Računanje i analiza matrične funkcije predznaka

Stopić, Petra

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI
FAKULTET
MATEMATIČKI ODSJEK

Petra Stopić

RAČUNANJE I ANALIZA MATRIČNE FUNKCIJE PREDZNAKA

Diplomski rad

Zagreb, 2016.

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI
FAKULTET
MATEMATIČKI ODSJEK

Petra Stopić

RAČUNANJE I ANALIZA MATRIČNE FUNKCIJE PREDZNAKA

Diplomski rad

Voditelj rada:
Doc.dr.sc. Nela Bosner

Zagreb, 2016.

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

1._____, predsjednik

2._____, član

3._____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1._____

2._____

3._____

Sadržaj

Uvod

Ovaj rad uključuje teoriju matrica, numeričku analizu, teoriju aproksimacija i razvoj algoritama. Pojam 'matrična funkcija' koji je korišten u narednom tekstu odnosi se na funkciju f koja uzima kvadratnu matricu $A \in \mathbf{C^{n\times n}}$ te $f(A)$ predstavlja matricu jednakih dimenzija kao što je i matrica A. Razne matrične funkcije uz pomoć matrične teorije, numeričke matematike te raznih algoritama koji doprinose računanju funkcije neke matrice, korisne su, ne samo za teoriju matrica, nego i u drugim primjenama.

Tema ovog rada je usko vezana samo uz jednu matričnu funkciju, a to je funkcija predznaka ili *sign* funkcija. Definirati matričnu sign funkciju može se na više načina. Jedan od najčešćih načina je onaj koji koriste razni programerski jezici - elementarni pristup, tj. primjenjuje se funkcija na svaki element matrice. Nadalje, postoje varijante kada matrična funkcija daje skalar kao rezultat, npr. trag matrice, uvjetovanost, determinanta ili pristup funkciji predznaka matrice kojoj je kodomena opet matrica ali nije izvedena pomoću neke skalarne funkcije. U daljnjem tekstu definiran je drugačiji način od svih navedenih koji se bazira na svojstvenim vrijednostima.

U teoriji kontrole funkcija predznaka primjenjuje se za rješavanje Lyapunove jednadžbe te Riccatijeve algebarske jednadžbe. Za oba rješenja je potrebno grupirati vrijednosti po tome nalaze li se lijevo ili desno od imaginarne osi u kompleksnoj ravnini što ćemo vidjeti u narednom tekstu da je usko vezano uz *sign* funkciju. Također se pomoću matrične *sign* funkcije može izbrojati koliko ima svojstvenih vrijednosti matrice u lijevom području kompleksne ravnine u odnosu na imaginarnu os, odnosno u desnom:

$$p = \frac{1}{2}(n - tr(sign(A))),\ q = \frac{1}{2}(n + tr(sign(A)))$$

gdje je p broj svojstvenih vrijednosti u lijevoj, q u desnoj poluravnini, n je dimenzija matrice A. U teorijskoj fizici čestica kod Diracovog operatora je potrebno računanje sustava jednadžbi koje sadrže *sign* funkciju:

$$(G - sign(H))x = b$$

gdje je $G = diag(\pm 1)$, a H kompleksna hermitska matrica.

U prvom poglavlju nalaze se sve definicije i korolari potrebni za daljnje razumjevanje rada. Drugo poglavlje sadrži općenitu teoriju matričnih funkcija koja će se kasnije primjenjivati na funkciju predznaka. Navode se dvije potrebne definicije matričnih funkcija, svojstva, uvjetovanost funkcija te neke metode. Nadalje, rad se ograničava na *sign* funkciju u trećem poglavlju u kojem detaljno objašnjava metode računanja funkcije predznaka. Za više numeričkih algoritama dala bi se analiza točnosti, stabilnosti i složenosti. Također se daje uvid u osjetljivost ovog problema baziran na Fréchetovoj derivaciji.

1 Potrebne definicije i korolari

Definicija 1.1 Matrica $A \in \mathbf{C^{n\times n}}$ je *regularna* ako postoji matrica $B \in \mathbf{C^{n\times n}}$ za koju vrijedi:

$$AB = BA = I,$$

gdje je I jedinična matrica. Ako postoji takva matrica, ona je jedinstvena i zove se *inverzna matrica* matrice A. Matrica je *singularna* ako nije regularna.

Definicija 1.2 *Svojstvena vrijednost* matrice $A \in \mathbf{C^{n\times n}}$ je $\lambda \in \mathbf{C}$ ako postoji vektor $x \in \mathbf{C^n}, x \neq 0$ t.d. vrijedi:

$$Ax = x.$$

Skup svih svojstvenih vrijednosti od A naziva se *spektar matrice* A (oznaka: $\sigma(A)$). *Spektralni radijus* matrice $A \in \mathbf{C^{n\times n}}$ je

$$\rho(A) := \max_{\lambda\in\sigma(A)} |\lambda|.$$

Definicija 1.3 *Karakteristični polinom* matrice $A \in \mathbf{C^{n\times n}}$ je

$$k_A(\lambda) = det(A - \lambda I).$$

Dakle, nultočke karakterističnog polinoma su svojstvene vrijednosti matrice A. Kratnost λ u tom polinomu je *algebarska kratnost* svojstvene vrijednosti λ. *Geometrijska kratnost* je dimenzija potprostora $Ker(A - \lambda I)$, gdje je Ker oznaka za jezgru.

Definicija 1.4 *Minimalni polinom* matrice $A \in \mathbf{C^{n\times n}}$ je jedinstveni normirani polinom p najmanjeg stupnja za koji vrijedi $p(A) = 0$.

Definicija 1.5 Pretpostavimo da su $\lambda_1, \lambda_2, ..., \lambda_d$ različite svojstvene vrijednosti matrice A, a n_k dimenzija najvećeg Jordanovog bloka u kojem se nalazi λ_k. Tada se n_k naziva *indeksom* svojstvene vrijednosti λ_k. Funkcija f je *definirana na spektru* matrice A ako postoje vrijednosti $f^{(p)}(\lambda_k)$, za sve $p = 0, ..., n_k - 1$ te za sve $k = 1, ..., d$.

Definicija 1.6 *Hermitski adjungirana matrica* $A^* \in \mathbf{C^{n\times n}}$ matrici A je matrica čiji su elementi

$$a_{ij} = \bar{a}_{ji},\ i, j = 1, ..., n \tag{1.1}$$

Normalna matrica $A \in \mathbf{C^{n\times n}}$ je kompleksna matrica za koju vrijedi:

$$AA^* = A^*A \tag{1.2}$$

Unitarna matrica $A \in \mathbf{C^{n\times n}}$ je kompleksna matrica za koju vrijedi:

$$AA^* = A^*A = I \tag{1.3}$$

Definicija 1.7 Matrica $B \in \mathbf{C^{n\times n}}$ je *slična matrica* matrici $A \in \mathbf{C^{n\times n}}$ ako vrijedi za neku regularnu matricu Z:

$$B = Z^{-1}AZ.$$

Definicija 1.8 *Matrična norma* na $\mathbf{C}^{m\times n}$ je funkcija $||\cdot|| : \mathbf{C}^{m\times n} \to \mathbf{R}$ koja zadovoljava slijedeće uvjete:

1. $||A|| \geq 0$.
2. $||A|| = 0 \iff A = 0$.
3. $||\alpha A|| = |\alpha|\ ||A||$ za sve $\alpha \in \mathbf{R}$, $A \in \mathbf{C}^{m\times n}$.
4. $||A + B|| \leq ||A|| + ||B||$ za sve $A, B \in \mathbf{C}^{m\times n}$.

Definicija 1.9 Matrična norma $||\cdot||$ na $\mathbf{C^{n\times n}}$ je *konzistentna* ako za sve A i B kvadratne kompleksne matrice vrijedi:

$$||AB|| \leq ||A||\,||B||$$

Za svaku konzistentnu matričnu normu postoji vektorska norma ν koja je konzistentna sa tom matričnom normom $||\cdot||$:

$$\nu(Ax) \leq ||A||\nu(x)$$

Definicija 1.10 Neka je dana neka vektorska norma ν na $\mathbf{C^n}$. Odgovarajuća *operatorska* norma je definirana kao:

$$||A|| := \max_{x \neq 0} \frac{\nu(Ax)}{\nu(x)}$$

Definicija 1.11 *Frobeniusova matrična norma* je $|| \cdot ||_F : \mathbf{C^{n \times n}} \to \mathbf{R}$,

$$||A||_F := \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{tr(A^*A)} \tag{1.4}$$

Spektralna matrična norma je $|| \cdot ||_2 : \mathbf{C^{n \times n}} \to \mathbf{R}$,

$$||A||_2 := \sqrt{\rho(A^*A)} \tag{1.5}$$

Matrična norma ∞ je $|| \cdot ||_\infty : \mathbf{C^{n \times n}} \to \mathbf{R}$,

$$||A||_\infty := \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| \tag{1.6}$$

Definicija 1.12 Neka je $A \in \mathbf{C^{nxn}}$. Korijeni svojstvenih vrijednosti matrice $A^*A \in \mathbf{C^{nxn}}$ zovu se *singularne vrijednosti* $\sigma_1, ..., \sigma_n$ od A. *Lijevi* u_i *i desni singularni vektor* v_i za $i = 1, \ldots, n$ zadovoljavaju:

$$Av_i = \sigma_i u_i$$

$$A^*u_i = \sigma_i v_i$$

Definicija 1.13 Ako je $B := o(\,||A||\,)$, onda $\lim_{||A|| \to 0} \frac{||B||}{||A||} = 0$

Definicija 1.14 Ako je $B := O(\,||A||\,)$, onda vrijedi:

$$||B|| \leq c||A||,$$

za neku konstantu c kada $||A|| \to 0$.

Definicija 1.15 Funkcija f je *idempotentna* ako vrijedi $(f \circ f)(x) = f(x)$.

Definicija 1.16 Ako je $z = x + iy$ kompleksni broj, onda je njegov zapis u *polarnim koordinatama* $z := re^{i\alpha} = r(\cos\alpha + i\sin\alpha)$, gdje je:

$$\tan\alpha = \frac{y}{x}$$

$$r^2 = x^2 + y^2$$

Definicija 1.17 Ako je $z = r(\cos\alpha + i\sin\alpha)$ polarni oblik kompleksnog broja, tada vrijedi:

$$\cos 2\alpha = (\cos\alpha)^2 - (\sin\alpha)^2$$

$$\sin 2\alpha = 2\sin\alpha\cos\alpha$$

$$z^n = r^n(\cos n\alpha + i\sin n\alpha)$$

$$z^{1/n} = r^{1/n}\left(\cos\frac{\alpha+2k\pi}{n} + i\sin\frac{\alpha+2k\pi}{n}\right),\ k = 0,\ 1,\ldots,\ n-1$$

Definicija 1.18 *Cayleyova metrika* definira udaljenost nekog kompleksnog broja x od $sign(x)$ kao udaljenost $\dfrac{x - sign(x)}{x + sign(x)}$ od ishodišta, tj.:

$$C(x, sign(x)) := \begin{cases} \left|\dfrac{x-1}{x+1}\right|,\ Rex > 0 \\ \left|\dfrac{x+1}{x-1}\right|,\ Rex < 0 \end{cases} = \begin{cases} \sqrt{\dfrac{(Rex-1)^2 + (Imx)^2}{(Rex+1)^2 + (Imx)^2}},\ Rex > 0 \\ \sqrt{\dfrac{(Rex+1)^2 + (Imx)^2}{(Rex-1)^2 + (Imx)^2}},\ Rex < 0 \end{cases}$$

Definicija 1.19 Taylorov red oko 0 za funkciju $f(x) = (1-x)^\alpha$, $\alpha \in \mathbf{C}$ je:

$$(1-x)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} (-1)^n x^n, \quad |x| < 1$$

gdje je:

$$\binom{\alpha}{0} := 1,\ \binom{\alpha}{n} = \frac{\alpha(\alpha-1)\ldots(\alpha-n+1)}{n!}$$

Korolar 1.20 *$A \in \mathbf{C^{n\times n}}$ je regularna matrica ako i samo ako 0 nije svojstvena vrijednost od A.*

Dokaz:

A singularna $\iff$ jezgra od A nije trivijalna $\iff$ postoji $x \neq 0$ t.d. $Ax = 0 \iff Ax = 0x$, $x \neq 0 \iff 0$ je svojstvena vrijednost

□

Korolar 1.21 *Ako su A i $B \in \mathbf{C^{n\times n}}$ slične matrice, onda one imaju jednake svojstvene vrijednosti*

Dokaz:

$$\begin{aligned} det(A - \lambda I) &= det(Z^{-1}BZ - \lambda I) = \\ &= det(Z^{-1}(B - \lambda I)Z) = \\ &= (Binet - Cauchyev\, teorem) = \\ &= det(Z^{-1})\, det(B - \lambda I)\, det(Z) = \\ &= det(Z^{-1})\, det(Z)\, det(B - \lambda I) = \\ &= det(B - \lambda I) \end{aligned}$$

□

Korolar 1.22 *Ako je $A \in \mathbf{C^{n\times n}}$ sa svojstvenim vrijednostima $\lambda_1, \ldots, \lambda_n$, onda vrijedi:*

$$det(A) = \prod_{i=1}^{n} \lambda_i$$

Dokaz:

Iz Definicije 1.3 znamo da su nultočke karakterističnog polinoma matrice A svojstvene vrijednosti od A, tj. $k_A(\lambda) = \prod_{i=1}^{n} (\lambda - \lambda_i)$.

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

$$\prod_{i=1}^{n} (\lambda - \lambda_i) = \begin{vmatrix} \lambda - a_{11} & -a_{12} & \ldots & -a_{1n} \\ -a_{21} & \lambda - a_{22} & \ldots & -a_{2n} \\ \vdots & \ldots & \ddots & \vdots \\ -a_{n1} & -a_{n2} & \ldots & \lambda - a_{nn} \end{vmatrix}$$

Kao slobodni član s lijeve strane je $\prod_{i=1}^{n} \lambda_i$, a s desne $det(A)$ što je jednako zbog jednakosti dva polinoma.

□

Korolar 1.23 *Za proizvoljnu konzistentnu matričnu normu* $||\cdot||$ *i spektralni radijus vrijedi nejednakost:*

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

Dokaz:

Ako je λ svojstvena vrijednost od A, tada postoji $x \neq 0$ t.d. $Ax = \lambda x$. Unutar Definicije 1.9 je jedan rezultat koji ovdje koristimo:

$$\nu(Ax) = |\lambda|\,\nu(x) \;\; i \;\; \nu(Ax) \leq ||A||\,\nu(x) \quad \Rightarrow \quad |\lambda|\,\nu(x) \leq ||A||\,\nu(x)$$

Budući x nije 0, možemo podjeliti sa $\nu(x)$ te dobimo: $|\lambda| \leq ||A||$, što vrijedi za svaku svojstvenu vrijednost pa tako i za onu najveću što je upravo $\rho(A)$.

□

2 TEORIJA MATRIČNIH FUNKCIJA

2.1 Definicije matrične funkcije

Matrična funkcija je definirana sa: $f : \mathbf{C^{n\times n}} \to \mathbf{C^{n\times n}}$.

U ovom radu se koriste dvije specifične definicije matričnih funkcija. Prva je preko Jordanove kanonske forme matrice, dok je druga definicija matrične funkcije pomoću interpolacijskog polinoma.

2.1.1 Jordanova forma

Ova definicija matrične funkcije se bazira na rezultatu da se svaka matrica $A \in \mathbf{C^{n\times n}}$ može izraziti u Jordanovoj kanonskoj formi J ($A = ZJZ^{-1}$).

$$J = Z^{-1}AZ = diag(J_1, \ldots, J_p), \tag{2.1}$$

gdje je Z regularna matrica, $m_1 + m_2 + \cdots + m_p = n$ te $J_k \in \mathbf{C^{m_k x m_k}}$:

$$J_k(\lambda_k) = \begin{bmatrix} \lambda_k & 1 & 0 & \ldots & 0 \\ 0 & \lambda_k & 1 & \ldots & 0 \\ \vdots & \ldots & \ddots & \ldots & \vdots \\ 0 & 0 & \ldots & \ddots & 1 \\ 0 & 0 & \ldots & 0 & \lambda_k \end{bmatrix}$$

Jordanova matrica J je jedinstvena do na redoslijed blokova J_k, a Z nije jedinstvena.

Zahtijevamo da je funkcija matrice A definirana na spektru (Definicija

1.5) od A. Definiramo funkciju matrice A, $f : \mathbf{C^{n\times n}} \to \mathbf{C^{n\times n}}$, pomoću Jordanove forme (2.1):

$$f(A) := Zf(J)Z^{-1} = Zdiag(f(J_1), \ldots, f(J_p))Z^{-1}, \tag{2.2}$$

gdje je svaka $f(J_k) \in \mathbf{C^{m_k x m_k}}$ u obliku:

$$f(J_k(\lambda_k)) = \begin{bmatrix} f(\lambda_k) & f'(\lambda_k) & \frac{f"(\lambda_k)}{2} & \ldots & \frac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ 0 & f(\lambda_k) & f'(\lambda_k) & \ldots & \frac{f^{(m_k-2)}(\lambda_k)}{(m_k-2)!} \\ \vdots & \ldots & \ddots & \ldots & \vdots \\ 0 & 0 & 0 & \ddots & f'(\lambda_k) \\ 0 & 0 & 0 & \ldots & f(\lambda_k) \end{bmatrix}$$

Ako je A dijagonalizabilna matrica kojoj su sve svojstvene vrijednosti različite, tj.oblika $A = ZDZ^{-1}$, onda je pripadna Jordanova forma (2.1) dekompozicija gdje je D dijagonalna matrica sa svojstvenim vrijednostima na dijagonali, a Z je matrica čiji su stupci svojstveni vektori od tih svojstvenih vrijednosti. Po (2.2) vrijedi da je

$$f(A) = ZDZ^{-1} = Zdiag(f(\lambda_1), \ldots, f(\lambda_d))Z^{-1}.$$

Očito je da matrica $f(A)$ ima iste svojstvene vektore kao A te svojstvene vrijednosti jednake vrijednostima funkcije f u svojstvenim vrijednostima od A.

Pomoću gornje definicije matrice dokazujemo idući korolar koji ćemo koristiti u narednim poglavljima.

Korolar 2.1 *Ako je $\rho(A) < 1$ za neku kompleksnu matricu A, tada vrijedi* $\lim_{i\to\infty} A^i = 0$.

Dokaz:

U Jordanovoj formi je $A = ZJZ^{-1}$ pa je $A^i = ZJ^iZ^{-1}$. K-ti Jordanov

blok dimenzije $m_k \times m_k$ na potenciju n je u obliku:

$$J_k^n = \begin{bmatrix} \lambda_k^n & n\,\lambda_k^{n-1} & \dots & \frac{n\cdots(n-m_k+2)}{(m_k-1)!}\lambda_k^{n-m_k+1} \\ 0 & \lambda_k^n & \dots & \frac{n\cdots(n-m_k+3)}{(m_k-2)!}\lambda_k^{n-m_k+2} \\ \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \ddots & n\,\lambda_k^{n-1} \\ 0 & 0 & \dots & \lambda_k^n \end{bmatrix}$$

Po pretpostavci teorema sve λ_k su po apsolutnoj vrijednosti manje od 1 pa kada ih potenciramo na n idu prema 0 za veliki n, puno veći od k.

□

2.1.2 Interpolacija polinomom

Dimenzija prostora kompleksnih matrica je n^2. Promotrimo niz matrica:

$$I,\ A,\ A^2, \dots,\ A^{n^2}.$$

Budući da je dimenzija prostora n^2, taj niz je zavisan. Zbog $I \neq 0$ vrijedi da je niz I nezavisan. Dakle, postoji neki $1 \leq m \leq n^2$ takav da je niz matrica

$$I,\ A,\ A^2, \dots,\ A^m$$

zavisan, a niz

$$I,\ A,\ A^2, \dots,\ A^{m-1}$$

linearno nezavisan. Odnosno, A^m je linearna kombinacija ostalih. Tada vrijedi,

$$-\alpha_0 I - \alpha_1 A + \dots - \alpha_{m-1}A^{m-1} + \alpha_m A^m = 0$$

pa podjelivši sa $\alpha_m \neq 0$, dobivamo polinom kojeg A poništava:

$$p(X) = X^m - \dots - \beta_1 X - \beta_0 I, \quad \beta_i = \frac{\alpha_i}{\alpha_m}$$

Teorem 2.2 *Neka je $A \in C^{n\times n}$. Tada je $p(X)$ iz dijela iznad teorema minimalni polinom matrice A. Vrijedi također: ako je $f(X)$ neki drugi polinom kojeg poništava matrica A, tada p dijeli f.*

Dokaz:

Kada bi postojao neki drugi netrivijalni polinom $f(X)$ kojeg A poništava stupnja $r < m$, tada bi vrijedilo

$$f(A) = \beta_0 I + \beta_1 A + \cdots + \beta_r A^r = 0.$$

Budući da je m odabran tako da je niz od stupnja m zavisan, a do stupnja $m-1$ nezavisan, gornji niz mora biti različit od 0 kao podniz nezavisnog niza, tj. dolazi do kontradikcije zbog izbora broja m. Jedinstvenost vrijedi jer ako bi uzeli neki drugi minimalni polinom $f(X)$ različit od p, vrijedilo bi da je polinom $f(X) - p(X) \neq 0$ polinom stupnja manjeg od m (minimalni polinom je normiran) kojeg A poništava, što je kontradikcija kao i u prvom dijelu.

Ako A poništava f, tada po prvom dijelu vrijedi da je f stupnja većeg ili jednakog m, te se dijeljenjem polinoma f sa p dobiva $f(X) = p(X)q(X) + r(X)$. Kada uvrstimo A, dobivamo $f(A) = p(A)q(A) + r(A)$. Budući da A poništava f i p, zaključujemo da poništava i r, a to može vrijediti samo za $r \equiv 0$ jer je p minimalni polinom i $deg(r) < deg(p)$. Dakle, p dijeli f.

□

Neka je f funkcija definirana na spektru od A. Različite svojstvene vrijednosti od A su $\lambda_k,\ k = 1, \ldots, d$. Uz oznaku n_k za dimenziju najvećeg Jordanovog bloka $k-te$ svojstvene vrijednosti, minimalni polinom od A je:

$$p(t) = \prod_{k=1}^{d} (t - \lambda_k)^{n_k} \tag{2.3}$$

Slijedeći teorem govori o svojstvu polinoma da je matrica $f(A)$ u potpunosti određena vrijednostima polinoma f na spektru od A.

Teorem 2.3 *Neka su z i q polinomi te $A \in C^{n\times n}$. Vrijedi: $z(A) = q(A)$ ako i samo ako z i q poprimaju iste vrijednosti na spektru od A.*

Dokaz:

$\Rightarrow$ Neka vrijedi da je $z(A) = q(A)$. Tada je $b(A) = z(A) - q(A) = 0$ polinom poništen matricom A. Po Teoremu 2.2 vrijedi da onda minimalni polinom p dijeli b. Iz (2.3) je očito da je minimalni polinom p jednak 0 na spektru od A. Tada i b poprima vrijednosti 0 na spektru (jer p dijeli b), tj z i q poprimaju iste vrijednosti na spektru.

$\Leftarrow$ Neka z i q poprimaju iste vrijednosti na spektru. Definiramo $b := z - q$. Tada je b jednak 0 na spektru od A pa ga p mora dijeliti zbog gornjih tvrdnji. Vrijedi: $b(A) = p(A)h(A) = 0$ za neki polinom h. Tada je $z(A) = q(A)$.

□

Sada možemo definirati matričnu funkciju na još jedan način.

$$f(A) := h(A), \tag{2.4}$$

gdje je h polinom za koji vrijedi

$$deg(h) < deg(p) = \sum_{k=1}^{d} n_k$$

uz uvjet interpolacije

$$h^{(l)}(\lambda_k) = f^{(l)}(\lambda_k),\ l = 0,\ 1, \ldots,\ n_k - 1,\ k = 1, \ldots,\ d. \tag{2.5}$$

Postoji jedinstven takav polinom h i naziva se Hermiteov interpolacijski polinom. Hermiteov interpolacijski polinom može se zapisati u Newtonovoj bazi. Čvorovi interpolacije su u ovom slučaju svojstvene vrijednosti, i svaka svojstvena vrijednost λ_k je n_k-struki čvor.

Podijeljena razlika u n_k-strukom čvoru:

$$f\left[\lambda_k, \ldots,\ \lambda_k\right] = \frac{f^{(n_k-1)}(\lambda_k)}{(n_k - 1)!}$$

Inače, vrijedi rekurzija:

$$f\left[\lambda_k,\ \lambda_{k+1}, \ldots,\ \lambda_{k+j}\right] = \frac{f\left[\lambda_{k+1}, \ldots,\ \lambda_{k+j}\right] - f\left[\lambda_k, \ldots,\ \lambda_{k+j-1}\right]}{\lambda_{k+j} - \lambda_k}$$

Hermiteov interpolacijski polinom je u obliku:

$$h(t) = \sum_{k=0}^{n_1-1} f_1^k (t - \lambda_1)^k + \sum_{k=0}^{n_2-1} f_{12}^k (t - \lambda_1)^{n_1} (t - \lambda_2)^k + \cdots + \sum_{k=0}^{n_d-1} f_{1d}^k (t - \lambda_1)^{n_1} \ldots (t - \lambda_d)^k$$

gdje su:

$$f_1^k := f\left[\lambda_1, \ldots, \lambda_1\right], k * \lambda_1$$

$$f_{12}^k := f\left[\lambda_1, \ldots, \lambda_1, \lambda_2, \ldots, \lambda_2\right], n_1 * \lambda_1, \; k * \lambda_2$$

$$f_{1j}^k := f\left[\lambda_1, \ldots, \lambda_1, \ldots, \lambda_j, \ldots, \lambda_j\right], \; n_1 * \lambda_1, \ldots, \; n_{j-1} * \lambda_{j-1}, \; k * \lambda_j$$

Ako bi računali na taj način funkciju neke matrice, to baš i ne bi bilo praktično zbog dva razloga. Prvi je taj da je zahtjeva $O(n)$ množenja matrica dok množenje dvije kvadratne matrice zahtjeva $O(n^3)$ operacija. Na kraju je to $O(n^4)$ operacija da bi dobili f(A), čak ako je h jednočlan polinom ili u obliku podijeljenih razlika, dok ostale metode imaju složenost $O(n^3)$. Drugi razlog je neizvjesna numerička stabilnost te kompliciranа analiza veličine pogrešaka koje bi opravdale tu metodu pomoću polinoma interpolacije.

Teorem 2.4 *Definicije (2.2) pomoću Jordanove kanonske forme i (2.4) pomoću Hermiteova polinoma su ekvivalentne.*

Dokaz:

Iz (2.4) slijedi da $f(A) = h(A)$ gdje je h Hermiteov interpolacijski polinom koji zadovoljava (2.5). Ako je A u Jordanovoj formi (2.1), onda zbog osnovnih svojstava matričnih potencija u polinomu vrijedi

$$f(A) = h(A) = h(ZJZ^{-1}) = Zh(J)Z^{-1} = Zdiag(h(J_k))Z^{-1}.$$

Za najjednostavniji polinom $h(x) = x^n$ vrijedi:

$$h(J_k) = J_k^n =$$

$$= \begin{bmatrix} \lambda_k^n & n\,\lambda_k^{n-1} & \dots & \frac{n\cdots(n-m_k+2)}{(m_k-1)!}\lambda_k^{n-m_k+1} \\ 0 & \lambda_k^n & \dots & \frac{n\cdots(n-m_k+3)}{(m_k-2)!}\lambda_k^{n-m_k+2} \\ \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \ddots & n\,\lambda_k^{n-1} \\ 0 & 0 & \dots & \lambda_k^n \end{bmatrix} =$$

$$= \begin{bmatrix} h(\lambda_k) & h'(\lambda_k) & \frac{h"(\lambda_k)}{2} & \dots & \frac{h^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ 0 & h(\lambda_k) & h'(\lambda_k) & \dots & \frac{h^{(m_k-2)}(\lambda_k)}{(m_k-2)!} \\ \vdots & \dots & \ddots & \dots & \vdots \\ 0 & 0 & 0 & \ddots & h'(\lambda_k) \\ 0 & 0 & 0 & \dots & h(\lambda_k) \end{bmatrix}$$

Gornja jednakost vrijedi i za proizvoljan polinom pa i za Hermiteov s uvjetima interpolacije (2.5). Zbog tih uvjeta zaključujem $h(J_k) = f(J_k)$.

□

Teorem 2.5 o svojstvima matrične funkcije

Neka je $A \in C^{n\times n}$ te f matrična funkcija definirana na spektru od A. Vrijede slijedeća svojstva:

1. $f(A)A = Af(A)$
2. $f(X^{-1}AX) = X^{-1}f(A)X$, *$X$ kvadratna kompleksna regularna matrica*
3. *Ako je λ svojstvena vrijednost od A, onda je $f(\lambda)$ svojstvena vrijednost od $f(A)$.*
4. *Ako je $A = (A_{ij})$ blok trokutasta matrica, tada je $F := f(A)$ blok trokutasta matrica sa istom strukturom blokova kao A i $F_{ii} = f(A_{ii})$.*

Dokaz:

Svojstvo komutativnosti (1) slijedi iz (2.4) što implicira da je $f(A)$ polinom potencija od A:

$$f(A)A = h(A)A = Ah(A) = Af(A).$$

Svojstva (2) i (3) slijede direktno iz (2.2). Koristeći ponovno (2.4), $f(A) = h(A)$ je očito trokutasta blok matrica jer je potencija trokutaste matrice opet trokutasta matrica. I-ti dijagonalni blok je $h(A_{ii})$. Budući da h interpolira f na spektru od A, onda interpolira f i na spektru od svake A_{ii} jer je $\sigma(A) = \cup_i(A_{ii})$ pa je zbog definicije funkcije pomoću interpolacijskog polinoma (2.4) i uvjeta interpolacije (2.5), $h(A_{ii}) = f(A_{ii})$. Time je dokazano i zadnje svojstvo.

□

2.2 Fréchetova derivacija

Definicija 2.6 Za funkciju $f : \Omega \subseteq X \to Y$ kažemo da je Fréchet derivabilna u točci $x \in \Omega$ ako postoji linearan neprekidan operator $Df(x) \in \mathcal{L}(X, Y)$, takav da vrijedi:

$$\lim_{h\to 0} \frac{|| f(x+h) - f(x) - (D(f(x))(h) ||_Y}{|| h ||_X} = 0$$

Slijedi: $f(x+h) - f(x) - (Df(x))(h) = o(|| h ||_X)$.
Analogno se definira u slučaju matrica.

Definicija 2.7 Za linearni operator $L_f(A) := Df(A)$ iz $\mathcal{L}(\mathbf{C^{n\times n}}, \mathbf{C^{n\times n}})$ vrijedi:

$$f(X+H) - f(X) - L_f(X, H) = o(|| H ||) \tag{2.6}$$

$L_f(X, H)$ je *Fréchetova derivacija* u X primjenjena na matricu H ili u smjeru H.

Kada se želimo orijentirati na matricu u kojoj gledamo derivaciju, a ne vrijednost matrice u nekom smjeru, oznaka je $L_f(X)$ ili češće $L(X)$.
Norma Fréchetove derivacije je definirana sa:

$$||L(X)|| := \max_{Z\neq 0} \frac{||L(X,Z)||_F}{||Z||_F} \tag{2.7}$$

Teorem 2.8 *Ako su f i g funkcije koje su Fréchet derivabilne u X, tada vrijedi da su $f+g$, fg i $f \circ g$ Fréchet derivabilne u X i vrijedi:*

1. *Pravilo sume:*

$$L_{f+g}(X,H) = L_f(X,H) + L_g(X,H) \tag{2.8}$$

2. *Pravilo produkta:*

$$L_{fg}(X,H) = L_f(X,H)g(X) + f(X)L_g(X,H) \tag{2.9}$$

3. *Lančano pravilo: Ovdje zahtjevamo da je f Fréchet derivabilna u $g(X)$.*

$$L_{f\circ g}(X,H) = L_f(g(X), L_g(X,H)) \tag{2.10}$$

Dokaz:

Pravilo sume proizlazi odmah iz Definicije 2.7. Pravilo produkta slijedi uz korištenje (2.6):

$$\begin{aligned}(fg)(X+H) &= f(X+H)g(X+H) = \\ &= (f(X) + L_f(X,H) + o(||H||))\ (g(X) + L_g(X,H) + o(||H||)) = \\ &= (fg)(X) + L_f(X,H)g(X) + f(X)L_g(X,H) + o(||H||)\end{aligned}$$

Lančano pravilo uz dvostruku upotrebu (2.6):

$$\begin{aligned}(f\circ g)(X+H) - (f\circ g)(X) &= f(g(X+H)) - f(g(X)) = \\ &= f(g(X) + L_g(X,H) + o(||H||)) - f(g(X)) = \\ = f(g(X)) + L_f(g(X), L_g(X,H) + o(||H||)) + o(||H||) - f(g(X)) &= \\ &= L_f(g(X), L_g(X,H)) + o(||H||)\end{aligned}$$

□

2.3 Uvjetovanost matrične funkcije

Osjetljivost matričnih funkcija na perturbacije podataka je mjerena koeficijentom ili brojem uvjetovanosti. Ovo poglavlje će pokazati kako definirati koeficijente uvjetovanosti te kako ih učinkovito procijeniti. Oni mogu biti

izraženi u normi Fréchetove derivacije pa ćemo brojeve uvjetovanosti upoznavati kroz svojstva Fréchet-ove derivacije.

Standardna definicija relativnog broja uvjetovanosti za skalarnu funkciju $f : \mathbf{R} \to \mathbf{R}$ je:

$$cond_{rel}(f, x) := \lim_{\varepsilon \to 0} \sup_{|\Delta x| \leq \varepsilon |x|} \left| \frac{f(x + \Delta x) - f(x)}{\varepsilon f(x)} \right|$$

Objašnjenje gornjeg izraza je da taj broj mjeri koliko jako male promjene u podacima povećavaju promjenu funkcijske vrijednosti, u slučaju kad su obje promjene mjerene u relativnom smislu.

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

$$\Rightarrow f(x + \Delta x) - f(x) - \Delta x f'(x) = o(\Delta x)$$

$$\Rightarrow \frac{f(x + \Delta x) - f(x)}{f(x)} = \frac{f'(x)\Delta x}{f(x)} \frac{x}{x} + o(\Delta x)$$

$$\Rightarrow \left| \frac{f(x + \Delta x) - f(x)}{f(x)} \right| = \left| \left(\frac{f'(x)x}{f(x)} \right) \frac{\Delta x}{x} + o(\Delta x) \right|$$

$$\Rightarrow \left| \frac{f(x + \Delta x) - f(x)}{f(x)} \right| \leq \left| \frac{f'(x)x}{f(x)} \right| \left| \frac{\Delta x}{x} \right| + o(\Delta x) \tag{2.11}$$

Zbog svojstva supremuma uz uvjet $\frac{|\Delta x|}{|x|} \leq \varepsilon$, slijedi

$$\left| \frac{f(x + \Delta x) - f(x)}{f(x)\varepsilon} \right| \leq \left| \frac{f'(x)x}{f(x)} \right| \frac{\varepsilon}{\varepsilon} + o(\Delta x)$$

pa za relativni broj uvjetovanosti uz $o(\Delta x) = o(\varepsilon)$ za $\sup_{|\Delta x| \leq \varepsilon |x|}$ vrijedi:

$$cond_{rel}(f, x) = \lim_{\varepsilon \to 0} \sup_{|\Delta x| \leq \varepsilon |x|} \left| \left(\frac{f'(x)x}{f(x)\varepsilon} \right) \frac{\Delta x}{x} + \frac{o(\Delta x)}{\varepsilon} \right| = \left| \frac{f'(x)x}{f(x)} \right| \tag{2.12}$$

Definicija 2.9 *Relativni broj uvjetovanosti* matrične funkcije f uz bilo koju matričnu normu te perturbacije matrice X koje označavamo sa H je:

$$cond_{rel}(f, X) := \lim_{\varepsilon \to 0} \sup_{||H|| \leq \varepsilon ||X||} \frac{|| f(X + H) - f(X) ||}{|| \varepsilon f(X) ||} \tag{2.13}$$

Primjenivši matrice na (2.11) i (2.12) slijedi granica približne perturbacije H:

$$\frac{||f(X+H)-f(X)||}{||f(X)||} \leq cond_{rel}(f,X)\frac{||H||}{||X||} + o(||H||) \tag{2.14}$$

Definicija 2.10 *Apsolutni broj uvjetovanosti* matrične funkcije f uz bilo koju matričnu normu te perturbacije matrice X koje označavamo sa H je:

$$cond_{abs}(f,X) := \lim_{\varepsilon \to 0} \sup_{||H|| \leq \varepsilon} \frac{||\, f(X+H)-f(X)\,||}{\varepsilon} \tag{2.15}$$

Relativni i apsolutni brojevi uvjetovanosti razlikuju se konstantom:

$$\begin{aligned}
cond_{rel}(f,X) &= \lim_{\varepsilon \to 0} \sup_{||H|| \leq \varepsilon||X||} \frac{||\, f(X+H)-f(X)\,||}{||\, \varepsilon f(X)\,||} \\
&= \lim_{\varepsilon \to 0} \sup_{||H|| \leq \varepsilon||X||} \frac{||\, f(X+H)-f(X)\,||}{||\, \varepsilon f(X)\,||}\frac{||X||}{||X||} \\
&= \lim_{\varepsilon \to 0} \sup_{||H|| \leq \varepsilon||X||} \frac{||\, f(X+H)-f(X)\,||}{||\, \varepsilon X\,||}\frac{||X||}{||f(X)||} \\
&= \lim_{\eta \to 0} \sup_{||H|| \leq \eta} \frac{||\, f(X+H)-f(X)\,||}{\eta}\ \frac{||X||}{||f(X)||} \\
&= cond_{abs}(f,X)\frac{||X||}{||f(X)||}
\end{aligned} \tag{2.16}$$

Slijedeći teorem govori o izražavanju brojeva uvjetovanosti pomoću norme Fréchetove derivacije (2.7).

Teorem 2.11 *Apsolutni i relativni brojevi uvjetovnosti su dani formulama:*

$$cond_{abs}(f,X) = ||L(X)|| \tag{2.17}$$

$$cond_{rel}(f,X) = \frac{||L(X)||||X||}{||f(X)||} \tag{2.18}$$

Dokaz:

Dovoljno je dokazati prvu formulu (2.17) jer iz (2.16) onda odmah slijedi i druga formula (2.18). Prva jednakost slijedi iz Definicije 2.10, druga iz (2.6). Treća je linearnost od L. Za posljednju se koristi Definicija 1.13, tj.

$||H|| = o(\varepsilon)$ pa je $o(O(\varepsilon)) = o(\varepsilon)$ te (2.7) i činjenica da neprekidna funkcija postiže maksimum na kompaktnom skupu u rubu:

$$\begin{aligned} cond_{abs}(f, X) &= \lim_{\varepsilon \to 0} \sup_{||H|| \leq \varepsilon} \frac{|| f(X+H) - f(X) ||}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \sup_{||H|| \leq \varepsilon} \frac{||L(X,H) + o(||H||)||}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \sup_{||H|| \leq \varepsilon} \left|\left| L\left(X, \frac{H}{\varepsilon}\right) + \frac{o(||H||)}{\varepsilon} \right|\right| \\ &= \sup_{||Z|| \leq 1} ||L(X,Z)|| \\ &= ||L(X)|| \end{aligned}$$

□

2.4 Schurova dekompozicija

Korolar 1.21 nam olakšava pronalaženje svojstvenih vrijednosti matrice A, tako da transformiramo A u neku njoj sličnu matricu koja je jednostavnije strukture (dijagonalna, gornje trokutasta) te joj je tako lakše pronaći svojstvene vrijednosti.

Teorem 2.12 *Neka je $A \in C^{n \times n}$ matrica sa svojstvenim vrijednostima λ_1, $\lambda_2, \ldots, \lambda_n$. Tada postoji unitarna matrica U te gornje trokutasta matrica T takve da vrijedi $A = UTU^*$ i na dijagonali od T su svojstvene vrijednosti od A, tj. $[t_{ii}] = \lambda_i$, za svaki i.*

Dokaz:

Dokazujemo pomoću matematičke indukcije.
Baza: $n = 1 \Rightarrow A = 1A1^*$
Pretpostavka: Neka tvrdnja vrijedi za $A \in C^{(n-1)\times(n-1)}$
Korak:
Gledamo: $Au_1 = \lambda_1 u_1$, $||u_1||_2 = 1$. Skup $\{u_1\}$ nadopunimo do ortonormirane baze u C^n, $\{u_1, , u_2, \ldots, u_n\}$ $(u_i^T u_i = 1,\ u_i^T u_j = 0)$.
Definirajmo ortonormiranu matricu koja je ortogonalna u odnosu na u_1:

$$V_2 := [u_2 \ \ldots \ u_n] \in C^{nx(n-1)}$$

Tada je matrica $U_1 := [u_1\ V_2] \in C^{n\times n}$ unitarna matrica t.d.

$$U_1^* A U_1 = \begin{bmatrix} u_1^* \\ V_2^* \end{bmatrix} \begin{bmatrix} Au_1 & AV_2 \end{bmatrix} =$$

$$= \begin{bmatrix} u_1^* \\ V_2^* \end{bmatrix} \begin{bmatrix} \lambda_1 u_1 & AV_2 \end{bmatrix} =$$

$$= \begin{bmatrix} \lambda_1 & u_1^* A V_2 \\ 0 & A_2 \end{bmatrix}, \text{gdje je } A_2 = V_2^* A V_2 \text{ slična sa A}$$

Zbog Korolara 1.21 vrijedi:

$$\prod_{i=1}^{n} (\lambda_i - \lambda) = det(A - \lambda I) = det(U_1^* A U_1 - \lambda I) = (\lambda_1 - \lambda) det(A_2 - \lambda I)$$

Slijedi da su $\lambda_2, \ldots, \lambda_n$ svojstvene vrijednosti od A_2.
Po pretpostavci indukcije postoje unitarna matrica $U_2 \in C^{(n-1)\times(n-1)}$ i gornje trokutasta matrica $T_2 \in C^{(n-1)\times(n-1)}$ sa $\lambda_2, \ldots, \lambda_n$ na dijagonali t.d.

$$A_2 = U_2 T_2 U_2^*.$$

Definiramo:

$$U := U_1 \begin{bmatrix} 1 & 0 \\ 0 & U_2 \end{bmatrix} \in \mathbf{C^{n\times n}}$$

$$\Rightarrow \mathrm{U}^* U = \begin{bmatrix} 1 & 0 \\ 0 & U_2^* \end{bmatrix} U_1^* U_1 \begin{bmatrix} 1 & 0 \\ 0 & U_2 \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & U_2^* U_2 \end{bmatrix} = I$$

Očito je U unitarna matrica i vrijedi:

$$\begin{aligned}
U^*AU &= \begin{bmatrix} 1 & 0 \\ 0 & U_2^* \end{bmatrix} U_1^* A U_1 \begin{bmatrix} 1 & 0 \\ 0 & U_2 \end{bmatrix} = \\
&= \begin{bmatrix} 1 & 0 \\ 0 & U_2^* \end{bmatrix} \begin{bmatrix} \lambda_1 & u_1^* A V_2 \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & U_2 \end{bmatrix} = \\
&= \begin{bmatrix} \lambda_1 & u_1^* A V_2 U_2 \\ 0 & U_2^* A_2 U_2 \end{bmatrix} = \\
&= \begin{bmatrix} \lambda_1 & u_1^* A V_2 U_2 \\ 0 & T_2 \end{bmatrix} = \\
&= \begin{bmatrix} \lambda_1 & & & u_1^* A V_2 U_2 \\ 0 & \lambda_2 & \dots & * \\ \dots & \dots & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_n \end{bmatrix} = \\
&= T
\end{aligned}$$

□

2.5 Matrične iteracije

Matrične iteracije se identificiraju pomoću rekurzije. Slijedeća iteracija je funkcija prethodne iteracije:

$$g : \mathbf{C^{n\times n}} \to \mathbf{C^{n\times n}},\ X_{i+1} = g(X_i) \tag{2.19}$$

Potreban je početni uvjet, koji je u ovom slučaju matrica najčešće $X_0 = A$ ili $X_0 = I$. Iterativna funkcija g može i ne mora ovisiti o A. Uzevši u obzir kompliciranost računanja, najbolje je da je funkcija g polinom ili racionalna funkcija. Racionalne funkcije dovode do računanja inverzne matrice ili rješenja sustava sa kompliciranijom desnom stranom. Na modernim računalima puno se brže izračunaju potencije matrica, tj. višestruko množenje matrica, nego računanje sustava ili traženje inverzne matrice. To znači da su polinom poželjniji od racionalnih funkcija.

Najstandardnija metoda izvođenja iteracija je Newtonova. Newtonova

iteracija za $f : \mathbf{R} \to \mathbf{R}$ je:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

Metoda služi traženju nultočke funkcije f iz iteracije. Ovisno o kojoj se funkciji radi, iteracija se razvija pomoću deriviranja i sređivanja gornje jednadžbe. Iteracija za realnu funkciju je primjenjiva na matrične funkcije, tj. funkciju g iz (2.19).

2.5.1 Red kovergencije

Ako je (X_i) niz koji konvergira prema $\overline{X}$ kažemo da je red konvergencije najveći broj p za koji vrijedi:

$$||\overline{X} - X_{i+1}|| \leq c\,||\overline{X} - X_i||^p, \tag{2.20}$$

gdje je c neka pozitivna konstanta, a i je dovoljno veliki indeks.

Iteracija je reda p ako niz koji ju generira ima konvergenciju reda p. Linearna kovergencija je reda 1, kvadratna je reda 2, a super linearna konvergencija je ona za koju vrijedi:

$$\lim_{i\to\infty} \frac{||\overline{X} - X_{i+1}||}{||\overline{X} - X_i||} = 0 \tag{2.21}$$

Konvergencija niza se dijeli na dva dijela. Prvi je početni dio u kojem se greška smanjuje ispod 1. Druga faza je asimptotska u kojoj (2.20) garantira da će niz konvergirati prema nuli. Red konvergencije se odnosi na asimptotsku fazu, dok za početnu fazu ne znamo koliki je broj iteracija potreban. Često se u praksi iteracije skaliraju što znači da se članovi niza množe sa nekim određenim skalarom u svrhu skraćivanja broja iteracija u početnoj fazi. Što je veći red konvergencije, to je teže skalirati iteraciju.

2.5.2 Kriterij zaustavljanja

Jedno od važnijih pitanja u iteracijama je kada se zaustaviti? Ako iteracija (X_i) konvergira prema $\overline{X}$, kriterij zaustavljanja uz neku zadanu toleranciju *tol* može biti (uz bilo koju matričnu normu):

1. Zaustaviti iteraciju kada X_i ima relativnu grešku manju *tol*:

$$\frac{||X_i - \overline{X}||}{||\overline{X}||} \leq tol \tag{2.22}$$

2. Zaustaviti iteraciju kada je apsolutna greška ispod tol:

$$||X_i - \overline{X}|| \leq tol \tag{2.23}$$

3. Zaustavljanje iteracije bazirano na relativnoj razlici između dvije susjedne iteracije koja je manja od tol:

$$\frac{||X_{i+1} - X_i||}{||X_{i+1}||} \leq tol \tag{2.24}$$

Treći kriterij zapravo aproksimira relativnu grešku u X_i iz prvog kriterija u kojoj je $\overline{X} = \lim_{i \to \infty} X_i$. Zadnji kriterij je najčešće korišten pa označimo taj kriterij sa:

$$\delta_{i+1} := \frac{||X_{i+1} - X_i||}{||X_{i+1}||}$$

Zaista, $X_{i+1} - \overline{X} = (X_{i+1} - X_i) + (X_i - \overline{X})$ pa kada greška naglo pada očito iteracije brzo konvergiraju prema $\overline{X}$ te će tada norme $||X_{i+1} - X_i||$ i $||X_i - \overline{X}||$ biti otprilike jednake jer vrijedi $||X_{i+1} - \overline{X}|| \ll ||X_i - \overline{X}||$.
Promotrimo kriterij zaustavljanja u kvadratnoj konvergenciji:

$$||X_{i+1} - \overline{X}|| \leq c\,||X_i - \overline{X}||^2 \tag{2.25}$$

Pomoću nejednakosti trokuta i kvadratne konvergencije vrijedi:

$$\begin{aligned} ||X_i - \overline{X}|| &\leq ||X_i - X_{i+1}|| + ||X_{i+1} - \overline{X}|| \\ &\leq ||X_i - X_{i+1}|| + c\,||X_i - \overline{X}||^2 \end{aligned} \tag{2.26}$$

pa je

$$||X_i - \overline{X}|| \leq \frac{||X_{i+1} - X_i||}{1 - c\,||X_i - \overline{X}||} \tag{2.27}$$

Rješavajući realnu kvadratnu nejednažbu (2.26) za dovoljno mali $||X_i - \overline{X}||$ takav da je

$$||X_i - \overline{X}|| \leq \frac{1}{2\,c} \tag{2.28}$$

Koristeći (2.25), (2.27) i (2.28) slijedi:

$$\begin{aligned} ||X_{i+1} - \overline{X}|| &\leq c\,||X_i - \overline{X}||^2 \\ &\leq c\left(\frac{||X_{i+1} - X_i||}{1 - c\,||X_i - \overline{X}||}\right)^2 \\ &\leq 2\,c\,||X_{i+1} - X_i||^2 \end{aligned} \tag{2.29}$$

Usporedbom (2.25) i (2.29) zaključujemo da bi se zaustavili na iteraciji X_{i+1}, za kriterij relativne razlike trebat će nam možda jedna iteracija više u odnosu na kriterij relativne greške.

2.5.3 Numerička stabilnost

Ako je $X_0 = A$, svaka iteracija u obliku (2.19) je funkcija od A pa po Teoremu 2.5, svaka iteracija komutira sa A. Svojstvo komutativnosti je često korišteno u dokazivanju konvergencije te razvijanju iteracije. U aritmetici konačne preciznosti, zaokruživanje greške uzrokuje gubitak svojstva komutativnosti što se očituje kao numerička nestabilnost. Također, proizvoljna greška može proširiti nestabilnost iz iteracije u iteraciju.

Napomena 2.13 I-tu potenciju Fréchetove derivacije u X označavamo sa $L^i(X)$, a definiramo kao i-tu kompoziciju. Za drugu potenciju vrijedi:

$$L^2(X, H) = (L \circ L)(X, H) := L(X, L(X, H))$$

Analogno vrijedi za ostale potencije, uz $L^0(X, H) := H$.

Definicija 2.14 Promatramo iteraciju $X_{i+1} = g(X_i)$ sa fiksnom točkom X. Pretpostavimo da je g Fréchet derivabilna u X. *Iteracija je stabilna* u okolini od X ako Fréchetova derivacija od g ima ograničene potencije, tj. postoji konstanta c t.d.

$$||L_g^i(X)|| \leq c, \; za \; sve \; i > 0.$$

Ako gledamo derivaciju u nekom smjeru, $(L_g^i(X))(H)$, u definiciji stabilnosti je tada $||(L_g^i(X))(H)|| \leq c||H||$.

Neka je $X_0 = X + H_0$ perturbiran oko fiksne točke X sa H_0 prilično male norme te $H_i := X_i - X$. Po Definiciji 2.7:

$$\begin{aligned} X_i = g(X_{i-1}) = g(X + H_{i-1}) = \\ = g(X) + L_g(X, H_{i-1}) + o(||H_{i-1}||) \end{aligned} \tag{2.30}$$

Uz gornju jednakost i $g(X) = X$ slijedi slijedeće:

$$\begin{aligned} \Rightarrow \quad \mathrm{H}_i = X_i - X = \\ = g(X) + L_g(X, H_{i-1}) + o(||H_{i-1}||) - X = \\ = L_g(X, H_{i-1}) + o(||H_{i-1}||) \end{aligned} \tag{2.31}$$

Pomoću gornje jednakosti za H_i te linearnosti operatora L_g razvijanjem iteracije dobivamo:

$$\begin{aligned}
H_i &= L_g(X, H_{i-1}) + o(||H_{i-1}||) = \\
&= L_g(X, L_g(X, H_{i-2}) + o(||H_{i-2}||)) + o(||H_{i-1}||) = \\
&= L_g(X, L_g(X, H_{i-2})) + L_g(X, o(||H_{i-2}||)) + o(||H_{i-1}||) = \\
&= L_g^2(X, H_{i-2}) + L_g(X, o(||H_{i-2}||)) + o(||H_{i-1}||) = \\
&= L_g^2(X, L_g(X, H_{i-3}) + o(||H_{i-3}||)) + L_g(X, o(||H_{i-2}||)) + o(||H_{i-1}||) = \\
&= L_g^3(X, H_{i-3}) + L_g^2(X, o(||H_{i-3}||)) + L_g(X, o(||H_{i-2}||)) + o(||H_{i-1}||) = \\
&= \dots
\end{aligned}$$

Na kraju H_i ima oblik:

$$H_i = L_g^i(X, H_0) + \sum_{k=0}^{i-1} L_g^k(X, o(||H_{i-1-k}||) \tag{2.32}$$

U slučaju kad je iteracija stabilna, iz Definicije 2.14 i (2.32) slijedi:

$$\begin{aligned}
||H_i|| &= ||L_g^i(X, H_0) + \sum_{k=0}^{i-1} L_g^k(X, o(||H_{i-1-k}||)|| \leq \\
&\leq c||H_0|| + c\sum_{k=0}^{i-1} o(||H_{i-1-k}||) \leq \\
&\leq c||H_0|| + i\, c\, o(||H_0||)
\end{aligned} \tag{2.33}$$

Zadnja nejednakost govori kako u stabilnoj iteraciji, približno male greške u okolini fiksne točke su ograničene pomoću izraza koji ovisi o prvoj grešci.

Za skalarne iteracije $g : \mathbf{R} \to \mathbf{R},\ g(x_i) = x_{i+1}$ koje su konvergentne, vrijedi da konvergiraju u fiksnu točku,

$$\lim_{i\to\infty} g(x_i) = x,\ x\ t.d.\ g(x) = x.$$

Za takve iteracije, konvergencija implicira stabilnost jer je takvoj superlinearno konvergentnoj iteraciji derivacija u fiksnoj točci x 0, tj. ograničena je. To pokazujemo uz pomoć definicije derivacije funkcije i činjenice $g(x) = x$:

$$0 = \lim_{i\to\infty} \frac{|g(x_i) - x|}{|x_i - x|} = \lim_{i\to\infty} \frac{|g(x_i) - g(x)|}{|x_i - x|} = g'(x)$$

Dakle,

$$g'(x) = 0$$

U slučaju matrica, Fréchetova derivacija iteracijske funkcije u fiksnoj točci ne mora biti 0, stoga nam trebaju slijedeća dva teorema da bi bolje utvrdili stabilnost matričnih iteracija.

Teorem 2.15 *Neka je f idempotentna funkcija koja je Fréchet derivabilna u $X = f(X)$. Tada je $L_f(X)$ idempotentna.*

Dokaz:

Neka je $h(t) = f(f(t))$. Iz Teorema 2.8, lančanog pravila (2.10) slijedi:

$$L_h(X, H) = L_f(f(X), L_f(X, H)) = L_f(X, L_f(X, H))$$

Zbog gornje jednkosti i idempotentnosti od f vrijedi:

$$\begin{aligned}((L_f \circ L_f)(X))(H) &= L_f(X, L_f(X, H)) = \\ &= L_h(X, H) = \\ &= L_f(X, H) = \\ &= (L_f(X))(H),\end{aligned}$$

što znači da je L_f idempotentna.

□

Stabilnost je utvrđena Fréchetovim derivacijama iteracijske funkcije, a ne funkcije f. Slijedeći teorem govori o vezi f i g.

Teorem 2.16 *Neka je f idempotentna Fréchet derivabilna u $X = f(X)$ sa Fréchetovom derivacijom $L_f(X)$, $g(X_i) = X_{i+1}$ iteracije koje superlinearno konvergiraju prema $f(X_0)$, gdje je X_0 dovoljno blizu X i neka je g neovisna o izboru X_0. Tada je Fréchetova derivacija od g u X $L_g(X) = L_f(X)$.*

Dokaz:

Za dovoljno mali H, $X_0 = X + H$, $f(X_0) = f(X + H)$ postoji po Definiciji 2.7 pa vrijedi:

$$f(X + H) = f(X) + L_f(X, H) + o(||H||) = X + L_f(X, H) + o(||H||)$$

$$\Rightarrow f(X+H) - (X+H) = L_f(X,H) - H + o(||H||) = O(||H||)$$

Superlinearna kovergencija iteracije $g(X_i) = X_{i+1}$ prema $f(X_0)$ iz (2.21) povlači:

$$\lim_{i\to\infty} \frac{||f(X_0) - X_{i+1}||}{||f(X_0) - X_i||} = \lim_{i\to\infty} \frac{||f(X_0) - g(X_i)||}{||f(X_0) - X_i||} = 0$$

Budući da $g(X_i)$ konvergiraju prema $f(X_0)$, slijedeća iteracija nakon limesa je upravo limes:

$$g(\lim_{i\to\infty} g(X_i)) = \lim_{i\to\infty} g(X_i),\ tj.$$

$$g(f(X_0)) = f(X_0).$$

Za $H = 0$, slijedi da i g ima fiksnu točku u X jer f ima fiksnu točku u X. Zbog svega navedenog slijedi:

$$||f(X_0) - g(X_0)|| = o(||f(X_0) - X_0||) = o(O(||H||)) = o(||H||)$$

$$\Rightarrow g(X_0) - g(X) = g(X_0) - X = f(X_0) - X + o(||H||) = L_f(X,H) + o(||H||)$$

Zaključujemo, $L_f(X,H) = L_g(X,H)$.

□

$L_f(X)$ je konačnodimenzionalni linearni operator pa vrijedi da je ograničen. Budući da je L_f idempotentna funkcija, prvi teorem implicira da su sve potencije nje ograničene. Uz drugi teorem imamo da je diferencijal iteracijske funkcije jednak $L_f(X)$. Iz tog slijedi da sve iteracije koje superlinearno konvergiraju su numerički stabilne i ne trebamo računati Fréchetove derivacije, niti testirati granice njihovih potencija.

3 MATRIČNA FUNKCIJA PREDZNAKA

3.1 Uvod u funkciju predznaka

Neka je $\mathbf{I}$ skup brojeva koji leže na imaginarnoj osi ($0 \in \mathbf{I}$). Funkcija $sign : \mathbf{C} \setminus \mathbf{I} \to \{-1, 1\}$ je definirana za $z \in \mathbf{C} \setminus \mathbf{I}$ kao:

$$sign(z) := \begin{cases} 1\ ,\ Re(z) > 0 \\ -1\ ,\ Re(z) < 0 \end{cases} \tag{3.1}$$

gdje je $Re(z)$ realni dio kompleksnog broja z.

Matrična funkcija predznaka može biti dobivena iz bilo koje od dvije definicije s početka (2.2) i (2.4). U definicijama baziranim na Jordanovoj formi i interpolaciji polinomom potrebne su derivacije funkcije. Lako je zaključiti da su derivacije funkcije $sign$ jednake nuli, $sign^{(k)}(z) = 0, \forall k \geq 1\ i\ \forall z \in \mathbf{C} \setminus \mathbf{I}$.

Pretpostavljamo da matrice koje koristimo iz $C^{n\times n}$ imaju svojstvene vrijednosti iz skupa kompleksnih brojeva bez čisto imaginarnih brojeva (imaginarne osi) pa je funkcija $sign$ dobro definirana u smislu (2.2) i (2.4) u kojima su nam potrebne vrijednosti funkcija i njenih derivacija u svojstvenim vrijednostima. Slijedi da ne postoji svojstvena vrijednost 0 jer je $0 \in \mathbf{I}$, stoga su takve matrice regularne (Korolar 1.20).

Iz (2.2) imamo jedan način primjene funkcije f na neku matricu. Neka je $A \in C^{n\times n}$ u Jordanovoj formi, $A = ZJZ^{-1}$, $J = diag(J_1, J_2)$. Budući da svojstvene vrijednosti matrice A nisu na imaginarnoj osi, neka su u $J_1 \in C^{p\times p}$ svojstvene vrijednosti koje se nalaze na lijevoj polovici ravnine (p algebarska kratnost tih svojstvenih vrijednosti) dok se u $J_2 \in C^{q\times q}$ nalaze svojstvene vrijednosti na desnoj polovici (q algebarska kratnost svojstvenih vrijednosti s desne polovice). Funkcija $sign$ primijenjena na te svojstvene vrijednosti

daje 1, odnosno -1, dok su derivacije 0. Zato je funkcija predznaka u obliku:

$$sign(A) := Z \begin{bmatrix} -I_p & 0 \\ 0 & I_q \end{bmatrix} Z^{-1} \tag{3.2}$$

Promotrimo sada skalarnu funkciju *sign* iz (3.1). Budući da je z kompleksan broj, može biti u polarnim koordinatama u obliku: $z = r(\cos\alpha + i\sin\alpha)$. Nadalje, pomoću Definicije (1.17):

$$\begin{aligned} z^2 &= r^2(\cos\alpha + i\sin\alpha)^2 = \\ &= r^2((\cos\alpha)^2 - (\sin\alpha)^2 + i2\cos\alpha\sin\alpha) = \\ &= r^2(\cos 2\alpha + i\sin 2\alpha) \end{aligned}$$

Kada računamo drugi korijen iz kompleksnog broja (1.17), on će imati dva rješenja. U ovom slučaju, ograničavamo rješenja na ona koja se nalaze s desne strane u odnosu na imaginarnu os, odnosno tražimo glavni drugi korijen iz z^2:

$$\begin{aligned} (z^2)^{\frac{1}{2}} &= \begin{cases} r(\cos\alpha + i\sin\alpha) & , \alpha \in \left\langle -\frac{\pi}{2}, \frac{\pi}{2} \right\rangle \\ r(\cos(\alpha+\pi) + i\sin(\alpha+\pi)) & , \alpha \in \left\langle \frac{\pi}{2}, \frac{3\pi}{2} \right\rangle \end{cases} \\ &= \begin{cases} r(\cos\alpha + i\sin\alpha) & , \alpha \in \left\langle -\frac{\pi}{2}, \frac{\pi}{2} \right\rangle \\ r(-\cos\alpha - i\sin\alpha) & , \alpha \in \left\langle \frac{\pi}{2}, \frac{3\pi}{2} \right\rangle \end{cases} \end{aligned} \tag{3.3}$$

Na kraju imamo:

$$\frac{z}{(z^2)^{\frac{1}{2}}} = \begin{cases} 1\,, \alpha \in \left\langle -\frac{\pi}{2}, \frac{\pi}{2} \right\rangle \\ -1\,, \alpha \in \left\langle \frac{\pi}{2}, \frac{3\pi}{2} \right\rangle \end{cases} = \begin{cases} 1\,, Re(z) > 0 \\ -1\,, Re(z) < 0 \end{cases} \tag{3.4}$$

Generaliziramo li gornju funkciju na matrice, dobivamo:

$$sign(A) := A(A^2)^{\frac{-1}{2}} \tag{3.5}$$

Teorem 3.1 *Neka $A \in C^{n\times n}$ nema čisto imaginarne svojstvene vrijednosti i neka je $S = sign(A)$. Vrijede slijedeća svojstva:*

1. S je involutorna matrica, tj. $S^2 = I$.
2. S je dijagonalizabilna sa svojstvenim vrijednostima -1 i 1.

3. $SA = AS$

4. Ako je λ svojstvena vrijednost od A, onda je i $sign(\lambda)$ svojstvena vrijednost od $sign(A)$.

Dokaz:

Za dokazivanje koristimo (3.2). Prvo svojstvo:

$$\begin{aligned} S^2 &= (Z\ sign(J)\ Z^{-1})^2 \\ &= (Z\ diag(-I_p,\ I_q)\ Z^{-1})^2 = \\ &= (Z\ diag(-I_p,\ I_q)\ Z^{-1})\ (Z\ diag(-I_p,\ I_q)\ Z^{-1}) = \\ &= Z\ diag((-I_p)^2,\ I_q^2)\ Z^{-1} = \\ &= Z\ diag(I_p^2,\ I_q^2)\ Z^{-1} = \\ &= I \end{aligned}$$

Drugo svojstvo slijedi direktno iz (3.2), dok su svojstva (3) i (4) dokazana za općenite matrične funkcije u Teoremu 2.5.

□

Ako je spektar od A cijeli u pozitivnoj poluravnini, odnosno negativnoj, tada je $sign(A) = ZZ^{-1} = I$, odnosno $sign(A) = -ZZ^{-1} = -I$, što vrijedi intuitivno iz Teorema 3.1 (svojstva (1) i (2)). Inače, ako je spektar matrice s obje strane imaginarne osi, ne vrijedi da je $sign(A)$ primarni drugi korijen jedinične matrice, I ili $-I$ (primarni drugi korijen preslikava svaki element 1 u sve vrijednosti 1, ili pak u sve -1, a kod neprimarnog postoji neka vrijednost 1 preslikana u -1 a druga u 1).

3.2 Schurova metoda

Neka je $A \in C^{n\times n}$. Schurova dekompozicija matrice A je QTQ^*, gdje je T gornje trokutasta matrica, a Q je unitarna. Po Teoremu 2.5, drugom svojstvu slijedi:

$$sign(A) = Qsign(T)Q^*.$$

Budući da je T gornje trokutasta matrica, onda je po Teoremu 2.5, četvrtom svojstvu, i $sign(T)$ gornje trokutasta matrica $U := sign(T)$ za koju vrijedi:

$$u_{ii} = sign(t_{ii}) = \pm 1$$

jer su t_{ii} svojstvene vrijednosti od matrice T (gornje trokutasta matrica ima svojstvene vrijednosti na dijagonali). Ostali u_{ij} se dobiva iz:

$$U^2 = I \ i \ UT = TU.$$

Elementi matrice U^2 koji se nalaze u i-tom retku i j-tom stupcu su u obliku:

$$\sum_{k=i}^{j} u_{ik}u_{kj} = (u_{ii} + u_{jj})u_{ij} + \sum_{k=i+1}^{j-1} u_{ik}u_{kj}$$

Kada njih izjednačimo sa nulama u jediničnoj matrici, dobit ćemo, u slučaju da je $u_{ii} + u_{jj} \neq 0$:

$$u_{ij} = -\frac{\sum_{k=i+1}^{j-1} u_{ik}u_{kj}}{u_{ii} + u_{jj}}$$

Za slučaj kada vrijedi $u_{ii} + u_{jj} = 0$ koristimo jednadžbu $UT = TU$. Izjednačavamo elemente matrica s obje strane:

$$\sum_{k=i}^{j} u_{ik}t_{kj} = \sum_{k=i}^{j} t_{ik}u_{kj}$$

$$(t_{ii} - t_{jj})u_{ij} - (u_{ii} - u_{jj})t_{ij} = \sum_{k=i+1}^{j-1} (u_{ik}t_{kj} - t_{ik}u_{kj})$$

Na kraju dobivamo:

$$u_{ij} = \frac{(u_{ii} - u_{jj})t_{ij}}{t_{ii} - t_{jj}} + \frac{\sum_{k=i+1}^{j-1} (u_{ik}t_{kj} - t_{ik}u_{kj})}{t_{ii} - t_{jj}}$$

Ako vrijedi $t_{ii} - t_{jj} = 0$, onda je $sign(t_{ii}) = sign(t_{jj})$. Iz tog slijedi da je $u_{ii} + u_{jj} = \pm 2 \neq 0$. Vrijedi i obrnuto, $t_{ii} - t_{jj} \neq 0 \iff u_{ii} + u_{jj} = 0$. Dakle, imamo dva disjunktna slučaja.

Algoritam:

▷ Svedi A na Schurovu dekompoziciju QTQ^*

▷ $u_{ii} = sign(t_{ii})$, $i = 1, \ldots, n$

▷ za $j = 2, \ldots, n$

za $i = j - 1, \ldots, 1$

$$u_{ij} = \begin{cases} -\dfrac{\sum_{k=i+1}^{j-1} u_{ik}\, u_{kj}}{u_{ii} + u_{jj}} & , u_{ii} + u_{jj} \neq 0 \\ \dfrac{(u_{ii} - u_{jj})\, t_{ij}}{t_{ii} - t_{jj}} + \dfrac{\sum_{k=i+1}^{j-1} (u_{ik}\, t_{kj} - t_{ik}\, u_{kj})}{t_{ii} - t_{jj}} & , u_{ii} + u_{jj} = 0 \end{cases}$$

kraj

kraj

▷ $S = QUQ^*$

Složenost algoritma je $O(n^3)$. Sveukupno je potrebno oko $\frac{86}{3}n^3$ operacija da bi se došlo do rješenja.

3.3 Newtonova metoda

U ovom slučaju gdje je *sign* funkcija involutorna matrica, tj.$sign(A)^2 = I$, slijedi da je *sign* nultočka funkcije $g(x) = x^2 - 1$. Kada tu funkciju uvrstimo u Newtonovu iteraciju, dobivamo:

$$\begin{aligned} x_{n+1} &= x_n - \frac{g(x_n)}{g'(x_n)} = \\ &= x_n - \frac{x_n^2 - 1}{2x_n} = \\ &= \frac{1}{2}\left(x_n + \frac{1}{x_n}\right) \end{aligned} \tag{3.6}$$

Uvrstimo li matricu, dobit ćemo oblik:

$$X_{i+1} = \frac{1}{2}\left(X_i + X_i^{-1}\right),\ X_0 = A \tag{3.7}$$

Za slučaj matrica se gledaju njihove svojstvene vrijednosti jer je iteracija (3.7) izvedena iz (3.6) za svaku pojedinu svojstvenu vrijednost.
Slijedeći teorem govori o konvergenciji Newtonove iteracije prema *sign*.

Teorem 3.2 *Neka matrica $A \in C^{n\times n}$ nema svojstvene vrijednosti na imaginarnoj osi. Tada vrijedi:*

1. *Iteracije X_i iz (3.7) konvergiraju prema $S := sign(A)$.*

2. *Kvadratna brzina konvergencije za svaku konzistentnu normu (Definicija 1.9):*
$$||X_{i+1} - S|| \leq \frac{1}{2} \, ||X_i^{-1}|| \, ||X_i - S||^2.$$

3. *Za $k \geq 1$ te $G_0 := (A - S)(A + S)^{-1}$ vrijedi:*
$$X_i = (I - G_0^{2^i})^{-1}(I + G_0^{2^i})S$$

Dokaz:

Kompleksni broj $z = Re(z) + iIm(z)$ se po Definiciji 1.16 može zapisati u polarnim koordinatama, $z = re^{i\alpha}$. Neka je $\lambda = re^{i\alpha}$. Iz Definicije 1.17 slijedi da je:

$$\begin{aligned}\frac{\lambda + \lambda^{-1}}{2} &= \frac{r}{2}(\cos\alpha + i\sin\alpha) + \frac{r^{-1}}{2}(\cos(-\alpha) + i\sin(-\alpha)) = \\ &= \frac{r}{2}(\cos\alpha + i\sin\alpha) + \frac{r^{-1}}{2}(\cos\alpha - i\sin\alpha) = \\ &= \frac{r + r^{-1}}{2}\cos\alpha + \frac{r - r^{-1}}{2} i\sin\alpha\end{aligned}$$

Ako je λ svojstvena vrijednost za neku matricu X_i iz iteracije (3.7), onda po Teoremu 2.5 imamo da je svojstvena vrijednost za X_{i+1}:

$$\mu := \frac{\lambda + \lambda^{-1}}{2}.$$

Usporedimo dva kompleksna broja u odnosu na poluravnine u kojima leže:

$$\lambda = r\,cos\alpha + r\,sin\alpha,\ \mu = \frac{r + r^{-1}}{2}\cos\alpha + \frac{r - r^{-1}}{2} i\sin\alpha$$

Ako je λ u jednoj od poluravnina, onda će μ ovisno $\frac{1}{2}(r + r^{-1})$ i $\frac{1}{2}(r - r^{-1})$ promjeniti poluravninu ili ostati u istoj. Budući da je $r + r^{-1}$ pozitivan broj, zaključujemo da je μ u istoj poluravnini kao i λ bez obzira je li $r - r^{-1}$ negativan ili pozitivan broj jer taj skalar mijenja samo okomitu komponentu smjera kompleksnog broja.

Počnimo od početka, tj. λ je svojstvena vrijednost od A. Budući da A nema svojstvene vrijednosti na imaginarnoj osi, po Korolaru 1.20 vrijedi da

je A regularna. Svaka slijedeća iteracija X_i nema svojstvene vrijednosti na imaginarnoj osi jer kada bi imala:

$$\mu = ci \iff \frac{\lambda + \lambda^{-1}}{2} = ci \iff \frac{\lambda^2 - 2ic\lambda + 1}{\lambda} = 0 \iff \lambda = ci \pm i\sqrt{c^2+1}$$

Dakle, svojstvena vrijednost od iteracije X_{i+1}, μ neće nikada biti na imaginarnoj osi jer bi onda svojstvena vrijednost od X_i trebala biti čisto imaginarni broj, a to ne vrijedi. Zadnja činjenica se lako dokazuje pomoću indukcije unatrag gdje dodjemo u kontradikciju s pretpostavkom da $X_0 = A$ nema imaginarne svojstvene vrijednosti. Iz Korolara 1.20 zaključujemo da je X_i dobro definirana na spektru bez imaginarne osi i regularna matrica, za svaki $i > 0$.

$$\begin{aligned} X_{i+1} \pm S &= \frac{1}{2}\left(X_i + X_i^{-1} \pm 2S\right) = \\ &= \frac{1}{2} X_i^{-1}\left(X_i^2 \pm 2X_i S + I\right) = \\ &= \frac{1}{2} X_i^{-1}\left(X_i \pm S\right)^2 \end{aligned} \tag{3.8}$$

Uz 2.2 imamo:

$$\begin{aligned} X_1 &= \frac{1}{2}(A + A^{-1}) = \\ &= \frac{1}{2}(ZJZ^{-1} + (ZJZ^{-1})^{-1}) = \\ &= \frac{1}{2}(ZJZ^{-1} + ZJ^{-1}Z^{-1}) = \\ &= Z\left(\frac{J + J^{-1}}{2}\right)Z^{-1} \end{aligned} \tag{3.9}$$

Kako su λ i μ sa iste strane imaginarne osi, istog su predznaka pa je:

$$\begin{aligned} sign(X_1) &= sign\left(Z\left(\frac{J + J^{-1}}{2}\right)Z^{-1}\right) = \\ &= Z sign\left(\frac{J + J^{-1}}{2}\right)Z^{-1} = \\ &= Z sign(J) Z^{-1} = \\ &= sign(A) = \\ &= S \end{aligned} \tag{3.10}$$

Analogno, za svaku slijedeću iteraciju vrijedi: $sign(X_i) = sign(A) = sign(X_0)$. Svojstvene vrijednosti od matrice $X_i + sign(X_i)$ će biti u obliku $\lambda_i + sign(\lambda_i)$, što je uvijek različito od 0. Po Korolaru 1.20, $X_i + S$ je regularna matrica, $\forall i \geq 0$ pa njen inverz postoji. Također, jer su X_i racionalne funkcije od A, po Teoremu 2.5, komutiraju sa A. Pomoću definicije funkcije s polinomom i primjenivši da A komutira sa S vrijedi $Sf(A) = Sp(A) = p(A)S = f(A)S$. Dakle, X_i komutira sa S.
Iz (3.8) imamo:

$$(X_{i+1} - S)(X_{i+1} + S)^{-1} = ((X_i - S)(X_i + S)^{-1})^2 \tag{3.11}$$

Uz oznaku $G_i := (X_i - S)(X_i + S)^{-1}$, vrijedi: $G_{i+1} = G_i^2 = \cdots = G_0^{2^{i+1}}$.

U slučaju kada je λ svojstvena vrijednost od A, po Teoremu 2.5, G_0 definiran u iskazu teorema ima svojstvene vrijednosti u obliku:

$$\frac{\lambda - sign(\lambda)}{\lambda + sign(\lambda)}.$$

Udaljenosti tih svojstvenih vrijednosti od ishodišta (C (λ,$sign(\lambda)$) iz Definicije 1.18 o Cayleyovoj metrici) za one u desnoj poluravnini, odnosno one u lijevoj su:

$$\frac{\sqrt{(Re\lambda - 1)^2 + (Im\lambda)^2}}{\sqrt{(Re\lambda + 1)^2 + (Im\lambda)^2}}, \quad \frac{\sqrt{(Re\lambda + 1)^2 + (Im\lambda)^2}}{\sqrt{(Re\lambda - 1)^2 + (Im\lambda)^2}}.$$

Zbog većeg nazivnika od brojnika u oba slučaja, svojstvene vrijednosti od G_0 su unutar jediničnog kruga oko 0 u kompleksnoj ravnini, tj. za spektralni radijus vrijedi:

$$\rho(G_0) = \max_{\mu \in \sigma(G_0)} |\mu| = \left(\max_{\lambda \in \sigma(A)} \left| \frac{\lambda - sign(\lambda)}{\lambda + sign(\lambda)} \right| \right) < 1$$

Dakle, $G_i = G_0^{2^i}$ i $\rho(G_0) < 1$ pa iz Korolara 1.23 i 2.1 slijedi: $\lim_{i \to \infty} G_i = 0$.
Napokon, dobivamo tvrdnje teorema (1) i (3):

$$\Rightarrow \; G_i(X_i + S) = (X_i - S)$$

$$\Rightarrow \; G_i X_i + G_i S = X_i - S$$

$$\Rightarrow \; X_i = (I - G_i)^{-1}(I + G_i)S$$

$$\Rightarrow \lim_{i \to \infty} X_i = \lim_{i \to \infty} (I - G_i)^{-1}(I + G_i)S = S$$

Tvrdnja (2) slijedi iz (3.8).

□

Iz gornjeg teorema i Korolara 1.23 i 2.1 slijedi:

$$||G_0^{2^i}|| \geq \rho(G_0^{2^i}) = \left(\max_{\lambda \in \sigma(A)} \left| \frac{\lambda - sign(\lambda)}{\lambda + sign(\lambda)} \right| \right)^{2^i} \tag{3.12}$$

Konvergencija iteracija prema $sign(A)$ je po trećoj tvrdnji Teorema 3.2 analogna konvergenciji potencija G_0 prema 0. Očito će konvergencija $G_0^{2^i}$ prema 0 biti spora ako su:

- svojstvene vrijednosti od A blizu imaginarne osi
- $\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda| \gg 1$

jer će tada $\left| \frac{\lambda - sign(\lambda)}{\lambda + sign(\lambda)} \right| \approx 1$.

Algoritam:

Maksimalni broj iteracija u algoritmu označimo sa m, a *kriterij zaustavljanja* koji je na početku $TRUE$ je neki račun iz Poglavlja 2.5.2.

▷ $X_0 = A$ i $i = 0$

▷ za $i = 1, \ldots, m-1$

$B_i = X_{i-1}^{-1}$

$X_i = \frac{1}{2}\left(X_{i-1} + B_{i-1} \right)$

ako je *kriterij zaustavljanja* $== TRUE$, onda $i = i + 1$

kraj

▷ $S = X_{i+1}$

Složenost algoritma je također $O(n^3)$, ali je broj potrebnih operacija otprilike $2in^3$ za i koraka Newtonovih iteracija.

3.4 Skalirane Newtnove iteracije

Newtonova iteracija iz Poglavlja 3.3 za *sign* funkciju dobivena iz traženja nultočke za funkciju $(sign(z))^2 - 1$ je

$$x_{i+1} = \frac{1}{2}(x_i + \frac{1}{x_i}),\ x_0 = a.$$

Ona kvadratično konvergira prema $sign(a) = \pm 1$.

Može se gledati na način da je to Newtonova iteracija za funkciju drugog korijena od 1. Jednom dok je pogreška mala, slijedeće uzastopne pogreške se brzo smanjuju. Svaka slijedeća pogreška je približno jednaka kvadratu prethodne pogreške (vidi (3.8)).

Za $|x_i| \gg 1$ je $x_i \approx \frac{1}{2}x_{i-1}$, iteracija je zapravo jako spor način običnog dijeljenja s 2. Također iz Teorema 3.2 i dijela gdje je objašnjena nejednakost (3.12) vidimo da iteracija sporo konvergira kada je a kompleksni broj blizu imaginarne osi jer a možemo identificirati sa svojstvenom vrijednosti matrice 1×1. Traži se način za ubrzavanje konvergencije u ta dva slučaja.

Teorem 3.3 *Za Newtonovu iteraciju (3.7), ako X_i ima svojstvene vrijednosti ± 1, za neki i, onda je $X_{i+p} = sign(A)$ za $2^p \geq m$, gdje je m dimenzija najvećeg Jordanovog bloka od X_i.*

Dokaz:

Neka su ± 1 svojstvene vrijednosti od X_i. Neka je Jordanova forma $X_i = ZJ_iZ^{-1}$, $J_i = D + N_i$, $D = diag(\pm 1)$, N_i nilpotentna strogo gornje trokutasta matrica sa jedinicama i nulama iznad glavne dijagonale. Indeks nilpotentnosti je m, tj. $N_i^m = 0,\ N_i^{m-1}, \ldots,\ I \neq 0$. Promatramo konvergenciju niza koji počinje sa J_i, a završava sa D i možemo staviti $Z = I$ bez smanjenja općenitosti. Iteracija $X_{i+1} = D + N_{i+1}$ zadovoljava po (3.8):

$$X_{i+1} - D = \frac{1}{2}X_i^{-1}(X_i - D)^2 \iff N_{i+1} = \frac{1}{2}X_i^{-1}N_i^2$$

Budući da je m indeks nilpotentnosti od N_i, N_{i+1} mora imati indeks nilpotentnosti $\lceil \frac{m}{2} \rceil$. Primjenjujući tu činjenicu više puta, za $2^p \geq m$, N_{i+p} ima indeks nilpotentnosti 1, slijedi da je $X_{i+p} = D = sign(A)$ za $2^p \geq m$.

□

Učinkoviti način da se poveća brzina konvergencije je skaliranje iteracija: u (3.7) zamijenimo X_i sa $\mu_i X_i$:

$$X_{i+1} = \frac{1}{2}(\mu_i X_i + \mu_i^{-1} X_i^{-1}),\ X_0 = A \tag{3.13}$$

Dok god je μ_i pozitivan i realan broj, predznak svake iteracije je sačuvan. Predložene su tri vrste skaliranja:

- skaliranje pomoću determinante:

$$\mu_i := |det(X_i)|^{-\frac{1}{n}} \tag{3.14}$$

- skaliranje pomoću spektralnog radijusa:

$$\mu_i := \sqrt{\frac{\rho(X_i^{-1})}{\rho(X_i)}} \tag{3.15}$$

- skaliranje pomoću norme:

$$\mu_i := \sqrt{\frac{||X_i^{-1}||}{||X_i||}} \tag{3.16}$$

Pogledajmo slučaj skaliranja pomoću determinante. Neka su $\lambda_1^{(i)}, \ldots,\ \lambda_n^{(i)}$ svojstvene vrijednosti od X_i. Iz Korolara 1.22 znamo da je umnožak svojstvenih vrijednosti matrice jednak determinanti matrice. Za skaliranje pomoću determinante (3.14) vrijedi da je apsolutna vrijednost umnoška svojstvenih vrijednosti skalirane iteracije jednaka 1:

$$\begin{aligned}
\left|\prod_{k=1}^{n}\left(\mu_i \lambda_k^{(i)}\right)\right| &= \left|det(\mu_i X_i)\right| = \\
&= \left|det\left(|det(X_i)|^{-\frac{1}{n}} X_i\right)\right| = \\
&= \left|\left(|det(X_i)|^{-\frac{1}{n}}\right)^n det(X_i)\right| = \\
&= 1
\end{aligned} \tag{3.17}$$

Neka su $\lambda_1, \ldots,\ \lambda_n$ svojstvene vrijednosti od X. Definiramo funkciju $t : \mathbf{C^{n\times n}} \to \mathbf{R}^+$:

$$t(X) := \sum_{k=1}^{n} (\log|\lambda_k|)^2$$

Faktor skaliranja μ_i iz (3.14) minimizira $t(\mu_i X_i)$. To dokazujemo uz osnovna svojstva logaritma te Korolara 1.22:

$$g(\mu) := t(\mu X) = \sum_{k=1}^{n} (\log|\mu\lambda_k|)^2 = \sum_{k=1}^{n} (\log\mu + \log|\lambda_k|)^2$$

$$\Rightarrow \quad g'(\mu) = \frac{2}{\mu}\sum_{k=1}^{n}(\log\mu + \log|\lambda_k|) = \frac{2}{\mu}\left(n\log\mu + \sum_{k=1}^{n}\log|\lambda_k|\right) = 0$$

$$\Rightarrow \quad \log\mu^n = \log|\lambda_1\lambda_2\ldots\lambda_n|^{-1} \Rightarrow \quad \mu = \left|\prod_{k=1}^{n}\lambda_k\right|^{-\frac{1}{n}} = |det(X)|^{-\frac{1}{n}}$$

Promotrimo skaliranje pomoću spektralnog radijusa. Za X_i promatramo svojstvene vrijednosti $\lambda_n^{(i)}, \ldots, \lambda_1^{(i)}$ redom od najmanje prema najvećoj udaljenosti od ishodišta te spektralni radijus $\rho(X_i)$. Tada je iz (3.15):

$$\mu_i = |\,\lambda_1^{(i)}\lambda_n^{(i)}\,|^{-\frac{1}{2}}$$

pa matrica $\mu_i X_i$ ima svojstvene vrijednosti najmanje i najveće veličine:

$$|\mu_i\lambda_n^{(i)}| = \left|\frac{\lambda_n^{(i)}}{\lambda_1^{(i)}}\right|^{\frac{1}{2}} \quad i \quad |\mu_i\lambda_1^{(i)}| = \left|\frac{\lambda_1^{(i)}}{\lambda_n^{(i)}}\right|^{\frac{1}{2}}$$

Dakle, vrijedi:

$$|\mu_i\lambda_n^{(i)}| = \frac{1}{|\mu_i\lambda_1^{(i)}|}$$

Ako su $\lambda_1^{(i)}$ i $\lambda_n^{(i)}$ realni brojevi postoje 4 slučaja uz $|\lambda_n^{(i)}| < |\lambda_1^{(i)}|$ te Cayleyevu metriku iz Definicije 1.18:

(1) $\lambda_1^{(i)} > \lambda_n^{(i)} > 0 \quad \Rightarrow \quad \mu_i\lambda_n^{(i)} = \dfrac{1}{\mu_i\lambda_1^{(i)}}$

$$\frac{|\,\mu_i\lambda_n^{(i)} - 1|}{|\,\mu_i\lambda_n^{(i)} + 1|} = \frac{\left|\frac{1}{\mu_i\lambda_1^{(i)}} - 1\right|}{\left|\frac{1}{\mu_i\lambda_1^{(i)}} + 1\right|} = \frac{|\,\mu_i\lambda_1^{(i)} - 1|}{|\,\mu_i\lambda_1^{(i)} + 1|}$$

(2) $\lambda_1^{(i)} > 0 > \lambda_n^{(i)} \quad \Rightarrow \quad \mu_i\lambda_n^{(i)} = -\dfrac{1}{\mu_i\lambda_1^{(i)}}$

$$\frac{|\,\mu_i\lambda_n^{(i)} + 1|}{|\,\mu_i\lambda_n^{(i)} - 1|} = \frac{\left| -\frac{1}{\mu_i\lambda_1^{(i)}} + 1\right|}{\left| -\frac{1}{\mu_i\lambda_1^{(i)}} - 1\right|} = \frac{|\,\mu_i\lambda_1^{(i)} - 1|}{|\,\mu_i\lambda_1^{(i)} + 1|}$$

(3) $\lambda_n^{(i)} > 0 > \lambda_1^{(i)} \;\Rightarrow\; \mu_i\lambda_n^{(i)} = -\dfrac{1}{\mu_i\lambda_1^{(i)}}$

$$\frac{|\,\mu_i\lambda_n^{(i)} - 1|}{|\,\mu_i\lambda_n^{(i)} + 1|} = \frac{\left| -\frac{1}{\mu_i\lambda_1^{(i)}} - 1 \right|}{\left| -\frac{1}{\mu_i\lambda_1^{(i)}} + 1 \right|} = \frac{|\,\mu_i\lambda_1^{(i)} + 1|}{|\,\mu_i\lambda_1^{(i)} - 1|}$$

(4) $0 > \lambda_n^{(i)} > \lambda_1^{(i)} \;\Rightarrow\; \mu_i\lambda_n^{(i)} = \dfrac{1}{\mu_i\lambda_1^{(i)}}$

$$\frac{|\,\mu_i\lambda_n^{(i)} + 1|}{|\,\mu_i\lambda_n^{(i)} - 1|} = \frac{\left| \frac{1}{\mu_i\lambda_1^{(i)}} + 1 \right|}{\left| \frac{1}{\mu_i\lambda_1^{(i)}} - 1 \right|} = \frac{|\,\mu_i\lambda_1^{(i)} + 1|}{|\,\mu_i\lambda_1^{(i)} - 1|}$$

Po Definiciji 1.18 o Cayleyovoj metrici sva četiri slučaja impliciraju:

$$C(\mu_i\lambda_n^{(i)},\, sign(\mu_i\lambda_n^{(i)})) = C(\mu_i\lambda_1^{(i)},\, sign(\mu_i\lambda_1^{(i)}))$$

Budući da je μ_i pozitivan realan broj, $sign(\mu_i\lambda_k^{(i)}) = sign(\lambda_k^{(i)}),\ k = 1, 2$ imamo:

$$C(\mu_i\lambda_n^{(i)},\, sign(\lambda_n^{(i)})) = C(\mu_i\lambda_1^{(i)},\, sign(\lambda_1^{(i)}))$$

Znamo da Newtonove iteracije (3.7) teže prema $sign(A)$ po Teoremu 3.2. Gledajući po svojstvenim vrijednostima od A, neka je λ njena svojstvena vrijednost, tada svojstvena vrijednost svake iteracije X_i dobivena iz prethodne iteracije, $\lambda^{(i)}$ teži prema $sign(\lambda)$. Zaključujemo da spektralno skaliranje izjednačava greške u Cayleyovoj metrici najmanje i najveće svojstvene vrijednosti.

Teorem 3.4 (Barraud) *Neka je matrica $A \in C^{n\times n}$ regularna sa realnim svojstvenim vrijednostima i neka je $S := sign(A)$. Za spektralno skalirane Newtonove iteracije (3.13), $X_{d+p-1} = sign(A)$, d je broj različitih svojstvenih vrijednosti od A, a $2^p \geq m$, gdje je m dimenzija najvećeg Jordanovog bloka od A.*

Dokaz:

Koristit ćemo svojstva iterativne funkcije $f(x) = \frac{1}{2}\left(x + \frac{1}{x}\right)$:

(1) $f(x) = f\left(\frac{1}{x}\right)$

(2) $0 \leq x_2 \leq x_1 \leq 1$ ili $1 \leq x_1 \leq x_2 \Rightarrow 1 \leq f(x_1) \leq f(x_2)$

Za $X_0 = A$ promatramo svojstvene vrijednosti $\lambda_n, \ldots, \lambda_1$ redom od najmanje prema najvećoj udaljenosti od ishodišta. Tada, iz (3.15) imamo $\mu_0 = |\lambda_n \lambda_1|^{-\frac{1}{2}}$ pa $\mu_0 X_0$ ima svojstvene vrijednosti kojima su moduli između:

$$|\mu_0 \lambda_n| = \left|\frac{\lambda_n}{\lambda_1}\right|^{\frac{1}{2}} \quad i \quad |\mu_0 \lambda_1| = \left|\frac{\lambda_1}{\lambda_n}\right|^{\frac{1}{2}}$$

Te su vrijednosti recipročne i svojstvene vrijednosti su realne ($|\lambda_i| = \pm\lambda_i$) pa uz svojstvo (1) vrijedi:

$$f(|\mu_0 \lambda_n|) = f\left(\left|\frac{1}{\mu_0 \lambda_n}\right|\right) = f(|\mu_0 \lambda_1|)$$

Svojstvo (2) govori da su te vrijednosti svojstvene vrijednosti najvećeg modula matrice X_1:

$$\{\, \mu_0\lambda_i \,:\, 0 \leq |\mu_0\lambda_n| \leq |\mu_0\lambda_i| \leq 1 \,\} \quad \Rightarrow \quad f(\,|\mu_0\lambda_n|\,) \geq f(\,|\mu_0\lambda_i|\,)$$

$$\{\, \mu_0\lambda_j \,:\, 1 \leq |\mu_0\lambda_j| \leq |\mu_0\lambda_1| \,\} \quad \Rightarrow \quad f(\,|\mu_0\lambda_1|\,) \geq f(\,|\mu_0\lambda_j|\,)$$

Zato X_1 ima svojstvene vrijednosti koje zadovoljavaju:

$$|\,\lambda_n^{(1)}\,| \leq |\,\lambda_{n-1}^{(1)}\,| \leq \cdots \leq |\,\lambda_2^{(1)}\,| = |\,\lambda_1^{(1)}\,|$$

U svakoj slijedećoj iteraciji se povećava broj najvećih po modulu svojstvenih vrijednosti barem za 1, sve dok nakon $d-1$ iteracija X_{d-1} ima sve svojstvene vrijednosti jednakog modula. Vrijedi da su svojstvene vrijednosti za $\mu_{d-1}X_{d-1}$ po modulu između:

$$|\,\mu_{d-1}\lambda_n^{(d-1)}\,| = \left|\frac{\lambda_n^{(d-1)}}{\lambda_1^{(d-1)}}\right|^{\frac{1}{2}} = 1 \quad i \quad |\,\mu_{d-1}\lambda_1^{(d-1)}\,| = \left|\frac{\lambda_1^{(d-1)}}{\lambda_n^{(d-1)}}\right|^{\frac{1}{2}} = 1$$

Dakle, svojstvene vrijednosti matrice $\mu_{d-1}X_{d-1}$ su jednake ± 1. Također, iz (3.13) slijedi da su svojstvene vrijednosti od X_d jednake ± 1:

$$\lambda_i^{(d)} = \frac{1}{2}\left(\mu_{d-1}\lambda_i^{(d-1)} + \left(\mu_{d-1}\lambda_i^{(d-1)}\right)^{-1}\right) = \frac{1}{2}(\pm 1 \pm 1) = \pm 1$$

Znači, μ_d je jednako 1 po definiciji. Također, svaki slijedeći μ_k, $k > d$ je 1 pa se skalirane iteracije poklapaju sa običnim iteracijama. Po Teoremu 3.3 slijedi da će iteracija nakon p koraka biti jednaka $sign(A)$.

□

Algoritam:

Potrebnan je skalar *kriterij skaliranja* koji regulira kada se prebacuje sa skaliranih na neskalirane iteracije. Drugi kriterij je račun *kriterij zaustavljanja* za zaustavljanje ako su dvije susjedne iteracije dovoljno blizu.

▷ $X_0 = A$ i $scale = TRUE$

▷ za $i = 1,\ 2, \ldots$

$B_i = X_i^{-1}$

ako je $scale == TRUE$

onda μ_i postaje jedan od (3.14)-(3.16)

inače, $\mu_i = 1$

$$X_{i+1} = \frac{1}{2}\left(\mu_i X_i + \mu_i^{-1} B_i\right)$$

$$\delta_{i+1} = \frac{||X_{i+1} - X_i||_F}{||X_{i+1}||_F}$$

ako je $scale == TRUE$ i $\delta_{i+1} \leq kriterij\ skaliranja$

onda $scale = FALSE$

ako je $kriterij\ zaustavljanja == TRUE$ idi na kraj

inače, $i = i + 1$

▷ $X = X_{i+1}$

Potrebno je $2in^3$ operacija za i iteracija.

3.5 Padéove iteracije

Definicija 3.5 Za skalarnu funkciju $f(x)$, racionalna funkcija dva polinoma p i q $r_{km}(x) = \dfrac{p_{km}(x)}{q_{km}(x)}$, $k \geq deg(p)$, a $m \geq deg(q)$ je $[k/m]$ *Padéova aproksimacija funkcije* f ako:

- $r_{km} \in R_{k,m} = \left\{\text{racionalne funkcije } \frac{f}{g} :\ deg(f) \leq k,\ deg(g) \leq m\right\}$

- $q_{km}(0) = 1$
- $f(x) - r_{km}(x) = O\left(x^{k+m+1}\right)$

Za kompleksni broj koji nije na imaginarnoj osi z iz Poglavlja 3.1 i (3.4) znamo:

$$sign(z) = \frac{z}{(z^2)^{\frac{1}{2}}} = \frac{z}{(1-(1-z^2))^{\frac{1}{2}}} = \{\xi := 1 - z^2\} = \frac{z}{(1-\xi)^{\frac{1}{2}}}$$

Definiramo funkciju:

$$h(\xi) := (1-\xi)^{-\frac{1}{2}}$$

Funkcija h je specijalani slučaj hipergeometrijske funkcije, pa imaju smisla Padéove aproksimacije funkcije h:

$$r_{lm}(\xi) = \frac{p_{lm}(\xi)}{q_{lm}(\xi)}$$

Kenney i Laub su imali ideju definirati familiju iteracija ako stavimo $z = x_k$, tj. $\xi = 1 - x_k^2$:

$$x_{k+1} = f_{lm}(x_k) := x_k \frac{p_{lm}(1-x_k^2)}{q_{lm}(1-x_k^2)},\ x_0 = a$$

Pomoću tog rezultata, definiramo Padéove iteracije matrica:

$$X_{k+1} = X_k\, p_{lm}\,(I - X_k^2)\, q_{lm}\,(I - X_k^2)^{-1},\ X_0 = A \tag{3.18}$$

Teorem 3.6 *Neka matrica $A \in \mathbf{C^{n\times n}}$ nema svojstvene vrijednosti na imaginarnoj osi. Uzimajući u obzir Padéove iteracije (3.18) uz $l + m > 0$ i bilo koju konzistentnu matričnu normu vrijedi:*

1. *Za $l \geq m - 1$, ako je $||I - A^2|| < 1$, onda*

$$\lim_{k\to\infty} X_k = sign(A)\ \ i\ \ ||I - X_k^2|| < ||I - A^2||^{(l+m+1)^k}.$$

2. *Za $l = m - 1$ i $l = m$, $S := sign(A)$,*

$$(S - X_k)(S + X_k)^{-1} = ((S - A)(S + A)^{-1})^{(l+m+1)^k}$$

$$pa\ \ je\ \ \lim_{k\to\infty} X_k = S.$$

Gornji teorem o konvergenciji govori da je konvergencija u (2) globalna, dok je u slučaju (1) lokalna. Brzina konvergencije je stupnja $l + m + 1$ u oba slučaja. Iteracije u slučaju (2) se nazivaju osnovne Padéove iteracije.

3.6 Numerička stabilnost i konačnost iteracija

3.6.1 Numerička stabilnost

Pitanje stabilnosti je objašnjeno u Poglavlju 2.5.3. Ovdje to primjenjujemo na funkciju $f = sign$.

Teorem 3.7 *Neka je $A \in \mathbf{C^{n\times n}}$ matrica koja nema svojstvene vrijednosti na imaginarnoj osi i $S := sign(A)$. Neka je $X_{i+1} = g(X_i)$ iteracija koja super linearno konvergira prema $sign(X_0) = sign(A)$, za svaki X_0 dovoljno blizu S i pretpostavimo da je g neovisna o X_0. Tada su te iteracije numerički stabilne, Fréchetova derivacija od g u S je idempotentna i za neku perturbaciju H vrijedi:*

$$L_g(S, H) = L_{sign}(S, H) = \frac{1}{2}(H - SHS).$$

Dokaz:

Funkcija predznaka je idempotentna, tj. $sign(sign(A)) = sign(A)$ što ćemo pokazati pomoću 2.2 i svojstava matrične funkcije iz Teorema 2.5:

$$sign(sign(A)) = sign(Zsign(J)Z^{-1}) = Zsign(sign(J))Z^{-1}$$

Zbog toga što je $sign(sign(z)) = sign(z)$, $\forall z \in \mathbf{C} \setminus \mathbf{I}$ vrijedi da je gornji izraz jednak:

$$Zsign(J)Z^{-1} = sign(A)$$

Iz Teorema 2.16 slijedi jednakost $L_g(S) = L_{sign}(S)$, zatim zbog te jednakosti i Teorema 2.15 slijedi idempotentost od L_g u S. Iteracije $g(X_i) = X_{i+1}$ zadovoljavaju dva spomenuta Teorema pa su zato numerički stabilne. Označimo

$$F := I - (S + H)^2 = I - S^2 - SH - HS - H^2 = -(SH + HS + H^2)$$

Iz gornje relacije, za neke pozitivne konstante c i k slijedi:

$$||F|| \leq c\,||H||\ /^2$$

$$||F||^2 \leq c^2\,||H||^2$$

$$\Rightarrow O(||F||^2) \leq k\,||F||^2 \leq kc^2\,||H||^2$$

$$\Rightarrow\ O(\,||F||^2\,) = O(\,||H||^2\,)$$

Tada dobivamo za $||F|| < 1$:

$$\begin{aligned}
sign(S+H) &= (S+H)\ ((S+H)^2)^{-\frac{1}{2}} = \\
&= (S+H)\ (I-(I-(S+H)^2)\ )^{-\frac{1}{2}} = \\
&= (S+H)\ (I-F)^{-\frac{1}{2}} = \\
&= (S+H)\ (\ I+\frac{1}{2}F+O(||F||^2)\ ) = \\
&= S+H+\frac{1}{2}(S+H)F+O(||F||^2) = \\
&= S+H-\frac{1}{2}(S+H)\ (\ SH+HS+H^2)+O(||F||^2) = \\
&= S+H-\frac{1}{2}(\ H+SHS+SH^2+HSH+H^2S+H^3)+O(||F||^2) = \\
&= S+\frac{1}{2}(\ H-SHS+O(||H||^2)\ )+O(||F||^2) = \\
&= S+\frac{1}{2}(H-SHS)+O(||H||^2)
\end{aligned}$$

Prva jednakost vrijedi iz (3.5) dok u trećoj primjenjujemo Definiciju 1.19 na matrice.

□

Ako S i H ne komutiraju, uz osnovna svojstva norme:

$$||L_{sign}(S,H)|| = ||\ \frac{1}{2}(H-SHS)\ || \leq \frac{1}{2}(\ ||H||+||SHS||\ ) = \frac{1}{2}(1+||S||^2)\ ||H||$$

Za ograničenu uvjetovanost matrice $(S) = ||S||\,||S^{-1}|| = ||S||^2$, možemo zaključiti da su $L_g(S,H) = L_{sign}(S,H) \leq c||S||^2\,||H||$, za neku konstantu c, ograničene pa je iteracija stabilna.
U slučaju kada komutiraju H i S te uz involutornost matrice S dobivamo:

$$||L_{sign}(S,H)|| = ||\frac{1}{2}(H-SHS)|| = ||\frac{1}{2}(H-HS^2)|| = ||\frac{1}{2}(H-HI)|| = 0$$

iz čeg slijedi također ograničenost Fréchetovof diferencijala L_g pa je iteracija stabilna.

3.6.2 Konačnost iteracija

Najvažnije što se tiče iteracija je znati koliko je iteracija potrebno da bi se došlo do rješenja.

Lema 3.8 *Neka je matrica $A \in \mathbf{C^{n\times n}}$ sa svojstvenim vrijednostima izvan imaginarne osi i neka je $S := sign(A)$. Za neku operatorsku matričnu normu $||\cdot||$ vrijedi:*

ako je $||S(A-S)|| = \varepsilon < 1$, onda je

$$\frac{1-\varepsilon}{2+\varepsilon}||A-A^{-1}|| \leq ||A-S|| \leq \frac{1+\varepsilon}{2-\varepsilon}||A-A^{-1}|| \tag{3.19}$$

i

$$\frac{||A^2-I||}{||S||\,(\,||A||+||S||\,)} \leq \frac{||A-S||}{||S||} \leq ||A^2-I|| \tag{3.20}$$

Dokaz:

Neka je $H := A - S$. Budući da je S involutorna,

$$A = S + H = (I + HS)S.$$

Zbog toga što A i S komutiraju po Teoremu 3.1, komutiraju i H i S:

$$HS = (A-S)S = AS - S^2 = SA - S^2 = S(A-S) = SH$$

pa vrijedi:

$$\begin{aligned} H(2I+HS) &= (A-S)(I+(I+HS)) = \\ &= A + A(I+HS) - S - S(I+HS) = \\ &= A + A(I+HS) - S - (I+HS)S = \\ &= A + A(I+HS) - S - A = \\ &= A(I+HS) - S = \\ &= (A - S(I+HS)^{-1})(I+HS) = \\ &= (A - A^{-1})(I+HS) \end{aligned}$$

Pomnožimo li zadnju jednakost sa $(2I+HS)^{-1}$ te primijenimo normu i nejednakost trokuta, dobit ćemo gornju granicu u (3.19), a ako pomnožimo sa $(I+HS)^{-1}$ dobit ćemo donju granicu u (3.19).

Za drugu nejednakost koristimo

$$(A^2 - I) = (A - S)(A + S).$$

Donju granicu dobit ćemo kada na tu jednakost primijenimo normu, upotrijebimo nejednakost trokuta te podijelimo sa normom od S. Također iz gornje jednakosti slijedi da je

$$A - S = (A^2 - I)(A + S)^{-1}$$

te uz

$$A + S = 2S + A - S = 2S + S^2(A - S) = 2S(I + \frac{1}{2}S(A - S))$$

Po pretpostavci teorema vrijedi $||S(A - S)|| = \varepsilon < 1$, tj. $1 - \frac{1}{2}\varepsilon > \frac{1}{2}$ pa koristeći Definiciju 1.19 imamo:

$$\begin{aligned}
||(A + S)^{-1}|| &= \left|\left|\left(2S\left(I + \frac{1}{2}S(A - S)\right)\right)^{-1}\right|\right| = \\
&= \left|\left|\left(I + \frac{1}{2}S(A - S)\right)^{-1}\frac{1}{2}S^{-1}\right|\right| = \\
&= \left|\left|\left(I - \frac{1}{2}S(A - S) + O(\,\varepsilon^2\,)\right)\frac{1}{2}S\right|\right| = \\
&= \left|\left| I - \frac{1}{2}S(A - S) + O(\,\varepsilon^2\,)\right|\right|\frac{1}{2}||S|| = \\
&\leq \left(1 + \frac{1}{2}\varepsilon + O(\,\varepsilon^2\,)\right)\frac{1}{2}||S|| = \\
&= \frac{\frac{1}{2}||S||}{1 - \frac{1}{2}\varepsilon} = \\
&< \frac{\frac{1}{2}||S||}{\frac{1}{2}} = \\
&= ||S||
\end{aligned}$$

Nakon svega dobivamo:

$$\frac{||\,A - S\,||}{||S||} = \frac{||\,(A^2 - I)(A + S)^{-1}\,||}{||S||} \leq ||\,(A^2 - I)\,||$$

□

Gornja lema daje neke granice koje pomažu izabrati kriterij zaustavljanja za rezidualnu pogrešku i relativnu pogrešku iteracije. Primjenjiva je kao kriterij zaustavljanja jer vrijedi da svaka iteracija X_i za dovoljno veliki i u ovom radu zadovoljava $sign(X_i) = sign(A)$. Tada u gornjoj lemi zamjenimo A sa X_i.

3.7 Osjetljivost i uvjetovanost

Dekompozicija matrice pomoću predznaka je:

$$A = SN, \quad S = sign(A), \quad N = (A^2)^{\frac{1}{2}} \tag{3.21}$$

Uz komutativnost A i S te involutornosti matrice S (Teorem 3.1) vrijedi:

$$N = S^{-1}A = SA \iff N^2 = SASA = ASSA = A^2$$

Uočimo još da su svojstvene vrijednosti od SA desno od imaginarne osi jer će biti u obliku $sign(\lambda)\,\lambda$. Napokon, zbog svojstvenih vrijednosti desno od imaginarne osi, $N = (A^2)^{\frac{1}{2}}$ je dobro definirana matrica jer se to slaže sa razmatranjima u (3.5).

Faktor u dekompoziciji matrice pomoću predznaka, N, je koristan u karakterizaciji Fréchetove derivacije matrične funkcije predznaka.

Pretpostavimo da je funkcija predznaka definirana na kugli radijusa $||\Delta A||$ oko A i $S + \Delta S := sign(A + \Delta A)$. Uzevši u obzir da je $L(A, \Delta A)$ Fréchetova derivacija matrične funkcije predznaka u A u smjeru ΔA po Definiciji 2.7:

$$\Delta S - L(A, \Delta A) = o(||\Delta A||) \tag{3.22}$$

Iz gornje jednakosti uz nejednakost trokuta i definciju norme linearnog operatora dobivamo:

$$\begin{aligned} ||\Delta S|| \leq\ & || L(A, \Delta A) || + o(||\Delta A||) \leq \\ & \leq || L(A) || \, || \Delta A || + o(||\Delta A||) = \\ & = O(\, ||\Delta A||\,) + O(\, ||\Delta A||^2) = \\ & = O(\, ||\Delta A||\,) \end{aligned} \tag{3.23}$$

Promotrimo:

$$\begin{aligned} &(A + \Delta A)(S + \Delta S) = (S + \Delta S)(A + \Delta A) \\ &A\Delta S - \Delta SA = S\Delta A - \Delta AS + \Delta S\Delta A - \Delta A\Delta S \\ &A\Delta S - \Delta SA = S\Delta A - \Delta AS + o(||\Delta A||) \end{aligned} \tag{3.24}$$

Nadalje, zbog involutornosti funkcije $sign(A + \Delta A)$ je

$$(S + \Delta S)^2 = (sign(A + \Delta A))^2 = I \tag{3.25}$$

pa (3.23) i (3.25) povlače:

$$S\Delta S + \Delta SS = -(\Delta S)^2 = o(||\Delta A||). \tag{3.26}$$

Pomnožimo (3.24) slijeva sa S, uzmimo u obzir da je $N = SA$ te (3.26):

$$SA\Delta S - S\Delta SA = S^2\Delta A - S\Delta AS + o(||\Delta A||)$$

$$N\Delta S - S\Delta SA = \Delta A - S\Delta AS + o(||\Delta A||)$$

Uvrstimo izraz za $S\Delta S$ iz (3.26):

$$N\Delta S - (o(||\Delta A||) - \Delta SS)A = \Delta A - S\Delta AS + o(||\Delta A||)$$

$$N\Delta S + \Delta SN = \Delta A - S\Delta AS + o(||\Delta A||) \tag{3.27}$$

Teorem 3.9 *Fréchetova derivacija $L = L_{sign}(A, \Delta A)$ matrične funkcije predznaka zadovoljava jednakost:*

$$NL + LN = \Delta A - S\Delta AS, \tag{3.28}$$

gdje je A u dekompoziciji (3.21).

Dokaz:

Budući da su svojstvene vrijednosti od N desno od imaginarne osi, Sylvesterova jednadžba (3.28) ima jedinstveno rješenje L kao linearna funkcija od ΔA. Po (3.22) L se razlikuje od ΔS za $o(||\Delta A||)$, a uz (3.27) dobivamo da je L točno $L_{sign}(A, \Delta A)$.

□

Uvjetovanost matrične *sign* funkcije mjeri osjetljivost $sign(A)$ na greške u podacima matrice A. Svrha uvjetovanosti je da daje odgovor na pitanje koju točnost rezultata možemo očekivati pri točnom računanju s malo perturbiranim podacima. Broj uvjetovanosti funkcije *sign* pomoću (2.13) je:

$$\kappa_{sign}(A) := cond_{rel}(sign, A) =$$

$$= \lim_{\varepsilon \to 0} \sup_{||\Delta A||_F \leq \varepsilon ||A||_F} \frac{|| \, sign(A + \Delta A) - sign(A) \, ||_F}{\varepsilon \, || \, sign(A) \, ||_F} \tag{3.29}$$

Jedna od glavnih upotreba κ_{sign} je da određuje osjetljivost funkcije $sign(A)$ na perturbacije od A pomoću ograde (2.14):

$$\frac{||\, sign(A+\Delta A) - sign(A)\,||_F}{||\, sign(A)\,||_F} \leq cond_{rel}(sign, A)\frac{||\Delta A||_F}{||A||_F} + o(||\Delta A||_F) \tag{3.30}$$

Slijedeći teorem daje ogradu na uvjetovanost $S := sign(A)$.

Teorem 3.10 *Neka je $A \in \mathbf{C^{n\times n}}$ matrica koja nema svojstvene vrijednosti na imaginarnoj osi te $S := sign(A)$. Ako je $||(A-S)S||_2 < 1$, onda vrijedi:*

$$\frac{||S||_2^2 - 1}{2(1+||(A-S)S||_2)} \leq \kappa_{sign}(A)\frac{||S||_F}{||A||_F} \leq \frac{||S||_2^2 + 1}{2(1-||(A-S)S||_2)} \tag{3.31}$$

Posebno vrijedi:

$$\frac{||S||_2^2 - 1}{2} \leq \kappa_{sign}(S) \leq \frac{||S||_2^2 + 1}{2} \tag{3.32}$$

Dokaz:

Kada uvrstimo $A = S$ u (3.31), dobivamo (3.32). Dokazujemo prvu nejednakost teorema. Iz Teorema 2.11 i jednakosti (2.18) u njemu vrijedi:

$$||L_{sign}(A)||_F = \kappa_{sign}(A)\frac{||S||_F}{||A||_F}$$

Označimo $\Delta S := L = L_{sign}(A, \Delta A)$ i $G := AS - S^2 = N - I$. Vrijede relacije za omeđenost operatora:

$$||\Delta S||_F = ||L_{sign}(A, \Delta A)||_F \leq c\,||\Delta A||_F\ , \quad ||L_{sign}(A)|| \leq c$$

Tražimo c, tj. omeđujemo ΔS. Uz Teorem 3.9 i jednakost (3.28) dobivamo:

$$\begin{aligned} N\Delta S + \Delta S N &= \Delta A - S\Delta A S \\ (G+I)\Delta S + \Delta S(G+I) &= \Delta A - S\Delta A S \\ 2\Delta S &= \Delta A - S\Delta A S - G\Delta S - \Delta S G \end{aligned} \tag{3.33}$$

Primijenimo Frobeniusovu normu na gornju jednakost, upotrijebimo nejednakost trokuta i nejednakost: $||ABC||_F \leq ||A||_2||B||_F||C||_2$, za kvadratne

matrice A, B i C:

$$\begin{aligned} 2\,||\Delta S||_F &= ||\Delta A - S\Delta AS - G\Delta S - \Delta SG||_F \leq \\ &\leq ||\Delta A||_F + ||S\Delta AS||_F + ||G\Delta SI||_F + ||I\Delta SG||_F \leq \\ &\leq ||\Delta A||_F + ||S||_2^2\,||\Delta A||_F + ||G||_2\,||\Delta S||_F + ||\Delta S||_F\,||G||_2 = \\ &= ||\Delta A||_F\,(\,1 + ||S||_2^2\,) + 2\,||G||_2\,||\Delta S||_F \end{aligned} \tag{3.34}$$

Uočimo, po pretpostavci teorema $||G||_2 < 1$, tj. $(\,1 - ||G||_2\,) > 0$. Iz gornje relacije slijedi:

$$\begin{aligned} 2\,||\Delta S||_F(\,1 - ||G||_2) &\leq ||\Delta A||_F(\,1 + ||S||_2^2\,) \\ \Rightarrow\ ||\Delta S||_F &\leq \frac{(1 + ||S||_2^2\,)\,||\Delta A||_F}{2\,(\,1 - ||G||_2)} \end{aligned} \tag{3.35}$$

Konačno dobivamo:

$$||L_{sign}(A)||_F = \kappa_{sign}(A)\frac{||S||_F}{||A||_F} \leq \frac{(1 + ||S||_2^2\,)}{2\,(\,1 - ||G||_2)} \tag{3.36}$$

Iz definicije spektralne norme (1.5) i Definicije 1.12 je $\alpha := ||S||_2$ singularna vrijednost. Nadalje, vrijedi:

$$Sv = \alpha u$$

$$u^*S = \alpha v^*$$

gdje su $u \in \mathbf{C^{n\times 1}}$ i $v \in \mathbf{C^{n\times 1}}$ lijevi i desni singularni vektori. Neka je $\Delta A := vu^*$. Dakle, po definiciji Frobeniusove norme (1.4) slijedi:

$$||\Delta A||_F = ||vu^*||_F = ||uv^*||_F.$$

Uvrstimo takav ΔA u (3.33) i slijedi:

$$\begin{aligned} 2\Delta S &= vu^* - Svu^*S - G\Delta S - \Delta S = \\ &= vu^* - \alpha^2 uv^* - G\Delta S - \Delta SG \end{aligned} \tag{3.37}$$

Upotrebljavajući svojstvo norme:

$$||X - Y||_F \geq ||X||_F - ||Y||_F$$

te opet nejednakost $-||ABC||_F \geq -||A||_2||B||_F||C||_2$ imamo:

$$\begin{aligned} 2||\Delta S\,||_F &= ||\,vu^* - \alpha^2 uv^* - G\Delta S - \Delta SG\,||_F \geq \\ &\geq ||\,vu^* - \alpha^2 uv^*\,||_F - ||G\Delta SI\,||_F - ||I\Delta SG\,||_F \geq \\ &\geq ||\,\alpha^2 uv^* - vu^*||_F - 2\,||G||_2\,||\Delta S||_F \geq \\ &\geq \alpha^2\,||uv^*||_F - ||vu^*||_F - 2\,||G||_2\,||\Delta S||_F \geq \\ &\geq ||S||_2^2\,||\Delta A||_F - ||\Delta A||_F - 2\,||G||_2\,||\Delta S||_F \end{aligned} \tag{3.38}$$

Sredimo gornju relaciju:

$$\begin{aligned} 2\,||\Delta S\,||_F\,(1 + ||G||_2) &\geq (\,||S||_2^2\, - 1)\,||\Delta A||_F \\ \Rightarrow \quad ||\Delta S\,||_F\, &\geq \frac{(\,||S||_2^2\, - 1)\,||\Delta A||_F}{2\,(1 + ||G||_2)} \end{aligned} \tag{3.39}$$

Iz gornje nejednakosti se može naslutiti donja ograda Analogno kao i za gornju granicu, dobivamo donju:

$$||L_{sign}(A)||_F = \kappa_{sign}(A)\frac{||S||_F}{||A||_F} \geq \frac{||S||_2^2 - 1}{2\,(1 + ||G||_2)} \tag{3.40}$$

□

Literatura

[1] D. Bakić, *Linearna algebra*, Školska knjiga, Zagreb, 2008.

[2] N. Bosner, *Sustavi linearnih jednadžbi, Problem svojstvenih vrijednosti*, dostupno na `https://web.math.pmf.unizg.hr/~nela/nmfmpredavanja.pdf`

[3] N. J. Higham, *Functions of Matrices: Theory and Computation*, Philadelphia, SIAM., 2008.

[4] G. Muić i M. Primc, *Vektorski prostori*, dostupno na `https://web.math.pmf.unizg.hr/~gmuic/predavanja/vp.pdf`

Sažetak

Matrična predznak funkcija je objašnjena pomoću dvije definicije. Pristup dvjema definicijama se temelji na svojstvenim vrijednostima. Prva je definicija preko Jordanove forme, dok je druga pomoću interpolacijskog polinoma. Koordinatna ravnina koja se koristi je kompleksna pa je važno za $sign(A)$ s koje se strane imaginarne osi nalaze svojstvene vrijednosti od A. Funkcija $S := sign(A)$ ima specifična svojstva kao što su involutornost, dijagonalizabilnost i komutativnost sa A.

Schurov algoritam za računanje matrične funkcije predznaka koristi Schurovu dekompoziciju matrice A te primjenu $sign$ funkcije na gornjetrokutastu matricu iz dekompozicije. Složenost mu je $O(n^3)$, a sveukupno za algoritam je potrebno oko $\frac{86}{3}n^3$ operacija da bi se došlo do rješenja.

Slijedeća metoda opisana u radu je bila Newtonova. Newtonova metoda koristi Newtonove iteracije te je dokazan teorem o konvergenciji Newtonovih iteracija prema $sign(A)$. Također taj teorem pokazuje da je brzina konvergencije kvadratna te pomoću njega vidimo da će konvergencija biti sporija ako su svojstvene vrijednosti od A blizu imaginarne osi te ako je spektralni radijus $\rho(A)$ puno veći od 1. Sam algoritam metode zahtjeva otprilike $2in^3$ za i koraka Newtonovih iteracija. Složenost mu je također $O(n^3)$. Učinkoviti način da se poveća brzine kovnergencije može ponekad biti skaliranje iteracija. Zato postoje i Newtonove skalirane iteracije koje su slične običnima, osim što se kako im ime kaže množe sa nekim određenim skalarom. U tekstu su korištena tri skalara: pomoću determinante, spektralnog radijusa te norme. Dokazan je teorem o kovergenciji spektralno skaliranih iteracija prema $sign(A)$ funkciji. Naime, konvergiraju nakon $d+p-1$ koraka gdje je d broj različitih svojstvenih vrijednosti od A, a p je određen dimenzijom najvećeg Jordanovog bloka od A. Algoritam skalirane metode zahtjeva $2in^3$ za i iteracija.

Postoji još jedna vrsta iteracija koja je izvedena iz formule:

$$sign(A) := A(A^2)^{-\frac{1}{2}}$$

a to su Padéove iteracije. Temelje se na racionalnim polinomnim funkcijama. Matrična iteracija ovisi o kvadratnoj potenciji i inverzu matrice. Teorem o konvergenciji Padéoveih iteracija govori o dva slučaja: kada je konvergencija iteracija lokalna i globalna. Brzina u oba slučaja je $l + m + 1$ gdje su l i m stupnjevi polinoma koji se koriste u toj Padéoveoj aproksimaciji.

Numerička stabilnost iteracija je objašnjena pomoću teorema koji daje rezultat da su iteracije koje superlinearno konvergiraju prema $sign(A)$ numerički stabilne, Fréchetova derivacija od iteracijske funkcije u S je idempotentna i vrijedi da je Fréchetova derivacija od iteracijske funkcije jednaka Fréchetovoj derivacija od $sign$ funkcije te jednaka

$$\frac{1}{2}(H - SHS)$$

gdje je H perturbacijska matrica. Zbog tih rezultata zaključujem da za ograničenu uvjetovanost matrice $S := sign(A)$ iteracija iz gornjeg teorema je stabilna. Također, isto vrijedi ako H i S komutiraju.

Što se tiče iteracija, analizirana je i njihova konačnost, odnosno koliko je iteracija potrebno da bi se došlo do rješenja. Dokazan je teorem o granicama za rezidualnu pogrešku $||X_i - S||$ te relativnu pogrešku $\frac{||X_i - S||}{||S||}$ za iteracije X_i što nam pomaže u odabiru kriterija zaustavljanja.

Osjetljivost i uvjetovanost matrice su objašnjene pomoću relativnog broja uvjetovanosti funkcije $sign$. Teorem daje rezultat o donjoj i gornjoj granici za broj uvjetovanosti $sign$ funkcije u odnosu na matricu A te matricu S.

Summary

Matrix sign function is defined in two ways. Both definitions require values of sign function on the spectrum of A. First definition is about Jordan canonial form and the other with polynomial interpolation. Everything is based on the complex coordinate plane so for sign function it's important to know if eigenvalues are on the right or left side of the plane. Function $S := sign(A)$ has some useful properties: involution, diagonalizable matrix and commutation with A.

Schur algorithm is based on Schur decomposition. The problem is therefore to computing $sign(T)$ where matrix T is triangular matrix from decomposition. The complexity of the algorithm is $O(n^3)$. In total, $\frac{86}{3}n^3$ flops are needed to get the solution.

The next method is Newton's. It uses Newton's iterations. In that chapter, theorem about quadratically convergence of Newton's sign iterations is proven. The other result from that theorem is that iterations will converge slower if the spectral radius is much greater than 1 and also if eigenvalues of A are very close to the imaginary axe. The method requires $2in^3$ where i is the number of used iterations. The complexity of the algorithm is also $O(n^3)$. An effective way to enhance the initial speed of convergence is to scale the iterations. That's why scaled Newton's iterations exist. They are very similar to the original Newton's iterations. The only difference is that scalar is multiplied by original iteration. There are three types of that positive and real scalar: determinantal, spectral and norm. There is the theorem which tells that scaled Newton's iterations converge to $sign(A)$. The finite iteration will be after $d + p - 1$ steps where d is the number of distinct eigenvalues and p is determined with dimension of the largest Jordan block. Algorithm needs $2in^3$ flops for i iterations.

There is one more kind of the iterations which is derivatived from formula:

$$sign(A) := A(A^2)^{-\frac{1}{2}}$$

and their name is Padé iterations. They are determined on racional polynomial functions. The theorem about convergence of Padé iterations contains two cases: global and local convergence. Speed of the convergence in both cases is $l + m + 1$ where l and m are degrees of polynomials which are used in that Padé approximation.

Numerical stability of iterations is explained by theorem which gives the result that iterations which superlineary converge to $sign(A)$ are stable. Also, Fréchet derivation of the iteration function in S is idempotent and it's equal to Fréchet derivation of $sign$ function. They are both equal to:

$$\frac{1}{2}(H - SHS)$$

where H is small perturbation. Because of these results, I conclude that for bounded condition of the matrix $S := sign(A)$, iteration is stable. That is also valid when H i S commute.

Also, in relation with iterations, stopping criteria is analyzed. The theorem about residual error and relative error boundaries is also proven. Sensitivity and condition of matrices are explained by relative condition number of the function $sign$. The result from that part gives the boundaries for condition number of $sign$ function.

europass Životopis

OSOBNE INFORMACIJE

Stopić Petra

Ulica Bogoslava Šuleka 15, 10000 Zagreb (Hrvatska)

(+385) 989976975

stopicpetra.8@gmail.com

Spol Žensko | Datum rođenja 24/08/1992 | Državljanstvo hrvatsko

OBRAZOVANJE I OSPOSOBLJAVANJE

rujan 2014–danas — visoka stručna sprema

Prirodoslovno-matematički fakultet, Zagreb (Hrvatska)

Diplomski sveučilišni studij Financijska i poslovna matematika

rujan 2011–srpanj 2014 — visoka stručna sprema

sveučilišna prvostupnica matematike; univ. bacc. math.

Prirodoslovno-matematički fakultet, Zagreb (Hrvatska)

Preddiplomski sveučilišni studij Matematika

rujan 2007–srpanj 2011 — srednja stručna sprema

XV. gimnazija, Zagreb (Hrvatska)

Informatičko-matematički smjer

OSOBNE VJEŠTINE

Materinski jezik Hrvatski

Ostali jezici

	RAZUMIJEVANJE		GOVOR		PISANJE
	Slušanje	Čitanje	Govorna interakcija	Govorna produkcija	
engleski	C1	B2	B1	B2	C1

Stupnjevi: A1 i A2: Početnik - B1 i B2: Samostalni korisnik - C1 i C2: Iskusni korisnik
Zajednički europski referentni okvir za jezike

Digitalna kompetencija Jako dobro poznavanje rada na računalu, MS Office
Programiranje u: C, R, Matlab
Baze podataka: MySQL

Ostale vještine Profesionalno bavljenje rukometom u prvoligaškom hrvatskom klubu od 2010. do 2016.

Vozačka dozvola B