)}}{\tilde{a}_{kk}^{(k)}}\right]^T.$$

Odavde se lako vidi da je $A^{(k)}$ matrica koja se dobije iz $\tilde{A}^{(k)}$ eliminiranjem elemenata iz k-tog stupca, uz pomoć elemenata iz k-tog retka, dok je $\tilde{A}^{(k)}$ dobivena iz $A^{(k-1)}$ zamjenom elemenata u k-tom stupcu i retku, čiji su indeksi u $\mathcal{P}$, sa 0.

Sada, $A^{(0)} = A$ je M-matrica, pa je $R^{(1)} \geq 0$. Iz Leme 3.4.2 slijedi da je $\tilde{A}^{(1)}$ M-matrica i, zbog toga je i $L^{(1)} \geq 0$. Iz Leme 3.4.1 slijedi da je $A^{(1)}$ M-matrica. Prema napomeni iza Leme 3.4.1 donji $(n-1) \times (n-1)$ dijagonalni blok od $A^{(1)}$ je također M-matrica, a sljedeći korak Gaussovih eliminacija će se izvesti upravo nad tim blokom. Tako možemo nastaviti redom za $k = 2,3,\ldots$, odnosno možemo, na prethodan način pokazati, da je za svako $k = 2,3,\ldots$ donji $(n-k) \times (n-k)$ dijagonalni blok matrice $A^{(k)}$ M-matrica. Trebamo još provjeriti kakva je situacija sa samom matricom $A^{(k)}$. Pretpostavimo, da je za $j = 1,\ldots,k-1$ $A^{(j)}$ M-matrica, te zapišimo $A^{(k)}$ u blok obliku

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

pri čemu je $A_{11}^{(k)}$ gornje trokutasti $k \times k$ blok, a $A_{22}^{(k)}$ je $(n-k) \times (n-k)$ blok. Kako su prvih k redaka matrice $A^{(k)}$ identični istim recima matrice $A^{(k-1)}$, osim k-tog u kojemu se neki elementi zamijenjeni sa nulom, a indukcijom smo pokazali da je i $A_{22}^{(k)}$ M-matrica, to znači da su svi dijagonalni elementi matrice $A^{(k)}$ pozitivni, a vandijagonalni elementi nepozitivni. Lako se može provjeriti da je inverz matrice $A^{(k)}$ oblika

$$(A^{(k)})^{-1} = \begin{bmatrix} (A_{11}^{(k)})^{-1} & -(A_{11}^{(k)})^{-1} A_{12}^{(k)} (A_{22}^{(k)})^{-1} \\ 0 & (A_{22}^{(k)})^{-1} \end{bmatrix},$$

Kako je je gornji $k \times k$ dijagonalni blok od $(A^{(k)})^{-1}$ jednak istom bloku od $(A^{(k-1)})^{-1}$, to znači da je $(A_{11}^{(k)})^{-1} \geq 0$. Nadalje je $A_{12}^{(k)} \leq 0$, a prema gore pokazanom $(A_{22}^{(k)})^{-1} \geq 0$, pa možemo zaključiti da je i $(A^{(k)})^{-1} \geq 0$, odnosno da je $A^{(k)}$ M-matrica. Analogno kao gore je i $\tilde{A}^{(k)}$ M-matrica, i $L^{(k)} \geq 0$, te $R^{(k)} \geq 0$ za $k = 1, \ldots, n-1$. Iz definicija odmah slijedi

$$\begin{aligned} L^{(k)} R^{(m)} &= R^{(m)} \quad \text{ako } k < m, \\ A^{(n-1)} &= L^{(n-1)} \tilde{A}^{(n-1)} = L^{(n-1)} A^{(n-2)} + L^{(n-1)} R^{(n-1)} = \cdots = \\ &= \left(\prod_{j=1}^{n-1} L^{(n-j)} \right) A^{(0)} + \sum_{i=1}^{n-1} \left(\prod_{j=1}^{n-i} L^{(n-j)} \right) R^{(i)}. \end{aligned}$$

Kombinirajući te dvije jednakosti dobivamo

$$A^{(n-1)} = \left(\prod_{j=1}^{n} L^{(n-j)} \right) \left(A + \sum_{i=1}^{n-1} R^{(i)} \right).$$

Definirajmo sada matrice $U = A^{(n-1)}$, $L = (\prod_{j=1}^{n} L^{(n-j)})^{-1}$, i $R = \sum_{i=1}^{n-1} R^{(i)}$. Tada su, prema već prije pokazanim svojstvima matrica $L^{(k)}$ i $R^{(k)}$, $LU = A+R$, $L^{-1} \geq 0$. U je M-matrica pa $U^{-1} \geq 0$, odakle slijedi da je $(LU)^{-1} \geq 0$, i na kraju $R \geq 0$, tako da je rastav $A = LU - R$ regularan. Preostalo nam je još pokazati jedinstvenost faktora L i U, što ćemo dobiti ako provjerimo samo jedinstvenost elemenata čiji indeksi nisu u $\mathcal{P}$, jer su oni, čiji indeksi jesu u $\mathcal{P}$, jednaki 0. To se lako može provjeriti matematičkom indukcijom po k-tom retku i stupcu, koristeći činjenicu da su elementi od A i LU jednaki za $(i,j) \notin \mathcal{P}$, da L ima jediničnu dijagonalu, te da je U M-matrica pa su joj svi dijagonalni elementi veći od 0. □

Korolar 3.4.4. *Ako je A simetrična M-matrica, tada za svaki podskup $\mathcal{P}$ vandijagonalnih indeksa, sa svojstvom da za $(i,j) \in \mathcal{P}$ slijedi $(j,i) \in \mathcal{P}$, postoji jedinstvena donje trokutasta matrica L sa $l_{ij} = 0$ ako je $(i,j) \in \mathcal{P}$ za $i > j$, takva da je $A = LL^T - R$, gdje je $r_{ij} = 0$ ako je $(i,j) \notin \mathcal{P}$. Rastav $A = LL^T - R$ je regularan rastav.*

Dokaz: Prema Teoremu 3.4.3 postoje jedinstvene matrice L_1, koja je donjetrokutasta s jediničnom dijagonalom, i U, koja je gornje trokutasta sa pozitivnom dijagonalom, takve da je $A = L_1 U - R$, i da su zadovoljena svojstva iz istog teorema. Ovdje treba naglasiti da je $u_{ij} = 0$ ako je $(j,i) \in \mathcal{P}$. Definirajmo $D = \text{diag}(U)$, koja je također jedinstvena. Tada je

$$A = A^T = U^T L_1^T - R^T = (U^T D^{-1})(D L_1^T) - R^T,$$

još jedan rastav kao u Teoremu 3.4.3, jer je $U^T D^{-1}$ donje trokutasta matrica sa jediničnom dijagonalom, koja ima 0 na istim pozicijama kao i L_1, a DL_1^T je gornje trokutasta matrica, koja ima 0 na istim pozicijama kao i U. Zbog jedinstvenosti faktora vrijedi $L_1 = U^T D^{-1}$, odnosno $U = DL_1^T$, pa ako sada definiramo $L = L_1 D^{1/2}$, imamo

$$A = L_1 DL_1^T - R = LL^T - R,$$

pri čemu je, zbog jedinstvenosti matrica L_1 i D, L jedinstven. □

U ovom slučaju radi se o *nekompletnoj faktorizaciji Choleskog* (IC).

Iako je nekompletna faktorizacija Choleskog regularni rastav, ona se ne može usporediti sa matricama prekondicioniranja, kao što su dijagonala od A ili njen donje trokutasti dio (Korolar 3.3.10, i Teorem 3.3.16), zbog toga što su neki elementi matrice prekondicioniranja nekompletene faktorizacije Choleskog $M = LL^T$ bliže odgovarajućim elementima iz A, nego što su odgovarajući elementi od diag(A) ili njenog donje trokutastog dijela, ali neki elementi su puno udaljeniji. Ipak, mnogi numerički primjeri pokazuju da nekompletna faktorizacija Choleskog primijenjena na CG algoritam, često zahtijeva značajno manje iteracija od dijagonalnog prekondicioniranja. S druge strane, svaka iteracija zahtijeva više operacija, pa se ukupni profit upotrebe nakompletne faktorizacije Choleskog treba gledati kroz omjer smanjenja broja iteracija i porasta broja operacija u svakoj iteraciji.

ILU(0)

Nekompletna LU faktorizacija bez popunjavanja, koju označavamo sa ILU(0), je ona u kojoj se podskup vandijagonalnih indeksa $\mathcal{P}$ točno poklapa sa podskupom indeksa vandijagonalnih elemenata koji su, kod matrice A, jednaki 0. Drugim riječima, raspored nula kod L i U, matrica koje su dobivene nakompletnom LU faktorizacijom, je ekvivalentan rasporedu nula kod matrice A.

ILU(p)

Točnost ILU(0) nekompletne faktorizacije, može biti nedovoljna za postizanje željene stope konvergencije, budući da, na primjer, kod rijetko popunjenih matrica, produkt LU, može imati puno više netrivijalnih elemenata od polazne matrice A. Točnije verzije nepotpune LU faktorizacije dozvoljavaju određeno "popunjavanje". Naime, za razliku od ILU(0), takve faktorizacije dozvoljavaju da matrice L i U imaju netrivijalne elemente na pozicijama, na kojima A ima 0, odnosno na pozicijama u skupu $\mathcal{P}$. U tu svrhu uvodimo pojam *stupnja popunjavanja*. Stupanj popunjavanja se pridružuje svakom elementu tokom procesa Gaussovih eliminacija, a da li će taj element biti sveden na nulu ili ne, ovisi o njegovoj vrijednosti. Pri tome vrijedi pravilo: što je stupanj veći, to je element manji.

Definicija 3.4.5. *Početna vrijednost stupnja popunjavanja elementa a_{ij} rijetko popunjene matrice A je definirana sa*

$$lev_{ij} = \begin{cases} 0 & \text{ako } a_{ij} \neq 0, \text{ ili } i = j \\ \infty & \text{inače.} \end{cases}$$

Nakon k-tog koraka Gaussovih eliminacija u kojem se element na poziciji (i, j) promijenio, njegov stupanj popunjavanja se također treba ponovo izračunati sa

$$lev_{ij} = \min\{lev_{ij},\ lev_{ik} + lev_{kj} + 1\}. \tag{3.60}$$

Primijetimo da se stupanj popunjavanja nikad neće povećati tokom eliminacija. Prema tome, ako je $a_{ij} \neq 0$ u originalnoj matrici A, tada će element na poziciji (i, j) imati stupanj popunjavanja jednak nuli, kroz cijeli postupak Gaussovih eliminacija. Gornja definicija daje prirodnu strategiju za transformaciju elemenata u nulu. Kod ILU(p) faktorizacije, svi elementi kojima stupanj popunjavanje ne nadmašuje p, su zadržani. Dakle, raspored nula za ILU(p) faktorizaciju je dan sa

$$\mathcal{P}_p = \{(i, j) : \ lev_{ij} > p\},$$

pri čemu je lev_{ij} stupanj popunjavanja nakon što su sve transformacije (3.60) izvršene. ILU(0) je podudaran sa slučajem kada je $p = 0$.

Postoje mnogi nedostaci ovakve nepotpune faktorizacije. Prvo, broj operacija za izvršavanje ILU(p) nije predvidiv za $p > 0$. Drugo, trošak za računanje stupnjeva popunjavanja može biti visok. Što je najvažnije, stupanj popunjavanja za indefinitne matrice ne mora biti dobar pokazatelj elemenata koje treba svesti na nulu. Može se dogoditi da algoritam ILU(p) faktorizacije odbaci veliki element (svede ga na 0), i kao rezultat dobivamo netočnu nekompletnu faktorizaciju, u smislu da $R = LU - A$ nije malen. U prosjeku, to će uzrokovati velikim brojem iteracija iterativne metode, sa kojom se primijenjuje nekompletna faktorizacija, do postignuća konvergencije. Da bi se popravili ovi nedostaci, uvedene su tehnike koje daju nekompletnu faktorizaciju sa malom greškom R i sa kontroliranim brojem elemenata koji nisu svedeni na 0.

ILUT

Nekompletne faktorizacije koje se oslanjaju na stupanj popunjavanja su slijepe na numeričke vrijednosti, budući da elementi koji se svode na nulu ovise o strukturi matrice A. Postoje nekoliko alternativnih metoda baziranih na Gaussovim eliminacijama i svođenju elemenata na nulu i to na osnovu njihove veličine a ne pozicije. Kod tih tehnika raspored nula se određuje dinamički.

Generička *ILU faktorizacija sa pragom* može se dobiti iz Gaussovih eliminacija, uz dodatak skupa pravila za svođenje malih elemenata na nulu, što znači da će element biti sveden na nulu ako zadovoljava određeni skup kriterija. Tako dobivenu faktorizaciju označavamo sa ILUT(p,τ), i kod nje su korištena sljedeća pravila.

1. U k-tom koraku Gaussovih eliminacija, pivotni element $a_{ik} = a_{ik}/a_{kk}$, za $i > k$ se svodi na nulu ako je manji od relativne tolerancije τ_i, dobivene množenjem τ sa normom originalnog i-tog retka.

2. Nakon prvih $k - 1$ koraka Gaussovih eliminacija, k-ti redak se neće više mijenjati, pa su u njemu smješteni gotovi elementi k-tih redaka matrica L i U. Prvo treba svesti na nulu sve elemente u k-tom retku koji su manji od relativne tolerancije τ_k. Tada, treba zadržati samo p najvećih elemenata u L dijelu retka, i p najvećih elemenata u U dijelu retka, uz dijagonalni element koji se uvijek zadržava. Svi ostali elementi se svode na nulu.

Cilj drugog koraka svođenja elemenata na nulu je kontrola broja netrivijalnih elemenata po retku. Grubo rečeno, na p možemo gledati kao na parametar koji pomaže u kontroli troška memorije, dok τ reducira troškove računanja.

3.4.2 Nekompletan Gram–Schmidt i IQR

Kod prekondicioniranja normalnih jednadžbi, postavlja se pitanje kako prekondicionirati rezultirajuće iteracije. ILU matrica prekondicioniranja može se izračunati za matricu A, i onda pomoću nje riješiti prekondicioniranu normalnu jednadžbu

$$A^T(LU)^{-T}(LU)^{-1}Ax = A^T(LU)^{-T}(LU)^{-1}b.$$

Međutim, znamo da ILU nekompletna faktorizacija ne mora uvijek postojati, a i ako postoji, dobivena matrica prekondicioniranja ne mora biti kvalitetna. Zato ćemo promatrati još jedan način prekondicioniranja, svojstven normalnim jednadžbama.

Razmotrimo općenitu rijetko popunjenu matricu A i sa $a_1, \ldots, a_n$ označimo njene stupce. Ako definiramo (kompletnu) QR faktorizaciju matrice A sa

$$A = QR,$$

gdje je R gornje trokutasta matrica, a Q je unitarna. Tada je R faktor gornje faktorizacije faktor faktorizacije Choleskog matrice A^TA, jer ako je $A = QR$, pri čemu R ima pozitivne dijagonalne elemente, tada je

$$B = A^TA = R^TQ^TQR = R^TR.$$

Iz jedinstvenosti faktorizacije Choleskog sa faktorom R koji ima pozitivne dijagonalne elemente, slijedi da je R zaista jednak faktoru faktorizacije Choleskog matrice B. Taj odnos se može iskoristiti za dobivanje matrice prekondicioniranja za normalne jednadžbe.

Postoje dva načina dobivanja matrice R. Prvi je eksplicitno formiranje matrice B i izračunavanje njene faktorizacije Choleskog. To znači formiranje matrice A^TA koja može biti puno gušće popunjenija od matrice A. Drugi pristup je korištenje Gramm-Schmidtovog postupka. Na prvi pogled ovo bi mogao biti nepoželjan izbor, zbog gubitka numeričke točnosti kod ortogonalizacije velikog broja vektora. Međutim, kako vektori ostaju rijetko popunjeni kod nekompletne QR faktorizacije, takvi gubici točnosti će biti svedeni na minimum. Sa strane memorije, Gram-Schmidt je optimalan jer ne zahtijeva dodatnu memoriju za spremanje dodatnih podataka.

Kako bi definirali nekompletnu QR faktorizaciju IQR, moramo definirati strategiju po kojoj ćemo svoditi elemente na nulu, sličnu onoj kod ILU faktorizacija. To se može napraviti na sljedeći način. Neka su $\mathcal{P}_Q$ i $\mathcal{P}_R$ podskupovi indeksa koji će predstavljati raspored nula za matrice Q i R. Jedina restrikcija za R je

$$\mathcal{P}_R \subset \{(i,j) : \ i \neq j\}.$$

Što se tiče skupa $\mathcal{P}_Q$, za svaki stupac matrice Q mora postojati barem jedan netrivijalan element, odnosno

$$\{j : \ (i,j) \in \mathcal{P}_Q\} \neq \{1, 2, \ldots, n\}, \qquad \text{za } i = 1, \ldots, n.$$

Ova dva skupa se mogu definirati na različite načine, kao na primjer kod ILUT faktorizacije, dinamički.

Svodjenje elemenata na nulu, čiji indeksi se nalaze u ova dva skupa, izvodi se na sljedeći način. U k-tom koraku Gram–Schmidtovog postupka, nakon što se izračunaju vandijagonalni elementi k -tog stupca matrice R (skalarni produkti koji predstavljaju koeficijente ortogonalizacije), na nulu se svode elementi tog stupca čiji se indeksi nalaze

u skupu $\mathcal{P}_R$. Zatim, nakon izračunavanja vektora q_k, sa dobivenim koeficijentima ortogonalizacije, na nulu se svode one komponenete tog vektora, čiji indeksi se nalaze u skupu $\mathcal{P}_Q$. Tada imamo

$$a_k = \sum_{j=1}^{k} r_{jk} q_j + p_k, \tag{3.61}$$

gdje je p_k stupac elemenata koji su izbačeni iz stupca q_k. Gornja jednakost se svodi na

$$A = QR + P, \tag{3.62}$$

gdje je P matrica kojoj je k-ti stupac jednak p_k. Slučaj u kojem je $\mathcal{P}_Q$ prazan skup je posebno zanimljiv, jer je tada $P = 0$, pa imamo egzaktnu jednakost $A = QR$. U svakom slučaju Q nije općenito unitarna, jer su neki elementi u R svedeni na nulu. Kad bi u k-tom koraku bilo $r_{kk} = 0$, tada bi iz (3.61) slijedilo da je a_k linearna kombinacija stupaca $q_1, \ldots, q_{k-1}$, odakle indukcijom slijedi da je a_k linearna kombinacija stupaca $a_1, \ldots, a_{k-1}$, što ne može biti istina za regularnu matricu A. Kao rezultat, možemo iznijeti sljedeći teorem.

Teorem 3.4.6 ([32]). *Ako je A regularna matrica i ako je $\mathcal{P}_Q = \emptyset$ tada nekompletna QR faktorizacija $A = QR$ postoji, u kojoj je Q regularna matrica, a R je gornje trokutasta matrica sa pozitivnim dijagonalnim elementima.*

3.5 Aproksimacije inverza matrice

Načini prekondicioniranja navedeni u prethodnom odjeljku ponekad mogu zakazati, ako npr. nekompletna LU faktorizacija ne postoji, ili je dobivena matrica prekondicioniranja daleko od optimalne. Iz tog razloga, nastojalo se pronaći način prekondicioniranja koji ne zahtijeva direktno rješavanje linearnog sustava, već se polazni sustav može prekondicionirati matricom M^{-1} koja je direktna aproksimacija inverza od A.

Jednostavna tehnika pronalaženja aproksimacije inverza proizvoljne rijetko popunjene matrice je pronalaženje rijetko popunjene matrice $N = M^{-1}$ koja minimizira Frobeniusovu normu rezidualne matrice $I - AN$,

$$F(N) = \|I - AN\|_F^2. \tag{3.63}$$

Matricu N, čija vrijednost $F(N)$ je dovoljna mala, možemo uzeti kao aproksimaciju desnog inverza (kao matricu desnog prekondicioniranja) od A. Slično, aproksimacija lijevog inverza može se definirati korištenjem objektne funkcije

$$F(N) = \|I - NA\|_F^2. \tag{3.64}$$

Na kraju, možemo tražiti par L i U, takav da minimizira

$$F(L, U) = \|I - LAU\|_F^2. \tag{3.65}$$

Nadalje, ćemo promatrati samo slučaj (3.63), dok se ostala dva slučaja mogu analizirati na sličan način.

Pristup *globalnih iteracija* tretira matricu N kao nepoznatu, rijetko popunjenu matricu, i metodom tipa najbržeg silaska minimizira objektnu funkciju (3.63). Ta funkcija

je kvadratna funkcija definirana na prostoru $n \times n$ matrica, kojeg možemo identificirati sa $\mathbb{R}^{n^2}$. Skalarni produkt na prostoru matrica je definiran sa

$$\langle X, Y \rangle = \mathrm{tr}(Y^T X), \tag{3.66}$$

a kvadrat odgovarajuće norme je upravo u skladu sa definicijom objektne funkcije (3.63), jer je $\|X\|_F^2 = \langle X, X \rangle$.

Definicija 3.5.1. Poljska reprezentacija n^2*-dimenzionalnog vektora* $X = [x_1\ x_2\ \ldots\ x_{n^2}]^T$ *predstavlja* $n \times n$ *matricu*

$$\begin{bmatrix} x_1 & x_{n+1} & \ldots & x_{n^2-n+1} \\ x_2 & x_{n+2} & \ldots & x_{n^2-n+2} \\ \vdots & \vdots & & \vdots \\ x_n & x_{2n} & \ldots & x_{n^2} \end{bmatrix}.$$

U algoritmu silaska nova aproksimacija N_k se definira izvođenjem koraka duž smjera G_{k-1}, odnosno

$$N_k = N_{k-1} + \alpha_{k-1} G_{k-1},$$

gdje je α_{k-1} izabran tako da minimizira objektnu funkciju $F(N_k)$. Kao što smo vidjeli kod izvoda Orthomin metode, minimiziranje norme reziduala je ekvivalentno zahtjevu da je $R_k = I - AN_k = R_{k-1} - \alpha_{k-1} AG_{k-1}$ okomito na AG_{k-1}, obzirom na skalarni produkt $\langle \cdot, \cdot \rangle$, pri čemu je $R_{k-1} = I - AN_{k-1}$. Prema tome, optimalno α_{k-1} je dano sa

$$\alpha_{k-1} = \frac{\langle R_{k-1}, AG_{k-1} \rangle}{\langle AG_{k-1}, AG_{k-1} \rangle} = \frac{\mathrm{tr}(R_{k-1}^T AG_{k-1})}{\mathrm{tr}((AG_{k-1})^T AG_{k-1})}. \tag{3.67}$$

Rezultirajuća matrica N_k će postajati sve gušća i gušća nakon svakog koraka metode silaska, zato je važno primijeniti neku strategiju svođenja elemenata na nulu. Međutim, u tom slučaju je svojstvo silaska narušeno, odnosno, nećemo više moći garantirati da je $F(N_k) \leq F(N_{k-1})$.

Najjednostavniji izbor za smjer silaska G_{k-1}, je da on bude jednak matrici reziduala $R_{k-1} = I - AN_{k-1}$, gdje je N_{k-1} zadnja aproksimacija. Osim koraka u kojem se elementi svode na nulu, odgovarajući algoritam silaska liči upravo na Orthomin(1) metodu, primijenjenu na $n^2 \times n^2$-dimenzionalni linearni sustav $AN = I$. *Metoda globalnog minimalnog reziduala* ima sljedeći oblik.

Algoritam 3.5.2. ALGORITAM GLOBALNOG MINIMALNOG REZIDUALA

Dana je početna iteracija N_0.

Za $k = 1, 2, \ldots$

$G_{k-1} = I - AN_{k-1}$,

Izračunaj AG_{k-1},

$$\alpha_{k-1} = \frac{\mathrm{tr}(G_{k-1}^T AG_{k-1})}{\|AG_{k-1}\|_F^2},$$

$N_k = N_{k-1} + \alpha_{k-1} G_{k-1}$,

Svedi odgovarajuće elemente od N_k *na nulu.*

Drugi izbor za matricu G_{k-1} bi bio da je ona jednaka smjeru najbržeg silaska, odnosno smjeru suprotnom gradijentu funkcije (3.63), s obzirom na varijablu N_{k-1}. Ako su svi vektori prezentirani kao 2-dimenzionalna $n \times n$ polja, tada se gradijent može definirati kao matrica G_{k-1}, koja zadovoljava sljedeću relaciju, za malu perturbaciju E,

$$F(N_{k-1} + E) = F(N_{k-1}) + \langle G_{k-1}, E\rangle + \mathcal{O}(\|E\|^2). \tag{3.68}$$

Time možemo gradijent shvatiti kao operator nad matricama, a ne kao opereator nad n^2-dimenzionalnim vektorima.

Teorem 3.5.3 ([32]). *Poljska reprezentacija gradijenta funkcije F, definirane sa (3.63), je matrica*

$$G_{k-1} = -2A^T R_{k-1}$$

gdje je $R_{k-1} = I - AN_{k-1}$ matrica reziduala.

Dokaz: Za bilo koju matricu E imamo

$$\begin{aligned} F(N_{k-1} + E) - F(N_{k-1}) &= \operatorname{tr}[(I - A(N_{k-1} + E))^T (I - A(N_{k-1} + E))] - \\ &\quad - \operatorname{tr}[(I - AN_{k-1})^T (I - AN_{k-1})] = \\ &= \operatorname{tr}[(R_{k-1} - AE)^T (R_{k-1} - AE) - R_{k-1}^T R_{k-1}] = \\ &= -\operatorname{tr}[(AE)^T R_{k-1} + R_{k-1}^T AE - (AE)^T (AE)] = \\ &= -2\operatorname{tr}(R_{k-1}^T AE) + \operatorname{tr}[(AE)^T (AE)] = \\ &= -2\langle A^T R_{k-1}, E\rangle + \langle AE, AE\rangle. \end{aligned}$$

Kako je

$$\langle AE, AE\rangle = \|AE\|_F^2 \leq \|A\|_F^2 \|E\|_F^2,$$

iz (3.68) slijedi traženi rezultat. □

Dakle, algoritam najbržeg silaska imati će $G_{k-1} = A^T R_{k-1}$ i ima sljedeći oblik.

Algoritam 3.5.4. ALGORITAM GLOBALNOG NAJBRŽEG SILASKA

Dana je početna iteracija N_0.

Za $k = 1, 2, \ldots$

$R_{k-1} = I - AN_{k-1}$,
$G_{k-1} = A^T R_{k-1}$,
izračunaj AG_{k-1},
$\alpha_{k-1} = \dfrac{\|G_{k-1}\|_F^2}{\|AG_{k-1}\|_F^2}$,
$N_k = N_{k-1} + \alpha_{k-1} G_{k-1}$,
Svedi odgovarajuće elemente od N_k na nulu.

Prvo teoretsko pitanje, koje se postavlja, je da li aproksimacije inverza, dobivene gornjim algoritmom, mogu biti singularne. Ne može se dokazati da je N_k regularna, ukoliko aproksimacija nije dovoljno točna. Taj zahtjev može biti u konfliktu sa zahtjevom da je N_k rijetko popunjena matrica.

Teorem 3.5.5 ([32]). *Pretpostavimo da je A regularna matrica i da rezidual aproksimacije inverza N zadovoljava relaciju*

$$\|I - AN\| < 1, \tag{3.69}$$

gdje je $\|\cdot\|$ bilo koja konzistentna norma. Tada je N regularna.

Dokaz: Rezultat odmah slijedi iz jednakosti

$$AN = I - (I - AN) = I - R.$$

Budući da je $\|R\| < 1$, tada prema Lemi 1.5.9 slijedi da je $I - R$ regularna matrica, pa je prema tome i N regularna. □

Ponekad se dogodi da je R velik, pa skaliranje može dati manju normu. Rezultat prethodnog teorema, zato, lako možemo proširiti na sljedeći način. Ako je A regularna matrica, i ako postoje dvije regularne dijagonalne matrice D_1 i D_2, takve da je

$$\|I - D_1 A D_2\| < 1,$$

gdje je $\|\cdot\|$ bilo koja konzistentna norma, tada je N regularna matrica.

Sljedeće, ispitajmo popunjenost matrice N za slučaj kada za svaki stupac od N vrijedi pretpostavka

$$\|\xi_j - An_j\|_1 \leq \tau_j, \tag{3.70}$$

gdje je ξ_j j-ti jedinični vektor, a n_j j-ti stupac od N.

Teorem 3.5.6 ([32]). *Neka je $B = A^{-1}$, i pretpostavimo da dani element b_{ij} od B zadovoljava nejednakost*

$$|b_{ij}| > \tau_j \max_{k=1,\ldots,n} |b_{ik}|, \tag{3.71}$$

tada je element n_{ij} različit od nule.

Dokaz: Iz jednakosti $AN = I - R$ imamo da je $N = A^{-1} - A^{-1}R$, pa odavde slijedi

$$n_{ij} = b_{ij} - \sum_{k=1}^{n} b_{ik} r_{kj},$$

za r_{kj} (k, j)-ti element od R. Zato je

$$\begin{aligned} |n_{ij}| &\geq |b_{ij}| - \sum_{k=1}^{n} |b_{ik} r_{kj}| \geq \\ &\geq |b_{ij}| - \max_{k=1,\ldots,n} |b_{ik}| \|r_j\|_1 \geq \\ &\geq |b_{ij}| - \max_{k=1,\ldots,n} |b_{ik}| \tau_j. \end{aligned}$$

Sada iz uvjeta (3.71) slijedi $|n_{ij}| > 0$. □

Prema ovom teoremu možemo zaključiti da, ukoliko je R dovoljno mali, odnosno kad su konstante τ_j dovoljno male, tada su elementi od N različiti od nule smješteni na pozicijama, koje odgovaraju pozicijama većih elemenata inverza od A. Sljedeći negativni rezultat je direktna poslijedica prethodnog teorema.

Korolar 3.5.7 ([32]). *Neka je* $\tau = \max_{j=1,\ldots,n} \tau_j$. *Ako netrivijalni elementi od* $B = A^{-1}$ *zadovoljavaju nejednakost*

$$|b_{ij}| > \tau \max_{k,l=1,\ldots,n} |b_{kl}|,$$

tada raspored netrivijalnih elemenata od N *sadrži i raspored netrivijalnih elemenata od* A^{-1}. *Naročito, ako je* A^{-1} *gusto popunjena, i ako za njene netrivijalne elemente vrijedi gornja nejednakost, tada je i* N *također gusto popunjena.*

Što je manji τ, to je vjerojatnije da će uvjet korolara biti ispunjen. Drugi način na koji možemo iznijeti tvrdnju korolara je, da se točna i rijetko popunjena aproksimacija inverza može izračunati samo ako elementi pravog inverza jako variraju u veličini. Na žalost, to je jako teško unaprijed provjeriti.

3.6 Primjer: difuzijska jednadžba

Kod proučavanja matrica prekondicioniranja, nakon što smo naveli različite načine na koje se ona može izabrati, korisno je pogledati kako se to može napraviti na konkretnom primjeru. Numeričko rješavanje difuzijske jednadžbe dati će simetričan pozitivno definitan linearan sustav, koji je značajan sa fizikalnog gledišta i koji daje prilično dobru ilustraciju principa koje smo naveli u ovom poglavlju.

Određeni broj različitih fizikalnih procesa može se opisati difuzijskom jednadžbom

$$\frac{\partial u}{\partial t} - \nabla \cdot (a \, \nabla u) = f \quad \text{ na } \Omega. \tag{3.72}$$

Ovdje $u(x,t)$ može, na primjer, predstavljati distribuciju temperature po vremenu t u objektu Ω, na koji djeluje vanjski izvor topline $f(x,t)$. Pozitivan koeficijent $a(x)$ je toplinski konduktivitet materijala. Da bi odredili temperaturu u vremenu t, trebamo znati početnu distribuciju topline $u(x,0)$ i neki rubni uvjet, recimo

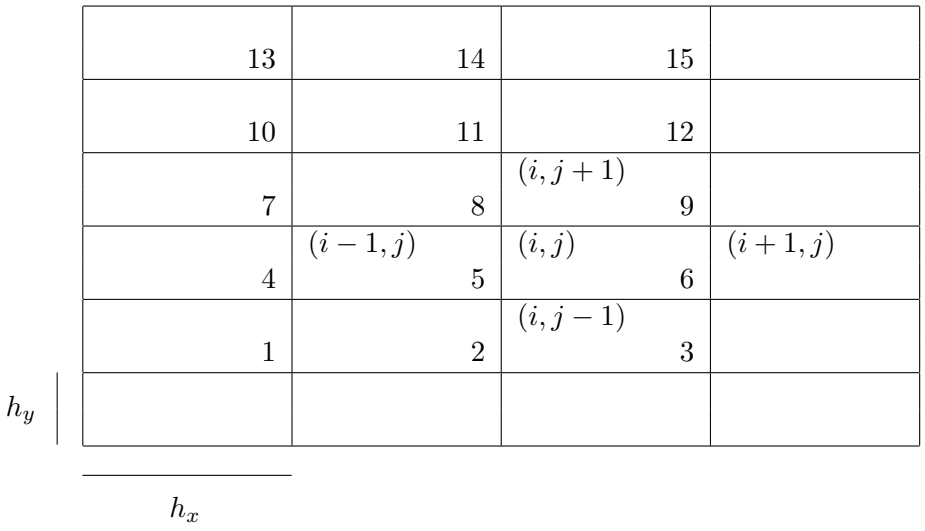
$$u(x,t) = 0 \qquad \text{na } \partial\Omega, \tag{3.73}$$

što odgovara situaciji kada je rub područja Ω držan na fiksnoj temperaturi.

Standardna metoda za dobivanje aproksimacija rješenja parcijalne diferencijalne jednadžbe, kao što je (3.72), je metoda *konačnih diferencija*. U toj metodi iz danog područja Ω izabran je skup točaka koji čini mrežu, i u svakoj točki mreže od Ω derivacija u (3.72) zamijenjuje se sa kvocijentom koji se približava pravoj derivaciji kada mreža postaje sve finija.

Na primjer, pretpostavimo da je područje Ω jedinični kvadrat $[0,1] \times [0,1]$. Uvedimo uniformnu mrežu $\{x_i, y_j : \; i = 0,1,\ldots,n_x+1, j = 0,1,\ldots,n_y+1\}$ sa koracima $h_x = 1/(n_x+1)$ u x smjeru, i $h_y = 1/(n_y+1)$ u smjeru y, kao što je pokazano u Slici 3.2 za $n_x = 3$, $n_y = 5$. Standardna aproksimacija parcijalne derivacije u smjeru x iz (3.72), pomoću centralnih diferencija je

$$\left(\frac{\partial}{\partial x} a \frac{\partial u}{\partial x}\right)(x_i, y_j) \approx \frac{a_{i+1/2,j}(u_{i+1,j} - u_{i,j}) - a_{i-1/2,j}(u_{i,j} - u_{i-1,j})}{h_x^2},$$



Slika 3.2: Diskretizacija konačnim diferencijama, prirodni poredak.

gdje je $a_{i\pm1/2,j} = a(x_i \pm h_x/2, y_j)$, a $u_{i,j}$ predstavlja aproksimaciju od $u(x_i, y_j)$. Analogan izraz može se ostvariti za parcijalnu derivaciju u y smjeru:

$$\left(\frac{\partial}{\partial y} a \frac{\partial u}{\partial y}\right)(x_i, y_j) \approx \frac{a_{i,j+1/2}(u_{i,j+1} - u_{i,j}) - a_{i,j-1/2}(u_{i,j} - u_{i,j-1})}{h_y^2},$$

gdje je $a_{i,j\pm1/2} = a(x_i, y_j \pm h_y/2)$. Ukoliko $a(x, y)$ ima diskontinuitet u nekoj liniji mreže, tada se uzima aritmetička sredina lijevog i desnog limesa u traženoj točki, obzirom na traženi smjer.

Za aproksimiranje derivacije po vremenu, često se koriste centralne diferencije ili diferencije u nazad. Na primjer, ako je poznata aproksimacija rješenja $u_{i,j}^l$ u vremenu t_l, i ako koristimo diferencije u nazad po vremenu, tada da bi dobili aproksimaciju rješenja $u_{i,j}^l$ u vremenu $t_{l+1} = t_l + \Delta t$, moramo riješiti sljedeći sustav linearnih jednadžbi:

$$\frac{u_{i,j}^{l+1} - u_{i,j}^l}{\Delta t} - \left(\frac{a_{i+1/2,j}\left(u_{i+1,j}^{l+1} - u_{i,j}^{l+1}\right) - a_{i-1/2,j}\left(u_{i,j}^{l+1} - u_{i-1,j}^{l+1}\right)}{h_x^2} + \right.$$

$$\left. + \frac{a_{i,j+1/2}\left(u_{i,j+1}^{l+1} - u_{i,j}^{l+1}\right) - a_{i,j-1/2}\left(u_{i,j}^{l+1} - u_{i,j-1}^{l+1}\right)}{h_y^2} \right) = f_{i,j}^{l+1},$$

$$i = 1, \ldots, n_x,\ j = 1, \ldots, n_y. \tag{3.74}$$

Da bismo napisali jednadžbu (3.74) u matričnom obliku, moramo izabrati neki poredak jednadžbi i nepoznanica. Uobičajeni izbor je tzv. *prirodni poredak* u kojem su točke mreže numerirane s lijeva na desno, i od dole prema gore, preko pridruživanja $(i, j) \longmapsto (j-1)n_x + i$, kao što se vidi u Slici 3.2. Sa ovakvim poretkom jednadžbe (3.74) mogu se napisati u obliku

$$Au = f, \tag{3.75}$$

gdje je A blok tridijagonalna matrica sa n_y dijagonalnih blokova, od kojih je svaki dimenzije $n_x \times n_x$, u je $n_x n_y$-dimenzionalni vektor vrijednosti funkcije u, kod kojeg je

$u_{i,j}$ smješten na poziciji $(j-1)n_x + i$, a f je $n_x n_y$-dimenzionalan vektor desne strane sustava kod kojeg je $f_{i,j}$ smješten na poziciji $(j-1)n_x + i$. Definirajmo

$$d_{i,j} = \frac{1}{\Delta t} + \frac{a_{i+1/2,j} + a_{i-1/2,j}}{h_x^2} + \frac{a_{i,j+1/2} + a_{i,j-1/2}}{h_y^2}, \tag{3.76}$$

$$b_{i+1/2,j} = \frac{-a_{i+1/2,j}}{h_x^2}, \qquad c_{i,j+1/2} = \frac{-a_{i,j+1/2}}{h_y^2}. \tag{3.77}$$

Matrica sustava se tada može napisati u obliku

$$A = \begin{bmatrix} S_1 & T_{3/2} & & \\ T_{3/2} & \ddots & \ddots & \\ & \ddots & \ddots & T_{n_y-1/2} \\ & & T_{n_y-1/2} & S_{n_y} \end{bmatrix}, \tag{3.78}$$

gdje su

$$S_j = \begin{bmatrix} d_{1,j} & b_{3/2,j} & & \\ b_{3/2,j} & \ddots & \ddots & \\ & \ddots & \ddots & b_{n_x-1/2,j} \\ & & b_{n_x-1/2,j} & d_{n_x,j} \end{bmatrix}, \tag{3.79}$$

$$T_{j+1/2} = \begin{bmatrix} c_{1,j+1/2} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & c_{n_x,j+1/2} \end{bmatrix}. \tag{3.80}$$

Vektor desne strane je oblika

$$f = \begin{bmatrix} f_{1,1} \\ \vdots \\ f_{n_x,1} \\ f_{1,2} \\ \vdots \\ f_{n_x,n_y-1} \\ f_{1,n_y} \\ \vdots \\ f_{n_x,n_y} \end{bmatrix}, \tag{3.81}$$

gdje je

$$f_{i,j} = f_{i,j}^{l+1} + \frac{u_{i,j}^l}{\Delta t}. \tag{3.82}$$

Teorem 3.6.1 ([12]). *Pretpostavimo da je $a(x,y) \geq \alpha > 0$ na $\langle 0,1\rangle \times \langle 0,1\rangle$. Tada je matrica A definirana sa (3.76–3.80) simetrična i pozitivno definitna.*

Dokaz: Simetričnost je očita. Za dokaz pozitivne definitnosti, razlikujemo dva slučaja. Prvi slučaj je situacija opisana prije teorema, odnosno situacija u kojoj je jednadžba (3.72) nestacionarna. Tada je matrica $A = (A_{ij})$ strogo dijagonalno dominantna, sa

pozitivnim dijagonalnim elementima, pa prema Gerschgorinovom teoremu postoji $i = 1, \ldots, n$, takav da je

$$A_{ii} - \lambda \leq |\lambda - A_{ii}| \leq \sum_{j \neq i} |A_{ij}| < |A_{ii}| = A_{ii},$$

za proizvoljnu svojstvenu vrijednost λ od A. Prema tome vrijedi da je $\lambda > 0$ za svaku svojstvenu vrijednost λ od A, pa je matrica A pozitivno definitna.

Drugi slučaj je kada je jednadžba (3.72) stacionarna, pa se gubi parcijalna derivacija po vremenu. U tom slučaju, na dijagonali matrice A nestaje sumand $1/\Delta t$, a na desnoj strani $u^l_{i,j}/\Delta t$. Tako dobivena matrica je tada slabo dijagonalno dominantna, i to apsolutna vrijednost dijagonalnog elementa je strogo veća od sume apsolutnih vrijednosti preostalih elemenata u retku samo ako je odgovarajući čvor mreže, reprezentiran tim retkom, susjedan rubu. Za sve ostale retke vrijedi jednakost. Primjenom Gerschgorinovog teorema sada dobivamo da je $\lambda \geq 0$ za svaku svojstvenu vrijednost λ od A. Pretpostavimo da postoji vektor v različit od nule, takav da je $Av = 0$, i pretpostavimo da komponenta od v sa najvećom apsolutnom vrijednosti odgovara čvoru mreže (i, j). Možemo izabrati predznak od v tako da je ta komponenta pozitivna. Prema definiciji od A i pretpostavci da je $a(x, y) > 0$, slijedi da se, raspisivanjem jednakosti $Av = 0$ po komponentama, $v_{i,j}$ može napisati kao težinska suma okolnih vrijednosti od v:

$$v_{i,j} = w_{i-1,j} v_{i-1,j} + w_{i+1,j} v_{i+1,j} + w_{i,j-1} v_{i,j-1} + w_{i,j+1} v_{i,j+1},$$

$$w_{i\pm1,j} = \frac{1}{d_{i,j}} \frac{a_{i\pm1/2,j}}{h_x^2}, \qquad w_{i,j\pm1} = \frac{1}{d_{i,j}} \frac{a_{i,j\pm1/2}}{h_y^2},$$

pri čemu su izrazi koji odgovarju rubnim čvorovima jednaki nuli. Težine $w_{i\pm1,j}$ i $w_{i,j\pm1}$ su pozitivne i suma im je jednaka 1. Odatle slijedi, da ako su svi susjedi od točke mreže (i, j) unutarnje točke, tada sve vrijednosti od v na tim točkama moraju imati istu maksimalnu vrijednost, budući da niti jedna ne može imati vrijednost veću od $v_{i,j}$. Ako ponavljamo ovakav princip zaključivanja na susjedne točke mreže, kad tad naći ćemo točku mreže sa istom maksimalnom vrijednošću za v, koja ima bar jednog susjeda na rubu. Međutim, vrijednost od v u toj točki je tada težinska suma vrijednosti susjednih unutarnjih čvorova, ali sa sumom težina manjom od 1. Tad slijedi da vrijednost od v u jednom od ovih ostalih unutarnjih čvorova mora biti veća od $v_{i,j}$, ako je $v_{i,j} > 0$, što je kontradikcija. Zbog toga, jedini vektor v za koji je $Av = 0$, je nul-vektor, pa su sve svojstvene vrijednosti od A pozitivne, i s time je A pozitivno definitna matrica. □

Drugi poretci jednadžbi i nepoznanica su također mogući, kao na primjer ako se čvorovi u Slici 3.2 pobojaju kao na šahovskoj tabli, gdje su crveni čvorovi grupirani oko crnih, i obratno, te ako crvene čvorove poredamo najprije, a zatim crne, tada matrica A poprima oblik

$$A = \begin{bmatrix} D_1 & B \\ B^T & D_2 \end{bmatrix},$$

gdje su D_1 i D_2 dijagonalne matrice.

Matrica oblika (3.76–3.80) se ponekad naziva i *aproksimacijom s 5 točaka*, budući da se druge derivacije u točki mreže (i, j) aproksimiraju sa izrazima koji sadrže vrijednosti funkcije u toj točci i njezinih četiri susjeda.

3.6.1 Poissonova jednadžba

U specijalnom slučaju kada je koeficijent difuzije $a(x, y)$ konstantan, recimo da je $a(x, y) = 1$ na Ω, i kada je jednadžba (3.72) stacionarna, tada se takva difuzijska jednadžba naziva *Poissonovom jednadžbom*. Matrica sustava A u tom slučaju ima poseban oblik:

$$A = \begin{bmatrix} S & T & & \\ T & \ddots & \ddots & \\ & \ddots & \ddots & T \\ & & T & S \end{bmatrix}, \tag{3.83}$$

gdje je $T = (-1/h_y^2)I$ i

$$S = \begin{bmatrix} d & b & & \\ b & \ddots & \ddots & \\ & \ddots & \ddots & b \\ & & b & d \end{bmatrix}, \qquad d = \frac{2}{h_x^2} + \frac{2}{h_y^2}, \quad b = \frac{-1}{h_x^2}. \tag{3.84}$$

Ovaj oblik je poznat pod imenom *blok-TST* matrica, gdje je "TST" kratica za Toeplitzovu (konstantna duž svih dijagonala), simetričnu i tridijagonalnu matricu. Ovakva matrica je blok-TST matrica jer su blokovi duž bilo koje blok dijagonale jednaki, simetrična je i blok-tridijagonalna je. Što više, svaki od njenih blokova je TST matrica. Svojstvene vrijednosti i svojstveni vektori takve matrice su eksplicitno poznati.

Lema 3.6.2 ([12]). *Neka je G $m \times m$ TST matrica sa dijagonalnim elementima α i vandijagonalnim elementima β. Tada su svojstvene vrijednosti od G dane sa*

$$\lambda_k = \alpha + 2\beta \cos\left(\frac{k\pi}{m+1}\right), \quad k = 1, \ldots, m, \tag{3.85}$$

a odgovarajući ortonormirani svojstveni vektori dani su, po komponentama, sa

$$q_l^{(k)} = \sqrt{\frac{2}{m+1}} \sin\left(\frac{lk\pi}{m+1}\right), \quad l, k = 1, \ldots, m. \tag{3.86}$$

Dokaz: Pretpostavimo da je $\beta \neq 0$, jer bi inače G bio multipl od identitete, čime je tvrdnja leme trivijalna. Pretpostavimo da je λ svojstvena vrijednost od G sa odgovarajućim svojstvenim vektorom q. Ako definiramo $q_0 = q_{m+1} = 0$, jednakost $Aq = \lambda q$ možemo napisati u obliku

$$\beta q_{l-1} + (\alpha - \lambda) q_l + \beta q_{l+1} = 0, \quad l = 1, \ldots, m. \tag{3.87}$$

To je jednadžba linearnih diferencija, i može se riješiti na sličan način kao odgovarajuća linearna diferencijalna jednadžba. Preciznije, razmatramo karakteristični polinom

$$\chi(z) = \beta + (\alpha - \lambda) z + \beta z^2.$$

Ako korijene ovog polinoma označimo sa z_1 i z_2, tada općenito rješenje jednadžbe diferencija (3.87) ima oblik

$$q_l = c_1 z_1^l + c_2 z_2^l, \quad \text{za konstante } c_1, c_2,$$

pri čemu su konstante određene iz rubnih uvijeta $q_0 = q_{m+1} = 0$.

Korijeni od $\chi(z)$ su

$$z_{1,2} = \frac{\lambda - \alpha \pm \sqrt{(\lambda-\alpha)^2 - 4\beta^2}}{2\beta}, \tag{3.88}$$

a iz uvjeta $q_0 = 0$ slijedi da je $c_1 + c_2 = 0$. S druge strane, iz uvjeta $q_{m+1} = 0$ slijedi:

$$c_1 z_1^{m+1} + c_2 z_2^{m+1} = 0,$$

odnosno, zbog prethodne napomene je $c_2 = -c_1$, a kako nas interesira netrivijalno rješenje slijedi $c_1 \neq 0$, imamo

$$z_1^{m+1} = z_2^{m+1}.$$

Postoji $m+1$ rješenja ove jednadžbe, i to

$$z_2 = z_1 \exp\left(\frac{2\pi k\iota}{m+1}\right), \quad k = 0, 1, \ldots, m, \ \iota = \sqrt{-1}, \tag{3.89}$$

međutim, slučaj za $k = 0$ se može odbaciti, jer je tada $z_2 = z_1$, a odatle je

$$q_l = c_1 z_1^l + c_2 z_2^l = c_1 z_1^l - c_1 z_1^l = 0.$$

Množeći (3.89) sa $\exp(-\pi k\iota/(m+1))$ i uvrštavajući vrijednosti za z_1 i z_2 iz (3.88) dobivamo

$$\left(\lambda - \alpha + \sqrt{(\lambda-\alpha)^2 - 4\beta^2}\right) \exp\left(\frac{-\pi k\iota}{m+1}\right) =$$

$$= \left(\lambda - \alpha - \sqrt{(\lambda-\alpha)^2 - 4\beta^2}\right) \exp\left(\frac{\pi k\iota}{m+1}\right).$$

Nakon potrebnih skračivanja, i dijeljenja jednakosti sa 2 imamo

$$\sqrt{(\lambda-\alpha)^2 - 4\beta^2} \cos\left(\frac{k\pi}{m+1}\right) = (\lambda - \alpha)\iota \sin\left(\frac{k\pi}{m+1}\right),$$

a nakon kvadriranja obaju strana jednakosti i rješavanja kvadratne jednadžbe po λ

$$\lambda^2 - 2\alpha\lambda + \alpha^2 - 4\beta^2 \cos^2\left(\frac{\pi k}{m+1}\right) = 0,$$

dobivamo rješenja

$$\lambda_{1,2} = \alpha \pm 2\beta \cos\left(\frac{\pi k}{m+1}\right).$$

Ako uzmemo rješenje sa znakom plus dobivamo (3.85), dok rješenje sa znakom minus samo ponavlja te iste vrijednosti, pa se stoga može odbaciti.

Uvrštavajući (3.85) u (3.88) dobivamo

$$z_{1,2} = \cos\left(\frac{k\pi}{m+1}\right) \pm \iota \sin\left(\frac{k\pi}{m+1}\right),$$

pa je zbog toga

$$q_l^{(k)} = c_1(z_1^l - z_2^l) = 2c_1\iota \sin\left(\frac{\pi kl}{m+1}\right), \quad k, l = 1, \ldots, m.$$

Ako uzmemo da je $c_1 = -(\iota/2)\sqrt{2/(m+1)}$, kao kod (3.86), onda se lako provjeri da svaki od vektora $q^{(k)}$ ima normu jedan, jer je

$$\begin{aligned}\sum_{l=1}^{m}\left(q_l^{(k)}\right)^2 &= \frac{2}{m+1}\sum_{l=1}^{m}\sin^2\left(\frac{lk\pi}{m+1}\right) = \frac{2}{m+1}\cdot\frac{1}{2}\sum_{l=1}^{m}\left[1-\cos\left(\frac{2lk\pi}{m+1}\right)\right] = \\ &= \frac{2}{m+1}\left[\frac{m}{2}-\frac{1}{2}\sum_{l=1}^{m}\cos\left(\frac{2lk\pi}{m+1}\right)\right]\end{aligned}$$

odakle se, prelaskom sa kosinusa na realni dio kompleksnog broja, dobiva da je suma kosinusa u zadnjoj jednakosti jednaka -1, čime smo dobili traženi rezultat. Svojstveni vektori su ortogonalni, budući da je matrica simetrična. □

Korolar 3.6.3 ([12]). *Sve $m \times m$ TST matrice međusobno komutiraju.*

Dokaz: Prema (3.86) sve takve matrice imaju iste ortonormirane svojstvene vektore. Ako su $G_1 = Q\Lambda_1 Q^T$ i $G_2 = Q\Lambda_2 Q^T$, tada je

$$G_1G_2 = Q\Lambda_1\Lambda_2 Q^T = Q\Lambda_2\Lambda_1 Q^T = G_2G_1.$$

□

Teorem 3.6.4 ([12]). *Svojstvene vrijednosti matrice A, definirane sa (3.83–3.84), su*

$$\lambda_{j,k} = \frac{4}{h_x^2}\sin^2\left(\frac{j\pi}{2(n_x+1)}\right) + \frac{4}{h_y^2}\sin^2\left(\frac{k\pi}{2(n_y+1)}\right),$$
$$j = 1,\ldots,n_x, \quad k = 1,\ldots,n_y, \tag{3.90}$$

a odgovarajući svojstveni vektori su

$$u_{m,l}^{(j,k)} = \frac{2}{\sqrt{(n_x+1)(n_y+1)}}\sin\left(\frac{mj\pi}{n_x+1}\right)\sin\left(\frac{lk\pi}{n_y+1}\right),$$
$$m,j = 1,\ldots,n_x, \quad l,k = 1,\ldots,n_y, \tag{3.91}$$

gdje $u_{m,l}^{(j,k)}$ označava komponentu koja odgovara točki mreže (m,l) svojstvenog vektora, pridruženog svojstvenoj vrijednosti $\lambda_{j,k}$.

Dokaz: Neka je λ neka svojstvena vrijednost od A sa odgovarajućim svojstvenim vektorom u, koji se može particionirati u oblik

$$u = \begin{bmatrix} u_1 \\ \vdots \\ u_{n_y}\end{bmatrix}, \quad u_l = \begin{bmatrix} u_{1,l} \\ \vdots \\ u_{n_x,l}\end{bmatrix}, \quad l = 1,\ldots,n_y.$$

Jednakost $Au = \lambda u$ može se tada napisati u obliku

$$Tu_{l-1} + (S-\lambda I)u_l + Tu_{l+1} = 0, \quad l = 1,\ldots,n_y, \tag{3.92}$$

gdje smo postavili da je $u_0 = u_{n_y+1} = 0$. Prema Lemi 3.6.2, možemo napisati da je $S = Q\Lambda_S Q^T$ i $T = Q\Lambda_T Q^T$, gdje su Λ_S i Λ_T dijagonalne matrice, sa j-tim dijagonalnim elementima jednakim

$$\lambda_{S,j} = \frac{2}{h_x^2} + \frac{2}{h_y^2} - \frac{2}{h_x^2}\cos\left(\frac{j\pi}{n_x+1}\right), \qquad \lambda_{T,j} = \frac{-1}{h_y^2}.$$

m-ti element j-tog stupca od Q jednak je

$$q_m^{(j)} = \sqrt{\frac{2}{n_x+1}} \sin\left(\frac{mj\pi}{n_x+1}\right), \quad m,j = 1,\ldots,n_x.$$

Pomnožimo (3.92) sa Q^T slijeva kako bi dobili

$$\Lambda_T y_{l-1} + (\Lambda_S - \lambda I) y_l + \Lambda_T y_{l+1} = 0, \quad y_l = Q^T u_l, \quad l = 1,\ldots,n_y.$$

Budući da su ovdje sve matrice dijagonalne, jednakosti duž *vertikalnih* linija u mreži imaju oblik

$$\lambda_{T,j} y_{j,l+1} + \lambda_{S,j} y_{j,l} + \lambda_{T,j} y_{j,l-1} = \lambda y_{j,l}, \quad j = 1,\ldots,n_x. \tag{3.93}$$

Ako je, za fiksnu vrijednost od j, vektor $[y_{j,1} \ \ldots \ y_{j,n_y}]^T$ svojstveni vektor TST matrice

$$\begin{bmatrix} \lambda_{S,j} & \lambda_{T,j} & & \\ \lambda_{T,j} & \ddots & \ddots & \\ & \ddots & \ddots & \lambda_{T,j} \\ & & \lambda_{T,j} & \lambda_{S,j} \end{bmatrix},$$

sa odgovarajućom svojstvenom vrijednošću λ, i ako su sve ostale komponenete vektora y jednake 0, tada će jednakost (3.93) biti zadovoljena. Ponovo prema Lemi 3.6.2, svojstvene vrijednosti ove matrice jednake su

$$\begin{aligned} \lambda_{j,k} &= \lambda_{S,j} + 2\lambda_{T,j}\cos\left(\frac{k\pi}{n_y+1}\right) = \\ &= \frac{2}{h_x^2} + \frac{2}{h_y^2} - \frac{2}{h_x^2}\cos\left(\frac{j\pi}{n_x+1}\right) - \frac{2}{h_y^2}\cos\left(\frac{k\pi}{n_y+1}\right) = \\ &= \frac{4}{h_x^2}\sin^2\left(\frac{j\pi}{2(n_x+1)}\right) + \frac{4}{h_y^2}\sin^2\left(\frac{k\pi}{2(n_y+1)}\right), \end{aligned}$$

za $k = 1,\ldots,n_y$. Pripadajući svojstveni vektori su

$$y_{j,l}^{(j,k)} = \sqrt{\frac{2}{n_y+1}} \sin\left(\frac{lk\pi}{n_y+1}\right).$$

Budući da je l-ti blok od $u^{(j,k)}$ jednak Q puta l-ti blok od y, i budući da je samo j-ta komponenta l-tog bloka od y različita od nule, imamo

$$u_{m,l} = q_m^{(j)} y_{j,l}^{(j,k)} = \frac{2}{\sqrt{(n_x+1)(n_y+1)}} \sin\left(\frac{mj\pi}{n_x+1}\right)\sin\left(\frac{lk\pi}{n_y+1}\right).$$

Dobivanjem svojstvenih vrijednosti $\lambda_{j,k}$ i odgovarajućih svojstvenih vektora $u^{(j,k)}$ za svako $j = 1,\ldots,n_x$, sada imamo sve $n_x n_y$ svojstvene parove od A. □

Kao poslijedica ovog teorema, vidimo da je matrica dobivena aproksimacijom s 5 točaka Poissonove jednadžbe M matrica, prema *(iv)* tvrdnji Teorema 3.3.12.

Korolar 3.6.5 ([12]). *Pretpostavimo da je $h_x = h_y = h$. Tada se najmanja i najveća svojstvena vrijednost od A u (3.83–3.84) ponašaju kao*

$$2\pi^2 + \mathcal{O}(h^2) \qquad i \qquad 8h^{-2} + \mathcal{O}(1) \tag{3.94}$$

kada $h \to 0$, tako da je uvjetovanost matrice A jednaka

$$\frac{4}{\pi^2} h^{-2} + \mathcal{O}(1).$$

Dokaz: Najmanja svojstvena vrijednost od A je ona sa $j = k = 1$, a najveća je ona sa $j = k = n_x = n_y$ u (3.90):

$$\lambda_{min} = 8h^{-2} \sin^2\left(\frac{\pi h}{2}\right), \qquad \lambda_{max} = 8h^{-2} \sin^2\left(\frac{\pi}{2} - \frac{\pi h}{2}\right).$$

Razvojem funkcija $\sin(x)$ i $\sin(\pi/2 - x) = \cos(x)$ u Taylorov red dobivamo traženi rezultat (3.94), a dijeljenjem λ_{max} sa λ_{min} dobivamo ocjenu za broj uvjetovanosti. □

3.6.2 Prekondicioniranje sustava Poissonove jednadžbe

Promotrit ćemo neke konkretne načine prekondicioniranja linearnog sustava dobivenog iz Poissonove jednadžbe. Najprije napomenimo da je ova matrica specifičnog oblika: blok TST matrica, da je za $a > 0$ pozitivno definitna, i da ako numeriramo čvorove mreže na crveno–crni način, kao na šahovskoj tabli, tada ćemo dobiti matricu oblika (3.33). Može se pokazati da za bilo koju numeraciju čvorova mreže matrica zadovoljava svojstvo (3.35). Svojstvene vrijednosti ove matrice su eksplicitno zadane iz tvrdnji Teorema 3.6.4.

Jacobi, Gauss–Seidel, SOR

Sada ćemo pretpostaviti da je $h_x = h_y = h$, odnosno $n_x = n_y = m$ i $n = m^2$. Jacobijeva matrica je tada oblika

$$G_J = I - \frac{h^2}{4} A,$$

tako da su joj svojstvene vrijednosti oblika

$$\lambda_{i,k}(G_J) = 1 - \sin^2\left(\frac{i\pi}{2(m+1)}\right) - \sin^2\left(\frac{k\pi}{2(m+1)}\right), \quad i,k = 1,\ldots,m.$$

Prema dokazu Korolara 3.6.5, imamo

$$\begin{aligned} \rho(G_J) &= \max_{i,k} \left| 1 - \sin^2\left(\frac{i\pi}{2(m+1)}\right) - \sin^2\left(\frac{k\pi}{2(m+1)}\right) \right| = \\ &= 1 - \frac{\pi^2}{2} h^2 + \mathcal{O}(h^4). \end{aligned} \tag{3.95}$$

Ako znamo vrijednost od $\rho(G_J)$, Teorem 3.2.21 će nam dati optimalnu vrijednost za ω, kao i stopu konvergencije SOR metode za proizvoljnu vrijednost ω. Iz tog teorema slijedi

$$\omega_{opt} = \frac{2}{1 + \sqrt{\pi^2 h^2 + \mathcal{O}(h^4)}} = 2(1 - \pi h) + \mathcal{O}(h^2),$$

pa je zbog toga

$$\rho(G_{SOR,\omega_{opt}}) = \omega_{opt} - 1 = 1 - 2\pi h + \mathcal{O}(h^2). \tag{3.96}$$

Prema Korolaru 3.2.20, za Gauss–Seidelovu metodu imamo

$$\rho(G_{GS}) = [\rho(G_J)]^2 = 1 - \pi^2 h^2 + \mathcal{O}(h^4). \tag{3.97}$$

Uspoređujući (3.95–3.97) uz ignoriranje potencije od h višeg reda, možemo zaključiti da SOR metoda sa optimalnim parametrom ω daleko najbrže konvergira, dok je Jacobijeva metoda najslabija.

Nekompletne faktorizacije

Povjesno gledano, matrice prekondicioniranja temeljene na nekompletnim faktorizacijama razvijene su najprije za matrice sa specifičnom pravilnom strukturom, kao što je matrica dobivena diskretizacijom Poissonove jednadžbe. Kao osnovna značajka ovako strukturiranih matrica je posjedovanje malog broja netrivijalnih dijagonala. Ako gledamo izgled matrice u ovisnosti o čvorovima mreže, tada vidimo da su u i-tom retku jedini netrivijalni vandijagonalni elementi na pozicijama $i+1$ i $i-1$ za horizontalne susjede čvora mreže reprezentiranim ovim retkom, i na pozivijama $i+m$ i $i-m$ za vertikalne susjede. Ako i-ti redak reprezentira čvor mreže (k,l), tada se, kao što smo već prije vidjeli, radi o susjednim čvorovima $(i+1,j)$, $(i-1,j)$, $(i,j+1)$ i $(i,j-1)$. Ako izvršimo nekompletnu faktorizaciju Choleskog nad tom matricom, tada će faktor L imati istu strukturu netrivijalni elemenata kao donji trokut polazne matrice. Međutim produkt LL^T ima još dodatne netrivijalne $m-1$-e sporedne dijagonale u gornjem i donjem trokutu, što se lako provjeri uvrštavanjem stvarnih vrijednosti. Dakle, u i-tom retku pojavit će se još dodatni netrivijalni elementi na pozicijama $i+m-1$ i $i-m+1$, koje odgovaraju čvorovima mreže $(i-1,j+1)$ i $(i+1,j-1)$. Ovo dakako vrijedi za nekompletnu faktorizaciju tipa IC(0). Ako promatramo, nadalje, IC(1) faktorizaciju, tada se vidi da je struktura netrivijalni elemenata od L jednaka strukturi netrivijalnih elemenata donjeg trokuta LL^T produkta IC(0) faktorizacije.

IC(0) faktorizaciju ove matrice pogodno je promatrati kao nekompletnu faktorizaciju oblika LDL^T, gdje je D dijagonalna matrica, i pri čemu je $d_{ii} = l_{ii}^{-1}$. Označimo sa $\mathbf{a}$ glavnu dijagonalu od A, sa $\mathbf{b}$ prvu donju sporednu dijagonalu, i sa $\mathbf{c}$ m-tu donju sporednu dijagonalu. Nadalje, označimo sa $\mathbf{e}$ glavnu dijagonalu matrice L, sa $\mathbf{f}$ njenu prvu donju sporednu dijagonalu, i sa $\mathbf{g}$ njenu m-tu donju sporednu dijagonalu, te sa $\mathbf{d}$ glavnu dijagonalu od D. Tada imamo

$$\mathbf{f} = \mathbf{b}, \qquad \mathbf{g} = \mathbf{c},$$

$$e_i = d_i^{-1} = a_i - f_{i-1}^2 d_{i-1} - g_{i-m}^2 d_{i-m}, \quad i = 1, \ldots, n.$$

Produkt $M = LDL^T$ ima i-ti redak oblika

$$\cdots\ 0\ c_{i-m}\ r_{i-m+1}\ 0\ \cdots\ 0\ b_{i-1}\ a_i\ b_i\ 0\ \cdots\ 0\ r_i\ c_i\ 0\ \cdots,$$

gdje je

$$r_i = \frac{b_{i-1}c_{i-1}}{e_{i-1}}.$$

Promatrajmo sada matricu $A = A_h = (a_{ij})$ definiranu sa (3.76–3.80), koju dobivamo iz aproksimacije s 5 točaka stacionarne difuzijske jednadžbe, ili općenitije, promatrajmo

bilo koju matricu A_h, dobivenu aproksimiranjem pomoću konačnih diferencija eliptične diferencijalne jednandžbe drugog stupnja, sa korakom mreže h,

$$\mathcal{L}u = -\frac{\partial}{\partial x}\left(\alpha_1 \frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(\alpha_2 \frac{\partial u}{\partial y}\right) = f \tag{3.98}$$

definirane na području $\Omega \subset \mathbb{R}^2$, sa odgovarajućim rubnim uvjetima na $\partial\Omega$. Pretpostavimo da je $\alpha_i = \alpha_i(x,y) \geq \alpha > 0$, $i = 1, 2$. U tom slučaju je prema Teoremu 3.6.1 matrica A simetrična i pozitivno definitna.

Ovakva matrica obično ima nekoliko posebnih svojstava. Prvo je lokalno svojstvo, tj. ako je $a_{ij} \neq 0$, tada je udaljenost od čvora i do čvora j zadane mreže ograničena sa produktom konstante (neovisne o h) i parametra h. To možemo označiti sa $\mathcal{O}(h)$, a ovo svojstvo je vidljivo i iz definicije matrice (3.78). Drugo, budući da svaki produkt Av aproksimira $\mathcal{L}v(x,y)$, gdje je $v(x,y)$ funkcija reprezentirana vektorom v, i budući da djelovanje operatora $\mathcal{L}$ na konstantnu funkciju v daje 0, slijedi da je suma elemenata iz svakog retka od A jednaka 0, osim možda za retke koji predstavljaju točke mreže, koje su susjedne rubu od Ω. Pretpostavimo da je A skalirana (pomnožena sa h^2), tako da su netrivijalni elementi od A reda veličine $\mathcal{O}(1)$. Dimenzija matrice A je $n = m^2 = \mathcal{O}(h^{-2})$. Tipičan primjer je aproksimacija s 5 točaka Poissonove jednadžbe (pomnožene sa h^2), za koju je

$$A = \begin{bmatrix} T & -I & & \\ -I & T & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & T \end{bmatrix}, \quad T = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & \end{bmatrix}. \tag{3.99}$$

Ako je $A = M - R$ rastav od A, i ako je $M = LL^T$, tada su najveća i najmanja svojstvena vrijednost $\lambda_{max}(M^{-1}A)$ i $\lambda_{min}(M^{-1}A)$ prekondicionirane matrice $M^{-1}A$ jednake odgovarajućim svojstvenim vrijednostima od $L^{-1}AL^{-T}$, zbog sličnosti tih dviju matrica. Vrijedi

$$\begin{aligned} \lambda_{max}(M^{-1}A) &= \max_{w\neq 0} \frac{\langle L^{-1}AL^{-T}w, w\rangle}{\langle w, w\rangle} = \max_{w\neq 0} \frac{\langle AL^{-T}w, L^{-T}w\rangle}{\langle w, w\rangle} = \\ &= \max_{v=L^{-T}w\neq 0} \frac{\langle Av, v\rangle}{\langle L^Tv, L^Tv\rangle} = \max_{v\neq 0} \frac{\langle Av, v\rangle}{\langle Mv, v\rangle}, \end{aligned}$$

pri čemu je $v = L^{-T}w$. Analogno je

$$\lambda_{min}(M^{-1}A) = \min_{v\neq 0} \frac{\langle Av, v\rangle}{\langle Mv, v\rangle}.$$

Nadalje, je

$$\frac{\langle Av, v\rangle}{\langle Mv, v\rangle} = \frac{1}{1 + \langle Rv, v\rangle / \langle Av, v\rangle}. \tag{3.100}$$

Pretpostavimo da vektor v predstavlja funkciju $v(x,y) \in C_0^1(\Omega)$, gdje je $C_0^1(\Omega)$ prostor neprekidno diferencijabilnih funkcija sa 0 na rubu od Ω. Sumirajući po elementima i koristeći simetričnost matrice A, imamo

$$\langle Av, v\rangle = \sum_i \sum_j a_{ij} v_i v_j =$$

$$\begin{aligned} &= \sum_i a_{ii} v_i^2 + 2 \sum_i \sum_{j>i} a_{ij} v_i v_j = \\ &= -\sum_i \sum_{j>i} a_{ij} (v_i - v_j)^2 + \sum_i \sum_j a_{ij} v_i^2. \end{aligned} \tag{3.101}$$

Zbog svojstva da je suma elemenata u retku matrice A jednaka 0, imamo da je $\sum_j a_{ij} v_i^2 = 0$, osim ako je čvor i susjed rubu od Ω, a to se događa samo ako je udaljenost od čvora i do ruba ograničena sa $\mathcal{O}(h)$. Kako je $v(x,y) \in C_0^1(\Omega)$, slijedi da je u takvim točkama, prema Taylorovom teoremu srednje vrijednosti,

$$v(x,y) = v(x_r, y_r) + \left\langle \nabla v(x_r + \theta(x - x_r), y_r + \theta(y - y_r)), \begin{bmatrix} x - x_r \\ y - y_r \end{bmatrix} \right\rangle,$$

za točku na rubu (x_r, y_r), i neki $\theta \in \langle 0, 1 \rangle$, pri čemu su (x, y) koordinate čvora i. Kako je $v(x_r, y_r) = 0$, imamo

$$|v_i| \leq \|\nabla v(x_r + \theta(x - x_r), y_r + \theta(y - y_r))\|_2 \left\| \begin{bmatrix} x - x_r \\ y - y_r \end{bmatrix} \right\|_2 \leq \mathcal{O}(h).$$

Kako su netrivijalni elementi od A reda veličine $\mathcal{O}(1)$, druga suma u (3.101) je ograđen sa veličinom jednakom broju čvorova i koji su susjedni sa rubom $\partial\Omega$ puta $\mathcal{O}(h^2)$.

Zbog lokalnog svojstva matrice A, slijedi da za čvorove i i j takve da je $a_{ij} \neq 0$, udaljenost između čvorova i i j je $\mathcal{O}(h)$. Ako sa (x_i, y_i) i (x_j, y_j) označimo točke u $\mathbb{R}^2$ koje predstavljaju čvorove i i j, tada ponovo zbog Taylorovog teorema srednje vrijednosti imamo da je

$$|v_i - v_j| \leq \|\nabla v(x_j + \theta(x_i - x_j), y_j + \theta(y_i - y_j))\|_2 \left\| \begin{bmatrix} x_i - x_j \\ y_i - y_j \end{bmatrix} \right\|_2 \leq \mathcal{O}(h),$$

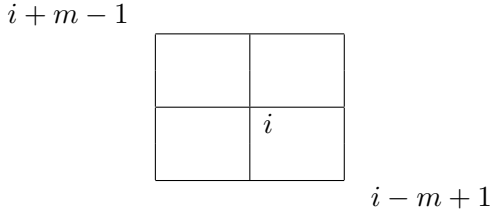
za neko $\theta \in \langle 0, 1 \rangle$. Kako u prvoj sumi u (3.101) sumiramo po svim netrivijalnim elementima gornjeg trokuta matrice A, (to su 1. i m-ta gornja sporedna dijagonala) sumanada u toj sumi ima $\mathcal{O}(n) = \mathcal{O}(h^{-2})$, i svaki je reda veličine $\mathcal{O}(h^2)$, vrijedi

$$\left| \sum_i \sum_{j>i} a_{ij} (v_i - v_j)^2 \right| \leq \mathcal{O}(1).$$

I za matricu reziduala R, također možemo pisati

$$\langle Rv, v \rangle = -\sum_i \sum_{j>i} r_{ij} (v_i - v_j)^2 + \sum_i \sum_j r_{ij} v_i^2. \tag{3.102}$$

Pretpostavimo da i matrica R ima isto svojstvo da netrivijalni elementi r_{ij} odgovaraju samo čvorovima i i j koji su udaljeni za $\mathcal{O}(h)$, i pretpostavimo sa su netrivijalni elementi od R reda veličine $\mathcal{O}(1)$ (ali možda manji od reda veličine elemenata od A). To jest slučaj kod nekompletne faktorizacije Choleskog matrice dobivene iz difuzijske jednadžbe. r_{ij} je različit od nule samo ako je $j = i + m - 1$ ili $j - m + 1$. Ti čvorovi odgovaraju čvorovima koji su od čvora i udaljeni za $\sqrt{2}h$.



Prema tome, zbog istih argumenata korištenih kod matrice A, prva suma u (3.102) je ograničena po apsolutnoj vrijednosti sa $\mathcal{O}(1)$.

Ograda drugog izraza u (3.101) ovisi o svojstvu matrice A da je suma elemenata u svakom retku jednaka 0. Medjutim, ovo svojstvo ne mora vrijediti za matricu R, što više u našem slučaju ne vrijedi, budući da je prema Korolaru 3.4.4 nekompletna dekompozicija regularni rastav, pa su svi elementi od R nenegativni. Stoga ova druga suma može biti prilično velika. Ona je ograničena sa brojem netrivijalnih elemenata od R u recima koji odgovaraju čvorovima koji nisu na rubu, što je reda veličine $\mathcal{O}(n) = \mathcal{O}(h^{-2})$ (gornja i donja $m-1$-a sporedna dijagonala), puta najveća vrijednost od r_{ij}, što je reda veličine $\mathcal{O}(1)$, puta kvadrat najveće vrijednosti funkcije $v(x,y)$ na unutrašnjim čvorovima područja Ω, što je reda veličine $\mathcal{O}(1)$. Prema tome, druga suma u (3.102) može biti reda veličine $\mathcal{O}(h^{-2})$. Dakle, imamo:

$$\langle Av, v\rangle = \mathcal{O}(1) + \mathcal{O}(h^2) = \mathcal{O}(1),$$

$$\langle Rv, v\rangle = \mathcal{O}(1) + \mathcal{O}(h^{-2}) = \mathcal{O}(h^{-2}),$$

jer znamo da je $h \ll 1$. Za vektore v koji predstavljaju C_0^1-funkciju, kvocijent $\langle Rv, v\rangle / \langle Av, v\rangle$ u (3.100) je stoga reda veličine $\mathcal{O}(h^{-2})$, i još ako je $\langle Rv, v\rangle$ pozitivan, (kako je $R \geq 0$, to će vrijediti za bilo koji vektor $v \geq 0$), tada je kvocijent $\langle Av, v\rangle / \langle Mv, v\rangle$ u (3.100) reda veličine $\mathcal{O}(h^2)$. Prema tome je $\lambda_{min}(M^{-1}A)$ najviše reda veličine $\mathcal{O}(h^2)$. S druge strane, ako uzmemo prvi jedinični vektor ξ_1, tada je $\langle A\xi_1, \xi_1\rangle / \langle M\xi_1, \xi_1\rangle = 1$, pa je $\lambda_{max}(M^{-1}A)$ barem reda $\mathcal{O}(1)$. Slijedi da je broj uvjetovanosti prekondicionirane matrice $M^{-1}A$ najmanje $\mathcal{O}(h^{-2})$, što je istog reda veličine kao i $\kappa(A)$, prema Korolaru 3.6.5, odakle se vidi da nismo postigli zadovoljavajuće poboljšanje.

Modificirana nekompletna faktorizacija Choleskog

Zbog prethodno iznesenog razmatranja, sljedeći nam je zadatak pronaći matricu prekondicioniranja $M = LL^T$, takvu da je $A = M - R$ i da je $|\langle Rv, v\rangle| = \mathcal{O}(h^{-1})$ za $v(x,y) \in C_0^1$, kako bismo mogli dobiti prekondicioniranu matricu sa brojem uvjetovanosti reda veličine $\mathcal{O}(h^{-1})$ umjesto $\mathcal{O}(h^{-2})$. Pretpostavimo da je A napisana u obliku $A = M - R$, gdje je

$$R = \hat{R} + E \tag{3.103}$$

i gdje je $\hat{R}$ negativno semidefinitna matrica ($\langle \hat{R}v, v\rangle \leq 0$, $\forall v$), $\sum_j \hat{r}_{ij} = 0$ $\forall i$, i E je pozitivno definitna dijagonalna matrica. Pretostavimo također da $\hat{R}$ ima netrivijalne elemente samo na pozicijama (i,j) koje odgovaraju čvorovima i i j koji su udaljeni za $\mathcal{O}(h)$ jedan od drugoga. Izbor matrice E ovisi o rubnim uvjetima. Za Dirichletov problem, na primjer, izabrat ćemo $E = \eta h^2 \operatorname{diag}(A)$, gdje je $\eta > 0$ parametar.

Kao i u dokazu da je apsolutna vrijednost prve sume u (3.101) reda veličine $\mathcal{O}(1)$, imamo da je zbog gornje pretpostavke lokalnog svojstva matrice $\hat{R}$,

$$\left| \sum_i \sum_{j>i} r_{ij}(v_i - v_j)^2 \right| = \left| \sum_i \sum_{j>i} \hat{r}_{ij}(v_i - v_j)^2 \right| \leq \mathcal{O}(1),$$

pa je prema (3.102)

$$\langle Rv, v\rangle = \sum_i \sum_j r_{ij}|v_i|^2 + \mathcal{O}(1),$$

kada je $v(x,y) \in C_0^1(\Omega)$. Kako su sume elemenata u svakom retku od $\hat{R}$ jednake nula i kako su netrivijalni elementi od E reda veličine $\mathcal{O}(h^2)$, imamo

$$\langle Rv, v\rangle = \sum_i \sum_j \hat{r}_{ij}|v_i|^2 + \sum_i e_{ii}|v_i|^2 + \mathcal{O}(1) = \mathcal{O}(h^2) + \mathcal{O}(1) = \mathcal{O}(1),$$

čime je nužan uvijet $|\langle Rv, v\rangle| \leq \mathcal{O}(h^{-1})$ sigurno zadovoljen. Teorem koji slijedi, i kojeg je dokazao Gustafsson, daje dovoljne uvjete za dobivanje prekondicionirane matrice sa brojem uvjetovanosti $\mathcal{O}(h^{-1})$.

Teorem 3.6.6 ([12]). *Neka je $A = M - R$, gdje je R oblika (3.103), $\hat{R}$ je negativno semidefinitna, zatim, suma elemenata u svakom retku od $\hat{R}$ je jednaka nuli, i $\hat{R}$ zadovoljava lokalno svojstvo, te neka je E pozitivno definitna dijagonalna matrica sa dijagonalnim elementima reda veličine $\mathcal{O}(h^2)$. Tada je dovoljan uvjet za dobivanje $\lambda_{max}(M^{-1}A)/\lambda_{min}(M^{-1}A) = \mathcal{O}(h^{-1})$:*

$$0 \leq -\langle \hat{R}v, v\rangle \leq (1+ch)^{-1}\langle Av, v\rangle \qquad \forall v, \tag{3.104}$$

gdje je $c > 0$ neovisan o h.

Dokaz: Najprije napomenimo da u ovom slučaju promatramo matricu dobivenu aproksimacijom s 5 točaka Poissonove jednadžbe pomnoženu sa h^2. Tada, prema Korolaru 3.6.5 postoje konstante c_1 i c_2, neovisne o h, takve da je

$$c_1 h^2 \leq \lambda_{min}(A) \leq \frac{\langle Av, v\rangle}{\langle v, v\rangle} \leq \lambda_{min}(A) \leq c_2,$$

za proizvoljni $v \neq 0$. Kako su E i A pozitivno definitne, i dijagonalni elementi od E su reda veličine $\mathcal{O}(h^2)$, slijedi da je

$$0 < \frac{\langle Ev, v\rangle}{\langle Av, v\rangle} \leq \frac{\hat{c}_3 h^2}{c_1 h^2} \leq c_3$$

za neku konstantu c_3. Iz (3.100) i činjenice da je E pozitivno definitna, a $\hat{R}$ negativno semidefinitna, te iz pretpostavke (3.104), slijedi da je

$$\begin{aligned}\frac{\langle Av, v\rangle}{\langle Mv, v\rangle} &\geq \frac{\langle Av, v\rangle}{\langle Mv, v\rangle - \langle \hat{R}v, v\rangle} = \frac{\langle Av, v\rangle}{\langle Av, v\rangle + \langle Ev, v\rangle} = \\ &= \frac{1}{1 + \langle Ev, v\rangle/\langle Av, v\rangle} \geq (1+c_3)^{-1} = \mathcal{O}(1),\end{aligned}$$

i s druge strane

$$\begin{aligned}\frac{\langle Av, v\rangle}{\langle Mv, v\rangle} &\leq \frac{\langle Av, v\rangle}{\langle Av, v\rangle + \langle \hat{R}v, v\rangle} = \frac{1}{1 + \langle \hat{R}v, v\rangle/\langle Av, v\rangle} \leq \\ &\leq \frac{1+ch}{ch} = \frac{1}{ch} + 1 = \mathcal{O}(h^{-1}),\end{aligned}$$

pa je prema tome

$$\lambda_{min}(M^{-1}A) = \mathcal{O}(1), \qquad \lambda_{max}(M^{-1}A) = \mathcal{O}(h^{-1}).$$

Dakle možemo zaključiti da je broj uvjetovanosti prekondicionirane matrice, $\lambda_{max}(M^{-1}A)/\lambda_{min}(M^{-1}A)$ reda veličine $\mathcal{O}(h^{-1})$. □

Kada je A matrica koja dolazi od diskretizacije jednadžbe (3.98), jednostavna modifikacija nekompletne faktorizacije Choleskog, poznata pod imenom *modificirana nekompletna faktorizacija Choleskog* (MIC), rezultira matricom prekondicioniranja M kojoj je $\lambda_{max}(M^{-1}A)/\lambda_{min}(M^{-1}A) = \mathcal{O}(h^{-1})$. Neka je L donje trokutasta matrica, sa nulama na pozicijama koje odgovaraju indeksima iz nekog skupa $\mathcal{P}$. Izaberimo netrivijalne elemente od L tako da se $M = LL^T$ poklapa sa A na pozicijama koje nisu u $\mathcal{P}$, osim glavne dijagonale. Definirajmo $E = \eta h^2 \,\mathrm{diag}(A)$, i postavimo elemente od $\hat{R} = LL^T - (A + E)$ tako da u svakom retku daju sumu jednaku nuli. Može se pokazati, slično kao i za nemodificiranu nekompletnu faktorizaciju Choleskog, da takva faktorizacija postoji za općenite M- matrice A, i da su vandijagonalni elementi od $\hat{R}$ nenegativni, dok su dijagonalni elementi negativni (minus suma elemenata u retku). I u ovom slučaju je najpopularniji izbor za skup $\mathcal{P}$ skup pozicija u kojima A ima nule, tako da L ima isti raspored nula kao i donji trokut od A.

Kada A ima raspored nula kao aproksimacija s 5 točaka (3.76–3.80), to se može postići na sljedeći način. Ponovo je pogodno modificiranu nekompletnu faktorizaciju Choleskog napisati u obliku LDL^T, gdje je D dijagonalna matrica. Označimo sa **a** glavnu dijagonalu od A, sa **b** prvu donju sporednu dijagonalu, i sa **c** m-tu donju sporednu dijagonalu od A. Nadalje, označimo sa **e** glavnu dijagonalu od L, sa **f** prvu donju sporednu dijagonalu, i sa **g** m-tu donju sporednu dijagonalu od L, te sa **d** označimo glavnu dijagonalu od D, a sa **r** $m-1$-u donju sporednu dijagonalu od R. Tada imamo

$$\mathbf{f} = \mathbf{b}, \qquad \mathbf{g} = \mathbf{c},$$

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

$$e_i = d_i^{-1} = a_i(1 + \eta h^2) - f_{i-1}^2 d_{i-1} - g_{i-m}^2 d_{i-m} - r_{i-1} - r_{i-m}, \tag{3.105}$$

$$r_i = f_i g_i d_i, \tag{3.106}$$

pri čemu su oni elementi koji ovdje nisu definirani jednaki nuli. Matrica $\hat{R}$ u (3.103) zadovoljava

$$\hat{r}_{i,i+m-1} = \hat{r}_{i+m-1,i} = r_{i-1}, \qquad \hat{r}_{i,i} = -r_{i-1} - r_{i-m},$$

a svi ostali elementi od $\hat{R}$ su jednaki nuli.

Sada ćemo pokazati rezultat koji je dobio Gustafsson za matricu A dobivenu iz aproksimacije s 5 točaka Poissonove jednadžbe, koji tvrdi da prekondicionirana matrica $L^{-1}AL^{-T}$ ima broj uvjetovanosti reda veličine $\mathcal{O}(h^{-1})$.

Lema 3.6.7 ([12]). *Neka su r_i, $i = 1, \ldots, n-m$ elementi definirani sa (3.105–3.106) za matricu A dobivenu iz aproksimacije s 5 točaka Poissonove jednadžbe. Tada je*

$$0 \le r_i \le \frac{1}{2(1 + ch)},$$

gdje je $c > 0$ neovisan o h.

Dokaz: Prvo ćemo pokazati da je

$$e_i \ge 2(1 + \sqrt{2\eta} h), \quad \forall i.$$

Za zadani model, jednakosti (3.105–3.106) mogu se, uz uvrštavanje vrijednosti od f_{i-1}, g_{i-m}, d_{i-1}, d_{i-m}, r_{i-1}, i r_{i-m}, napisati u obliku

$$e_i = a_i(1 + \eta h^2) - b_{i-1}(b_{i-1} + c_{i-1})/e_{i-1} - c_{i-m}(c_{i-m} + b_{i-m})/e_{i-m},$$

a nakon uvrštavanja $a_i = 4$, $b_i = c_i = -1$, $\forall i$, imamo

$$e_i = 4(1 + \eta h^2) - 2/e_{i-1} - 2/e_{i-m}.$$

Za $i = 1$, imamo da je

$$e_1 = 4(1 + \eta h^2) \geq 2(1 + \sqrt{2\eta}h).$$

Pretpostavimo da je $e_j \geq 2(1 + \sqrt{2\eta}h)$, za sve $j = 1, \ldots, i-1$. Tada imamo

$$\begin{aligned} e_i &\geq 4(1 + \eta h^2) - 2/(1 + \sqrt{2\eta}h) \geq \\ &\geq 4(1 + \eta h^2) - 2(1 - \sqrt{2\eta}h + 2\eta h^2) = \\ &= 2 + 2\sqrt{2\eta}h. \end{aligned}$$

Budući da je $r_i = b_i c_i / e_i$, dobivamo

$$0 \leq r_i \leq \frac{1}{2(1 + \sqrt{2\eta}h)}.$$

□

Teorem 3.6.8 ([12]). *Neka je $M = LL^T$ sa netrivijalnim elementima definiranim u (3.105–3.106), i neka je A matrica dobivena iz aproksimacije s 5 točaka Poissonove jednadžbe. Tada je $\lambda_{max}(M^{-1}A)/\lambda_{min}(M^{-1}A) = \mathcal{O}(h^{-1})$.*

Dokaz: Za ovaj model, koristeći izraz (3.101), imamo

$$\langle Av, v\rangle \geq -\sum_i [b_i(v_i - v_{i+1})^2 + c_i(v_i + v_{i+m})^2],$$

za bilo koji vektor v. Međutim, kako je u našem slučaju $b_i = c_i = -1$, $\forall i$, dobivamo

$$\langle Av, v\rangle \geq \sum_{i:b_i,c_i \neq 0} [(v_i - v_{i+1})^2 + (v_i - v_{i+m})^2]. \tag{3.107}$$

Analogan izraz za $\langle \hat{R}v, v\rangle$, budući da je suma elemenata u svakom retku jednaka nuli, daje

$$\langle \hat{R}v, v\rangle = -\sum_i \sum_{j>i} \hat{r}_{ij}(v_i - v_j)^2 = -\sum_i r_{i-1}(v_i - v_{i+m-1})^2.$$

Budući da je je $\hat{R}$ simetrična matrica, sa nenegativnim elementima van dijagonale, i sa sumom elemenata u svakom retku jednakom nuli, prema Gerschgorinovom teoremu imamo, da za svaku svojstvenu vrijednost λ od $\hat{R}$, postoji indeks i takav da je

$$\lambda - \hat{r}_{ii} \leq \sum_{j \neq i} \hat{r}_{ij} = -\hat{r}_{ii},$$

odakle je $\lambda \leq 0$, odnosno možemo zaključiti da je $\hat{R}$ negativno semidefinitna. Prema Lemi 3.6.7 slijedi da je

$$-\langle \hat{R}v, v\rangle \leq \frac{1}{2(1 + ch)} \sum_{i:r_{i-1} \neq 0} (v_i - v_{i+m-1})^2. \tag{3.108}$$

Koristeći nejednakost $\frac{1}{2}(a-b)^2 \leq (a-c)^2 + (c-b)^2$, koja vrijedi za bilo koje realne brojeve a, b, i c (slijedi iz $\frac{1}{2}[(a+b)-2c]^2 \geq 0$), nejednakost (3.108) može se napisati u obliku

$$-\langle Rv, v\rangle \leq \frac{1}{1+ch} \sum_{i:r_{i-1}\neq 0} [(v_i - v_{i-1})^2 + (v_{i-1} - v_{i+m-1})^2],$$

odnosno

$$-\langle Rv, v\rangle \leq (1+ch)^{-1} \sum_{i:r_i\neq 0} [(v_{i+1} - v_i)^2 + (v_i - v_{i+m})^2].$$

Budući da je r_i različit od nule samo kada su i b_i i c_i različiti od nule, ovu nejednakost kombiniramo sada sa (3.107) kako bi dobili

$$-\langle \hat{R}v, v\rangle \leq (1+ch)^{-1}\langle Av, v\rangle.$$

Traženi rezultat slijedi iz Teorema 3.6.6. □

Dakle, za dovoljno malu vrijednost od h, jasno je da MIC(0) daje bolju uvjetovanost prekondicionirane matrice od IC(0), u slučaju aproksimacije s 5 točaka Poissonove jednadžbe. Prema tome očekujemo da će CG metoda prekondicionirana sa MIC(0) bolje konvergirati od one prekondicionirane sa IC(0).

Glava 4

Multigrid metode

Do sada smo se bavili načinima prekondicioniranja za općenite klase matrica. Originalan problem, koji je aproksimiran problemom rješavanja linearnog sustava, nije igrao nikakvu ulogu u konstrukciji matrice prekondicioniranja. Metode, koje će biti predstavljene u ovom i sljedećem poglavlju, dizajnirane su za dobivanje prekondicioniranja sustava koji su dobiveni diskretizacijom diferencijalnih jednadžbi.

Originalno, multigrid metode razvijene su za rješavanje rubnih problema, smještenih na određenim prostornim domenama. Takvi problemi su se diskretizirali izborom skupa točaka iz domene problema, koje su tvorile mrežu (grid). Rezultirajući diskretni problem postaje problem rješavanja sistema linearnih jednadžbi, pridružen izabranim točkama mreže. Na taj način, fizička mreža se pojavljuje kao prirodan faktor u formulaciji zadanih rubnih problema. Kao što ćemo vidjeti u daljnim razmatranjima, nakon potrebne analize može se pokazati da rješavanje tog diskretnog sustava na raznim mrežama (sve grubljim i grubljim) predstavlja temelj za metodu sa dobrim svojstvima konvergencije. Ovakve metode se mogu koristiti za široku klasu problema, i nisu ograničene samo za određenu diferencijalnu jednadžbu, pa se čak koriste i za probleme koji nisu povezani ni sa kakvom fizičkom mrežom. Polazni multigrid pristup u sadašnjici se apstrahirao i proširio se na puno širi skup problema, u obliku kojeg nazivamo algebarske multigrid metode.

4.1 Osnove multigrid metoda

Multigrid metode nisu originalno bile opisane kao kombinacija neke iterativne metode i prekondicioniranja, ali one se, kao što ćemo pokazati, mogu gledati na taj način. Prve multigrid metode koristile su jednostavne iteracije

$$x_k = x_{k-1} + M^{-1}(b - Ax_{k-1}),$$

čija je greška $e_k = A^{-1}b - x_k$, dana tada sa

$$e_k = (I - M^{-1}A)e_{k-1}.$$

Da bi najbolje shvatili osnovne karakteristike multigrid metoda, najbolje da najprije razradimo jedan konkretan primjer, koji je detaljno obrađen u [3]. Radi se ponovo o difuzijskoj jednadžbi, opisanoj u odjeljku 3.6, samo što ćemo se više koncentrirati na jednodimenzijalnom slučaju. Dakle, krećemo od problema stacionarne distribucije temperature duž uniformnog štapa, koji je dan diferencijalnom jednadžbom drugog reda,

sa rubnim uvjetima

$$\begin{aligned} -u''(x) &= f(x), \qquad x \in \langle 0,1\rangle, \\ u(0) &= u(1) = 0. \end{aligned}$$

Problem ćemo ponovo rješavati aproksimativno pomoću konačnih diferencija. Domena $\langle 0,1\rangle$ dijeli se na $n+1$ podintervala (pri čemu je, zbog potreba algoritma, $n+1$ paran, ili još bolje potencija broja 2) duljine $h = 1/(n+1)$, tako da svaku točku mreže možemo označiti sa $x_i = ih$, $i = 0,\ldots,n+1$, a samu mrežu s Ω^h. Ponovo s u_i označimo aproksimaciju vrijednosti $u(x_i)$, s $f_i = f(x_i)$ za $i = 1,\ldots,n$, čime dobivamo sustav sa n linearnih jednadžbi oblika

$$\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i, \qquad i = 1,\ldots,n,$$

$$u_0 = u_{n+1} = 0.$$

U matričnom obliku, ovaj sustav ima matricu

$$A = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}, \tag{4.1}$$

koja je pozitivno definitna TST matrica, čije su nam svojstvene vrijednosti i svojstveni vektori poznati. Prema Lemi 3.6.2, svojstvene vrijednosti ove matrice su

$$\lambda_k(A) = \frac{2}{h^2}\left[1 - \cos\left(\frac{k\pi}{n+1}\right)\right] = \frac{4}{h^2}\sin^2\left(\frac{k\pi}{2(n+1)}\right), \quad k = 1,\ldots,n, \tag{4.2}$$

a j-ta komponenta odgovarajućeg svojstvenog vektora je

$$q_j^{(k)} = \sin\left(\frac{jk\pi}{n+1}\right), \qquad j,k = 1,\ldots,n. \tag{4.3}$$

Jedan od koraka multigrid metode čine iteracije neke od standardnih iterativnih metoda. Za početak koristit ćemo se jednostavnim iteracijama, uz prekondicioniranje sa jednom od klasičnih iterativnih metoda. Pokazat će se da je to sasvim dovoljno za postizanje zadovoljavajuće konvergencije. Najčešće su to JOR metoda i standardna Gauss–Seidelova metoda ili Gauss–Seidelova metoda kojoj su točke kvadratne mreže numerirane kao na šahovskoj ploči. U ovom modelu promatrat ćemo iteracije JOR metode, budući da ona ima neka interesantna svojstva. JOR metoda je detaljno obrađena u odjeljku 3.2. Ako particioniramo matricu $A = D - L - U$, pri čemu je D dijagonalna matrica, a L i U su strogo donja, odnosno, gornja trokutasta matrica, tada znamo da je JOR matrica oblika

$$G_{JOR,\omega} = (1-\omega)I + D^{-1}(L+U) = I - \omega D^{-1}A,$$

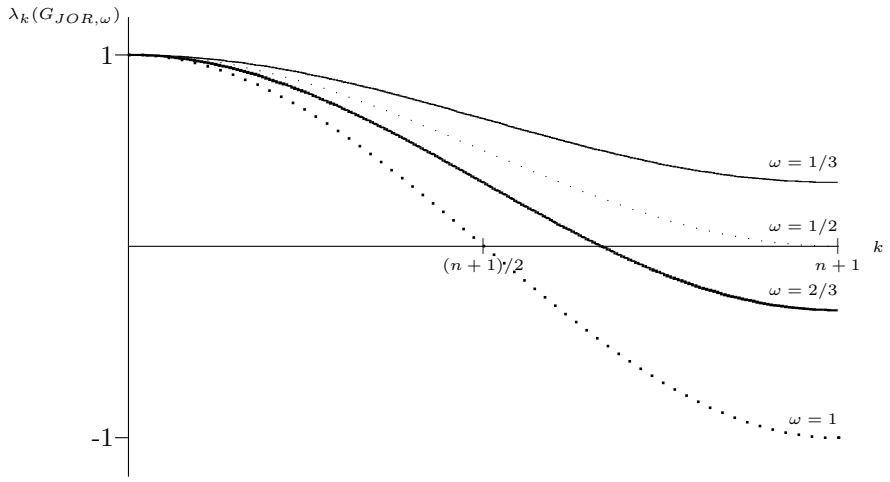
a u našem slučaju je

$$G_{JOR,\omega} = I - \frac{\omega h^2}{2}A.$$

Zato su njene svojstvene vrijednosti dane sa

$$\lambda_k(G_{JOR,\omega}) = 1 - 2\omega \sin^2\left(\frac{k\pi}{2(n+1)}\right), \qquad k = 1, \ldots, n,$$

a svojstveni vektori su jednaki svojstvenim vektorima matrice A. Važno je još napomenuti, da se iz izraza za svojstenu vrijednost vidi, kako za slučaj kada je $0 < \omega \leq 1$, vrijedi da je $|\lambda_k(G_{JOR,\omega})| < 1$, odnosno da JOR metoda konvergira. Preostalo nam je još pronaći ω za koji se postiže najbolja stopa konvergencije, odnosno takav ω, za koji je $|\lambda_k(G_{JOR,\omega})|$ najmanji mogući za sve $k = 1, \ldots, n$. Slika 4.1 je prikaz svojstvenih vrijednosti $\lambda_k(G_{JOR,\omega})$ za različite vrijednosti ω, koje su nacrtane kao kontinuirana varijabla na intervalu $k \in \langle 0, n+1\rangle$.



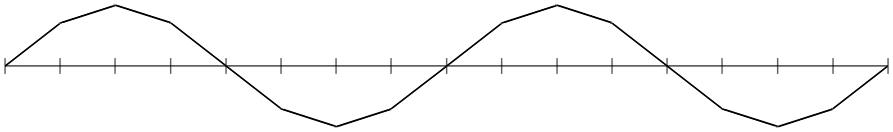
Slika 4.1: Svojstvene vrijednosti JOR matrice $G_{JOR,\omega}$ za $\omega = \frac{1}{3},\ \frac{1}{2},\ \frac{2}{3},\ 1$.

Za nastavak analize potrebno nam je uvesti nekoliko pojmova. Kad bismo izraz za koordinate pojedinih svojstvenih vektora (4.3) prikazali na domeni $\langle 0, 1\rangle$, i između diskretnih vrijednosti $j = 0, 1, \ldots, n+1$, provukli linearni interpolant, tada bismo uočili da tako dobivena po dijelovima linearna funkcija predstavljaju malo ili jako oscilatorne valove, s obzirom da li je parametar k bliži 1 ili je bliži n. Radi se o *Fourierovim modovima*, odnosno vektorima čije koordinate možemo definirati preko jednakosti (4.3), a čije oscilacije ovise o *frekvenciji* k. Zato se pojavljuje prirodna potreba da Fourijerove modove podijelimo u dvije skupine. Modove, za koje je $1 \leq k < \frac{n+1}{2}$, nazivamo *niskofrekventnim* ili *glatkim*, dok one sa $\frac{n+1}{2} \leq k \leq n$ nazivamo *visokofrekventnim* ili *oscilatornim* modovima. Poslije ćemo vidjeti da te dvije skupine imaju i različitu ulogu u multigrid procesu.

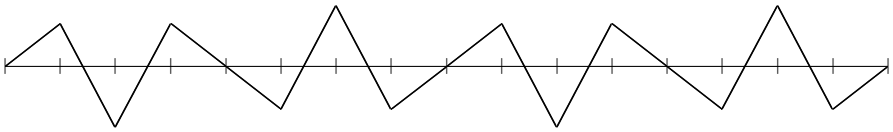
Primijetimo da je za sve vrijednosti od $\omega \in \langle 0, 1]$

$$\lambda_1(G_{JOR,\omega}) = 1 - 2\omega \sin^2\left(\frac{\pi}{2(n+1)}\right) = 1 - 2\omega \sin^2\left(\frac{\pi h}{2}\right) \approx 1 - \frac{\omega\pi^2 h^2}{2}. \qquad (4.4)$$

Odavde slijedi da će $\lambda_1(G_{JOR,\omega})$, svojstvena vrijednost pridružena najglatkijem Fourierovom modu, uvijek biti blizu 1. Budući da za iteriranje greške vrijedi $e_k = G^k_{JOR,\omega} e_0$,



Slika 4.2: Koordinate glatkog Fourierovog moda $q^{(k)}$ za $n = 15$ i $k = 4$.



Slika 4.3: Koordinate oscilatornog Fourierovog moda $q^{(k)}$ za $n = 15$ i $k = 12$.

i u slučaju da početnu grešku predstavimo kao linearnu kombinaciju svojstvenih vektora od $G_{JOR,\omega}$

$$e_0 = \sum_{i=1}^{n} c_i q^{(i)},$$

tada ćemo nakon k koraka JOR metode imati

$$e_k = \sum_{i=1}^{n} c_i \lambda_i^k (G_{JOR,\omega}) q^{(i)}.$$

Zbog toga, niti jedna vrijednost od ω neće moći uspješno eliminirati komponente vektora greške koje bi bile u smjeru glatkih modova. Što više, što je manji korak mreže h, u svrhu postizanja što točnije diskretizacije, to je $\lambda_1(G_{JOR,\omega})$ bliži 1, i konvergencija glatkih komponenata greške je sve gora.

Budući da se moramo pomiriti sa činjenicom da niti jedna vrijednost od ω neće na zadovoljavajući način eliminirati glatke komponente greške, postavlja se pitanje koje vrijednosti od ω će na najbolji način eliminirati oscilatorne komponente. Iz svojstava krivulja prikazanih u Slici 4.1, ovaj uvjet možemo predstaviti kao zahtjev da je

$$\lambda_{\frac{n+1}{2}}(G_{JOR,\omega}) = -\lambda_{n+1}(G_{JOR,\omega}). \tag{4.5}$$

Rješavanjem ove jednadžbe dobivamo optimalnu vrijednost $\omega = \frac{2}{3}$.

Za svojstvene vrijednosti pridružene oscilatornim Fourierovim modovima zbog uvjeta (4.5) vrijedi da je

$$\max_{\frac{n+1}{2} \le k \le n+1} |\lambda_k(G_{JOR,\frac{2}{3}})| = \max\{|\lambda_{\frac{n+1}{2}}(G_{JOR,\frac{2}{3}})|, |\lambda_{n+1}(G_{JOR,\frac{2}{3}})|\},$$

a dalje je,

$$|\lambda_{\frac{n+1}{2}}(G_{JOR,\frac{2}{3}})| = 1 - \frac{4}{3} \cdot \frac{1}{2} = \frac{1}{3},$$

i

$$\begin{aligned}|\lambda_{n+1}(G_{JOR,\frac{2}{3}})| &= -1+\frac{4}{3}\sin^2\left(\frac{n\pi}{2(n+1)}\right) = -1+\frac{2}{3}\left[1-\cos\left(\frac{n\pi}{n+1}\right)\right] = \\ &= -1+\frac{2}{3}\left[1+\cos\left(\frac{\pi}{n+1}\right)\right] = -\frac{1}{3}+\frac{2}{3}\cos\left(\frac{\pi}{n+1}\right) < \\ &< \frac{1}{3},\end{aligned}$$

pa možemo zaključiti da za $\omega = \frac{2}{3}$ vrijedi $|\lambda_k(G_{JOR,\omega})| \leq \frac{1}{3}$ za sve $\frac{n+1}{2} \leq k \leq n$. To znači da se oscilatorne komponente greške reduciraju za najmanje faktor 3 prilikom izvršavanje svake iteracije JOR metode. Ovaj reducirajući faktor za oscilatorne modove je jako važno svojstvo klasičnih iterativnih metoda i zovemo ga *izglađujućim faktorom* sheme. Još jedno važno svojstvo je to da je on neovisan o koraku mreže h.

Dakle da rezimiramo, iz Slike 4.1 vidimo da se za $\omega = \frac{2}{3}$ iteriranjem JOR metode glatki modovi sporo reduciraju, dok oscilatorni modovi brzo konvergiraju k nuli. Za razliku od toga, za $\omega = 1$ i vrlo glatki i vrlo oscilatorni modovi se sporo reduciraju, a samo modovi sa frekvencijama blizu $\frac{n+1}{2}$ vrlo brzo nestaju.

Do sada smo se detaljnije koncentrirali na JOR metodu, jer se lako dade analizirati, a dijeli i sva osnovna svojstva sa ostalim klasičnim iterativnim metodama. Može se pokazati da slično vrijedi i za Gauss–Seidelovu metodu (odjeljak 3.2). Za Gauss–Seidelovu matricu, svojstvene vrijednosti su oblika

$$\lambda_k(G_{GS}) = \cos^2\left(\frac{k\pi}{n+1}\right), 1 \leq k \leq n.$$

Vidimo da kada je k blizu 1 ili n , odgovarajuće svojstvene vrijednosti su blizu 1, i konvergencija komponenti u smjeru pridruženih svojstvenih vektora k nuli je spora. S druge strane, svojstveni vektori od G_{GS} su dani sa

$$q_j^{(k)} = \left[\cos\left(\frac{k\pi}{n+1}\right)\right]^j \sin\left(\frac{jk\pi}{n+1}\right),$$

za $j, k = 1, \ldots, n$. Ovi svojstveni vektori se ne poklapaju sa svojstvenim vektorima od A, zato $\lambda_k(G_{GS})$ daje stopu konvergencije komponente u smjeru k-tog svojstvenog vektora matrice G_{GS}, a ne matrice A.

Uglavnom, konvergencijska svojstva JOR metode, dijele i ostale klasične iterativne metode. Sve te metode funkcioniraju jako dobro u prvih nekoliko iteracija, kada se greška naglo smanjuje zbog efikasne eliminacije oscilatornih komponenti greške. Nakon nekog vremena, konvergencija se usporava, jer nakon što su odstranjene oscilatorne komponente, ostaju samo glatke komponente, čija je redukcija prilično neefikasna. Ovakvo svojstvo eliminacije oscilatornih modova i ostavljanje glatkih modova nazivamo *izglađujuće svojstvo*, i predstavlja značajan nedostatak klasičnih iterativnih metoda. Ispravljanje tog nedostatka vodi prema multigrid metodama. Treba još napomenuti da za razvoj multigrid metode, ne trebamo koristiti komplicirane iterativne metode, već su, kao što ćemo vidjeti, za efektnu multigrid metodu dovoljne i jednostavne metode poput JOR-a i Gauss–Seidela.

Jedan način da se poboljša rezultat klasičnih iterativnih metoda je uzimanje dobre početne iteracije, čija greška ne sadrži glatke komponente ili su one vrlo male. Poznata

tehnika za dobivanje bolje početne iteracije je izvođenje preliminarnih iteracija na grubljoj mreži. Iteriranje na grubljoj mreži je jeftinije jer imamo manji broj nepoznanica koje trebamo izračunati. Osim toga, prema (4.4) stopa konvergencije, koja je ekvivalentna spektralnom radijusu matrice klasične iterativne metode, ponaša se kao $1-\mathcal{O}(h^2)$, pa će zbog toga grublja mreža imati bolju stopu konvergencije. Odatle dolazi ideja o daljnjem razmatranju grublje mreže.

Kako nam nakon određenog broja iteracija klasične iterativne metode ostaju samo glatke komponente, postavlja se pitanje kako ti glatki modovi izgledaju na grubljoj mreži. Najprije uvedimo Ω^h kao oznaku za mrežu sa korakom h, koja je definirana na domeni Ω. Ω^{2h} je tada oznaka za grublju mrežu, sa dvostruko većim korakom, i u našem slučaju jednodimenzionalne difuzijske jednadžbe, točke grublje mreže su točke na finoj mreži Ω^h označene sa parnim brojevima. Promotrimo k-ti mod na finoj mreži ali samo u parnim točkama mreže. Ako $1 \leq k < \frac{n+1}{2}$, njegove komponente se mogu napisati kao

$$q_{2j}^{(k)^h} = \sin\left(\frac{2jk\pi}{n+1}\right) = \sin\left(\frac{jk\pi}{(n+1)/2}\right) = q_j^{(k)^{2h}}, \qquad 1 \leq j,k < \frac{n+1}{2},$$

pri čemu h ili $2h$ označava na kojoj mreži je vektor definiran. Zbog toga Ω^h možemo poistovjetiti sa vektorskim prostorom svih vektora v^h definiranih na toj mreži. Iz ove jednakosti vidimo da k-ti mod na Ω^h postaje k-ti mod na Ω^{2h}, a tamo ih ima duplo manje nego na finijoj mreži Ω^h. Važna posljedica ove činjenice je da prelaskom s finije mreže na grublju, mod postaje više oscilatoran. Naročito za $k \geq \frac{n+1}{4}$, kada doslovce postaju oscilatorni modovi na Ω^{2h}. $\frac{n+1}{2}$-ti mod na Ω^h prelazi u nul-vektor na Ω^{2h}, jer je $q_j^{((n+1)/2)^{2h}} = \sin(j\pi) = 0$.

Oscilatorni modovi na finoj mreži sa $k > \frac{n+1}{2}$, prelaskom na grublju mrežu, prolaze kroz ozbiljniju transformaciju. Definirajmo $k' = n+1-k$, i tada je $k' < \frac{n+1}{2}$, pa stoga imamo

$$\begin{aligned} q_{2j}^{(k)^h} &= \sin\left(\frac{2j(n+1-k')\pi}{n+1}\right) = \sin\left(2j\pi - \frac{jk'\pi}{(n+1)/2}\right) = \\ &= \sin\left(-\frac{jk'\pi}{(n+1)/2}\right) = -\sin\left(\frac{jk'\pi}{(n+1)/2}\right) = \\ &= -q_j^{(n+1-k)^{2h}}. \end{aligned} \tag{4.6}$$

Dakle, k-ti mod na Ω^h postaje $(n+1-k)$-ti mod na Ω^{2h}, kada je $k > \frac{n+1}{2}$. Drugim riječima, oscilatorni modovi na Ω^h postaju relativno glatki modovi na Ω^{2h}.

Ono važno, što sada možemo zaključiti, je to da glatki modovi na finoj mreži izgledaju manje glatki na grubljoj mreži. To nam daje sugestiju da se, kada iteracije klasičnih iterativnih metoda naglo uspore svoju konvergenciju zbog dominacije glatkih modova, prebacimo na grublju mrežu. Tamo će glatki modovi brže konvergirati jer izgledaju više oscilatorno.

Dakle, glavna ideja multigrid metoda je iteriranje neke iterativne metode na raznim mrežama, odnosno prebacivanje početnog problema sa polazne na grublju mrežu, kako bi se izvršila korekcija aproksimacije rješenja. Sada se postavlja pitanje komunikacije među mrežama, odnosno kako prebaciti vektore i matrice s jedne mreže na drugu. U razmatranju prijenosa među mrežama, uzimat ćemo u obzir samo one slučajeve kada grublja mreža ima korak dva puta veći od prve finije mreže. Najprije samo trebamo istaknuti jedan dogovor o oznakama vektora: vektore koji odgovaraju mreži Ω^h označavamo sa

gornjim indeksom h (v^h). Prva klasa prijenosa među mrežama je prebacivanje vektora sa fine mreže na grublju. Oni se općenito zovu operatori *restrikcije* i označavaju sa I_h^{2h}. Najočigledniji operator restrikcije je *injekcija*, a definiran je sa $I_h^{2h}v^h = v^{2h}$, gdje je

$$v_j^{2h} = v_{2j}^h, \qquad j = 1, \ldots, \frac{n-1}{2}$$

Drugim riječima, primjenom injekcije vektor na grubljoj mreži dobiva svoje komponenete direktno iz odgovarajućeg vektora na finoj mreži.

Alternativni operator restrikcije, kojeg nazivamo *potpunim težinskim sumiranjem*, definiran je sa $I_h^{2h}v^h = v^{2h}$, gdje je

$$v_j^{2h} = \frac{1}{4}(v_{2j-1}^h + 2v_{2j}^h + v_{2j+1}^h), \qquad j = 1, \ldots, \frac{n-1}{2}.$$

Kao što vidimo iz definicije, komponente vektora na grubljoj mreži su težinske sume, bolje reći konveksne kombinacije, komponenti koje odgovaraju susjednim točkama pripadne točke na finoj mreži. Mi ćemo od sada pa na dalje koristiti ovaj operator restrikcije iz razloga koje ćemo kasnije navesti. Potpuno težinsko sumiranje je linearan operator sa $\mathbb{R}^n$ u $\mathbb{R}^{\frac{n-1}{2}}$, ima puni rang jednak $\frac{n-1}{2}$ i jezgru dimenzije $\frac{n+1}{2}$. Matrični mu je oblik

$$I_h^{2h} = \frac{1}{4}\underbrace{\left[\begin{array}{ccccccccccc} 1 & 2 & 1 & & & & & & & & \\ & & 1 & 2 & 1 & & & & & & \\ & & & & 1 & 2 & 1 & & & & \\ & & & & & & \ddots & & & & \\ & & & & & & & & 1 & 2 & 1 \end{array}\right]}_{n} \Bigg\} \frac{n-1}{2}.$$

U dvodimenzionalnom slučaju, operator potpunog težinskog sumiranja je ponovo uzimanje konveksne kombinacije komponenti vektora koje odgovaraju susjednim točkama fine mreže. Ako je $I_h^{2h}v^h = v^{2h}$, tada je

$$\begin{aligned} v_{ij}^{2h} &= \frac{1}{16}[v_{2i-1,2j-1}^h + v_{2i-1,2j+1}^h + v_{2i+1,2j-1}^h + v_{2i+1,2j+1}^h + \\ &\quad +2(v_{2i,2j-1}^h + v_{2i,2j+1}^h + v_{2i-1,2j}^h + v_{2i+1,2j}^h) + 4v_{2i,2j}^h], \end{aligned}$$

$$i, j = 1, \ldots, \frac{n-1}{2}.$$

Budući da je ideja multigrida korekcija aproksimacije rješenja sustava pomoću iteracija na grubljoj mreži od polazne, tada je potrebno tu korekciju prikazati ponovo na polaznoj, finoj mreži. Zato uvodimo drugu klasu prijenosa među mrežama za prebacivanje vektora sa grube mreže na finu. Takvi operatori se općenito zovu operatori *interpolacije*, i označavaju se sa I_{2h}^h. U našem slučaju koristit ćemo najjednostavniji oblik interpolacije, a to je *linearna interpolacija*, zato što je ona dovoljno djelotvorna. Ovaj operator uzima vektor sa grube mreže, i pretvara ga u vektor fine mreže prema pravilu $I_{2h}^h v^{2h} = v^h$, gdje je

$$\begin{aligned} v_{2j}^h &= v_j^{2h}, \\ v_{2j+1}^h &= \frac{1}{2}(v_j^{2h} + v_{j+1}^{2h}), \qquad j = 0, \ldots, \frac{n-1}{2}. \end{aligned}$$

Komponente koje odgovaraju parnim točkama fine mreže se direktno prebacuju sa Ω^{2h} na Ω^h. Komponente koje odgovaraju neparnim točkama fine mreže, dobivaju se kao aritmetička sredina komponenti koje odgovaraju susjednim točkama na gruboj mreži. Operator linearne interpolacije je linearni operator sa $\mathbb{R}^{\frac{n-1}{2}}$ u $\mathbb{R}^n$, punog ranga i trivijalne jezgre. Matrični mu je oblik

$$I_{2h}^h = \frac{1}{2}\underbrace{\begin{bmatrix} 1 & & & & \\ 2 & & & & \\ 1 & 1 & & & \\ & 2 & & & \\ & 1 & 1 & & \\ & & 2 & & \\ & & 1 & \ddots & \\ & & & & 1 \\ & & & & 2 \\ & & & & 1 \end{bmatrix}}_{\frac{n-1}{2}} \Bigg\} n.$$

Za dvodimenzionalan slučaj, operator interpolacije se može definirati na sličan način. Ako je $I_{2h}^h v^{2h} = v^h$, tada su komponente od v^h dane sa

$$\begin{aligned} v_{2i,2j}^h &= v_{ij}^{2h}, \\ v_{2i+1,2j}^h &= \frac{1}{2}(v_{ij}^{2h} + v_{i+1,j}^{2h}), \\ v_{2i,2j+1}^h &= \frac{1}{2}(v_{ij}^{2h} + v_{i,j+1}^{2h}), \\ v_{2i+1,2j+1}^h &= \frac{1}{4}(v_{ij}^{2h} + v_{i+1,j}^{2h} + v_{i,j+1}^{2h} + v_{i+1,j+1}^{2h}), \end{aligned}$$

$$i, j = 0, \ldots, \frac{n-1}{2}.$$

Sada promotrimo neka osnovna svojstva interpolacije. Pretpostavimo, najprije, da je prava greška (koja nam nije egzaktno poznata) gladak vektor kada se prikaže na finoj mreži. Pretpostavimo također da je nađena aproksimacija greške na gruboj mreži, i ta je aproksimacija po definiciji egzaktna u točkama grube mreže. Kada se ta aproksimacija na gruboj mreži interpolira na finu mrežu, interpolant je također gladak, jer linearno interpoliranje ima izglađujući efekt. Zbog toga očekujemo da je on dobra aproksimacija greške na finoj mreži. Naprotiv, ako je prava greška oscilatorna, čak i vrlo dobra aproksimacija na gruboj mreži, može proizvesti gladak interpolant, koji neće biti dobra aproksimacija greške na finoj mreži. Prema tome interpolacija je najefektnija kada je greška glatka, i s time predstavlja sretnu nadopunu klasičnim iterativnim metodama, koje su najefektnije kada je greška oscilatorna. Zato se obično najprije izvršava iteriranje klasiče iterativne metode, koja eliminira oscilatorne komponente greške i ostavlja samo glatke, a zatim se prelazi na korekciju na grubljoj mreži, na kojoj se rješava rezidualna jednadžba kako bi se greška točno izračunala, budući da će ona pri ponovnoj intepolaciji na finu mrežu biti precizno prenijeta, postupak ima smisla.

Nadalje, trebamo obratiti pažnju na ovakav izbor operatora redukcije i interpolacije. Naime, jedan razlog za odabir potpunog težinskog sumiranja kao operatora redukcije je važnost činjenice

$$I_{2h}^h = c(I_h^{2h})^T, \qquad c \in \mathbb{R}.$$

Činjenica da je operator linearnog interpoliranja jednak transponiranom operatoru potpunog težinskog sumiranja, do na konstantu, zove se *varijacijsko svojstvo*, i uskoro će se pokazati kao svojstvo od velike važnosti.

Kao što smo već prije spomenuli, nakon nekog broja iteracija iterativne metode na finoj mreži, cijeli se problem prebacuje na grublju mrežu, na kojoj se ponovo vrši određeno iteriranje. Definirali smo kako vektore možemo prebacivati s jedne mreže na drugu, međutim ostalo nam je još otvoreno pitanje vezano uz operator, odnosno matricu kao njegovu reprezentaciju na grubljoj mreži. Trebamo definirati matricu A^{2h} na mreži Ω^{2h} kao neku verziju originalne matrice A^h na mreži Ω^h.

Analiza koja slijedi temelji se na pretpostavci da radimo sa problemom $-u''(x) = f(x)$, i da je odgovarajući diskretni operator A^h. Neka je v^h izračunata aproksimacija egzaktnog rješenja u^h sustava $A^h u^h = f^h$ dobivenog diskretizacijom na mreži Ω^h. Za sada pretpostavimo da greška te aproksimacije $e^h = u^h - v^h$ leži u potpunosti unutar slike interpolacije, koju označavamo sa $\mathcal{R}(I_{2h}^h)$. To znači da je za neki vektor g^{2h}, koji je definiran na mreži Ω^{2h}, $e^h = I_{2h}^h g^{2h}$. Prema tome, rezidualnu jednadžbu na Ω^h možemo napisati kao

$$r^h = A^h e^h = A^h I_{2h}^h g^{2h}. \tag{4.7}$$

Budući da u ovoj jednakosti A^h djeluje na vektor koji leži u slici interpolacije, možemo zaključiti kako A^h djeluje na $\mathcal{R}(I_{2h}^h)$. Ako promatramo vektor $I_{2h}^h g^{2h}$, koji se nalazi u $\mathcal{R}(I_{2h}^h)$, tada su koordinate tog vektora koje odgovaraju točkama fine mreže $2j$, $2j+1$, i $2j+2$, za $j = 0, \ldots, (n-1)/2$, jednake redom: g_j^{2h}, $(g_j^{2h} + g_{j+1}^{2h})/2$, i g_{j+1}^{2h}. Ako gledamo $(2j+1)$-u koordinatu vektora $A^h I_{2h}^h g^{2h}$, tada prema definiciji matrice $A^h = A$ (4.1), zaključujemo da u njenom dobivanju sudjeluju samo prethodno navedene koordinate vektora $I_{2h}^h g^{2h}$, i vrijedi

$$(A^h I_{2h}^h g^{2h})_{2j+1} = \frac{1}{h^2}\left(-1 \cdot g_j^{2h} + 2 \cdot \frac{g_j^{2h} + g_{j+1}^{2h}}{2} - 1 \cdot g_{j+1}^{2h}\right) = 0.$$

Dakle, zaključujemo da su one komponente od $A^h I_{2h}^h g^{2h}$, koje odgovaraju neparnim točkama mreže Ω^h, jednake nuli. Taj efekt je analogan uzimanju druge derivacije od po dijelovima linearne funkcije.

Sada možemo zaključiti da su neparni reci matrice $A^h I_{2h}^h$ u (4.7) jednaki nuli. S druge strane, parni reci iste matrice odgovaraju točkama grube mreže Ω^{2h}. Zbog toga možemo naći verziju rezidualne jednadžbe na gruboj mreži tako da izbacimo neparne retke u (4.7). To formalno možemo postići upotrebom operatora restrikcije I_h^{2h} na obje strane u (4.7). Time rezidualna jednadžba prelazi u oblik

$$\underbrace{I_h^{2h} A^h I_{2h}^h}_{A^{2h}} g^{2h} = I_h^{2h} r^h.$$

Ova analiza daje nam logičnu definiciju za matricu na gruboj mreži:

$$A^{2h} = I_h^{2h} A^h I_{2h}^h.$$

Elementi od A^{2h} se mogu eksplicitno izračunati kao što je prikazano u Tablici 4.1. Ovdje su prikazani netrivijalni elementi j-tog stupca od A^{2h}, koji se dobiju primjenom $I_h^{2h} A^h I_{2h}^h$ na j-ti jedinični vektor ξ_j^{2h}, definiran na Ω^{2h}. Također možemo zaključiti da

	$j-1$		j		$j+1$
ξ_j^{2h}	0		1		0
$I_{2h}^h \xi_j^{2h}$	0	$\frac{1}{2}$	1	$\frac{1}{2}$	0
$A^h I_{2h}^h \xi_j^{2h}$	$-\frac{1}{2h^2}$	0	$\frac{1}{h^2}$	0	$-\frac{1}{2h^2}$
$I_h^{2h} A^h I_{2h}^h \xi_j^{2h}$	$-\frac{1}{4h^2}$		$\frac{1}{2h^2}$		$-\frac{1}{2h^2}$

Tablica 4.1: Dobivanje j-tog stupca operatora $A^{2h} = I_h^{2h} A^h I_{2h}^h$.

je j-ti stupac od A^{2h} jednak j-tom stupcu matrice

$$\frac{1}{(2h)^2} \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix},$$

koju bismo dobili da smo originalni problem diskretizirali na gruboj mreži Ω^{2h}. Prema tome, po ovoj definiciji A^{2h} je zaista verzija od A^h samo na mreži Ω^{2h}.

Prethodna razmatranja bila su bazirana na pretpostavci da greška e^h leži u slici interpolacije. To općenito nije slučaj, međutim gornja definicija za A^{2h} u svakom slučaju ima smisla. Time možemo upotpuniti *varijacijska svojstva*, koja su sada dana sa

$$A^{2h} = I_h^{2h} A^h I_{2h}^h \qquad \text{(Garlekinov uvjet)}, \tag{4.8}$$

$$I_h^{2h} = c(I_{2h}^h)^T, \qquad c \in \mathbb{R}. \tag{4.9}$$

Sada smo pobrojali sve elemente multigrid metode, pa nam je preostalo još samo da ih uklopimo u jednu shemu. Postoje dva načina na koji možemo iskoristiti korekciju greške pomoću grublje mreže.

Najprije ćemo iznijeti strategiju koja koristi grube mreže kako bi ostvarila bolje početne iteracije za neku od klasičnih iterativnih metoda.

- Primijeni iterativnu metodu nad $A^{ph}u = f^{ph}$ na vrlo gruboj mreži Ω^{ph} kako bismo dobili početnu iteraciju za sljedeću finiju mrežu.

$$\vdots$$

- Primijeni iterativnu metodu nad $A^{4h}u = f^{4h}$ na Ω^{4h} kako bismo dobili početnu iteraciju za Ω^{2h}.

- Primijeni iterativnu metodu nad $A^{2h}u = f^{2h}$ na Ω^{2h} kako bismo dobili početnu iteraciju za Ω^h.

- Primijeni iterativnu metodu nad $A^h u = f^h$ na Ω^h kako bismo dobili konačnu aproksimaciju rješenja.

Prelaskom s jedne mreže na drugu, aproksimaciju rješenja prebacujemo pomoću operatora interpolacije, a na svakoj mreži koristimo odgovarajuću matricu A^{ih}. Ova ideja koristi grublje mreže kako bi generirala poboljšane početne iteracije, i ona je baza strategije koju nazivamo *ugniježđene iteracije*. Iako je ovakav pristup vrlo privlačan, ipak ostavlja neka otvorena pitanja. Na primjer, što će se dogoditi, kada jednom dođemo do fine mreže, a greška i dalje sadrži neke glatke komponente? Mi smo možda ostvarili neko poboljšanje korištenjem grubih mreža, ali konačno iteriranje će opet zakazati zbog prisutnosti glatkih komponenti. Poslije ćemo vidjeti, da postoji odgovor na to pitanje, koje će nam omogućiti korištenje ugniježđenih iteracija na vrlo moćan način.

Druga strategija se sastoji od ideje korištenja rezidualne jednadžbe i primjene iterativne metode nad njom, kako bi se dobila aproksimacija greške.

- Primijeni iterativnu metodu nad $A^h u = f^h$ na Ω^h kako bismo dobili aproksimaciju v^h.

- Izračunaj rezidual $r^h = f^h - A^h v^h$.

  Primijeni iterativnu metodu nad $A^{2h}e = r^{2h}$ na Ω^{2h} kako bismo dobili aproksimaciju greške e^{2h}.

- Korigiraj aproksimaciju rješenja dobivenu na Ω^h sa aproksimacijom greške dobivenom na Ω^{2h}: $v^h := v^h + e^h$.

Ponovo, za prelaz sa fine mreže na grublju koristimo operator restrikcije, kako bismo dobili r^{2h}, a sa grube na finu koristimo operator interpolacije, kako bismo dobili e^h. Ova procedura je baza *korektivne sheme*. Primijenimo iterativnu metodu na finoj mreži dok konvergencije ne počne usporavati, zatim primijenimo iterativnu metodu nad rezidualnom jednadžbom na grubljoj mreži kako bismo dobili aproksimaciju greške. Tada se vraćamo na finu mrežu kako bismo korigirali aproksimaciju koju smo najprije dobili. Ponovo postoje dobri razlozi za upotrebu ove sheme, ali opet ostaju i neka pitanja. Na primjer, koju početnu iteraciju upotrijebiti za iteriranje nad rezidualnom jednadžbom? I na to ćemo ubrzo odgovoriti.

Ovo su bile samo glavne ideje shema. Budući da su nam sad na raspolaganju svi alati koji su potrebni za multrigrid metodu, ove sheme možemo i preciznije napisati.

Korektivna shema na dvije mreže

$$v^h = MG(v^h, f^h).$$

- Iteriraj ν_1 puta iterativnu metodu nad $A^h u^h = f^h$ na Ω^h sa početnom iteracijom v^h.

- Izračunaj rezidual na finoj mreži $r^h = f^h - A^h v^h$ i prenesi ga na grubu mrežu sa $r^{2h} = I_h^{2h} r^h$.

- Riješi $A^{2h}e^{2h} = r^{2h}$ na Ω^{2h}.
- Interpoliraj grešku sa grube mreže na finu mrežu sa $e^h = I_{2h}^h e^{2h}$ i korigiraj aproksimaciju na finoj mreži sa $v^h := v^h + e^h$.
- Iteriraj ν_2 puta iterativnu metodu nad $A^h u^h = f^h$ na Ω^h sa početnom iteracijom v^h.

U ovoj proceduri, iteriramo na finoj mreži tako dugo dok se to isplati, u praksi ν_1 je obično 1,2, ili 3. Nenegativni cijeli brojevi ν_1 i ν_2 su parametri sheme koji kontroliraju broj iteracija iterativne metode prije i poslije obilaska grube mreže. Oni su obično određeni na početku, na osnovu teoretskih razmatranja ili prethodnih eksperimentalnih rezultata. Kada rješavamo rezidualnu jednadžbu na gruboj mreži, egzaktno rješenje nam također neće biti dostupno, već samo njegova aproksimacija, kojom korigiramo aproksimaciju rješenja na finoj mreži.

Važno je ponovo naglasiti komplementarnost koja dolazi do izražaja u ovoj proceduri. Iteracije iterativne metode na finoj mreži eliminiraju oscilatorne komponente greške, ostavljajući realativno glatku grešku. Ako pretpostavimo da rezidualnu jednadžbu možemo riješiti egzaktno na Ω^{2h}, još je važno točno prebaciti grešku na finu mrežu. Budući da je greška glatka, interpolacija bi je trebala vrlo precizno prebaciti na Ω^h, i korekcija aproksimacije rješenja na finoj mreži bi trebala biti djelotvorna.

Korektivna shema na dvije mreže ostavlja jedno pitanje, a to je koji je najbolji način rješavanja problema $A^{2h}e^{2h} = r^{2h}$ na gruboj mreži? Problem na gruboj mreži nije puno drugačiji od originalnog problema. Zbog toga možemo upotrijebiti korektivnu shemu na dvije mreže i za rezidualnu jednadžbu na Ω^{2h}, što znači: iteriranje na toj mreži, a zatim prebacivanje na Ω^{4h} za korekciju. Ovaj proces možemo ponavljati na sve grubljim i grubljim mrežama, dok direktno rješavanje rezidualne jednadžbe ne postane moguće.

Još jednu stvar moramo riješiti, a to je početna iteracija za rješavanje rezidualnog problema na Ω^{2h}. Budući nemamo nikakve informacije o njenom rješenju e^{2h}, jednostavno ćemo uzeti početnu iteraciju jednaku 0. Ionako očekujemo da greška bude što bliža nuli, pa je to logičan izbor.

Sada imamo potpunu sliku multigrid metode. Budući da se korektivna shema na dvije mreže izvodi na sve grubljim mrežama, s time da se na početku iterira nad polaznim problemom, a u ostalim koracima nad rezidualnim problemu, i da su ti problemi vrlo slični, potrebna je ekonomizacija oznaka kao kod računalne implementacije. Desnu stranu rezidualne jednadžbe također ćemo označavati sa f^{2h}, a ne sa r^{2h}, a rješenje rezidualne jednadžbe e^{2h} sa u^{2h}. v^{2h} možemo onda iskoristiti za oznaku aproksimacije od u^{2h}.

Ono što slijedi je korektivna shema na dvije mreže, ali koja ponovo poziva samu sebe, rekurzivno. Nazivamo je *shema V-ciklusa*. Pretpostavljamo da postoji $l > 1$ mreža sa koracima h, $2h$, $4h$,...,$Lh = 2^{l-1}h$.

Shema V-ciklusa

$$v^h = V^h(v^h, f^h).$$

- Iteriraj ν_1 puta iterativnu metodu nad $A^h u^h = f^h$ sa početnom iteracijom v^h.
- Izračunaj $f^{2h} = I_h^{2h} r^h$.

- Iteriraj ν_1 puta iterativnu metodu nad $A^{2h}u^{2h} = f^{2h}$ sa početnom iteracijom $v^{2h} = 0$.
- Izračunaj $f^{4h} = I_{2h}^{4h} r^{2h}$.
  - Iteriraj ν_1 puta iterativnu metodu nad $A^{4h}u^{4h} = f^{4h}$ sa početnom iteracijom $v^{4h} = 0$.
  - Izračunaj $f^{8h} = I_{4h}^{8h} r^{4h}$.

    $\vdots$

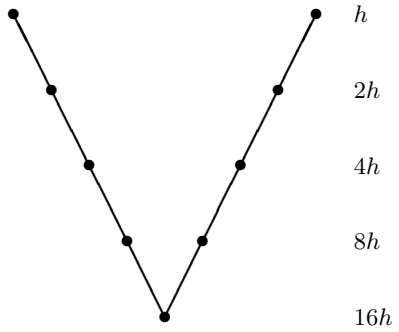
    - Riješi $A^{Lh}u^{Lh} = f^{Lh}$.

    $\vdots$

  - Korigiraj $v^{4h} := v^{4h} + I_{8h}^{4h} v^{8h}$.
  - Iteriraj ν_2 puta iterativnu metodu nad $A^{4h}u^{4h} = f^{4h}$ sa početnom iteracijom v^{4h}.
- Korigiraj $v^{2h} := v^{2h} + I_{4h}^{2h} v^{4h}$.
- Iteriraj ν_2 puta iterativnu metodu nad $A^{2h}u^{2h} = f^{2h}$ sa početnom iteracijom v^{2h}.

- Korigiraj $v^h := v^h + I_{2h}^h v^{2h}$.
- Iteriraj ν_2 puta iterativnu metodu nad $A^h u^h = f^h$ sa početnom iteracijom v^h.

Algoritam ide od najfinije mreže prema najgrubljoj, koja se može sastojati od jedne ili nekoliko unutarnjih točaka, a zatim se vraća ponovo prema najfinijoj mreži. Slika 4.4 pokazuje raspored posjećivanja mreža. Zbog svog oblika, ovaj algoritam se zove V-ciklus i on je prvi predstavnik multigrid metode.



Slika 4.4: Raspored posjeta mreža na 5 nivoa za V-ciklus.

Zbog svoje definicije, V-ciklus ima kompaktnu rekurzivnu definiciju, koja je dana na sljedeći način.

Algoritam 4.1.1. REKURZIVNA SHEMA V-CIKLUSA $v^h = V^h(v^h, f^h)$

1. *Iteriraj ν_1 puta iterativnu metodu nad $A^h u^h = f^h$ sa danom početnom iteracijom v^h.*
2. *Ako je Ω^h najgrublja mreža, tada idi na korak 4.*

   Inače

$$\begin{aligned} f^{2h} &= I_h^{2h}(f^h - A^h v^h), \\ v^{2h} &= 0, \\ v^{2h} &:= V^{2h}(v^{2h}, f^{2h}). \end{aligned}$$

3. *Korigiraj $v^h := v^h + I_{2h}^h v^{2h}$.*
4. *Iteriraj ν_2 puta iterativnu metodu nad $A^h u^h = f^h$ sa početnom iteracijom v^h.*

V-ciklus je samo jedan algoritam iz familije multigrid cikličkih shema. Cijela familija se zove *μ-ciklus* metoda i rekurzivno je definirana na sljedeći način.

Algoritam 4.1.2. REKURZIVNA SHEMA μ-CIKLUSA $v^h = M\mu^h(v^h, f^h)$

1. *Iteriraj ν_1 puta iterativnu metodu nad $A^h u^h = f^h$ sa danom početnom iteracijom v^h.*
2. *Ako je Ω^h najgrublja mreža, tada idi na korak 4.*

   Inače

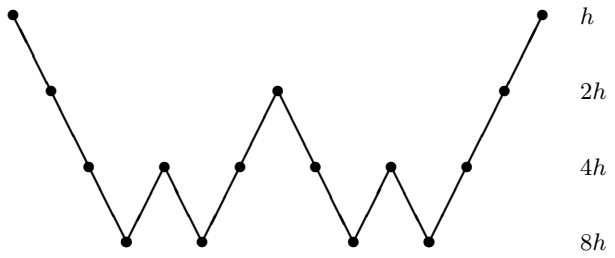
$$\begin{aligned} f^{2h} &= I_h^{2h}(f^h - A^h v^h), \\ v^{2h} &= 0, \\ v^{2h} &:= V^{2h}(v^{2h}, f^{2h}) \ \mu \text{ puta.} \end{aligned}$$

3. *Korigiraj $v^h := v^h + I_{2h}^h v^{2h}$.*
4. *Iteriraj ν_2 puta iterativnu metodu nad $A^h u^h = f^h$ sa početnom iteracijom v^h.*

U praksi se koriste samo sheme sa $\mu = 1$ (V-ciklus), i $\mu = 2$. Slika 4.5 pokazuje raspored posjeta mreža za $\mu = 2$, što rezultira *W-ciklusom.*

Još ćemo uvesti nekoliko oznaka. V-ciklus sa ν_1 iteracija prije korektivnog koraka i sa ν_2 iteracija nakon korektivnog koraka označavamo kao V(ν_1,ν_2)-ciklus. Isto vrijedi iza W(ν_1,ν_2)-ciklus.

Do sada smo samo razrađivali ideju korektivne sheme, a sada ćemo razmotriti ugniježđene iteracija. Ugniježđene iteracije koriste grublju mrežu za dobivanje početne



Slika 4.5: Raspored posjeta mreža na 4 nivoa za W-ciklus sa $\mu = 2$.

iteracije za problem na finoj mreži. Ako s druge strane promatramo V-ciklus, postoji problem odabira početne iteracije za problem na finoj mreži. Ako te dvije sheme udružimo dobivamo *potpuni multigrid V-ciklus (FMG)*, koji je definiran na sljedeći način.

Potpuni multigrid V-ciklus

$$v^h = FMG^h(f^h).$$

Inicijaliziraj $f^{2h} = I_h^{2h} f^h$, $f^{4h} = I_{2h}^{4h} f^{2h}$,...

- Riješi ili primijeni iterativnu metodu na najgrubljoj mreži.

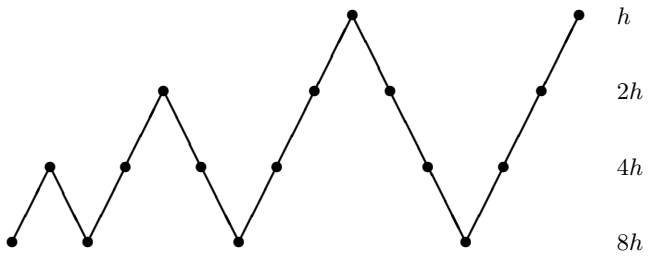
$$\vdots$$

  - $v^{4h} = I_{8h}^{4h} v^{8h}$.
  - $v^{4h} := V^{4h}(v^{4h}, f^{4h})$ ν_0 puta.
- $v^{2h} = I_{4h}^{2h} v^{4h}$.
- $v^{2h} := V^{2h}(v^{2h}, f^{2h})$ ν_0 puta.

- $v^h = I_{2h}^h v^h$.
- $v^h := V^h(v^h, f^h)$ ν_0 puta.

Desne strane problema na grubim mrežama dobivaju se prebacivanjem s mreže na mrežu vektora f^h, počevši od fine mreže. Druga mogućnost je korištenje originalne funkcije f. Parametar ciklusa ν_0, određuje broj V-ciklusa u svakom nivou. Obično se određuje na temelju prethodnih eksperimenata, ali najčešći izbor je $\nu_0 = 1$. Slika 4.6 pokazuje raspored posjeta mreža za FMG sa $\nu_0 = 1$. Svakom V-ciklusu prethodi V-ciklus na grubljoj mreži, koji mu osigurava najbolju moguću početnu iteraciju.

Izražen rekurzivno, algoritam ima sljedeći kompaktan oblik.



Slika 4.6: Raspored posjeta mreža na 4 nivoa za FMG shemu sa $\nu_0 = 1$.

Algoritam 4.1.3. REKURZIVNI POTPUNI MULTIGRID V-CIKLUS $v^h = FMG^h(f^h)$

1. Ako je Ω^h najgrublja mreža, postavi $v^h = 0$ i idi na korak 3.

Inače

$$\begin{aligned} f^{2h} &= I_h^{2h} f^h, \\ v^{2h} &= FMG^{2h}(f^{2h}). \end{aligned}$$

2. Korigiraj $v^h = I_{2h}^h v^{2h}$.

3. $v^h := V^h(v^h, f^h)$ ν_0 puta.

Kao zaključak, možemo primijetiti da su multigrid metode sinteza ideja i tehnika koje su zasebno bile već dugo poznate i upotrebljavane. Ako ih gledamo svaku za sebe, mnoge od tih ideja imaju ozbiljne nedostatke. Potpuni multigrid je tehnika koja ih sve integrira, tako da one mogu zajedno funkcionirati, i to na način koji prerasta njihova ograničenja. Rezultat je vrlo moćan algoritam.

4.2 Spektralna i algebarska slika multigrid metoda: uvod u teoriju konvergencije

Analiza konvergencije multigrid metoda je vrlo komplicirana i predstavlja, još uvijek otvoreno pitanja u numeričkoj matematici. Konvergencija multigrid metoda, koje su primijenjene na dobro uvjetovane probleme, poput skalarnog eliptičnog problema, je rigorozno dokazana, čime je potvrđeno da u tom slučaju multigrid metode funkcioniraju vrlo uspješno. Za općenite probleme, kod kojih još ne postoje analitički rezultati, postoje mnogi eksperimentalni dokazi o njihovoj djelotvornosti.

Heurističku argumentaciju ukratko možemo opisati na sljedeći način. Kao što smo vidjeli, izglađujući faktor (stopa konvergencije oscilatornih modova) za klasične iterativne metode je mala i ne ovisi o koraku mreže h. Budući da glatki modovi, koji ostanu nakon primjene iterativne metode, izgledaju više oscilatorno na grubljim mrežama, prebacivanjem na sve grublju i grublju mrežu, sve će komponente greške kad-tad izgledati

oscilatorno i moći će biti eliminirane pomoću iterativne metode. Odavde slijedi da će konačna stopa konvergencije dobre multigrid sheme biti mala i neovisna o h.

Posvetit ćemo sada više pažnje korektivnoj shemi na dvije mreže, jer je V-ciklus samo ugniježđena primjena korektivne sheme na dvije mreže, a FMG metoda je ponavljana primjena V-ciklusa na raznim mrežama. Zbog toga je razumijevanje korektivne sheme na dvije mreže važno za shvaćanje osnovnih multigrid metoda.

Započet ćemo sa detaljnjijom analizom operatora za prijenos među mrežama. Razmotrimo najprije operator (restrikcije) potpunog težinskog sumiranja I_h^{2h}. To je operator koji preslikava $\mathbb{R}^n \to \mathbb{R}^{\frac{n-1}{2}}$, koji ima rang jednak $\frac{n-1}{2}$ i jezgru $\mathcal{N}(I_h^{2h})$ dimenzije $\frac{n+1}{2}$. Postavlja se pitanje kako I_h^{2h} djeluje na modove originalne matrice A^h?

U našem jednodimenzionalnom primjeru modovi od A^h su dani sa

$$q_j^{(k)^h} = \sin\left(\frac{jk\pi}{n+1}\right), \qquad j,k = 1,\ldots,n.$$

Primijenimo operator potpunog težinskog sumiranja na te vektore. Djelovanje operatora I_h^{2h} na glatke modove rezultira sa

$$\begin{aligned}
(I_h^{2h} q^{(k)^h})_j &= \frac{1}{4}\left[\sin\left(\frac{(2j-1)k\pi}{n+1}\right) + 2\sin\left(\frac{2jk\pi}{n+1}\right) + \sin\left(\frac{(2j+1)k\pi}{n+1}\right)\right] = \\
&= \frac{1}{4}\left[2\sin\left(\frac{2jk\pi}{n+1}\right)\cos\left(\frac{k\pi}{n+1}\right) + 2\sin\left(\frac{2jk\pi}{n+1}\right)\right] = \\
&= \frac{1}{2}\sin\left(\frac{jk\pi}{(n+1)/2}\right)\left[\cos\left(\frac{k\pi}{n+1}\right) + 1\right] = \\
&= \sin\left(\frac{jk\pi}{(n+1)/2}\right)\cos^2\left(\frac{k\pi}{2(n+1)}\right) = \cos^2\left(\frac{k\pi}{2(n+1)}\right) q_j^{(k)^{2h}},
\end{aligned}$$

odnosno

$$I_h^{2h} q^{(k)^h} = \cos^2\left(\frac{k\pi}{2(n+1)}\right) q^{(k)^{2h}}, \qquad 1 \le k < \frac{n+1}{2}.$$

To znači da djelovanjem I_h^{2h} na k-ti, glatki mod od A^h dobivamo konstantu puta k-ti mod od A^{2h}, kada je $1 \le k < \frac{n+1}{2}$. Za $k = \frac{n+1}{2}$ je $I_h^{2h} q^{(\frac{n+1}{2})^h} = 0$, dok za oscilatorne modove, sa $\frac{n+1}{2} < k' \le n$ i $k' = n+1-k$, prema dokazu (4.6) imamo

$$\begin{aligned}
I_h^{2h} q^{(k')^h} &= \cos^2\left(\frac{(n+1-k)\pi}{2(n+1)}\right) q^{(n+1-k)^{2h}} = -\cos^2\left(\frac{\pi}{2} - \frac{k\pi}{2(n+1)}\right) q^{(k)^{2h}} = \\
&= -\sin^2\left(\frac{k\pi}{2(n+1)}\right) q^{(k)^{2h}}, \qquad 1 \le k < \frac{n+1}{2}.
\end{aligned}$$

Odavde možemo zaključiti da kada I_h^{2h} djeluje na $(n+1-k)$-ti, oscilatorni mod od A^h dobivamo konstantu puta k-ti mod od A^{2h}. Operator I_h^{2h} transformira oscilatorne modove na Ω^h u relativno glatke modove na Ω^{2h}.

Na kraju možemo zaključiti da i k-ti i $(n+1-k)$-ti mod na Ω^h, djelovanjem operatora potpunog težinskog sumiranja I_h^{2h}, postaju k-ti mod na Ω^{2h}. Par modova na finoj mreži $\{q^{(k)^h}, q^{(n+1-k)^h}\}$ nazivamo *komplementarnim modovima*. Imamo

$$I_h^{2h}(\mathrm{span}\{q^{(k)^h}, q^{(n+1-k)^h}\}) = \mathrm{span}\{q^{(k)^{2h}}\}.$$

Kako je

$$\sin\left(\frac{j(n+1-k)\pi}{n+1}\right) = \sin\left(j\pi - \frac{jk\pi}{n+1}\right) = (-1)^{j+1}\sin\left(\frac{jk\pi}{n+1}\right),$$

vrijedi da je

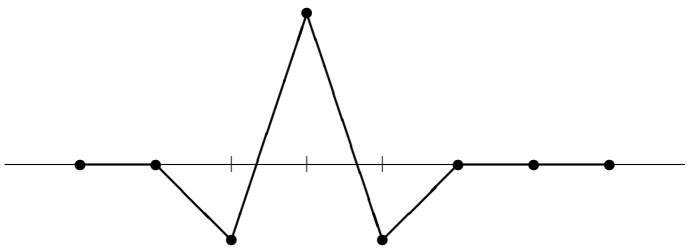
$$q_j^{(n+1-k)^h} = (-1)^{j+1} q_j^{(k)^h}.$$

Time smo dobili spektralna svojstva od I_h^{2h}.

Kao što već znamo od prije, operator potpunog težinskog sumiranja ima netrivijalnu jezgru $\mathcal{N}(I_h^{2h})$. Tvrdimo da je jezgra razapeta vektorima $n_j = A^h \xi_j^h$, gdje je j-neparan broj, a ξ_j^h je j-ti jedinični vektor na Ω^h. Zapravo radi se o neparnim stupcima matrice A^h. Dokažimo tu tvrdnju. Najprije promatramo produkt matrice I_h^{2h} sa neparnim stupcima od A^h, i zanima nas koji neparni stupci od A^h mogu dati netrivijalni skalarni produkt sa i-tim retkom od I_h^{2h}, $i = 1, \ldots, \frac{n-1}{2}$. Jedini kandidati su $(2i-1)$-i i $(2i+1)$-i stupac, za ostale stupce skalarni produkt u svakom sumandu ima produkt s nulom. Promotrimo komponente danog retka od I_h^{2h} i stupaca od A^h u sljedećoj tablici

	$2i-2$	$2i-1$	$2i$	$2i+1$	$2i+2$
i-ti redak od I_h^{2h}	0	1	2	1	0
$(2i-1)$-i stupac od A^h	-1	2	-1	0	0
$(2i+1)$-i stupac od A^h	0	0	-1	2	-1

odakle vidimo da je skalarni produkt i-tog retka od I_h^{2h} sa $(2i-1)$-im stupcem od A^h jednak $2-2=0$, a sa $(2i+1)$-im stupcem jednak $-2+2=0$. Dakle, budući da smo uzeli proizvoljni redak, možemo zaključiti da je skalarni produkt bilo kojeg retka od I_h^{2h} i bilo kojeg neparnog stupca od A^h jednak nuli, odnosno produkt matrice I_h^{2h} sa bilo kojim neparnim stupcem od A^h je jednak nulvektoru. Stoga se vektori n_j, za neparni j, nalaze u $\mathcal{N}(I_h^{2h})$. Lako se vidi da su oni i linearno nezavisni, ima ih $\frac{n+1}{2}$, pa stoga čine bazu od $\mathcal{N}(I_h^{2h})$. Kao što vidimo u Slici 4.7 vektori baze n_j izgledaju prilično oscilatorno. Međutim oni se ne poklapaju sa oscilatornim modovima od A^h. Rastav vektora n_j po



Slika 4.7: Tipični vektor baze jezgre operatora potpunog težinskog sumiranja $\mathcal{N}(I_h^{2h})$

modovima od A^h zahtijeva sve modove, a ne samo oscilatorne. Zato jezgra od I_h^{2h} sadrži i oscilatorne i glatke modove od A^h, samo što su komponente u smjerovima glatkih modova male. Sada ćemo sličnu analizu napraviti i za operator interpolacije I_{2h}^h. To je operator koji preslikava $\mathbb{R}^{\frac{n-1}{2}} \to \mathbb{R}^n$ i ima puni rang. Kako bismo dobili spektralna

svojstva i od I_{2h}^h, trebamo naći kako I_{2h}^h djeluje na modovima od A^{2h}. Neka su

$$q_j^{(k)^{2h}} = \sin\left(\frac{jk\pi}{(n+1)/2}\right), \qquad j,k = 1,\ldots,\frac{n-1}{2},$$

modovi na Ω^{2h}. Pokazat ćemo da I_{2h}^h ne čuva te modove. Posebno ćemo gledati parne, a posebno neparne čvorove od $I_{2h}^h q^{(k)^{2h}}$. Neka je $j = 0,\ldots,\frac{n-1}{2}$, $k = 1,\ldots,\frac{n-1}{2}$ i $k' = n+1-k$, tada imamo

$$\begin{aligned}
\left(I_{2h}^h q^{(k)^{2h}}\right)_{2j+1} &= \frac{1}{2}\left(q_j^{(k)^{2h}} + q_{j+1}^{(k)^{2h}}\right) = \\
&= \frac{1}{2}\left[\sin\left(\frac{2jk\pi}{n+1}\right) + \sin\left(\frac{2(j+1)k\pi}{n+1}\right)\right] = \\
&= \sin\left(\frac{(2j+1)k\pi}{n+1}\right)\cos\left(\frac{k\pi}{n+1}\right) = \\
&= \sin\left(\frac{(2j+1)k\pi}{n+1}\right)\left[\cos^2\left(\frac{k\pi}{2(n+1)}\right) - \sin^2\left(\frac{k\pi}{2(n+1)}\right)\right] = \\
&= \cos^2\left(\frac{k\pi}{2(n+1)}\right) q_{2j+1}^{(k)^h} - (-1)^{2j+2}\sin^2\left(\frac{k\pi}{2(n+1)}\right) q_{2j+1}^{(k)^h} = \\
&= \cos^2\left(\frac{k\pi}{2(n+1)}\right) q_{2j+1}^{(k)^h} - \sin^2\left(\frac{k\pi}{2(n+1)}\right) q_{2j+1}^{(k')^h}
\end{aligned}$$

i

$$\begin{aligned}
\left(I_{2h}^h q^{(k)^{2h}}\right)_{2j} &= q_j^{(k)^{2h}} = \sin\left(\frac{2jk\pi}{n+1}\right) = q_{2j}^{(k)^h} = \\
&= \left[\cos^2\left(\frac{k\pi}{2(n+1)}\right) + \sin^2\left(\frac{k\pi}{2(n+1)}\right)\right] q_{2j}^{(h)^h} = \\
&= \cos^2\left(\frac{k\pi}{2(n+1)}\right) q_{2j}^{(k)^h} - (-1)^{2j+1}\sin^2\left(\frac{k\pi}{2(n+1)}\right) q_{2j}^{(k)^h} = \\
&= \cos^2\left(\frac{k\pi}{2(n+1)}\right) q_{2j}^{(k)^h} - \sin^2\left(\frac{k\pi}{2(n+1)}\right) q_{2j}^{(k')^h}
\end{aligned}$$

odakle možemo zaključiti da je

$$I_{2h}^h q^{(k)^{2h}} = \cos^2\left(\frac{k\pi}{2(n+1)}\right) q^{(k)^h} - \sin^2\left(\frac{k\pi}{2(n+1)}\right) q^{(k')^h}, \qquad k = 1,\ldots,\frac{n-1}{2}.$$

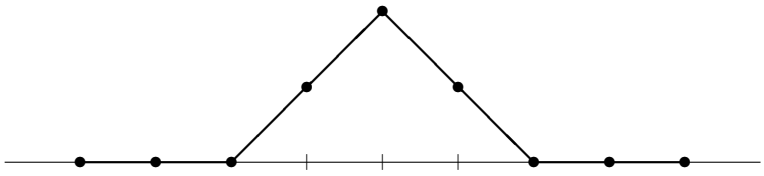
Vidimo da djelovanje od I_{2h}^h na k-ti mod na Ω^{2h} proizvodi ne samo k-ti mod na Ω^h, već i komplementarni k'-ti mod. To otkriva zanimljivo svojstvo, a to je da interpolacije glatkih modova na Ω^{2h} ponovo uvodi oscilatorne modove na Ω^h. Primijetimo, da je za jako glatki mod na Ω^{2h} sa $k \ll \frac{n-1}{2}$

$$I_{2h}^h q^{(k)^{2h}} = \left[1 - \mathcal{O}\left(\frac{k^2}{(n+1)^2}\right)\right] q^{(k)^h} + \mathcal{O}\left(\frac{k^2}{(n+1)^2}\right) q^{(k')^h}.$$

U tom slučaju, rezultat interpolacije se u većini poklapa sa odgovarajućim glatkim modom na Ω^h, uz vrlo malo smetnje od strane komplementarnog oscilatornog moda.

Već smo se uvjerili u važnost slike interpolacije $\mathcal{R}(I_{2h}^h)$. Baza za $\mathcal{R}(I_{2h}^h)$ je dana sa stupcima od I_{2h}^h. Dok vektori baze izgledaju prilično glatko, kao što Slika 4.8 pokazuje,

oni se ne poklapaju sa glatkim modovima od A^h. Može se pokazati, da bilo koji od ovih vektora baze zahtijeva sve modove od A^h za svoj rastav po komponentama. Drugim riječima, slika interpolacije sadrži i glatke i oscilatorne modove od A^h, samo što glatke komponente prevladavaju.



Slika 4.8: Tipični vektor baze slike operatora interpolacije $\mathcal{R}(I_{2h}^h)$

S ovom analizom operatora prijenosa među mrežama, ponovo se vraćamo korektivnoj shemi na dvije mreže. Kao što smo već prije iznijeli, iteracija klasične iterativne metode može se izraziti kao

$$v_k^h = v_{k-1}^h + (M_G^h)^{-1}(f^h - A^h v_{k-1}^h) = (I - (M_G^h)^{-1}A^h)v_{k-1}^h + (M_G^h)^{-1}f^h,$$

a ako identificiramo $G^h = I - (M_G^h)^{-1}A^h$ kao matricu iterativne metode, dobivamo

$$v_k^h = G^h v_{k-1}^h + (M_G^h)^{-1}f^h.$$

Induktivno se lako vidi da je

$$v_k^h = (G^h)^k (A^h)^{-1}(A^h v_0^h - f^h) + (A^h)^{-1}f^h,$$

ili jednostavnije

$$v_k^h = (G^h)^k v_0^h + (I - (G^h)^k)(A^h)^{-1}f^h.$$

Koraci korektivne sheme na dvije mreže, sa egzaktnim rješenjem na gruboj mreži, dani su sljedećom shemom.

- Iteriraj ν puta na Ω^h sa matricom iteracije G^h:
  $v^h := (G^h)^\nu v^h + (I - (G^h)^\nu)(A^h)^{-1}f^h$.
- Prenesi, pomoću operatora potpunog težinskog sumiranja, r^h na Ω^{2h}: $f^{2h} = I_h^{2h}(f^h - A^h v^h)$.
- Egzaktno riješi rezidualnu jednadžbu: $v^{2h} = (A^{2h})^{-1}f^{2h}$.
- Korigiraj aproksimaciju na Ω^h: $v^h := v^h + I_{2h}^h v^{2h}$.

Ako sa v_{stari}^h označimo staru vrijednost aproksimacije rješenja na Ω^h, prije primjene korektivne sheme na dvije mreže, i ako sa v_{novi}^h označimo njenu novu vrijednost, nakon korektivne sheme na dvije mreže, tada cijelu proceduru možemo napisati kao jedan korak sa

$$\begin{aligned} v_{novi}^h &= (G^h)^\nu v_{stari}^h + (I - (G^h)^\nu)(A^h)^{-1}f^h + \\ &\quad + I_{2h}^h (A^{2h})^{-1} I_h^{2h}[f^h - A^h((G^h)^\nu v_{stari}^h + (I - (G^h)^\nu)(A^h)^{-1}f^h)]. \end{aligned}$$

Ako ovaj glomazni izraz malo sredimo dobit ćemo konačni oblik

$$v^h_{novi} = v^h_{stari} + [I - (I - I^h_{2h}(A^{2h})^{-1} I^{2h}_h A^h)(G^h)^\nu](A^h)^{-1}(f^h - A^h v^h_{stari}).$$

Ako označimo sa

$$(M^h)^{-1} = [I - (I - I^h_{2h}(A^{2h})^{-1} I^{2h}_h A^h)(G^h)^\nu](A^h)^{-1},$$

tada imamo

$$v^h_{novi} = v^h_{stari} + (M^h)^{-1}(f^h - A^h v^h_{stari}),$$

čime smo pokazali da je jedno izvršavanje korektivne sheme na dvije mreže ekvivalentno jednom koraku jednostavnih iteracija, nad sustavom $A^h u^h = f^h$, i sa matricom prekondicioniranja jednakom M^h. Ovo je i razlog zašto multigrid metode možemo smatrati jednom vrstom prekondicioniranja, kod primjene jednostavnih iteracija. Kako je izraz za grešku $e^h = u^h - v^h$ jednostavnih iteracija jednak

$$e^h_{novi} = (I - (M^h)^{-1} A^h) e^h_{stari},$$

tada za operator korektivne sheme na dvije mreže, kojeg ćemo označiti sa $TG = I - (M^h)^{-1}A^h$, vrijedi

$$e^h_{novi} = [I - I^h_{2h}(A^{2h})^{-1} I^{2h}_h A^h](G^h)^\nu e^h_{stari} = TG e^h_{stari}. \tag{4.10}$$

Kao što smo već prije napomenuli, grešku možemo izraziti kao linearnu kombinaciju modova od A^h, zato nas to vodi do pitanja kako TG djeluje na modove od A^h. Međutim, TG se sastoji od G^h (koja je u našem primjeru jednaka G_{JOR}), A^h, $(A^{2h})^{-1}$, I^{2h}_h, i I^h_{2h}, a za svaki od tih operatora znamo kako djeluje na modove od A^h. Za trenutak, pretpostavimo da se radi o korektivnoj shemi na dvije mreže bez iteracija iterativne metode, odnosno da je $\nu = 0$. Koristeći sva spektralna svojstva, koja smo do sada dobili, za $1 \le k < \frac{n+1}{2}$ i $k' = n + 1 - k$ imamo

$$\begin{aligned}
TGq^{(k)^h} &= [I - I^h_{2h}(A^{2h})^{-1} I^{2h}_h A^h] q^{(k)^h} = \\
&= q^{(k)^h} - \frac{4}{h^2}\sin^2\left(\frac{k\pi}{2(n+1)}\right) I^h_{2h}(A^{2h})^{-1} I^{2h}_h q^{(k)^h} = \\
&= q^{(k)^h} - \frac{4}{h^2}\sin^2\left(\frac{k\pi}{2(n+1)}\right)\cos^2\left(\frac{k\pi}{2(n+1)}\right) I^h_{2h}(A^{2h})^{-1} q^{(k)^{2h}} = \\
&= q^{(k)^h} - \frac{4}{h^2}\sin^2\left(\frac{k\pi}{2(n+1)}\right)\cos^2\left(\frac{k\pi}{2(n+1)}\right) \cdot \\
&\quad \cdot h^2 \sin^{-2}\left(\frac{k\pi}{n+1}\right) I^h_{2h} q^{(k)^{2h}} = \\
&= q^{(k)^h} - I^h_{2h} q^{(k)^{2h}} = \\
&= q^{(k)^h} - \cos^2\left(\frac{k\pi}{2(n+1)}\right) q^{(k)^h} + \sin^2\left(\frac{k\pi}{2(n+1)}\right) q^{(k')^h} = \\
&= \sin^2\left(\frac{k\pi}{2(n+1)}\right)(q^{(k)^h} + q^{(k')^h}).
\end{aligned}$$

i, na sličan način je

$$TGq^{(k')^h} = q^{(k')^h} + \frac{4}{h^2}\sin^2\left(\frac{k'\pi}{2(n+1)}\right)\sin^2\left(\frac{k\pi}{2(n+1)}\right) \cdot$$

$$
\begin{aligned}
& \cdot h^2 \sin^{-2}\left(\frac{k\pi}{n+1}\right) I_{2h}^h q^{(k)^{2h}} = \\
= \; & q^{(k')^h} + 4\cos^2\left(\frac{k\pi}{2(n+1)}\right)\sin^2\left(\frac{k\pi}{2(n+1)}\right)\cdot \\
& \cdot \sin^{-2}\left(\frac{k\pi}{n+1}\right) I_{2h}^h q^{(k)^{2h}} = \\
= \; & q^{(k')^h} + I_{2h}^h q^{(k)^{2h}} = \\
= \; & q^{(k')^h} + \cos^2\left(\frac{k\pi}{2(n+1)}\right) q^{(k)^h} - \sin^2\left(\frac{k\pi}{2(n+1)}\right) q^{(k')^h} = \\
= \; & \cos^2\left(\frac{k\pi}{2(n+1)}\right)(q^{(k)^h} + q^{(k')^h}),
\end{aligned}
$$

jer je

$$
\sin\left(\frac{k'\pi}{2(n+1)}\right) = \sin\left(\frac{\pi}{2} - \frac{k\pi}{2(n+1)}\right) = \cos\left(\frac{k\pi}{2(n+1)}\right).
$$

Za $k = \frac{n+1}{2}$ je $I_h^{2h} q^{(k)^h} = 0$ pa je zbog toga $TGq^{(\frac{n+1}{2})^h} = q^{(\frac{n+1}{2})^h}$. S druge strane je

$$
\sin^2\left(\frac{k\pi}{2(n+1)}\right) = \cos^2\left(\frac{k\pi}{2(n+1)}\right) = \frac{1}{2},
$$

i $q^{(k')^h} = q^{(k)^h}$. Zato je za $k = \frac{k+1}{2}$

$$
TGq^{(k)^h} = \sin^2\left(\frac{k\pi}{2(n+1)}\right)(q^{(k)^h} + q^{(k')^h}) = \cos^2\left(\frac{k\pi}{2(n+1)}\right)(q^{(k)^h} + q^{(k')^h}).
$$

Dakle, za ovako definiranu korektivnu shemu na dvije mreže, operator TG je invarijantan na potprostor $\text{span}\{q^{(k)^h}, q^{(k')^h}\}$, jer vrijedi

$$
TGq^{(k)^h} = \sin^2\left(\frac{k\pi}{2(n+1)}\right)(q^{(k)^h} + q^{(k')^h}), \tag{4.11}
$$

$$
TGq^{(k')^h} = \cos^2\left(\frac{k\pi}{2(n+1)}\right)(q^{(k)^h} + q^{(k')^h}), \tag{4.12}
$$

$$
1 \le k \le \frac{n+1}{2}, \; k' = n+1-k.
$$

Odavde slijedi, da kada se TG primijeni na glatki ili oscilatorni mod, rezultat je kombinacija istog moda i njegovog komplementa. Ali važno je obratiti pažnju na amplitude rezultirajućih modova. Pretpostavimo da TG djeluje na vrlo gladak mod i na vrlo oscilatorni mod, sa $k \ll n+1$. Tada (4.11) i (4.12) postaju

$$
\begin{aligned}
TGq^{(k)^h} &= \mathcal{O}\left(\frac{k^2}{(n+1)^2}\right)(q^{(k)^h} + q^{(k')^h}), \\
TGq^{(k')^h} &= \left[1 - \mathcal{O}\left(\frac{k^2}{(n+1)^2}\right)\right](q^{(k)^h} + q^{(k')^h}), \quad 1 \le k \le \frac{n+1}{2}, k' = n+1-k.
\end{aligned}
$$

Kada TG djeluje na glatkim modovima, kao rezultat daje glatke i oscilatorne modove, sa vrlo malim amplitudama. Zbog toga je korektivna shema djelotvorna kod eliminacije glatkih komponenti greške, i to što je vektor glađi to bolje. Međutim, kada TG djeluje na

vrlo oscilatornim modovima, proizvodi glatke i oscilatorne modove sa $\mathcal{O}(1)$ amplitudom. Zbog toga, korektivna shema na dvije mreže, bez iterativne metode, ne može eliminirati oscilatorne modove.

Sada uvodimo iterativnu metodu u shemu. Budući da znamo njezina spektralna svojstva, očekujemo da će ona jako dobro izbalansirati djelovanje operatora TG bez iterativne metode. Uključit ćemo ν koraka iterativne metode, sa matricom G^h, uz pretpostavku da G^h ne miješa modove od A^h. Neka je $\lambda_k(G^h)$ svojstvena vrijednost od G^h, koja je pridružena k-tom modu $q^{(k)^h}$. Kombinirajući prethodnu analizu sa (4.10), djelovanje operatora TG sa iterativnom metodom dano je sa

$$TGq^{(k)^h} = \lambda_k(G^h)^\nu \sin^2\left(\frac{k\pi}{2(n+1)}\right)(q^{(k)^h} + q^{(k')^h}), \tag{4.13}$$

$$TGq^{(k')^h} = \lambda_{k'}(G^h)^\nu \cos^2\left(\frac{k\pi}{2(n+1)}\right)(q^{(k)^h} + q^{(k')^h}), \tag{4.14}$$

$$1 \le k \le \frac{n+1}{2},\ k' = n+1-k.$$

Znamo da izglađujuće svojstvo klasičnih iterativnih metoda ima najbolji efekt na oscilatornim modovima. To se reflektira kroz izraz $\lambda_{k'}(G^h)^\nu$, koji je mali. S druge strane, sama korektivna shema na dvije mreže, bez iterativne metode, eliminira glatke modove. To se reflektira kroz izraz $\sin^2(k\pi/((2(n+1)))$ koji je mali. Zbog toga su svi izrazi u (4.13) i (4.14) mali. Rezultat je kompletan proces koji uspješno eliminira i glatke i oscilatorne modove. Time smo dovršili takozvanu spektralnu sliku multigrid metode.

Postoji još jedan način na koji možemo promatrati korektivnu shemu na dvije mreže, pomoću kojeg ćemo dobiti uvid u algebarsku sliku multigrid metode. Sa spektralnom i algebarskom slikom moći ćemo dati solidno objašnjenje funkcioniranja multigrid metoda.

Daljnja razmatranja temelje se na varijacijskim svojstvima (4.8) i (4.9). Kao što smo vidjeli, slika interpolacije $\mathcal{R}(I_{2h}^h)$, i jezgra potpunog težinskog sumiranja $\mathcal{N}(I_h^{2h})$ leže na Ω^h i imaju redom dimenzije $\frac{n-1}{2}$ i $\frac{n+1}{2}$. Prema fundamentalnom teoremu linearne algebre, znamo da je

$$\mathcal{N}[(I_{2h}^h)^T] \perp \mathcal{R}(I_{2h}^h),$$

i

$$\Omega^h = \mathcal{R}(I_{2h}^h) \oplus \mathcal{N}[(I_{2h}^h)^T].$$

Prema drugom varijacijskom svojstvu, slijedi

$$\mathcal{N}(I_h^{2h}) \perp \mathcal{R}(I_{2h}^h),$$

i

$$\Omega^h = \mathcal{R}(I_{2h}^h) \oplus \mathcal{N}(I_h^{2h}).$$

što predstavlja važno svojstvo. Sada ćemo koristiti pojam A-ortogonalnosti kako bi na drugačiji način napisali gornju relaciju. Činjenica da je $\mathcal{N}(I_h^{2h}) \perp \mathcal{R}(I_{2h}^h)$ znači da je $\langle v^h, w^h\rangle = 0$ kad god je $v^h \in \mathcal{R}(I_{2h}^h)$ i $I_h^{2h} w^h = 0$. To je ekvivalentno uvjetu da je $\langle v^h, A^h w^h\rangle = 0$ za $v^h \in \mathcal{R}(I_{2h}^h)$ i $I_h^{2h} A^h w^h = 0$. Ovaj zadnji uvjet može se napisati kao

$$\mathcal{N}(I_h^{2h} A^h) \perp_{A^h} \mathcal{R}(I_{2h}^h),$$

odnosno, jezgra od $I_h^{2h}A^h$ je A^h-ortogonalna slici interpolacije. To možemo napraviti jer je u našem primjeru A^h pozitivno definitna matrica. Ovo svojstvo ortogonalnosti omogučava dekompoziciju prostora Ω^h

$$\Omega^h = \mathcal{R}(I_{2h}^h) \oplus_{A^h} \mathcal{N}(I_h^{2h}A^h).$$

To znači da ako je e^h vektor iz Ω^h, tada se on uvijek može izraziti kao

$$e^h = s^h + t^h,$$

gdje su $s^h \in \mathcal{R}(I_{2h}^h)$, i $t^h \in \mathcal{N}(I_h^{2h}A^h)$.

Sada bi bilo korisno bolje opisati vektore s^h i t^h. Budući da je s^h element iz $\mathcal{R}(I_{2h}^h)$, mora zadovoljavati $s^h = I_{2h}^h q^{2h}$, za neki vektor q^{2h} iz Ω^{2h}. Mi smo primijetili da postoji izglađujući efekt interpolacije i da vektori baze od $\mathcal{R}(I_{2h}^h)$ imaju glatki izgled. Zbog tog razloga, vektoru s^h pridružit ćemo u večini glatke komponente greške e^h. Također smo primijetili oscilatorni izgled vektora baze od $\mathcal{N}(I_h^{2h})$, a s druge strane bazu od $\mathcal{N}(I_h^{2h}A^h)$ čine neparni jedinični vektori. Zato ćemo vektoru t^h pridružiti uglavnom oscilatorne komponente greške.

Sada ćemo ponovo razmotriti djelovanje operatora korektivne sheme na dvije mreže, sa naglaskom na ove potprostore. Taj operator, bez iterativne metode ima oblik, kao što već znamo

$$TG = I - I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h.$$

Tada imamo da je

$$TGs^h = [I - I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h] I_{2h}^h q^{2h}.$$

Međutim, prema prvom varijacijskom svojstvu je $I_h^{2h}A^h I_{2h}^h = A^{2h}$. Zato imamo da je

$$TGs^h = 0.$$

To nam daje važan rezultat, koji kaže, da bilo koji vektor iz slike interpolacije, također leži i u jezgri operatora korektivne sheme na dvije mreže, odnosno

$$\mathcal{N}(TG) \supset \mathcal{R}(I_{2h}^h).$$

U drugu ruku imamo

$$TGt^h = [I - I_{2h}^h (A^{2h})^{-1} I_h^{2h} A^h] t^h.$$

Kako je $I_h^{2h}A^h t^h = 0$, zaključujemo daje

$$TGt^h = t^h.$$

To znači da je TG identiteta kada djeluje na $\mathcal{N}(I_h^{2h}A^h)$. To znači da je dimenzija slike od TG veća ili jednaka od dimenzije $\mathcal{N}(I_h^{2h}A^h)$. Prema toremu o rangu i defektu vrijedi

$$\begin{aligned} \dim \mathcal{N}(TG) &\leq n - \dim \mathcal{N}(I_h^{2h}A^h) = n - \dim \mathcal{N}(I_h^{2h}) = \\ &= n - \frac{n+1}{2} = \frac{n-1}{2} = \dim \mathcal{R}(I_{2h}^h). \end{aligned}$$

Dakle možemo zaključiti da je

$$\mathcal{N}(TG) = \mathcal{R}(I_{2h}^h).$$

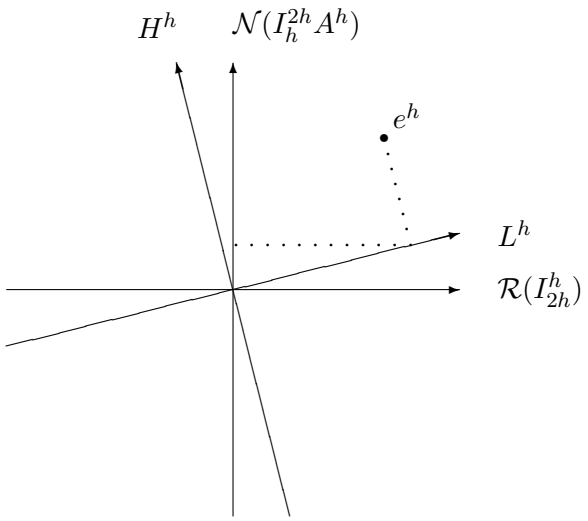
Ovom argumentacijom dobili smo uvid u djelovanje operatora korektivne sheme na dvije mreže, na dva ortogonalna potprostora od Ω^h, što čini njenu algebarsku sliku. Ako sada stavimo i spektralnu i algebarsku sliku zajedno, vidimo da prostor vektora na finoj mreži Ω^h možemo dekomponirati na dva nezavisna načina. Imamo spektralnu dekompoziciju

$$\begin{aligned}\Omega^h &= L^h \oplus H^h = \\ &= \text{span}\left\{q^{(k)^h} : \ 1 \le k < \frac{n+1}{2}\right\} \oplus \ \text{span}\left\{q^{(k)^h} : \ \frac{n+1}{2} \le k \le n\right\},\end{aligned}$$

i dekompoziciju s potprostorima

$$\Omega^h = \mathcal{R}(I_{2h}^h) \oplus_{A^h} \mathcal{N}(I_h^{2h} A^h).$$

Kako se u standardnom multigrid algoritmu izmjenjuju iteracije iterativne metode i korektivna shema na dvije mreže, to vidimo da se najprije eliminiraju oscilatorne komponente greške, a zatim komponente u potprostoru $\mathcal{R}(I_{2h}^h)$, koje se uvelike poklapaju sa glatkim modovima. Ovo je ukratko slika djelovanja standardne multigrid metode, u kojoj su dani argumenti o postupnoj eliminaciji svih komponenti greške, i o konvergenciji metode, a ilustrirana je na Slici 4.9.



Slika 4.9: Redukcija komponenata greške e^h u jednom izvođenju korektivne sheme na dvije mreže.

Do sad smo razmatrali samo svojstvene vrijednosti i vektore matrice A^h ili G^h. Pogledajmo još na kraju kako izgleda spektar same matrice TG, ali u najopćenitijem slučaju. Pretpostavimo da se korektivna shema na dvije mreže sastoji od ν_1 iteracija iterativne metode prije korekcije na gruboj mreži, i od ν_2 iteracija nakon korekcije. U tom slučaju, iz (4.13) i (4.14) slijedi

$$TGq^{(k)^h} = \lambda_k^{\nu_1+\nu_2} s_k q^{(k)^h} + \lambda_k^{\nu_1} \lambda_{k'}^{\nu_2} s_k q^{(k')^h}, \tag{4.15}$$

$$TGq^{(k')^h} = \lambda_{k'}^{\nu_1} \lambda_k^{\nu_2} c_k q^{(k)^h} + \lambda_{k'}^{\nu_1+\nu_2} c_k q^{(k')^h}, \tag{4.16}$$

$$1 \le k \le \frac{n+1}{2},\ k' = n+1-k,$$

pričemu su $\lambda_k = \lambda_k(G^h)$, $s_k = \sin^2\left(\frac{k\pi}{2(n+1)}\right)$ i $c_k = \cos^2\left(\frac{k\pi}{2(n+1)}\right)$. Ako definiramo matricu

$$Q^h = \sqrt{\frac{2}{n+1}} \cdot [q^{(1)^h}\ q^{(2)^h} \dots\ q^{(n)^h}],$$

čije stupce čine ortonormirani svojstveni vektori simetrične matrice A^h, onda se radi o ortogonalnoj matrici. Iz (4.15) i (4.16) slijedi da matrica

$$Z = (Q^h)^T\, TG\, Q^h$$

ima oblik

$$Z = \begin{bmatrix} \lambda_1^{\nu_1+\nu_2} s_1 & & & & & & \lambda_1^{\nu_2}\lambda_n^{\nu_1} c_1 \\ & \lambda_2^{\nu_1+\nu_2} s_2 & & & & \lambda_2^{\nu_2}\lambda_{n-1}^{\nu_1} c_2 & \\ & & \ddots & & \cdot & & \\ & & & \lambda_{\frac{n+1}{2}}^{\nu_1+\nu_2} & & & \\ & & \cdot & & \ddots & & \\ & \lambda_2^{\nu_1}\lambda_{n-1}^{\nu_2} s_2 & & & & \lambda_{n-1}^{\nu_1+\nu_2} c_2 & \\ \lambda_1^{\nu_1}\lambda_n^{\nu_2} s_1 & & & & & & \lambda_n^{\nu_1+\nu_2} c_1 \end{bmatrix},$$

i budući da je slična matrici TG, imaju isti spektar. Direktan račun daje

$$\det(Z - \lambda I) = \lambda_{\frac{n-1}{2}} \prod_{k=1}^{\frac{n+1}{2}} [\lambda - (\lambda_k^{\nu_1+\nu_2} s_k + \lambda_{k'}^{\nu_1+\nu_2} c_k)],$$

pa za svojstvene vrijednosti μ_k matrice Z, a ujedno i matrice TG, imamo

$$\mu_k = \begin{cases} \lambda_k^{\nu_1+\nu_2} s_k + \lambda_{k'}^{\nu_1+\nu_2} c_k, & 1 \le k \le \frac{n+1}{2}, \\ 0, & \frac{n+1}{2} < k \le n. \end{cases}$$

Budući da nas zapravo interesira spektralni radijus matrice TG, kao stope konvergencije korektivne sheme na dvije mreže, promatrat ćemo samo njene netrivijalne svojstvene vrijednosti, i to u tri različita slučaja, za $G^h = G_{JOR,\frac{2}{3}}$.

1. $\nu_1 = \nu_2 = 0$
   To je slučaj kada korektivnu shemu na dvije mreže koristimo bez iterativne metode. Tada je,
   $$\mu_k = s_k + c_k = 1, \qquad 1 \le k \le \frac{n+1}{2},$$
   odakle se vidi da, prema stopi konvergencije $\rho(TG) = 1$, konvergencija nije zagarantirana.

2. $\nu_1 \ge 1,\ \nu_2 = 0$
   To je slučaj kada se u korektivnoj mreži, prije korekcije na gruboj mreži izvede ν_1

iteracija iterativne metode. Tada je, zbog toga što je $\lambda_k < 1$ za JOR metodu za naš primjer, kao što smo već prije ustanovili,

$$\begin{aligned}
\mu_k &= \lambda_k^{\nu_1} s_k + \lambda_{k'}^{\nu_1} c_k \leq \lambda_k s_k + \lambda_{k'} c_k = \\
&= \left(1 - \frac{4}{3} s_k\right) s_k + \left(1 - \frac{4}{3} c_k\right) c_k = \\
&= 1 - \frac{4}{3}(s_k^2 + c_k^2) = 1 - \frac{4}{3}[(s_k + c_k)^2 - 2 s_k c_k] = \\
&= 1 - \frac{4}{3}\left[1 - \frac{1}{2}\sin^2\left(\frac{k\pi}{n+1}\right)\right] = -\frac{1}{3} + \frac{2}{3}\sin^2\left(\frac{k\pi}{n+1}\right) \leq \\
&\leq -\frac{1}{3} + \frac{2}{3} = \frac{1}{3}, \quad 1 \leq k \leq \frac{n+1}{2},
\end{aligned}$$

odnosno $\rho(TG) \leq \frac{1}{3}$, što znači da se sa svakim izvršavanjem ovakve korektivne metode na dvije mreže greška reducira za faktor koji je barem 3, i koji ne ovisi o h.

3. $\nu_1 \geq 1,\ \nu_2 \geq 1$
   Ovo je najopćenitiji slučaj korektivne sheme na dvije mreže, kada se prije i poslije korekcije na gruboj mreži izvršavaju barem po jedna iteracija iterativne metode. Vrijedi

$$\begin{aligned}
\mu_k &= \lambda_k^{\nu_1+\nu^2} s_k + \lambda_{k'}^{\nu_1+\nu^2} c_k \leq \lambda_k^2 s_k + \lambda_{k'}^2 c_k = \\
&= \left(1 - \frac{4}{3} s_k\right)^2 s_k + \left(1 - \frac{4}{3} c_k\right)^2 c_k = \\
&= 1 - \frac{8}{3}(s_k^2 + c_k^2) + \frac{16}{9}(s_k^3 + c_k^3) = \\
&= 1 - \frac{8}{3}(s_k^2 + c_k^2) + \frac{16}{9}(s_k + c_k)(s_k^2 - s_k c_k + c_k^2) = \\
&= 1 - \frac{8}{9}(s_k^2 + c_k^2) - \frac{16}{9} s_k c_k = 1 - \frac{8}{9}(s_k + c_k)^2 = \\
&= 1 - \frac{8}{9} = \frac{1}{9}, \quad 1 \leq k \leq \frac{n+1}{2},
\end{aligned}$$

   odakle je $\rho(TG) \leq \frac{1}{9}$. Ovaj slučaj je najbolji, jer garantira da će se svakom primjenom korektivne sheme na dvije mreže sa $\nu_1, \nu_2 \geq 1$ greška reducirati za faktor koji je barem 9, i koji ne ovisi o h.

Iz prethodne analize vidimo kako se klasična iterativna metoda i korektivna shema na dvije mreže vrlo dobro upotpunjavaju. Kada se ta dva algoritma primjene zajedno, rezultat je vrlo djelotvoran algoritam.

Trebamo još napomenuti, da se cijela ova analiza bavila samo korektivnom shemom na dvije mreže. V-ciklus koristi tu shemu na svim nivoima, osiguravajući da iteracije iterativne metode budu usmjerene na oscilatorne modove tekuće mreže. Korektivna shema na dvije mreže bez iterativne metode, brine se o glatkim komponentema greške na tekućoj greški, tako da su, ponovo, na svakom nivou postupno eliminirane sve komponente mreže.

Djelotvornost korektivne sheme na dvije mreže se dalje pojačava FMG metodom. Ta metoda koristi ugniježđene V-cikluse, kako bi dobili točne početne iteracije na grubljoj

mreži, prije nego što se djeluje sa iterativnom metodom na finoj mreži. To osigurava, da se odgovarajući problem na gruboj mreži riješi najtočnije što se može, prije nego što se pristupi skupljim iteracijama iterativne metode na finoj mreži. Međutim, čak i u ovim kompliciranijim algoritmima, glavnu snagu daje kombinacija iterativne metode, i korekcije na grubljoj mreži.

Osim klasičnih iterativnih metoda, koje smo koristili kao jedan dio multigrid metode, mogu se koristiti i druge iterativne metode kao npr. CG, GMRES, QMR, ili BICGSTAB, za ubrzavanje konvergencije. Isto tako, multigrid možemo upotrijebiti i kao postupak prekondicioniranja iterativnih metoda koje koriste aproksimacije iz Krylovljevih potprostora. Za rješavanje jednadžbe $Mz = r$, sa matricom prekondicioniranja M, dobivenom iz multigrid V-ciklusa, treba jednostavno izvesti jedan V-cikus nad sustavom čija je matrica jednaka matrici polaznog sustava, desna strana jednaka r, a početna iteracija z_0 jednaka nuli. Tada nakon jednog V-ciklusa imamo $z_1 = 0 + M^{-1}(r - M \cdot 0) = M^{-1}r$, pa je z_1 traženo rješenje. Odavde se vidi da primjena multigrid metoda, kao i njen razvoj, mogu ići u raznim smjerovima.

Glava 5

Metode dekompozicije domene

Simulacijski se problemi često sastoje od kompliciranih struktura, kao na primjer trupovi aviona ili automobila. U tom slučaju pojavljuje se potreba za novim metodama koje se mogu uhvatiti u koštac sa nepravilnim domenama u dvodimenzionalnom i trodimenzionalnom prostoru, na kojima je definiran problem, i sa veličinom problema. Zbog ograničenosti računalne memorije i vremena, ovakvi problemi se često ne mogu riješiti od jednom, već se problem razbija na više djelova koji se posebno rješavaju. Ako se više tih dijelova može riješiti nezavisno jedan od drugoga, i zatim, ako se sva ta rješenja mogu nekako slijepiti u rješenje cijelog problema, tada se rješavanje tog problema može izvoditi na paralelnim računalima. Najbrže metode koje smo do sada spominjali, multigrid metode, funkcioniraju najbolje na prilično regularnim problemima kao što je rješavanje Poissonove jednadžbe, definiranim na pravilnim domenama, poput pravokutnika. S druge strane domena problema kojeg želimo riješiti može biti nepravilnog oblika, kao kod modela krila aviona. Također, možemo imati kompliciranije jednadžbe od Poissonove, ili možemo imati različite jednadžbe na različitim dijelovima domene. Ponekad je problem vrlo velik, bez obzira na regularnost, i ne može stati u memoriju računala. U svim tim slučajevima, traži se mogućnost razbijanja domene na manje dijelove na kojima problem nije kompliciran za rješavanje, i na kojima se može primijeniti jednostavnija metoda, recimo, razbijanje na pravokutnike na kojima se onda primjenjuje multigrid metoda. Rješavanje tih potproblema može se izvoditi jedno po jedno, ili paralelno. Ako se rješenja potproblema mogu iskombinirati na pametan način, pri čemu bi se dobilo rješenje kompletnog problema, tada ćemo dobiti bržu i paralelizabilnu metodu za dobivanje rješenja, od primjene standardne iterativne metode na cijeli, veliki problem. Vidjet ćemo da je ovaj pristup ponovo ekvivalentan prekondicioniranju sustava, samo što ćemo u ovom slučaju imati rješavanje manjih problema na poddomenama. Općenito postoji mnogo načina pomoću kojih možemo razbiti domenu na poddomene, mnogo načina na koje možemo riješiti potprobleme, i mnogo načina na koji možemo dobiti konačno rješenje kompletnog problema. Teorija metoda dekompozicije domene ne nudi općeniti recept kako odabrati sve te načine, već daje razumne mogućnosti koje treba isprobati.

Metode dekompozicije domena dijele se u dvije klase, u jednoj se koriste *neprekla-pajuće poddomene*, a u drugoj *preklapajuće poddomene*. U nastavku ćemo razmotriti po nekoliko predstavnika metoda dekompozicije domene, koje su raspoređene u obje klase metoda.

5.1 Metode sa nepreklapajućim poddomenama

Ovakve metode nazivaju se još *substrukturalne* ili *metode sa Schurovim komplementom*. Ovakve metode već se dugo koriste za razbijanje velikih problema na manje, koji bi stali u memoriju računala.

Neka je $\mathcal{L}$ diferencijalni operator definiran na domeni Ω, i pretpostavimo da želimo riješiti rubni problem

$$\begin{aligned} \mathcal{L}u &= f, \qquad \text{na } \Omega, \\ u &= u_\Gamma \qquad \text{na } \Gamma = \partial\Omega. \end{aligned} \tag{5.1}$$

Domena Ω je otvoreni skup u ravnini ili 3-dimenzionalnom prostoru, a $\partial\Omega$ označava rub od Ω. Ovdje smo uzeli Dirichletove rubne uvjete, ali mogu se isto tako staviti Neumannovi ili Robinovi rubni uvjeti.

Prepostavimo da je naš problem bio diskretiziran centralnim diferencijama, i da je na domeni problema Ω definirana mreža. Prepostavimo da je domena podijeljena na s poddomena Ω_i, $i = 1, \ldots, s$, odnosno da vrijedi

$$\Omega = \bigcup_{i=1}^{s} \Omega_i,$$

pri čemu se poddomene ne preklapaju, što znači da susjedne domene dijele samo rub. Nazovimo rub izmedju Ω_i i Ω_j, $i \neq j$ granicom Γ_{ij} između i-te i j-te poddomene. Te se granice mogu ili ne moraju poklapati sa bridovima mreže, ali općenito možemo pretpostaviti da svaka granica Γ_{ij} sadrži neke točke mreže. Čvorove mreže možemo sada poredati tako da grupiramo, redom, čvorove koji se nalaze u unutrašnjosti poddomena, najprije poddomene Ω_1, zatim Ω_2, sve do Ω_s. Na kraju poredamo čvorove granica. Kao rezultat matrica pridružena tom problemu imat će oblik

$$\begin{bmatrix} B_1 & & & & E_1 \\ & B_2 & & & E_2 \\ & & \ddots & & \vdots \\ & & & B_s & E_s \\ F_1 & F_2 & \cdots & F_s & C \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_s \\ y \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_s \\ g \end{bmatrix}, \tag{5.2}$$

gdje x_i predstavlja podvektor nepoznanica koje se odnose na točke u unutrašnjosti poddomene Ω_i, a y predstavlja vektor svih nepoznanica koje se odnose na točke koje pripadaju graničnom području. Primijetimo da su blokovi na pozicijama (i, j), za $i, j = 1, \ldots, s$, $i \neq j$ jednaki nuli, zato što niti jedna točka iz unutrašnjosti jedne domene nije direktan susjed niti jedne točke iz unutrašnjosti bilo koje druge domene. Jedino može biti susjedna nekoj točki iz graničnog područja.

5.1.1 Blok-Gaussove eliminacije i Schurov komplement

Sustav (5.2) možemo napisati u jednostavnijoj formi, koju će namo koristiti u daljnoj analizi.

$$A \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \qquad \text{sa} \quad A = \begin{bmatrix} B & E \\ F & C \end{bmatrix}. \tag{5.3}$$

Pretpostavimo da je B regularna matrica, što je za očekivati, jer su B_i regularne, kao matrice lokalnih problema na Ω_i. Tada iz prve jednadžbe možemo izraziti x kao

$$x = B^{-1}(f - Ey). \tag{5.4}$$

Uvrštavajući to u drugu jednadžbu, dobivamo *reducirani sustav* na nepoznanicama granice

$$(C - FB^{-1}E)y = g - FB^{-1}f. \tag{5.5}$$

Matrica

$$S = C - FB^{-1}E \tag{5.6}$$

zove se matrica *Schurovog komplementa* sustava, po nepoznanici y. Ako bi se ta matrica mogla izračunati, i ako se može riješiti sustav (5.5), dobili bi vrijednosti za sve varijable graničnog područja. Ostale nepoznanice mogu se tada izračunati preko (5.4). Zbog posebne strukture matrice B, koja je zapravo blok-dijagonalna, rješavanje linearnog sustava s njom je ekvivalentno rješavanju s nezavisnih i manjih sustava. Ovdje do izražaja može doći paralelno računanje.

Metoda za rješavanje sustava bazirana na ovom pristupu sasatoji se od četiri koraka:

- Izračunaj desnu stranu reduciranog sustava (5.5).
- Izračunaj matricu Schurovog komplementa (5.6).
- Riješi reducirani sustav (5.5).
- Pomoću (5.4) izračunaj ostale nepoznanice.

Jedno rješavanje sustava sa matricom B može se uštedjeti preformuliranjem algoritma u pogodniji oblik. Definirajmo

$$E' = B^{-1}E, \quad \text{i} \quad f' = B^{-1}f.$$

Matrica E' i vektor f' su potrebni u prvom i drugom koraku algoritma. Četvrti korak, tada možemo napisati kao

$$x = B^{-1}f - B^{-1}Ey = f' - E'y,$$

što nam daje sljedeći algoritam.

Algoritam 5.1.1. Algoritam blok-Gaussovih eliminacija

Riješi $BE' = E$, *i* $Bf' = f$ *po* E' *i* f'.

Izračunaj $g' = g - Ff'$.

Izračunaj $S = C - FE'$.

Riješi $Sy = g'$.

Izračunaj $x = f' - E'y$.

U konkretnoj implementaciji, rješavanje sustava sa matricom B svodi se na rješavanje s sustava $B_iE'_i = E_i$ i $B_if'_i = f_i$. Treba još napomenuti, da su mnogi stupci od E_i jednaki nuli, i to oni koji se odnose na one granice koje ne ograničavaju poddomenu i.

Sada ćemo pogledati koja je veza između Schurovog komplementa i Gaussovih eliminacija. Započnimo sa blok-LU faktorizacijom matrice A,

$$A = \begin{bmatrix} B & E \\ F & C \end{bmatrix} = \begin{bmatrix} I & 0 \\ FB^{-1} & I \end{bmatrix} \begin{bmatrix} B & E \\ 0 & S \end{bmatrix},$$

što se lako provjeri. Na Schurov komplement možemo onda gledati kao na $(2,2)$-blok U faktora LU faktorizacije matrice A. Iz gornje jednakosti slijedi da ako je A regularna, tada su to i njezini faktori, odakle slijedi da je i S regularna. Uzimajući inverz od A dobivamo

$$\begin{aligned} \begin{bmatrix} B & E \\ F & C \end{bmatrix}^{-1} &= \begin{bmatrix} B^{-1} & -B^{-1}ES^{-1} \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -FB^{-1} & I \end{bmatrix} = \\ &= \begin{bmatrix} B^{-1} + B^{-1}ES^{-1}FB^{-1} & -B^{-1}ES^{-1} \\ -S^{-1}FB^{-1} & S^{-1} \end{bmatrix}. \end{aligned} \tag{5.7}$$

Primijetimo da je S^{-1} $(2,2)$-blok blok-inverza od A. Posebno, ako je originalna matrica A simetrična pozitivno definitna matrica tada je to i A^{-1} (svojstvene vrijednosti od A^{-1} su inverzi svojstvenih vrijednosti od A, pa su pozitivne). Kao posljedica toga je i S simetrična i pozitivno definitna, jer je glavna minora matrice A^{-1} (svojstvene vrijednosti od S^{-1} su između najmanje i najveće svojstvene vrijednosti od A^{-1}, koje su pozitivne).

Ova svojstva, koja smo jednostavno provjerili, sakupljena su u sljedećem teoremu.

Teorem 5.1.2 ([32]). *Neka je A regularna matrica, particionirana kao u (5.3), takva da je podmatrica B regularna, i neka je I_y operator restrikcije na varijable granice, odnosno operator definiran sa*

$$I_y \begin{bmatrix} x \\ y \end{bmatrix} = y.$$

Tada vrijede sljedeća svojstva.

(i) Matrica Schurovog komplementa S je regularna.

(ii) Ako je A simetrična pozitivno definitna matrica, tada je to i S.

(iii) Za bilo koji y je $S^{-1}y = I_y A^{-1} \begin{bmatrix} 0 \\ y \end{bmatrix}$.

Prvo svojstvo pokazuje da će metoda koja koristi blok-Gaussove eliminacije biti izvediva, budući da je S regularna. Poslijedica drugog svojstva je, da kad je A pozitivno definitna, algoritma poput konjugiranih gradijenata može se primijeniti kod rješnavanja reduciranog sustava (5.5).

Dakle, rješavanje početnog sustava svodi se na množenje vektora sa matricom A^{-1}, koje se opet satoji od množenja podvektora blok-elementima od A^{-1}. U tom množenju, osim standardnog množenja matrica–vektor, imamo i rješavanje sustava sa matricama B i S. Rješavanje sustava sa matricom B je jeftino jer se svodi na rješavanj sustava sa B_i, kojima su domene tako određene da se lako rješavaju npr. multigridom. Ostaje još problem rješavanje sustava s matricom S, odnosno množenje sa S^{-1}. Budući da je u graničnom području mnogo manje točaka mreže nego u unutrašnjosti poddomena, S ima mnogo manju dimenziju od bilo kojeg B_i. Ako je matrica A simetrična pozitivno definitna, pa je onda to i S, S možemo eksplicitno izračunati preko (5.6), što uključuje

rješavanje sustava s matricom B, te zatim izvesti faktorizaciju Choleskog, i pomoću nje riješiti sustav. Međutim, to je prilično skuplje od samog množenja vektora sa matricom S. Zbog tog nam se metode aproksimacije iz Krylovljevih potprostora, kao npr konjugirani gradijenti, čine vrlo primjenljivim u ovom slučaju. Može se pokazati da je matrica S puno bolje uvjetovana od matrice A, $\mathcal{O}(h^{-1})$ umjesto $\mathcal{O}(h^{-2})$, što znači da će i konvergencija biti bolja. (Za reference vidi [7].)

5.2 Metode sa preklapajućim poddomenama

Sada ćemo promatrati metode koje dozvoljavaju da se poddomene preklapaju, kao što su aditivna i multiplikativna Schwarzova metoda. Ove metode potekle su od alternirajuće procedure za rješavanje diferencijalnih jednadžbi, koju je Schwarz još 1870. godine opisao. Ona se sastoji od tri djela: alterniranja između dviju preklapajućih domena, rješavanja Dirichletovog rubnog problema na jednoj od domena u svakoj pojedinačnoj iteraciji, i korištenja rubnih uvjeta, koji su dobiveni iz aproksimacije rješenja, ostvarenog u zadnjoj iteraciji na drugoj domeni.

Pogledajmo spomenutu *alternirajuću Schwarzovu metodu* malo detaljnjije. Pretpostavimo da polazimo od problema (5.1), i da je domena Ω podijeljena na s poddomena Ω_i. Pretpostavimo da svaki par susjednih poddomena ima neprazni presjek. Rub poddomene Ω_i koji je uključen u poddomenu j označen je sa $\Gamma_{i,j}$, budući da se svaka poddomena prostire izvan svojih originalnih granica i prodire u područje susjednih poddomena. Označimo sa Γ_i rub domene Ω_i, koji se sastoji od djela originalnog ruba ukupne domene $\partial\Omega_i \cap \partial\Omega$, označenog sa $\Gamma_{i,0}$, i svih $\Gamma_{i,j}$. Označimo sa $u_{j,i}$ restrikciju rješenja u na rub $\Gamma_{j,i}$. Alternirajuća Schwarzova metoda može se tada opisati na sljedeći način.

Algoritam 5.2.1. Alternirajuća Schwarzova metoda

Izaberi početnu aproksimaciju rješenja u.

Dok nismo postigli željenu konvergenciju, radi:

Za $i = 1, \ldots, s$

Riješi $\mathcal{L}u = f$ na Ω_i, sa $u = u_{ij}$ na $\Gamma_{i,j}$.
Korigiraj vrijednost u na $\Gamma_{j,i}$ za $\forall j$, sa novoizračunatim vrijednostima na Ω_i.

Dakle, algoritam prolazi kroz s poddomena i na svakoj od njih rješava originalni problem, koristeći rubne uvjete koji su dobiveni od aproksimacija rješenja iz prethodnih iteracija. Budući da će se potproblemi najvjerojatnije rješavati iterativnom metodom, za početnu aproksimaciju danog potproblema uzima se zadnja dobivena aproksimacija.

U ovom slučaju varijable ćemo najprije poredati u grupe, tako da se u jednoj grupi nalaze sve varijable koje korespondiraju točkama jedne poddomene izvan preklapanja. Nakon toga ćemo na kraju poredati varijable koje odgovaraju točkama u preklopljenim područjima. Matrica, dobivena diskretizacijom, i koja je pridružena ovakvoj particiji domene, imat će ponovo oblik kao u relaciji (5.2). Treba još napomenuti da će od sada

pa nadalje Ω predstavljati vektorski prostor, kod kojeg vektori sadrže komponente koje odgovaraju svim točkama mreže, dok će Ω_i predstavljati vekorski prostor vektora čije komponente odgovaraju točkama mreže unutar poddomene Ω_i.

5.2.1 Multiplikativna Schwarzova metoda

Za definiciju multiplikativne Schwarzove metode potrebna nam je definicija projektora na poddomene. Neka je $\mathcal{S}_i$ skup indeksa

$$\mathcal{S}_i = \{j_1, j_2, \ldots, j_{n_i}\},$$

gdje indeksi j_l označavaju n_i čvorova mreže unutrašnjosti poddomene Ω_i. Svi skupovi $\mathcal{S}_i$ zajedno čine kolekciju skupova indeksa takvih da je

$$\bigcup_{i=1,\ldots,n} \mathcal{S}_i = \{1, \ldots, n\},$$

pri čemu $\mathcal{S}_i$ nisu nužno disjunktni. Neka je I_i operator restrikcije sa Ω na Ω_i. Prema definiciji, $I_i v$ se nalazi u Ω_i, i zadržava samo one komponente proizvoljnog vektora v, koje su u Ω_i. Operator je reprezentiran $n_i \times n$ matricom koja se sastoji od jedinica i nula, čiji raspored ovisi o poretku čvorova. Matrica I_i je $n_i \times n$ matrica, čiji reci su dobiveni iz transponiranih jediničnih vektora ξ_j, za $j \in \mathcal{S}_i$. Operator I_i^T je operator interpolacije, koja uzima vektor iz Ω_i i ekspandira ga u ekvivalentan vektor u Ω. Matrica

$$A_i = I_i A I_i^T$$

dimenzije $n_i \times n_i$ definira restrikciju od A na Ω_i. Sa ovom notacijom multiplikativna Schwarzova metoda može se opisati na sljedeći način

Algoritam 5.2.2. MULTIPLIKATIVNA SCHWARZOVA METODA, MATRIČNI OBLIK

Izaberi početnu aproksimaciju rješenja u.

Dok nismo postigli željenu konvergenciju, radi:

Za $i = 1, \ldots, s$

$$u := u + I_i^T A_i^{-1} I_i (b - Au),$$

Počevši od početne, globalne aproksimacije u_0, čiji je vektor greške definiran sa $e_0 = u - u_0$, svaka poditeracija daje grešku, koja zadovoljava relaciju

$$e_i = e_{i-1} - I_i^T A_i^{-1} I_i A e_{i-1},$$

za $i = 1, \ldots, s$. Primijetimo samo da varijable u_i i e_i označavaju vektore definirane na cijeloj domeni Ω, a ne na Ω_i, samo što sudjeluju u i-toj poditeraciji. Kao rezultat, imamo

$$e_i = (I - P_i) e_{i-1},$$

gdje je

$$P_i = I_i^T A_i^{-1} I_i A. \tag{5.8}$$

Primijetimo da je tako definiran operator P_i projektor, jer vrijedi

$$P_i^2 = (I_i^T A_i^{-1} I_i A)^2 = I_i^T A_i^{-1} (I_i A I_i^T) A_i^{-1} I_i A = I_i^T A_i^{-1} I_i A = P_i.$$

Prema tome, jedan prolaz kroz unutarnju petlju algoritma multiplikativne Schwarzove metode zadovoljava relaciju

$$e_s = (I - P_s)(I - P_{s-1}) \cdots (I - P_1) e_0. \tag{5.9}$$

U nastavku koristit ćemo oznaku

$$Q_s = (I - P_s)(I - P_{s-1}) \cdots (I - P_1). \tag{5.10}$$

Zbog ekvivalencije multiplikativne Schwarzove metode i blok-Gauss–Seidelove metode, moguće je jedan prolaz kroz unutarnju petlju multiplikativne Schwarzove metode napisati u obliku globalne iteracije fiksne točke $u_{novi} = G u_{stari} + f$. Naime, multiplikativna Schwarzova metoda korigira aproksimaciju rješenja po poddomenama na isti način kao što to čini blok Gauss–Seidel po blok koordinatama. Prisjetimo se da je u slučaju primjene jednostavnih iteracija na prekondicionirani sustav $M^{-1}Au = M^{-1}b$ matrica G jednaka $G = I - M^{-1}A$, i $e_i = G e_{i-1}$. Iz (5.9) slijedi $e_s = Q_s e_0$. Ako sada sa u_{novi} označimo novu vrijednost aproksimacije rješenja nakon jednog prolaska kroz unutarnju petlju multiplikativne Schwarzove metode, tada imamo

$$u_{novi} = Q_s u_{stari} + (I - Q_s) A^{-1} b,$$

pa je zbog toga

$$G = Q_s, \qquad f = (I - Q_s) A^{-1} b.$$

Prema tome je prekondicionirana matrica jednaka $M^{-1}A = I - Q_s$. Ovaj rezultat sadržan je u sljedećem teoremu.

Teorem 5.2.3 ([32]). *Multiplikativna Schwarzova metoda je ekvivalentna iteraciji fiksne točke za prekondicionirani sustav*

$$M^{-1}Au = M^{-1}b,$$

u kojem je

$$M^{-1}A = I - Q_s, \tag{5.11}$$
$$M^{-1}b = (I - Q_s) A^{-1} b. \tag{5.12}$$

Desna strana prekondicioniranog sustava iz prethodnog teorema nije exsplicitno poznata jer sadrži egzaktno rješenje u svom izrazu. Ipak može se naći procedura koja će ju izračunati. Drugim riječima, moguće je upotrebljavati M^{-1} bez računanja A^{-1}. Primijetimo da je $M^{-1} = (I - Q_s)A^{-1}$. Kao što sljedeća lema pokazuje M^{-1} i $M^{-1}A$ se mogu rekurzivno izračunati i iskoristiti za prekondicioniranje bilo koje iterativne metode za rješavanje sustava.

Lema 5.2.4 ([32]). *Definirajmo matrice*

$$\begin{aligned} Z_i &= I - Q_i && (5.13)\\ M_i &= Z_i A^{-1} && (5.14)\\ T_i &= P_i A^{-1} = I_i^T A_i^{-1} I_i && (5.15) \end{aligned}$$

za $i = 1, \ldots, s$. *Tada je* $M^{-1} = M_s$, $M^{-1}A = Z_s$, *a matrice* Z_i *i* M_i *zadovoljavaju rekurzije*

$$\begin{aligned} Z_1 &= P_1, \\ Z_i &= Z_{i-1} + P_i(I - Z_{i-1}), \quad i = 2, \ldots, s && (5.16) \end{aligned}$$

i

$$\begin{aligned} M_1 &= T_1, \\ M_i &= M_{i-1} + T_i(I - AM_{i-1}), \quad i = 2, \ldots, s. && (5.17) \end{aligned}$$

Dokaz: Iz definicija (5.13) i (5.14) jasno je da je da je $M_s = M^{-1}$, $Z_s = M^{-1}A$, i da je $Z_1 = P_1$, $M_1 = T_1$. Za slučajeve, kada je $i > 1$, iz definicija za Q_i i Q_{i-1} slijedi

$$Z_i = I - (I - P_i)(I - Z_{i-1}) = P_i + Z_{i-1} - P_i Z_{i-1}, \tag{5.18}$$

što daje rekurziju (5.16). Množeći (5.18) sa A^{-1} slijeva dobivamo

$$M_i = T_i + M_{i-1} - P_i M_{i-1}.$$

Ako u gornju relaciju umjesto P_i ubacimo $T_i A$, dobivamo i rekurziju (5.17). □

Primijetimo da direktno iz (5.16) slijedi važna relacija

$$Z_i = \sum_{j=1}^{i} P_j Q_{j-1}. \tag{5.19}$$

S druge strane, pogledajmo kako onda izgleda matrica M^{-1}. Imamo

$$\begin{aligned} M^{-1} &= (I - Q_s)A^{-1} = [I - (I - P_s)(I - P_{s-1}) \cdots (I - P_1)]A^{-1} = \\ &= \sum_i T_i - \sum_{i>j} P_i T_j + \sum_{i>j>k} P_i P_j T_k + \cdots + (-1)^{s-1} P_s \cdots P_2 T_1 = \\ &= \sum_i T_i - \sum_{i>j} T_i A T_j + \sum_{i>j>k} T_i A T_j A T_k + \cdots \\ &\quad + (-1)^{s-1} T_s A T_{s-1} \cdots T_2 A T_1, \end{aligned}$$

odakle se vidi opravdanje za naziv "multiplikativna" metoda.

Ako rekurziju (5.17) pomnožimo s desna sa vektorom v, i ako vektor $M_i v$ označimo sa z_i, tada dobivamo sljedeću rekurziju

$$z_i = z_{i-1} + T_i(v - Az_{i-1}).$$

Kako je $z_s = M_s v = M^{-1}v$, vektor $M^{-1}v$ za proizvoljni vektor v može se izračunati sljedećom procedurom.

Djelovanje matrice prekondicioniranja multiplikativne Schwarzove metode na vektor

Ulaz: v, izlaz: $z = M^{-1}v$.

- $z = T_1 v$.
- Za $i = 2, \ldots, s$
  - $z := z + T_i(v - Az)$.

Sličnom argumentacijom, možemo naći proceduru koja računa vektore oblika $z = M^{-1}Av$. U tom slučaju, procedura je oblika

Djelovanje prekondicionirane matrice multiplikativne Schwarzove metode na vektor
Ulaz: v, izlaz: $z = M^{-1}Av$.

- $z = P_1 v$.
- Za $i = 2, \ldots, s$
  - $z := z + P_i(v - z)$.

Na kraju možemo reći da je multiplikativna Schwarzova metoda ekvivalentna rješavanju prekondicioniranog linearnog sustava

$$(I - Q_s)u = g, \tag{5.20}$$

gdje se operacije $z = (I - Q_s)v$ i $g = M^{-1}b$ mogu izračunati iz gornjih algoritama. Te procedure se mogu dalje koristiti unutar neke od iterativnih metoda, kao na primjer GMRES.

5.2.2 Aditivna Schwarzova metoda

Aditivna Schwarzova metoda je slična blok-Jacobijevoj metodi, i sastoji se od korigiranja svih novih komponenti pomoću istog reziduala. Zbog toga se i razlikuje od multiplikativne metode, jer komponente svake poddomene se ne mijenjaju dok cijeli ciklus korekcija kroz sve poddomene nije dovršen. Osnovna aditivna Schwarzova metoda stoga ima oblik:

Algoritam 5.2.5. ADITIVNA SCHWARZOVA METODA, MATRIČNI OBLIK

Izaberi početnu aproksimaciju rješenja u.

Dok nismo postigli željenu konvergenciju, radi:

Za $i = 1, \ldots, s$

Izračunaj $\delta_i = I_i^T A_i^{-1} I_i (b - Au_{stari})$.

$u := u + \sum_{i=1}^{s} \delta_i$

Nova aproksimacija rješenja u_{novi}, dobivena nakon što je primijenjen cijeli ciklus od s iteracija gornje petlje na u_{stari}, jednaka je

$$u_{novi} = u_{stari} + \sum_{i=1}^{s} I_i^T A_i^{-1} I_i (b - Au_{stari}).$$

Svako izvođenje ove petlje redefinira sve komponente nove aproksimacije, i ne postoji nikakava ovisnost između potproblema vezanih uz poddomene. Ako ovu relaciju rastavimo po blok-komponentama, dobit ćemo

$$u_{i,novi} = u_{i,stari} + A_i^{-1} r_{i,stari} = A_i^{-1} \left(- \sum_{j \neq i} A_{ij} u_{j,stari} + b_i \right),$$

što je upravo korak blok-Jacobijeve metode primijenjen na globalni sustav.

Prekondicioniranu matricu za aditivnu Schwarzovu metodu je vrlo jednostavno dobiti. Koristeći oznake i definicije, iz multiplikativne Schwarzove metode, nova aproksimacija zadovoljava relaciju

$$u_{novi} = u_{stari} + \sum_{i=1}^{s} T_i (b - A u_{stari}) = \left(I - \sum_{i=1}^{s} P_i \right) u_{stari} + \sum_{i=1}^{s} T_i b.$$

Prema tome, ovakve iteracije korespondiraju iteracijama fiksne točke $u_{novi} = G u_{stari} + f$, gdje je

$$G = I - \sum_{i=1}^{s} P_i, \qquad f = \sum_{i=1}^{s} T_i b.$$

Uz pomoć relacije $G = I - M^{-1}A$, vidimo da je

$$M^{-1} A = \sum_{i=1}^{s} P_i,$$

i

$$M^{-1} = \sum_{i=1}^{s} P_i A^{-1} = \sum_{i=1}^{s} T_i,$$

odakle se vidi razlog za ime metode: "aditivna" metoda. Sada je jasno kako izgleda procedura primjene matrice prekondicioniranja M^{-1} na vektor.

Djelovanje matrice prekondicioniranja aditivne Schwarzove metode na vektor
Ulaz: v, izlaz: $z = M^{-1} v$.

- Za $i = 2, \ldots, s$
  - Izračunaj $z_i = T_i v$.
- $z = z_1 + z_2 + \cdots + z_s$.

Primijetimo da se ova petlja može paralelizirati. Procedura za računanje $M^{-1}Av$ je identična gornjoj procedura, samo što se T_i treba zamijeniti sa P_i.

5.2.3 Konvergencija metoda sa preklapajućim domenama

U analizi konvergencije, pretpostavit ćemo da je A simetrična pozitivno definitna matrica. Projektori P_i definirani sa (5.8) imaju važnu ulogu u teoriji konvergencije za aditivnu i multiplikativnu Schwarzovu metodu. Prvu važnu stvar koju moramo primijetiti je ta da je P_i hermitski operator, obzirom na A-skalarni produkt, što je uz svojstvo $P_i^2 = P_i$ dovoljan uvjet da za P_i možemo reći da je A-ortogonalan projektor. Vrijedi

$$\langle P_i v, w \rangle_A = \langle A I_i^T A_i^{-1} I_i A v, w \rangle = \langle A v, I_i^T A_i^{-1} I_i A w \rangle = \langle v, P_i w \rangle_A.$$

Promotrimo operator

$$A_J = \sum_{i=1}^{s} P_i, \tag{5.21}$$

koji predstavlja prekondicioniranu matricu $M^{-1}A$ aditivne Schwarzove metode. Budući da je svaki P_i A-hermitski operator, tada je to i A_J. Zbog toga A_J ima realne svojstvene vrijednosti. Direktna posljedica činjenice da su operatori P_i projektori, iskazana je u sljedećem teoremu.

Teorem 5.2.6 ([32]). *Najveća svojstvena vrijednost od A_J je takva da vrijedi*

$$\lambda_{max}(A_J) \leq s,$$

gdje je s broj poddomena.

Dokaz: Za svaku matričnu normu je $\lambda_{max}(A_J) \leq \|A_J\|$. Posebno za A-normu, imamo

$$\lambda_{max}(A_J) \leq \sum_{i=1}^{s} \|P_i\|_A.$$

A-norma svakog P_i je jednaka 1, budući da je P_i A-ortogonalan projektor. Time smo dokazali tvrdnju teorema. □

Kako bismo ocijenili najmanju svojstvenu vrijednost prekondicionirane matrice A_J, moramo navesti jednu pretpostavku koja se tiče dekompozicije proizvoljnog vektora v na komponente od Ω_i.

Pretpostavka 1. Postoji konstanta K_0 takva da je nejednakost

$$\sum_{i=1}^{s} \langle Au_i, u_i \rangle \leq K_0 \langle Au, u \rangle,$$

zadovoljena za svako $u \in \Omega$, ako je u reprezentiran kao

$$u = \sum_{i=1}^{s} u_i, \qquad u_i \in \Omega_i.$$

Teorem 5.2.7 ([32]). *Ako vrijedi Pretpostavka 1, tada je*

$$\lambda_{min}(A_J) \geq \frac{1}{K_0}.$$

Dokaz: Započnimo sa proizvoljnim vektorom u, koji je predstavljen rastavom $u = \sum_{i=1}^{s} u_i$, tada, zbog toga što je P_i A-ortogonalan projektor na Ω_i, imamo

$$\langle u, u \rangle_A = \sum_{i=1}^{s} \langle u_i, u \rangle_A = \sum_{i=1}^{s} \langle P_i u_i, u \rangle_A = \sum_{i=1}^{s} \langle u_i, P_i u \rangle_A.$$

Korištenjem Cauchy–Schwarzove nejednakosti dobivamo

$$\langle u,u\rangle_A = \sum_{i=1}^{s}\langle u_i, P_i u\rangle_A \le \left(\sum_{i=1}^{s}\langle u_i,u_i\rangle_A\right)^{1/2}\left(\sum_{i=1}^{s}\langle P_i u, P_i u\rangle_A\right)^{1/2}.$$

Prema Pretpostavci 1, dalje slijedi

$$\|u\|_A^2 \le K_0^{1/2}\|u\|_A\left(\sum_{i=1}^{s}\langle P_i u, P_i u\rangle_A\right)^{1/2},$$

odakle, nakon kvadriranja, slijedi

$$\|u\|_A^2 \le K_0\sum_{i=1}^{s}\langle P_i u, P_i u\rangle_A.$$

Napokon, primijetimo da, budući da je P_i A-ortogonalan projektor, vrijedi

$$\sum_{i=1}^{s}\langle P_i u, P_i u\rangle_A = \sum_{i=1}^{s}\langle P_i u, u\rangle_A = \left\langle \sum_{i=1}^{s} P_i u, u\right\rangle_A = \langle A_J u, u\rangle_A.$$

Prema tome, za svaki u vrijedi nejednakost

$$\langle A_J u, u\rangle_A \ge \frac{1}{K_0}\langle u,u\rangle_A,$$

što daje traženu ogradu prema Teoremu 1.6.6. □

Dakle kad bi simetrično prekondicionirani sustav, sa matricom prekondicioniranja dobivenom iz aditivne Schwarzove metode, rješavali pomoću CG metode, tada bi konvergencija te metode ovisila kvocijentu najveće i najmanje svojstvene vrijednosti matrice $A_J = M^{-1}A$. U tom slučaju bilo bi

$$\kappa(A_J) \le sK_0,$$

što ne ovisi o koraku mreže h.

Sada prelazimo na analizu multiplikativne Schwarzove metode. Započinjemo sa već poznatom relacijom za grešku

$$e_s = Q_s e_0.$$

Želimo naći gornju ogradu za $\|Q_s\|_A$. Prvo primijetimo da iz relacije (5.16) Leme 5.2.4 slijedi da je

$$Q_i = Q_{i-1} - P_i Q_{i-1},$$

odakle je, zbog toga što je P_i A-ortogonalan projektor

$$\|Q_i v\|_A^2 = \|Q_{i-1}v\|_A^2 - \|P_i Q_{i-1}v\|_A^2,$$

za proizvoljan vektor v. Gornja jednakost vrijedi i za $i=1$, ako uzmemo da je $Q_0 = I$. Ako sada sumiramo ove jednakosti od 1 do s, dobit ćemo kao rezultat

$$\|Q_s v\|_A^2 = \|v\|_A^2 - \sum_{i=1}^{s}\|P_i Q_{i-1}v\|_A^2. \tag{5.22}$$

Ovime smo pokazali da A-norma greške neće rasti u svakom potkoraku unutarnje petlje multiplikativne Schwarzove metode.

Za dokazivanje sljedeće leme, moramo uvesti još jednu pretpostavku.

Pretpostavka 2. Za bilo koji podskup S skupa $\{1, 2, \ldots, s\}^2$ i $u_i, v_j \in \Omega$, vrijedi sljedeća nejednakost:

$$\sum_{(i,j)\in S} \langle P_i v_i, P_j v_j \rangle_A \leq K_1 \left(\sum_{i=1}^{s} \|P_i u_i\|_A^2 \right)^{1/2} \left(\sum_{j=1}^{s} \|P_j v_j\|_A^2 \right)^{1/2}. \tag{5.23}$$

Lema 5.2.8 ([32]). *Ako su zadovoljene Pretpostavke 1 i 2, tada vrijedi sljedeća nejednakost:*

$$\sum_{i=1}^{s} \|P_i v\|_A^2 \leq (1 + K_1)^2 \sum_{i=1}^{s} \|P_i Q_{i-1} v\|_A^2. \tag{5.24}$$

Dokaz: Zbog toga što je P_i A-ortogonalni projektor, imamo

$$\langle P_i v, P_i v \rangle_A = \langle P_i v, P_i Q_{i-1} v \rangle_A + \langle P_i v, (I - Q_{i-1}) v \rangle_A,$$

odakle uz pomoć (5.19) slijedi

$$\sum_{i=1}^{s} \|P_i v\|_A^2 = \sum_{i=1}^{s} \langle P_i v, P_i Q_{i-1} v \rangle_A + \sum_{i=1}^{s} \sum_{j=1}^{i-1} \langle P_i v, P_j Q_{j-1} v \rangle_A. \tag{5.25}$$

Za prvi izraz na desnoj strani, koristimo Cauchy–Schwarzovu nejednakost, kako bismo dobili

$$\sum_{i=1}^{s} \langle P_i v, P_i Q_{i-1} v \rangle_A \leq \left(\sum_{i=1}^{s} \|P_i v\|_A^2 \right)^{1/2} \left(\sum_{i=1}^{s} \|P_i Q_{i-1} v\|_A^2 \right)^{1/2}.$$

Za drugi izraz desne strane od (5.25) primjenjujemo (5.23) iz Pretpostavke 2, da bi dobili

$$\sum_{i=1}^{s} \sum_{j=1}^{i-1} \langle P_i v, P_j Q_{j-1} v \rangle_A \leq K_1 \left(\sum_{i=1}^{s} \|P_i v\|_A^2 \right)^{1/2} \left(\sum_{i=1}^{s} \|P_i Q_{i-1} v\|_A^2 \right)^{1/2}.$$

Zbrajajući ove dvije nejednakosti, kvadrirajući rezultat uz korištenje (5.25) dobivamo traženi rezultat. □

Iz (5.22) možemo zaključiti, da ako Pretpostavka 2 vrijedi tada je

$$\|Q_s v\|_A^2 \leq \|v\|_A^2 - \frac{1}{(1 + K_1)^2} \sum_{i=1}^{s} \|P_i v\|_A^2. \tag{5.26}$$

Sada možemo iskoristiti i Pretpostavku 1 kako bismo dobili donju ogradu za $\sum_{i=1}^{s} \|P_i v\|_A^2$.

Teorem 5.2.9 ([32]). *Pretpostavimo da vrijede Pretpostavke 1 i 2. Tada je*

$$\|Q_s\|_A \leq \left[1 - \frac{1}{K_0 (1 + k_1)^2} \right]^{1/2}.$$

Dokaz: Uz pomoćnu relaciju $\|P_i v\|_A^2 = \langle P_i v, v\rangle_A$, imamo

$$\sum_{i=1}^{s} \|P_i v\|_A^2 = \left\langle \sum_{i=1}^{s} P_i v, v \right\rangle_A = \langle A_J v, v\rangle_A.$$

Prema Teoremu 5.2.7 je $\lambda_{min}(A_J) \geq \frac{1}{K_0}$, odakle slijedi da je $\langle A_J v, v\rangle_A \geq \langle v, v\rangle_A / K_0$ za proizvoljni v. Zbog toga je

$$\sum_{i=1}^{s} \|P_i v\|_A^2 \geq \frac{\langle v, v\rangle_A}{K_0},$$

koje kad ubacimo u (5.26), daje nejednakost

$$\frac{\|Q_s v\|_A^2}{\|v\|_A^2} \leq 1 - \frac{1}{K_0(1+K_1)^2}.$$

Rezultat slijedi iz uzimanja maksimuma po svim vektorima v. □

Dakle možemo zaključiti da je za multiplikativnu Schwarzovu metodu, za jedno izvođenje unutarnje petlje po poddomenama,

$$\|e_s\|_A \leq \left[1 - \frac{1}{K_0(1+K_1)^2}\right]^{1/2} \|e_0\|_A.$$

Konvergencija ponovo ne ovisi o koraku mreže h.

Ovi rezultati daju informacije o konvergenciji Schwarzovih metoda uz dvije važne pretpostavke. Te pretpostavke se ne mogu provjeriti samo uz pomoć argumenata linearne algebre. Ako uzmemo neki linarni sustav, vrlo teško ćemo provjeriti da li te pretpostavke vrijede za njega. Međutim te pretpostavke vrijede za sustav dobiven iz diskretizacije pomoću konačnih elemenata eliptične parcijalne diferencijalne jednadžbe.

5.3 Velik broj poddomena i korištenje grube mreže

Ako definiramo mrežu na domeni problema Ω sa korakom h, tada prema prethodnim rezultatima uz određene uvjete, stopa konvergencije metoda dekompozicije domene sa preklapajućim domenama ne ovisi o h. Sada pretpostavimo da imamo puno poddomena Ω_i, svaka veličine $H \gg h$. Drugim riječima, na Ω_i možemo gledati kao na područje ograđeno grubom mrežom sa korakom H, plus dodatne ćelije fine mreže iza te granice, kako bi dobili preklapanje. Tada, na primjer, za aditivnu Schwarzovu metodu, kao što se vidi iz Teorema 5.2.6 i 5.2.7, uvjetovanost prekondicionirane matrice ovisi o broji poddomena s, odnosno ono je reda veličine $\mathcal{O}(H^{-1})$. Prema tome, ako se radi o pozitivno definitnoj matrici, ili barem simetričoj, uz simetrično prekondicioniranje, i ako za rješavanje sustava s takvom matricom koristimo neku od standardnih iterativnih metoda aproksimacije iz Krylovljevih potprostora, tada će i konvergencija te metode ovisiti o uvjetovanosti prekondicionirane matrice, odnosno o $\mathcal{O}(H^{-1})$. U slučaju da nismo zadovoljni sa ovakvom konvergencijom, i ako želimo imati stopu konvergencije kao multigrid, koji ne ovisi ni o h, niti o H, možemo se koristiti upravo idejom multigrid metoda. Dryja i Widlund, predložili su da, uz rješavanje problema na poddomenama, globalni problem treba riješiti na grubljoj mreži, uz standardni operator restrikcije I_h^H i interpolacije I_H^h.

Drugim riječima, treba još izvršiti korekciju na gruboj mreži. Kao što smo već prije vidjeli, obje Schwarzove metode ekvivalentne su jednostavnim iteracijama oblika

$$u_{novi} = (I - M^{-1}A)u_{stari} + M^{-1}b,$$

što se poklapa sa jednim prolaskom kroz sve poddomene. Kod multiplikativne Scwarzove metode korekcija tada izgleda

$$\begin{aligned} u_{korigirani} &= u_{novi} + I_H^h (A^H)^{-1} I_h^H r_{novi} = \\ &= [I - I_H^h (A^H)^{-1} I_h^H A] u_{novi} + I_H^h (A^H)^{-1} I_h^H b = \\ &= [I - I_H^h (A^H)^{-1} I_h^H A](I - M^{-1}A) u_{stari} + f = \\ &= \{I - [I_H^h (A^H)^{-1} I_h^H + (I - I_H^h (A^H)^{-1} I_h^H A) M^{-1}]A\} u_{stari} + f, \end{aligned}$$

pri čemu je f vektor

$$f = [I_H^h (A^H)^{-1} I_h^H + (I - I_H^h (A^H)^{-1} I_h^H A) M^{-1}]b,$$

a A^H je restrikcija operatora A na grubu mrežu. Vidimo da smo ponovo dobili jednostavne iteracija odakle se vidi da je matrica prekondicioniranja M_{tl} ovakvog algoritma sa korekcijom na gruboj mreži

$$\begin{aligned} M_{tl} &= I_H^h (A^H)^{-1} I_h^H + (I - I_H^h (A^H)^{-1} I_h^H A) \cdot \\ &\quad \cdot \left(\sum_{i=1}^{s} T_i + \sum_{i>j} T_i A T_j + \cdots + (-1)^s T_s A T_{s-1} \cdots T_2 A T_1 \right), \end{aligned}$$

gdje je $T_i = I_i^T A_i^{-1} I_i$.

Kod aditivne Schwarzove metode korekcija je

$$\begin{aligned} u_{korigirani} &= u_{novi} + I_H^h (A^H)^{-1} I_h^H r_{stari} = \\ &= u_{stari} + [M^{-1} + I_H^h (A^H)^{-1} I_h^H] r_{stari} = \\ &= \{I - [M^{-1} + I_H^h (A^H)^{-1} I_h^H] A\} u_{stari} + f, \end{aligned}$$

gdje je

$$f = [M^{-1} + I_H^h (A^H)^{-1} I_h^H] b.$$

Ponovo se radi o jednostavnim iteracijama, kod kojih je matrica prekondicioniranja M_{tl} jednostavnijeg oblika

$$M_{tl} = \sum_{i=1}^{s} T_i + I_H^h (A^H)^{-1} I_h^H.$$

Ovakva metoda se zove *two-level* metoda, kod koje je pokazano da konvergencija ne ovisi niti o koraku mreže h, niti o veličini poddomena H, uz uvjet da je širina područja preklapanja $\mathcal{O}(H)$. Dakle, možemo zaključiti da, sa samo nekoliko velikih poddomena, rješavanje problema na poddomenama će biti preskupo, dok sa mnogo malih poddomena, rješavanje na gruboj mreži će biti preskupo. Zato treba naći dobru ravnotežu između broja i veličine poddomena.

Glava 6

Numerički primjeri

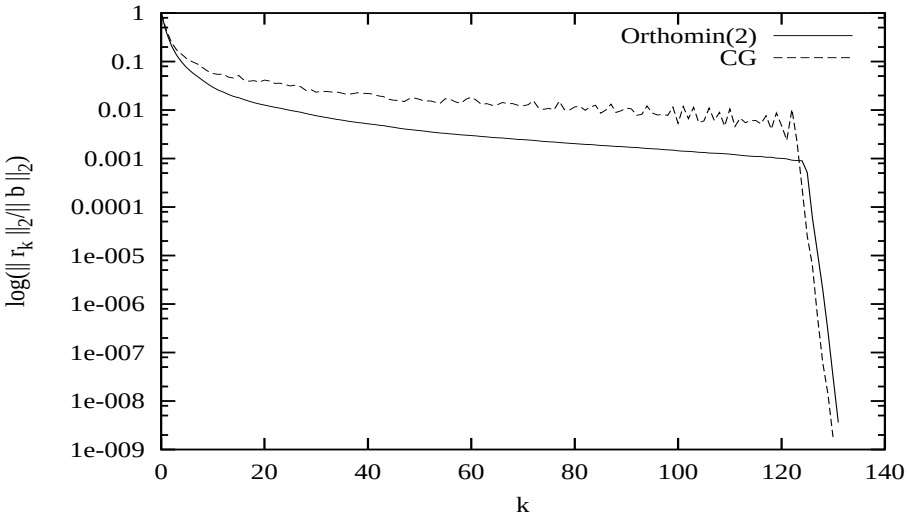
U ovom zadnjem poglavlju demonstrirat ćemo rad svih iterativnih metoda za rješavanje linearnih sustava, koje su spomenute u ovoj radnji. Osim što će se moći vidjeti da li metoda konvergira ili ne konvergira za određeni linearni sustav, potvrdit će se također i neki teoretski rezultati vezani uz iterativne metode i prekondicioniranje. U svim primjerima, osim u nekoliko njih, kada će to biti posebno naglašeno, dimenzija matrice A je 100×100, početna iteracija je $x_0 = [0\ 0\ \dots\ 0]^T$, a desna strana sustava b je određena tako da rješenje sustava bude jednako $x = [1\ 1\ \dots\ 1]^T$, odnosno da je $b = A \cdot x$. Sve metode programirane su u *Matlabu*, uz mašinsku točnost $\epsilon = 2.2204 \cdot 10^{-16}$, pri čemu se u svakom koraku k kontrolira relativna norma reziduala $\|r_k\|_2/\|b\|_2$ i iteriranje se zaustavlja kada je ona manja od $tol = 10^{-8}$.

6.1 Primjer 1

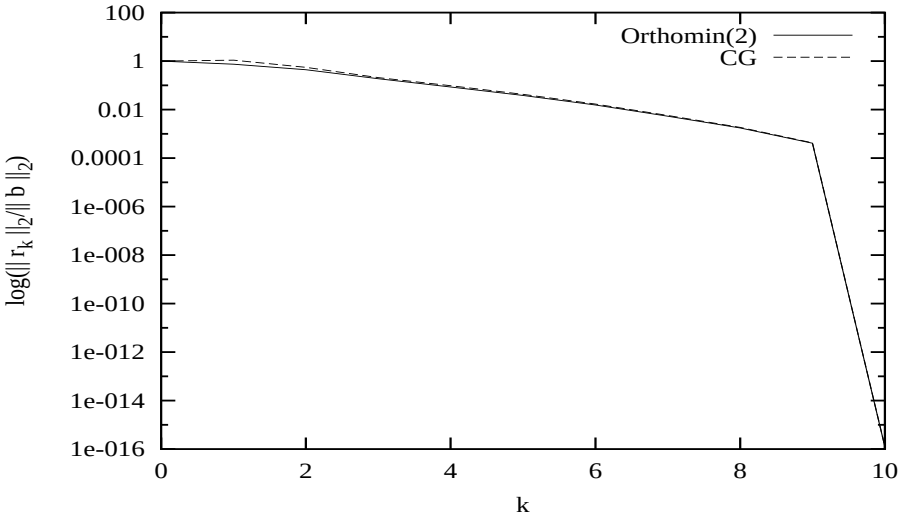
Matrica sustava u ovom primjeru je simetrična i pozitivno definitna, sa svojstvenim vrijednostima $\lambda(A) \in \{1, 4, 9, \dots, 10000\}$, a dobivena je kao produkt $A = Q\Lambda Q^T$, pri čemu je Λ dijagonalna matrica svojstvenih vrijednosti, a Q slučajna ortogonalna matrica. Uvjetovanost joj je jednaka $\kappa(A) = 10^4$. Za rješavanje ovog sustava korištene su metode CG i Orthomin(2), koje bi u egzaktnoj aritmetici trebale konvergirati u najviše 100 iteracija, međutim budući da je matrica A loše uvjetovana, i budući da broj iteracija do postizanja željene točnosti ovisi o $\mathcal{O}(\sqrt{\kappa})$, broj iteracija do postizanja tolearncije tol u aritmetici konačne preciznosti je veći od 100.

6.2 Primjer 2

Situacija u ovom primjeru je slična prethodnom, samo što pozitivno definitna matrica A ima deset različitih svojstvenih vrijednosti, svaka od njih kratnosti 10. Dakle, $A = Q\Lambda Q^T$, gdje je Λ dijagonalna matrica svojstvenih vrijednosti $\lambda(A) \in \{1, 2, \dots, 10\}$, a Q slučajna ortogonalna matrica. Uvjetovanost matrice A iznosi $\kappa(A) = 10$. Budući da, konvergencija CG i Orthomin(2) metoda ovisi o stupnju polinoma koji minimizira vrijednost polinoma u različitim svojstvenim vrijednostima matrice A, takav polinom može imati minimalno stupanj 10, pa je za konvergenciju dovoljno 10 iteracija gore navedenih metoda.



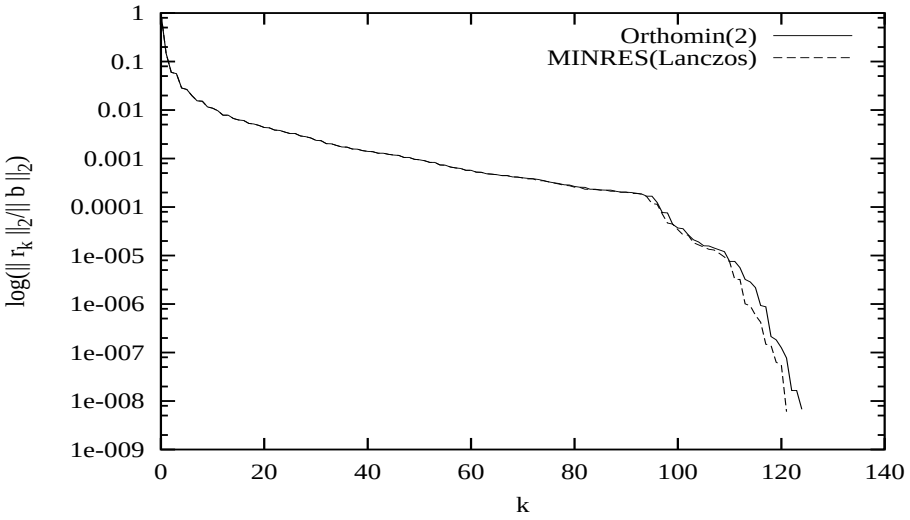
Slika 6.1: Primjer 1.



Slika 6.2: Primjer 2.

6.3 Primjer 3

Matrica sustava A i u ovom primjeru ostaje simetrična, medjutim ona je indefinitna, odnosno svojstvene vrijednosti mogu biti i negativne, $\lambda(A) \in \{-50, -49, \ldots, -1, 2, 4, \ldots, 100\}$. Isto je dobivena kao produkt $A = Q\Lambda Q^T$, dijagonalne matrice svojstvenih vrijednosti Λ i slučajne ortogonalne matrice Q, a uvjetovanost joj je jednaka $\kappa(A) = 100$. Budući da matrica nije pozitivno definitna, ne možemo upotrijebiti CG metodu, već koristimo dvije varijante istog algoritma: Orthomin(2) i MINRES metodu dobivenu preko Lanczosovog algoritma. Pokazuje se da je ova druga metoda malo stabilnija od prve.



Slika 6.3: Primjer 3.

6.4 Primjer 4

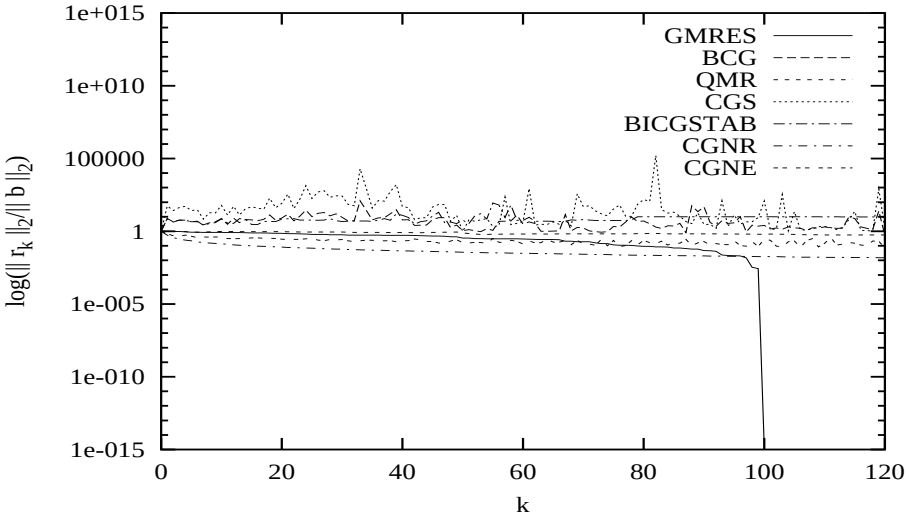
Sljedećih šest primjera bavi se nehermitskim matricama, odnosno metodama za rješavanje nehermitskih sustava. Matrica sustava A ovog primjera je ne-normalna, nedijagonalizabilna matrica, kojoj se polje vrijednosti ne može smijestiti u kružnicu koja ne sadrži ishodište. Zato ne možemo ništa reći o konvergenciji GMRES metode primijenjene na taj sustav, osim da mora konvergirati do 100-te iteracije. Singularne vrijednosti matrice A su $\sigma(A) \in \{1, 4, 9, \ldots, 10000\}$, a dobivena je kao $A = U\Sigma V^T$, pri čemu je Σ dijagonalna matrica singularnih vrijednosti, a U i V su različite slučajne ortogonalne matrice. Uvjetovanost joj iznosi $\kappa(A) = 10000$. Niti ostale metode nemaju nikakvu garanciju da će konvergirati u manje od 100 koraka, dapače pokazuje se da su vrlo nestabilne i vrlo slabo konvergiraju. Konvergencija metoda koje rješavaju normalne jednadžbe ovisi o uvjetovanosti matrice koja je u ovom slučju velika, pa ponovo nemamo garanciju o dobroj konvergenciji.

6.5 Primjer 5

([29]) Matrica sustava ovog primjera je vrlo jednostavna. Ona je nehermitska, ali je ortogonalna, stoga se singularne vrijednosti svode na jednu jedinu $\sigma(A) \in \{1\}$, a uvjetovanost joj je jednaka 1. Matrica A ima oblik

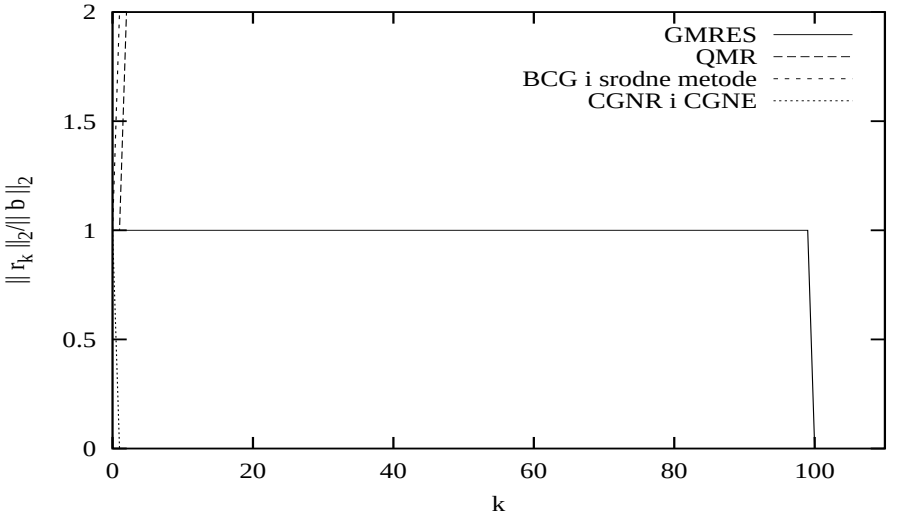
$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Desna strana u ovom slučaju je jednaka $b = \xi_{100}$, odakle slijedi da je rješenje sustava $x = \xi_1$. ξ_i je i-ti jedinični vektor. Ako ponovo uzmemo da je x_0 nul-vektor, tada je



Slika 6.4: Primjer 4.

$r_0 = \xi_{100}$, $Ar_0 = \xi_{99}$, $A^2 r_0 = \xi_{98}$,...,$A^{98} r_0 = \xi_2$ i $A^{99} r_0 = \xi_1$. Kod primjene GMRES metode na ovaj sustav imamo situaciju da $r_0 \perp A\mathcal{K}_k(A, r_0)$ za $k = 1, \ldots, 99$, odakle slijedi da je $r_0 = r_1 = \cdots = r_{99}$, odnosno GMRES metoda neće konvergirati prije poslijednjeg, 100-og koraka. Osim toga, vektor rješenja x se ne nalazi niti u jednom Krylovljevom potprostoru i okomit je na njih, osim $\mathcal{K}_{100}(A, r_0)$, pa su svi ostali Krylovljevi potprostori loša aproksimacija za rješenje. Kod metoda koje se zasnivaju na dvostranom Lanczosovom algoritmu, za $\hat{r}_0 = r_0$ imamo da je $v_1 = w_1 = \xi_{100}$, $v_2 = \xi_{99}$, $\tilde{w}_2 = \xi_1$, pa je $\beta_1 = \langle v_2, \tilde{w}_2 \rangle = 0$ i dolazi do ozbiljnog sloma algoritma. Metode koje rješavaju normalne jednadže su ekvivalentne minimiziranju polinoma na skupu singularnih vrijednosti, koji ima jedan element. Zato to možemo postići polinomom stupnja jedan, i takve metode će konvergirati u jednom koraku.



Slika 6.5: Primjer 5.

6.6 Primjer 6

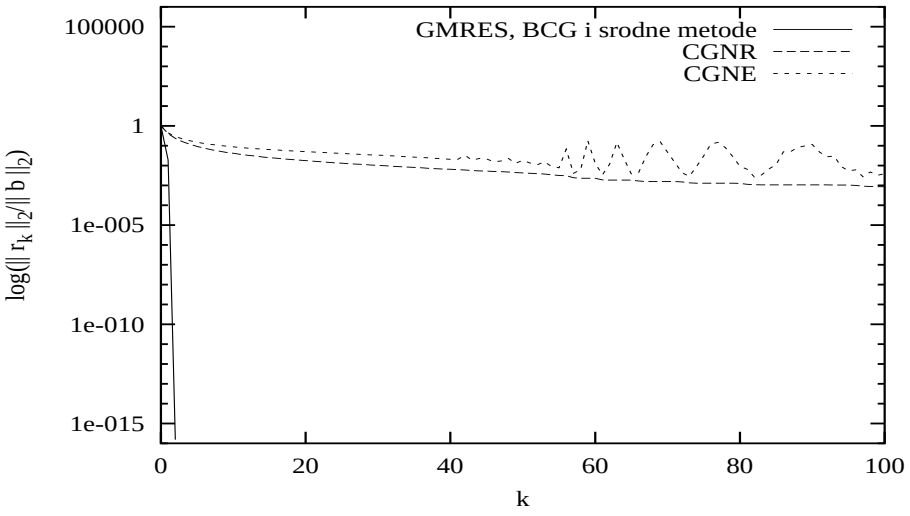
([29]) U ovom primjeru matrica sustava A je ne-normalna, i ima oblik

$$A = \begin{bmatrix} M_1 & & & \\ & M_2 & & \\ & & \ddots & \\ & & & M_{50} \end{bmatrix},$$

gdje je

$$M_i = \begin{bmatrix} 1 & i-1 \\ 0 & 1 \end{bmatrix}, \qquad i = 1, \ldots, 50.$$

Singularne vrijednosti matrice A nalaze se u rasponu otprilike od 0.2 do 50, a s druge strane vidimo da je njezin minimalni polinom stupnja 2. To znači da postoji polinom p_2 stupnja 2, takav da je $p_2(A) = 0$, odnosbo da je $p_2(A)r_0 = 0$ za bilo koji početni rezidual. Iz ovoga slijedi da će GMRES algoritam konvergirati nakon dvije iteracije. Slično vrijedi i za metode bazirane na Lanczosovom algoritmu, jer ispada da je $r_2 = const \cdot v_3 = 0$, jer je $v_3 = p_2(A)r_0 = 0$. S druge strane, za metode koje rješavaju normalne jednadžbe konvergencija ovisi o uvjetovanosti koja je jednaka $\kappa(A) = 2.4 \cdot 10^3$, a one nastoje pronaći polinom koje će se minimizirati na skupu kvadrata singularnih vrijednosti. Budući da njih ima 100 različitih na intervalu $\langle 0.2, 50 \rangle$, tim metodama treba veliki broj iteracija da bi dostigle konvergenciju.



Slika 6.6: Primjer 6.

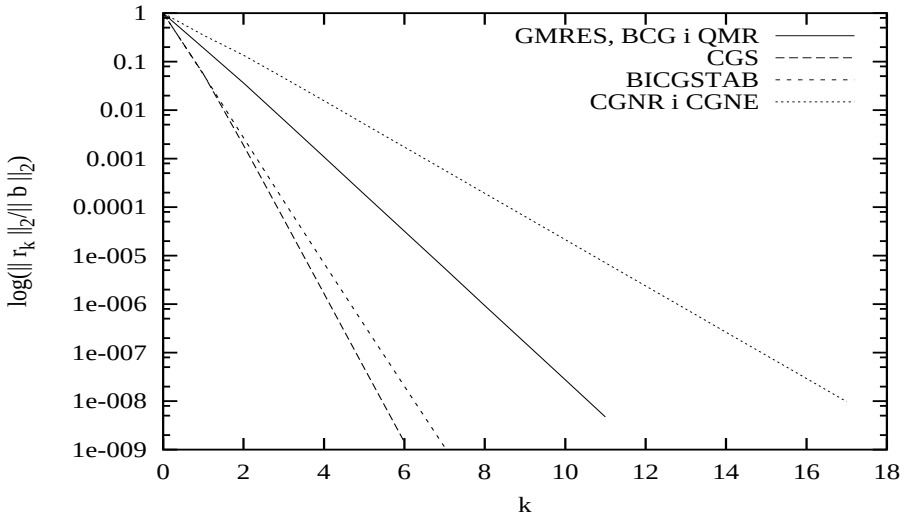
6.7 Primjer 7

([29]) Za razliku od prethodnih primjera, u ovom slučaju matrica sustava A je dijagonalna matrica, ali demonstrira tipično ponašanje iterativnih metoda za rješavanje nehermitskih sustava. Matrica je oblika $A = \mathrm{diag}(d_1, d_2, \ldots, d_{100})$, gdje d_i predstavlja

ekstremne točke Čebiševljevog polinoma na intervalu $[1, 2]$,

$$d_i = 1 + \frac{1}{2}\left(\cos\left(\frac{(i-1)\pi}{99}\right) + 1\right), \qquad i = 1, \ldots, 100.$$

Uvjetovanost matrice je $\kappa(A) = 2$. GMRES i QMR metoda ponašaju se tada kao MINRES metoda, koja je ekvivalentna minimiziranju polinoma nad skupom svojstvenih vrijednosti. Budući da je taj skup predstavlja točke gusto poredane u intervalu $[1, 2]$, može se naći polinom stupnja oko 10, koji dobro minimizira skup. Metoda BCG ponaša se kao CG, i nalazi se u istoj situaciji, jer polinomi koji minimiziraju A-normu greške i normu reziduala su jednaki u svakoj iteraciji. Metoda CGS trebala bi imati dva puta bržu konvergenciju od gornjih metoda, a BICGSTAB je prema pretpostavci blizu njene krivulje konvergencije. Metode koje rješavaju normalne jednadžbe očekivano lošije konvergiraju od GMRES, jer njihova konvergencija ovisi o $\mathcal{O}(\kappa(A))$, a ne o $\mathcal{O}(\sqrt{(}\kappa(A)))$ kao kod GMRES metode.



Slika 6.7: Primjer 7.

6.8 Primjer 8

([29]) U ovom primjeru ponovo se radi o ne-normalnoj matrici sustava A. Ona ima oblik

$$A = \begin{bmatrix} N_1 & & & \\ & N_2 & & \\ & & \ddots & \\ & & & N_{50} \end{bmatrix},$$

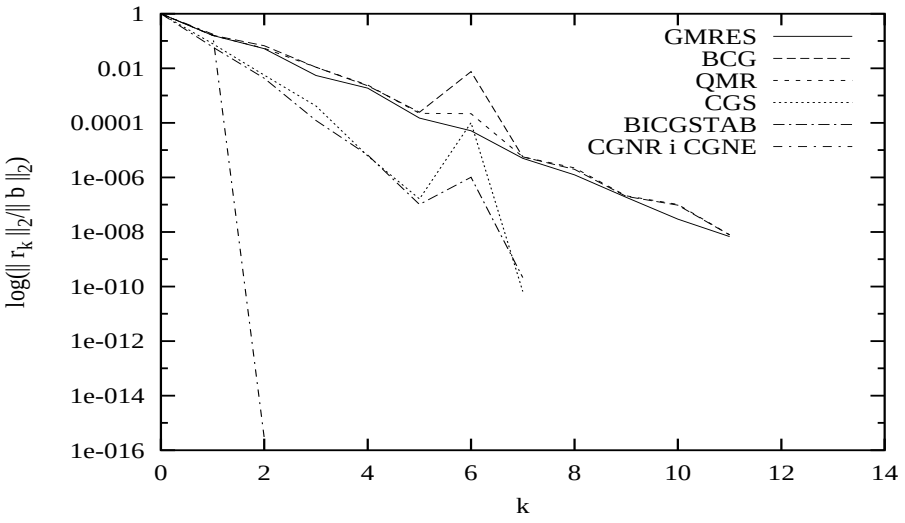
gdje je

$$N_i = \begin{bmatrix} d_i & c_i \\ 0 & 2/d_i \end{bmatrix}, \qquad i = 1, \ldots, 50,$$

i

$$d_i = 1 + \frac{1}{2}\left(\cos\left(\frac{(i-1)\pi}{49}\right) + 1\right), \quad c_i = \left(5 - d_i^2 - \frac{4}{d_i^2}\right)^{1/2}.$$

Svojstvene vrijednosti ove matrice raspoređene su po cijelom intervalu $[1, 2]$, dok su singularne vrijednosti svake matrice N_i, pa stoga i matrice A, jednake 1 i 2. Zato metode koje rješavaju normalne jednadžbe konvergiraju u 2 koraka, a ostale metode u puno više. Budući da GMRES i BCG slično konvergiraju, CGS i BICGSTAB imaju skoro dvostruko bržu konvergenciju.



Slika 6.8: Primjer 8.

6.9 Primjer 9

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

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

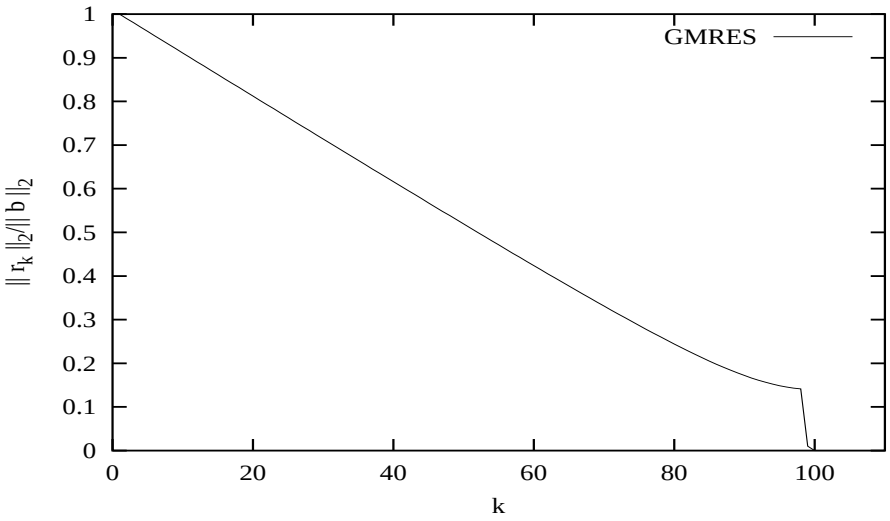
Matrica $V = [v_1 \ v_2 \ \cdots \ v_{100}]$ je slučajna ortogonalna matrica, i ako definiramo $b = \sum_{i=1}^{100} g(i)v_i$, tada je $\|b\|_2 = f(0)$. U nastavku, definirajmo još matricu $B = [b \ v_1 \ \cdots \ v_{99}]$ i izračunajmo koeficijente polinoma

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

i na kraju konstruirajmo matricu A kao

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

Prema teoretskim rezultatima, ovako definirani sustav $Ax = b$ ima matricu sa gore definiranim svojstvenim vrijednostima i sa rezidualima, takvim da nakon svake iteracije GMRES metode vrijedi $\|r_k\|_2 = f(k)$. Dobiveni rezultati u Slici 6.9 to potvrđuju, do na numeričke greške.



Slika 6.9: Primjer 9.

6.10 Primjer 10

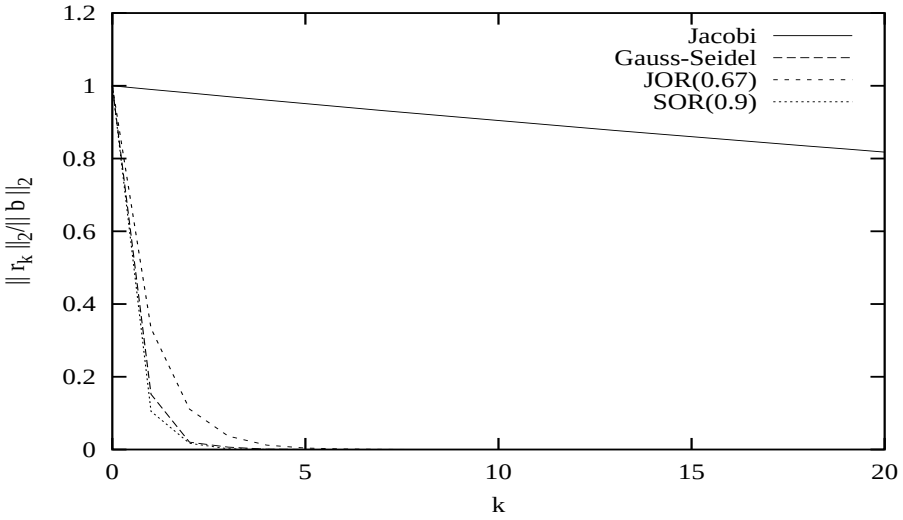
Sljedeća dva primjera demonstriraju različite načine prekondicioniranja. U ovom primjeru, riječ je o dijagonalno dominantnoj matrici sustava A, oblika

$$A = \begin{bmatrix} 1 & 0.01 & \cdots & 0.01 & 0.01 \\ 0.02 & 2 & \cdots & 0.02 & 0.02 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0.99 & 0.99 & \cdots & 99 & 0.99 \\ 1 & 1 & \cdots & 1 & 100 \end{bmatrix}.$$

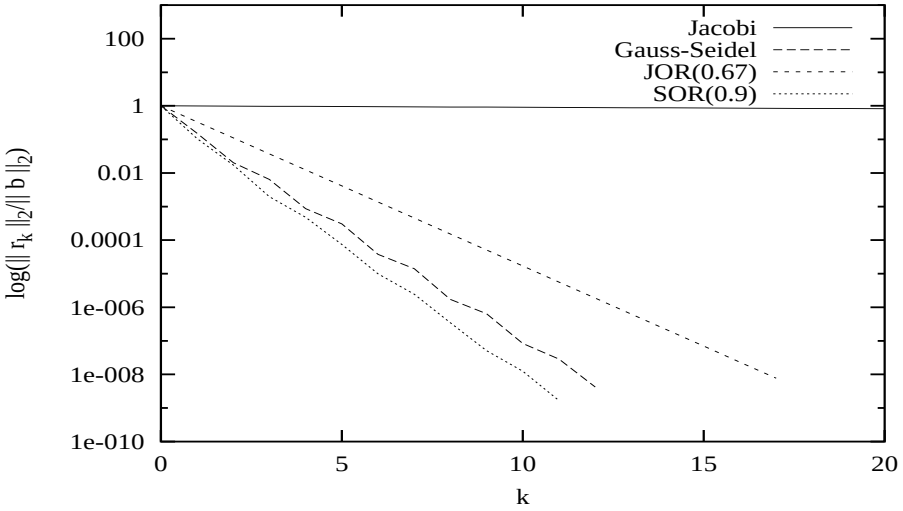
Za Jacobijevu i Gauss–Seidelovu metodu primijenjenu na ovakav sustav stope konvergencije iznose $\rho(G_J) = 0.99$ i $\rho(G_{GS}) = 0.2144$, dok je eksperimentalno utvrđeno da za JOR i SOR metodu, optimalni parametri iznose $\omega_{JOR} = 0.67$ i $\omega_{SOR} = 0.9$, za koje vrijedi $\rho(G_{JOR,0.67}) = 0.3367$ i $\rho(G_{SOR,0.9}) = 0.1713$. Ove teorijske pretpostavke potvrđuju i numerički rezultati prikazani na Slici 6.10, sa dekadskom skalom. Na primjer u prvom koraku vrijedi: za Jacobijevu metodu $\|r_1\|_2/\|r_0\|_2 = 0.99$, za Gauss–Seidelovu metodu $\|r_1\|_2/\|r_0\|_2 = 0.15184$, za JOR metodu $\|r_1\|_2/\|r_0\|_2 = 0.3333$ i za SOR metodu $\|r_1\|_2/\|r_0\|_2 = 0.105038$.

6.11 Primjer 11

Matrica sustava ovog primjera je rijetko popunjena Stieltjesova matrica, čiji raspored netrivijalnih elemenata je dan u Slici 6.12. Svojstvene vrijednosti ove matrice nalaze se

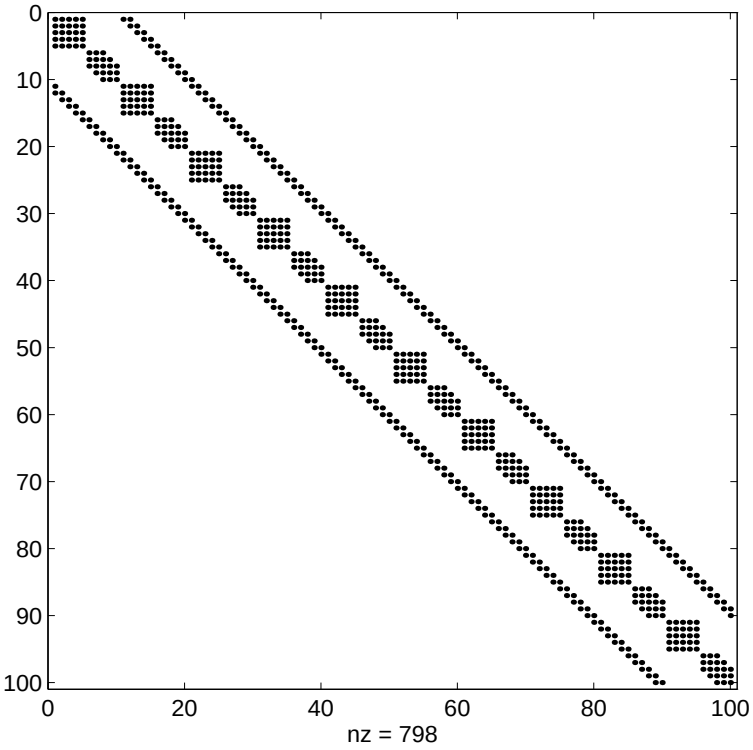


Slika 6.10: Primjer 10, dekadska skala.



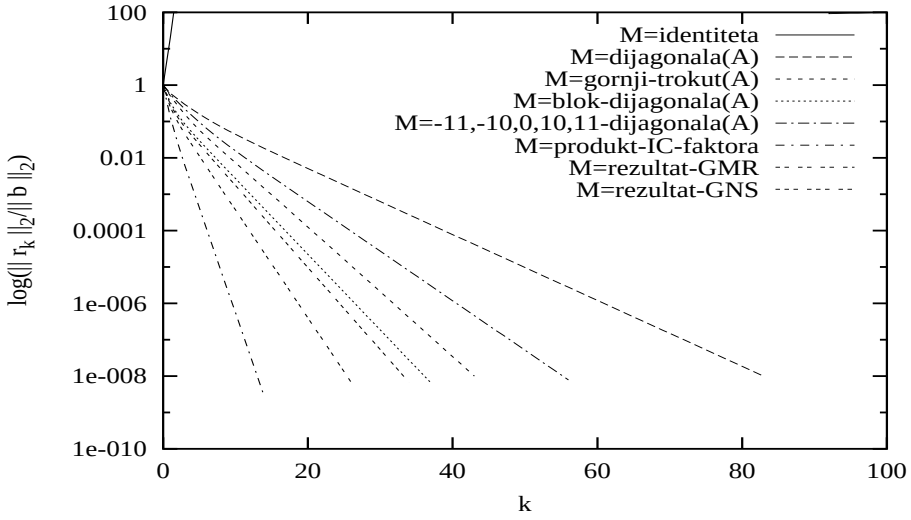
Slika 6.11: Primjer 10, logaritamska skala.

u intervalu $\lambda(A) \in \langle 3.23, 47.07\rangle$, i mnoge su vrlo blizu jedne drugima, a uvjetovanost iznosi $\kappa(A) = 14.5627$. Izvršene su dvije klase pokusa. U prvoj se izvode jednostavne iteracije sa različitim matricama prekondicioniranja M. Prema teorijskim rezultatima stopa konvergencije Gauss–Seidelove metode (M=gornji–trokut) je bolja od Jacobijeve metode (M=dijagonala), a prekondicioniranja u kojem je matrica M=blok-dijagonali matrice A i M=-11-oj, -10-oj, 0-oj, 10-oj i 11-oj dijagonali matrice A, također bolje konvergiraju od Jacobijeve metode, što se potvrđuje u numeričnkim rezultatima. Također, vidimo da su od prethodno navedenih prekondicioniranja, bolja prekondicioniranja koja su matricu M dobila iz metoda aproksimiranja inverza preko algoritama globalnog minimalnog reziduala i globalnog najbržeg silaska. Najbolje od svih je prekondicioniranje pomoću nekompletne faktorizacije Choleskog. U drugoj klasi pokusa, ista prekondicioniranja (osim Gauss-Seidelovog) korištena su uz primjenu CG metode. Ponovo prema

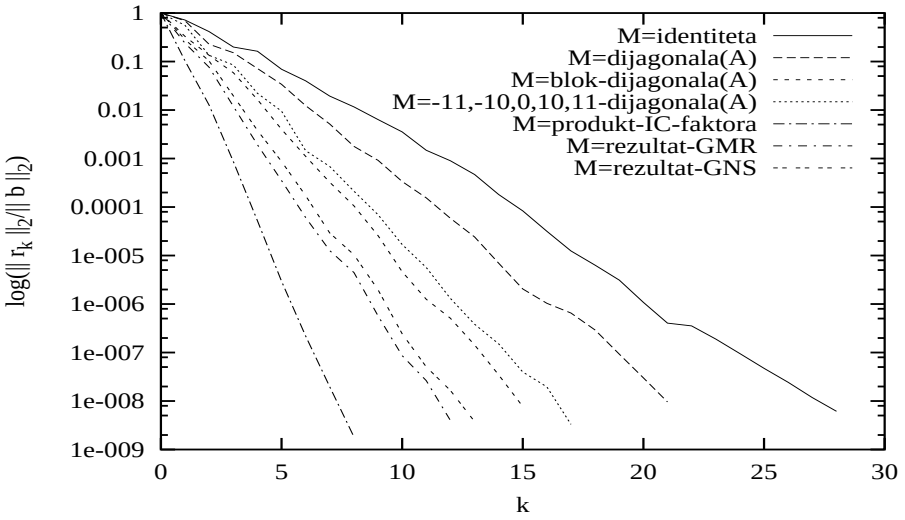


Slika 6.12: Raspored netrivijalnih elemenata matrice sustava iz Primjera 11.

teorijskim rezultatima, prekondicioniranja kod kojih se matrica M poklapa sa različitim dijagonalama matrice A, imaju, do na faktor 2, bolju konvergenciju od prekondicioniranja samo sa glavnom dijagonalom matrice A. Također, ponovo najbolju konvergenciju ima prekondicioniranje pomoću nekompletne faktorizacije Choleskog.



Slika 6.13: Primjer 11, jednostavne iteracije.



Slika 6.14: Primjer 11, konjugirani gradijenti.

6.12 Primjer 12

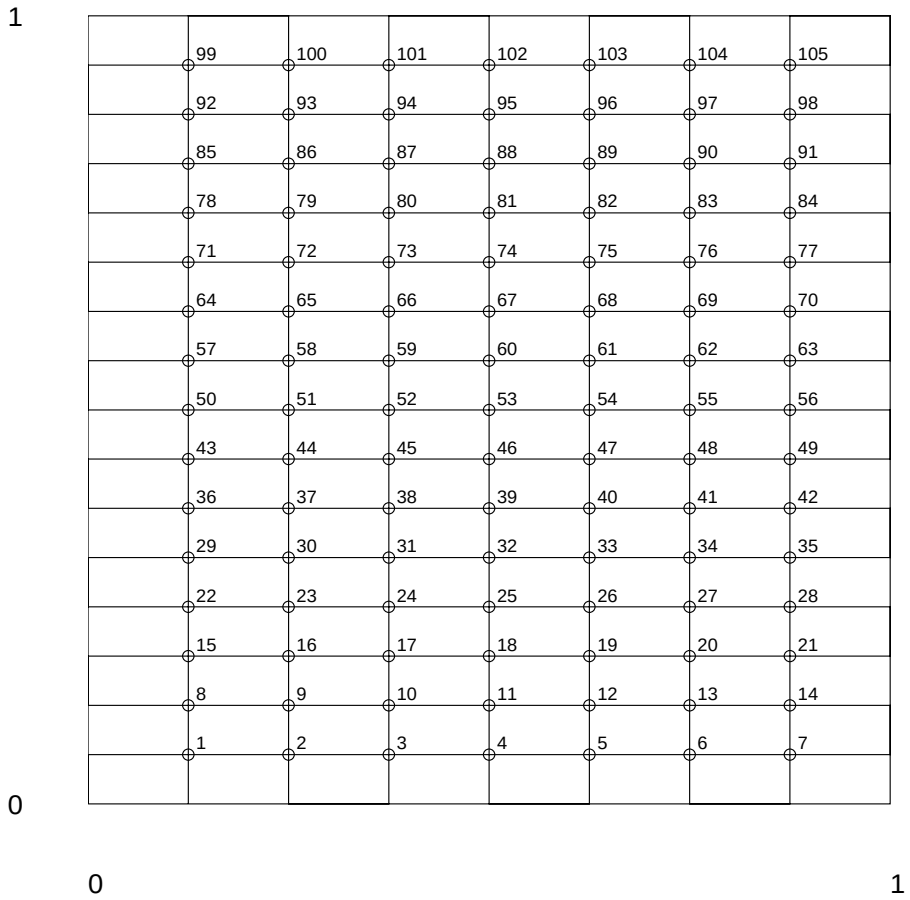
U ovom primjeru obrađena je primjena multigrid metode na sustav dobiven diskretizacijom dvodimenzionalne difuzijske jednadžbe. Uzeti su parametri $n_x = 7$ i $n_y = 15$, tako da mreža na kvadratu $[0,1] \times [0,1]$ izgleda kao na Slici 6.15. Raspored netrivijalnih elemenata matrice A dobivene iz diskretizacije, izgleda kao na Slici 6.16. Napravljena su dva pokusa multigrid metodama. U prvom je uzeta Poissonova jednadžba, na kojoj su izvršeni V i W ciklusi, pri čemu su za JOR metodu korištenu u multigrid iteraciji, upotrebljeni parametri $\omega = 2/3$ i $\omega = 4/5$. Budući da je ovaj drugi parametar optimalan za dvodimenzionalni slučaj, tada je i konvergencija brža. V i W ciklusi primijenjeni su i u slučaju kada su za početnu iteraciju uzeti rezultati dobiveni nakon FMG algoritma. Za FMG rezultat dobivamo da je $\|r\|_2/\|b\|_2 = 0.07$, i pri tom smo za dostizanje iste tolerancije kao i kad je početna iteracija bila jednaka nul-vektoru, uštedjeli samo jedan V-ciklus, a potrošili smo vrijeme za računanje samog FMG algoritma. Drugi pokus napravljen je za difuzijsku jednadžbu, kod koje je toplinski konduktivitet materijala definiran sa

$$a(x,y) = \begin{cases} x^2 + y^2, & x, y \in \langle 0, 0.5] \\ x^2, & x \in \langle 0.5, 1\rangle,\ y \in \langle 0, 0.5] \\ y^2, & x \in \langle 0, 0.5],\ y \in \langle 0.5, 1\rangle \\ 100, & x, y \in \langle 0.5, 1\rangle \end{cases}$$

Konvergencija V-ciklusa uz $\omega = 4/5$ je slična kao i kod Poissonove jednadžbe.

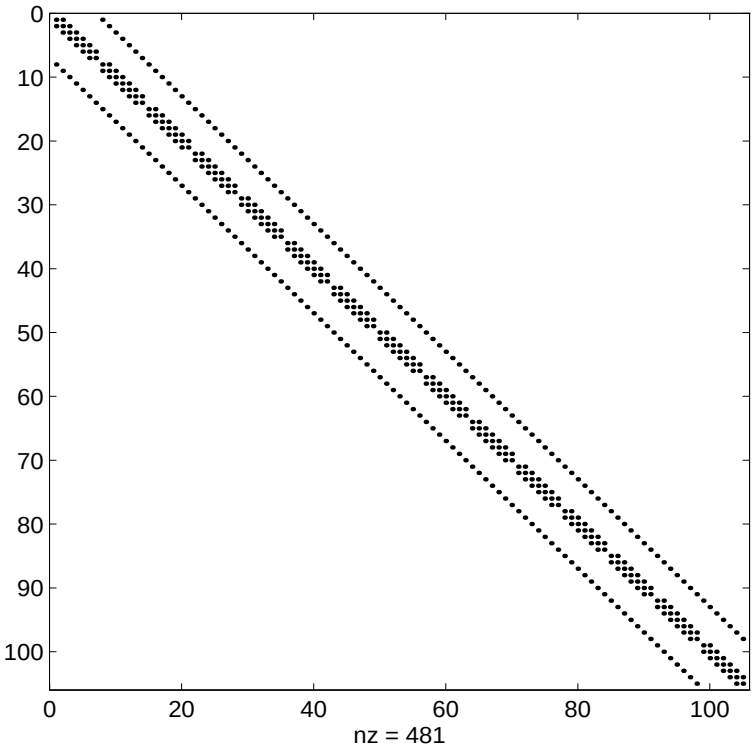
6.13 Primjer 13

U zadnjem primjeru napravljena je demonstracija rada metoda dekompozicije domene. Kao primjer domene uzet je L-oblik, koji je podijeljen u tri pravokutne poddomene, i to na dva načina: jednom bez preklapanja i drugi put sa preklapanjima. U oba slučaja mreža domene se sastojala od $n = 94$ točaka. Jednadžba koja se rješava na toj domeni je Poissonova jednadžba. Prvi pokus napravljen je sa dekompozicijom domene bez preklapanja, kada mreža na domeni izgleda kao u Slici 6.19. Matrica A koja je dobi-

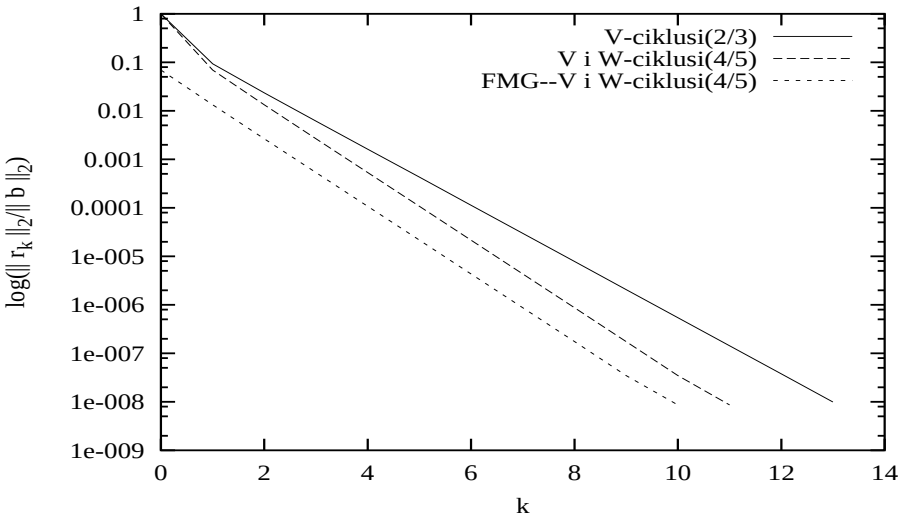


Slika 6.15: Primjer 12, mreža na intervalu $[0,1] \times [0,1]$.

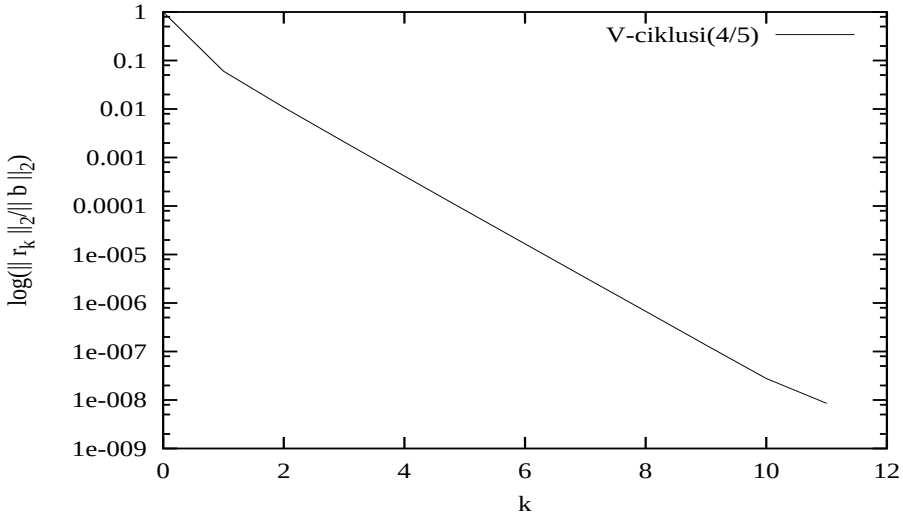
vena diskretizacijom ovog problema, uz zadani raspored točaka mreže, je simetrična i pozitivno definitna, a raspored njenih netrivijalnih elemenata prikazan je u Slici 6.20. Spektar matrice je sadržan u intervalu $\lambda(A) \in \langle 0.2494, 7.7506 \rangle$, tako da uvjetovanost iznosi $\kappa(A) = 31.077$. Problem je rješavan metodom blok-Gaussovih eliminacija, pri čemu se za invertiranje podmatrice i Schurovog komplementa koristila CG metoda. Dobivena aproksimacija rješenja imala je rezidual sa $\|r\|_2/\|b\|_2 = 1.4429 \cdot 10^{-6}$. Drugi pokus rješava problem dekompozicije domene sa preklapanjem, a mreža na domeni ima oblik kao u Slici 6.21. I u ovom slučaju matrica A dobivena diskretizacijom Poissonove jednadžbe na ovoj domeni, uz zadani raspored točaka, je ponovo simetrična i pozitivno definitna, a raspored njenih netrivijalnih elemenata je prikazan na Slici 6.22. Također spektar matrice A se nalazi u istom intervalu, sa istom uvjetovanošću, kao i u slučaju dekompozicije domene bez preklapanja, jer se radi o istoj matrici kojoj su ispermutirani određeni reci i stupci. Ovaj pokus izvršen je primjenom multiplikativne i aditivne Schwarzove metode, a rješavanje potproblema na poddomenama izvršeno je CG metodom. Dok multiplikativna metoda konvergira u 11 koraka, aditivna metoda divergira jer spektralni radijus njene matrice iteracija iznosi $\rho(G_{AS}) = 6.702 > 1$.



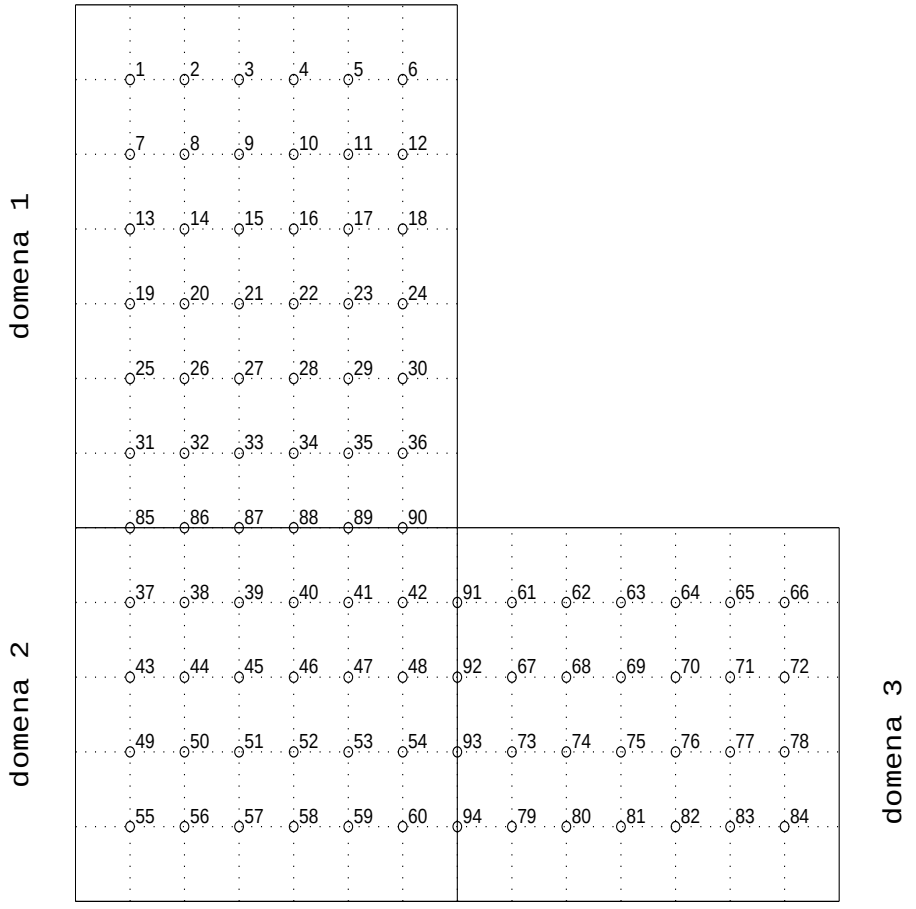
Slika 6.16: Raspored netrivijalnih elemenata matrice sustava iz Primjera 12.



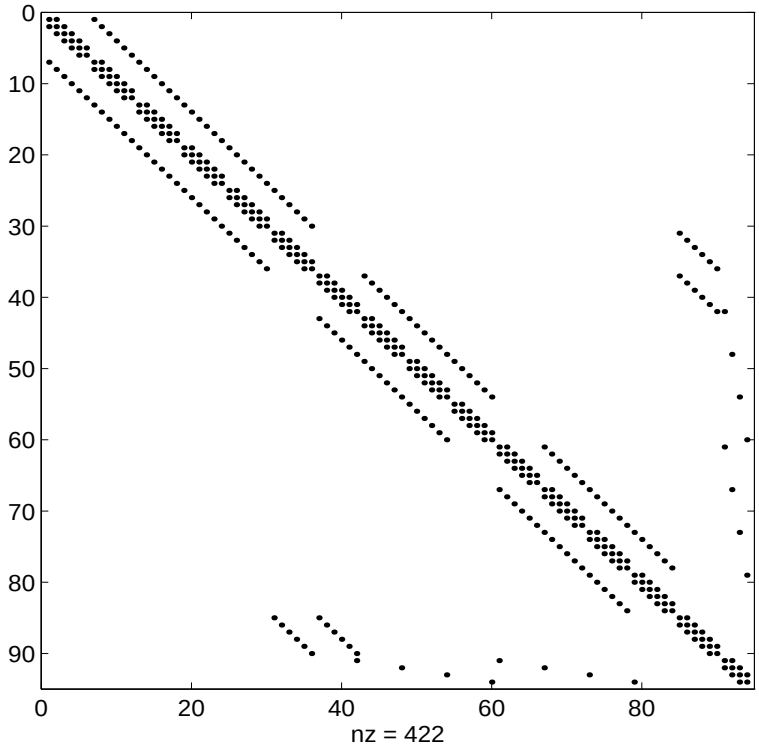
Slika 6.17: Primjer 12, Poissonova jednadžba.



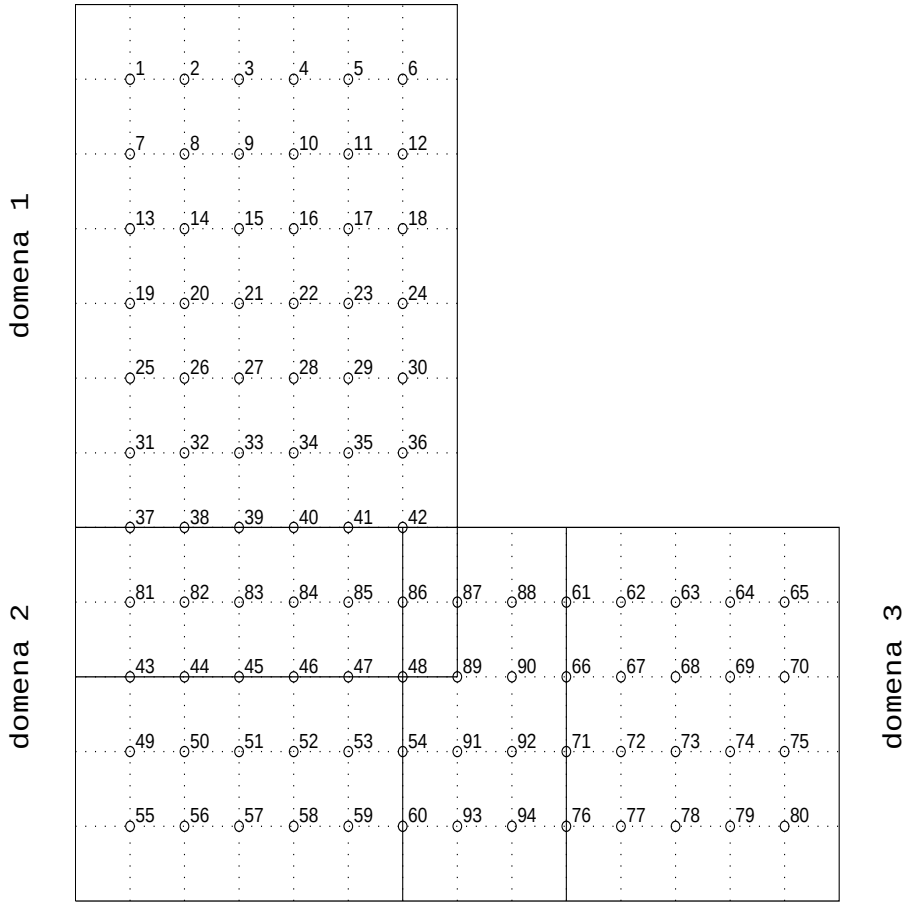
Slika 6.18: Primjer 12, difuzijska jednadžba.



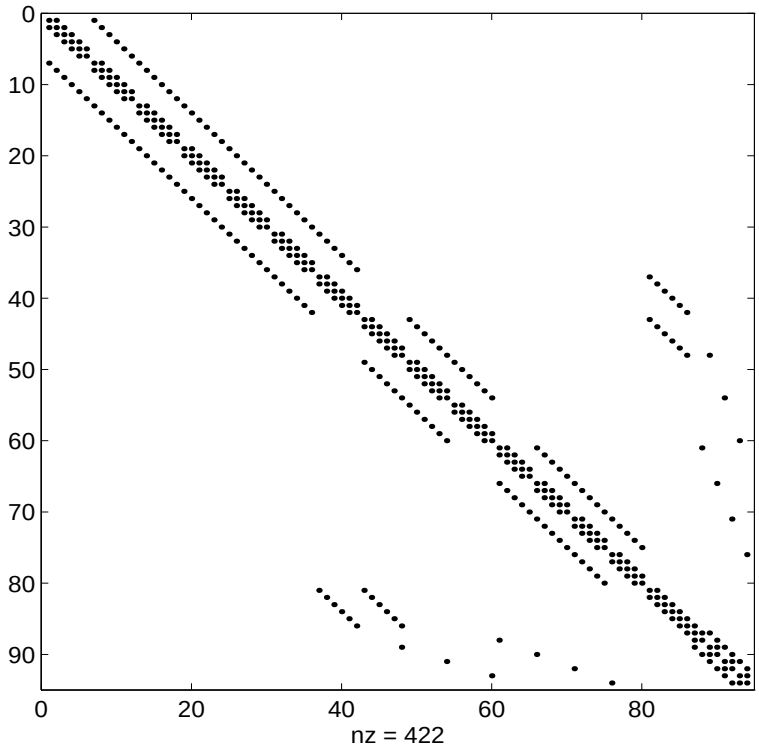
Slika 6.19: Primjer 13, mreža za dekompoziciju domene bez preklapanja.



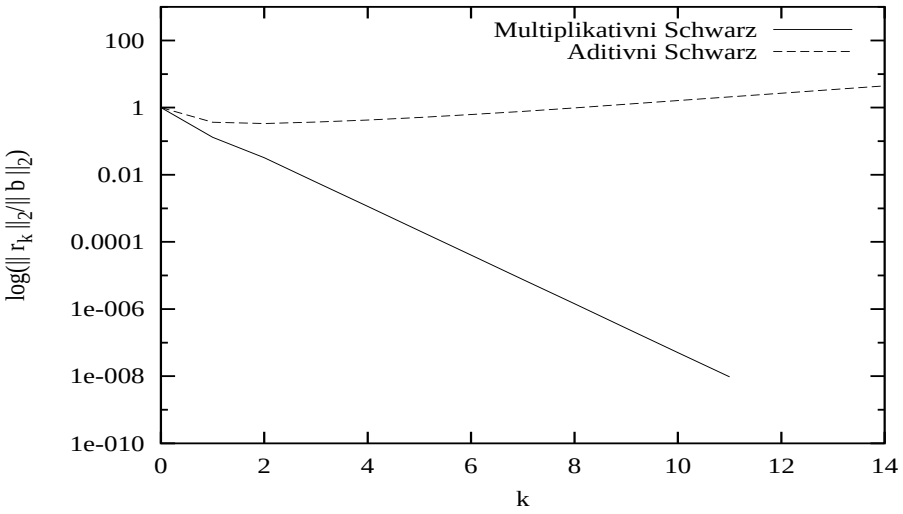
Slika 6.20: Raspored netrivijalnih elemenata matrice sustava iz Primjera 13, za dekompoziciju domene bez preklapanja.



Slika 6.21: Primjer 13, mreža za dekompoziciju domene sa preklapanjem.



Slika 6.22: Raspored netrivijalnih elemenata matrice sustava iz Primjera 13, za dekompoziciju domene sa preklapanjem.



Slika 6.23: Primjer 13, metode dekompozicije domene sa preklapanjem.

Bibliografija

[1] R. Barett, M. Berry, T. F. Chan, J. Demmel, J. M. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and V. vad der Vorst. **Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods**. SIAM, Philadelphia, 1994.

[2] Å . Björck and C. C. Paige. **Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm**. *SIAM J. Matrix Anal. Appl.*, 13(1): pp. 176–190, January 1992.

[3] W. L. Briggs, V. E. Henson, and S. F. McCormick. **A Multigrid Tutorial**. SIAM, Philadelphia, Second edition, 2000.

[4] P. N. Brown. **A theoretical comparision of the Arnoldi and GMRES algorithms**. *SIAM J. Sci. Stat. Comput.*, 12(1): pp. 58–78, January 1991.

[5] J. Cullum and A. Greenbaum. **Relations between Garlekin and norm-minimizing iterative methods for solving linear systems**. *SIAM J. Matrix Anal. Appl.*, 17(2): pp. 223–247, April 1996.

[6] J. Demmel. **The condition number of equivalence transformations that block diagonalize matrix pencil**. *SIAM J. Numer. Anal.*, 20(3): pp. 599–610, June 1983.

[7] J. Demmel. **Applied Numerical Linear Algebra**. SIAM, Philadelphia, 1997.

[8] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš. **Numerical Stability of GMRES**. *BIT*, 35: pp. 309–330, 1995.

[9] R. W. Freund and N. M. Nachtigal. **QMR: a quasi-minimal residual method for non-Hermitian linear systems**. *Numerische Mathematik*, 60: pp. 315–339, 1991.

[10] A. Greenbaum. **Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences**. *Linear Algebra and its Applications*, 113: pp. 7–63, 1989.

[11] A. Greenbaum. **Estimating the attainable accuracy of recursively computed residual methods**. *SIAM J. Matrix Anal. Appl.*, 18(3): pp. 535–551, July 1997.

[12] A. Greenbaum. **Iterative Methods for Solving Linear Systems**. SIAM, Philadelphia, 1997.

[13] A. Greenbaum and L. Gurvits. **Max–min properties of matrix factor norms**. *SIAM J. Sci. Comput*, 15(2): pp. 348–358, March 1994.

[14] A. Greenbaum, V. Pták, and Z. Strakoš. **Any nonincreasing convergence curve is possible for GMRES**. *SIAM J. Matrix Anal. Appl.*, 17(3): pp. 465–469, July 1996.

[15] A. Greenbaum and Z. Strakoš. **Predicting the behavior of finite precision Lanczos and conjugate gradient computations**. *SIAM J. Matrix Anal. Appl.*, 13(1): pp. 123–137, January 1992.

[16] A. Greenbaum and Z. Strakoš. **Matrices that generate the same Krylov residual spaces**. In G. Golub, A. Greenbaum and M. Luskin, editor, *Recent Advances in Iterative Methods*, pages 95–118. Springer–Verlag, Berlin, New York, 1994.

[17] A. Greenbaum and L. N. Trefethen. **GMRES/CR and Arnoldi/Lanczos as matrix approximation problems**. *SIAM J. Sci. Comput*, 15(2): pp. 359–368, March 1994.

[18] V. Hari. **Numerička linearna algebra**. Zabilješke sa predavanja iz kolegija *Numeričke linearne algebre* na Matematičkom odjelu PMF-a, Zagreb.

[19] V. Hari. **Matrična teorija perturbacija**. Sveučilište u Zagrebu, 1996.

[20] M. R. Hestenes and E. Stiefel. **Methods of Conjugate Gradients for Solving Linear Systems**. *Journal of Research of the National Bureau of Standards*, 49(6): pp. 409–436, December 1952.

[21] N. J. Higham. **Accuracy and Stability of Numerical Algorithms**. SIAM, Philadelphia, 1996.

[22] R. A. Horn and C. R. Johnson. **Matrix Analysis**. Cambridge University Press, Cambridge, 1985.

[23] R. A. Horn and C. R. Johnson. **Topics in Matrix Analysis**. Cambridge University Press, Cambridge, 1991.

[24] A. Iserles. **A First Course in the Numerical Analysis of Differential Equations**. Cambridge University Press, London, 1996.

[25] W. Joubert. **A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems**. *SIAM J. Sci. Comput*, 15(2): pp. 427–439, March 1994.

[26] M. A. Kowolski, K. A. Sikorski, and F. Stenger. **Selected Topics in Approximation and Computation**. Oxford University Press, New York, 1995.

[27] S. Kurepa. **Konačno dimenzionalni vektorski prostori i primjene**. Sveučilišna naklada Liber, Zagreb, 1979.

[28] S. Kurepa. **Funkcionalna Analiza, elementi teorije operatora**. Školska knjiga, Zagreb, 1990.

[29] N. M. Nachtigal, S. Reddy, and L. N. Trefethen. **How fast are nonsymmetric matrix iterations?** *SIAM J. Matrix Anal. Appl.*, 13(3): pp. 778–795, July 1992.

[30] B. N. Parlett. **The Symmetric Eigenvalue Problem**. SIAM, Philadelphia, 1998.

[31] B. N. Parlett, D. R. Taylor, and Z. A. Liu. **A look-ahead Lanczos algorithm for unsymmetric matrices**. *Mathematics of Computation*, 44(169): pp. 105–124, January 1985.

[32] Y. Saad. **Iterative Methods for Sparse Linear Systems**, 2000.

[33] J. R. Shewchuk. **An Introduction to the Conjugate Gradient Method Without the Agonizing Pain**. School of Computer Science Carnegie Mellon University, Pittsburgh, August 1994.

[34] Z. Stojaković i D. Herceg. **Numeričke metode linearne algebre**. IRO Građevinska knjiga, Beograd, 1985.

[35] K. C. Toh. **GMRES vs. ideal GMRES**. *SIAM J. Matrix Anal. Appl.*, 18(1): pp. 30–36, January 1997.

[36] H. A. van der Vorst. **Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems**. *SIAM J. Sci. Stat. Comput.*, 13(2): pp. 631–644, March 1992.

[37] R. S. Varga. **Matrix Iterative Analysis**. Prentice–Hall, Englewood, 1962.