Desanka P. Radunović

N U M E R I Č K E M E T O D E

AKADEMSKA MISAO
Beograd, 2003

Predgovor

Materijal koji obuhvata ova knjiga se uglavnom zasniva na programima jednosemestralnog kursa "Uvod u numeričku matematiku" i dvosemestralnog kursa "Numeričke metode", koji se predaju na Matematičkom fakultetu u Beogradu. Stoga je ona napisana, u prvom redu, kao udžbenik za studente Matematičkog fakulteta. Težište je stavljeno na metode koje su pogodne za kompjutersku primenu, i koje su osnov većine programskih paketa iz ove oblasti.

Izloženi materijal je podeljen u devet poglavlja. Prvo je uvodnog karaktera i definiše osnovne pojmove – pojam i vrste grešaka, približne brojeve, greške funkcija. U sledećem poglavlju se obrađuju različiti vidovi interpolacije, detaljno Lagrangeova, Hermiteova i interpolacija splajnovima, a informativno racionalna, trigonometrijska i interpolacija funkcija više promenljivih. Takođe je objašnjena primena interpolacionih polinoma za rešavanje problema inverzne interpolacije i numeričkog diferenciranja. Primena interpolacionih polinoma u približnom izračunavanju integrala data je, zbog svog značaja, u posebnom poglavlju (trećem). Izvedene su Newton–Cotessove formule (trapezna, Simpsonova,...) i dat opšti algoritam izvođenja kvadraturnih formula Gaussovog tipa. Data je ocena greške ovih formula, i ukazano na mogućnosti rešavanja nesvojstvenih integrala formulama Gaussovog tipa. Opštiji pristup aproksimaciji funkcija razmatra se u četvrtom poglavlju. Posebno se razmatraju srednjekvadratna i ravnomerna aproksimacija, kao i metoda najmanjih kvadrata i diskretna Fourierova transformacija. U okviru ovoga poglavlja, ukratko su date matematičke osnove Brze Fourierove transformacije (FFT).

Metode linearne algebre su obrađene u petom i šestom poglavlju. Za rešavanje sistema linearnih jednačina, izračunavanje determinanti i inverznih matrica dati su Gaussova eliminacija (posebno za sisteme sa trodijagonalnim matricama), LU dekompozicija i dekompozicija Choleskog. Posebna pažnja je posvećena problemu numeričke stabilnosti. Za rešavanje loše uslovljenih, ili čak singularnih sistema, prikazana je metoda singularne dekompozicije. Metode za rešavanje problema sopstvenih vrednosti matrica izdvojene su u posebno poglavlje (šesto). Pored klasičnih metoda za rešavanje potpunog i delimičnog problema, obrađene su i savremene metode, kao na primer Givensova, Jacobijeva, Householderova, LR i QR-algoritam.

Sedmo poglavlje je posvećeno rešavanju nelinearnih jednačina i sistema. Analizirana je konvergencija i ocena greške familije dvoslojnih iterativnih metoda, sa posebnim osvrtom na njihovu primenu na sisteme linearnih jednačina. Detaljno je obrađena i metoda Newtona. Razmotrene su metode regula falsi, sečice i polovljenja intervala, kao posebne metode za rešavanje jedne jednačine. Specijalno, za rešavanje

algebarskih jednačina data je metoda Bairstowa. Ukazano je na mogućnost primene metoda minimizacije u rešavanju ovih problema, i dat kratak prikaz.

Osmo i deveto poglavlje su posvećeni metodama za rešavanje običnih diferencijalnih jednačina. Zbog svoje različite prirode, metode za rešavanje Cauchyevih problema i metode za rešavanje graničnih problema su razdvojene u posebna poglavlja. U osmom poglavlju se razmatraju numeričke metode za rešavanje Cauchyevih problema: aproksimativne metode, metode tipa Runge–Kutta i prediktor –korektor metode. Posebna pažnja je posvećena problemima tačnosti i numeričke stabilnosti. Poslednje poglavlje ove knjige, deveto, je posvećeno metodama za rešavanje graničnih problema za obične diferencijalne jednačine. Razmatraju se metode gađanja, metode konačnih razlika i varijacione metode. U okviru ovih poslednjih, posebna pažnja je posvećena metodi konačnog elementa.

Brojnim jednostavnim primerima ilustrovane su teorijske i numeričke karakteristike metoda.

Obzirom da se metode Numeričke matematike koriste sve više u raznim oblastima nauke i prakse, a da ima malo literature na srpskom jeziku iz ove oblasti, smatram da će ova knjiga korisno poslužiti i studentima drugih fakulteta koji izučavaju Numeričku matematiku, kao i stručnjacima koji se u svakodnevnom radu njome koriste. Čitanje knjige ne zahteva posebno predznanje, osim osnova Matematičke analize i Linearne algebre.

Koristim ovu priliku da se zahvalim kolegi prof. dr Bošku Jovanoviću, na pažljivom čitanju rukopisa knjige i korisnim primedbama i sugestijama.

Beograd, novembra 1990. D. P. Radunović

Predgovor drugom izdanju

U drugom izdanju knjige ispravljene su sve greške koje su uočene u prvom izdanju. Četvrto poglavlje, koje se odnosi na aproksimaciju funkcija, dopunjeno je odeljkom o talasićima, pošto se oni poslednjih desetak godina sve više koriste u različitim oblastima primene matematike.

Beograd, septembra 1997. D. P. Radunović

Predgovor trećem izdanju

U trećem izdanju knjige dorađen je samo deo četvrtog poglavlja, koji se odnosi na talasiće (§4.6), i, svakako, ispravljene novouočene greške.

Beograd, novembra 2003. D. P. Radunović[1]

[1] dradun@matf.bg.ac.yu

Sadržaj

1

Uvodni pojmovi o numeričkoj matematici

Sve veći broj realnih problema u svim oblastima života danas se rešava matematičkim modeliranjem, zahvaljujući pre svega intenzivnom razvoju računarske tehnike. Umesto da se vrši veliki broj eksperimenata, što je često dug i skup put, formira se matematički model kojim se simulira određeni proces ili pojava. Model se obično sastoji od skupa jednačina kojima treba da su opisane sve važnije pojave ili procesi značajni za postavljeni problem. Karakteristike sredine ili objekata izražene su kroz koeficijente jednačina.

Sledeći korak je nalaženje rešenja formulisanog modela matematičkim metodama. U slučaju prostih i dosta grubih modela, rešenje se često može odrediti analitički. Međutim, dobri modeli su najčešće vrlo složeni, te se rešenja ne mogu naći analitičkim metodama. Tada se koriste metode numeričke matematike. Od kakvog su one značaja govori i činjenica da su se njima bavili i mnogi istaknuti matematičari, kao što su Newton, Euler, Gauss, Lagrange, Hermite i drugi. Posebno intenzivan razvoj ova oblast matematike doživljava pojavom elektronskih računskih mašina (1940. godine). Mogućnost da se veliki broj računskih operacija realizuje za kratko vreme dozvoljava numeričko rešavanje novih klasa zadataka, na primer onih opisanih parcijalnim diferencijalnim jednačinama.

I dok je u klasičnoj matematici osnovni cilj utvrditi pod kojim uslovima postoji rešenje nekog zadatka i koje su osobine tog rešenja, zadatak numeričke matematike je efektivno nalaženje rešenja sa zadatom tačnošću. Ta tačnost treba da bude nešto veća od tačnosti koju obezbeđuje matematički model, ali ne ni suviše visoka, jer se tačnost približnog rešenja i tako neće povećati s obzirom na usvojeni model.

1.1 Pojam i vrste grešaka

Šta znači numeričko rešavanje zadatka i greška rešenja? Simbolički se problem određivanja neke veličine y na osnovu date veličine x može zapisati u obliku

$$y = A(x).$$

Ako je operator A toliko složen da se rešenje ne može eksplicitno napisati ili tačno izračunati, zadatak rešavamo približno. Na primer, neka operator A predstavlja integral,

$$y = \int_a^b x(t)\,dt,$$

pri čemu ovaj integral nije moguće izračunati analitički. Možemo zameniti x polinomom ili nekom drugom funkcijom $\overline{x}$ čiji se integral može izračunati, ili pak, možemo zameniti integral sumom $\sum_i x(t_i)\Delta t_i$, koju možemo izračunati. Znači, u ovom slučaju, približna metoda se sastoji u zameni date veličine $x(t)$ njoj bliskom veličinom $\overline{x}$ i (ili) u zameni operatora A bliskim operatorom $\overline{A}$, kako bi se vrednost $\overline{y} = \overline{A}(\overline{x})$ mogla izračunati. Greškom se ocenjuje koliko je približno rešenje $\overline{y}$ blisko tačnom rešenju y. Šta se podrazumeva pod pojmom "blisko" zavisi od prostora u kome je definisan problem i u njemu uvedene metrike.

Uzroci greške mogu biti različiti i, s obzirom na poreklo greške, ona može biti
- neotklonjiva greška,
- greška metode ili greška odsecanja, i
- računska greška ili greška zaokrugljivanja.

Neotklonjiva greška nastaje zbog nedostataka matematičkog modela ili grešaka ulaznih podataka. Neotklonjiva je u tom smislu da ne zavisi od primenjenog matematičkog aparata.

Greška metode nastaje usled toga što se operator ili ulazne veličine zamenjuju približnim veličinama (izvod–razlikom, funkcija–polinomom, itd.), ili što se beskonačni iterativni proces zamenjuje konačnim algoritmom. Numeričke metode se obično konstruišu tako da u njima postoji neki parametar čijim izborom se može menjati greška metode, u tom smislu da greška teži nuli kada taj parametar teži određenoj granici. Detaljnije će biti reči o ovim greškama kada budu izložene konkretne metode.

Sada će biti reči o računskoj grešci.

1.2 Približni brojevi

Neki brojevi, na primer π, $\sqrt{2}$, e, $\frac{2}{3}, \ldots$, ne mogu da se zapišu pomoću konačnog broja cifara. Stoga smo prinuđeni da u izračunavanjima koristimo samo njihove približne vrednosti, tj. brojeve koji su određeni odgovarajućim konačnim nizom cifara. Kada se za obradu podataka koriste računske mašine, zbog načina zapisa

brojeva u njima, i rezultati računskih operacija sa tačnim brojevima mogu biti približni brojevi.

Naime, digitalni računari za interni zapis broja koriste fiksirani broj mesta n. Taj broj se naziva dužina reči i zavisi od tehničkih karakteristika računara. I pri fiksiranoj dužini reči, postoje različiti načini zapisa broja. Zapis *u fiksnom zarezu* je definisan prirodnim brojevima n_1 i n_2, $n_1 + n_2 = n$, tako da se broj zapisuje sa n_1 cifara ispred i n_2 cifara iza decimalne tačke (ili binarne, ako se koristi binarni sistem). Položaj decimalne (binarne) tačke je fiksiran.

PRIMER 1. Ako je $n = 9$, $n_1 = 4$ i $n_2 = 5$, onda je reč sa dekadnim zapisom u fiksnom zarezu broja 31.207

| 0031 \| 20700 |.

Mnogo češće se koristi zapis broja *u pokretnom zarezu*. Položaj decimalne (binarne) tačke nije fiksiran, već se on u odnosu na prvu cifru zapisa određuje zadavanjem eksponenta. Drugim rečima, svaki realan broj se prikazuje u obliku

$$a = p \cdot 10^q \quad (a = p \cdot 2^q), \qquad |p| < 1,\ q \text{ ceo broj}, \tag{1}$$

gde je p mantisa, a q eksponent. Brojevi m, broj cifara mantise, i e, broj cifara eksponenta, su fiksirani i $m + e = n$.

PRIMER 2. Ako je $n = 10$, $m = 7$ i $e = 3$, zapisi u pokretnom zarezu broja iz primera 1 mogu biti

| 3120700 \| 002 |, | 0312070 \| 003 |, ...

Očigledno je da zadavanjem brojeva m i e zapis broja u pokretnom zarezu nije jednoznačno određen. Stoga se definiše *normalizovani* zapis broja u pokretnom zarezu – zapis u kome prva cifra mantise mora biti različita od nule, tj. u (1) je $|p| \geq 10^{-1}$ ($|p| \geq 2^{-1}$ u slučaju binarnog zapisa). U primeru 2 prvi navedeni zapis je normalizovani zapis. Dakle, u najvećem broju slučajeva, svaki broj u računaru je predstavljen normalizovanim zapisom u pokretnom zarezu. Ukoliko broj ima više od m cifara, njegov normalizovani zapis u računaru predstavlja samo približnu vrednost datog broja, tj. vrednost broja datu sa određenom greškom. Greška će imati uticaja na izračunavanja u kojima učestvuje ovaj broj, te ćemo je detaljnije analizirati.

Ako je a tačna vrednost neke veličine, a $\overline{a}$ njena približna vrednost onda je veličina $|a - \overline{a}|$ apsolutna, a $|a - \overline{a}|/|a|$ relativna greška, i

$$\begin{aligned} |a - \overline{a}| &\leq \Delta(\overline{a}) \qquad \textit{granica apsolutne greške}, \\ \left|\tfrac{a-\overline{a}}{a}\right| &\leq \delta(\overline{a}) \qquad \textit{granica relativne greške}. \end{aligned} \tag{2}$$

U praksi, poznate su samo granice apsolutne ili relativne greške približnog broja $\overline{a}$, te se često $\Delta(\overline{a})$ naziva apsolutnom, a $\delta(\overline{a})$ relativnom greškom približnog broja.

Procentualna greška je $\delta(\overline{a}) \cdot 100$, a *promilna* greška je $\delta(\overline{a}) \cdot 1000$.

Pošto tačna vrednost a obično nije poznata u praksi se kao granica relativne greške koristi količnik

$$\delta(\overline{a}) = \frac{\Delta(\overline{a})}{|\overline{a}|}.$$

Značajne cifre broja su sve cifre njegovog zapisa, polazeći od prve nenula cifre sa leve strane. To znači, ako je u dekadnom zapisu broja $\overline{a}$,

$$\overline{a} = \pm(\alpha_1 10^n + \cdots + \alpha_k 10^{n-k+1} + \cdots + \alpha_m 10^{n-m+1}), \tag{3}$$

cifra $\alpha_1 \neq 0$, onda su sve cifre $\alpha_1, \ldots, \alpha_m$ značajne.

PRIMER 3. U broju $\overline{a} = 0.03120700$ sve cifre, izuzev prve dve nule, su značajne. Prve dve nule nisu značajne cifre jer broj može da se napiše i bez njih, na primer u obliku $\overline{a} = 3.120700 \cdot 10^{-2}$. Poslednje dve nule su značajne cifre jer ukazuju na tačnost sa kojom je broj dat.

Za značajnu cifru broja se kaže da je *sigurna cifra* ako apsolutna greška broja nije veća od dekadnog činioca koji odgovara toj cifri, tj. α_k je sigurna cifra ako je

$$\Delta(\overline{a}) \leq \omega \cdot 10^{n-k+1}, \qquad 0 < \omega \leq 1. \tag{4}$$

Pri tome, ako je $\omega \leq \frac{1}{2}$ cifra je sigurna u užem smislu, a ako je $\frac{1}{2} < \omega \leq 1$ ona je sigurna u širem smislu. Ako je cifra α_k sigurna, onda su i sve cifre $\alpha_1, \ldots, \alpha_{k-1}$ sigurne cifre.

PRIMER 4. Ako se zna da je $\Delta(\overline{a}) = 0.5 \cdot 10^{-5}$ apsolutna greška približnog broja $\overline{a} = 0.03120700$, onda su, s obzirom na (4), sigurne cifre $3, 1, 2$ i 0. Poslednje tri cifre $(7, 0, 0)$ nisu sigurne, jer u broju a čija je $\overline{a}$ približna vrednost, umesto ovih cifara mogu stajati i ma koje druge. Naime, s obzirom na definiciju (2) apsolutne greške, a se nalazi u intervalu

$$\begin{aligned} 0.03120700 - 0.5 \cdot 10^{-5} \leq & a \leq 0.03120700 + 0.5 \cdot 10^{-5} \qquad \text{tj.} \\ 0.03120200 \leq & a \leq 0.03121200, \end{aligned}$$

te se poslednje tri cifre brojeva a i $\overline{a}$ mogu razlikovati.

Stoga cifre koje nisu sigurne ne treba ni pisati, jer nepotrebno opterećuju izračunavanja. Pri odbacivanju cifara koje nisu sigurne, poslednja sigurna cifra broja se menja tako da bude sigurna u užem smislu. Naime, poslednja sigurna cifra α_k se neće menjati ako je $\alpha_{k+1} < 5$ i ako je $\alpha_{k+1} = 5$, a α_k parno. U ostalim slučajevima se α_k povećava za jedan. U primeru 4, posle odbacivanja cifara koje nisu sigurne, biće $\overline{a} = 0.03121$.

Između broja sigurnih cifara i relativne greške postoji sledeća veza:

$$\frac{\omega}{(\alpha_1 + 1)10^k} < \delta(\overline{a}) \leq \frac{\omega}{\alpha_1 10^{k-1}}, \qquad 0 < \omega \leq 1,$$

gde je k broj sigurnih cifara broja $\overline{a}$, a α_1 njegova prva sigurna cifra. Zaista, s obzirom da je cifra α_k poslednja sigurna cifra broja $\overline{a}$, prema (4) je

$$\omega 10^{n-k} < \Delta(\overline{a}) \leq \omega 10^{n-k+1}.$$

Deljenjem ove dvostruke nejednakosti sa $|\overline{a}| \neq 0$ i korišćenjem reprezentacije (3), dobijamo

$$\frac{\omega 10^{n-k}}{\alpha_1 10^n + \cdots + \alpha_k 10^{n-k+1}} < \delta(\overline{a}) \leq \frac{\omega 10^{n-k+1}}{\alpha_1 10^n + \cdots + \alpha_k 10^{n-k+1}}.$$

Kako je $0 \leq \alpha_2 10^{n-1} + \cdots + \alpha_k 10^{n-k+1} < 10^n$, to je

$$\frac{\omega 10^{n-k}}{\alpha_1 10^n + 10^n} < \delta(\overline{a}) \leq \frac{\omega 10^{n-k+1}}{\alpha_1 10^n},$$

odakle sledi tvrđenje.

Stoga, dok apsolutna greška ukazuje na broj sigurnih decimalnih cifara približnog broja, relativna greška ukazuje na ukupan broj njegovih sigurnih cifara.

PRIMER 5. U primeru 4, u broju 0.03120700 datom sa tačnošću $\Delta(\overline{a}) = 0.5 \cdot 10^{-5}$ sigurne cifre su, kao što smo već pokazali, 3, 1, 2 i 0, pri čemu se 0 menja u 1 posle odbacivanja cifara koje nisu sigurne. Dakle, s obzirom na zadatu tačnost je $\overline{a} = 0.03121$, tj. broj ima četiri sigurne cifre. Njegova relativna greška je $\delta(\overline{a}) = 1.6 \cdot 10^{-4}$.

Računanje sa približnim brojevima utiče na grešku konačnog rezultata. Ako se računska greška ne akumulira, kažemo da je numerički algoritam *stabilan*. U protivnom, algoritam je *nestabilan* i zbog akumuliranja računske greške javlja se velika greška konačnog rezultata. Konstrukcija stabilnih algoritama je jedan od osnovnih zadataka teorije numeričkih metoda.

PRIMER 6. Potrebno je izračunati vrednosti integrala

$$I_n = \int_0^1 \frac{x^n}{x+10}\, dx, \qquad n = 0, 1, 2, \ldots .$$

Jedan od načina da se to uradi je pomoću rekurentne formule

$$I_0 = \ln 1.1, \qquad I_n + 10\, I_{n-1} = \frac{1}{n}, \quad n = 1, 2, \ldots . \tag{5}$$

I_0 može biti izračunato samo približno, tj. sa određenom greškom, te će I_1 biti izračunato sa deset puta većom greškom, jer je $I_1 = 1 - 10\, I_0$, I_2 sa sto puta većom greškom, itd. Dakle, rekurentnom formulom (5) definisan je nestabilan algoritam, iako nikakva aproksimacija nije vršena.

Sa druge strane, algoritam

$$I_n = \frac{0.1}{n+1} - \frac{0.01}{n+2} + \frac{0.001}{n+3} - \cdots, \qquad n = 0, 1, \ldots, \tag{6}$$

koji je dobijen razvojem podintegralne funkcije u red

$$\frac{x^n}{10(1+\frac{x}{10})} = 0.1 \left(x^n - \frac{x^{n+1}}{10} + \frac{x^{n+2}}{100} - \cdots \right),$$

je stabilan. Štaviše, alternativni red (6) brzo konvergira, te se sa svega nekoliko sabiraka može postići zadovoljavajuća tačnost. Poređenja radi, u sledećoj tabeli je dato nekoliko vrednosti integrala izračunatih formulama (5) i (6):

n	2	8	13
form.(5)	0.03102	0.00977	42.92151
form.(6)	0.03103	0.01020	0.00654

Često je uzrok nestabilnosti numeričkih algoritama gubitak sigurnih cifara do koga dolazi oduzimanjem bliskih brojeva.

PRIMER 7. Manji koren kvadratne jednačine $x^2 - 140x + 1 = 0$ je prema formuli jednak $x_2 = 70 - \sqrt{4899}$. Ako se brojevi zapisuju sa četiri sigurne cifre, onda je $\sqrt{4899} = 69.99$, te je približna vrednost korena $\overline{x}_2 = 70 - 69.99 = 0.01$. Dakle, rezultat je dobijen sa samo jednom sigurnom cifrom, tj. relativnom greškom $\delta(\overline{x}_2) = 1 = 100\%$, što znači da je korišćeni algoritam nestabilan.

Stabilan algoritam za izračunavanje ovog korena

$$x_2 = \frac{70^2 - 4899}{70 + \sqrt{4899}} = \frac{1}{70 + \sqrt{4899}} \approx \frac{1}{70 + 69.99} = \frac{1}{140.0} = 0.007143$$

omogućava dobijanje rezultata takođe na četiri sigurne cifre, tj. sa relativnom greškom $1.4 \cdot 10^{-4}$.

Primerima 6 i 7 ilustrovani su nestabilni i stabilni numerički algoritmi. Moguće je, međutim, da i sam matematički model bude nestabilan, tj. da male promene ulaznih podataka dovode do velikih promena rezultata. Za takve modele se kaže da su *loše uslovljeni*.

PRIMER 8. Opšte rešenje diferencijalne jednačine $y''(x) = y(x)$ je

$$y(x) = \tfrac{1}{2}\big(y(0) + y'(0)\big)e^x + \tfrac{1}{2}\big(y(0) - y'(0)\big)e^{-x}, \tag{7}$$

dok je partikularno rešenje koje zadovoljava uslove $y(0) = 1$, $y'(0) = -1$ jednako $y(x) = e^{-x}$. Međutim, mala greška u ulaznim podacima $y(0)$ i $y'(0)$ može dovesti do toga da se prvi sabirak u izrazu (7) ne anulira, te se u rešenju pojavljuje i član oblika $\epsilon\, e^x$, koji za veće vrednosti x unosi veliku grešku u približno rešenje.

Posebno nepogodni za numeričko rešavanje su tzv. *nekorektni zadaci*.

PRIMER 9. Rešenje Cauchyevog zadatka

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \qquad u(x,0) = 0, \quad \frac{\partial u}{\partial y}(x,0) = \varphi(x)$$

u poluravni $y \geq 0$ je trivijalno rešenje $\overline{u}(x,y) \equiv 0$ ako je $\varphi(x) \equiv \overline{\varphi}(x) \equiv 0$. Ako je $\varphi(x) \equiv \varphi_n(x) = \frac{1}{n}\cos nx$, onda je rešenje zadatka $u_n(x,y) = \frac{1}{n^2}\cos nx \sinh ny$. Očigledno je da $\varphi_n(x)$ ravnomerno teži ka $\overline{\varphi}(x)$ kada $n \to \infty$; pri tome, ako je $y \neq 0$, $u_n(x,y)$ postaje neograničeno i ne teži ka $\overline{u}(x,y)$.

Neposredna primena numeričkih metoda na ovakve zadatke je besmislena, jer će se greška, koja se neminovno javlja u toku izračunavanja, uvećati toliko da će se dobiti neupotrebljiv rezultat. Posebne metode za rešavanje nekorektnih zadataka se zasnivaju na rešavanju ne polaznog, već bliskog njemu pomoćnog zadatka koji je korektno postavljen. Pri tome, pomoćni zadatak zavisi od nekog parametra tako da, kada ovaj teži nuli, rešenje pomoćnog zadatka teži rešenju polaznog zadatka. Ovaj postupak naziva se *regularizacija* nekorektnog zadatka.

1.3 Greške približnih vrednosti funkcija

Neka je y funkcija parametara $(a_1, \ldots, a_n) \in G$, $y = y(a_1, \ldots, a_n)$, i neka je $\overline{y}$ približna vrednost za y. *Apsolutna greška* veličine $\overline{y}$ je

$$A(\overline{y}) = \sup_{(a_1,\ldots,a_n)\in G} |y(a_1, \ldots, a_n) - \overline{y}|, \tag{8}$$

a *relativna greška* je $\frac{A(\overline{y})}{|\overline{y}|}$.

Ako je oblast G n-dimenzioni pravougaonik

$$|a_k - \overline{a}_k| \le \Delta(\overline{a}_k), \qquad k = 1, \ldots, n,$$

$\overline{y} = y(\overline{a}_1, \ldots, \overline{a}_n)$, i ako je y neprekidno diferencijabilna funkcija svojih argumenata, prema Lagrangeovoj formuli je

$$y(a_1, \ldots, a_n) - \overline{y} = \sum_{k=1}^{n} \frac{\partial y}{\partial a_k}\big(\overline{a}_1 + \theta(a_1 - \overline{a}_1), \ldots, \overline{a}_n + \theta(a_n - \overline{a}_n)\big)\,(a_k - \overline{a}_k),$$

$$0 \le \theta \le 1.$$

Stoga je, na osnovu (8),

$$A(\overline{y}) \le \sum_{k=1}^{n} \sup_{G} |\tfrac{\partial y}{\partial a_k}(a_1, \ldots, a_n)|\, \Delta(\overline{a}_k). \tag{9}$$

U praksi se umesto ocene (9) koristi tzv. *linearna ocena* apsolutne greške funkcije

$$\Delta(\overline{y}) = \sum_{k=1}^{n} |\tfrac{\partial y}{\partial a_k}(\overline{a}_1, \ldots, \overline{a}_n)|\, \Delta(\overline{a}_k). \tag{10}$$

Pri tome je ([2])

$$\Delta(\overline{y}) + \epsilon_1(\rho) \le A(\overline{y}) \le \Delta(\overline{y}) + \epsilon_2(\rho),$$

gde je

$$\rho = \sqrt{[\Delta(\overline{a}_1)]^2 + \cdots + [\Delta(\overline{a}_n)]^2} \qquad \text{i} \quad \epsilon_j = o(\rho), \quad j = 1, 2,$$

što znači da je ocena (10) zadovoljavajuća za male apsolutne greške argumenata.

PRIMER 10. Odrediti grešku vrednosti funkcije $y = a^{10}$ za $\overline{a} = 1$ i $\Delta(\overline{a}) = 10^{-3}$.

Kako je $\overline{y} = 1$, $\sup_{|a-1|\leq 10^{-3}} |\frac{dy}{da}(a)| = 10.09$ i $\frac{dy}{da}(1) = 10$, to je, prema (8), apsolutna greška funkcije

$$A(\overline{y}) = \sup_{|a-1|\leq 10^{-3}} |a^{10} - 1| = 1.001^{10} - 1 = 0.010045,$$

ocena ove greške izrazom (9)

$$A(\overline{y}) \leq \sup_{|a-1|\leq 10^{-3}} |\frac{dy}{da}(a)|\, \Delta(\overline{a}) = 10.09 \cdot 10^{-3} = 0.01009,$$

a linearna ocena greške (10) je

$$\Delta(\overline{y}) = |\frac{dy}{da}(1)|\, \Delta(\overline{a}) = 1 \cdot 10^{-3} = 0.01.$$

U ovom slučaju nema značajnije razlike između navedenih ocena.

Ako je, međutim, $\Delta(\overline{a}) = 10^{-1}$, apsolutna greška je $A(\overline{y}) = 1.5$, ocena ove greške izrazom (9) je $A(\overline{y}) \leq 2.3$, a linearna ocena greške je $\Delta(\overline{y}) = 1$. Kada je relativna greška približne vrednosti funkcije velika (u ovom slučaju je preko 100%), razlike u pojedinim ocenama su veće.

Iz opšteg izraza za grešku funkcije se mogu oceniti greške koje nastaju pri standardnim operacijama sa približnim brojevima.

Linearna ocena apsolutne greške zbira ili razlike jednaka je zbiru apsolutnih grešaka argumenata. Zaista, ova funkcija se može predstaviti izrazom

$$y = \gamma_1 a_1 + \cdots + \gamma_n a_n,$$

gde su γ_k, $k = 1, \ldots, n$, konstante ± 1. Kako je $\frac{\partial y}{\partial a_k}(a_1, \ldots, a_n) = \gamma_k$ za svako $(a_1, \ldots, a_n)$, to je $\Delta(\overline{y}) = \sum_{k=1}^{n} \Delta(\overline{a}_k)$.

Linearna ocena relativne greške proizvoda ili količnika jednaka je sumi relativnih grešaka argumenata. Uzmimo opštiji oblik funkcije

$$y = a_1^{e_1} \cdot \cdots \cdot a_n^{e_n}, \tag{11}$$

pri čemu su, u slučaju proizvoda ili količnika, vrednosti e_k, $k = 1, \ldots, n$, jednake ± 1. Tada je $\frac{\partial y}{\partial a_k}(\overline{a}_1, \ldots, \overline{a}_n) = e_k \overline{y}/\overline{a}_k$, pa je

$$\Delta(\overline{y}) = \sum_{k=1}^{n} |e_k||\overline{y}| \frac{\Delta(\overline{a}_k)}{|\overline{a}_k|},$$

tj., deljenjem sa $|\overline{y}| \neq 0$, dobijamo da je

$$\delta(\overline{y}) = \frac{\Delta(\overline{y})}{|\overline{y}|} = \sum_{k=1}^{n} |e_k| \delta(\overline{a}_k). \tag{12}$$

Kada je y proizvod ili količnik $|e_k| = 1$, te je tvrđenje dokazano. Očigledno, ocena (12) važi i za opštiji oblik stepene funkcije (11).

1.4 Obratan problem greške

Pod obratnim problemom greške se podrazumeva nalaženje granica dopustivih grešaka argumenata pri kojima greška funkcije ne prelazi dozvoljenu vrednost. Zadatak je jednoznačno rešiv samo za funkciju jednog argumenta $y = y(a)$. Ako je ta funkcija diferencijabilna, onda je

$$y = \overline{y} + y'(\xi)\,(a - \overline{a}) \qquad \text{gde je} \;\; \xi = \overline{a} + \theta(a - \overline{a}), \quad 0 \le \theta \le 1,$$

te je, za $y'(\xi) \neq 0$,

$$a - \overline{a} = \frac{y - \overline{y}}{y'(\xi)}.$$

Približno, granica apsolutne greške argumenta je određena relacijom

$$\Delta(\overline{a}) = \frac{\Delta(\overline{y})}{|y'(\overline{a})|}, \qquad \text{za} \quad y'(\overline{a}) \neq 0.$$

Ako y zavisi od više argumenata, $y = y(a_1, \ldots, a_n)$, onda se zadavanjem greške funkcije zadaje samo jedna veza između n nepoznatih $\Delta(\overline{a}_1), \ldots, \Delta(\overline{a}_n)$. Ako je zadata linearna ocena apsolutne greške funkcije (10), dodatni uslovi koje apsolutne greške argumenata treba da zadovoljavaju obično se definišu na jedan od sledećih načina:

(*i*) *Princip jednakih uticaja*

$$|\tfrac{\partial y}{\partial a_1}(\overline{a}_1, \ldots, \overline{a}_n)|\Delta(\overline{a}_1) = \cdots = |\tfrac{\partial y}{\partial a_n}(\overline{a}_1, \ldots, \overline{a}_n)|\Delta(\overline{a}_n).$$

Onda je

$$\Delta(\overline{y}) = n|\tfrac{\partial y}{\partial a_k}(\overline{a}_1, \ldots, \overline{a}_n)|\Delta(\overline{a}_k),$$

te je

$$\Delta(\overline{a}_k) = \frac{\Delta(\overline{y})}{n|\frac{\partial y}{\partial a_k}(\overline{a}_1, \ldots, \overline{a}_n)|}, \qquad k = 1, \ldots, n.$$

(*ii*) *Princip jednakih apsolutnih grešaka*

$$\Delta(\overline{a}_1) = \cdots = \Delta(\overline{a}_n).$$

Iz (10) je

$$\Delta(\overline{y}) = \Delta(\overline{a}_k) \sum_{j=1}^{n} |\tfrac{\partial y}{\partial a_j}(\overline{a}_1, \ldots, \overline{a}_n)|,$$

odakle sledi da je

$$\Delta(\overline{a}_k) = \frac{\Delta(\overline{y})}{\sum_{j=1}^{n} |\frac{\partial y}{\partial a_j}(\overline{a}_1, \ldots, \overline{a}_n)|}, \qquad k = 1, \ldots, n.$$

(*iii*) *Princip jednakih relativnih grešaka*

$$\delta(\overline{a}_1) = \cdots = \delta(\overline{a}_n).$$

Sada, (10) može da se napiše u obliku

$$\Delta(\overline{y}) = \frac{\Delta(\overline{a}_k)}{|\overline{a}_k|} \sum_{j=1}^{n} |\overline{a}_j \frac{\partial y}{\partial a_j}(\overline{a}_1, \ldots, \overline{a}_n)|,$$

pa je

$$\Delta(\overline{a}_k) = \frac{\Delta(\overline{y})|\overline{a}_k|}{\sum_{j=1}^{n} |\overline{a}_j \frac{\partial y}{\partial a_j}(\overline{a}_1, \ldots, \overline{a}_n)|}, \qquad k = 1, \ldots, n.$$

2

Interpolacija

Zadati funkciju $y = f(x)$ znači svakoj dopustivoj vrednosti argumenta x pridružiti odgovarajuću vrednost funkcije y. No vrlo često je određivanje vrednosti y skopčano sa mnogim poteškoćama – na primer, y se određuje kao rešenje komplikovanog zadatka ili skupim eksperimentom. Stoga je moguće dobiti samo neveliku tablicu vrednosti funkcije. Osim toga, funkcija može učestvovati u nekim složenijim matematičkim ili tehničkim izračunavanjima koja je nemoguće egzaktno realizovati zbog složenosti reprezentacije te funkcije.

Stoga je pogodno, ili čak neophodno, zameniti funkciju $f(x)$ približnom formulom, tj. funkcijom $g(x)$ koja je bliska u nekom smislu funkciji $f(x)$, a čije vrednosti se mogu jednostavno izračunati; kaže se da funkcija $g(x)$ aproksimira funkciju $f(x)$, $f(x) \approx g(x)$. Kako će biti definisana bliskost ovih funkcija zavisiće od metrike uvedene u prostoru kome pripadaju funkcije, te stoga imamo različite tipove zadataka teorije aproksimacija. Optimalna bliskost funkcija $f(x)$ i $g(x)$ se postiže odgovarajućim izborom slobodnih parametara $(c_0, \ldots, c_n)$ funkcije $g(x)$. Ako je $g(x)$ linearna funkcija parametara c_k, $k = 0, \ldots, n$, aproksimacija je linearna, u protivnom ona je nelinearna. Pri linearnoj aproksimaciji funkcija $g(x)$ se traži u obliku generalisanog polinoma,

$$g(x) = c_0\phi_0(x) + \cdots + c_n\phi_n(x), \tag{1}$$

gde su $\phi_0(x), \ldots, \phi_n(x)$ linearno nezavisne funkcije koje čine tzv. osnovni sistem funkcija. Na primer, ako osnovni sistem čine celi nenegativni stepeni promenljive x, $\phi_0(x) = 1, \ldots, \phi_n(x) = x^n$, $g(x) = c_0 + \cdots + c_n x^n$ je algebarski polinom stepena n; ako je $\phi_0(x) = 1, \phi_1(x) = \cos x, \phi_2(x) = \sin x, \ldots, \phi_{2n-1}(x) = \cos nx$, $\phi_{2n}(x) = \sin nx$, $g(x) = a_0 + a_1 \cos x + b_1 \sin x + \cdots + a_n \cos nx + b_n \sin nx$ je trigonometrijski polinom reda n.

Kada se parametri aproksimacije $c_0, \ldots, c_n$ određuju tako da su vrednosti funkcija $f(x)$ i $g(x)$ jednake na diskretnom skupu tačaka $x_0, \ldots, x_n$,

$$f(x_k) = g(x_k), \qquad k = 0, \ldots, n,$$

onda se taj oblik aproksimacije naziva *interpolacija*. Tačke $x_0, \ldots, x_n$ se nazivaju *čvorovi interpolacije*. Ako se funkcija $g(x)$ traži u obliku generalisanog polinoma (1),

parametri interpolacije se direktno mogu odrediti tzv. metodom neodređenih koeficijenata, tj. rešavanjem sistema linearnih jednačina

$$\sum_{i=0}^{n} c_i \phi_i(x_k) = f(x_k), \qquad k = 0, \ldots, n.$$

2.1 Interpolacioni polinom Lagrangea

Kada je u reprezentaciji (1) $\phi_k(x) \equiv x^k$, $k = 0, \ldots, n$, interpolaciona funkcija $g(x)$ se naziva interpolacioni polinom,

$$L_n(x) = \sum_{i=0}^{n} c_i x^i. \tag{2}$$

TEOREMA 1. *Postoji jedinstveno određen polinom $L_n(x)$ stepena n koji u $(n+1)$-oj različitoj tački x_k, $k = 0, \ldots, n$, zadovoljava uslove*

$$L_n(x_k) = f(x_k), \qquad k = 0, \ldots, n. \tag{3}$$

DOKAZ: Dokažimo prvo da, ukoliko taj polinom postoji, on je jedinstveno određen. Pretpostavimo suprotno, tj. da postoje dva različita polinoma $L_n^1(x)$ i $L_n^2(x)$ takva da je

$$L_n^1(x_k) = L_n^2(x_k) = f(x_k), \qquad k = 0, \ldots, n.$$

Polinom $L_n^1(x) - L_n^2(x)$ je polinom najviše stepena n i ima bar $(n+1)$-nu različitu nulu $x_k, k = 0, \ldots, n$, što je nemoguće. Dakle, polinom (2) je uslovima (3) jedinstveno određen.

Njegovu egzistenciju ćemo dokazati konstruišući ga. Odredimo prvo polinome $l_i(x)$, $i = 0, \ldots, n$, stepena n, takve da je

$$l_i(x_j) = \delta_{ij}, \qquad i, j = 0, \ldots, n, \tag{4}$$

gde je δ_{ij} Kronekerov delta simbol,

$$\delta_{ij} = \begin{cases} 1, & \text{za } i = j \\ 0, & \text{za } i \neq j. \end{cases}$$

Na osnovu prvog dela dokaza teoreme, oni su jedinstveno određeni. Kako su tačke x_k, $k = 0, \ldots, i-1, i+1, \ldots, n$, nule polinoma $l_i(x)$, ovaj se može napisati u obliku

$$l_i(x) = a_i \prod_{\substack{j=0 \\ j \neq i}}^{n} (x - x_j),$$

pri čemu konstantu a_i određujemo iz uslova da je $l_i(x_i) = 1$. Tako dobijamo da je

$$a_i = \Big(\prod_{\substack{j=0\\ j\neq i}}^{n}(x_i - x_j)\Big)^{-1},$$

te je

$$l_i(x) = \prod_{\substack{j=0\\ j\neq i}}^{n}\frac{x - x_j}{x_i - x_j}. \tag{5}$$

S obzirom na (4), polinom

$$L_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i) \tag{6}$$

je polinom stepena n koji zadovoljava uslove (3). ∎

Uvrstimo (5) u (6) i dobijamo izraz za *interpolacioni polinom Lagrangea*

$$L_n(x) = \sum_{i=0}^{n}\Big(\prod_{\substack{j=0\\ j\neq i}}^{n}\frac{x - x_j}{x_i - x_j}\Big) f(x_i). \tag{7}$$

Ako uvedemo oznaku

$$\omega_{n+1}(x) = \prod_{j=0}^{n}(x - x_j),$$

onda je

$$\omega'_{n+1}(x) = \sum_{k=0}^{n}\big(\prod_{\substack{j=0\\ j\neq k}}^{n}(x - x_j)\big),$$

pa je

$$\omega'_{n+1}(x_i) = \prod_{\substack{j=0\\ j\neq i}}^{n}(x_i - x_j).$$

Stoga je

$$\prod_{\substack{j=0\\ j\neq i}}^{n}\frac{x - x_j}{x_i - x_j} = \frac{\prod_{j=0}^{n}(x - x_j)}{(x - x_i)\prod_{\substack{j=0\\ j\neq i}}^{n}(x_i - x_j)} = \frac{\omega_{n+1}(x)}{(x - x_i)\omega'_{n+1}(x_i)},$$

te se Lagrangeov polinom može zapisati i u sledećem obliku

$$L_n(x) = \sum_{i=0}^{n}\frac{\omega_{n+1}(x) f(x_i)}{(x - x_i)\omega'_{n+1}(x_i)}. \tag{8}$$

Greška polinomijalne interpolacije. Greška polinomijalne interpolacije u tački x je razlika vrednosti funkcije i interpolacionog polinoma u toj tački, $f(x) - L_n(x)$.

TEOREMA 2. *Ako je funkcija $f(x)$ diferencijabilna $(n+1)$ puta, tada za svaki argument $\overline{x}$ postoji tačka ξ, koja pripada minimalnom intervalu koji sadrži sve tačke $x_0, \ldots, x_n, \overline{x}$, takva da je*

$$f(\overline{x}) - L_n(\overline{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(\overline{x}), \tag{9}$$

gde je

$$\omega_{n+1}(x) = \prod_{j=0}^{n}(x - x_j).$$

DOKAZ: Pretpostavimo da je $\overline{x} \neq x_j$, $j = 0, \ldots, n$, jer je u protivnom tvrđenje očigledno. Konstruišimo funkciju $F(x)$ takvu da je

$$F(x) \equiv f(x) - L_n(x) - K\omega_{n+1}(x),$$

pri čemu je konstanta K određena tako da je $\overline{x}$ nula funkcije $F(x)$,

$$K = \frac{f(\overline{x}) - L_n(\overline{x})}{\omega_{n+1}(\overline{x})}. \tag{10}$$

Nule funkcije $F(x)$ su i tačke x_k, $k = 0, \ldots, n$, jer je $f(x_k) = L_n(x_k)$ i $\omega_{n+1}(x_k) = 0$ za svako k. Stoga funkcija $F(x)$ ima bar $(n+2)$ nule, $x_0, \ldots, x_n, \overline{x}$. Uzastopnom primenom Rolleove teoreme zaključujemo da $F'(x)$ ima bar $(n+1)$-u nulu, $F''(x)$ bar n nula,..., i da $F^{(n+1)}(x)$ ima bar jednu nulu ξ na intervalu određenom tačkama $x_0, \ldots, x_n, \overline{x}$. Kako je $L_n^{(n+1)}(x) \equiv 0$ i $\omega_{n+1}^{(n+1)}(x) \equiv (n+1)!$ za svako x, to je

$$F^{(n+1)}(\xi) = f^{(n+1)}(\xi) - K(n+1)! = 0,$$

tj.

$$K = \frac{f^{(n+1)}(\xi)}{(n+1)!}. \tag{11}$$

Tvrđenje teoreme sledi iz (10) i (11). ∎

PRIMER 1. Interpolacionim polinomom drugog stepena određenim tačkama $(100, 10)$, $(121, 11)$ i $(144, 12)$ izračunajmo približnu vrednost funkcije $f(x) = \sqrt{x}$ za $x = 115$ i ocenimo tačnost dobijenog rešenja.

Lagrangeov interpolacioni polinom (7) je u ovom slučaju

$$\begin{aligned} L_2(x) &= \tfrac{(x-121)(x-144)}{(100-121)(100-144)}10 + \tfrac{(x-100)(x-144)}{(121-100)(121-144)}11 + \tfrac{(x-100)(x-121)}{(144-100)(144-121)}12 \\ &= -0.00009411x^2 + 0.06842x + 4.099, \end{aligned}$$

te je $L_2(115) = 10.723$. S obzirom da je

$$\max_{(100,144)} |f'''(x)| = 0.375 \cdot 10^{-5},$$

to je apsolutna greška dobijenog rezultata, na osnovu (9),

$$|f(115) - L_2(115)| \leq \tfrac{1}{6} \max_{(100,144)} |f'''(x)||\omega_3(115)| = 0.16 \cdot 10^{-2}.$$

Stoga je, zapisano samo sigurnim ciframa, $L_2(115) = 10.72$.

Formula (7) nije pogodna za izračunavanje vrednosti polinoma u datoj tački, posebno za veće vrednosti n. Umesto da se direktno računa vrednost interpolacionog polinoma određenog svim zadatim čvorovima, može se poći od rešavanja problema interpolacije na manjem skupu čvorova i zatim ovaj proširivati, dok se ne uzmu u obzir svi zadati čvorovi ili ne postigne zahtevana tačnost. Algoritam koji sledi je jedan od efikasnih algoritama ovoga tipa.

Algoritam Nevillea. Označimo sa $L_{i,i+1,\ldots i+k}(x)$ interpolacioni polinom stepena k određen čvorovima x_{i+j}, $0 \leq i + j \leq n$, $j = 0, \ldots, k$, tj. takav da je

$$L_{i,i+1,\ldots i+k}(x_{i+j}) = f_{i+j}, \qquad j = 0, \ldots, k, \tag{12}$$

gde je $f_{i+j} = f(x_{i+j})$. Među ovako definisanim polinomima važi rekurentna veza

$$L_i(x) \equiv f_i \tag{13}$$

$$L_{i,\ldots i+k}(x) = \frac{1}{x_{i+k} - x_i}\big((x - x_i)L_{i+1,\ldots i+k}(x) - (x - x_{i+k})L_{i,\ldots i+k-1}(x)\big). \tag{14}$$

Zaista, identitet (13) je očigledan jer je, prema (12), $L_i(x)$ polinom nultog stepena koji u tački x_i treba da ima vrednost f_i. Dalje, dokažimo da je desna strana relacije (14), koju možemo radi kratkoće označiti sa $P(x)$, identična polinomu $L_{i,\ldots i+k}(x)$. $P(x)$ je polinom stepena ne većeg od k, jer su $L_{i+1,\ldots i+k}(x)$ i $L_{i,\ldots i+k-1}(x)$ polinomi stepena ne većeg od $k - 1$. Još je

$$\begin{aligned}
P(x_i) &= L_{i,\ldots i+k-1}(x_i) = f_i, \\
P(x_{i+k}) &= L_{i+1,\ldots i+k}(x_{i+k}) = f_{i+k}, \\
P(x_{i+j}) &= \frac{1}{x_{i+k} - x_i}\big((x_{i+j} - x_i)f_{i+j} - (x_{i+j} - x_{i+k})f_{i+j}\big) = f_{i+j}, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = 1, \ldots, k-1.
\end{aligned}$$

Dakle, polinom $P(x)$, koji je najviše stepena k, zadovoljava $(k + 1)$ uslova interpolacije u čvorovima $x_i, \ldots, x_{i+k}$. Na osnovu teoreme 1 ovakav polinom je jedinstveno određen, te je

$$P(x) \equiv L_{i,\ldots i+k}(x).$$

Koristeći algoritam (13),(14) možemo formirati tablicu vrednosti polinoma $L_{i,\dots i+k}(x)$ za dato $\overline{x}$:

$$\begin{array}{lllll} x_0 & f_0 = T_{00} & & & \\ & & T_{11} & & \\ x_1 & f_1 = T_{10} & & T_{22} & \\ & & T_{21} & & T_{33} \\ x_2 & f_2 = T_{20} & & T_{32} & \\ & & T_{31} & & \\ x_3 & f_3 = T_{30} & & & \\ \vdots & \vdots & & & \end{array}$$

Radi kraćeg označavanja stavili smo da je

$$T_{i+k,k} \equiv L_{i,i+1,\dots i+k}(\overline{x}),$$

pri čemu prvi indeks ukazuje na poslednji čvor, a drugi na stepen polinoma. Koristeći nove oznake, formule (13) i (14) se mogu zapisati u obliku

$$\begin{aligned} T_{i0} &= f_i \\ T_{ik} &= \frac{1}{x_i - x_{i-k}}\big((\overline{x} - x_{i-k})T_{i,k-1} - (\overline{x} - x_i)T_{i-1,k-1}\big) \\ &= T_{i,k-1} + \big(T_{i,k-1} - T_{i-1,k-1}\big)\frac{\overline{x} - x_i}{x_i - x_{i-k}} \qquad\qquad i = 0, \dots, n. \\ &= T_{i-1,k-1} + \big(T_{i,k-1} - T_{i-1,k-1}\big)\frac{\overline{x} - x_{i-k}}{x_i - x_{i-k}}, \end{aligned} \tag{15}$$

Ako uvedemo oznake

$$\begin{aligned} C_{i0} = D_{i0} = f_i, \\ C_{ik} = T_{ik} - T_{i,k-1}, \qquad D_{ik} = T_{ik} - T_{i-1,k-1}, \qquad 1 \le k \le i, \end{aligned} \tag{16}$$

formule (15) mogu da se napišu u obliku

$$\begin{aligned} C_{i0} = D_{i0} = f_i, \\ C_{ik} = \tfrac{\overline{x} - x_i}{x_i - x_{i-k}}\big(D_{i,k-1} - C_{i-1,k-1}\big), \\ D_{ik} = \tfrac{\overline{x} - x_{i-k}}{x_i - x_{i-k}}\big(D_{i,k-1} - C_{i-1,k-1}\big), \end{aligned} \qquad 1 \le k \le i,\ i = 0, \dots n. \tag{17}$$

Stoga, s obzirom da je iz (16)

$$T_{ik} = C_{ik} + T_{i,k-1} = C_{ik} + C_{i,k-1} + T_{i,k-2} = \dots = \sum_{j=0}^{k} C_{ij},$$

vrednost $L_n(\overline{x}) \equiv T_{nn}$ se računa izrazom

$$T_{nn} = \sum_{j=0}^{n} C_{nj},$$

pri čemu se C_{nj}, $j = 0, \ldots, n$ računaju pomoću rekurentne formule (17).

Nevilleovim algoritmom se efikasno računa vrednost interpolacionog polinoma u nekoj tački. Ako nam je potreban sam polinom, pogodnije je, posebno za veće n, koristiti Newtonovu interpolacionu formulu.

2.2 Interpolacioni polinom Newtona sa podeljenim razlikama

Lagrangeov interpolacioni polinom (7) se može zapisati i u drugom obliku, koji ukazuje da se ovaj polinom može smatrati uopštenjem parcijalne sume Taylorovog reda. Pri tome se kao generalizacija pojma izvoda definišu *podeljene razlike*. Podeljena razlika nultog reda je jednaka vrednosti funkcije u čvoru, a zatim se rekurentno definišu razlike prvog reda pomoću razlika nultog reda, razlike drugog reda pomoću razlika prvog reda, ..., razlike k-tog reda pomoću razlika $(k-1)$-og reda na sledeći način:

$$
\begin{aligned}
f[x_{i_0}] &= f(x_{i_0}) \\
f[x_{i_0}, x_{i_1}] &= \frac{f[x_{i_1}] - f[x_{i_0}]}{x_{i_1} - x_{i_0}} \\
f[x_{i_0}, x_{i_1}, x_{i_2}] &= \frac{f[x_{i_1}, x_{i_2}] - f[x_{i_0}, x_{i_1}]}{x_{i_2} - x_{i_0}} \\
&\vdots \\
f[x_{i_0}, \ldots, x_{i_k}] &= \frac{f[x_{i_1}, \ldots, x_{i_k}] - f[x_{i_0}, \ldots, x_{i_{k-1}}]}{x_{i_k} - x_{i_0}}
\end{aligned} \tag{18}
$$

Da bismo pojednostavili indeksiranje, čvorove koji određuju podeljenu razliku reda k označimo sa $x_0, \ldots, x_k$, pri čemu ovaj skup tačaka ne mora biti uređen.

LEMA 1. *Podeljena razlika reda k se, pomoću vrednosti funkcije u čvorovima kojima je određena, izražava formulom*

$$
f[x_0, \ldots, x_k] = \sum_{i=0}^{k} \frac{f(x_i)}{\prod_{\substack{j=0 \\ j \neq i}}^{k} (x_i - x_j)}. \tag{19}
$$

DOKAZ: Izraz (19) ćemo dokazati matematičkom indukcijom. Za $k = 1$ je

$$
f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0} = \sum_{i=0}^{1} \frac{f(x_i)}{\prod_{\substack{j=0 \\ j \neq i}}^{1} (x_i - x_j)}.
$$

Pretpostavimo da (19) važi za $k \leq n-1$ i dokažimo da važi za $k = n$:

$$
\begin{aligned}
f[x_0,\dots,x_n] &= \frac{f[x_1,\dots,x_n]-f[x_0,\dots,x_{n-1}]}{x_n-x_0}\\
&= \frac{1}{x_n-x_0}\left(\sum_{i=1}^{n}\frac{f(x_i)}{\prod_{\substack{j=1\\ j\neq i}}^{n}(x_i-x_j)}-\sum_{i=0}^{n-1}\frac{f(x_i)}{\prod_{\substack{j=0\\ j\neq i}}^{n-1}(x_i-x_j)}\right)\\
&= \frac{1}{x_n-x_0}\frac{f(x_n)}{\prod_{j=1}^{n-1}(x_n-x_j)}+\frac{1}{x_n-x_0}\sum_{i=1}^{n-1}\frac{f(x_i)}{\prod_{\substack{j=1\\ j\neq i}}^{n}(x_i-x_j)}\\
&\quad-\frac{1}{x_n-x_0}\frac{f(x_0)}{\prod_{j=1}^{n-1}(x_0-x_j)}-\frac{1}{x_n-x_0}\sum_{i=1}^{n-1}\frac{f(x_i)}{\prod_{\substack{j=0\\ j\neq i}}^{n-1}(x_i-x_j)}\\
&= \frac{f(x_0)}{\prod_{j=1}^{n}(x_0-x_j)}+\frac{1}{x_n-x_0}\sum_{i=1}^{n-1}\frac{(x_i-x_0)-(x_i-x_n)}{\prod_{\substack{j=0\\ j\neq i}}^{n}(x_i-x_j)}f(x_i)\\
&\qquad+\frac{f(x_n)}{\prod_{j=0}^{n-1}(x_n-x_j)}=\sum_{i=0}^{n}\frac{f(x_i)}{\prod_{\substack{j=0\\ j\neq i}}^{n}(x_i-x_j)}.
\end{aligned}
$$

∎

Iz leme 1 neposredno slede osobine podeljenih razlika:

(*i*) podeljena razlika je linearni operator

$$(\alpha_1 f_1+\alpha_2 f_2)[x_0,\dots,x_n]=\alpha_1 f_1[x_0,\dots,x_n]+\alpha_2 f_2[x_0,\dots,x_n];$$

(*ii*) podeljena razlika je simetrična funkcija svojih argumenata, što znači da redosled čvorova nije bitan.

Radi preglednijeg zapisa i lakšeg korišćenja, obično se podeljene razlike pišu u tabeli

$$
\begin{array}{cccccc}
x_0 & f(x_0) & & & & \\
 & & f[x_0,x_1] & & & \\
x_1 & f(x_1) & & f[x_0,x_1,x_2] & & \\
 & & f[x_1,x_2] & & & \\
x_2 & f(x_2) & & & \dots & f[x_0,\dots,x_k]\\
\vdots & \vdots & \vdots & & & \\
 & & f[x_{k-1},x_k] & & & \\
x_k & f(x_k) & & & &
\end{array}
$$

Interpolacioni polinom (7) se može zapisati i na drugi način, pomoću podeljenih razlika. Da bismo do toga došli, izrazimo najpre grešku interpolacije odgovarajućom podeljenom razlikom. Koristeći izraz (8), imamo da je

$$f(x) - L_k(x) = f(x) - \omega_{k+1}(x) \sum_{i=0}^{k} \frac{f(x_i)}{(x - x_i) \prod_{\substack{j=0 \\ j \neq i}}^{k} (x_i - x_j)}$$
$$= \omega_{k+1}(x) \Bigg(\frac{f(x)}{\prod_{j=0}^{k} (x - x_j)} + \sum_{i=0}^{k} \frac{f(x_i)}{(x_i - x) \prod_{\substack{j=0 \\ j \neq i}}^{k} (x_i - x_j)} \Bigg),$$

te je na osnovu leme 1 greška interpolacije polinomom $L_k(x)$

$$f(x) - L_k(x) = \omega_{k+1}(x) f[x, x_0, \ldots, x_k]. \tag{20}$$

Polinom $L_n(x)$ može da se napiše u obliku

$$L_n(x) = L_0(x) + \big(L_1(x) - L_0(x)\big) + \cdots + \big(L_n(x) - L_{n-1}(x)\big), \tag{21}$$

gde je $L_m(x)$ interpolacioni polinom određen čvorovima $x_0, \ldots, x_m$. Razlika $L_m(x) - L_{m-1}(x)$ je polinom stepena m, koji je nula u tačkama $x_0, \ldots, x_{m-1}$ jer je $L_{m-1}(x_j) = L_m(x_j) = f(x_j)$, $j = 0, \ldots, m-1$. Stoga je

$$L_m(x) - L_{m-1}(x) \equiv a_m \omega_m(x), \tag{22}$$

gde je a_m konstanta koju treba odrediti, a $\omega_m(x) = \prod_{j=0}^{m-1} (x - x_j)$. Ako u izraz (22) stavimo $x = x_m$ i uzmemo u obzir da je $L_m(x_m) = f(x_m)$, dobijamo da je

$$f(x_m) - L_{m-1}(x_m) = a_m \omega_m(x_m),$$

što, poređenjem sa (20) za $x = x_m$ i $k = m - 1$, daje

$$a_m = f[x_0, \ldots, x_m].$$

Konačno, zamenom dobijenog izraza za a_m u (22) imamo da je

$$L_m(x) - L_{m-1}(x) \equiv f[x_0, \ldots, x_m]\, \omega_m(x),$$

pa je na osnovu (21)

$$\begin{aligned} L_n(x) = {} & f(x_0) + f[x_0, x_1](x - x_0) + \cdots \\ & + f[x_0, \ldots, x_n](x - x_0) \cdots (x - x_{n-1}). \end{aligned} \tag{23}$$

Interpolacioni polinom zapisan u obliku (23) naziva se *Newtonov interpolacioni polinom sa podeljenim razlikama*, i može se smatrati uopštenjem parcijalne sume Taylorovog reda funkcije $f(x)$.

PRIMER 2. Tabela podeljenih razlika za funkciju $f(x) = \sqrt{x}$ zadatu u primeru 1 je

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$
100	10		
		0.04762	
121	11		-9.411·10^{-5}
		0.04348	
144	12		

Zamenom u formuli (23) za $n = 2$ i $x_0 = 100$, dobijamo da je Newtonov interpolacioni polinom sa podeljenim razlikama za funkciju $f(x) = \sqrt{x}$

$$L_2(x) = 10 + 0.04762(x - 100) - 0.00009411(x - 100)(x - 121)$$

$$= -0.00009411x^2 + 0.06842x + 4.099,$$

što je identično polinomu dobijenom u primeru 1.

Poredeći izraze (9) i (20) zaključujemo da je

$$f[x_0, \ldots, x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}, \tag{24}$$

odakle sledi još jedna osobina podeljenih razlika. Naime, ako je funkcija $f(x)$ polinom stepena k,

$$P_k(x) \equiv \sum_{i=0}^{k} b_i x^i,$$

tada je na osnovu (24)

$$P_k[x_0, \ldots, x_n] = \begin{cases} b_n, & \text{za } k = n \\ 0, & \text{za } k < n, \end{cases}$$

za proizvoljno $x_0, \ldots, x_n$.

2.3 Interpolacioni polinomi sa ravnomerno raspoređenim čvorovima

Kada su čvorovi x_i ravnomerno raspoređeni sa korakom h, $x_i = x_0 + i\,h$, umesto podeljenih koriste se *konačne razlike*. Razlika $f_{i+1} - f_i$, gde je $f_i = f(x_i)$, naziva se konačnom razlikom prvog reda. U zavisnosti od potrebe, označavamo je sa

$$\begin{aligned} f_{i+1} - f_i &= \Delta f_i && \textit{razlika unapred},\\ &= \nabla f_{i+1} && \textit{razlika unazad},\\ &= \delta f_{i+\frac{1}{2}} && \textit{centralna razlika}. \end{aligned} \tag{25}$$

Razlike višeg reda se definišu rekurentnim relacijama

$$\begin{aligned} \Delta^k f_i &= \Delta(\Delta^{k-1} f_i) = \Delta^{k-1} f_{i+1} - \Delta^{k-1} f_i,\\ \nabla^k f_i &= \nabla(\nabla^{k-1} f_i) = \nabla^{k-1} f_i - \nabla^{k-1} f_{i-1},\\ \delta^k f_i &= \delta(\delta^{k-1} f_i) = \delta^{k-1} f_{i+\frac{1}{2}} - \delta^{k-1} f_{i-\frac{1}{2}}. \end{aligned} \tag{26}$$

Veza između konačnih razlika različitog tipa određenih istim skupom čvorova je

$$\Delta^k f_i = \nabla^k f_{i+k} = \delta^k f_{i+\frac{k}{2}}. \tag{27}$$

I konačne razlike se, radi lakšeg korišćenja, zapisuju u tabeli

$$\begin{array}{llllll} x_0 & f_0 & & & & \\ & & \Delta f_0 & & & \\ x_1 & f_1 & & \Delta^2 f_0 & & \\ & & \Delta f_1 & & \Delta^3 f_0 & \\ x_2 & f_2 & & \Delta^2 f_1 & & \dots \\ & & \Delta f_2 & & \Delta^3 f_1 & \\ x_3 & f_3 & & \Delta^2 f_2 & & \\ & & \Delta f_3 & & & \\ x_4 & f_4 & & & & \\ \vdots & \vdots & & & & \end{array}$$

LEMA 2. *Konačne razlike reda k izražavaju se pomoću vrednosti funkcije u čvorovima formulom*

$$\Delta^k f_i = \sum_{j=0}^{k} (-1)^j C_k^j f_{i+k-j}, \qquad C_k^j = \binom{k}{j}. \tag{28}$$

DOKAZ: Izraz (28) ćemo dokazati matematičkom indukcijom. Za $k = 1$ on se svodi na $\Delta f_i = f_{i+1} - f_i$, što predstavlja definiciju (25). Pretpostavimo da (28) važi za

$k \le n$ i dokažimo da važi za $k = n+1$. Na osnovu (26) i indukcijske hipoteze je

$$\begin{aligned}\Delta^{n+1} f_i &= \Delta^n f_{i+1} - \Delta^n f_i \\ &= \sum_{j=0}^{n} (-1)^j C_n^j f_{i+1+n-j} - \sum_{j=0}^{n} (-1)^j C_n^j f_{i+n-j} \\ &= C_n^0 f_{i+n+1} + \sum_{j=0}^{n-1} (-1)^{j+1} C_n^{j+1} f_{i+n-j} - \sum_{j=0}^{n-1} (-1)^j C_n^j f_{i+n-j} - (-1)^n C_n^n f_i.\end{aligned}$$

Kako je

$$(-1)^{j+1} C_n^{j+1} - (-1)^j C_n^j = (-1)^{j+1} C_{n+1}^{j+1}, \quad C_n^0 = 1 = C_{n+1}^0, \quad C_n^n = 1 = C_{n+1}^{n+1},$$

to je dalje

$$\begin{aligned}\Delta^{n+1} f_i &= C_{n+1}^0 f_{i+n+1} + \sum_{j=0}^{n-1} (-1)^{j+1} C_{n+1}^{j+1} f_{i+n-j} - (-1)^n C_{n+1}^{n+1} f_i \\ &= C_{n+1}^0 f_{i+n+1} + \sum_{j=1}^{n} (-1)^j C_{n+1}^j f_{i+n-(j-1)} + (-1)^{n+1} C_{n+1}^{n+1} f_i \\ &= \sum_{j=0}^{n+1} (-1)^j C_{n+1}^j f_{i+n+1-j},\end{aligned}$$

što je i trebalo dokazati. ∎

Veza podeljenih i konačnih razlika data je sledećom lemom:

LEMA 3. *Ako je $x_i = x_0 + i\,h$, onda je*

$$f[x_i, \ldots, x_{i+k}] = \frac{\Delta^k f_i}{h^k\, k!}. \tag{29}$$

DOKAZ: I ovo tvrđenje dokažimo indukcijom. Za $k = 1$, na osnovu definicija podeljene i konačne razlike prvog reda sledi tvrđenje, jer je

$$f[x_i, x_{i+1}] = \frac{f_{i+1} - f_i}{x_{i+1} - x_i} = \frac{\Delta f_i}{h}.$$

Dalje, neka (29) važi za $k \le n$. Tada je

$$\begin{aligned}f[x_i, \ldots, x_{i+n+1}] &= \frac{f[x_{i+1}, \ldots, x_{i+n+1}] - f[x_i, \ldots, x_{i+n}]}{x_{i+n+1} - x_i} \\ &= \frac{1}{(n+1)h} \left(\frac{\Delta^n f_{i+1}}{h^n\, n!} - \frac{\Delta^n f_i}{h^n\, n!} \right) = \frac{\Delta^{n+1} f_i}{h^{n+1}\,(n+1)!}.\end{aligned}$$

∎

Iz (29), (24) i (27) dobijamo vezu između konačnih razlika i izvoda funkcije:

$$\Delta^k f_i = \nabla^k f_{i+k} = \delta^k f_{i+\frac{k}{2}} = h^k f^{(k)}(\xi), \qquad x_i \le \xi \le x_{i+k}. \tag{30}$$

Posledica ove veze je da su konačne razlike reda n polinoma stepena n konstantne i jednake $h^n b_n n!$, gde je b_n koeficijent polinoma uz najviši stepen.

Uzimajući u obzir ravnomernu raspoređenost čvorova interpolacije, polinom (23) možemo zapisati na različite načine, što zavisi od položaja u odnosu na čvorove interpolacije tačke x u kojoj računamo vrednost polinoma. Tako dobijamo različite interpolacione formule sa ravnomerno raspoređenim čvorovima.

Neka je sa x_0 uvek označen čvor koji je najbliži tački x, a ostali čvorovi x_i imaju pozitivan ili negativan indeks u zavisnosti od njihovog položaja u odnosu na taj osnovni čvor, tj. $x_i = x_0 + ih$, $i = 0, \pm 1, \ldots$. Definišimo novu promenljivu q,

$$x = x_0 + q\,h, \tag{31}$$

koja pripada intervalu $(-1, 1)$, s obzirom na način izbora čvora x_0.

Ako se x nalazi na početku tabele, odnosno svi čvorovi imaju pozitivan indeks i, onda se polinom (23) može zapisati u obliku tzv. *Newtonovog interpolacionog polinoma za interpolaciju unapred*

$$L_n(x_0+qh) = f_0 + q\Delta f_0 + \frac{q(q-1)}{2!}\Delta^2 f_0 + \cdots + \frac{q(q-1)\cdots(q-n+1)}{n!}\Delta^n f_0, \tag{32}$$

jer je

$$\omega_k(x_0 + qh) = h^k \prod_{j=0}^{k-1}(q-j), \tag{33}$$

a podeljene razlike se izražavaju pomoću konačnih iz veze (29). Greška interpolacije (9), s obzirom na (33), je

$$f(x_0 + qh) - L_n(x_0 + qh) = \frac{q(q-1)\cdots(q-n)}{(n+1)!} h^{n+1} f^{(n+1)}(\xi), \qquad x_0 \le \xi \le x_n,$$

i može se, imajući u vidu (30), oceniti izrazom $\frac{q(q-1)\cdots(q-n)}{(n+1)!}\Delta^{n+1} f$, u kome se obično koristi srednja vrednost konačnih razlika reda $n+1$.

Ako se x nalazi pri kraju tabele, indeksi čvorova kojima je određen polinom su negativni, i čvorovi se, prema njihovom rastojanju od x, uključuju u polinom sledećim redosledom: $x_0, x_{-1}, \ldots, x_{-n}$. Koristeći vezu (29) i smenu (31), polinom (23) je sada oblika

$$L_n(x_0 + qh) = f_0 + q\Delta f_{-1} + \frac{q(q+1)}{2!}\Delta^2 f_{-2} + \cdots + \frac{q(q+1)\cdots(q+n-1)}{n!}\Delta^n f_{-n},$$

i naziva se *Newtonov interpolacioni polinom za interpolaciju unazad*. Greška interpolacije je

$$f(x_0+qh)-L_n(x_0+qh)=\frac{q(q+1)\cdots(q+n)}{(n+1)!}h^{n+1}f^{(n+1)}(\xi),\qquad x_{-n}\le\xi\le x_0,$$

i može se, ako je teško oceniti $f^{(n+1)}(x)$, grubo oceniti izrazom $\frac{q(q+1)\cdots(q+n)}{(n+1)!}\Delta^{n+1}f$.

Pomenuta dva Newtonova interpolaciona polinoma se koriste i za *ekstrapolaciju*, tj. nalaženje približne vrednosti funkcije $f(x)$ u tački x koja ne pripada intervalu određenom čvorovima interpolacije. Ako je za svako k $x<x_k$ koristi se Newtonov polinom za interpolaciju unapred, a ako je $x>x_k$ koristi se Newtonov polinom za interpolaciju unazad. Greška ekstrapolacije je veća od greške interpolacije, te ekstrapolaciju treba izbegavati kad god je moguće.

Kada se interpolacija vrši ma kojim od pomenuta dva Newtonova polinoma, koriste se informacije o funkciji koju interpolišemo samo sa jedne strane tačke x. Veća tačnost će se postići ako se za formiranje interpolacionog polinoma koriste čvorovi sa obe strane tačke x, prema njihovoj udaljenosti od te tačke. To je moguće ako tačka x nije blizu početka ili kraja tabele.

Ako $x\in(x_0,x_0+\frac{h}{2}]$, pogodno je čvorove interpolacije izabrati sledećim redom: x_0, $x_1=x_0+h$, $x_{-1}=x_0-h$, $x_2=x_0+2h$, $x_{-2}=x_0-2h$, ... Na skupu od $2n+2$ čvora $x_0,x_1,x_{-1},\dots,x_n,x_{-n},x_{n+1}$ polinom (23) je

$$\begin{aligned}L_{2n+1}(x)&=f(x_0)+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_{-1}](x-x_0)(x-x_1)\\&\quad+f[x_0,x_1,x_{-1},x_2](x-x_0)(x-x_1)(x-x_{-1})+\cdots\\&\quad+f[x_0,x_1,\dots,x_{-n},x_{n+1}](x-x_0)\cdots(x-x_n)(x-x_{-n}).\end{aligned}$$

Uvodeći u poslednji polinom smenu (31) i koristeći vezu (29) dobijamo *Gaussov interpolacioni polinom za interpolaciju unapred*

$$\begin{aligned}L_{2n+1}(x_0+qh)&=f_0+q\Delta f_0+\frac{q(q-1)}{2!}\Delta^2f_{-1}+\frac{q(q^2-1)}{3!}\Delta^3f_{-1}+\\&\quad\cdots+\frac{q(q^2-1)\cdots(q^2-n^2)}{(2n+1)!}\Delta^{2n+1}f_{-n}.\end{aligned}\tag{34}$$

Ako $x\in[x_0+\frac{h}{2},x_1)$ tački x je najbliži čvor x_1, pa napišimo polinom (23) na skupu od $2n+2$ čvora $x_1,x_0,x_2,x_{-1},\dots,x_{n+1},x_{-n}$:

$$\begin{aligned}L_{2n+1}(x)&=f(x_1)+f[x_1,x_0](x-x_1)+f[x_1,x_0,x_2](x-x_1)(x-x_0)\\&\quad+f[x_1,x_0,x_2,x_{-1}](x-x_1)(x-x_0)(x-x_2)+\cdots\\&\quad+f[x_1,x_0,\dots,x_{n+1},x_{-n}](x-x_1)\cdots(x-x_{-(n-1)})(x-x_{n+1}).\end{aligned}$$

Uvodeći smenu (31) i koristeći vezu (29) dobijamo *Gaussov interpolacioni polinom za interpolaciju unazad*

$$\begin{aligned} L_{2n+1}(x_0+qh) &= f_1 + (q-1)\Delta f_0 + \frac{q(q-1)}{2!}\Delta^2 f_0 \\ &+ \frac{q(q-1)(q-2)}{3!}\Delta^3 f_{-1} + \cdots \\ &+ \frac{q(q^2-1)\cdots(q^2-(n-1)^2)(q-n)(q-(n+1))}{(2n+1)!}\Delta^{2n+1} f_{-n}. \end{aligned} \tag{35}$$

Gaussov interpolacioni polinom za interpolaciju unazad može da se dobije i u drugom obliku, ukoliko se čvorovi prenumerišu tako da $x \in [x_0 - \frac{h}{2}, x_0)$. Polinom (23) se tada piše na skupu od $2n+2$ čvora $x_0, x_{-1}, x_1, x_{-2}, x_2, \ldots, x_n, x_{-(n+1)}$:

$$\begin{aligned} L_{2n+1}(x) &= f(x_0) + f[x_0, x_{-1}](x-x_0) + f[x_0, x_{-1}, x_1](x-x_0)(x-x_{-1}) \\ &+ f[x_0, x_{-1}, x_1, x_{-2}](x-x_0)(x-x_{-1})(x-x_1) + \cdots \\ &+ f[x_0, x_{-1}, \ldots, x_n, x_{-(n+1)}](x-x_0)\cdots(x-x_{-n})(x-x_n). \end{aligned}$$

Uvodeći smenu (31) i koristeći vezu (29) dobijamo drugi oblik Gaussovog interpolacionog polinoma za interpolaciju unazad

$$\begin{aligned} L_{2n+1}(x_0+qh) &= f_0 + q\Delta f_{-1} + \frac{q(q+1)}{2!}\Delta^2 f_{-1} + \frac{q(q^2-1)}{3!}\Delta^3 f_{-2} \\ &+ \cdots + \frac{q(q^2-1)\cdots(q^2-n^2)}{(2n+1)!}\Delta^{2n+1} f_{-(n+1)}. \end{aligned} \tag{36}$$

Aritmetička sredina polinoma (34) i (36) je *Stirlingov interpolacioni polinom*

$$\begin{aligned} L_{2n+1}(x_0+qh) &= f_0 + q\frac{\Delta f_{-1} + \Delta f_0}{2} + \frac{q^2}{2!}\Delta^2 f_{-1} + \frac{q(q^2-1)}{3!}\,\frac{\Delta^3 f_{-2} + \Delta^3 f_{-1}}{2} \\ &+ \cdots + \frac{q(q^2-1)\cdots(q^2-n^2)}{(2n+1)!}\,\frac{\Delta^{2n+1} f_{-(n+1)} + \Delta^{2n+1} f_{-n}}{2}, \end{aligned}$$

koji se, s obzirom na način kako je izveden, obično koristi kada je $|q| \leq 0.25$.

Aritmetička sredina polinoma (34) i (35) je *Besselov interpolacioni polinom*

$$\begin{aligned} L_{2n+1}(x_0+qh) &= \frac{f_0+f_1}{2} + (q-\frac{1}{2})\Delta f_0 + \frac{q(q-1)}{2!}\,\frac{\Delta^2 f_{-1} + \Delta^2 f_0}{2} \\ &+ \cdots + \frac{q(q^2-1)\cdots(q^2-(n-1)^2)(q-n)(q-\frac{1}{2})}{(2n+1)!}\Delta^{2n+1} f_{-n}, \end{aligned}$$

i obično se koristi kada je $0.25 \leq q \leq 0.75$.

Greška interpolacije ovim polinomima se može izvesti pomoću opšteg izraza za grešku interpolacije (9).

PRIMER 3. Na osnovu zadatih vrednosti funkcije $f(x) = e^x$ u tačkama 1.10, 1.15 i 1.20 izračunajmo $f(1.14)$ i ocenimo grešku dobijenog rezultata.

Čvorovi su ravnomerno raspoređeni sa korakom $h = 0.05$. Najbliži tački $x = 1.14$ je čvor $x_0 = 1.15$, te je $q = \frac{1.14-1.15}{0.05} = -0.2$. Stoga ćemo $f(1.14)$ izračunati približno pomoću Stirlingove interpolacione formule

x_i	$f(x_i)$	Δf_i	$\Delta^2 f_i$
1.10	3.00417		
		0.15402	
1.15	3.15819		0.00791
		0.16193	
1.20	3.32012		

$$L_2(1.14) = 3.15819 - 0.2\,\frac{0.15402 + 0.16193}{2} + \frac{0.2^2}{2}\,0.00791 = 3.12675.$$

Kako je $\max_{(1.10,1.20)} |f'''(x)| = 3.32012$, to je, prema (9), greška izračunate vrednosti

$$|f(1.14) - L_2(1.14)| \le \frac{1}{6} \max_{(1.10,1.20)} |f'''(x)|\,|h^3 q(q^2-1)| = 0.24 \cdot 10^{-4}.$$

Dakle, $f(1.14) = 3.1268$ sa greškom ne većom od $0.5 \cdot 10^{-4}$.

2.4 Interpolacioni polinom Hermitea

Ako su u nekim od čvorova interpolacije i izvodi interpolacionog polinoma jednaki odgovarajućim izvodima funkcije koju interpolišemo, onda se takav interpolacioni polinom naziva Hermiteov interpolacioni polinom. Drugim rečima, neka su dati realni brojevi x_i i $f_i^k = f^{(k)}(x_i)$, $k = 0, \ldots, n_i - 1$, $i = 0, \ldots, m$, pri čemu su svi x_i različiti među sobom. Polinom $P_n(x)$ stepena n,

$$n = \sum_{i=0}^{m} n_i - 1, \tag{37}$$

koji zadovoljava uslove interpolacije

$$P_n^{(k)}(x_i) = f_i^k, \qquad k = 0, \ldots, n_i - 1, \quad i = 0, \ldots, m, \tag{38}$$

je Hermiteov interpolacioni polinom. Lagrangeov interpolacioni polinom je specijalan slučaj Hermiteovog interpolacionog polinoma za $n_i = 1$, $i = 0, \ldots, m$.

TEOREMA 3. *Za proizvoljan skup realnih brojeva* x_i *i* f_i^k, $k = 0, \ldots, n_i - 1, i = 0, \ldots, m$, *uz uslov da je za svako* $0 \leq i, j \leq m$ $x_i \neq x_j$, *postoji tačno jedan polinom* $P_n(x)$ *stepena* $n = \sum_{i=0}^{m} n_i - 1$ *koji zadovoljava uslove* $P_n^{(k)}(x_i) = f_i^k$, $k = 0, \ldots, n_i - 1$, $i = 0, \ldots, m$.

DOKAZ: Dokažimo prvo da, ukoliko taj polinom postoji, on je jedinstveno određen. Pretpostavimo da postoje dva različita polinoma $P_n^1(x)$ i $P_n^2(x)$ koji zadovoljavaju uslove (38). Tada je njihova razlika $Q(x) \equiv P_n^1(x) - P_n^2(x)$ polinom najviše stepena n za koji važi da je $Q^{(k)}(x_i) = 0$, $k = 0, \ldots, n_i - 1$, $i = 0, \ldots, m$. Dakle, tačka x_i je bar n_i-tostruki koren polinoma $Q(x)$, te prema (37) polinom $Q(x)$ ima bar $n+1$ koren, što je nemoguće.

Egzistencija Hermiteovog polinoma je posledica dokazane jedinstvenosti. Uslovima (38) je određen sistem od $(n+1)$-e linearne jednačine po $(n+1)$-om nepoznatom koeficijentu c_j polinoma $P_n(x) \equiv \sum_{j=0}^{n} c_j x^j$. Matrica sistema nije singularna, jer je dokazano da ovaj ne može imati više od jednog rešenja (u protivnom bi za neki izbor vektora desne strane sistem imao više rešenja). A sistem sa regularnom matricom uvek ima rešenja, dakle za proizvoljan izbor desne strane f_i^k. ∎

Konstruišimo sada taj polinom. Neka je $\epsilon > 0$ mali broj i neka su x_{ik}^{ϵ}, $k = 0, \ldots, n_i - 1$, $i = 0, \ldots, m$ nizovi tačaka takvi da su sve tačke x_{ik}^{ϵ} među sobom različite i da je $\lim_{\epsilon \to 0} x_{ik}^{\epsilon} = x_i$. Jedan od načina izbora ovih tačaka je

$$x_{ik}^{\epsilon} = x_i + k\epsilon.$$

Formirajmo za funkciju $f(x) \in \mathcal{C}^{n+1}[a, b]$ interpolacioni polinom $P_n^{\epsilon}(x)$ stepena n sa čvorovima interpolacije x_{ik}^{ϵ}. Tabela podeljenih razlika je

$$\begin{array}{cccc}
f(x_{00}^{\epsilon}) & & & \\
 & f[x_{00}^{\epsilon}, x_{01}^{\epsilon}] & & \\
f(x_{01}^{\epsilon}) & & f[x_{00}^{\epsilon}, x_{01}^{\epsilon}, x_{02}^{\epsilon}] & \\
 & f[x_{01}^{\epsilon}, x_{02}^{\epsilon}] & & \\
\vdots & & & \\
f(x_{0,n_0-1}^{\epsilon}) & & & \\
 & f[x_{0,n_0-1}^{\epsilon}, x_{10}^{\epsilon}] & \cdots & f[x_{00}^{\epsilon}, x_{01}^{\epsilon}, \ldots, x_{m,n_m-1}^{\epsilon}] \\
f(x_{10}^{\epsilon}) & & & \\
\vdots & & & \\
f(x_{m,n_m-1}^{\epsilon}) & & &
\end{array} \tag{39}$$

te je Newtonov interpolacioni polinom sa podeljenim razlikama

$$\begin{aligned} P_n^{\epsilon}(x) =& a_0^{\epsilon} + a_1^{\epsilon}(x - x_{00}^{\epsilon}) + a_2^{\epsilon}(x - x_{00}^{\epsilon})(x - x_{01}^{\epsilon}) \\ &+ \cdots + a_n^{\epsilon}(x - x_{00}^{\epsilon}) \cdots (x - x_{m,n_m-2}^{\epsilon}), \end{aligned} \tag{40}$$

pri čemu je

$$a_0^\epsilon = f(x_{00}^\epsilon), \quad a_1^\epsilon = f[x_{00}^\epsilon, x_{01}^\epsilon], \quad a_2^\epsilon = f[x_{00}^\epsilon, x_{01}^\epsilon, x_{02}^\epsilon], \quad \ldots, \quad a_n^\epsilon = f[x_{00}^\epsilon, x_{01}^\epsilon, \ldots, x_{m,n_m-1}^\epsilon]. \tag{41}$$

Prema (24), podeljene razlike se mogu izraziti odgovarajućim izvodima

$$f[x_{ik}^\epsilon, \ldots, x_{il}^\epsilon] = \frac{f^{(l-k)}(\xi_{ikl}^\epsilon)}{(l-k)!}, \qquad \xi_{ikl}^\epsilon \in [\min_{k\le j\le l}(x_{ij}^\epsilon), \max_{k\le j\le l}(x_{ij}^\epsilon)]. \tag{42}$$

Kada u (42) pustimo da $\epsilon \to 0$, s obzirom da je

$$\lim_{\epsilon\to 0} f[x_{ik}^\epsilon, \ldots, x_{il}^\epsilon] = f[\underbrace{x_i, \ldots, x_i}_{(l-k+1)\text{ puta}}] \qquad \lim_{\epsilon\to 0} \xi_{ikl}^\epsilon = x_i, \tag{43}$$

dobijamo

$$f[\underbrace{x_i, \ldots, x_i}_{(p+1)\text{ puta}}] = \frac{f^{(p)}(x_i)}{p!}, \tag{44}$$

gde smo radi kratkoće zapisa stavili $p = l - k$.

Indukcijom dokažimo da sve podeljene razlike koje se javljaju u tabeli (39) imaju graničnu vrednost kada $\epsilon \to 0$. Razlike nultog reda su jednake vrednostima funkcije, te je za njih tvrđenje tačno. Pretpostavimo da je tvrđenje tačno za podeljene razlike reda $q-1$. Podeljena razlika reda q se može izraziti pomoću podeljenih razlika reda $q-1$,

$$f[x_{ik}^\epsilon, \ldots, x_{jl}^\epsilon] = \frac{f[x_{i,k+1}^\epsilon, \ldots, x_{jl}^\epsilon] - f[x_{ik}^\epsilon, \ldots, x_{j,l-1}^\epsilon]}{x_{jl}^\epsilon - x_{ik}^\epsilon}, \tag{45}$$

pri čemu ćemo, ne umanjujući opštost dokaza, pretpostaviti da je $k < n_i - 1$ i $l > 0$. Ako je $i = j$, egzistencija granične vrednosti podeljenih razlika reda q sledi iz (43). Ako je $i \neq j$, granična vrednost imenioca desne strane izraza (45) je $x_i - x_j \neq 0$, a granična vrednost brojioca istog izraza postoji na osnovu indukcijske hipoteze. Dakle, granična vrednost izraza (45) uvek postoji, što je i trebalo pokazati.

Stoga su granične vrednosti koeficijenata a_k^ϵ, $k = 0, \ldots, n$, kada u (41) $\epsilon \to 0$, jednake

$$a_0 = f(x_0), \quad a_1 = f[x_0, x_0], \quad \ldots \quad a_{n_0-1} = f[\underbrace{x_0, \ldots, x_0}_{n_0\text{ puta}}],$$

$$a_{n_0} = f[\underbrace{x_0, \ldots, x_0}_{n_0\text{ puta}}, x_1], \quad \ldots \quad a_n = f[\underbrace{x_0, \ldots, x_0}_{n_0\text{ puta}}, \ldots, \underbrace{x_m, \ldots, x_m}_{n_m\text{ puta}}].$$

Uzimajući ovo u obzir, kada u polinomu (40) pustimo da $\epsilon \to 0$ dobijamo Hermiteov interpolacioni polinom stepena n

$$
\begin{aligned}
P_n(x) &= f(x_0) + f[x_0, x_0](x - x_0) + f[x_0, x_0, x_0](x - x_0)^2 + \cdots \\
&+ f[\underbrace{x_0, \ldots, x_0}_{n_0 \text{ puta}}](x - x_0)^{n_0 - 1} + f[\underbrace{x_0, \ldots, x_0}_{n_0 \text{ puta}}, x_1](x - x_0)^{n_0} \\
&+ f[\underbrace{x_0, \ldots, x_0}_{n_0 \text{ puta}}, x_1, x_1](x - x_0)^{n_0}(x - x_1) + \cdots \\
&+ f[\underbrace{x_0, \ldots, x_0}_{n_0 \text{ puta}}, \ldots, \underbrace{x_m, \ldots, x_m}_{n_m \text{ puta}}](x - x_0)^{n_0} \cdots (x - x_m)^{n_m - 1}.
\end{aligned} \tag{46}
$$

Koeficijenti a_k, $k = 0, \ldots, n$, se računaju tako što se prvo podeljene razlike definisane različitim čvorovima izraze pomoću podeljenih razlika definisanih samo sa po jednim čvorom, a zatim iskoristi veza (44) i zadati podaci.

PRIMER 4. Napišimo Hermiteov interpolacioni polinom za funkciju $f(x)$ zadatu tabelom

x_i	$f(x_i)$	$f'(x_i)$	$f''(x_i)$
0	-1	-2	–
1	0	10	40

Čvor $x_0 = 0$ je dvostruki čvor, a $x_1 = 1$ trostruki, te je ovaj polinom oblika

$$
\begin{aligned}
P_4(x) &= f(x_0) + f[x_0, x_0](x - x_0) + f[x_0, x_0, x_1](x - x_0)^2 \\
&+ f[x_0, x_0, x_1, x_1](x - x_0)^2(x - x_1) + f[x_0, x_0, x_1, x_1, x_1](x - x_0)^2(x - x_1)^2.
\end{aligned}
$$

Koristeći definiciju podeljenih razlika (18) i vezu (44), imamo da je

$$
\begin{aligned}
f[x_0, x_0] &= f'(x_0) = -2 \\
f[x_0, x_0, x_1] &= \frac{1}{x_1 - x_0}\left(\frac{f(x_1) - f(x_0)}{x_1 - x_0} - f'(x_0)\right) = 3 \\
f[x_0, x_0, x_1, x_1] &= \frac{1}{x_1 - x_0}\left(\frac{1}{x_1 - x_0}\left(f'(x_1) - \frac{f(x_1) - f(x_0)}{x_1 - x_0}\right) - f[x_0, x_0, x_1]\right) = 6 \\
f[x_0, x_0, x_1, x_1, x_1] &= \frac{1}{x_1 - x_0}\left(\frac{1}{x_1 - x_0}\left(\frac{f''(x_1)}{2} - \frac{f'(x_1)}{x_1 - x_0} + \frac{f(x_1) - f(x_0)}{(x_1 - x_0)^2}\right)\right. \\
&\qquad \left. - f[x_0, x_0, x_1, x_1]\right) = 5
\end{aligned}
$$

pa je

$$
P_4(x) = -1 - 2x + 3x^2 + 6x^2(x - 1) + 5x^2(x - 1)^2 = 5x^4 - 4x^3 + 2x^2 - 2x - 1.
$$

Ocenu greške interpolacije polinomom (46) možemo izvesti pomoću Lagrangeove ocene greške (9). Pretpostavimo da su nizovi x_{ij}^{ϵ} izabrani tako da $x_{ij}^{\epsilon} \in [y_1, y_2]$, gde je $y_1 = \min(x, x_0, \ldots, x_m)$ i $y_2 = \max(x, x_0, \ldots, x_m)$. Greška interpolacije polinomom (40) je

$$f(x) - P_n^{\epsilon}(x) = \frac{f^{(n+1)}(\eta_\epsilon)}{(n+1)!}\omega_{n+1}^{\epsilon}(x), \qquad \eta_\epsilon \in [y_1, y_2], \tag{47}$$

gde je

$$\omega_{n+1}^{\epsilon}(x) = \prod_{i=0}^{m}\prod_{j=0}^{n_i-1}(x - x_{ij}^{\epsilon}). \tag{48}$$

Kada $\epsilon \to 0$ imamo da je

$$\lim_{\epsilon\to 0}\omega_{n+1}^{\epsilon}(x) = \omega_{n+1}(x) \equiv \prod_{i=0}^{m}(x - x_i)^{n_i}, \tag{49}$$

$$\lim_{\epsilon\to 0} f^{(n+1)}(\eta_\epsilon) = z(x). \tag{50}$$

Kako $f \in \mathcal{C}^{n+1}[y_1, y_2]$, to je

$$m_{n+1} = \min_{x\in[y_1,y_2]} f^{(n+1)}(x) \le f^{(n+1)}(\eta_\epsilon) \le \max_{x\in[y_1,y_2]} f^{(n+1)}(x) = M_{n+1},$$

te puštajući da $\epsilon \to 0$ dobijamo ocenu

$$m_{n+1} \le z(x) \le M_{n+1}.$$

Vrednost $z(x)$ je između najmanje i najveće vrednosti $(n+1)$-og izvoda funkcije f, te prema teoremi o srednjoj vrednosti postoji tačka $\eta \in [y_1, y_2]$ takva da je

$$z(x) = f^{(n+1)}(\eta). \tag{51}$$

Uzimajući u obzir (49), (50) i (51), kada u (47) pustimo da $\epsilon \to 0$ dobijamo ocenu greške Hermiteovog interpolacionog polinoma (46)

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\eta)}{(n+1)!}\,\omega_{n+1}(x), \qquad \eta \in [y_1, y_2]. \tag{52}$$

Grešku interpolacije Hermiteovim interpolacionim polinomom možemo izraziti i pomoću podeljenih razlika sa višestrukim čvorovima. Naime, greška interpolacije polinomom (40) je, prema (20),

$$f(x) - P_n^{\epsilon}(x) = f[x, x_{00}^{\epsilon}, \ldots, x_{m,n_m-1}^{\epsilon}]\,\omega_{n+1}^{\epsilon}(x),$$

što, kada $\epsilon \to 0$, daje

$$f(x) - P_n(x) = f[x, \underbrace{x_0, \ldots, x_0}_{n_0 \text{ puta}}, \ldots, \underbrace{x_m, \ldots, x_m}_{n_m \text{ puta}}]\,\omega_{n+1}(x). \tag{53}$$

Iz (52) i (53) sledi da je

$$f[x, \underbrace{x_0, \ldots, x_0}_{n_0 \text{ puta}}, \ldots, \underbrace{x_m, \ldots, x_m}_{n_m \text{ puta}}] = \frac{f^{(n+1)}(\eta)}{(n+1)!}. \tag{54}$$

Jednakost (54) važi i kada $x \to x_k$, te je možemo zapisati u opštijem obliku

$$f[x_0, \ldots, x_N] = \frac{f^{(N)}(\eta)}{N!}, \tag{55}$$

pri čemu ne moraju svi x_i, $i = 0, \ldots, N$, biti među sobom različiti. To znači da (24) važi i kada su čvorovi višestruki.

2.5 Splajn interpolacija

Lagrangeov interpolacioni polinom višeg stepena osciluje između čvorova. Da bi se ove oscilacije izbegle, pri interpolaciji se koriste tzv. splajnovi.

Označimo sa Δ podelu intervala $[a, b]$ na n podintervala,

$$\Delta = \{a = x_0 < x_1 < \cdots < x_n = b\}. \tag{56}$$

DEFINICIJA. *Splajn reda m definisan podelom (56), $S_\Delta^m(x)$, je realna funkcija na intervalu $[a, b]$, koja ima sledeće osobine*

(*i*) $S_\Delta^m \in \mathcal{C}^{m-1}[a, b]$,

(*ii*) *S_Δ^m je polinom stepena m na svakom od intervala $[x_i, x_{i+1}]$, $i = 0, \ldots, n-1$, podele (56).*

Najčešće se koristi kubni splajn, funkcija koja je na svakom od intervala $[x_i, x_{i+1}]$, $i = 0, \ldots, n-1$, polinom trećeg stepena; u tački spajanja dva takva polinoma, čvoru x_i, $i = 1, \ldots, n-1$, ovi polinomi, kao i njihovi prvi i drugi izvodi imaju jednake vrednosti. U daljem tekstu će biti reči samo o kubnim splajnovima, te će u oznaci biti izostavljen indeks m.

Interpolacioni splajn funkcije $f(x)$ određen na skupu čvorova podele (56) označićemo sa $S_\Delta(f; x)$, i on u čvorovima zadovoljava uslove interpolacije

$$S_\Delta(f; x_i) = f(x_i), \qquad i = 0, \ldots, n.$$

Sa $(n+1)$-im uslovom interpolacije i uslovima neprekidnosti prvog i drugog izvoda u čvorovima x_i, $i = 1, \ldots, n-1$, splajn $S_\Delta(f; x)$ nije jednoznačno određen. Na svakom od intervala $[x_i, x_{i+1}]$ on je oblika $P_{3,i}(x) = c_{0,i}x^3 + c_{1,i}x^2 + c_{2,i}x + c_{3,i}$, što znači da ima ukupno $4n$ slobodnih parametara, a broj zadatih uslova je $4n-2$, jer uslova interpolacije ima $2n$, a glatkosti $2(n-1)$. Zadavanjem preostala dva uslova na granici interpolacioni splajn je jedinstveno određen, što se dokazuje sledećom teoremom.

TEOREMA 4. *Neka je podelom* $\Delta = \{a = x_0 < x_1 < \cdots < x_n = b\}$ *intervala* $[a, b]$ *definisan interpolacioni kubni splajn funkcije* $f(x) \in \mathcal{C}^2[a, b]$, *koji zadovoljava jedan od uslova na granici*

(i) $S''_\Delta(f; a) = S''_\Delta(f; b) = 0$,

(ii) $f(x)$ *i* $S_\Delta(f; x)$ *su periodične funkcije na* $[a, b]$,

(iii) $f'(a) = S'_\Delta(f; a), \quad f'(b) = S'_\Delta(f; b)$.

Tada je splajn $S_\Delta(f; x)$ jedinstveno određen, i važi da je

$$\|f\|^2 \geq \|S_\Delta\|^2, \tag{57}$$

ili, tačnije,

$$\|f - S_\Delta\|^2 = \|f\|^2 - \|S_\Delta\|^2 \geq 0, \tag{58}$$

gde je

$$\|f\|^2 = \int_a^b \big(f''(x)\big)^2\, dx. \tag{59}$$

Splajn koji zadovoljava na granici uslove (*i*) teoreme naziva se *prirodni kubni splajn*, a onaj koji zadovoljava uslove (*ii*) naziva se *periodični kubni splajn*.

Nejednakošću (57) je izraženo tzv. *svojstvo minimalnosti* kubnog splajna. To znači da među svim funkcijama $f \in \mathcal{C}^2[a, b]$ koje imaju date vrednosti u čvorovima podele (56), kubni splajn je ona funkcija koja ima minimalnu vrednost integrala (59). Kako je krivina funkcije f data izrazom $f''(1 + f'^2)^{-3/2}$ i približno je jednaka f'' ako je f' malo, znači da je kubni splajn interpolaciona funkcija najveće glatkosti.

Da bismo dokazali teoremu 4, potrebna nam je sledeća lema:

LEMA 4. *Ako je* $f \in \mathcal{C}^2[a, b]$ *i* $S_\Delta(x)$ *kubni splajn određen podelom (56), onda je*

$$\|f - S_\Delta\|^2 = \|f\|^2 - \|S_\Delta\|^2 - 2\bigg(\big(f'(x) - S'_\Delta(x)\big)S''_\Delta(x)\Big|_a^b - \sum_{i=1}^{n}\big(f(x) - S_\Delta(x)\big)S'''_\Delta(x)\Big|_{x_{i-1}^+}^{x_i^-}\bigg),$$

gde je

$$g(x)\Big|_{x_{i-1}^+}^{x_i^-} = \lim_{x \to x_i - 0} g(x) - \lim_{x \to x_{i-1} + 0} g(x).$$

DOKAZ: Prema (59) je

$$\begin{aligned}\|f - S_\Delta\|^2 &= \int_a^b \big(f''(x) - S''_\Delta(x)\big)^2\, dx \\ &= \int_a^b \big(f''(x)\big)^2\, dx - 2\int_a^b f''(x)S''_\Delta(x)\, dx + \int_a^b \big(S''_\Delta(x)\big)^2\, dx \\ &= \|f\|^2 - 2\int_a^b \big(f''(x) - S''_\Delta(x)\big)S''_\Delta(x)\, dx - \|S_\Delta\|^2.\end{aligned}$$

Parcijalnom integracijom na intervalu $[x_{i-1}, x_i]$, $i = 1, \ldots, n$, dobijamo da je

$$\begin{aligned}\int_{x_{i-1}}^{x_i} \big(f''(x) - S''_\Delta(x)\big) S''_\Delta(x)\, dx \\ &= \big(f'(x) - S'_\Delta(x)\big) S''_\Delta(x)\big|_{x_{i-1}}^{x_i} - \int_{x_{i-1}}^{x_i} \big(f'(x) - S'_\Delta(x)\big) S'''_\Delta(x)\, dx \\ &= \big(f'(x) - S'_\Delta(x)\big) S''_\Delta(x)\big|_{x_{i-1}}^{x_i} - \big(f(x) - S_\Delta(x)\big) S'''_\Delta(x)\big|_{x_{i-1}^+}^{x_i^-} \\ &\qquad + \int_{x_{i-1}}^{x_i} \big(f(x) - S_\Delta(x)\big) S^{(4)}_\Delta(x)\, dx.\end{aligned}$$

$f'(x)$, $S'_\Delta(x)$ i $S''_\Delta(x)$ su neprekidne funkcije na $[a, b]$, a na intervalima $[x_{i-1}, x_i]$, $i = 1, \ldots, n$, je $S_\Delta(x)$ polinom trećeg stepena te je $S^{(4)}_\Delta(x) \equiv 0$. Stoga je

$$\begin{aligned}\int_a^b \big(f''(x) - S''_\Delta(x)\big) S''_\Delta(x)\, dx &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} \big(f''(x) - S''_\Delta(x)\big) S''_\Delta(x)\, dx \\ &= \sum_{i=1}^n \big(f'(x) - S'_\Delta(x)\big) S''_\Delta(x)\big|_{x_{i-1}}^{x_i} - \sum_{i=1}^n \big(f(x) - S_\Delta(x)\big) S'''_\Delta(x)\big|_{x_{i-1}^+}^{x_i^-} \\ &= \big(f'(x) - S'_\Delta(x)\big) S''_\Delta(x)\big|_a^b - \sum_{i=1}^n \big(f(x) - S_\Delta(x)\big) S'''_\Delta(x)\big|_{x_{i-1}^+}^{x_i^-},\end{aligned}$$

što dokazuje tvrđenje leme. ∎

Dokažimo sada teoremu 4.

DOKAZ: Pri ma kojoj od pretpostavki (*i*), (*ii*) ili (*iii*) je

$$\big(f'(x) - S'_\Delta(x)\big) S''_\Delta(x)\big|_a^b - \sum_{i=0}^n \big(f(x) - S_\Delta(x)\big) S'''_\Delta(x)\big|_{x_{i-1}^+}^{x_i^-} = 0,$$

te na osnovu leme 4 sledi (58), tj. svojstvo minimalnosti kubnog splajna (57).

Dokažimo još da je, ako kubni splajn određen uslovima teoreme postoji, on jedinstveno određen. Pretpostavimo suprotno, tj. da postoje dva različita splajna $S^1_\Delta(f; x)$ i $S^2_\Delta(f; x)$ koji zadovoljavaju pretpostavke teoreme. U izrazu (58) zamenimo $f(x)$ sa $S^2_\Delta(f; x)$, pa dobijamo da je

$$\|S^2_\Delta - S^1_\Delta\|^2 = \|S^2_\Delta\|^2 - \|S^1_\Delta\|^2 \geq 0,$$

a kada S^2_Δ i S^1_Δ zamene mesta, imamo da je

$$\|S^1_\Delta - S^2_\Delta\|^2 = \|S^1_\Delta\|^2 - \|S^2_\Delta\|^2 \geq 0.$$

Stoga je

$$\|S^2_\Delta - S^1_\Delta\|^2 = \int_a^b \big(S^2_\Delta{}''(f; x) - S^1_\Delta{}''(f; x)\big)^2\, dx = 0,$$

odakle, zbog neprekidnosti funkcija $S_\Delta^1{}''(f;x)$ i $S_\Delta^2{}''(f;x)$, sledi da je

$$S_\Delta^2{}''(f;x) = S_\Delta^1{}''(f;x),$$

tj.

$$S_\Delta^2(f;x) = S_\Delta^1(f;x) + cx + d.$$

Kako je $S_\Delta^2(f;x) = S_\Delta^1(f;x)$ za $x = a$ i $x = b$, mora biti $c = d = 0$.

Egzistencija kubnog splajna će biti dokazana konstruktivnim algoritmom koji sledi. ∎

Neka su zadati podela Δ intervala $[a,b]$, i realni brojevi $f_k = f(x_k)$, $k = 0,\dots,n$. Označimo dužine intervala sa h_i,

$$h_i = x_i - x_{i-1}, \qquad i = 1,\dots,n,$$

a sa M_i označimo *momente* traženog splajna $S_\Delta(f;x)$,

$$M_i = S''_\Delta(f;x_i), \qquad i = 0,\dots,n.$$

Na svakom od intervala $[x_i, x_{i+1}]$, $i = 0,\dots,n-1$, funkcija $S''_\Delta(f;x)$ je linearna, te je potpuno određena svojim vrednostima M_i i M_{i+1} na krajevima intervala,

$$S''_\Delta(f;x) = M_i\frac{x_{i+1}-x}{h_{i+1}} + M_{i+1}\frac{x-x_i}{h_{i+1}}, \qquad x \in [x_i, x_{i+1}]. \tag{60}$$

Integraljenjem izraza (60) dobijamo da je na intervalu $[x_i, x_{i+1}]$

$$\begin{aligned} S'_\Delta(f;x) &= -M_i\frac{(x_{i+1}-x)^2}{2h_{i+1}} + M_{i+1}\frac{(x-x_i)^2}{2h_{i+1}} + A_i, \\ S_\Delta(f;x) &= M_i\frac{(x_{i+1}-x)^3}{6h_{i+1}} + M_{i+1}\frac{(x-x_i)^3}{6h_{i+1}} + A_i(x-x_i) + B_i, \end{aligned} \tag{61}$$

gde su A_i i B_i konstante integracije. One su rešenja sistema linearnih jednačina

$$\begin{aligned} M_i\frac{h_{i+1}^2}{6} + B_i &= f_i \\ M_{i+1}\frac{h_{i+1}^2}{6} + A_i h_{i+1} + B_i &= f_{i+1} \end{aligned}$$

određenog uslovima $S_\Delta(f;x_i) = f_i$ i $S_\Delta(f;x_{i+1}) = f_{i+1}$, tj.

$$\begin{aligned} A_i &= \frac{f_{i+1}-f_i}{h_{i+1}} - \frac{h_{i+1}}{6}(M_{i+1}-M_i) \\ B_i &= f_i - M_i\frac{h_{i+1}^2}{6}\,. \end{aligned} \tag{62}$$

Uobičajeni način zapisa splajna na intervalu $[x_i, x_{i+1}]$ je

$$S_\Delta(f;x) = \alpha_i + \beta_i(x-x_i) + \gamma_i(x-x_i)^2 + \delta_i(x-x_i)^3,$$

gde je, na osnovu (60), (61) i (62),

$$\begin{aligned}
\alpha_i &= S_\Delta(f;x_i) = f_i \\
\beta_i &= S'_\Delta(f;x_i) = -M_i\frac{h_{i+1}}{2} + A_i = \frac{f_{i+1}-f_i}{h_{i+1}} - \frac{h_{i+1}}{6}(2M_i + M_{i+1}) \\
\gamma_i &= \tfrac{1}{2}S''_\Delta(f;x_i) = \tfrac{1}{2}M_i \\
\delta_i &= \tfrac{1}{6}S'''_\Delta(f;x_i) = \frac{1}{6h_{i+1}}(M_{i+1}-M_i).
\end{aligned}$$

Svi koeficijenti splajna su izraženi u funkciji momenata, pa ove treba odrediti.

Do sada još nije korišćen uslov neprekidnosti funkcije $S'_\Delta(f;x)$ u unutrašnjim čvorovima,

$$\lim_{x\to x_i-0} S'_\Delta(f;x) = \lim_{x\to x_i+0} S'_\Delta(f;x) \qquad i = 1,\dots,n-1. \tag{63}$$

Uslovi (63) određuju $(n-1)$-u jednačinu po momentima M_i. Na osnovu (61) i (62) je na intervalu $[x_i, x_{i+1}]$

$$S'_\Delta(f;x) = -M_i\frac{(x_{i+1}-x)^2}{2h_{i+1}} + M_{i+1}\frac{(x-x_i)^2}{2h_{i+1}} + \frac{f_{i+1}-f_i}{h_{i+1}} - \frac{h_{i+1}}{6}(M_{i+1}-M_i),$$

pa je

$$\lim_{x\to x_i-0} S'_\Delta(f;x) = \frac{f_i-f_{i-1}}{h_i} + \frac{h_i}{6}M_{i-1} + \frac{h_i}{3}M_i,$$

$$\lim_{x\to x_i+0} S'_\Delta(f;x) = \frac{f_{i+1}-f_i}{h_{i+1}} - \frac{h_{i+1}}{3}M_i - \frac{h_{i+1}}{6}M_{i+1}.$$

Uslovi (63) se stoga svode na sistem linearnih jednačina po M_i

$$\frac{h_i}{6}M_{i-1} + \frac{h_i+h_{i+1}}{3}M_i + \frac{h_{i+1}}{6}M_{i+1} = \frac{f_{i+1}-f_i}{h_{i+1}} - \frac{f_i-f_{i-1}}{h_i}, \qquad i = 1,\dots,n-1, \tag{64}$$

koji može da se zapiše i u sledećem obliku

$$\mu_i M_{i-1} + 2M_i + \nu_i M_{i+1} = \lambda_i, \qquad i = 1,\dots,n-1, \tag{65}$$

gde je

$$\begin{aligned}
&\nu_i = \frac{h_{i+1}}{h_i+h_{i+1}}, \qquad \mu_i = \frac{h_i}{h_i+h_{i+1}} = 1-\nu_i, \\
&\lambda_i = \frac{6}{h_i+h_{i+1}}\left(\frac{f_{i+1}-f_i}{h_{i+1}} - \frac{f_i-f_{i-1}}{h_i}\right) = 6f[x_{i-1},x_i,x_{i+1}],
\end{aligned} \qquad i = 1,\dots,n-1. \tag{66}$$

Sistem (65) je sistem od $(n-1)$-e jednačine sa $(n+1)$-om nepoznatom, te se preostale dve veze između momenata dobijaju iz graničnih uslova navedenih u teoremi 4.

(*i*) Kod prirodnog splajna je $S''_\Delta(f;a) = S''_\Delta(f;b) = 0$, pa je $M_0 = M_n = 0$.

(*ii*) Pretpostavka o periodičnosti povlači da je $S'_\Delta(f;a) = S'_\Delta(f;b)$, odnosno

$$\lim_{x\to a+0} S'_\Delta(f;x) = \lim_{x\to b-0} S'_\Delta(f;x),$$

što daje jednačinu

$$\frac{h_1}{3}M_0 + \frac{h_1}{6}M_1 + \frac{h_n}{6}M_{n-1} + \frac{h_n}{3}M_n = \frac{f_1 - f_0}{h_1} - \frac{f_n - f_{n-1}}{h_n}.$$

Kako je zbog periodičnosti $f_0 = f_n$ i $M_0 = M_n$, poslednja jednačina postaje

$$\frac{h_n}{6}M_{n-1} + \frac{h_n + h_1}{3}M_n + \frac{h_1}{6}M_1 = \frac{f_1 - f_n}{h_1} - \frac{f_n - f_{n-1}}{h_n}.$$

i oblika je (64) za $i = n$, ako stavimo da je $h_{n+1} = h_1$, $M_{n+1} = M_1$ i $f_{n+1} = f_1$.

(*iii*) Zadavanjem vrednosti prvog izvoda splajna na levoj i desnoj granici, $S'_\Delta(f;a) = f'(a) = f'_0$ i $S'_\Delta(f;b) = f'(b) = f'_n$, dobijamo jednačine

$$\frac{h_1}{3}M_0 + \frac{h_1}{6}M_1 = \frac{f_1 - f_0}{h_1} - f'_0,$$

$$\frac{h_n}{6}M_{n-1} + \frac{h_n}{3}M_n = f'_n - \frac{f_n - f_{n-1}}{h_n}.$$

Objedinjujući zapis jednačina dobijenih iz uslova (*i*) ili (*iii*) na granici sa jednačinama u unutrašnjim čvorovima, sledi da (65) važi i za $i = 0$ i $i = n$, tako da je pri graničnim uslovima (*i*)

$$\mu_0 = \nu_0 = \lambda_0 = 0, \qquad \text{i} \qquad \mu_n = \nu_n = \lambda_n = 0,$$

a pri graničnim uslovima (*iii*)

$$\begin{aligned} &\mu_0 = 0, \qquad \nu_0 = 1, \qquad \lambda_0 = \frac{6}{h_1}\left(\frac{f_1 - f_0}{h_1} - f'_0\right), \\ &\mu_n = 1, \qquad \nu_n = 0, \qquad \lambda_n = \frac{6}{h_n}\left(f'_n - \frac{f_n - f_{n-1}}{h_n}\right). \end{aligned} \tag{67}$$

Matrični zapis sistema linearnih jednačina po momentima M_i, $i = 0, \ldots, n$, u slučaju važenja graničnih uslova (*i*) ili (*iii*), je

$$\begin{pmatrix} 2 & \nu_0 & & & \\ \mu_1 & 2 & \nu_1 & 0 & \\ & \mu_2 & 2 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & 0 & & 2 & \nu_{n-1} \\ & & & \mu_n & 2 \end{pmatrix} \cdot \begin{pmatrix} M_0 \\ M_1 \\ M_2 \\ \cdots \\ M_{n-1} \\ M_n \end{pmatrix} = \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_{n-1} \\ \lambda_n \end{pmatrix}. \tag{68}$$

U slučaju periodičnih graničnih uslova (*ii*) sistem ima n jednačina (jer je $M_0 = M_n$), tj. (65) važi i za $i = n$, pri čemu je

$$\nu_n = \frac{h_1}{h_n + h_1}, \qquad \mu_n = \frac{h_n}{h_n + h_1} = 1 - \nu_n, \tag{69}$$
$$\lambda_n = \frac{6}{h_n + h_1}\left(\frac{f_1 - f_n}{h_1} - \frac{f_n - f_{n-1}}{h_n}\right).$$

Matrični zapis ovoga sistema je

$$\begin{pmatrix} 2 & \nu_1 & & & \mu_1 \\ \mu_2 & 2 & \nu_2 & 0 & \\ & \mu_3 & 2 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & 0 & & 2 & \nu_{n-1} \\ \nu_n & & & \mu_n & 2 \end{pmatrix} \cdot \begin{pmatrix} M_1 \\ M_2 \\ M_3 \\ \cdots \\ M_{n-1} \\ M_n \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \cdots \\ \lambda_{n-1} \\ \lambda_n \end{pmatrix}. \tag{70}$$

Koeficijenti sistema (68) i (70), dati izrazima (66),(67) i (69), očigledno zadovoljavaju uslove

$$\nu_i \geq 0, \qquad \mu_i \geq 0, \qquad \nu_i + \mu_i = 1, \tag{71}$$

i zavise samo od izabrane podele (56), a ne zavise od zadatih brojeva f_i, i f_0' i f_n' u graničnom slučaju (*iii*). Ovi zaključi će biti korišćeni u dokazu leme koja sledi, a na osnovu koje se konačno garantuje egzistencija kubnog splajna pri ma kom izboru graničnih uslova (*i*), (*ii*) ili (*iii*).

LEMA 5. *Sistemi linearnih jednačina (68) i (70) imaju jedinstveno rešenje pri proizvoljnoj podeli (56) intervala* $[a, b]$.

DOKAZ: Dokažimo nesingularnost matrice

$$A = \begin{pmatrix} 2 & \nu_0 & & & \\ \mu_1 & 2 & \nu_1 & 0 & \\ & \mu_2 & 2 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & 0 & & 2 & \nu_{n-1} \\ & & & \mu_n & 2 \end{pmatrix}$$

dimenzije $(n+1) \times (n+1)$ sistema (68). Da bismo to pokazali, dokažimo da ona ima sledeću osobinu:

$$A\,\mathbf{z} = \mathbf{w} \qquad \Longrightarrow \qquad \max_{0 \leq i \leq n} |z_i| \leq \max_{0 \leq i \leq n} |w_i| \tag{72}$$

za svaki par vektora $\mathbf{z} = (z_0, \ldots, z_n)^T$ i $\mathbf{w} = (w_0, \ldots, w_n)^T$. Neka je

$$|z_m| = \max_{0 \leq i \leq n} |z_i|.$$

Iz $A\,\mathbf{z} = \mathbf{w}$ sledi da je

$$\mu_m z_{m-1} + 2z_m + \nu_m z_{m+1} = w_m,$$

pa je, obzirom na (71) i način izbora m,

$$\begin{aligned}\max_{0\le i\le n} |w_i| \ge |w_m| &\ge 2|z_m| - \mu_m|z_{m-1}| - \nu_m|z_{m+1}| \\ &\ge 2|z_m| - \mu_m|z_m| - \nu_m|z_m| = (2-\mu_m-\nu_m)|z_m| \\ &\ge |z_m| = \max_{0\le i\le n} |z_i|.\end{aligned}$$

Pretpostavimo, sada, da je matrica A singularna, tj. da postoji vektor $\mathbf{z} \neq \mathbf{0}$ takav da je $A\,\mathbf{z} = \mathbf{0}$. Tada, iz (72) sledi da je $0 < \max_{0\le i\le n} |z_i| \le 0$, što je nemoguće. Nesingularnost matrice sistema (70) se dokazuje analogno, pošto je ona iste strukture kao i matrica A, samo dimenzije $n \times n$. ∎

Sistem (68) se efikasno rešava Gaussovom metodom eliminacije, koja je data u §5.2. Momente računamo rekurentnom formulom

$$M_n = r_n, \qquad M_i = q_i M_{i+1} + r_i = -\frac{\nu_i}{p_i} M_{i+1} + \frac{\lambda_i - \mu_i r_{i-1}}{p_i}, \quad i = n-1, \dots, 0,$$

pri čemu su koeficijenti p_i, q_i i r_i prethodno određeni rekurentnim formulama

$$\begin{aligned} &p_0 = 2, && q_0 = -\tfrac{\nu_0}{p_0}, && r_0 = \tfrac{\lambda_0}{p_0} \\ &p_i = \mu_i q_{i-1} + 2, && q_i = -\tfrac{\nu_i}{p_i}, && r_i = \tfrac{\lambda_i - \mu_i r_{i-1}}{p_i}, \quad i = 1, \dots, n. \end{aligned} \tag{73}$$

Algoritam je numerički stabilan, jer je $p_i > 1$ za svako i, što možemo dokazati indukcijom. Prema (73) je $p_0 = 2$, i pretpostavimo da je $p_i > 1$ za svako $i \le m$. Tada je, na osnovu (73),

$$p_{m+1} = \mu_{m+1} q_m + 2 = \mu_{m+1}\big(-\frac{\nu_m}{p_m}\big) + 2 = 2 - \frac{\nu_m \mu_{m+1}}{p_m}.$$

Iz (71) sledi da je $0 \le \nu_i \le 1$ i $0 \le \mu_i \le 1$ za svako i, te je $\nu_m \mu_{m+1} \le 1$. Kako je po pretpostavci $p_m > 1$, to je $\nu_m \mu_{m+1}/p_m < 1$, te je $p_{m+1} > 1$.

PRIMER 5. Konstruišimo interpolacioni kubni splajn periodične, sa periodom tri, funkcije $f(x)$ zadate tabelom

x_i	0	1	2	3
$f(x_i)$	1	3	2	1

U skladu sa korišćenim oznakama, u ovom primeru je

$$n = 3, \qquad h_i = 1, \qquad \nu_i = \mu_i = \frac{1}{2}, \qquad i = 1, 2, 3$$

$$\lambda_i = 3(f_{i+1} - 2f_i + f_{i-1}) = \begin{cases} -9, & i = 1, \\ 0, & i = 2, \\ 9, & i = 3, \end{cases}$$

te je sistem linearnih jednačina po momentima (70)

$$2M_1 + \tfrac{1}{2}M_2 + \tfrac{1}{2}M_3 = -9$$

$$\tfrac{1}{2}M_1 + 2M_2 + \tfrac{1}{2}M_3 = 0$$

$$\tfrac{1}{2}M_1 + \tfrac{1}{2}M_2 + 2M_3 = 9.$$

Njegovo rešenje je $M_1 = -6$, $M_2 = 0$, $M_3 = 6$, tako da je traženi splajn

$$S_\Delta(f; x) = \begin{cases} 1 + x + 3x^2 - 2x^3, & x \in (0, 1), \\ 3 + (x-1) - 3(x-1)^2 + (x-1)^3, & x \in (1, 2), \\ 2 - 2(x-2) + (x-2)^3, & x \in (2, 3). \end{cases}$$

2.6 Drugi vidovi interpolacije

U ovom odeljku će biti reči ukratko o nekim drugim vidovima interpolacije koji se češće koriste.

Interpolacija racionalnim funkcijama. Interpolaciona funkcija se traži u obliku količnika dva polinoma,

$$R_{i,\dots,i+k}(x) = \frac{P_m(x)}{Q_n(x)} = \frac{p_0 + p_1 x + \cdots + p_m x^m}{q_0 + q_1 x + \cdots + q_n x^n}, \tag{74}$$

pri čemu je $R_{i,\dots,i+k}(x)$ određeno sa $(m + n + 1)$-im uslovom interpolacije

$$R_{i,\dots,i+k}(x_{i+j}) = f(x_{i+j}), \qquad j = 0, \dots, k, \quad k = m + n. \tag{75}$$

U reprezentaciji (74) se deljenjem brojioca i imenioca sa jednim od koeficijenata polinoma $P_m(x)$ ili $Q_n(x)$ broj nepoznatih parametara smanjuje za jedan, odnosno ima ih $m+n+1$, te su oni jednoznačno određeni uslovima (75). Stoga je racionalna interpolacija određena zadavanjem uređenog para (m, n) i niza realnih brojeva f_j, $j = 0, \dots, m + n$ koji predstavljaju vrednosti funkcije $f(x)$ u čvorovima interpolacije, $f_j = f(x_{i+j})$, $j = 0, \dots, m + n$.

Ukoliko je potrebno izračunati približnu vrednost funkcije $f(x)$ za dato x racionalnom interpolacijom (74), nije najpogodnije, posebno za veće m i n, to učiniti nalaženjem eksplicitnog izraza za funkciju $R_{i,\dots,i+k}(x)$ korišćenjem uslova (75). Pogodnije je, ukoliko je $m = n$ ili $m = n-1$, koristiti algoritam Nevilleovog tipa ([26])

$$R_{i+j}(x) \equiv f_j, \quad j = 0, \dots, m+n, \qquad (R_{i+1,i}(x) \stackrel{\text{def}}{=} 0),$$
$$R_{i,\dots,i+k}(x) = \frac{R_{i+1,\dots,i+k}(x) - R_{i,\dots,i+k-1}(x)}{\frac{x-x_i}{x-x_{i+k}}\left(1 - \frac{R_{i+1,\dots,i+k}(x)-R_{i,\dots,i+k-1}(x)}{R_{i+1,\dots,i+k}(x)-R_{i+1,\dots,i+k-1}(x)}\right) - 1} + R_{i+1,\dots,i+k}(x),$$
$$k = 1, \dots, m+n,$$

koji se može jednostavnije zapisati uvođenjem smena analognih formulama (17).

Interpolacija racionalnim funkcijama ima prednost nad polinomijalnom interpolacijom posebno kod interpolacije funkcija sa izrazitim ekstremima.

Interpolacija trigonometrijskim funkcijama. Za interpolaciju periodičnih funkcija bolje je koristiti trigonometrijske funkcije. Jedna od takvih formula za interpolaciju 2π-periodične funkcije je *formula Hermitea*

$$f(x) \approx \sum_{i=0}^{n}\left(\prod_{\substack{j=0\\ j\neq i}}^{n} \frac{\sin(x-x_j)}{\sin(x_i-x_j)}\right) f(x_i). \tag{76}$$

Ona odgovara Lagrangeovoj formuli (7) za neperiodične funkcije, i koristi se za proizvoljan raspored čvorova interpolacije.

Problem trigonometrijske interpolacije prvi je rešio Gauss, koji je izveo nekoliko formula sličnih Hermiteovoj. Formula koja se obično naziva *Gaussova formula*, razlikuje se od formule (76) u činiocu $\frac{1}{2}$ koji se javlja u argumentu sinusa. Interpolaciona funkcija se traži u obliku trigonometrijskog polinoma

$$Q_n(x) = a_0 + \sum_{k=1}^{n}(a_k \cos kx + b_k \sin kx), \tag{77}$$

čiji su koeficijenti određeni uslovima interpolacije $Q_n(x_i) = f(x_i)$, $i = 0, \dots, 2n$, u čvorovima $0 \le x_0 < x_1 < \cdots < x_{2n} < 2\pi$. Oni su, stoga, rešenja sistema linearnih jednačina

$$\begin{aligned} f(x_0) &= a_0 + \sum_{k=1}^{n}(a_k \cos kx_0 + b_k \sin kx_0)\\ f(x_1) &= a_0 + \sum_{k=1}^{n}(a_k \cos kx_1 + b_k \sin kx_1)\\ &\vdots\\ f(x_{2n}) &= a_0 + \sum_{k=1}^{n}(a_k \cos kx_{2n} + b_k \sin kx_{2n}). \end{aligned} \tag{78}$$

Determinanta sistema (78)

$$\begin{vmatrix} 1 & \cos x_0 & \sin x_0 & \dots & \cos nx_0 & \sin nx_0 \\ 1 & \cos x_1 & \sin x_1 & \dots & \cos nx_1 & \sin nx_1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & \cos x_{2n} & \sin x_{2n} & \dots & \cos nx_{2n} & \sin nx_{2n} \end{vmatrix} = 2^{n^2} \prod_{0 \le p < q \le 2n} \sin \frac{x_q - x_p}{2}$$

je različita od nule, jer je $0 < x_q - x_p < 2\pi$, pa je $\sin \frac{x_q - x_p}{2} \neq 0$ za svako p i q.

Sistemu (78) pridružimo jednačinu (77) napisanu za $x \neq x_k$, $k = 0, \dots, 2n$

$$\begin{aligned} -Q_n(x) + a_0 + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx) &= 0 \\ -f(x_0) + a_0 + \sum_{k=1}^{n} (a_k \cos kx_0 + b_k \sin kx_0) &= 0 \\ \vdots & \\ -f(x_{2n}) + a_0 + \sum_{k=1}^{n} (a_k \cos kx_{2n} + b_k \sin kx_{2n}) &= 0. \end{aligned}$$

Ako ovaj sistem posmatramo kao sistem od $2n+2$ jednačine po $2n+2$ koeficijenta a_0, a_k, b_k, $k = 1, \dots, 2n$, i po koeficijentu $c = -1$ uz $Q_n(x)$ i $f(x_k)$, $k = 0, \dots, 2n$, onda njegova determinanta mora biti jednaka nuli,

$$\begin{vmatrix} Q_n(x) & 1 & \cos x & \sin x & \dots & \cos nx & \sin nx \\ f(x_0) & 1 & \cos x_0 & \sin x_0 & \dots & \cos nx_0 & \sin nx_0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ f(x_{2n}) & 1 & \cos x_{2n} & \sin x_{2n} & \dots & \cos nx_{2n} & \sin nx_{2n} \end{vmatrix} = 0, \tag{79}$$

jer je sistem homogen a ima netrivijalno rešenje. Razvijanjem determinante po elementima prve kolone i izražavanjem $Q_n(x)$, iz (79) se dobija Gaussova formula za trigonometrijsku interpolaciju.

Inverzna interpolacija. Inverzna interpolacija je postupak kojim se određuje približna vrednost argumenta za koju funkcija $f(x)$ ima zadatu vrednost Y. Pretpostavlja se da je funkcija zadata svojim vrednostima y_i u diskretnim tačkama x_i, $f(x_i) = y_i$, $i = 0, \dots, n$, i da je Y između dve tabelirane vrednosti y_i. Ako postoji dovoljno glatka inverzna funkcija $g(y)$ funkciji $f(x)$, ona se može interpolisati polinomom koji zadovoljava uslove

$$L_n(y_i) = x_i, \qquad i = 0, \dots, n,$$

te će biti $g(Y) \approx L_n(Y)$.

Ako je funkcija $f(x)$ zadata u ravnomerno raspoređenim čvorovima, može se interpolisati nekim od polinoma sa konačnim razlikama $L_n(x)$, i umesto jednačine

$f(x) = Y$ rešavati jednačina $L_n(x) = Y$. Specijalno, ako je funkcija $f(x)$ zadata analitički, treba je tabelirati sa dovoljno malim korakom tako da se linearnom ili kvadratnom interpolacijom postiže željena tačnost u okolini tačke $f(x) = Y$. Jednačina $f(x) = Y$ se zamenjuje jednačinom $L_n(x) = Y$, za $n = 1$ ili $n = 2$, čiji se koreni mogu direktno izračunati.

Inverznom interpolacijom se može odrediti i tačka ekstrema neke funkcije. Prvo se tabelira funkcija sa većim korakom, da bi se lokalizovao položaj tačke ekstrema. Zatim se u okolini te tačke funkcija tabelira sa manjim korakom, pri čemu se korak bira tako da greška najviše kubne interpolacije date funkcije ne premaša dozvoljenu grešku. Pomoću ove tabele se odredi interpolacioni polinom funkcije. Nula izvoda tog interpolacionog polinoma, koja se direktno može izračunati jer je polinom najviše trećeg stepena, predstavlja aproksimaciju apscise tačke ekstrema sa željenom tačnošću.

2.7 Interpolacija funkcija više promenljivih

Jedan od načina da se interpolacijom izračuna u nekoj tački približna vrednost funkcije više promenljivih zadate na skupu diskretnih tačaka, je da se više puta primeni jednodimenziona interpolacija. Na primer, ako je potrebno izračunati vrednost funkcije dve promenljive $f(x, y)$ u tački (X, Y) na osnovu zadatih vrednosti $f(x_i, y_j)$, $i = 0, \ldots, m$, $j = 0, \ldots, n$, onda je to moguće učiniti na sledeći način: $(n+1)$ puta primenjujući jednodimenzionu interpolaciju po promenljivoj x, izračunaju se približno vrednosti $f(X, y_j)$, $j = 0, \ldots, n$, a zatim se pomoću ovih, interpolacijom po promenljivoj y, računa približna vrednost za $f(X, Y)$. Analogno bi se postupilo i u slučaju funkcije više promenljivih.

Drugi način interpolacije funkcije više promenljivih je pomoću višedimenzionih interpolacionih polinoma. Izvešćemo ovaj polinom za funkciju dva argumenta, što ne umanjuje opštost algoritma koji sledi.

Neka je dato $\frac{1}{2}(n+1)(n+2)$ čvorova, zapisanih sledećom tabelom

$$
\begin{array}{ccccc}
(x_0, y_0) & (x_1, y_0) & \ldots & (x_{n-1}, y_0) & (x_n, y_0) \\
(x_0, y_1) & (x_1, y_1) & \ldots & (x_{n-1}, y_1) & \\
\vdots & \vdots & & & \\
(x_0, y_{n-1}) & (x_1, y_{n-1}) & & & \\
(x_0, y_n) & & & &
\end{array} \tag{80}
$$

pri čemu je $x_i \neq x_j$, $y_i \neq y_j$ za $i \neq j$. Vrednosti x_i i y_j mogu biti proizvoljne, dakle raspored čvorova je dovoljno opšti. Zadatim vrednostima funkcije $f(x, y)$ u čvorovima tabele (80) određen je interpolacioni polinom stepena n,

$$L_n(x, y) = c_{00} + c_{10}x + c_{01}y + c_{20}x^2 + \cdots + c_{1,n-1}xy^{n-1} + c_{0n}y^n,$$

jer je $\frac{1}{2}(n+1)(n+2)$ koeficijenta c_{ij} određeno sa $\frac{1}{2}(n+1)(n+2)$ uslova interpolacije

$$L_n(x_i, y_j) = f(x_i, y_j), \qquad 0 \le i + j \le n. \tag{81}$$

Čvorovima za koje je $i+j < n$ je određen interpolacioni polinom $L_{n-1}(x, y)$ stepena najviše $n-1$. Razlika ova dva polinoma se može zapisati u obliku

$$\begin{aligned} L_n(x,y) - L_{n-1}(x,y) = {} & a_{n0}(x-x_0)(x-x_1)\cdots(x-x_{n-1}) \\ & + a_{n-1,1}(x-x_0)\cdots(x-x_{n-2})(y-y_0) \\ & + a_{n-2,2}(x-x_0)\cdots(x-x_{n-3})(y-y_0)(y-y_1) \\ & + \cdots + a_{0n}(y-y_0)\cdots(y-y_{n-1}), \end{aligned}$$

jer je to polinom stepena ne višeg od n čije su nule (x_i, y_j), $i+j < n$. Ako definišemo da je $x - x_{-1} = y - y_{-1} = 1$, sledi da je

$$L_n(x,y) = L_{n-1}(x,y) + \sum_{i=0}^{n} a_{n-i,i}(x-x_0)\cdots(x-x_{n-i-1})(y-y_0)\cdots(y-y_{i-1}).$$

Predstavljajući na isti način $L_{n-1}(x, y)$, zatim $L_{n-2}(x, y)$, ..., dobijamo da je

$$\begin{aligned} L_n(x,y) = {} & a_{00} + a_{10}(x-x_0) + a_{01}(y-y_0) + a_{20}(x-x_0)(x-x_1) \\ & + a_{11}(x-x_0)(y-y_0) + a_{02}(y-y_0)(y-y_1) + \cdots \\ & + a_{n0}(x-x_0)\cdots(x-x_{n-1}) + a_{n-1,1}(x-x_0)\cdots(x-x_{n-2})(y-y_0) \\ & + \cdots + a_{0n}(y-y_0)\cdots(y-y_{n-1}). \end{aligned} \tag{82}$$

Izrazimo sada koeficijente a_{ij} pomoću vrednosti funkcije u čvorovima $f(x_k, y_l)$. Stavljajući u (82) $x = x_0$ i $y = y_0$, na osnovu (81) je $a_{00} = f(x_0, y_0)$. Na isti način, za $x = x_1$ i $y = y_0$ dobija se da je

$$f(x_1, y_0) = a_{00} + a_{10}(x_1 - x_0),$$

pa je

$$a_{10} = \frac{f(x_1, y_0) - f(x_0, y_0)}{x_1 - x_0} = f[x_0, x_1; y_0].$$

Analogno, za $x = x_0$ i $y = y_1$ je

$$a_{01} = \frac{f(x_0, y_1) - f(x_0, y_0)}{y_1 - y_0} = f[x_0; y_0, y_1].$$

Ako u (82) stavimo $y = y_0$, dobijamo interpolacioni polinom po x

$$L_n(x, y_0) = a_{00} + a_{10}(x - x_0) + \cdots + a_{n0}(x - x_0)\cdots(x - x_{n-1}),$$

koji u tački (x_i, y_0) ima vrednost $f(x_i, y_0)$. Stoga je

$$a_{i0} = f[x_0, \dots, x_i; y_0]. \tag{83}$$

Stavljajući $y = y_1$ u (82), dobijamo polinom po x koji može da se zapiše u sledećem obliku

$$\begin{aligned} L_n(x, y_1) = {} & \big(a_{00} + a_{01}(y_1 - y_0)\big) + \big(a_{10} + a_{11}(y_1 - y_0)\big)(x - x_0) \\ & + \big(a_{20} + a_{21}(y_1 - y_0)\big)(x - x_0)(x - x_1) + \cdots \\ & + \big(a_{n-1,0} + a_{n-1,1}(y_1 - y_0)\big)(x - x_0) \cdots (x - x_{n-2}) \\ & + a_{n0}(x - x_0) \cdots (x - x_{n-1}). \end{aligned}$$

Ovaj polinom u tačkama (x_k, y_1), $k = 0, \dots, n-1$, treba da ima vrednosti $f(x_k, y_1)$. Poslednji sabirak se u svakoj od tih tačaka anulira, a koeficijenti u ostalim sabircima su

$$a_{i0} + a_{i1}(y_1 - y_0) = f[x_0, \dots, x_i; y_1],$$

tj., obzirom na (83),

$$a_{i1} = \frac{f[x_0, \dots, x_i; y_1] - f[x_0, \dots, x_i; y_0]}{y_1 - y_0} = f[x_0, \dots, x_i; y_0, y_1].$$

Analogno određujemo i ostale koeficijente a_{ij}, i, kada ih uvrstimo u (82), dobijamo Newtonov interpolacioni polinom (23) za funkciju dve promenljive

$$\begin{aligned} L_n(x, y) = \sum_{k=0}^{n} \sum_{i+j=k} {} & (x - x_0) \cdots (x - x_{i-1})(y - y_0) \cdots \\ & (y - y_{j-1})\, f[x_0, \dots, x_i; y_0, \dots, y_j]. \end{aligned} \tag{84}$$

Kada su tačke ravnomerno raspoređene, tj. kada je $x_i - x_{i-1} = h$ i $y_j - y_{j-1} = k$, za svako i i j, umesto podeljenih razlika koristimo konačne razlike, koje se za funkciju dva argumenta definišu na sledeći način

$$\begin{aligned} \Delta^{1+0} f(x_i, y_j) &= f(x_{i+1}, y_j) - f(x_i, y_j) \\ \Delta^{0+1} f(x_i, y_j) &= f(x_i, y_{j+1}) - f(x_i, y_j) \\ \Delta^{2+0} f(x_i, y_j) &= \Delta^{1+0} f(x_{i+1}, y_j) - \Delta^{1+0} f(x_i, y_j) \\ \Delta^{1+1} f(x_i, y_j) &= \Delta^{0+1} f(x_{i+1}, y_j) - \Delta^{0+1} f(x_i, y_j) \\ &= \Delta^{1+0} f(x_i, y_{j+1}) - \Delta^{1+0} f(x_i, y_j) \\ \Delta^{0+2} f(x_i, y_j) &= \Delta^{0+1} f(x_i, y_{j+1}) - \Delta^{0+1} f(x_i, y_j) \\ &\vdots \end{aligned}$$

Veza između podeljenih i konačnih razlika funkcije dve promenljive je

$$f[x_0,x_1;y_0]=\frac{1}{h}\Delta^{1+0}f(x_0,y_0),\qquad f[x_0;y_0,y_1]=\frac{1}{k}\Delta^{0+1}f(x_0,y_0),$$
$$f[x_0,x_1,x_2;y_0]=\frac{1}{2!h^2}\Delta^{2+0}f(x_0,y_0),\qquad f[x_0;y_0,y_1,y_2]=\frac{1}{2!k^2}\Delta^{0+2}f(x_0,y_0),$$
$$f[x_0,x_1;y_0,y_1]=\frac{1}{hk}\Delta^{1+1}f(x_0,y_0),$$
$$\vdots$$

Ako se u polinomu (84) uzmu u obzir ove veze i uvedu smene $p=\frac{x-x_0}{h}$ i $q=\frac{y-y_0}{k}$, gde je sa (x_0,y_0) označen čvor najbliži tački (x,y), dobija se Newtonova formula za interpolaciju unapred za funkciju dve promenljive

$$\begin{aligned}L_n(x_0+ph,y_0+qk)&=f(x_0,y_0)+p\Delta^{1+0}f(x_0,y_0)+q\Delta^{0+1}f(x_0,y_0)\\&+\frac{1}{2!}\big(p(p-1)\Delta^{2+0}f(x_0,y_0)+2pq\Delta^{1+1}f(x_0,y_0)+q(q-1)\Delta^{0+2}f(x_0,y_0)\big)\\&+\frac{1}{3!}\big(p(p-1)(p-2)\Delta^{3+0}f(x_0,y_0)+3p(p-1)q\Delta^{2+1}f(x_0,y_0)\\&+3pq(q-1)\Delta^{1+2}f(x_0,y_0)+q(q-1)(q-2)\Delta^{0+3}f(x_0,y_0)\big)+\cdots.\end{aligned}$$

Polinom (84) se može uopštiti i za funkciju više promenljivih.

2.8 Numeričko diferenciranje

Numeričko diferenciranje se koristi kada je funkciju teško diferencirati analitički, ili čak nemoguće – na primer, kada je ona zadata tabelarno. Ono je takođe neophodno pri rešavanju diferencijalnih jednačina metodama konačnih razlika.

Prostije formule numeričkog diferenciranja se dobijaju diferenciranjem interpolacionih polinoma, tj. uzima se da je

$$f^{(k)}(x)\approx L_n^{(k)}(x),$$

gde je $L_n(x)$ interpolacioni polinom određen čvorovima $x_0,\dots,x_n$. Te formule imaju jednostavniji oblik ako se koriste za izračunavanje približne vrednosti izvoda funkcije u čvoru. Na primer, diferenciranjem Newtonovog interpolacionog polinoma za interpolaciju unapred (32), dobijamo da je

$$\begin{aligned}f'(x_0)&\approx\left(\frac{d}{dx}L_n(x)\right)_{x=x_0}=\left(\frac{d}{dq}L_n(x_0+qh)\,\frac{dq}{dx}\right)_{q=0}\\&=\left(\frac{d}{dq}\big(f_0+\sum_{j=1}^{n}\frac{1}{j!}q(q-1)\cdots(q-j+1)\Delta^j f_0\big)\frac{1}{h}\right)_{q=0}\\&=\frac{1}{h}\sum_{j=1}^{n}\frac{(-1)^{j-1}(j-1)!}{j!}\Delta^j f_0=\frac{1}{h}\sum_{j=1}^{n}\frac{(-1)^{j-1}}{j}\Delta^j f_0.\end{aligned}\tag{85}$$

Slično, polazeći od Newtonovog interpolacionog polinoma za interpolaciju unazad dobija se da je

$$f'(x_0) \approx L_n'(x_0) = \frac{1}{h} \sum_{j=1}^{n} \frac{\Delta^j f_{-j}}{j}.$$

PRIMER 6. Odredimo koordinate tačke ekstrema funkcije zadate tabelom

x_i	$f(x_i)$	Δf_i	$\Delta^2 f_i$	$\Delta^3 f_i$
3.8	197.101			
		2.408		
4.0	199.509		-1.992	
		0.416		-0.006
4.2	199.925		-1.998	
		-1.582		
4.4	198.343			

Na osnovu promene znaka konačne razlike prvoga reda, uočavamo da se ekstrem nalazi u okolini čvora $x = 4.2$, i verovatnije u intervalu $(4.0, 4.2)$, jer je u tom intervalu Δf manje po apsolutnoj vrednosti nego u intervalu $(4.2, 4.4)$. Stoga aproksimiramo izvod date funkcije izvodom njenog Besselovog interpolacionog polinoma, i tražimo njegovu nulu (jer je u tački ekstrema prvi izvod funkcije jednak nuli),

$$0.416 - \frac{1.992 + 1.998}{2} \frac{2q-1}{2} - 0.006 \frac{3q^2 - 3q + 0.5}{6} = 0.$$

Koren ove kvadratne jednačine koji je po modulu manji od jedan je $q_e = 0.709$, te je apscisa tačke ekstrema približno $x_e = x_0 + q_e h = 4.142$. Već pomenutim Besselovim polinomom se sada dobija ordinata tačke ekstrema $f(4.142) \approx 199.932$.

Greška formule za numeričko diferenciranje, koja se dobija diferenciranjem interpolacionog polinoma, jednaka je odgovarajućem izvodu greške interpolacije (20),

$$\begin{aligned} f^{(k)}(x) - L_n^{(k)}(x) &= \big(f(x) - L_n(x)\big)^{(k)} = \big(f[x, x_0, \ldots, x_n]\omega_{n+1}(x)\big)^{(k)} \\ &= \sum_{j=0}^{k} C_k^j \big(f[x, x_0, \ldots, x_n]\big)^{(j)} \omega_{n+1}^{(k-j)}(x), \qquad C_k^j = \binom{k}{j}. \end{aligned} \tag{86}$$

Na osnovu formule (44), tretirajući $f[x, x_0, \ldots, x_n]$ kao funkciju po x sa parametrima $x_0, \ldots, x_n$, je

$$\big(f[x, x_0, \ldots, x_n]\big)^{(j)} = j!\, f[\underbrace{x, \ldots, x}_{j+1 \text{ puta}}, x_0, \ldots, x_n]. \tag{87}$$

Uvrstimo (87) u (86) i, uzimajući u obzir vezu podeljenih razlika i izvoda funkcije (55), dobijamo da je

$$f^{(k)}(x) - L_n^{(k)}(x) = \sum_{j=0}^{k} \frac{k!}{(k-j)!} f[\underbrace{x, \ldots, x}_{j+1 \text{ puta}}, x_0, \ldots, x_n]\, \omega_{n+1}^{(k-j)}(x)$$
$$= \sum_{j=0}^{k} \frac{k!}{(k-j)!} \frac{f^{(n+j+1)}(\xi)}{(n+j+1)!} \omega_{n+1}^{(k-j)}(x),$$

odakle sledi ocena

$$|f^{(k)}(x) - L_n^{(k)}(x)| \leq \sum_{j=0}^{k} \frac{k!}{(k-j)!(n+j+1)!} \max_{[y_1,y_2]} |f^{(n+j+1)}(\xi)|\, |\omega_{n+1}^{(k-j)}(x)|, \tag{88}$$

$y_1 = \min(x, x_0, \ldots, x_n)$ i $y_2 = \max(x, x_0, \ldots, x_n)$.

Pri ravnomernom rasporedu čvorova je $\omega_{n+1}^{(j)}(x) = O(h^{n+1-j})$, gde je $h = x_k - x_{k-1}$ za svako k. Stoga je, na osnovu (88),

$$f^{(k)}(x) - L_n^{(k)}(x) = O(h^{n+1-k}), \tag{89}$$

što znači da svako diferenciranje smanjuje red tačnosti za jedan. Na primer, greška formule (85) je

$$f'(x_0) - L_n'(x_0) = \frac{f^{(n+1)}(\xi)}{n+1}(-1)^n h^n. \tag{90}$$

Ocena (89) ukazuje da se greška formule za numeričko diferenciranje smanjuje sa smanjivanjem h. Međutim, smanjivanje koraka povećava uticaj računske greške, što uslovljava rast ukupne greške.

Ilustrujmo ovo na primeru formule (85). Za $n = 1$ aproksimacija prvog izvoda funkcije u čvoru se računa izrazom

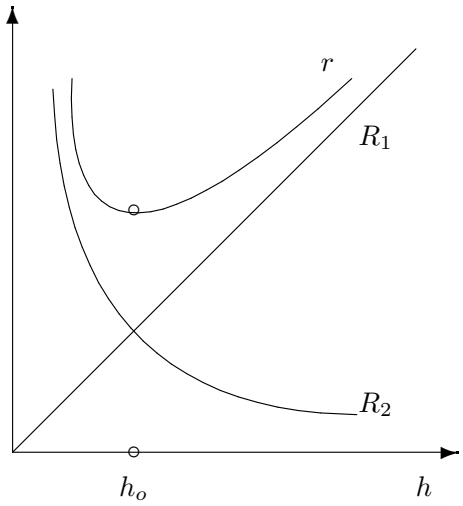
$$f'(x_0) \approx \frac{f(x_1) - f(x_0)}{h}, \tag{91}$$

a greška metode, prema (90), je

$$|R_1| = \left|\frac{f''(\xi)h}{2}\right| \leq \frac{M_2 h}{2},$$

gde je $|f''(\xi)| \leq M_2$. Ako su vrednosti funkcije $f(x_i)$ izračunate sa greškama ϵ_i, pri čemu je $\epsilon_i \leq E$, onda je računska greška izraza (91)

$$|R_2| = \frac{\epsilon_1 + \epsilon_2}{h} \leq \frac{2E}{h}.$$



Slika 2.1: Zavisnost greške numeričkog diferenciranja od koraka.

Ukupna greška aproksimacije izvoda izrazom (91) je

$$|R| = |R_1 + R_2| \leq |R_1| + |R_2| \leq \frac{M_2 h}{2} + \frac{2E}{h} \equiv r(h).$$

Funkcija $r(h)$ (slika 1.) sastoji se od linearnog i hiperboličkog dela, tako da sa smanjivanjem h opada prvi sabirak (greška metode R_1), ali raste drugi sabirak (računska greška R_2). Funkcija $r(h)$ ima minimum u tački

$$h_o = 2\sqrt{\frac{E}{M_2}},$$

i to je optimalna vrednost koraka h za koju je ukupna greška najmanja,

$$|R| \leq r(h_o) = 2\sqrt{M_2 E}. \tag{92}$$

Smanjivanjem koraka ispod optimalne vrednosti ukupna greška se povećava, jer raste računska greška.

Ocena (92) pokazuje da se u najboljem slučaju formulom (85) može izračunati vrednost $f'(x_0)$ sa dvostruko manje sigurnih decimalnih cifara nego što ih je dato za $f(x_i)$, jer je $R = O(E^{1/2})$ pri optimalnom izboru koraka h.

3

Numerička integracija

Često nije moguće integral izraziti pomoću elementarnih funkcija, a i kada je moguće, obično je još dosta truda potrebno uložiti da bi se dobio konačan rezultat u obliku broja. U takvim slučajevima se pribegava numeričkoj integraciji. Formule za numeričku integraciju, ili tzv. *kvadraturne formule*, se obično dobijaju zamenom podintegralne funkcije nekom njoj bliskom funkcijom čiji se integral relativno lako može izračunati.

Ilustrujmo to na primeru određenog integrala dovoljno opšteg oblika

$$I(f) = \int_a^b p(x)f(x)\,dx, \tag{1}$$

gde je $f(x)$ neprekidna funkcija na odsečku $[a, b]$, a tzv. *težinska funkcija* $p(x) > 0$ neprekidna na otvorenom intervalu (a, b). Funkcija $f(x)$ se obično aproksimira generalisanim interpolacionim polinomom,

$$f(x) = \sum_{i=0}^{n} f(x_i)\phi_i(x) + r(x), \tag{2}$$

gde su $\phi_k(x)$, $k = 0, \ldots, n$, date linearno nezavisne funkcije, a $r(x)$ greška interpolacije. Zamenom (2) u (1) dobijamo kvadraturnu formulu

$$S_n(f) = \sum_{i=0}^{n} c_i f(x_i), \qquad c_i = \int_a^b p(x)\phi_i(x)\,dx \tag{3}$$

u kojoj su x_i čvorovi, a c_i težinski koeficijenti. Očigledno je da čvorovi i koeficijenti ne zavise od funkcije $f(x)$. Greška kvadraturne formule je, s obzirom na (2),

$$R_n(f) = I(f) - S_n(f) = \int_a^b p(x)r(x)\,dx. \tag{4}$$

Najčešće korišćene kvadraturne formule su izvedene tako što je podintegralna funkcija $f(x)$ aproksimirana algebarskim interpolacionim polinomom, tj. u aproksimaciji (2) je $\phi_k(x) = x^k$, $k = 0, \ldots, n$.

3.1 Newton–Cotesove kvadraturne formule

U integralu (1) funkciju $f(x)$ aproksimirajmo Lagrangeovim interpolacionim polinomom $L_n(x)$ stepena n, te stoga zadajemo čvorove $x_k \in [a,b]$, $k = 0,\dots,n$. Da bismo dobili izraze za koeficijente c_k kvadraturne formule (3) koji ne zavise od intervala integracije, preslikajmo odsečak $[a,b]$ na odsečak $[-1,1]$

$$x = \frac{b+a}{2} + \frac{b-a}{2}t. \tag{5}$$

Označimo sa $t_k \in [-1,1]$, $k = 0,\dots,n$ slike čvorova x_k određene preslikavanjem (5). Lagrangeov interpolacioni polinom je invarijantan u odnosu na linearnu smenu promenljive (5), tj.

$$L_n(x) = \sum_{i=0}^{n}\Big(\prod_{\substack{j=0\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}\Big)f(x_i) = \sum_{i=0}^{n}\Big(\prod_{\substack{j=0\\ j\neq i}}^{n}\frac{t-t_j}{t_i-t_j}\Big)f(x_i). \tag{6}$$

Dakle, aproksimiranjem funkcije $f(x)$ njenim interpolacionim polinomom (6) u integralu (1) i korišćenjem smene (5), dobijamo

$$I(f) \approx \int_a^b p(x)L_n(x)\,dx = \frac{b-a}{2}\int_{-1}^{1}\bar{p}(t)L_n\big(\tfrac{b+a}{2}+\tfrac{b-a}{2}t\big)\,dt, \tag{7}$$

gde je $\bar{p}(t) \equiv p\big(\frac{b+a}{2}+\frac{b-a}{2}t\big) \equiv p(x)$. Zamenom (6) u (7), dobijamo opšti oblik Newton–Cotesovih kvadraturnih formula

$$S_n(f) = \frac{b-a}{2}\sum_{i=0}^{n}c_i f(x_i), \qquad c_i = \int_{-1}^{1}\bar{p}(t)\Big(\prod_{\substack{j=0\\ j\neq i}}^{n}\frac{t-t_j}{t_i-t_j}\Big)\,dt. \tag{8}$$

Očigledno je da koeficijenti c_k formule (8) ne zavise od intervala integracije $[a,b]$ i podintegralne funkcije $f(x)$, već samo od izbora čvorova i težinske funkcije $p(x)$.

Prema (4) i (2.9), greška kvadraturne formule (8) je

$$R_n(f) = \int_a^b p(x)\frac{f^{(n+1)}(\xi(x))}{(n+1)!}\omega_{n+1}(x)\,dx, \qquad \xi \in [a,b], \tag{9}$$

i, posle smene (5) u integralu (9), može da se oceni izrazom

$$|R_n(f)| \le \frac{1}{(n+1)!}\Big(\frac{b-a}{2}\Big)^{n+2}\max_{x\in[a,b]}|f^{(n+1)}(x)|\int_{-1}^{1}|\bar{p}(t)\,\bar{\omega}_{n+1}(t)|\,dt, \tag{10}$$

jer je

$$\omega_{n+1}(x) = \prod_{i=0}^{n}(x-x_i) = \Big(\frac{b-a}{2}\Big)^{n+1}\prod_{i=0}^{n}(t-t_i) \equiv \Big(\frac{b-a}{2}\Big)^{n+1}\bar{\omega}_{n+1}(t). \tag{11}$$

Jedna osobina koeficijenata kvadraturne formule, koja ima uticaja na grešku, data je sledećom lemom:

LEMA 1. *Ako je težinska funkcija $p(x)$ parna u odnosu na sredinu odsečka $[a,b]$, a čvorovi x_k su simetrično raspoređeni u odnosu na sredinu odsečka, tj. $t_k = -t_{n-k}$, onda su koeficijenti kvadraturne formule (8), koji odgovaraju simetričnim čvorovima, jednaki, tj. $c_k = c_{n-k}$.*

DOKAZ: Koeficijent c_k dat u (8) je, s obzirom na (11),

$$c_k = \int_{-1}^{1} \bar{p}(t) \frac{\bar{\omega}_{n+1}(t)}{(t-t_k)\bar{\omega}'_{n+1}(t_k)}\, dt,$$

ili, uvođenjem u integral smene $t = -\tau$ i korišćenjem pretpostavke o simetričnosti čvorova,

$$c_k = \int_{-1}^{1} \bar{p}(-\tau) \frac{\bar{\omega}_{n+1}(-\tau)}{-(\tau - t_{n-k})\bar{\omega}'_{n+1}(-t_{n-k})}\, d\tau. \tag{12}$$

Kako su čvorovi simetrično raspoređeni, polinom $\bar{\omega}_{n+1}(t)$ je oblika

$$\bar{\omega}_{n+1}(t) = \begin{cases} \prod_{i=0}^{\frac{n-1}{2}} (t^2 - t_i^2), & \text{ako je } n \text{ neparno} \\ t \prod_{i=0}^{\frac{n}{2}-1} (t^2 - t_i^2), & \text{ako je } n \text{ parno.} \end{cases}$$

Dakle, kada je n neparno, tj. broj čvorova paran, $\bar{\omega}_{n+1}(t)$ je parna funkcija, a njen izvod neparna; kada je n parno, tj. broj čvorova neparan, ova funkcija je neparna, a njen izvod je parna funkcija. Stoga je za svako n

$$\frac{\bar{\omega}_{n+1}(-\tau)}{\bar{\omega}'_{n+1}(-t_{n-k})} = -\frac{\bar{\omega}_{n+1}(\tau)}{\bar{\omega}'_{n+1}(t_{n-k})},$$

te, uzimajući i pretpostavku o parnosti funkcije $\bar{p}(t)$ u obzir, iz (12) sledi

$$c_k = \int_{-1}^{1} \bar{p}(\tau) \frac{\bar{\omega}_{n+1}(\tau)}{(\tau - t_{n-k})\bar{\omega}'_{n+1}(t_{n-k})}\, d\tau = c_{n-k},$$

što je i trebalo dokazati. ∎

Kvadraturna formula (8) sa simetrično raspoređenim čvorovima u odnosu na sredinu odsečka $[a,b]$ je tačna, ako je $f(x)$ neparna, a $p(x)$ parna funkcija u odnosu na sredinu tog odsečka. Ovo neposredno sledi, jer je na osnovu leme 1 i činjenice da je $f(\frac{a+b}{2}) = 0$

$$S_n(f) = \frac{b-a}{2} \sum_{i=0}^{n} c_i f(x_i) = \frac{b-a}{2} \sum_{i=0}^{[\frac{n}{2}]} (c_i - c_{n-i}) f(x_i) = 0,$$

a integral funkcije neparne u odnosu na sredinu intervala integracije je takođe nula. Formula (8) je, stoga, tačna za proizvoljan polinom $P_{n+1}(x)$ stepena $n+1$, ako je n parno, jer se ovaj može predstaviti u obliku

$$P_{n+1}(x) \equiv P_n(x) + c\big(x - \frac{a+b}{2}\big)^{n+1},$$

gde je $P_n(x)$ proizvoljan polinom stepena n, a c proizvoljna konstanta. Oba sabirka se tačno integrale kvadraturnom formulom – prvi, jer je polinom stepena najviše n, a drugi, jer je neparna funkcija u odnosu na sredinu intervala $[a, b]$.

Dakle, u opštem slučaju kvadraturna formula (8) je tačna za sve polinome najviše stepena n; ako su čvorovi simetrično raspoređeni i funkcija $p(x)$ parna u odnosu na sredinu intervala integracije, formula (8) je tačna i za polinome stepena $n+1$, ako je n parno. U ovom specijalnom slučaju središnji čvor $x_{\frac{n}{2}} = \frac{b+a}{2}$ možemo tretirati kao dvostruki čvor interpolacije, te se greška kvadraturne formule ocenjuje izrazom (10) u kome je n zamenjeno sa $n+1$, a

$$\bar{\omega}_{n+2}(t) = t^2 \prod_{i=0}^{\frac{n}{2}-1} (t^2 - t_i^2), \qquad \left(\prod_{i=0}^{\frac{n}{2}-1} (t^2 - t_i^2) \right)_{n=0} \stackrel{\text{def}}{=} 1. \tag{13}$$

U daljem tekstu će biti navedene neke poznate Newton–Cotesove formule ovoga tipa, kod kojih je $p(x) \equiv 1$.

Formula pravougaonika. Dobija se iz formule (8) za $n = 0$ i $t_0 = 0$. Koeficijent je

$$c_0 = \int_{-1}^{1} dt = 2,$$

te je

$$S_0(f) = (b-a) f(\tfrac{a+b}{2}). \tag{14}$$

$S_0(f)$ predstavlja površinu pravougaonika čija je jedna stranica jednaka dužini intervala integracije, a druga vrednosti funkcije $f(x)$ u središnjoj tački intervala, te otuda naziv formule. Kako ova formula spada u formule sa simetrično raspoređenim čvorovima kojih ima neparan broj (jedan), to se, s obzirom na (13), greška ocenjuje izrazom

$$|R_0(f)| \le \frac{1}{2!} \big(\frac{b-a}{2}\big)^3 \max_{x\in[a,b]} |f''(x)| \int_{-1}^{1} t^2 \, dt = \max_{x\in[a,b]} |f''(x)| \frac{(b-a)^3}{24}. \tag{15}$$

Formula trapeza. Za $n = 1$, $t_0 = -1$ i $t_1 = 1$ formula (8) je

$$S_1(f) = \frac{b-a}{2} \big(f(a) + f(b)\big), \tag{16}$$

što predstavlja površinu trapeza čije su osnovice $f(a)$ i $f(b)$, a visina $b - a$. Na osnovu (10), greška se ocenjuje izrazom

$$|R_1(f)| \le \frac{1}{2!} \big(\frac{b-a}{2}\big)^3 \max_{x\in[a,b]} |f''(x)| \int_{-1}^{1} |t^2 - 1| \, dt = \max_{x\in[a,b]} |f''(x)| \frac{(b-a)^3}{12}. \tag{17}$$

Poredeći ocene (15) i (17), vidimo da se formulom pravougaonika postiže čak veća tačnost, iako je ona dobijena aproksimiranjem podintegralne funkcije polinomom nižeg stepena (nula) u odnosu na stepen polinoma korišćenog u trapeznoj formuli (jedan).

Formula Simpsona. Dobija se iz (8) za $n = 2$, $t_0 = -1$, $t_1 = 0$ i $t_2 = 1$,

$$S_2(f) = \frac{b-a}{2}\,\frac{1}{3}\big(f(a) + 4f(\tfrac{a+b}{2}) + f(b)\big). \tag{18}$$

Ocena greške formule (18), s obzirom da je n parno, je

$$|R_2(f)| \le \frac{1}{4!}\big(\frac{b-a}{2}\big)^5 \max_{x\in[a,b]} |f^{(4)}(x)| \int_{-1}^{1} t^2|t^2-1|\,dt = \max_{x\in[a,b]} |f^{(4)}(x)| \frac{1}{90}\big(\frac{b-a}{2}\big)^5. \tag{19}$$

Kvadraturne formule mogu biti *formule zatvorenog tipa* ili *formule otvorenog tipa*, u zavisnosti od toga da li krajnje tačke intervala integracije jesu ili nisu čvorovi kvadraturne formule. Od navedenih, formula pravougaonika je formula otvorenog tipa, a formula trapeza i formula Simpsona su formule zatvorenog tipa.

Pomenute formule su tzv. *osnovne kvadraturne formule*. Iz ocena (15), (17) i (19) je očigledno da se tačnost može povećati smanjivanjem dužine intervala (a, b) na kome se primenjuje osnovna formula, ili korišćenjem formule dobijene aproksimiranjem funkcije interpolacionim polinomom višeg stepena n. Drugi način je manje pogodan, jer su formule sve složenije što je n veće. Stoga se obično tačnost povećava smanjivanjem dužine intervala na kome se primenjuje osnovna formula. Interval integracije (a, b) se podeli na izvestan broj podintervala jednake dužine h, i na svakom od njih se primeni osnovna formula. Kako je integral po intervalu (a, b) jednak sumi integrala po podintervalima dužine h, to se on aproksimira sumom osnovnih formula primenjenih na pomenute podintervale. Tako se dobijaju *opšte kvadraturne formule*.

Kada je interval (a, b) podeljen na m podintervala dužine h, opšta formula pravougaonika je, s obzirom na (14),

$$\int_a^b f(x)\,dx \approx h\sum_{i=1}^{m} f_{i-\frac{1}{2}} \equiv S_0^h(f),$$

gde je korišćena oznaka $f_p = f(x_p)$, $x_p = a + ph$, $0 \le p \le m$. Ova formula nastaje kao rezultat aproksimiranja funkcije $f(x)$ step funkcijom. Ocena greške formule se dobija sumiranjem ocena grešaka osnovnih formula (15)

$$\begin{aligned} |R_0^h(f)| &\le \sum_{i=1}^{m} \frac{h^3}{24} \max_{[x_{i-1},x_i]} |f''(\xi_i)| = \frac{mh^3}{24}\left(\frac{1}{m}\sum_{i=1}^{m} \max_{[x_{i-1},x_i]} |f''(\xi_i)|\right) \\ &\le \frac{mh^3}{24} \max_{[a,b]} |f''(\xi)| = \frac{(b-a)h^2}{24} \max_{[a,b]} |f''(\xi)|. \end{aligned} \tag{20}$$

Slično, opšta formula trapeza, dobijena sumiranjem izraza oblika (16), je

$$\begin{aligned} \int_a^b f(x)\,dx &\approx \frac{h}{2}\sum_{i=1}^{m}(f_{i-1} + f_i) \\ &= \frac{h}{2}(f_0 + 2\sum_{i=1}^{m-1} f_i + f_m) \equiv S_1^h(f), \end{aligned} \tag{21}$$

i rezultat je aproksimacije funkcije $f(x)$ deo po deo linearnom funkcijom. Ocena greške ove formule, dobijena na isti način kao i ocena (20), je

$$|R_1^h(f)| \leq \frac{(b-a)h^2}{12} \max_{[a,b]} |f''(\xi)|.$$

Kod uopštavanja formule Simpsona (18), interval (a,b) se mora podeliti na paran broj podintervala dužine h, $h = \frac{b-a}{2m}$, jer se osnovna formula definiše nad dva podintervala. Sumiranjem osnovnih formula dobija se

$$\begin{aligned} \int_a^b f(x)\,dx &= \sum_{i=1}^{m} \int_{x_{2(i-1)}}^{x_{2i}} f(x)\,dx \approx \frac{h}{3} \sum_{i=1}^{m} (f_{2(i-1)} + 4f_{2i-1} + f_{2i}) \\ &= \frac{h}{3}(f_0 + 4\sum_{i=1}^{m} f_{2i-1} + 2\sum_{i=1}^{m-1} f_{2i} + f_{2m}) \equiv S_2^h(f). \end{aligned} \tag{22}$$

Ocena greške opšte formule je, s obzirom na (19),

$$\begin{aligned} |R_2^h(f)| &\leq \sum_{i=1}^{m} \frac{h^5}{90} \max_{[x_{2i-2},x_{2i}]} |f^{(4)}(\xi_i)| = \frac{h^4}{90}\,\frac{b-a}{2m} \sum_{i=1}^{m} \max_{[x_{2i-2},x_{2i}]} |f^{(4)}(\xi_i)| \\ &\leq \frac{(b-a)h^4}{180} \max_{[a,b]} |f^{(4)}(\xi)|. \end{aligned}$$

Navedene ocene grešaka su dosta nepraktične, jer treba oceniti maksimum odgovarajućeg izvoda funkcije $f(x)$, ako je ona uopšte i zadata analitički. Stoga se često koristi tzv. *Rungeova ocena greške*. Pretpostavljajući da se odgovarajući izvod funkcije, koji figuriše u oceni greške, ne menja mnogo na intervalu (a,b), može se između tačne vrednosti integrala i vrednosti izračunatih kvadraturnom formulom sa korakom h i korakom $2h$, napisati sledeća veza

$$I(f) \approx S^h(f) + Mh^k \approx S^{2h}(f) + M(2h)^k,$$

iz koje je $M \approx \frac{S^h(f) - S^{2h}(f)}{h^k(2^k-1)}$. Stoga se greška vrednosti $S^h(f)$ ocenjuje izrazom

$$I(f) - S^h(f) \approx \frac{S^h(f) - S^{2h}(f)}{2^k - 1},$$

a popravljena približna vrednost integrala je

$$I(f) \approx S^h(f) + \frac{S^h(f) - S^{2h}(f)}{2^k - 1}.$$

Očigledna dobra osobina ovog algoritma je mogućnost realizacije na računaru.

PRIMER 1. Izračunajmo Simpsonovom kvadraturnom formulom sa tačnošću 10^{-3} integral

$$\int_0^{\pi} \frac{1}{x+\cos x}\,dx.$$

Tačnost ćemo oceniti Rungeovim kriterijumom. Integral prvo izračunamo Simpsonovom formulom sa najvećim mogućim korakom, $h = \frac{\pi}{2}$, a zatim smanjujemo korak polovljenjem, sve dok ne postignemo željenu tačnost. Tako dobijamo vrednosti

$$S^{h_1}(f) = 2.1014, \qquad S^{h_2}(f) = 2.0540, \qquad S^{h_3}(f) = 2.0441,$$

gde je $h_i = \pi/2^i$, $i = 1, 2, 3$. S obzirom da je $R = \frac{1}{15}(S^{h_3} - S^{h_2}) = -7 \cdot 10^{-4}$, tačnost je postignuta, te je

$$I(f) \approx S^{h_3} + R = 2.043.$$

Opšta Simpsonova formula (22) se može dobiti i kombinovanjem trapeznih formula (21) sa korakom h i korakom $2h$

$$\begin{aligned} S_2^h(f) &= \frac{h}{3}\big(f_0 + 4(f_1 + f_3 + \cdots + f_{2m-1}) + 2(f_2 + f_4 + \cdots + f_{2m-2}) + f_{2m}\big) \\ &= \frac{4}{3}h(\frac{f_0}{2} + f_1 + \cdots + f_{2m-1} + \frac{f_{2m}}{2}) \\ &\qquad - \frac{1}{3}\,2h(\frac{f_0}{2} + f_2 + f_4 + \cdots + f_{2m-2} + \frac{f_{2m}}{2}) \\ &= \frac{4}{3}S_1^h(f) - \frac{1}{3}S_1^{2h}(f). \end{aligned}$$

Ovo je posledica tzv. *Euler–Maclaurinove sumacione formule*([25])

$$\begin{aligned} \int_{x_0}^{x_m} f(x)\,dx &= h(\frac{1}{2}f_0 + f_1 + \cdots + f_{m-1} + \frac{1}{2}f_m) \\ &\quad - \big(B_2\frac{h^2}{2!}(f_m' - f_0') + \cdots + B_{2k}\frac{h^{2k}}{(2k)!}(f_m^{(2k-1)} - f_0^{(2k-1)}) + \cdots\big), \end{aligned} \tag{23}$$

gde su B_{2k} Bernoullievi brojevi definisani razvojem

$$\frac{x}{e^x - 1} = \sum_{i=0}^{\infty} B_i \frac{x^i}{i!}.$$

Formulom (23) je dat asimptotski razvoj greške trapezne kvadraturne formule, koji sadrži samo parne stepene h,

$$S_1^h(f) - \int_{x_0}^{x_m} f(x)\,dx = \sum_{i=1}^{\infty} a_i h^{2i}, \qquad a_i = \frac{B_{2i}}{(2i)!}(f_m^{(2i-1)} - f_0^{(2i-1)}). \tag{24}$$

Simpsonova formula je dobijena takvom kombinacijom trapeznih formula sa koracima h i $2h$, kojom se anulira koeficijent a_1 u (24). Slično se mogu, formiranjem odgovarajućih kombinacija trapeznih formula napisanih za različito h, u razvoju (24) anulirati i drugi koeficijenti i tako dobiti kvadraturne formule višeg reda tačnosti.

Ova ideja se može realizovati i na drugi način. $S_1^h(f)$ se može aproksimirati polinomom po parnim stepenima h, koji predstavlja parcijalnu sumu reda (24),

$$S_1^h(f) = \sum_{i=0}^{\infty} a_i h^{2i} \approx \sum_{i=0}^{k} a_i h^{2i} \equiv T_k(h),$$

gde je $a_0 = \int_{x_0}^{x_m} f(x)\,dx$. Polinom $T_k(h)$ odredimo tako da bude interpolacioni polinom stepena k po h^2 izraza $S_1^h(f)$ kao funkcije od h. Izračunamo trapeznom formulom (21) $S_1^h(f)$ za niz vrednosti koraka h,

$$h_0 = x_m - x_0, \quad h_1 = \frac{h_0}{n_1}, \quad h_2 = \frac{h_0}{n_2}, \quad \ldots, \quad h_k = \frac{h_0}{n_k}, \tag{25}$$

gde je n_j, $j = 1, \ldots, k$, strogo rastući niz pozitivnih celih brojeva, pa je tabelom $(h_j, S_1^{h_j}(f))$, $j = 0, \ldots, k$, zadat interpolacioni polinom $T_k(h)$. Ekstrapolacijom, na primer Nevilleovim algoritmom (§2.1), odredimo graničnu vrednost trapezne formule za $h = 0$, $S_1^0(f) \approx T_k(0)$.

Ako je u (25) $n_j = 2^j$, tj. $h_j = (x_m - x_0)/2^j$, metoda se naziva *Rombergova metoda*. S obzirom da je svaki korak dvostruko manji od prethodnog, računanje trapeznom formulom (21) se može ubrzati korišćenjem rezultata dobijenog sa prethodnim korakom, jer je

$$S_1^h(f) = \frac{1}{2} S_1^{2h}(f) + h(f_1 + f_3 + \cdots + f_{m-1}).$$

Rombergova metoda je primer vrlo opšte ideje u numeričkoj matematici, poznate pod nazivom Richardsonova ekstrapolacija – neki numerički algoritam se realizuje za razne vrednosti koraka h, i zatim se ekstrapolacijom za $h = 0$ odredi aproksimacija tražene vrednosti povećane tačnosti.

3.2 Kvadraturne formule Gaussovog tipa

U Newton–Cotesovim formulama čvorovi su zadati, a koeficijenti se određuju tako da formula bude tačna za polinome što je moguće višeg stepena. Formule pravougaonika (14) i Simpsona (18) ukazuju da se ravnomernim rasporedom neparnog broja čvorova može povećati tačnost Newton–Cotesove kvadraturne formule. Kvadraturne formule

$$S_n(f) = \frac{b-a}{2} \sum_{i=1}^{n} c_i f(x_i) \tag{26}$$

u kojima se, pored koeficijenata c_i, i čvorovi x_i biraju tako da formula bude tačna za polinome što je moguće višeg stepena, nazivaju se kvadraturne formule Gaussovog tipa. Obzirom da u formuli (26) ima $2n$ slobodnih parametara, maksimalni stepen polinoma za koga ona važi tačno je $2n-1$. Dakle, zahtevom da jednakost

$$I(P_m) \equiv \int_a^b p(x)P_m(x)\,dx = \frac{b-a}{2}\sum_{i=1}^{n} c_i P_m(x_i) \equiv S_n(P_m) \tag{27}$$

važi za proizvoljan polinom $P_m(x)$ stepena $m \leq 2n-1$, određena je kvadraturna formula (26). Uslov (27) se može zameniti sistemom od $2n$ jednačina

$$\frac{b-a}{2}\sum_{i=1}^{n} c_i x_i^k = \int_a^b p(x)x^k\,dx, \qquad k = 0,\ldots,2n-1, \tag{28}$$

što je posledica sledeće leme.

LEMA 2. *Da bi kvadraturna formula (26) bila tačna za proizvoljan polinom stepena m,*

$$P_m(x) = \sum_{k=0}^{m} a_k x^k,$$

potrebno je i dovoljno da ona bude tačna za sve funkcije x^k, $k = 0,\ldots,m$.

DOKAZ: Ako je formula (26) tačna za sve funkcije x^k, $k = 0,\ldots,m$, onda je

$$\begin{aligned} S_n(P_m) &= \frac{b-a}{2}\sum_{i=1}^{n} c_i\left(\sum_{k=0}^{m} a_k x_i^k\right) = \sum_{k=0}^{m} a_k\left(\frac{b-a}{2}\sum_{i=1}^{n} c_i x_i^k\right) \\ &= \sum_{k=0}^{m} a_k S_n(x^k) = \sum_{k=0}^{m} a_k I(x^k) = I(P_m). \end{aligned}$$

Sa druge strane, ako je formula (26) tačna za proizvoljan polinom $P_m(x)$ stepena najviše $m = 2n-1$, tj. $S_n(P_m) = I(P_m)$ za svaki izbor koeficijenata a_k, $k = 0,\ldots,m$, onda je

$$\sum_{k=0}^{m} a_k S_n(x^k) = \sum_{k=0}^{m} a_k I(x^k),$$

što je moguće samo ako je $S_n(x^k) = I(x^k)$ za svako $k = 0,\ldots,m$. ∎

Sistem (28) je nelinearan po x_i, te nije pogodan za nalaženje čvorova kvadraturne formule (26). Umesto da rešavamo ovaj sistem, za određivanje čvorova kvadraturne formule (26) koristimo rezultat naredne leme.

LEMA 3. *Ako su x_i, $i = 1, \ldots, n$, čvorovi kvadraturne formule (26), tačne za sve polinome stepena $2n-1$, onda je*

$$\int_a^b p(x)\omega_n(x)P_{n-1}(x)\,dx = 0, \tag{29}$$

gde je

$$\omega_n(x) \equiv \prod_{i=1}^{n}(x - x_i) \equiv x^n + b_1 x^{n-1} + \cdots + b_n, \tag{30}$$

a $P_{n-1}(x)$ je proizvoljan polinom stepena ne većeg od $n-1$.

DOKAZ: Kako je formula (26), po pretpostavci, tačna za proizvoljan polinom stepena do $2n-1$, to je ona tačna i za polinom $\omega_n(x)P_{n-1}(x)$. Stoga je, obzirom da su x_i, $i = 1, \ldots, n$ nule polinoma $\omega_n(x)$,

$$\int_a^b p(x)\omega_n(x)P_{n-1}(x)\,dx = \frac{b-a}{2}\sum_{i=1}^{n} c_i\omega_n(x_i)P_{n-1}(x_i) = 0.$$

∎

Jednakost (29), s obzirom na lemu 2, je ekvivalentna uslovima

$$\int_a^b p(x)\omega_n(x)x^k\,dx = 0, \qquad k = 0, \ldots, n-1, \tag{31}$$

koji određuju sistem linearnih jednačina po koeficijentima b_j polinoma $\omega_n(x)$. Nule ovog polinoma su realne, različite i pripadaju intervalu integracije $[a, b]$, jer je to polinom iz klase ortogonalnih polinoma u odnosu na težinsku funkciju $p(x)$.

Ortogonalni polinomi. Integralom (1) je u linearnom prostoru funkcija $\mathcal{L}_2(a, b)$ definisan skalarni proizvod

$$(f\ ,\ g) = \int_a^b p(x)f(x)g(x)\,dx. \tag{32}$$

Dve funkcije su ortogonalne, ako je njihov skalarni proizvod (32) jednak nuli. Sistem funkcija $f_k(x)$, $k = 0, 1, \ldots$, predstavlja ortogonalni sistem funkcija u odnosu na skalarni proizvod (32), ako je za svako $j \neq k$ $(f_j\ ,\ f_k) = 0$. Ako su funkcije $f_k(x)$ polinomi, tj. $f_k(x) \equiv Q_k(x)$, kaže se da je taj sistem funkcija sistem ortogonalnih polinoma u odnosu na težinsku funkciju $p(x)$.

Dokaz egzistencije i jedinstvenosti sistema normiranih (polinomi sa koeficijentom jedan uz najviši stepen), ortogonalnih polinoma za $p(x) \geq 0$, $x \in [a, b]$, i konstruktivni algoritam dati su sledećom teoremom.

TEOREMA 1. *Postoje normirani polinomi $Q_k(x)$, $k = 0, 1, \ldots$, takvi da je*

$$(Q_i\,,\,Q_j) = 0, \qquad i \neq j.$$

Ovi polinomi su jedinstveno određeni rekurentnom formulom

$$Q_0(x) \equiv 1$$

$$Q_{k+1} = \left(x - \frac{(xQ_k\,,\,Q_k)}{(Q_k\,,\,Q_k)}\right)Q_k(x) - \frac{(Q_k\,,\,Q_k)}{(Q_{k-1}\,,\,Q_{k-1})}Q_{k-1}(x), \quad k = 0, 1, \ldots, \tag{33}$$

pri čemu je drugi sabirak u (33) nula za $k = 0$.

DOKAZ: Polinomi se uzastopno mogu konstruisati Gram–Schmidtovim postupkom ortogonalizacije. Polazeći od polinoma $Q_0(x) = 1$, pretpostavimo da su jedinstveno određeni polinomi navedenih osobina za svako $j \leq k$. Ma koji normirani polinom $Q_{k+1}(x)$ se može jedinstveno predstaviti u obliku

$$Q_{k+1}(x) = (x - a_k)Q_k(x) + a_{k-1}Q_{k-1}(x) + a_{k-2}Q_{k-2}(x) + \cdots + a_0Q_0(x), \tag{34}$$

jer ovaj polinom i svi polinomi $Q_j(x)$, $j \leq k$, imaju koeficijent jedan uz najviši stepen. Koeficijente a_j, $j = 0, \ldots, k$, u (34) odredimo tako da je

$$(Q_{k+1}\,,\,Q_j) = 0, \qquad j = 0, \ldots, k.$$

Množenjem skalarno izraza (34) sa $Q_j(x)$, $j = 0, \ldots, k$, vodeći računa da je $(Q_i\,,\,Q_j) = 0$ za $i \neq j$ i $i, j \leq k$, dobijamo

$$\begin{aligned} (Q_{k+1}\,,\,Q_k) &= \big((x - a_k)Q_k\,,\,Q_k\big) = (xQ_k\,,\,Q_k) - a_k(Q_k\,,\,Q_k) = 0 \\ (Q_{k+1}\,,\,Q_j) &= \big((x - a_k)Q_k\,,\,Q_j\big) + a_j(Q_j\,,\,Q_j) \\ &= (xQ_k\,,\,Q_j) + a_j(Q_j\,,\,Q_j) = 0, \qquad j = 0, \ldots, k-1. \end{aligned} \tag{35}$$

S obzirom da je $(Q_j\,,\,Q_j) \equiv \int_a^b p(x)Q_j^2(x)\,dx = 0$ moguće samo ako je $Q_j(x) \equiv 0$, to jednačine (35) imaju jedinstvena rešenja

$$a_k = \frac{(xQ_k\,,\,Q_k)}{(Q_k\,,\,Q_k)} \tag{36}$$

$$a_j = -\frac{(xQ_k\,,\,Q_j)}{(Q_j\,,\,Q_j)}, \qquad j = 0, \ldots, k-1. \tag{37}$$

Na osnovu indukcijske hipoteze za $j \leq k - 1$ je

$$Q_{j+1}(x) = \left(x - \frac{(xQ_j\,,\,Q_j)}{(Q_j\,,\,Q_j)}\right)Q_j(x) - \frac{(Q_j\,,\,Q_j)}{(Q_{j-1}\,,\,Q_{j-1})}Q_{j-1}(x),$$

pa je

$$xQ_j(x) = Q_{j+1}(x) + \frac{(xQ_j\,,\,Q_j)}{(Q_j\,,\,Q_j)}Q_j(x) + \frac{(Q_j\,,\,Q_j)}{(Q_{j-1}\,,\,Q_{j-1})}Q_{j-1}(x).$$

Skalarnim množenjem sa $Q_k(x)$ poslednje relacije, dobijamo

$$(xQ_k\,,\,Q_j) = (xQ_j\,,\,Q_k) = (Q_{j+1}\,,\,Q_k), \qquad j = 0, \ldots, k-1,$$

što, kada se uvrsti u (37), daje

$$a_j = -\frac{(Q_{j+1}\,,\,Q_k)}{(Q_j\,,\,Q_j)} = \begin{cases} -\frac{(Q_k\,,\,Q_k)}{(Q_{k-1}\,,\,Q_{k-1})}, & \text{za } j = k-1 \\ 0, & \text{za } j < k-1. \end{cases} \tag{38}$$

Zamenom (36) i (38) u (34) dobijamo (33). ∎

Proizvoljan polinom se može predstaviti linearnom kombinacijom ortogonalnih polinoma $P_m(x) = \sum_{i=0}^{m} a_i Q_i(x)$, te je

$$(Q_n\,,\,P_m) \equiv \int_a^b p(x)Q_n(x)P_m(x)\,dx = 0, \qquad m < n. \tag{39}$$

što je saglasno tvrđenju (29) leme 3 za $Q_n(x) \equiv \omega_n(x)$. To znači da je polinom $\omega_n(x)$ definisan u lemi 3 relacijom (30), identičan polinomu $Q_n(x)$ iz sistema normiranih ortogonalnih polinoma u odnosu na težinsku funkciju $p(x)$. Stoga se on, umesto rešavanjem sistema (31), može odrediti rekurentnom formulom (33).

Nas interesuju osobine nula ortogonalnih polinoma, jer su to čvorovi kvadraturne formule (26).

TEOREMA 2. *Koreni x_i, $i = 1, \ldots, n$, normiranog ortogonalnog polinoma $Q_n(x)$ su realni i jednostruki. Svi pripadaju otvorenom intervalu (a, b).*

DOKAZ: Izdvojmo one korene polinoma $Q_n(x)$ iz intervala (a, b), koji su neparne višestrukosti,

$$a < x_1 < \cdots < x_j < b.$$

Polinom $Q_n(x)\,P(x)$, gde je $P(x) = \prod_{i=1}^{j}(x - x_i)$, ne menja znak na (a, b), te je

$$(Q_n\,,\,P) \equiv \int_a^b p(x)Q_n(x)P(x)\,dx \neq 0.$$

To znači da je stepen polinoma $P(x)$ jednak n, jer bi u protivnom, na osnovu (39), moralo biti $(Q_n\,,\,P) = 0$. Time je tvrđenje teoreme dokazano. ∎

Dakle, za svaku neprekidnu težinsku funkciju $p(x) > 0$ na intervalu (a, b), čvorovi kvadraturne formule (26) su nule polinoma stepena n ortogonalnog u odnosu na tu težinsku funkciju. Kada su čvorovi određeni, koeficijenti c_i, $i = 1, \ldots, n$, formule (26) se mogu odrediti kao rešenja prvih n jednačina sistema (28), koji je linearan po ovim koeficijentima.

PRIMER 2. U odnosu na težinsku funkciju $p(x) \equiv 1$ ortogonalni su Legendreovi polinomi

$$L_n(x) = \frac{1}{2^n\, n!} \frac{d^n}{dx^n}\big((x^2-1)^n\big), \qquad x \in [-1,1].$$

Kvadraturna formula Gaussovog tipa napisana za $p(x) \equiv 1$, tj. sa čvorovima koji su nule Legendreovog polinoma odgovarajućeg stepena, u literaturi se obično naziva *Gaussova kvadraturna formula*. Na primer, za $n = 3$ ova formula je na osnovnom intervalu $[-1, 1]$

$$\int_{-1}^{1} f(x)\,dx \approx \frac{1}{9}\left(5f(-\sqrt{\tfrac{3}{5}}) + 8f(0) + 5f(\sqrt{\tfrac{3}{5}})\right).$$

Kod kvadraturnih formula Gaussovog tipa ne može se naći formula oblika (26) koja je tačna i za sve polinome stepena $2n$. Naime, nijedna od ovakvih formula ne računa tačno integral polinoma $P_{2n}(x) = \prod_{i=0}^{n}(x - x_i)^2$, jer je, za $p(x) > 0$,

$$I(P_{2n}) = \int_a^b p(x)(x-x_1)^2 \cdots (x-x_n)^2\,dx > 0,$$
$$S_{2n-1}(P_{2n}) = \frac{b-a}{2}\sum_{i=1}^{n} c_i (x_i - x_1)^2 \cdots (x_i - x_i)^2 \cdots (x_i - x_n)^2 = 0.$$

Greška kvadraturnih formula Gaussovog tipa (26) ocenjuje se izrazom

$$|R_{2n-1}(f)| \le \frac{1}{(2n)!} \max_{[a,b]} |f^{(2n)}(\xi)| \int_a^b p(x)\omega_n^2(x)\,dx, \tag{40}$$

koji se dobija kada se u opštoj formuli za grešku (9) stavi $2n-1$ umesto n, i $\omega_{2n}(x) \equiv \omega_n^2(x)$, jer je svaki od čvorova kvadraturne formule dvostruki čvor interpolacije funkcije $f(x)$ polinomom $(2n-1)$-og stepena.

Prednosti kvadraturnih formula Gaussovog tipa u odnosu na Newton–Cotesove formule su veća tačnost koja se postiže sa istim brojem čvorova, i mogućnost približnog integraljenja funkcija sa singularitetima. Naime, ako se funkcija $g(x)$ loše, a funkcija $\frac{g(x)}{p(x)}$, $p(x) > 0$ dobro aproksimira polinomima, onda se kvadraturna formula Gaussovog tipa umesto na integral $\int_a^b g(x)\,dx$, primenjuje na integral oblika (1), gde je $f(x) \equiv \frac{g(x)}{p(x)}$.

PRIMER 3. U integralu $\int_0^1 \frac{\cos x}{\sqrt{1-x}}\,dx$ gornja granica je singularna tačka podintegralne funkcije. Integral se može približno izračunati kvadraturnom formulom Gaussovog tipa sa težinskom funkcijom $p(x) = 1/\sqrt{1-x}$, koja na primer za $n = 2$ ima oblik

$$\int_0^1 \frac{f(x)}{\sqrt{1-x}}\,dx = c_1 f(x_1) + c_2 f(x_2) + R(f).$$

Prema rekurentnoj vezi (33), u kojoj je $(f\,,\,g)=\int_0^1 \frac{f(x)g(x)}{\sqrt{1-x}}\,dx$, prva tri polinoma sistema normiranih polinoma ortogonalnih u odnosu na težinsku funkciju $p(x)=1/\sqrt{1-x}$ su

$$Q_0(x)=1, \qquad Q_1(x)=x-\frac{2}{3}, \qquad Q_2(x)=x^2-\frac{8}{7}x+\frac{8}{35}.$$

Nule polinoma $Q_2(x)$ su čvorovi tražene kvadraturne formule

$$x_1=\frac{20-\sqrt{120}}{35}=0.25844, \qquad x_2=\frac{20+\sqrt{120}}{35}=0.88441.$$

Koeficijente c_1 i c_2 određujemo tako da formula bude tačna za funkcije $f(x)=1$ i $f(x)=x$, tj. kao rešenje sistema linearnih jednačina

$$\begin{aligned} c_1+c_2&=2 \quad \left(=\int_0^1 \frac{1}{\sqrt{1-x}}\,dx\right)\\ c_1x_1+c_2x_2&=\frac{4}{3} \quad \left(=\int_0^1 \frac{x}{\sqrt{1-x}}\,dx\right). \end{aligned}$$

Tako dobijamo da je

$$c_1=1-\frac{\sqrt{30}}{18}=0.69571, \qquad c_2=1+\frac{\sqrt{30}}{18}=1.30429,$$

pa je tražena formula Gaussovog tipa

$$S(f)=0.69571f(0.25844)+1.30429f(0.88441).$$

Greška ove formule se može oceniti izrazom (40)

$$|R(f)|\le\frac{1}{4!}\max_{[0,1]}|f^{(4)}(x)|\int_0^1\frac{Q_2^2(x)}{\sqrt{1-x}}\,dx=0.5\cdot10^{-3}\max_{[0,1]}|f^{(4)}(x)|.$$

U konkretnom primeru je $f(x)=\cos x$, te je

$$\int_0^1\frac{\cos x}{\sqrt{1-x}}\,dx=1.4992\pm0.0005.$$

PRIMER 4. Ako funkcija $g(x)$ u sebi sadrži činilac $\frac{1}{\sqrt{1-x^2}}$, integral $\int_{-1}^1 g(x)\,dx$ će se uspešno izračunati ako se taj činilac izdvoji kao težinska funkcija, $p(x)\equiv\frac{1}{\sqrt{1-x^2}}$, i primeni kvadraturna formula Gaussovog tipa sa čvorovima koji su nule Čebiševljevog polinoma $T_n(x)=\cos(n\arccos x)$. Za $n=3$ ova formula je

$$\int_{-1}^1\frac{f(x)}{\sqrt{1-x^2}}\,dx\approx\frac{\pi}{3}\sum_{i=1}^3 f(\cos\tfrac{(2i-1)\pi}{6}).$$

Nedostatak kvadraturnih formula Gaussovog tipa u odnosu na Newton–Cotesove formule je što se njima ne može povećati tačnost automatskom regulacijom koraka.

4

Aproksimacija funkcija

Pod aproksimacijom funkcije $f(x)$ podrazumeva se zamena te funkcije nekom drugom funkcijom $g(x)$ koja je njoj bliska u nekom smislu. Razlozi tome mogu biti različiti: analitički izraz za funkciju $f(x)$ je suviše glomazan pa je nepraktično koristiti ga u raznim izračunavanjima, funkcija $f(x)$ je eksperimentalno određena samo na diskretnom skupu tačaka, itd.

Interpolacija je jedan vid aproksimacije kod koje se pod bliskošću dveju funkcija podrazumeva njihovo poklapanje na diskretnom skupu tačaka – tzv. čvorovima interpolacije. Međutim, ako su vrednosti funkcije $f(x)$ u čvorovima interpolacije određene sa nekom greškom (napr. zbog nepreciznosti merenja), onda nema mnogo opravdanja aproksimirati funkciju $f(x)$ funkcijom $g(x)$ određenom interpolacijom, tj. insistirati da funkcija $g(x)$ ima u čvorovima upravo te, neprecizno određene vrednosti. Logičnije je zahtevati da funkcija $g(x)$ bude bliska funkciji $f(x)$ u nekom drugom smislu, napr. bliska u srednjem ili ravnomerno bliska u svim tačkama oblasti definisanosti funkcije $f(x)$.

Kako ćemo definisati bliskost dveju funkcija umnogome će zavisiti i od izbora prostora u kome određujemo aproksimaciju.

4.1 Najbolja aproksimacija u linearnom normiranom prostoru

Neka je f element linearnog normiranog prostora $\mathcal{R}$ i neka su $g_1, \ldots, g_n$ dati linearno nezavisni elementi prostora $\mathcal{R}$. Naći *element najbolje aproksimacije* za f određen elementima $g_1, \ldots, g_n$ znači naći element $\sum_{i=1}^{n} c_i^\circ g_i$ takav da je

$$E_n(f) = \left\| f - \sum_{i=1}^{n} c_i^\circ g_i \right\| = \inf_{c_1,\ldots,c_n} \left\| f - \sum_{i=1}^{n} c_i g_i \right\|, \tag{1}$$

ako takav element postoji. Sa $\|\cdot\|$ je označena norma prostora $\mathcal{R}$.

TEOREMA 1. *U linearnom normiranom prostoru postoji element najbolje aproksimacije.*

DOKAZ: Označimo sa

$$F_f(c_1,\dots,c_n)=\left\|f-\sum_{i=1}^{n}c_ig_i\right\|$$

funkciju n promenljivih $c_1,\dots,c_n$. Kako je

$$\begin{aligned}\left|F_f(c_1^1,\dots,c_n^1)-F_f(c_1^2,\dots,c_n^2)\right| &= \left|\left\|f-\sum_{i=1}^{n}c_i^1g_i\right\|-\left\|f-\sum_{i=1}^{n}c_i^2g_i\right\|\right| \\ \le \left\|\left(f-\sum_{i=1}^{n}c_i^1g_i\right)-\left(f-\sum_{i=1}^{n}c_i^2g_i\right)\right\| &= \left\|\sum_{i=1}^{n}(c_i^1-c_i^2)g_i\right\| \le \sum_{i=1}^{n}|c_i^1-c_i^2|\,\|g_i\|,\end{aligned}$$

to je $F_f(c_1,\dots,c_n)$ neprekidna funkcija svojih argumenata za svaki element $f\in\mathcal{R}$. Dakle, i funkcija

$$F_\circ(c_1,\dots,c_n)=\|c_1g_1+\cdots+c_ng_n\|$$

je neprekidna funkcija svojih argumenata, pa je neprekidna i na jediničnoj sferi $\|\mathbf{c}\|_2=1$, gde je sa $\|\mathbf{c}\|_2$ označena euklidska norma vektora $\mathbf{c}=(c_1,\dots,c_n)$, tj. $\|\mathbf{c}\|_2^2=\sum_{i=1}^{n}c_i^2$. Stoga ona u nekoj tački $(\tilde{c}_1,\dots,\tilde{c}_n)$ jedinične sfere dostiže svoj minimum $\tilde{F}$ na jediničnoj sferi. Pri tome je $\tilde{F}\neq 0$, jer bi u protivnom iz

$$\tilde{F}=\|\tilde{c}_1g_1+\cdots+\tilde{c}_ng_n\|=0$$

sledilo da su $g_1,\dots,g_n$ linearno zavisni elementi, što je suprotno pretpostavci. Dalje, za proizvoljno $\mathbf{c}=(c_1,\dots,c_n)\neq(0,\dots,0)$ važi da je

$$\begin{aligned}F_\circ(c_1,\dots,c_n) &= \|c_1g_1+\cdots+c_ng_n\| \\ &= \|\mathbf{c}\|_2\left\|\frac{c_1}{\|\mathbf{c}\|_2}g_1+\cdots+\frac{c_n}{\|\mathbf{c}\|_2}g_n\right\| = \|\mathbf{c}\|_2\,F_\circ\left(\frac{c_1}{\|\mathbf{c}\|_2},\dots,\frac{c_n}{\|\mathbf{c}\|_2}\right)\ge\|\mathbf{c}\|_2\tilde{F}.\end{aligned}$$

Tako smo dobili sa donje strane ocenu norme elementa $\sum_{i=1}^{n}c_ig_i$,

$$\|c_1g_1+\cdots+c_ng_n\|\ge\|\mathbf{c}\|_2\tilde{F}. \tag{2}$$

van neke okoline tačke $\mathbf{c}=(0,\dots,0)$.

Sa druge strane je $F_f(0,\dots,0)=\|f\|$. Kako je $F_f(c_1,\dots,c_n)$ neprekidna funkcija svojih argumenata, ona je neprekidna i u nekoj okolini tačke $(0,\dots,0)$, tj. za $\|\mathbf{c}\|_2\le\gamma$. Uzmimo da za

$$\gamma>\frac{2\|f\|}{\tilde{F}}$$

funkcija F_f u okolini $\|\mathbf{c}\|_2 \leq \gamma$ postiže svoj minimum F° u tački $(c_1^\circ, \ldots, c_n^\circ)$. Tačka $(0, \ldots, 0)$ pripada lopti $\|\mathbf{c}\|_2 \leq \gamma$, pa je

$$\|f\| = F_f(0, \ldots, 0) \geq F^\circ. \tag{3}$$

Na osnovu (2) i (3) je, takođe, za $\|\mathbf{c}\|_2 > \gamma > \frac{2\|f\|}{\tilde{F}}$

$$\begin{aligned} F_f(c_1, \ldots, c_n) &= \|c_1 g_1 + \cdots + c_n g_n - f\| \geq \|c_1 g_1 + \cdots + c_n g_n\| - \|f\| \\ &\geq \|\mathbf{c}\|_2 \tilde{F} - \|f\| > \frac{2\|f\|}{\tilde{F}} \tilde{F} - \|f\| = \|f\| \geq F^\circ. \end{aligned}$$

Dakle, $F^\circ = F_f(c_1^\circ, \ldots, c_n^\circ)$ je minimum funkcije $F_f(c_1, \ldots, c_n)$ za svaki dopustivi vektor $(c_1, \ldots, c_n)$, tj.

$$F_f(c_1^\circ, \ldots, c_n^\circ) = \inf_{c_1, \ldots, c_n} F_f(c_1, \ldots, c_n)$$

čime je teorema dokazana. ∎

Ovom teoremom smo dokazali da u svakom linearnom normiranom prostoru postoji element najbolje aproksimacije, ali on, u opštem slučaju, ne mora biti jedinstveno odredjen.

TEOREMA 2. *Ako je prostor $\mathcal{R}$ strogo normiran linearni prostor, element najbolje aproksimacije je jedinstven.*

DOKAZ: Pretpostavimo suprotno, tj. da postoje u prostoru $\mathcal{R}$ dva elementa najbolje aproksimacije za f,

$$Q_1 \neq Q_2, \qquad Q_j = \sum_{i=1}^{n} c_i^j g_i, \quad j = 1, 2.$$

Veličina najbolje aproksimacije $E_n(f)$,

$$E_n(f) = \|f - Q_1\| = \|f - Q_2\|,$$

ne može biti jednaka nuli, jer bi u protivnom bilo $Q_1 = Q_2 = f$, što je suprotno pretpostavci. Osim toga je

$$\left\| f - \frac{Q_1 + Q_2}{2} \right\| = \left\| \frac{f - Q_1}{2} + \frac{f - Q_2}{2} \right\| \leq \frac{1}{2}\|f - Q_1\| + \frac{1}{2}\|f - Q_2\| = E_n(f),$$

tj.

$$\left\| f - \frac{Q_1 + Q_2}{2} \right\| \leq E_n(f). \tag{4}$$

Sa druge strane, s obzirom na (1), je

$$\left\| f - \frac{Q_1 + Q_2}{2} \right\| \geq E_n(f), \tag{5}$$

jer je $\frac{Q_1+Q_2}{2} = \frac{1}{2}\sum_{i=1}^{n}(c_i^1 + c_i^2)g_i$, pa iz (4) i (5) sledi da je

$$\left\| \frac{f - Q_1}{2} \right\| + \left\| \frac{f - Q_2}{2} \right\| = E_n(f) = \left\| \frac{f - Q_1}{2} + \frac{f - Q_2}{2} \right\|.$$

Prostor $\mathcal{R}$ je strogo normiran i poslednja jednakost je moguća samo ako je

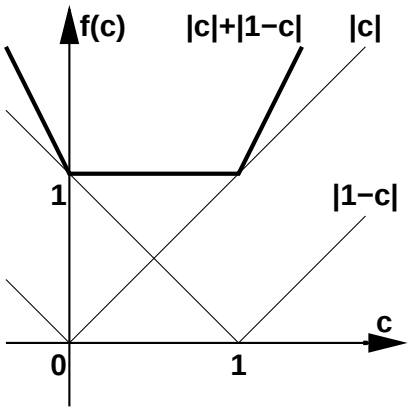
$$\frac{f - Q_1}{2} = \lambda \frac{f - Q_2}{2}, \quad \lambda > 0.$$

Za $\lambda = 1$ je $Q_1 = Q_2$, što je suprotno pretpostavci. Za $\lambda \neq 1$ je $f = \frac{Q_1 - \lambda Q_2}{1-\lambda}$. To znači da je element f linearna kombinacija elemenata $g_1, \ldots, g_n$ te je $E_n(f) = 0$, a to je prema već dokazanom nemoguće. ∎

Prostori koji nisu strogo normirani su npr. $\mathcal{L}[a, b]$ i $\mathcal{C}[a, b]$.

PRIMER 1. Odredimo najbolju aproksimaciju konstantom funkcije $f(x) = x$ na odsečku $[0, 1]$, ako je u prostoru linearnih funkcija norma zadana na sledeći način

$$\|f\| = |f(0)| + |f(1)|.$$



Slika 4.1: Grafik funkcije $|c| + |1 - c|$.

S obzirom da je za svaku konstantu c $\|x - c\| = |c| + |1 - c| \geq 1$, i da je $\|x - c\| = 1$ za $0 \leq c \leq 1$, to su, prema (1), traženi elementi najbolje aproksimacije $Q = c$, $0 \leq c \leq 1$. Očigledno, element najbolje aproksimacije nije jedinstveno

određen, što se i moglo očekivati jer posmatrani prostor nije strogo normiran. Na primer, za funkcije $f_1 = 1$ i $f_2 = x$ je

$$\|f_1 + f_2\| = \|1 + x\| = 3 = 2 + 1 = \|f_1\| + \|f_2\|,$$

iako je $f_1 \neq \lambda f_2$, $\lambda = \text{const}$.

4.2 Najbolja aproksimacija u Hilbertovom prostoru

U Hilbertovim prostorima norma je indukovana skalarnim proizvodom, te je veličina najbolje aproksimacije (1) (tj. njen kvadrat)

$$\big(E_n(f)\big)^2 = \left\| f - \sum_{i=1}^{n} c_i^\circ g_i \right\|^2 = \left(f - \sum_{i=1}^{n} c_i^\circ g_i \,,\, f - \sum_{i=1}^{n} c_i^\circ g_i \right).$$

Označimo sa $\mathcal{H}$ potprostor Hilbertovog prostora $\mathcal{R}$ čiju bazu čine elementi $g_1, \ldots, g_n$ i u kome određujemo element $Q_\circ = \sum_{i=1}^{n} c_i^\circ g_i$ najbolje aproksimacije za $f \in \mathcal{R}$. Osobine elementa $Q_\circ$ iskazuju se dvema lemama koje slede.

LEMA 1. *Neka je $Q_\circ$ element najbolje aproksimacije za f iz $\mathcal{H}$. Tada je razlika $f - Q_\circ$ ortogonalna na svim elementima potprostora $\mathcal{H}$, tj. $Q_\circ$ je ortogonalna projekcija elementa f na potprostor $\mathcal{H}$.*

DOKAZ: Pretpostavimo suprotno, tj. da postoji element $Q_1 \in \mathcal{H}$ takav da je

$$(f - Q_\circ, Q_1) = \alpha \neq 0. \tag{6}$$

Pretpostavimo još da je $\|Q_1\| = 1$; ako to nije ispunjeno uzećemo element $Q_1/\|Q_1\|$. Za element $Q_2 = Q_\circ + \alpha Q_1 \in \mathcal{H}$ važi da je

$$\begin{aligned} \|f - Q_2\|^2 &= (f - Q_\circ - \alpha Q_1, f - Q_\circ - \alpha Q_1) \\ &= (f - Q_\circ, f - Q_\circ) - \alpha(Q_1, f - Q_\circ) - \bar{\alpha}(f - Q_\circ, Q_1) + \alpha\bar{\alpha}(Q_1, Q_1). \end{aligned}$$

S obzirom na (6) imamo da je $(Q_1, f - Q_\circ) = \overline{(f - Q_\circ, Q_1)} = \bar{\alpha}$, pa je

$$\|f - Q_2\|^2 = \|f - Q_\circ\|^2 - |\alpha|^2 < \|f - Q_\circ\|^2,$$

što je u suprotnosti sa pretpostavkom da je $Q_\circ$ element najbolje aproksimacije. ∎

LEMA 2. *Ako je* $(f-Q_\circ, Q)=0$ *za proizvoljan element* $Q\in\mathcal{H}$, *onda je* $Q_\circ$ *element najbolje aproksimacije iz* $\mathcal{H}$ *za* f.

DOKAZ: Za proizvoljan element $Q\in\mathcal{H}$ je

$$\begin{aligned}\|f-Q\|^2 &= (f-Q_\circ+Q_\circ-Q, f-Q_\circ+Q_\circ-Q) = (f-Q_\circ, f-Q_\circ)\\ &\quad + (Q_\circ-Q, f-Q_\circ) + (f-Q_\circ, Q_\circ-Q) + (Q_\circ-Q, Q_\circ-Q).\end{aligned}$$

Kako $Q, Q_\circ\in\mathcal{H}$ to i $Q_\circ-Q\in\mathcal{H}$, te su na osnovu pretpostavke leme drugi i treći sabirak jednaki nuli. Stoga je

$$\|f-Q\|^2 = \|f-Q_\circ\|^2 + \|Q-Q_\circ\|^2, \tag{7}$$

odakle neposredno sledi tvrđenje leme

$$\|f-Q\|^2 > \|f-Q_\circ\|^2 \qquad \text{za } Q\neq Q_\circ.$$

Jedinstvenost ovog elementa sledi iz teoreme 2, jer je svaki Hilbertov prostor strogo normiran. ∎

Sada, koeficijente $c_1^\circ, \ldots, c_n^\circ$ u elementu najbolje aproksimacije $Q_\circ$ možemo odrediti pomoću navedenih lema. Naime, prema lemi 1 je

$$\left(f-\sum_{i=1}^n c_i^\circ g_i\,,\, g_j\right)=0, \qquad j=1,\ldots,n, \tag{8}$$

te je i za proizvoljne skalare c_j, $j=1,\ldots,n$,

$$\left(f-\sum_{i=1}^n c_i^\circ g_i\,,\, \sum_{j=1}^n c_j g_j\right) = \sum_{j=1}^n \bar{c}_j \left(f-\sum_{i=1}^n c_i^\circ g_i\,,\, g_j\right)=0.$$

Kako se svaki element $Q\in\mathcal{H}$ može predstaviti u obliku $Q=\sum_{j=1}^n c_j g_j$, to je

$$\left(f-\sum_{i=1}^n c_i^\circ g_i\,,\, Q\right)=0 \qquad \forall Q\in\mathcal{H}. \tag{9}$$

Sa druge strane, na osnovu leme 2 iz (9) sledi (8), što znači da su ovi uslovi ekvivalentni. Dakle, koeficijenti $c_1^\circ, \ldots, c_n^\circ$ su određeni sistemom linearnih jednačina (8), koji se može zapisati i u sledećem obliku

$$\sum_{i=1}^n c_i^\circ (g_i, g_j) = (f, g_j) \qquad j=1,\ldots,n. \tag{10}$$

Sistem, s obzirom na zaključak o egzistenciji i jedinstvenosti najbolje aproksimacije u Hilbertovom prostoru, ima jedinstveno rešenje. Determinanta sistema (9)

$$G(g_1,\ldots,g_n) = [(g_i\,,\, g_j)]$$

naziva se *Gramovom determinantom* sistema elemenata $g_1, \dots, g_n$. Ona je jednaka nuli ako i samo ako su elementi $g_1, \dots, g_n$ linearno zavisni. Mi smo pretpostavili da su elementi g_i, $i = 1, \dots, n$, linearno nezavisni, odakle takođe sledi jedinstvenost rešenja sistema (10).

Pošto je sistem (10) sve lošije uslovljen što je dimenzija sistema veća, poželjno je koristiti ortonormirane sisteme elemenata, tj. takve da je

$$(g_i \, , \, g_j) = \delta_{ij}, \qquad i, j = 1, \dots, n.$$

Tada je matrica sistema (10) jedinična matrica, i njegovo rešenje je

$$c_j^\circ = (f \, , \, g_j) \, , \qquad j = 1, \dots, n, \tag{11}$$

a element najbolje aproksimacije je

$$Q_\circ = \sum_{i=1}^{n} (f, g_i) \, g_i.$$

Veličinu najbolje aproksimacije $E_n(f)$ možemo dobiti iz relacije (7), stavljajući $Q = 0$:

$$\begin{aligned} \|f - Q_\circ\|^2 = \|f\|^2 - \|Q_\circ\|^2 &= (f \, , \, f) - \left(\sum_{i=1}^{n} c_i^\circ g_i \, , \, \sum_{j=1}^{n} c_j^\circ g_j \right) \\ = (f \, , \, f) - \sum_{i=1}^{n} \sum_{j=1}^{n} c_i^\circ \overline{c_j^\circ} (g_i, g_j) &= (f \, , \, f) - \sum_{i=1}^{n} |c_i^\circ|^2 = (f \, , \, f) - \sum_{i=1}^{n} |(f, g_i)|^2, \end{aligned}$$

tj.

$$\big(E_n(f)\big)^2 = \|f - Q_\circ\|^2 = \|f\|^2 - \sum_{i=1}^{n} |(f, g_i)|^2. \tag{12}$$

Iz (12) neposredno sledi *Besselova nejednakost*

$$\|f\|^2 \geq \sum_{i=1}^{n} |(f \, , \, g_i)|^2. \tag{13}$$

Ako polazni sistem elemenata $g_1, \dots, g_n$ nije ortonormiran, može se izvršiti njegova ortogonalizacija, napr. pomoću Gram–Schmidtovog postupka. Zbog nagomilavanja greške nije preporučljivo ortogonalizaciju vršiti numerički.

Pretpostavimo, dakle, da je $g_1, \dots, g_n, \dots$ ortonormirani sistem elemenata Hilbertovog prostora $\mathcal{R}$. Skalarni proizvodi $c_i^\circ = (f, g_i)$ se nazivaju *Fourierovim koeficijentima* elementa f po ortonormiranom sistemu $\{g_k\}$. Elementu f je pridružen red (ili konačan zbir, ako je ortonormirani sistem konačan)

$$f \sim c_1^\circ g_1 + \cdots + c_n^\circ g_n + \cdots \, ,$$

koji se naziva *Fourierovim redom* elementa f po ortonormiranom sistemu $\{g_k\}$.

Besselova nejednakost važi i za beskonačni ortonormirani sistem elemenata, jer nejednakost (13) važi za svako n, pa i kada $n \to \infty$

$$\sum_{i=1}^{\infty} |c_i^\circ|^2 \leq \|f\|^2, \qquad c_i^\circ = (f, g_i).$$

TEOREMA 3. *Fourierov red elementa f po ortonormiranom sistemu $\{g_k\}$ je konvergentan.*

DOKAZ: Neka je Fourierov red elementa f po ortonormiranom sistemu $\{g_k\}$

$$c_1^\circ g_1 + \cdots + c_n^\circ g_n + \cdots, \qquad c_i^\circ = (f, g_i)$$

i neka je njegova n-ta delimična suma

$$S_n = \sum_{i=1}^{n} c_i^\circ g_i.$$

Za $m < n$ je

$$\|S_n - S_m\|^2 = \left\| \sum_{i=m+1}^{n} c_i^\circ g_i \right\|^2 = \sum_{i=m+1}^{n} \sum_{j=m+1}^{n} c_i^\circ \overline{c_j^\circ} (g_i, g_j) = \sum_{i=m+1}^{n} |c_i^\circ|^2.$$

Pošto iz Besselove nejednakosti sledi da je red $\sum_{i=1}^{\infty} |c_i^\circ|^2$ konvergentan, to $\sum_{i=m+1}^{n} |c_i^\circ|^2 \to 0$ kada $m, n \to \infty$. Stoga je niz delimičnih suma $\{S_k\}$ Cauchyev niz, i on konvergira ka nekom elementu S iz Hilbertovog prostora $\mathcal{R}$,

$$\lim_{k \to \infty} S_k = S, \qquad S \in \mathcal{R}, \tag{14}$$

što je i trebalo dokazati. ∎

Naredna teorema formuliše uslov pod kojim Fourierov red konvergira upravo ka samom elementu. Pre navođenja teoreme, definišimo pojam potpunog ortonormiranog sistema. Ortonormirani sistem elemenata se naziva potpunim, ako ne postoji ni jedan drugi element Hilbertovog prostora $\mathcal{R}$ koji je različit od nule, a ortogonalan je na svim elementima sistema. U svakom Hilbertovom prostoru postoji najviše prebrojiv potpun ortonormirani sistem elemenata.

TEOREMA 4. *U Hilbertovom prostoru $\mathcal{R}$ Fourierov red proizvoljnog elementa po potpunom ortonormiranom sistemu elemenata konvergira ka tom elementu.*

DOKAZ: Dokazaćemo da je razlika $f - S$, gde je S suma Fourierovog reda elementa f, ortogonalna na sve elemente potpunog ortonormiranog sistema elemenata $\{g_k\}$.

Kako je za $n > k$

$$(S_n\,,\,g_k) = \left(\sum_{i=1}^{n} c_i^\circ g_i\,,\,g_k\right) = \sum_{i=1}^{n} c_i^\circ (g_i, g_k) = c_k^\circ,$$

a iz (11) je $(f, g_k) = c_k^\circ$, to je

$$\begin{aligned}(f - S\,,\,g_k) &= (f\,,\,g_k) - (S - S_n\,,\,g_k) - (S_n\,,\,g_k)\\ &= c_k^\circ - (S - S_n\,,\,g_k) - c_k^\circ = -(S - S_n\,,\,g_k).\end{aligned}$$

Na osnovu Cauchy–Schwarzove nejednakosti i relacije (14) je

$$0 \le |(f - S\,,\,g_k)| \le \|S - S_n\|\,\|g_k\| \overset{n\to\infty}{\longrightarrow} 0.$$

Pošto leva strana ne zavisi od n, poslednja nejednakost je moguća samo ako je

$$(f - S\,,\,g_k) = 0.$$

Indeks k je proizvoljan, pa ovaj zaključak važi za svako k. Po pretpostavci sistem $\{g_k\}$ je potpun, a $f - S \in \mathcal{R}$, to je moguće samo ako je $f = S$, tj. ako Fourierov red elementa f konvergira ka f. ∎

Iz relacije (12), za $Q_\circ = S_n$ i kada $n \to \infty$, na osnovu (14) i dokazane teoreme sledi *Parsevalova jednakost*

$$\|f\|^2 = \sum_{i=1}^{\infty} |c_i^\circ|^2, \qquad c_i^\circ = (f, g_i). \tag{15}$$

4.3 Srednjekvadratna aproksimacija

Ako se za Hilbertov prostor $\mathcal{R}$ uzme prostor funkcija $\mathcal{L}_2[a, b]$, tj. prostor funkcija integrabilnih sa kvadratom na odsečku $[a, b]$, u kome je norma definisana integralom

$$\|f\|^2 = \int_a^b f^2\,dx, \qquad f \in \mathcal{L}_2[a, b],$$

onda se element najbolje aproksimacije naziva elementom *najbolje srednjekvadratne aproksimacije*.

Ovaj pojam se može shvatiti i nešto opštije, ukoliko se u posmatranom prostoru skalarni proizvod definiše na sledeći način:

$$(f\,,\,g) = \int_a^b p(x) f(x) g(x)\,dx, \qquad p(x) > 0.$$

Funkcija $p(x)$ naziva se *težinskom funkcijom.* Definisana je na odsečku $[a, b]$ i zadovoljava uslov $p(x) > 0$ skoro svuda, tj. može biti jednaka nuli samo na skupu mere nula.

Prema napred rečenom, element najbolje srednjekvadratne aproksimacije postoji i jedinstveno je određen. To je funkcija $Q_\circ(x)$ iz potprostora prostora $\mathcal{L}_2[a, b]$ određenog linearno nezavisnim funkcijama $g_k(x)$, $k = 1, \ldots, n$,

$$Q_\circ(x) = c_1^\circ g_1(x) + \cdots + c_n^\circ g_n(x),$$

koja zadovoljava relaciju

$$\|f - Q_\circ\| = \inf_Q \|f - Q\| = \inf_Q \left(\int_a^b p(f - Q)^2 \, dx \right)^{\frac{1}{2}}.$$

Dakle, $Q_\circ(x)$ je ona funkcija iz skupa svih dopustivih funkcija $Q(x)$, kojom se postiže minimalno odstupanje u srednjem, tj. u nekom smislu minimalna veličina površine izmedju funkcija f i Q – u pojedinim tačkama intervala odstupanje funkcije $Q_\circ$ od f može biti veliko. Pomoću funkcije $p(x)$ se postiže različiti kvalitet aproksimacije u različitim delovima intervala. Naime, u delovima intervala gde je $p(x)$ veće, razlika $f(x) - Q_\circ(x)$ je množena većim koeficijentom, te sa većom težinom učestvuje u minimizaciji. Iz tog razloga je funkcija $p(x)$ nazvana težinskom.

Kao što je u prethodnom odeljku rečeno, najjednostavnije je ako se aproksimacija određuje pomoću ortonormiranog sistema, tj. ako je

$$(g_i \, , \, g_j) = \int_a^b p(x) g_i(x) g_j(x) \, dx = \delta_{ij}, \qquad i, j = 1, \ldots, n.$$

Za određene oblike težinske funkcije $p(x)$ ortonormirani sistemi su poznati – napr. sistem Legendreovih polinoma ($p(x) \equiv 1$), sistem Čebiševljevih polinoma prve vrste ($p(x) = 1/\sqrt{1 - x^2}$), sistem Hermiteovih polinoma ($p(x) = e^{-x^2}$), itd. Za druge oblike težinske funkcije ortonormirani sistemi polinoma se mogu odrediti napr. Gram–Schmidtovim postupkom ortogonalizacije, polazeći od proizvoljnog skupa linearno nezavisnih polinoma (najčešće $1, x, x^2, \ldots, x^n$). Sistem ortonormiranih polinoma u odnosu na datu težinsku funkciju je jednoznačno određen.

PRIMER 2. Ako je težinska funkcija $p(x) \equiv 1$, za nalaženje najbolje polinomijalne srednjekvadratne aproksimacije funkcije $f(x) \in \mathcal{L}_2[a, b]$ najbolje je koristiti Legendreove polinome, koji su na odsečku $[-1, 1]$ definisani formulom

$$L_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} \big((x^2 - 1)^n \big).$$

S obzirom da je

$$(L_i \, , \, L_j) = \begin{cases} 0, & i \neq j \\ \frac{2}{2j+1}, & i = j \end{cases}$$

koeficijenti tražene aproksimacije

$$Q_\circ(x) = \sum_{k=0}^{n} c_k^\circ L_k(x)$$

se direktno određuju po formuli

$$c_k^\circ = \frac{(f\,,\,L_k)}{\|L_k\|^2} = \frac{2k+1}{2}(f\,,\,L_k).$$

Na primer, ako je $f(x) = |x|$, $[a,b] \equiv [-1,1]$ i $n = 5$, onda je

$$Q_\circ(x) = \frac{1}{2} + \frac{5}{16}(3x^2-1) - \frac{3}{128}(35x^4 - 30x^2 + 3) = \frac{15}{128}(-7x^4 + 14x^2 + 1).$$

Ako se ortonormirani sistem sastoji od prebrojivo mnogo polinoma $P_k(x)$, $k = 0, 1, \ldots$, funkciji $f \in \mathcal{L}_2[a,b]$ možemo pridružiti njen Fourierov red po tom ortonormiranom sistemu:

$$f(x) \sim \sum_{i=0}^{\infty} c_i^\circ P_i(x), \qquad c_i^\circ = \int_a^b p(x) f(x) P_i(x)\,dx. \tag{16}$$

Sistem polinoma P_k je obrazovan od potpunog sistema $1, x, \ldots, x^n, \ldots$, te je i sam potpun. Stoga, prema teoremi 4, Fourierov red (16) konvergira ka funkciji $f(x)$ u smislu uvedene metrike, tj.

$$\lim_{n\to\infty} \int_a^b p(x) \left(f(x) - \sum_{i=0}^{n} c_i^\circ P_i(x) \right)^2 dx = 0.$$

Ako je $f(x)$ periodična funkcija, prirodno je zahtevati da i aproksimacija bude takva. Taj uslov će biti zadovoljen ako se sistem funkcija $\{g_k\}$ sastoji iz funkcija

$$1, \sin x, \cos x, \sin 2x, \cos 2x, \ldots, \sin nx, \cos nx, \ldots\,. \tag{17}$$

Ovaj sistem funkcija je ortogonalan sistem funkcija na odsečku $[-\pi, \pi]$ u odnosu na težinsku funkciju $p(x) \equiv 1$, jer je

$$\begin{aligned}
(\sin mx\,,\,\sin nx) &= \begin{cases} 0, & \text{za } m \neq n \\ \pi, & \text{za } m = n \end{cases} \\
(\cos mx\,,\,\cos nx) &= \begin{cases} 0, & \text{za } m \neq n \\ \pi, & \text{za } m = n \neq 0 \\ 2\pi, & \text{za } m = n = 0 \end{cases} \\
(\sin mx\,,\,\cos nx) &= 0 \qquad \text{za svako } m \text{ i } n.
\end{aligned} \tag{18}$$

Funkcija definisana pomoću prvih $2n+1$ funkcija sistema (17) naziva se trigonometrijskim polinomom reda n,

$$Q(x) = \tfrac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx).$$

Ovaj polinom će biti polinom najbolje srednjekvadratne aproksimacije funkcije f, periodične na intervalu $[-\pi, \pi]$, ako su koeficijenti a_k i b_k rešenja sistema (10). S obzirom na (18), matrica sistema je dijagonalna te su ovi, tzv. trigonometrijski Fourierovi koeficijenti funkcije f,

$$
\begin{aligned}
a_k &= \tfrac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx \, dx \qquad k = 0, \ldots, n, \\
b_k &= \tfrac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin kx \, dx \qquad k = 1, \ldots, n.
\end{aligned} \tag{19}
$$

Red određen svim funkcijama sistema (17),

$$\tfrac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

naziva se trigonometrijskim Fourierovim redom. Ako su njegovi koeficijenti određeni izrazima (19), red je trigonometrijski Fourierov red funkcije f i, s obzirom da je sistem funkcija (17) potpun ortogonalni sistem funkcija, prema teoremi 4 on konvergira u srednjem ka funkciji $f(x)$:

$$\lim_{n \to \infty} \int_{-\pi}^{\pi} \left(f - \tfrac{a_0}{2} - \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx) \right)^2 dx = 0.$$

Dakle, svaku dovoljno glatku periodičnu funkciju (u tom smislu da postoje integrali (19)) možemo predstaviti njenim Fourierovim redom

$$f(x) = \tfrac{a_0}{2} + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x + \cdots, \tag{20}$$

odnosno prikazati je kao linearnu kombinaciju čistih harmonika $\sin kx$ i $\cos kx$, $k = 1, 2, \ldots$, čija je učestanost oscilovanja k na intervalu 2π. Konstantni član $\frac{a_0}{2}$ je srednja vrednost funkcije $f(x)$ na intervalu $(-\pi, \pi)$

$$\frac{a_0}{2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \, dx,$$

te ostali sabirci u redu (20) osciluju oko nule a suma im je $f - \frac{a_0}{2}$.

Zamenimo u redu (20) poznate izraze za $\sin kx$ i $\cos kx$

$$\sin kx = \tfrac{1}{2\imath} \left(e^{\imath kx} - e^{-\imath kx} \right) \qquad \cos kx = \tfrac{1}{2} \left(e^{\imath kx} + e^{-\imath kx} \right)$$

i stavimo

$$a_0 = 2c_0, \qquad a_k = c_k + c_{-k}, \qquad b_k = \imath(c_{-k} - c_k), \qquad k = 1, 2, \ldots,$$

gde su $c_k, k = 0, \pm 1, \ldots$, nove konstante. Dobijamo kompleksni zapis Fourierovog reda (20)

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{-\imath kx}. \tag{21}$$

Koeficijenti c_k se mogu izračunati na osnovu njihove veze sa a_k i b_k i formula (19). Drugi način da se oni odrede je direktno korišćenjem reda (21). Sistem funkcija $\{e^{\imath kx}\}$ je sistem ortogonalnih funkcija, jer je

$$\int_{-\pi}^{\pi} e^{\imath kx} e^{-\imath lx}\, dx = \begin{cases} 0, & \text{za } k \neq l \\ 2\pi, & \text{za } k = l. \end{cases}$$

za svako $k, l = 0, \pm 1, \ldots$. Zbog toga, množenjem reda (21) sa $e^{\imath jx}$ i integraljenjem u granicama od $(-\pi, \pi)$ dobijamo da je

$$c_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{\imath jx}\, dx \qquad j = 0, \pm 1, \ldots . \tag{22}$$

PRIMER 3. Fourierov red generalisane funkcije $\delta(x)$ na odsečku $[-1, 1]$ je

$$\begin{aligned} \delta(x) &\sim \frac{1}{2\pi}(1 + e^{-\imath x} + e^{\imath x} + e^{-2\imath x} + e^{2\imath x} + \cdots) \\ &= \frac{1}{2\pi}(1 + 2\cos x + 2\cos 2x + \cdots), \end{aligned}$$

jer je, prema (22),

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} \delta(x) e^{\imath kx}\, dx = \frac{1}{2\pi}, \quad k = 0, \pm 1, \ldots .$$

Njegova n-ta parcijalna suma je

$$S_n(x) = \frac{1}{2\pi} \sum_{k=-n}^{n} e^{\imath kx} = \frac{1}{2\pi} e^{-\imath nx} \frac{e^{\imath(2n+1)x} - 1}{e^{\imath x} - 1} = \frac{1}{2\pi} \frac{\sin(n + \frac{1}{2})x}{\sin \frac{1}{2}x}.$$

Pored toga što čine potpun ortogonalni sistem, funkcije $e^{\imath kx}$ imaju i druge dobre osobine – napr. one su sopstvene funkcije operatora diferenciranja i operatora konačnih razlika

$$\frac{d}{dx} e^{\imath kx} = \imath k\, e^{\imath kx} \qquad \Delta\, e^{\imath kx} = \left(\frac{e^{\imath kh} - 1}{h} \right) e^{\imath kx},$$

te se stoga reprezentacija (21) često koristi u praksi. Red (21) sa koeficijentima datim formulom (22) je pridružen 2π-periodičnoj funkciji $f(x)$. Da bi se dobile odgovarajuće formule za funkciju periodičnu na intervalu T, u formuli (22) uvedimo smenu $x = \frac{2\pi}{T} t$,

$$c_k = \tfrac{1}{T} \int_{\frac{-T}{2}}^{\frac{T}{2}} f(\tfrac{2\pi}{T} t) e^{\imath k \frac{2\pi}{T} t}\, dt = \tfrac{1}{T} \int_{\frac{-T}{2}}^{\frac{T}{2}} f_T(t) e^{\imath Kt}\, dt,$$

gde je $f_T(t) \equiv f(\frac{2\pi}{T} t)$ periodična funkcija sa periodom T, i $K = k\frac{2\pi}{T} = k\Delta K$.

Fourierov red (21) funkcije $f_T(t)$ je

$$f_T(t) = \sum_{k=-\infty}^{\infty} c_k e^{-\imath Kt} = \sum_{k=-\infty}^{\infty} \tfrac{1}{T} e^{-\imath Kt} \left(\int_{\frac{-T}{2}}^{\frac{T}{2}} f_T(t) e^{\imath Kt}\, dt \right). \tag{23}$$

Što je T veće, funkcija f_T je jednaka funkciji f na većem intervalu, a suma u (23) teži integralu po K, jer $\frac{1}{T} = \frac{\Delta K}{2\pi} \to 0$ kada $T \to \infty$. Stoga, kada $T \to \infty$ izraz (23) postaje

$$f(t) = \int_{-\infty}^{\infty} \tfrac{dK}{2\pi} e^{-\imath Kt} \left(\int_{-\infty}^{\infty} f(t) e^{\imath Kt}\, dt \right). \tag{24}$$

Izraz u zagradi u relaciji (24) naziva se *Fourierovom transformacijom* funkcije $f(t)$ i funkcija je od K,

$$F(K) = \int_{-\infty}^{\infty} f(t) e^{\imath Kt}\, dt, \tag{25}$$

a izraz (24), kada se (25) uzme u obzir, je *inverzna Fourierova transformacija* kojom se funkcija $F(K)$ transformiše natrag u funkciju $f(t)$

$$f(t) = \tfrac{1}{2\pi} \int_{-\infty}^{\infty} F(K) e^{-\imath Kt}\, dK.$$

Parsevalova jednakost (15) u ovom graničnom slučaju postaje

$$2\pi \int_{-\infty}^{\infty} |f(t)|^2\, dt = \int_{-\infty}^{\infty} |F(K)|^2\, dK.$$

4.4 Metoda najmanjih kvadrata

U prethodnom odeljku je bilo reči o aproksimaciji funkcija zadatih skoro svuda na odsečku $[a, b]$. Neka je sada funkcija $f(x)$ poznata na konačnom skupu tačaka $x_0, x_1, \ldots, x_n$ odsečka $[a, b]$. Jedna od mogućih aproksimacija funkcije $f(x)$ je interpolacioni polinom određen zadatim podacima. Međutim, ako je n veliko ili su podaci dati sa određenom greškom, ovakav izbor aproksimacije nije najpogodniji.

Za funkcije zadate na diskretnom skupu tačaka, metoda koja odgovara srednjekvadratnoj aproksimaciji je metoda najmanjih kvadrata. Integral se zamenjuje sumom, te je skalarni proizvod definisan izrazom

$$(f\ ,\ g) = \sum_{i=0}^{n} p_i f(x_i) g(x_i), \qquad p_i > 0, \tag{26}$$

gde je $n+1$ broj tačaka u kojima je funkcija zadata. p_i su zadati pozitivni brojevi i nazivaju se *težinski koeficijenti*. U tački x_k, kojoj je pridružen veći težinski koeficijent p_k, odstupanje aproksimacije od funkcije će biti manje. Skalarnim proizvodom (26) definisana je norma

$$\|f\|^2 = (f\ ,\ f) = \sum_{i=0}^{n} p_i\big(f(x_i)\big)^2.$$

Neka su $g_0(x), g_1(x), \ldots, g_m(x)$, $m \leq n$, date linearno nezavisne funkcije na odsečku $[a, b]$. Ako je sistem funkcija g_k Čebiševljev sistem, tj. ako proizvoljan generalisani polinom po tom sistemu funkcija ima na $[a, b]$ ne više od m različitih nula, onda je element najbolje aproksimacije za f, određen ovim sistemom funkcija, jedinstven. Taj element tražimo u obliku

$$Q_\circ(x) = \sum_{i=0}^{m} c_i^\circ g_i(x),$$

pri čemu se koeficijenti c_i° određuju tako da izraz

$$\|f - Q_\circ\|^2 \equiv \sum_{i=0}^{n} p_i\big(f(x_i) - Q_\circ(x_i)\big)^2 \tag{27}$$

bude minimalan. Prema onome što je rečeno uopšte o aproksimaciji u Hilbertovim prostorima, koeficijenti c_k° su rešenja sistema linearnih jednačina

$$\sum_{i=0}^{m} c_i^\circ (g_i\ ,\ g_j) = (f\ ,\ g_j), \qquad j = 0, \ldots, m, \tag{28}$$

pri čemu je skalarni proizvod $(\cdot\ ,\ \cdot)$ određen izrazom (26). Determinanta sistema (28) je različita od nule jer su funkcije $g_k(x)$ linearno nezavisne, te sistem ima jedinstveno rešenje.

U slučaju da je $m = n$ izraz (27) će biti minimalan, tj. nula, ako je

$$f(x_k) = Q_\circ(x_k) \qquad k = 0, \ldots, m,$$

što znači da je $Q_\circ(x)$ interpolacioni polinom određen datim skupom tačaka. Kada je $n > m$, broj sabiraka u sumi (27) je veći od broja slobodnih parametara c_k°, te u opštem slučaju ovi ne mogu biti određeni tako da svi sabirci budu jednaki nuli. Stoga se postavlja zahtev da ukupno odstupanje $Q_\circ(x)$ od $f(x)$ u svim tačkama x_i, $i = 0, \ldots, n$ bude što je moguće manje. Kako razlike $Q_\circ(x_i) - f(x_i)$, $i = 0, \ldots, n$, mogu biti različitog znaka, minimizira se suma kvadrata ovih razlika, eventualno pomnoženih težinskim faktorima, tj. izraz (27). Zato se metoda i naziva metodom najmanjih kvadrata.

Najčešće se za funkcije $g_0, \ldots, g_m$ biraju algebarski ili trigonometrijski polinomi. Funkcije $1, x, \ldots, x^n$ obrazuju Čebiševljev sistem funkcija na proizvoljnom odsečku, te se oni ili ma koja njihova kombinacija, koja predstavlja sistem linearno nazavisnih

funkcija, mogu uzeti kao funkcije $g_k(x)$. Odgovarajućim kombinovanjem mogu se dobiti i polinomi koji su ortogonalni u smislu skalarnog proizvoda (26). Pri tome će sistem ortogonalnih polinoma očigledno zavisiti od izbora težinskih koeficijenata p_k i tačaka x_k.

U slučaju aproksimacije periodične funkcije zadate u tačkama

$$0 < x_1 < x_2 < \cdots < x_n \leq 2\pi,$$

najčešće se koristi Čebiševljev sistem funkcija

$$1, \quad \cos x, \quad \sin x, \quad \cos 2x, \quad \sin 2x, \quad \ldots \quad , \cos mx, \quad \sin mx, \tag{29}$$

pri čemu je $n \geq 2m+1$. Trigonometrijski polinom

$$Q_\circ(x) = \frac{a_0}{2} + \sum_{k=1}^{m}(a_k \cos kx + b_k \sin kx),$$

čiji su koeficijenti rešenja sistema jednačina (28), je trigonometrijski polinom najbolje aproksimacije određen u smislu metode najmanjih kvadrata i, s obzirom na prethodne opšte zaključke, jednoznačno je određen.

Ako su svi težinski koeficijenti p_k jednaki jedinici a tačke x_k ravnomerno raspoređene na intervalu $[0, 2\pi]$, sistem linearnih jednačina (28) je sistem sa dijagonalnom matricom, jer tada funkcije (29) čine ortogonalni sistem funkcija u odnosu na skalarni proizvod (26). Dokažimo ortogonalnost sistema funkcija (29) pri navedenim pretpostavkama. Neka je

$$x_1 = \alpha, \quad x_2 = 2\alpha, \quad \ldots \quad , x_n = n\alpha, \qquad \alpha = \tfrac{2\pi}{n}.$$

Tada je za $e^{\imath k\alpha} \neq 1$, tj. $k \neq Nn$, $N = 0, \pm 1, \ldots,$

$$\sum_{j=1}^{n} e^{\imath k x_j} = \sum_{j=1}^{n} e^{\imath kj\alpha} = e^{\imath k\alpha}\frac{e^{\imath kn\alpha} - 1}{e^{\imath k\alpha} - 1} = 0,$$

odakle sledi da je

$$\sum_{j=1}^{n} \cos kx_j = 0, \qquad \sum_{j=1}^{n} \sin kx_j = 0, \qquad \text{za } k \neq Nn, \ N \text{ ceo broj.}$$

Stoga je za $k, l = 0, 1, \ldots, m$ i $k \neq l$

$$\sum_{j=1}^{n} \cos kx_j \cos lx_j = \tfrac{1}{2}\sum_{j=1}^{n} \cos{(k+l)x_j} + \tfrac{1}{2}\sum_{j=1}^{n} \cos{(k-l)x_j} = 0,$$
$$\sum_{j=1}^{n} \sin kx_j \sin lx_j = \tfrac{1}{2}\sum_{j=1}^{n} \cos{(k-l)x_j} - \tfrac{1}{2}\sum_{j=1}^{n} \cos{(k+l)x_j} = 0,$$

jer zbog uslova $n \geq 2m+1$, $k+l$ i $k-l$ ne mogu biti jednaki umnošku broja n. Za $k=l$ je

$$\sum_{j=1}^{n} \cos^2 kx_j = \tfrac{1}{2}\sum_{j=1}^{n}(1+\cos 2kx_j) = \tfrac{n}{2},$$
$$\sum_{j=1}^{n} \sin^2 kx_j = \tfrac{1}{2}\sum_{j=1}^{n}(1-\cos 2kx_j) = \tfrac{n}{2}.$$

I na kraju, za svako k i l, $k,l = 0,\ldots,m$, je

$$\sum_{j=1}^{n} \sin kx_j \cos lx_j = \tfrac{1}{2}\sum_{j=1}^{n} \sin(k+l)x_j + \tfrac{1}{2}\sum_{j=1}^{n} \sin(k-l)x_j = 0.$$

Dobili smo, dakle, da je

$$(\sin kx\,,\,\sin lx) = \begin{cases} 0, & \text{za } k \neq l \\ \frac{n}{2}, & \text{za } k = l \end{cases} \qquad (\cos kx\,,\,\cos lx) = \begin{cases} 0, & \text{za } k \neq l \\ \frac{n}{2}, & \text{za } k = l \neq 0 \\ n, & \text{za } k = l = 0 \end{cases}$$
$$(\sin kx\,,\,\cos lx) = 0 \qquad \text{za svako } k \text{ i } l,$$

što je i trebalo pokazati.

Stoga su koeficijenti trigonometrijskog polinoma najbolje aproksimacije u smislu metode najmanjih kvadrata, tj. rešenja sistema (28),

$$a_k = \tfrac{2}{n}\sum_{j=1}^{n} f(x_j)\cos kx_j \qquad k = 0,\ldots,m,$$
$$b_k = \tfrac{2}{n}\sum_{j=1}^{n} f(x_j)\sin kx_j \qquad k = 1,\ldots,m.$$

Ove formule nazivaju se *Besselove formule* i predstavljaju trapeznim pravilom izračunate integrale kojima su dati Fourierovi trigonometrijski koeficijenti (19).

4.5 Diskretna Fourierova transformacija

Diskretna Fourierova transformacija je algoritam kojim se n-dimenzioni vektor $\mathbf{c} = (c_0,\ldots,c_{n-1})^T$ koeficijenata parcijalne sume reda (21), određuje tako da je u ekvidistantnim tačkama $\frac{2\pi j}{n}$, $j = 0,\ldots,n-1$, intervala $[0,2\pi]$ ova parcijalna suma jednaka vrednostima funkcije f u tim tačkama:

$$f(x_j) = \sum_{k=0}^{n-1} c_k e^{-\imath k\frac{2\pi j}{n}}, \qquad j = 0,\ldots,n-1. \tag{30}$$

Ako označimo sa $f_j = f(x_j)$, vektor $\mathbf{f} = (f_0, \ldots, f_{n-1})^T$ i

$$W = e^{\imath \frac{2\pi}{n}} = \sqrt[n]{e^{\imath 2\pi}}, \tag{31}$$

sistem linearnih jednačina (30) može da se zapiše i u sledećem obliku

$$\sum_{k=0}^{n-1} c_k \overline{W}^{kj} = f_j, \qquad j = 0, \ldots, n-1, \tag{32}$$

ili, u vektorskom obliku,

$$F^* \mathbf{c} = \mathbf{f}.$$

Matrica sistema (32) je konjugovana matrica *Fourierove matrice*

$$F = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & W & W^2 & \ldots & W^{n-1} \\ 1 & W^2 & W^4 & \ldots & W^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & W^{n-1} & W^{2(n-1)} & \ldots & W^{(n-1)^2} \end{pmatrix}, \tag{33}$$

gde je W dato u (31). Sistem (32) se može direktno rešiti, jer je

$$F\,F^* = F^*\,F = nI, \tag{34}$$

I je jedinična matrica. Zaista, element u j-toj vrsti i k-toj koloni matrice proizvoda F^*F je

$$1 + \overline{W}^j W^k + \cdots + \overline{W}^{(n-1)j} W^{(n-1)k}. \tag{35}$$

Kada je $j = k$, sabirci u zbiru (35) su oblika $(\overline{W}\,W)^{jl}$, $l = 0, \ldots, n-1$, te je zbir jednak n. Kada je $j \neq k$, uvodeći oznaku $r = \overline{W}^j W^k$, (35) može da se predstavi kao suma prvih n članova geometrijske progresije sa količnikom $r \neq 1$, te je $1 + r + r^2 + \cdots + r^{n-1} = \frac{r^n - 1}{r-1} = 0$, jer je, na osnovu (31), $r^n = (\overline{W}^n)^j (W^n)^k = 1$. Dakle, dijagonalni član matrice F^*F je n, a nedijagonalni 0, čime je (34) dokazano.

Izraz (34) možemo napisati i u obliku

$$\left(\tfrac{1}{\sqrt{n}} F\right) \left(\tfrac{1}{\sqrt{n}} F\right)^* = I$$

što znači da je matrica $\frac{1}{\sqrt{n}} F$ unitarna matrica.

Iz relacije (34) sledi da je inverzna matrica matrici sistema (32)

$$(F^*)^{-1} = \tfrac{1}{n} F$$

te je rešenje ovoga sistema, tj. diskretna Fourierova transformacija vektora $\mathbf{f} = (f_0, \ldots, f_{n-1})^T$, vektor $\mathbf{c}$ sa koordinatama

$$c_k = \frac{1}{n}\sum_{j=0}^{n-1} f_j W^{jk} = \frac{1}{n}\sum_{j=0}^{n-1} f_j e^{\imath\frac{2\pi jk}{n}}, \qquad k = 0,\dots,n-1 \tag{36}$$

Poredeći izraze (36) i (22), vidimo da formula (36) predstavlja približnu vrednost integrala (22) 2π-periodične funkcije, izračunatu trapeznim pravilom sa korakom $\frac{2\pi}{n}$. Ovo je, dakle, način da se izračunaju neki od koeficijenata (22) u razvoju (21), kada integrale nije moguće iz ma kojih razloga izračunati, i tako odredi aproksimacija funkcije $f(x)$.

Neke primene diskretne Fourierove transformacije ilustruju sledeći primeri.

PRIMER 4. Pomoću diskretne Fourierove transformacije može se realizovati množenje polinoma korišćenjem osobine diskretne konvolucije. *Diskretna konvolucija* dva vektora $\mathbf{f} = (f_0,\dots,f_{n-1})^T$ i $\mathbf{g} = (g_0,\dots,g_{n-1})^T$ se definiše kao vektor dimenzije n u oznaci $\mathbf{f}*\mathbf{g}$, čija je m-ta koordinata, $m = 0,\dots,n-1$, suma proizvoda $f_j g_k$ za $j+k = m$ ili $j+k = m+n$,

$$\mathbf{f}*\mathbf{g} = \begin{pmatrix} f_0g_0 + f_1g_{n-1} + f_2g_{n-2} + \cdots + f_{n-1}g_1 \\ f_0g_1 + f_1g_0 + f_2g_{n-1} + \cdots + f_{n-1}g_2 \\ \vdots \\ f_0g_{n-1} + f_1g_{n-2} + \cdots + f_{n-1}g_0 \end{pmatrix}.$$

Proizvod dva polinoma $P_1(x) = \sum_{j=0}^{n_1} f_j x^j$ i $P_2(x) = \sum_{j=0}^{n_2} g_j x^j$ je polinom stepena $n = n_1 + n_2$ čiji su koeficijenti određeni konvolucijom $(n+1)$-dimenzionih vektora

$$\mathbf{f} = (f_0, f_1, \dots, f_{n_1}, 0, \dots, 0)^T \quad \text{i} \quad \mathbf{g} = (g_0, g_1, \dots, g_{n_2}, 0, \dots, 0)^T.$$

Međutim, Fourierova transformacija diskretne konvolucije n-dimenzionih vektora $\mathbf{f}$ i $\mathbf{g}$ je jednaka proizvodu njihovih diskretnih Fourierovih transformacija pomnoženom sa n (dokazati),

$$\mathbf{f}*\mathbf{g} = n\,F^*(\hat{\mathbf{f}}\,\hat{\mathbf{g}}) \qquad \text{gde je} \quad \hat{\mathbf{f}} = \tfrac{1}{n}F\,\mathbf{f}, \quad \hat{\mathbf{g}} = \tfrac{1}{n}F\,\mathbf{g}.$$

Stoga je dovoljno naći Fourierove transformacije vektora koeficijenata polinoma činilaca dopunjenih nulama, izmnožiti po koordinatama dobijena dva vektora i naći inverznu Fourierovu transformaciju vektora proizvoda. Komponente dobijenog vektora pomnožene sa n predstavljaju koeficijente polinoma proizvoda.

PRIMER 5. Diskretna konvolucija vektora $\mathbf{f}$ i $\mathbf{g}$ je, ustvari, vektor koji se dobija kada se cikličnom matricom odredjenom vektorom $\mathbf{f}$ pomnoži sa leve strane vektor $\mathbf{g}$

$$\mathbf{f}*\mathbf{g} = \begin{pmatrix} f_0 & f_{n-1} & f_{n-2} & \cdots & f_1 \\ f_1 & f_0 & f_{n-1} & \cdots & f_2 \\ f_2 & f_1 & f_0 & \cdots & f_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{n-1} & f_{n-2} & f_{n-3} & \cdots & f_0 \end{pmatrix} \cdot \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_{n-1} \end{pmatrix}.$$

Stoga se pomoću diskretne Fourierove transformacije efikasno realizuje množenje cikličnim (Toeplitzovim) matricama.

Očigledno je da je od interesa da se diskretna Fourierova transformacija izračuna sa što manje računskih operacija.

Brza Fourierova transformacija (FFT). (Često se u literaturi koristi ova skraćenica, koja potiče od engleskog naziva Fast Fourier Transformation). Da bi se realizovala diskretna Fourierova transformacija, tj. da bi se izračunali svi koeficijenti c_k dati formulom (36), potrebno je izvršiti n^2 množenja u opštem slučaju kompleksnih brojeva $f_j W^{jk}$, $j,k = 0,\dots,n-1$, i još izvestan broj operacija radi nalaženja stepena broja W. U slučaju korišćenja FFT-algoritma broj operacija je $O(n\log_2 n)$, dakle skoro linearan sa n.

FFT se zasniva na poznatoj lemi Danielsona i Lanczosa (1942), kojom je pokazano da se diskretna Fourierova transformacija reda n može predstaviti sumom dve diskretne Fourierove transformacije reda $\frac{n}{2}$. Naime, ako je $n = 2m$, imamo da je

$$W_n^2 \equiv \left(e^{\imath\frac{2\pi}{n}}\right)^2 = e^{\imath\frac{2\pi}{n/2}} = e^{\imath\frac{2\pi}{m}} \equiv W_m, \tag{37}$$

što omogućava da se n-dimenzioni vektor $\mathbf{y} = F_n\mathbf{x}$ (F_n je Fourierova matrica (33) dimenzije n) generiše pomoću dva m-dimenziona vektora $\mathbf{y}^e$ i $\mathbf{y}^o$,

$$\mathbf{y}^e = F_m\mathbf{x}^e, \qquad \mathbf{y}^o = F_m\mathbf{x}^o,$$

gde je $\mathbf{x} = (x_0, x_1, \dots, x_{n-1})^T$, $\mathbf{x}^e = (x_0, x_2, \dots, x_{n-2})^T$ i $\mathbf{x}^o = (x_1, x_3, \dots, x_{n-1})^T$. j-ta komponenta vektora $\mathbf{y}$ je, s obzirom na (37),

$$\begin{aligned} y_j &= \sum_{k=0}^{n-1} W_n^{kj}x_k = \sum_{k=0}^{m-1} W_n^{2kj}x_{2k} + \sum_{k=0}^{m-1} W_n^{(2k+1)j}x_{2k+1} \\ &= \sum_{k=0}^{m-1} W_m^{kj}x_k^e + W_n^j \sum_{k=0}^{m-1} W_m^{kj}x_k^o = y_j^e + W_n^j y_j^o, \qquad j = 0,\dots,m-1 \end{aligned} \tag{38}$$

Dakle, prvih m komponenti vektora $\mathbf{y}$ računamo iz veze (38). Preostalih m komponenti y_{m+j}, $j = 0,\dots,m-1$, ćemo, vodeći računa da je zbog (37)

$$W_m^{k(m+j)} = W_m^{km}W_m^{kj} = W_m^{kj}, \qquad W_n^{m+j} = W_m^{\frac{m}{2}}W_n^j = -W_n^j,$$

dobiti kada u (38) umesto j stavimo $m + j$:

$$y_{m+j} = y_j^e - W_n^j y_j^o \qquad j = 0,\dots,m-1. \tag{39}$$

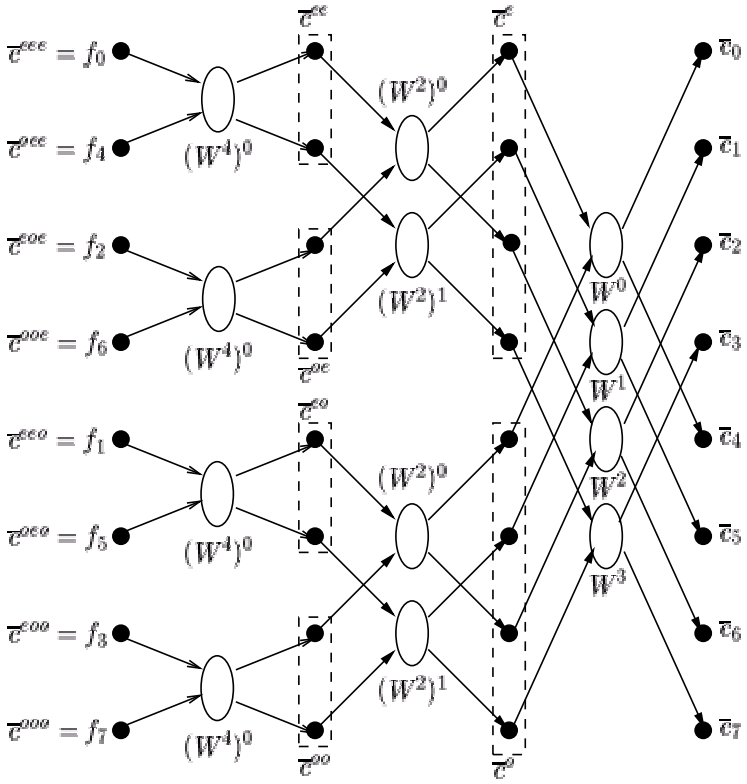
Vektor $\mathbf{y}$ treba još pomnožiti sa $\frac{1}{n}$ da bi, prema (36), predstavljao diskretnu Fourierovu transformaciju vektora $\mathbf{x}$.

Ponavljajući ovu ideju, izražavamo Fourierove transformacije reda m pomoću transformacija reda $\frac{m}{2}$, itd. U slučaju da je $n = 2^l$, gde je l prirodan broj, opisanim algoritmom dolazimo do transformacija reda $m = 1$. U tom slučaju je FFT najefikasnija – polazeći od samih komponenti vektora $\mathbf{x}$, koje su identične Fourierovim transformacijama reda jedan,

$$y_0^{eoeee\ldots oe} = x_k,$$

formulama (38) i (39) dobijamo transformacije reda $2, 4, 8, \ldots, n$.

Dakle, ukoliko su poznati koeficijenti diskretne Fourier-ove transformacije vektora parnih, odn. neparnih elemenata polaznog vektora, moguće je kombinovanjem odgovarajućih elemenata izračunati koeficijente Fourier-ove transformacije polaznog vektora. Pritom, to kombinovanje se može predstaviti tzv. "butterfly" strukturom koja je prikazana na slici 4.2.



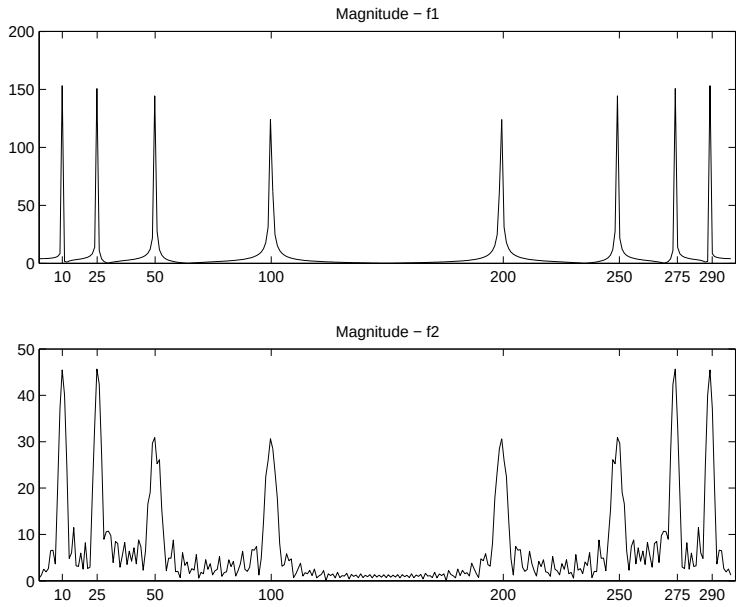
Slika 4.2: Butterfly struktura.

Broj računskih operacija koje je potrebno izvršiti da bi se algoritam realizovao je u ovom slučaju $\frac{1}{2}n \log_2 n = \frac{1}{2}nl$. Većina algoritama kojima se praktično realizuje FFT dopušta mogućnost da n bude proizvoljan složen broj. FFT nema nikakvog efekta ako je n prost broj.

4.6 Talasići

Fourierov red po trigonometrijskim funkcijama je nepodesan za predstavljanje funkcija sa diskontinuitetima i oštrim pikovima, jer predstavlja globalnu reprezentaciju funkcije po x (ili t, nezavisno promenljiva je obično vreme), a lokalnu po frekvencijama (Primer 3, f-ja $\delta(x)$).

U praksi često funkcija $f(x)$ nije zadata za svako x, već u obliku niza $f(n)$, tj. samo za diskretne vrednosti x. Takva funkcija se obično naziva *signal*. Dvodimenzioni signal naziva se slika. U Fourierovoj reprezentaciji nije moguće vremenski ograničiti pojavu neke frekvencije u složenom signalu, već se interferencijom sa drugim frekvencijama anulira njen efekat u signalu u nekom vremenskom periodu. Na primer, ako se neka nota jednom pojavi u muzičkoj temi, pri harmonijskoj (Fourierovoj) analizi će se pojaviti odgovarajuća frekvencija sa određenom amplitudom i fazom, ali ne lokalizovana u vremenu. Da li se ona čuje ili ne podešava se interferencijom pomoću bliskih frekvencija. Dakle, matematički je zapis teme Fourierovom reprezentacijom korektan, ali se odgovarajuća frekvencija (harmonik) pojavljuje u harmonijskoj analizi, iako fizički nije prisutna u signalu .



Slika 4.3: Fourierova analiza stacionarnog i nestacionarnog signala

PRIMER 6. Slika 4.3 predstavlja Fourierove koeficijente dva signala koji sadrže ista četiri harmonika. U stacionarnom signalu $f1(x)$ sva četiri harmonika traju sve vreme, a u nestacionarnom signalu $f2(x)$, u svakom od vremenskih intervala pojavljuje se samo jedan od tih harmonika.

$$f1(x) = \cos(2\pi * 10 * x) + \cos(2\pi * 25 * x) + \cos(2\pi * 50 * x) + \cos(2\pi * 100 * x)$$

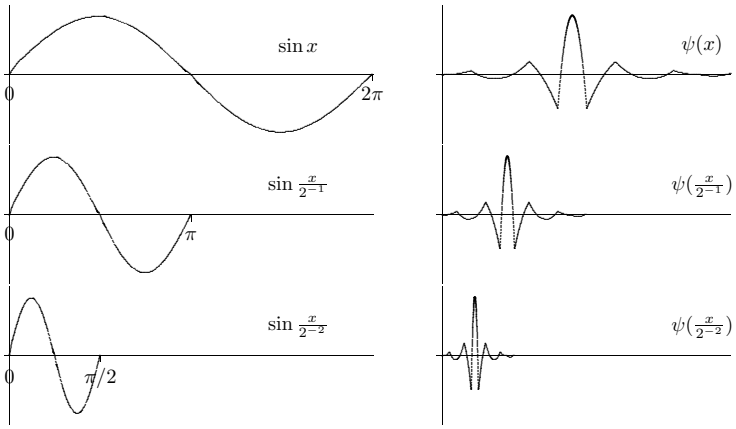
$$f2(x) = \begin{cases} \cos(2\pi * 10 * x), & 0 < x < 300 \\ \cos(2\pi * 25 * x), & 300 < x < 600 \\ \cos(2\pi * 50 * x), & 600 < x < 800 \\ \cos(2\pi * 100 * x), & 800 < x < 1000 \end{cases}$$

Slika jasno pokazuje da su Fourierove slike ovih signala vrlo slične, jedino se u slučaju drugog signala pojavljuje šum zbog prekidnosti funkcije.

Stoga se javlja potreba i za lokalizacijom po vremenu. Ideja koja leži u osnovi tzv. *kratkotrajne (ili prozor) Fourierove transformacije (STFT=Short Time Fourier Transformation)* je da se signal izdeli na male segmente po vremenu, pa se vrši harmonijska analiza svakog od ovih kratkih signala. Što je prozor uži bolja je vremenska, a lošija frekventna rezolucija, i obrnuto (princip neodređenosti). Beskonačna dužina prozora definiše standardnu Fourierovu transformaciju, koja daje savršenu frekventnu rezoluciju. Ekstremni slučajevi su signal $\sin x$ koji predstavlja jednu frekvenciju u beskonačnom intervalu, i Dirakova funkcija $\delta(x)$, koja predstavlja beskonačno mnogo frekvencija u jednom vremenskom trenutku. Segmentiranje signala se vrši pomoću prozor (window) funkcije, čija je širina jednaka segmentu u kome se signal može smatrati stacionarnim. Najjednostavnija prozor funkcija je box funkcija (karakteristična funkcija intervala), koja je jednaka jedan na segmentu, a van njega je jednaka nuli. Moguć je i drugačiji izbor – na primer, Gaussovo zvono $e^{-at^2/2}$, gde je a širina prozora. Kratkotrajna Fourierova transformacija signala se računa kao Fourierova transformacija proizvoda prozor funkcije i datog signala. Očigledno je da dobijena funkcija zavisi od frekvencije, ali i od vremena koje određuje poziciju prozora. Njen nedostatak je ista dužina vremenskih segmenata, bez obzira na oblik signala.

Sistem trigonometrijskih funkcija na kome se zasnivaju i Fourierova i kratkotrajna Fourierova transformacija očigledno ne može da zadovolji postavljene zahteve. Potrebno je u reprezentaciji koristiti bazisne funkcije sa kompaktnim nosačem (ograničenog vremenskog trajanja) koje su oscilatorne. Funkcija sa ovim osobinama naziva se *talasić* (eng. *wavelet*) - nazvana je talas zbog oscilatorne prirode, a mali zbog kratkog trajanja (slika 4.4).

Transformacija talasićima (WT=Wavelet Transformation) omogućava promenljivu rezoluciju po vremenu - više frekvencije su date sa boljom vremenskom rezolucijom (oštri, kratkotrajni pikovi), dok su niže frekvencije (glatke, sporo promenljive komponente signala) date sa lošijom vremenskom, ali boljom frekventnom rezolucijom. Značaj jednih i drugih možemo uočiti na primeru govora kao zvučnog signala. Ako obrišemo visokofrekventne komponente, govor će biti deformisan, ali prepoznatljiv. Ako obrišemo niskofrekventne komponente, svaki smisao se gubi. Dakle,



Slika 4.4: Fourierov bazis i bazis talasića

suštinu signala (aproksimaciju) definišu niskofrekventne komponente, a detalje visokofrekventne komponente.

Ovim dolazimo do pojma *multirezolucije*. To je dekompozicija Hilbertovog prostora $\mathcal{L}_2(R)$ na niz zatvorenih potprostora $\{\mathcal{V}_j\}_{j\in Z}$ takvih da je

$$(a) \qquad \cdots \subset \mathcal{V}_2 \subset \mathcal{V}_1 \subset \mathcal{V}_0 \subset \mathcal{V}_{-1} \subset \mathcal{V}_{-2} \subset \ldots$$

$$(b) \qquad \cap_{j\in Z}\, \mathcal{V}_j = \{0\}, \qquad \overline{\cup_{j\in Z}\mathcal{V}_j} = \mathcal{L}_2(R)$$

$$(c) \qquad \forall f \in \mathcal{L}_2(R) \ \text{ i } \ \forall j \in Z, \qquad f(x) \in \mathcal{V}_j \quad \Leftrightarrow \quad f(2x) \in \mathcal{V}_{j-1}$$

$$(d) \qquad \forall f \in \mathcal{L}_2(R) \ \text{ i } \ \forall k \in Z, \qquad f(x) \in \mathcal{V}_0 \quad \Leftrightarrow \quad f(x-k) \in \mathcal{V}_0$$

$$(e) \qquad \exists \varphi \in \mathcal{V}_0 \ \text{ tako da je } \ \{\varphi(x-k)\}_{k\in Z} \ \text{ Rieszov bazis u } \ \mathcal{V}_0.$$

Specijalno, u uslovu (e) se bazis može izabrati tako da bude ortonormirani bazis potprostora $\mathcal{V}_0$.

Funkcija $\varphi(x)$ naziva se *funkcija skaliranja*. Ona očigledno, s obzirom da je $\mathcal{V}_0 \subset \mathcal{V}_{-1}$, zadovoljava *dilatacionu jednačinu*

$$\varphi(x) = \sum_{k=0}^{N-1} c(k)\varphi_{-1,k}(x) = \sum_{k=0}^{N-1} c(k)\sqrt{2}\varphi(2x-k) \tag{40}$$

Radi definisanja jedinstvenog rešenja jednačine(40) (jer je i svaka funkcija $K\varphi(x)$, $K = const.$ takođe rešenje), zahteva se da to rešenje bude normalizovano, tj. da je

$$\int \varphi(x)\, dx = 1.$$

Ovaj zahtev dovodi integraljenjem jednačine (40)

$$\sqrt{2}\int \varphi(x)\,dx = \sum_{k=0}^{N-1} c(k) \int \varphi(2x-k)\,d(2x-k).$$

do uslova koji koeficijenti $c(k)$ treba da zadovoljavaju,

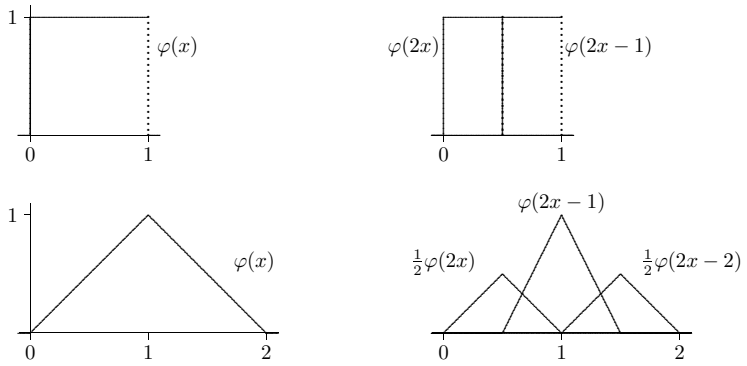
$$\sum_{k=0}^{N-1} c(k) = \sqrt{2}. \tag{41}$$

PRIMER 7. (slika 4.5)
(*i*) Ako je $c(0) = \sqrt{2}$ i $c(k) = 0$, $k \neq 0$ funkcija skaliranja je δ-funkcija,

$$\varphi(x) = \delta(x), \qquad \text{jer zadovoljava jednačinu} \quad \delta(x) = 2\delta(2x).$$

(*ii*) Pri izboru $c(0) = c(1) = 1/\sqrt{2}$ i $c(k) = 0$ za $k \neq 0,\ 1$, funkcija skaliranja je box funkcija,

$$\varphi(x) = \varphi(2x) + \varphi(2x-1), \qquad \varphi(x) = \begin{cases} 1, & x \in [0,1) \\ 0, & x \notin [0,1) \end{cases}$$



Slika 4.5: Box i krov funkcija skaliranja

(*iii*) Krov funkcija je funkcija skaliranja koja je određena izborom koeficijenata $c(0) = c(2) = 1/(2\sqrt{2})$ i $c(1) = 1/\sqrt{2}$, dok su ostali koeficijenti jednaki nuli,

$$\varphi(x) = \frac{1}{2}\varphi(2x) + \varphi(2x-1) + \frac{1}{2}\varphi(2x-2), \qquad \varphi(x) = \begin{cases} x, & x \in [0,1] \\ 2-x, & x \in [1,2] \\ 0, & x \notin [0,2] \end{cases}$$

Primeri pokazuju, a može se dokazati i u opštem slučaju, da je kompaktan nosač funkcije $\varphi(x)$ interval $[0, N-1]$, određen brojem sabiraka u jednačini (40).

Fukcije $\varphi_{j,k}(x) = 2^{-j/2}\varphi(2^{-j}x - k)$ obrazuju bazis potprostora $\mathcal{V}_j$. Koeficijent $2^{-j/2}$ predstavlja faktor normiranja. Indeks j određuje rastezanje ili sabijanje (dilataciju) funkcije, a indeks k pomeranje (translaciju) duž x-ose. Što je j manje,

nosač funkcije $\varphi_{jk}(x)$ je kraći, te ove funkcije opisuju brze promene (visoke frekvencije); i obrnuto, što je j veće, nosač funkcije $\varphi_{jk}(x)$ je duži, pa se ovim funkcijama opisuju spore promene (niske frekvencije).

Konačno dolazimo do talasića. Neka potprostor $\mathcal{W}_j$ predstavlja ortogonalnu dopunu potprostora $\mathcal{V}_j$ u $\mathcal{V}_{j-1}$ (s obzirom na osobinu(a)),

$$\mathcal{V}_{j-1} = \mathcal{V}_j \oplus \mathcal{W}_j.$$

Bazisne funkcije $\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k)$, $k \in Z$, potprostora $\mathcal{W}_j, j \in Z$, nazivaju se *talasići*. Oni se takođe mogu izraziti pomoću bazisa potprostora $\mathcal{V}_{j-1}$. Za $j = 0$ ova reprezentacija predstavlja jednačinu talasića

$$\psi(x) = \sum_{k=0}^{N-1} d(k)\varphi_{-1,k}(x) = \sum_{k=0}^{N-1} d(k)\sqrt{2}\varphi(2x - k) \tag{42}$$

Talasić $\psi(x)$ naziva se *talasić majka* (mother wavelet) jer se njegovim translacijama $\psi(x) \to \psi(x - k)$ i dilatacijama $\psi(x) \to \psi(2^{-j}x)$ generišu svi talasići, tj. bazisne funkcije svih potprostora $\mathcal{W}_j$, $j \in Z$.

Pokažimo da talasići i funkcija skaliranja sa svojim translacijama čine bazis prostora $\mathcal{L}_2$. Pođimo od potprostora

$$\mathcal{V}_J \oplus \mathcal{W}_J = \mathcal{V}_{J-1}, \qquad \mathcal{V}_{J-1} \oplus \mathcal{W}_{J-1} = \mathcal{V}_{J-2}.$$

Zamenom prve u drugoj jednakosti, predstavljamo $\mathcal{V}_{J-2}$ kao sumu tri uzajamno ortogonalna potprostora

$$\mathcal{V}_J \oplus \mathcal{W}_J \oplus \mathcal{W}_{J-1} = \mathcal{V}_{J-2}.$$

Dalje dodajući detalje dolazimo do potprostora $\mathcal{V}_{-(j+1)}$,

$$\mathcal{V}_J \oplus \mathcal{W}_J \oplus \mathcal{W}_{J-1} \oplus \cdots \oplus \mathcal{W}_{-j} = \mathcal{V}_{-(j+1)}. \tag{43}$$

Ako je f_j projekcija funkcije f na potprostor $\mathcal{V}_j$, onda je $\Delta f_j = f_{j-1} - f_j$ njena projekcija na $\mathcal{W}_j$. Na osnovu (43) sledi da je

$$\begin{aligned} f_{-(j+1)}(x) &= f_J + (f_{J-1} - f_J) + (f_{J-2} - f_{J-1}) + \cdots + (f_{-(j+1)} - f_{-j}) \\ &= f_J(x) + \Delta f_J(x) + \Delta f_{J-1}(x) + \cdots + \Delta f_{-j}(x), \end{aligned}$$

što predstavlja razlaganje funkcije $f_{-(j+1)}$ na aproksimaciju $f_J(x)$ i detalje $\Delta f_l(x)$, $l = J, \ldots, -j$ ("zumiranje" funkcije).

Ortogonalnost potprostora $\mathcal{W}_j$ na potprostor $\mathcal{V}_j$ je poželjna, ali ne i neophodna osobina. Ako su potprostori $\mathcal{W}_j$ i $\mathcal{V}_j$ ortogonalni, onda je $\mathcal{W}_j$ ortogonalan na sve $\mathcal{W}_k$, $k > j$, jer su svi ovi potprostori sadržani u $\mathcal{V}_j$. Uslov kompletnosti (b) je tada granični slučaj relacije (43),

$$\mathcal{V}_J \oplus \sum_{j=-\infty}^{J} \mathcal{W}_j = \mathcal{L}_2(R). \tag{44}$$

Iz (44) sledi da je

$$f(x) = f_J(x) + \sum_{-\infty}^{J} \Delta f_j(x), \qquad f(x) = \sum_{-\infty}^{\infty} \Delta f_j(x).$$

pri čemu je druga reprezentacija funkcije $f(x)$ dobijena za $J \to \infty$. Dakle, svaka funkcija se može proizvoljno dobro predstaviti svojom aproksimacijom f_J i detaljima na različitim nivoima rezolucije Δf_j (multirezolucija).

I kada potprostori $\mathcal{V}_j$ i $\mathcal{W}_j$ nisu ortogonalni, svaka funkcija f_{j-1} iz $\mathcal{V}_{j-1}$ ima jedinstveno razlaganje na $f_j + \Delta f_j$. Ovaj zaključak se koristi kod tzv. biortogonalnih talasića, gde su potprostori $\mathcal{W}_j$ ortogonalni na neke druge potprostore $\tilde{\mathcal{V}}_j$.

Pitanje je kada će funkcija skaliranja i talasići činiti ortonormirani sistem funkcija u $\mathcal{L}_2$. Zahtev da funkcija skaliranja bude ortogonalna na svoje translacije daje sledeći uslov:

(45)
$$\begin{aligned}
&\int_{-\infty}^{\infty} \varphi(x-m)\varphi(x-n)\,dx \\
&= \int \Big(\sqrt{2}\sum_k c(k)\varphi\big(2(x-m)-k\big)\Big)\Big(\sqrt{2}\sum_l c(l)\varphi\big(2(x-n)-l\big)\Big)\,dx \\
&= 2\int \Big(\sum_k c(k)\varphi(2x-2m-k)\Big)\Big(\sum_{l_1} c(l_1-2(n-m))\varphi(2x-2m-l_1)\Big)\,dx \\
&= \sum_k \sum_{l_1} c(k)c(l_1-2(n-m)) \int \varphi(2(x-m)-k)\,\varphi(2(x-m)-l_1)\,d(2x) \\
&= \sum_k c(k)c(k-2(n-m)) = \delta(n-m),
\end{aligned}$$

gde je uvedena smena $l_1 = l - 2(m-n)$.

Ortogonalnost funkcije skaliranja i talasića daje uslov

(46)
$$\begin{aligned}
&\int_{-\infty}^{\infty} \varphi(x-m)\psi(x-n)\,dx \\
&= \int \Big(\sqrt{2}\sum_k c(k)\varphi\big(2(x-m)-k\big)\Big)\Big(\sqrt{2}\sum_l d(l)\varphi\big(2(x-n)-l\big)\Big)\,dx \\
&= 2\int \Big(\sum_k c(k)\varphi(2x-2m-k)\Big)\Big(\sum_{l_1} d(l_1-2(n-m))\varphi(2x-2m-l_1)\Big)\,dx \\
&= \sum_k \sum_{l_1} c(k)d(l_1-2(n-m)) \int \varphi(2(x-m)-k)\,\varphi(2(x-m)-l_1)\,d(2x) \\
&= \sum_k c(k)d(k-2(n-m)) = 0.
\end{aligned}$$

Talasići na istom nivou rezolucije će biti ortogonalni ako je

$$\begin{aligned}&\int_{-\infty}^{\infty}\psi(x-m)\psi(x-n)\,dx\\ &=\int\Big(\sqrt{2}\sum_k d(k)\varphi\big(2(x-m)-k\big)\Big)\Big(\sqrt{2}\sum_l d(l)\varphi\big(2(x-n)-l\big)\Big)\,dx\\ &=\sum_k d(k)d(k-2(n-m))=\delta(n-m).\end{aligned}\tag{47}$$

Ortogonalnost talasića za različite nivoe rezolucije (različito j) sledi neposredno iz osobine multirezolucije. Neka je $j > J$. Tada je

$$\mathcal{V}_j \oplus \mathcal{W}_j = \mathcal{V}_{j-1} \subset \mathcal{V}_{j-2} \subset \cdots \subset \mathcal{V}_J, \qquad \mathcal{V}_J \perp \mathcal{W}_J$$

tj. $\mathcal{W}_j$ pripada potprostoru $\mathcal{V}_J$ koji je ortogonalan na potprostor $\mathcal{W}_J$, što znači da je $\mathcal{W}_j$ ortogonalan na $\mathcal{W}_J$. Treba uočiti da funkcije skaliranja sa različitih nivoa rezolucije (različito j) nisu ortogonalne, dok talasići jesu.

Uslovi (45), (46) i (47) mogu istovremeno biti ispunjeni samo ako je N parno, i ako je pri tome

$$d(k) = (-1)^k c(N-1-k), \qquad k = 0, \ldots, N-1, \quad N \text{ parno}.\tag{48}$$

PRIMER 8. Talasići pridruženi funkcijama skaliranja iz primera 7. su:
(*i*) Za $d(1) = \sqrt{2}$ i $d(k) = 0$, $k \neq 1$, talasić je $\psi(x) = \delta(x - 1/2)$ (δ-funkcija u tačkama $x = n + 1/2$, što je posledica jednakosti $\mathcal{W}_0 = \mathcal{V}_{-1} - \mathcal{V}_0$).
(*ii*) Pri izboru $d(0) = 1/\sqrt{2}$, $d(1) = -1/\sqrt{2}$ i $d(k) = 0$, $k \neq 0, 1$, jednačina talasića je

$$h(x) = h(2x) - h(2x-1).$$

Talasić koji odgovara box funkciji naziva se Haarov talasić. Prvi put su talasići i pomenuti u tezi A. Haara 1909. godine. Talasić majka je funkcija

$$h(x) = \begin{cases} 1, & x \in [0, 1/2) \\ -1, & x \in [1/2, 1) \\ 0, & x \notin [0, 1) \end{cases},$$

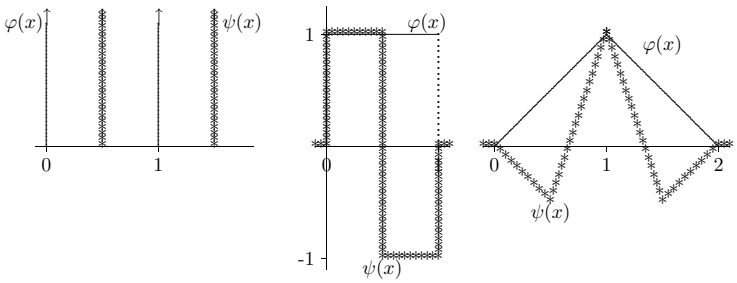
a sistem funkcija koje čine ortonormirani bazis u prostoru $\mathcal{L}_2[0,1]$ je

$$\begin{aligned} &h_0(x) = 1, && x \in [0,1)\\ &h_i(x) = 2^{j/2}\,h(2^j x - k), && x \in [k\,2^{-j}, (k+1)2^{-j})\\ &&& i = 2^j + k, \quad j \geq 0, \quad 0 \leq k < 2^j \end{aligned}$$

Najbolja srednjekvadratna aproksimacija funkcije $f(x)$ na ovom sistemu funkcija je funkcija deo po deo konstanta

$$Q_0(x) = \sum_{i=0}^{n} (f, h_i) h_i(x),$$

čije su vrednosti srednje vrednosti funkcije $f(x)$ na odgovarajućem intervalu. Prekidnost bazisnih funkcija, a time i aproksimacije je upravo nedostatak Haarovih talasića. Ipak, korišćenjem različito skaliranih Haarovih bazisnih funkcija u proučavanju malih komplikovanih detalja Brownovog kretanja fizičar Paul Levy (1930) je dobio mnogo bolje rezultate nego korišćenjem Fourierove analize.



Slika 4.6: Talasići δ, box i krov funkcije.

(*iii*) Talasići pridruženi krov funkciji određeni su koeficijentima $d(0) = d(2) = -1/(2\sqrt{2})$ i $d(1) = 1/\sqrt{2}$, dok su ostali koeficijenti jednaki nuli. Jednačina talasića je

$$\psi(x) = -\frac{1}{2}\,\psi(2x) + \psi(2x-1) - \frac{1}{2}\,\psi(2x-2).$$

Krov funkcija i njoj pridruženi talasići ne čine ortogonalni sistem funkcija.

Krov funkcija je linearni splajn. Uopšte, ako se za koeficijente $c(k)$ uzmu binomni koeficijenti, normirani tako da bude zadovoljen uslov (41), rešenje dilatacione jednačine je splajn odgovarajućeg reda. Na primer, za $c(0) = 1/8\sqrt{2}$, $c(1) = 4/8\sqrt{2}$, $c(2) = 6/8\sqrt{2}$, $c(3) = 4/8\sqrt{2}$, $c(4) = 1/8\sqrt{2}$, rešenje dilatacione jednačine je kubni B-splajn.

Ono što je važno istaći je da, za razliku od Fourierove analize koja se zasniva na jednom skupu funkcija (trigonometrijskim funkcijama), reprezentacija talasićima je moguća na beskonačno mnogo različitih bazisa. Izbor bazisa, a time i osobine aproksimacije, određen je izborom broja N i vrednostima koeficijenata $c(k)$, i zavisi od osobina problema koji želimo da analiziramo. Veliki pomak u teoriji talasića je usledio osamdesetih godina prošlog veka, kada je Ingrid Daubechies ([4]) povezala teoriju talasića sa filterima koji se koriste u obradi signala (funkcija skaliranja odgovara niskofrekventnom filteru, a talasić odgovara visokofrekventnom filteru). Tako je nastala cela klasa Daubechies ortonormiranih talasića. Na primer, za $N = 4$ koeficijenti kojima je određen ortonormirani bazis talasića su

$$d(0) = c(3) = (1-\sqrt{3})/(4\sqrt{2}), \qquad d(1) = -c(2) = -(3-\sqrt{3})/(4\sqrt{2}),$$

$$d(2) = c(1) = (3+\sqrt{3})/(4\sqrt{2}), \qquad d(3) = -c(0) = -(1+\sqrt{3})/(4\sqrt{2}).$$

Bazis je određen sledećim algoritmom. Na osnovu željenih osobina talasića (glatkost, broj iščezavajućih momenata,...) biraju se koeficijenti $c(k)$, $k = 0, \dots N-1$. Funkcija skaliranja nalazi se kao rešenje dilatacione jednačine (40). Ako je bazis ortogonalan, relacija (48) daje koeficijente jednačine talasića (42), kojom je određen talasić majka. Njegovim translacijama i dilatacijama definisan je bazis. Glavni problem u ovom algoritmu je malaženje funkcije skaliranja, tj. rešavanje dilatacione jednačine (40). Izuzev trivijalnih slučajeva, kao što su oni navedeni u primeru 7., rešenje se ne moze dobiti u analitičkom obliku. Vrednosti funkcije skaliranja računaju se u diadskim tačkama $(k2^{-j})$, čija gustina može biti proizvoljna.

Algoritmi koji se koriste za rešavanje dilatacione jednačine su:

Kaskadni algoritam. Rešenje se, ako postoji, nalazi kao granična funkcija $\varphi(x) = \lim_{i\to\infty} \varphi^{(n)}(x)$ niza funkcija

$$\varphi^{(n+1)}(x) = \sum_{k=0}^{N-1} c(k)\sqrt{2}\varphi^{(n)}(2x-k),\ i = 0, 1, \dots, \quad \varphi^{(0)}(x) \text{ je box f-ja.}$$

Algoritam zasnovan na Fourierovoj transformaciji jednačine (40)

$$\begin{aligned} \hat{\varphi}(\omega) &= \sum_{k=0}^{N-1} c(k)\sqrt{2} \int \varphi(2x-k)\, e^{\imath\omega x}\, dx \\ &= \left(\frac{1}{\sqrt{2}} \sum_{k=0}^{N-1} c(k) e^{\imath\frac{k\omega}{2}}\right) \int \varphi(y)\, e^{\imath\frac{y\omega}{2}}\, dy = H\big(\frac{\omega}{2}\big)\hat{\varphi}\big(\frac{\omega}{2}\big), \end{aligned} \tag{49}$$

gde je

$$\hat{\varphi}(\omega) = \int \varphi(y)\, e^{\imath y\omega}\, dy \qquad \text{i} \qquad H(\omega) = \frac{1}{\sqrt{2}} \sum_{k=0}^{N-1} c(k) e^{\imath k\omega}.$$

Treba uočiti da je $H(0) = 1$, s obzirom na uslov (41).

Ponavljajući transformaciju (49) u $\omega/2$, $\omega/4, \dots$ i uzimajući u obzir da je $\hat{\varphi}(0) = \int \varphi(x)\, dx = 1$, dobijamo da je

$$\hat{\varphi}(\omega) = \left(\prod_{j=1}^{n} H(\frac{\omega}{2^j})\right) \hat{\varphi}(\frac{\omega}{2^n}) \quad \stackrel{n\to\infty}{\longrightarrow} \quad \prod_{j=1}^{\infty} H(\frac{\omega}{2^j}).$$

Primenom inverzne Fourierove transformacije nalazimo funkciju skaliranja.

Algoritam zasnovan na rekurziji. Pretpostavimo da je funkcija $\varphi(x)$ zadata u celobrojnim tačkama $x = n$. Rekurzijom (40) definisane su vrednosti funkcije $\varphi(x)$ u polovinama celih brojeva. Koristeći dobijene vrednosti na isti način određujemo vrednosti funkcije $\varphi(x)$ u četvrtinama celih brojeva, i uopšte u tackama $x = k/2^j$. Vrednosti funkcije skaliranja u celobrojnim tačkama se određuju kao koordinate sopstvenog vektora definisanog jednačinom (40) za $x = 1, \dots, N-1$.

Kada je funkcija skaliranja određena, uzimajući u obzir vezu (48), talasići se određuju pomoću jednačine talasića (42).

Pošto smo odredili bazis talasića željenih osobina, sledeći korak je projektovanje funkcije u potprostor određen tim bazisom. To znači da treba odrediti koeficijente u reprezentacijama funkcije $f(x)$ oblika

$$f_j(x) = \sum_k a_{j,k}\varphi_{j,k}(x) \quad \text{ili} \quad f(x) = \sum_j \sum_k b_{j,k}\psi_{j,k}(x)$$

Piramidalni algoritam je algoritam po kome se određuju koeficijenti u aproksimaciji talasićima (*dekompozicija*), a kojim je dat i postupak kako na osnovu datih koeficijenata rekonstruisati funkciju (*rekonstrukcija*).

Ako funkcija $f_{-1}(x) \in \mathcal{V}_{-1}$, ona se može predstaviti kombinacijom bazisnih funkcija $\varphi_{-1,k}(x) = \sqrt{2}\varphi(2x-k)$ prostora $\mathcal{V}_{-1}$. Multirezolucijom se ovaj prostor razlaže na $\mathcal{V}_{-1} = \mathcal{V}_0 \oplus \mathcal{W}_0$, te se $f_{-1}(x)$ može predstaviti i kao kombinacija bazisnih funkcija $\varphi_{0,k}(x) = \varphi(x-k)$ prostora $\mathcal{V}_0$ i bazisnih funkcija $\psi_{0,k}(x) = \psi(x-k)$ prostora $\mathcal{W}_0$,

$$\begin{aligned} \sum_k a_{-1,k}\varphi_{-1,k}(x) &= \sum_k a_{0,k}\varphi_{0,k}(x) + \sum_k b_{0,k}\psi_{0,k}(x) \\ &= \sum_k a_{0,k}\varphi(x-k) + \sum_k b_{0,k}\psi(x-k) \end{aligned} \tag{50}$$

Želimo da nađemo koeficijente $a_{0,k}$ i $b_{0,k}$ ako znamo koeficijente $a_{-1,k}$. Pretpostavićemo da su bazisi ortonormirani, zbog jednostavnosti formula.

Da bi našli rekurziju, translirajmo promenljivu x u jednačinama (40) i (42) za k i stavimo $n = l - 2k$,

$$\varphi(x-k) = \sum_n c(n)\sqrt{2}\varphi(2x-2k-n) = \sum_l c(l-2k)\varphi_{-1,l}(x), \tag{51}$$

$$\psi(x-k) = \sum_n d(n)\sqrt{2}\varphi(2x-2k-n) = \sum_l d(l-2k)\varphi_{-1,l}(x), \tag{52}$$

Obe jednačine pomnožimo sa $f_{-1}(x)$ i integralimo po x,

$$\begin{aligned} \int f_{-1}(x)\varphi_{0,k}(x)\,dx &= \int f_{-1}(x)\varphi(x-k)\,dx \\ &= \sum_l c(l-2k)\int f_{-1}(x)\varphi_{-1,l}(x)\,dx, \\ \int f_{-1}(x)\psi_{0,k}(x)\,dx &= \int f_{-1}(x)\psi(x-k)\,dx \\ &= \sum_l d(l-2k)\int f_{-1}(x)\varphi_{-1,l}(x)\,dx. \end{aligned}$$

Bazisi su ortonormirani, te su Fourierovi koeficijenti $a_{j,l} = (f_{-1}, \varphi_{j,l})$, $j = -1, 0$. Uvodeći ove oznake u poslednje jednakosti, dobijamo da se koeficijenti računaju po formulama

$$a_{0,k} = \sum_l c(l-2k)a_{-1,l}, \qquad b_{0,k} = \sum_l d(l-2k)a_{-1,l} \tag{53}$$

To je osnov rekurzije. Uopšte, rekurzijom zbog koje je transformacija talasićima brza, na osnovu koeficijenata $a_{j-1,k}$ određujemo koeficijente $a_{j,k}$ i $b_{j,k}$,

TEOREMA 5. *Funkcija* $\sum a_{j-1,l}\varphi_{j-1,l}(t)$ *iz prostora* $\mathcal{V}_{j-1} = \mathcal{V}_j \oplus \mathcal{W}_j$ *ima koeficijente* $a_{j,k}$ *i* $b_{j,k}$ *po novom ortonormiranom bazisu* $\{\varphi_{j,k}(t),\, \psi_{j,k}(t)\}$,

$$a_{j,k} = \sum_l c(l-2k)a_{j-1,l}, \qquad b_{j,k} = \sum_l d(l-2k)a_{j-1,l} \tag{54}$$

U vektorskoj notaciji $\mathbf{a}_j = (\downarrow 2)C^\top \mathbf{a}_{j-1}$ *i* $\mathbf{b}_j = (\downarrow 2)D^\top \mathbf{a}_{j-1}$ *piramidalni algoritam je*

$$\begin{array}{ccccccccccc}
\mathbf{a}_{-j-1} & \overset{C^\top}{\longrightarrow} & \mathbf{a}_{-j} & \overset{C^\top}{\longrightarrow} & \mathbf{a}_{-j+1} & \longrightarrow & \cdots & \mathbf{a}_{-1} & \overset{C^\top}{\longrightarrow} & \mathbf{a}_0 \\
 & D^\top \searrow & & D^\top \searrow & & \searrow & & & D^\top \searrow & \\
 & & \mathbf{b}_{-j} & & \mathbf{b}_{-j+1} & & & & & \mathbf{b}_0
\end{array}$$

DOKAZ: Za $j = 0$ formula (54) se svodi na formulu (53). Uopštenje za proizvoljno j sledi iz dilatacione jednačine

$$\begin{aligned}
\varphi_{j,k}(x) &= 2^{-j/2}\varphi(2^{-j}x - k) = 2^{-j/2}\sum_n c(n)\sqrt{2}\varphi(2^{-j+1}x - 2k - n) \\
&= \sum_l c(l-2k)\varphi_{j-1,l}(x).
\end{aligned}$$

Da bi se odredili koeficijenti $b_{j,k}$ potrebno je koristiti jednačinu talasića (42), a to znači u prethodnom izrazu potrebno je samo koeficijente c zameniti koeficijentima d. Skalarni proizvodi ovih jednačina sa $f(x)$ daju rekurzije (54) za koeficijente $a_{j,k}$ i $b_{j,k}$. ∎

U procesu rekonstrukcije, potrebno je sa bazisa $\{\varphi_{j,k}(x),\, \psi_{j,k}(x)\}$ vratiti se natrag na bazis $\{\varphi_{j-1,l}(x)\}$ (suprotan proces).

TEOREMA 6. *Koeficijenti* $a_{j-1,l}$ *se računaju pomoću koeficijenata* $a_{j,k}$ *i* $b_{j,k}$ *po formuli*

$$a_{j-1,l} = \sum_k \big(c(l-2k)a_{j,k} + d(l-2k)b_{j,k}\big).$$

Šematski prikaz inverznog piramidalnog algoritma, koristeći vektorsku notaciju, je

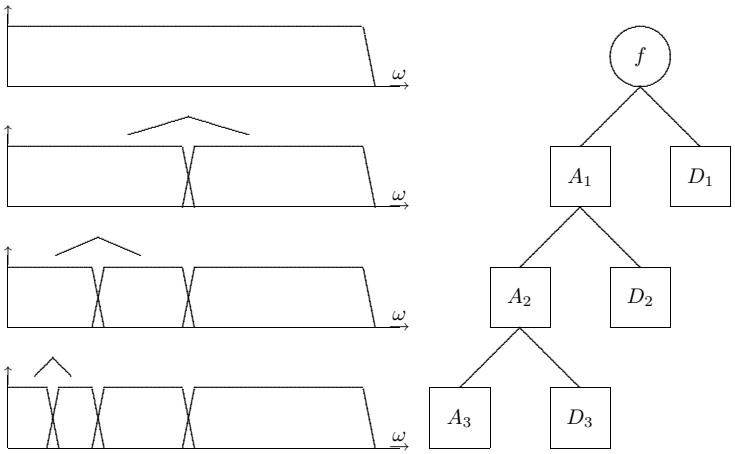
$$\begin{array}{ccccccccc}
\mathbf{a}_0 & \overset{C}{\longrightarrow} & \mathbf{a}_{-1} & \overset{C}{\longrightarrow} & \mathbf{a}_{-2} & \cdots & \mathbf{a}_{-j} & \overset{C}{\longrightarrow} & \mathbf{a}_{-j-1} \\
 & \nearrow D & & \nearrow D & & & & \nearrow D & \\
\mathbf{b}_0 & & \mathbf{b}_{-1} & & & & \mathbf{b}_{-j} & &
\end{array}$$

DOKAZ: Za $j = 0$ iz (50), (51) i (52) sledi da je

$$\sum_k a_{-1,k}\varphi_{-1,k}(x) = \sum_n a_{0,n}\varphi_{0,n}(x) + \sum_n b_{0,n}\psi_{0,n}(x)$$
$$= \sum_n a_{0,n}\Big(\sum_l c(l-2n)\varphi_{-1,l}(x)\Big) + \sum_n b_{0,n}\Big(\sum_l d(l-2n)\varphi_{-1,l}(x)\Big)$$
$$= \sum_l \Big(\sum_n \big(a_{0,n}c(l-2n) + b_{0,n}d(l-2n)\big)\Big)\varphi_{-1,l}(x)$$

Za ostale nivoe j dokaz analogno sledi. ∎

Brza transformacija talasićima (FWT). Mallat ([16]) je 1988 godine razvio efikasan postupak za brzo računanje koeficijenata a_{jk} i b_{jk}, odnosno za razlaganje signala na aproksimaciju (glatki deo) i detalje. Mallatov algoritam je, ustvari, klasična šema u obradi signala poznata pod nazivom dvokanalno slojno kodiranje korišćenjem konjugovanih kvadraturnih filtera ili kvadraturnih filtera ogledala (QMF).



Slika 4.7: Piramidalni algoritam

Taj postupak predstavlja matričnu varijantu piramidalnog algoritma. Predstavljanjem matrice *Diskretne transformacije talasićima* (DWT=Discrete Wavelet Transformation) proizvodom nekoliko retkih matrica, koristeći svojstvo samokonjugovanosti, ubrzava se realizacija piramidalnog algoritma. Potrebno je reda n računskih operacija za signal dužine n. Kreatori ovog algoritma, koji se naziva *Brza diskretna transformacija talasićima* (FWT=Fast Wavelet Transformation), su već pomenuti Stephane Malatt i Ingrid Daubechies. FWT algoritam predstavlja analogon FFT-u u Fourierovoj analizi. Za signal dužine n broj računskih operacija u FWT algoritmu je reda $O(n)$, dok je ovaj broj u FFT algoritmu reda $O(n \ln n)$. FWT algoritam je u potpunosti rekurzivan.

PRIMER 9. Neka je u četiri ekvidistantne tačke intervala [0,1) funkcija zadata svojim vrednostima $\mathbf{f} = (9, 1, 2, 0)^T$. Dakle, maksimalni broj nivoa razlaganja u ovom slučaju je $J = 2$. Reskaliranjem indeksa nivoa, početni uzorak će predstavljati nulti nivo. Radi jednostavnosti, koristićemo Haarov talasić određen box funkcijom jer je ortogonalan u odnosu na translaciju. U opštem slučaju, zbog male glatkosti Haarovih bazisnih funkcija aproksimacija sporo konvergira, te je bolje koristiti talasiće višeg reda (definisane većim brojem nenula koeficijenata $c(n)$).

Nivo 0:

$$\mathbf{a}_0 = \mathbf{f} = (9, 1, 2, 0)^T$$

Nivo 1:

$$\begin{pmatrix} \mathbf{a}_1 \\ \mathbf{b}_1 \end{pmatrix} = (\downarrow 2) \begin{pmatrix} C^T \\ D^T \end{pmatrix} \mathbf{a}_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 9 \\ 1 \\ 2 \\ 0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 10 \\ 2 \\ 8 \\ 2 \end{pmatrix}$$

jer je

$$C = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \qquad C^T = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$(\downarrow 2)C^T = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

$$D = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \qquad D^T = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$(\downarrow 2)D^T = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

Dakle, na nivou $j = 1$ aproksimacija je određena sa prve dve koordinate, a detalj sa druge dve koordinate dobijenog vektora,

$$\mathbf{a}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 10 \\ 2 \end{pmatrix} \qquad \mathbf{b}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 8 \\ 2 \end{pmatrix}$$

Nivo 2:

$$\begin{pmatrix} \mathbf{a}_2 \\ \mathbf{b}_2 \end{pmatrix} = (\downarrow 2) \begin{pmatrix} C^T \\ D^T \end{pmatrix} \mathbf{a}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 10 \\ 2 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

Stoga je na nivou $j = 2$ (poslednjem mogućem za ovaj obim uzorka)

$$\mathbf{a}_2 = (6), \qquad \mathbf{b}_2 = (4)$$

Dekompozicija signala $\mathbf{f}$ je određena koordinatama vektora $\mathbf{a}_2$, $\mathbf{b_2}$ i $\mathbf{b}_1$,

$$\mathbf{f} = \begin{pmatrix} 9 \\ 1 \\ 2 \\ 0 \end{pmatrix} = 3 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + 2 \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} + 4 \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} + 1 \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}$$

jer je

$$\mathbf{f} = a_{20}\varphi_{20} + b_{20}\psi_{20} + b_{10}\psi_{10} + b_{11}\psi_{11},$$

gde je $\mathbf{a}_j = \{a_{jk}\}$, $\mathbf{b}_j = \{b_{jk}\}$ i

$$\varphi_{20} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \psi_{20} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \quad \psi_{10} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \quad \psi_{11} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}$$

4.7 Ravnomerna aproksimacija

U mnogim slučajevima, npr. pri obradi eksperimentalnih rezultata, srednjekvadratna aproksimacija je zadovoljavajuća jer se njome izglađuju greške ulaznih podataka (nastale možda zbog grešaka merenja), te ona daje dovoljno tačnu opštu predstavu o procesu. Nekada je, međutim, potrebno da aproksimacija $Q(x)$ zadovoljava strožije uslove, npr. da na celom odsečku $[a, b]$ odstupanje funkcije $Q(x)$ od funkcije $f(x)$ ne bude veće od dozvoljenog. U ovom slučaju aproksimacija se naziva *ravnomerna aproksimacija*. Prostor $\mathcal{R}$ je $\mathcal{C}[a, b]$, tj. prostor neprekidnih funkcija na odsečku $[a, b]$, u kome je norma

$$\|f\| = \sup_{[a,b]} |f(x)|.$$

Funkcija

$$Q_\circ(x) = \sum_{j=0}^{n} c_j^\circ g_j(x)$$

je *element najbolje ravnomerne aproksimacije* za funkciju $f \in \mathcal{C}[a, b]$ po sistemu linearno nezavisnih neprekidnih funkcija $g_0(x), \ldots, g_n(x)$ ako je

$$E_n(f) \equiv \|f - Q_\circ\| \leq \|f - Q\| \tag{55}$$

za svaki generalisani polinom

$$Q(x) = \sum_{j=0}^{n} c_j g_j(x).$$

Pošto je $\mathcal{C}[a, b]$ linearni normirani prostor, prema teoremi 1 element najbolje ravnomerne aproksimacije postoji. Njegova jedinstvenost ne sledi iz teoreme 2 jer taj prostor nije strogo normiran. Uslovi pod kojima je ovaj element jedinstveno određen dati su sledećom teoremom

TEOREMA 7 (HAAR). *Da bi za proizvoljno zadatu funkciju $f \in \mathcal{C}[a,b]$ postojao jedinstveni generalisani polinom najbolje aproksimacije, potrebno je i dovoljno da funkcije $g_0(x), \dots, g_n(x)$ obrazuju Čebiševljev sistem funkcija, tj. da proizvoljan generalisani polinom po tom sistemu funkcija ima na odsečku $[a,b]$ ne više od n različitih nula.*

Dokaz ove teoreme se može naći u [3].

Najčešće se u praksi za ravnomernu aproksimaciju koriste algebarski polinomi, dakle sistem $\{g_k(x)\}$ je sistem funkcija $1, x, \dots, x^n$. Element najbolje ravnomerne aproksimacije je polinom $Q_\circ(x)$ stepena n koji zadovoljava uslov (55) za proizvoljan polinom $Q(x)$ stepena n. Ocenu veličine $E_n(f)$ najbolje ravnomerne aproksimacije polinomom, pod određenim pretpostavkama, daje sledeća teorema

TEOREMA 8 (DE LA VALLÉE POUSSIN). *Neka postoje $n+2$ tačke odsečka $[a,b]$ $x_0 < \cdots < x_{n+1}$ takve da je*

$$sign\Big(\big(f(x_i) - Q(x_i)\big)(-1)^i\Big) = const,$$

tj. da pri prelazu od tačke x_i ka tački x_{i+1} veličina $f(x)-Q(x)$ menja znak. Tada je

$$E_n(f) \geq \mu \equiv \min_{i=0,\dots,n+1} |f(x_i) - Q(x_i)|.$$

DOKAZ: Ako je $\mu = 0$ tvrđenje očigledno važi, jer je po definiciji (50) $E_n(f)$ nenegativna veličina. Neka je $\mu > 0$ i pretpostavimo da tvrđenje nije tačno, tj. neka je

$$\|f - Q_\circ\| \equiv E_n(f) < \mu.$$

Tada je u tačkama x_i

$$\begin{aligned} sign\big(Q(x_i) - Q_\circ(x_i)\big) &= sign\Big(\big(Q(x_i) - f(x_i)\big) - \big(Q_\circ(x_i) - f(x_i)\big)\Big) \\ &= sign\big(Q(x_i) - f(x_i)\big), \end{aligned}$$

jer je, s obzirom na pretpostavku, prvi sabirak veći po modulu od drugog. To znači, prema pretpostavci teoreme, da polinom $Q(x) - Q_\circ(x)$ stepena ne višeg od n menja znak $n+2$ puta, odnosno da ima $(n+1)$-nu nulu, što je nemoguće. Dakle, mora biti $E_n(f) \geq \mu$, što je i trebalo dokazati. ∎

Koristeći ovu teoremu, dokažimo sledeću važnu teoremu

TEOREMA 9 (ČEBIŠEV). *Da bi polinom $Q(x)$ bio polinom najbolje ravnomerne aproksimacije neprekidne funkcije $f(x)$, potrebno je i dovoljno da na odsečku $[a,b]$ postoji bar $n+2$ tačaka $x_0 < \cdots < x_{n+1}$ takvih da je*

$$f(x_i) - Q(x_i) = \alpha(-1)^i \|f - Q\|, \qquad i = 0, \dots, n+1,$$

pri čemu je $\alpha = 1$ ili $\alpha = -1$ istovremeno za sve i.

Tačke $x_0, \dots, x_{n+1}$, koje zadovoljavaju uslove teoreme, nazivaju se *tačkama Čebiševljeve alternanse.*

DOKAZ: Dokažimo prvo da je uslov dovoljan, tj. da je $Q(x)$ polinom najbolje ravnomerne aproksimacije, ukoliko postoji bar $n+2$ tačaka Čebiševljeve alternanse. Ako uvedemo oznaku

$$L = \|f - Q\|,$$

onda je prema pretpostavci teoreme

$$|f(x_i) - Q(x_i)| = L, \qquad i = 0, \ldots, n+1,$$

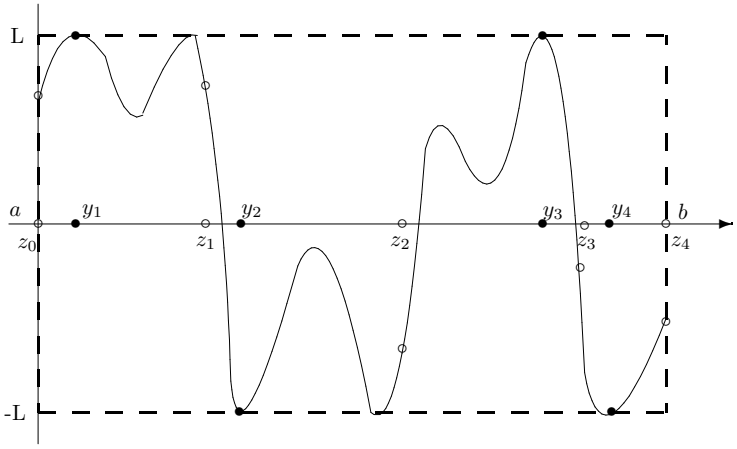
te je

$$\mu = L = \min_{i=0,\ldots,n+1} |f(x_i) - Q(x_i)|.$$

Kako je na osnovu teoreme 8. $L \leq E_n(f)$, a zbog (55) $L \geq E_n(f)$, to je

$$L = \|f - Q\| = E_n(f),$$

što znači da je $Q(x)$ polinom najbolje ravnomerne aproksimacije za $f(x)$.



Slika 4.8: Tačke Čebiševljeve alternanse.

Sada dokazujemo da je uslov potreban, tj. da iz pretpostavke da je $Q(x)$ polinom najbolje ravnomerne aproksimacije sledi da postoje bar $n+2$ tačke Čebiševljeve alternanse. Označimo sa y_1 (slika 4.8) donju granicu tačaka $x \in [a, b]$, takvih da je $|f(x) - Q(x)| = L$. Zbog neprekidnosti funkcije $f(x) - Q(x)$ je

$$|f(y_1) - Q(y_1)| = L;$$

radi određenosti, uzmimo da je

$$f(y_1) - Q(y_1) = L.$$

Dalje, neka je y_2 donja granica svih tačaka $x \in (y_1, b]$ u kojima je $f(x) - Q(x) = -L$,

$$f(y_2) - Q(y_2) = -L,$$

uopšte, y_{k+1} donja granica svih tačaka $x \in (y_k, b]$ u kojima je $f(x)-Q(x) = (-1)^k L$,

$$f(y_{k+1}) - Q(y_{k+1}) = (-1)^k L.$$

Neka je poslednja ovako određena tačka y_m, pri čemu je ili $y_m = b$ ili je y_m takvo da ni za jedno $x \in (y_m, b]$ nije $f(x) - Q(x) = (-1)^m L$. Tačke y_k su upravo tačke Čebiševljeve alternanse.

Ako je $m \geq n + 2$, tvrđenje teoreme je dokazano.

Pretpostavimo da je $m < n + 2$. Zbog neprekidnosti funkcije $f(x) - Q(x)$, za svako k, $k = 2, \ldots, m$, može se naći tačka z_{k-1} takva da je

$$|f(x) - Q(x)| < L, \qquad \text{za } z_{k-1} \leq x < y_k.$$

Ovom nizu tačaka pridružimo još tačke $z_0 = a$ i $z_m = b$. S obzirom na način kako su određene tačke z_k, na odsečcima $[z_{k-1}, z_k]$, $k = 1, \ldots, m$, postoje tačke (a to su bar y_k) u kojima je

$$f(x) - Q(x) = (-1)^{k-1} L,$$

i nema tačaka u kojima je

$$f(x) - Q(x) = (-1)^k L.$$

Tačaka z_k ima $m + 1$, što znači, obzirom na pretpostavku, da ih je najviše $n + 2$. Stoga su polinomi $P(x)$ i $Q_\epsilon(x)$,

$$P(x) = \prod_{j=1}^{m-1} (z_j - x), \qquad Q_\epsilon(x) = Q(x) + \epsilon P(x),$$

gde je $\epsilon > 0$, najviše stepena n.

Analizirajmo ponašanje razlike

$$f(x) - Q_\epsilon(x) = f(x) - Q(x) - \epsilon P(x)$$

na svakom od odsečaka $[z_{k-1}, z_k]$, $k = 1, \ldots, m$. Kada $x \in [z_0, z_1)$, polinom $P(x)$ je strogo pozitivna funkcija, te je

$$f(x) - Q_\epsilon(x) \leq L - \epsilon P(x) < L. \tag{56}$$

Ova ocena važi i u tački z_1, desnom kraju posmatranog intervala, obzirom da je $P(z_1) = 0$, a tačka z_1 je izabrana upravo tako da uslov $|f(z_1) - Q(z_1)| < L$ bude ispunjen. Sa druge strane, na tom odsečku je $f(x) - Q(x) > -L$, jer je prva tačka u kojoj ova razlika dostiže vrednost $-L$ tačka y_2. Stoga, ako izaberemo ϵ dovoljno malo, npr.

$$\epsilon < \epsilon_1 = \frac{\min_{x \in [z_0, z_1]} |f(x) - Q(x) + L|}{\max_{x \in [z_0, z_1]} |P(x)|},$$

biće

$$f(x) - Q_\epsilon(x) > -L. \tag{57}$$

Objedinjujući ocene (56) i (57), imamo da je za dovoljno mali pozitivan broj $\epsilon < \epsilon_1$ i $x \in [z_0, z_1]$

$$|f(x) - Q_\epsilon(x)| < L. \tag{58}$$

Analogna analiza na ostalim odsečcima $[z_{k-1}, z_k]$, $k = 2, \ldots, m$ takođe dovodi do ocene (58) za odgovarajuće ϵ_k, te ocena (58) važi na celom intervalu $[a, b]$ ukoliko je ϵ manje od svake konstante ϵ_k, $k = 1, \ldots, m$. To, pak, znači da je

$$\|f - Q_\epsilon\| < L,$$

što je u suprotnosti sa pretpostavkom da je $Q(x)$ polinom najbolje ravnomerne aproksimacije stepena n za funkciju $f(x)$. Dakle, pretpostavka da je broj tačaka Čebiševljeve alternanse $m < n + 2$ je neodrživa.

Time je teorema u potpunosti dokazana. ∎

Korišćenjem Čebiševljeve teoreme, može se dokazati jedinstvenost polinoma najbolje ravnomerne aproksimacije.

TEOREMA 10. *Polinom najbolje ravnomerne aproksimacije neprekidne funkcije je jedinstven.*

DOKAZ: Pretpostavimo da postoje dva polinoma $Q_1(x)$ i $Q_2(x)$ stepena n najbolje ravnomerne aproksimacije za $f(x)$,

$$Q_1(x) \neq Q_2(x), \qquad \|f - Q_1\| = \|f - Q_2\| \equiv E_n(f).$$

Tada je

$$\|f - \tfrac{Q_1+Q_2}{2}\| \leq \|\tfrac{f-Q_1}{2}\| + \|\tfrac{f-Q_2}{2}\| = E_n(f),$$

tj. i polinom $\frac{1}{2}(Q_1 + Q_2)$ je takođe polinom najbolje ravnomerne aproksimacije za $f(x)$. Ako su $x_0, \ldots, x_{n+1}$ tačke Čebiševljeve alternanse tog polinoma, onda je

$$|\tfrac{Q_1(x_k)+Q_2(x_k)}{2} - f(x_k)| = E_n(f), \qquad k = 0, \ldots, n+1,$$

odnosno

$$|\big(Q_1(x_k) - f(x_k)\big) + \big(Q_2(x_k) - f(x_k)\big)| = 2E_n(f). \tag{59}$$

Kako je $|Q_j(x_k) - f(x_k)| \leq \sup_x |Q_j(x) - f(x)| \equiv E_n(f)$, $j = 1, 2$, to je (59) moguće samo ako je

$$Q_1(x_k) - f(x_k) = Q_2(x_k) - f(x_k) = \pm E_n(f),$$

odakle sledi da je

$$Q_1(x_k) = Q_2(x_k), \qquad k = 0, \ldots, n+1.$$

Dobili smo da su dva različita polinoma stepena n jednaka u $n + 2$ tačke, što je nemoguće. Stoga je pretpostavka da postoje dva različita polinoma najbolje aproksimacije za funkciju $f(x)$ neodrživa. ∎

Jedinstvenost polinoma najbolje aproksimacije sledi i iz teoreme Haara (teorema 7), jer je sistem funkcija $1, x, \ldots, x^n$ Čebiševljev sistem funkcija.

Svaka neprekidna funkcija se može proizvoljno tačno aproksimirati polinomom.

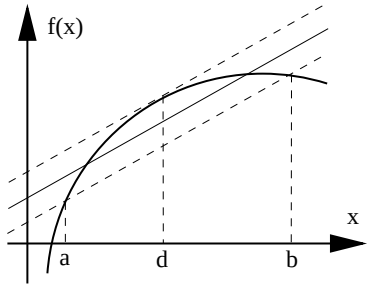
TEOREMA 11. (WEIERSTRASS). *Ako je $f(x) \in \mathcal{C}[a,b]$, tada za proizvoljno $\epsilon > 0$ postoji polinom $Q(x)$ takav da je*

$$|f(x) - Q(x)| < \epsilon \qquad \forall x \in [a,b].$$

Dokaz ove teoreme se može naći u [3].

Polinom najbolje ravnomerne aproksimacije je mnogo teže odrediti nego polinom najbolje srednjekvadratne aproksimacije, jer norma nije definisana skalarnim proizvodom. U izvesnoj meri konstruktivan algoritam za nalaženje polinoma najbolje ravnomerne aproksimacije daje Čebiševljeva teorema.

PRIMER 10. U slučaju aproksimacije konveksne funkcije pravom, tačke Čebiševljeve alternanse su krajevi intervala i tačka u kojoj razlika funkcije i njene aproksimacije dostiže maksimum. Stoga je polinom ravnomerne aproksimacije prvog stepena prava koja je paralelna sa sečicom određenom tačkama $(a, f(a))$ i $(b, f(b))$, a koja prolazi kroz sredinu rastojanja između ove sečice i njoj paralelne tangente na krivu $f(x)$ (slika 4.9).



Slika 4.9: Ravnomerna aproksimacija konkavne funkcije pravom.

Na primer, za element najbolje ravnomerne aproksimacije $Q(x) = a+bx$ funkcije $f(x) = |x|$ na odsečku $[-1,5]$, tačke Čebiševljeve alternanse su $x_0 = -1$, $x_1 = 0$ i $x_2 = 5$. Iz Čebiševljeve teoreme je

$$\begin{aligned} f(-1) - Q(-1) &= 1 - a + b = \alpha L \\ f(0) - (0) &= \quad -a \quad\ = -\alpha L \\ f(5) - Q(5) &= 5 - a - 5b = \alpha L, \end{aligned}$$

što daje $a = \frac{5}{6}$, $b = \frac{2}{3}$, $\alpha = 1$, $L = \frac{5}{6}$, te je $Q(x) = \frac{1}{6}(4x+5)$, $\quad E_1(f) = \frac{5}{6}$.

Polinom najbolje ravnomerne aproksimacije se u opštem slučaju određuje iterativnim postupkom. Problem se može pojednostaviti ako je neprekidna funkcija $f(x)$ parna ili neparna u odnosu na sredinu odsečka $[a,b]$, jer je tada i polinom parna, tj. neparna, funkcija u odnosu na sredinu odsečka.

Polinomi Čebiševa. Poseban značaj u ravnomernoj aproksimaciji imaju polinomi Čebiševa $T_n(x)$. Kao što je pomenuto u odeljku 3 ovoga poglavlja, ovi polinomi čine ortogonalan sistem polinoma na intervalu $[-1,1]$ u odnosu na težinsku

funkciju $p(x) = \frac{1}{\sqrt{1-x^2}}$,

$$(T_n\,,\,T_m) \equiv \int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\,dx = \begin{cases} 0, & \text{za } m \neq n \\ \frac{\pi}{2}, & \text{za } m = n \neq 0 \\ \pi, & \text{za } m = n = 0. \end{cases}$$

Jedan od načina njihovog zadavanja na intervalu $[-1, 1]$ je

$$T_n(x) = \cos(n \arccos x). \tag{60}$$

Koristeći trigonometrijsku transformaciju

$$\cos\big((n+1)\arccos x\big) + \cos\big((n-1)\arccos x\big) = 2\cos(n\arccos x)\cos(\arccos x)$$

i oznaku (60), dobijamo rekurentnu relaciju kojom su takođe definisani Čebiševljevi polinomi

$$\begin{aligned} T_0(x) &= 1, \qquad T_1(x) = x \\ T_{n+1}(x) &= 2x\,T_n(x) - T_{n-1}(x), \quad n = 1, 2, \ldots . \end{aligned} \tag{61}$$

Iz rekurentne relacije (61) matematičkom indukcijom se neposredno dokazuje da je koeficijent uz x^n polinoma $T_n(x)$ jednak 2^{n-1}, i da su polinomi $T_{2n}(x)$ parne, a $T_{2n+1}(x)$ neparne funkcije za svaki prirodan broj n. Rešavanjem diferencijske jednačine (61), dobija se još jedan izraz za Čebiševljeve polinome

$$T_n(x) = \frac{(x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^n}{2}$$

Iz (60) sledi da su koreni polinoma $T_n(x)$

$$x_k = \cos\tfrac{(2k+1)\pi}{2n}, \qquad k = 0, \ldots, n-1, \tag{62}$$

a tačke ekstrema

$$x_k = \cos\tfrac{k\pi}{n}, \qquad T_n(x_k) = (-1)^k, \qquad k = 0, \ldots, n. \tag{63}$$

Ovi polinomi su posebno značajni za ravnomernu aproksimaciju, jer se pomoću njih definišu *polinomi najmanjeg odstupanja od nule*

$$\overline{T}_n(x) = 2^{1-n} T_n(x). \tag{64}$$

To su polinomi sa koeficijentom jedan uz najviši stepen koji imaju sledeću osobinu:

LEMA 3. *Ako je $P_n(x)$ polinom sa koeficijentom jedan uz najviši stepen, onda je*

$$\max_{x\in[-1,1]} |P_n(x)| \geq \max_{x\in[-1,1]} |\overline{T}_n(x)| = 2^{1-n}.$$

DOKAZ: Pretpostavimo suprotno tvrđenju leme, tj. da postoji polinom $P_n(x)$ sa koeficijentom jedan uz najviši stepen, takav da je $|P_n(x)| < 2^{1-n}$ za svako x. Tada je u tačkama ekstrema (63) polinoma $T_n(x)$, obzirom na (64),

$$\operatorname{sign}\big(\overline{T}_n(x_k) - P_n(x_k)\big) = \operatorname{sign}\big((-1)^k 2^{1-n} - P_n(x_k)\big) = (-1)^k, \qquad k = 0, \ldots, n.$$

To znači da polinom $\overline{T}_n(x) - P_n(x)$ stepena $n-1$ menja znak u $(n+1)$-oj tački x_k, $k = 0, \ldots, n$, te u svakom od intervala (x_{k-1}, x_k), $k = 1, \ldots, n$ ima bar jednu nulu. Dakle, došli smo do zaključka da polinom stepena $n-1$, koji nije identički jednak nuli (jer je u tačkama x_k različit od nule), ima bar n nula, što je nemoguće. Time smo dokazali da je za svaki polinom $P_n(x)$ sa koeficijentom jedan uz najviši stepen $\max_{x\in[-1,1]} |P_n(x)| \geq 2^{1-n}$, te iz (63) i (64) sledi tvrđenje leme. ∎

Pretpostavka da $x \in [-1, 1]$ u prethodnoj lemi ne umanjuje opštost zaključka, jer se linearnom smenom

$$t = \tfrac{b+a}{2} + \tfrac{b-a}{2}x, \qquad \text{tj.} \quad x = \tfrac{2t-(b+a)}{b-a} \tag{65}$$

interval $[-1, 1]$ preslikava na interval $[a, b]$. Polinom $\overline{T}_n(x)$ se transformiše u polinom $\overline{T}_n(\frac{2t-(b+a)}{b-a})$ sa koeficijentom $\left(\frac{2}{b-a}\right)^n$ uz t^n. Da bi se dobio polinom najmanjeg odstupanja od nule na intervalu $[a, b]$ sa koeficijentom jedan uz najviši stepen, treba ovaj poslednji polinom pomnožiti sa $\left(\frac{b-a}{2}\right)^n$,

$$\overline{T}_n^{[a,b]}(t) = \left(\tfrac{b-a}{2}\right)^n \overline{T}_n(\tfrac{2t-(b+a)}{b-a}) = (b-a)^n 2^{1-2n} T_n(\tfrac{2t-(b+a)}{b-a}), \qquad t \in [a, b].$$

Sada je, prema lemi 3, za proizvoljan polinom $P(x)$ sa koeficijentom jedan uz najviši stepen

$$\max_{x\in[a,b]} |P_n(x)| \geq \max_{x\in[a,b]} |\overline{T}_n^{[a,b]}(x)| = (b-a)^n 2^{1-2n}. \tag{66}$$

PRIMER 11. Lagrangeov interpolacioni polinom funkcije $f(x)$ sa minimalnom greškom interpolacije za dati broj čvorova, se dobija ako se za čvorove interpolacije uzmu nule polinoma $\overline{T}_n^{[a,b]}(t)$, tj. polinoma $T_n(\frac{2t-(b+a)}{b-a})$:

$$t_k = \tfrac{b+a}{2} + \tfrac{b-a}{2} \cos \tfrac{(2k+1)\pi}{2n}, \quad k = 0, \ldots, n-1,$$

što sledi iz (62) i (65). Greška interpolacije je

$$|f(x) - L_{n-1}(x)| = |\tfrac{1}{n!} f^{(n)}(\xi) \overline{T}_n^{[a,b]}(x)| \leq \tfrac{1}{n!} \max_{x\in[a,b]} |f^{(n)}(x)| (b-a)^n 2^{1-2n}.$$

Obzirom na (66), to je optimalna ocena greške interpolacije date funkcije polinomom n-tog stepena.

Za ravnomernu aproksimaciju periodičnih neprekidnih funkcija pogodnije je, kao i u slučaju srednjekvadratne aproksimacije, koristiti Čebiševljev sistem funkcija

$$1, \sin x, \cos x, \ldots, \sin nx, \cos nx.$$

5

Sistemi linearnih jednačina

Četiri osnovna zadatka linearne algebre su:

(1) rešavanje sistema linearnih jednačina $A\mathbf{x} = \mathbf{b}$,

(2) izračunavanje determinanti matrica $det\, A$,

(3) nalaženje inverznih matrica A^{-1},

(4) određivanje sopstvenih vrednosti i sopstvenih vektora matrica, tj. nalaženje netrivijalnih rešenja sistema $A\mathbf{x} = \lambda\mathbf{x}$,

gde je A matrica, $\mathbf{x}$ i $\mathbf{b}$ su vektori, a λ je skalar.

Formalno gledano, nema poteškoća u rešavanju ovih zadataka, jer postoje egzaktni algoritmi koji teorijski dovode do tačnog rešenja. Na primer, sistem linearnih jednačina se može rešiti primenom Cramerovog pravila. Međutim, za rešavanje sistema sa n promenljivih potrebno je izvršiti $O(n!n)$ aritmetičkih operacija, što dovoljno govori o neprihvatljivosti ove metode za rešavanje sistema većih dimenzija. I u rešavanju sistema manjih dimenzija, ozbiljan nedostatak metode je veliki uticaj računske greške na konačan rezultat. Slični zaključci važe i za druge klasične metode za rešavanje zadataka linearne algebre. Stoga je primena numeričkih metoda u ovoj oblasti nužna. Razvojem računarske tehnike one doživljavaju intenzivan razvoj, s obzirom da se diskretizacijom različitih problema upravo dobijaju veliki sistemi linearnih jednačina.

Numeričke metode za rešavanje zadataka linearne algebre mogu se podeliti na dve osnovne grupe – direktne i iterativne. Direktnim metodama određuje se tačno rešenje sa konačno mnogo računskih operacija, pod pretpostavkom da su svi parametri dati tačno i da se sve računske operacije realizuju tačno. Iterativnim metodama rešenje je određeno graničnom vrednošću niza uzastopnih aproksimacija, koje se računaju nekim jednoobraznim algoritmom. Ovim metodama se mogu rešavati i sistemi nelinearnih jednačina, te će o njima biti reči u poglavlju 7.

5.1 Osnovni pojmovi i stavovi o matricama

U ovom odeljku su navedeni neki osnovni pojmovi i stavovi u vezi sa matricama, koji će biti korišćeni u izlaganju materijala o numeričkim metodama za rešavanje problema linearne algebre (poglavlja 5 i 6).

Matrica A je regularna, ukoliko je njena determinanta različita od nule. U protivnom, ona se naziva singularnom. Matrica B je slična matrici A, $A \sim B$,ukoliko postoji regularna matrica T takva da je $B = T^{-1}\,A\,T$.

Neka su sa a_{ij} označeni elementi matrice A, tj. $A = (a_{ij})$. Tada je
$A^T = (a_{ij}^T)$ transponovana matrica matrici A ako je $a_{ij}^T = a_{ji}$,
$A^* = (a_{ij}^*)$ konjugovana matrica matrici A ako je $a_{ij}^* = \overline{a}_{ji}$.

Specijalno, matrica A se naziva

Hermiteova ako je $A^* = A$,
simetrična ako je $A^T = A$,
unitarna ako je $A^* = A^{-1}$,
ortogonalna ako je $A^T = A^{-1}$,
normalna ako je $A^*\,A = A\,A^*$.

Ako je matrica A realna, pojam simetrična identičan je pojmu Hermiteova, a pojam ortogonalna pojmu unitarna matrica, jer je $A^* \equiv A^T$.

Matrica A je *pozitivno definisana* ako je Hermiteova i zadovoljava uslov da je $\mathbf{x}^* A\mathbf{x} > 0$ za svaki n-dimenzioni vektor $\mathbf{x} \neq 0$.

Sopstvene vrednosti realne ili kompleksne kvadratne matrice A reda n su one vrednosti skalara λ za koje sistem

$$A\,\mathbf{x} = \lambda\,\mathbf{x} \tag{1}$$

ima netrivijalna rešenja. Pomenuta netrivijalna rešenja nazivaju se sopstvenim vektorima. Sistem (1) se može napisati u obliku homogenog sistema

$$(A - \lambda\,I)\mathbf{x} = \mathbf{0}, \tag{2}$$

I je jedinična matrica, koji ima netrivijalno rešenje ukoliko mu je determinanta jednaka nuli,

$$D(\lambda) \equiv det(A - \lambda\,I) = 0. \tag{3}$$

$D(\lambda)$ je polinom n-tog stepena po λ i naziva se karakteristični polinom matrice A. Dakle, sopstvene vrednosti matrice su koreni njenog karakterističnog polinoma.

Navedimo neke pomoćne stavove koji se odnose na sopstvene vrednosti i sopstvene vektore matrica.

LEMA 1. *Sopstveni vektori koji odgovaraju različitim sopstvenim vrednostima su linearno nezavisni.*

LEMA 2. *Ako je λ_k koren reda n_k karakterističnog polinoma matrice A, onda njemu odgovara najviše n_k linearno nezavisnih sopstvenih vektora.*

LEMA 3. *Slične matrice imaju jednake sopstvene vrednosti.*

LEMA 4. *Matrice A, αA i A^k, $k = 1, 2, \ldots$, imaju jednake sopstvene vektore, a za sopstvene vrednosti važe relacije*

$$\lambda[\alpha A] = \alpha\,\lambda[A], \qquad \lambda[A^k] = (\lambda[A])^k.$$

LEMA 5. *Matrice A i A^T imaju jednake sopstvene vrednosti, a sopstvene vrednosti matrica A i A^* su uzajamno konjugovane.*

LEMA 6. *Za svaku Hermiteovu matricu A dimenzije $n \times n$ postoji unitarna matrica U dimenzije $n \times n$ takva da je*

$$U^*\,A\,U = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Sopstvene vrednosti λ_k, $k = 1, \ldots, n$, matrice A su realne. Vektori kolona $\mathbf{u}_k$ matrice U su sopstveni vektori matrice A, tj. $A\,\mathbf{u}_k = \lambda_k \mathbf{u}_k$. Stoga su sopstveni vektori ortogonalni i čine bazu prostora $\mathcal{C}^n$.

LEMA 7. *Za svaku matricu A dimenzije $n \times n$ postoji unitarna matrica U dimenzije $n \times n$ takva da je*

$$U^*\,A\,U = \begin{pmatrix} \lambda_1 & \ldots & * \\ & \ddots & \vdots \\ 0 & & \lambda_n \end{pmatrix},$$

pri čemu su sa $*$ označene u opštem slučaju nenula vrednosti, a λ_k, $k = 1, \ldots, n$, su sopstvene vrednosti matrice A.

Norme vektora i matrica. U vektorskom prostoru $\mathcal{C}^n$ norma se može uvesti na različite načine. Familija normi ovoga prostora definisana je izrazom

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}. \tag{4}$$

Specijalno, izrazom (4) za $p = 1$ je definisana *apsolutna* norma vektora

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|, \tag{5}$$

za $p = 2$ *euklidska* norma vektora

$$\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}, \tag{6}$$

a kada $p \to \infty$ norma $\|x\|_p$ prelazi u *uniformnu* normu vektora

$$\|\mathbf{x}\|_\infty = \max_{1\le i\le n} |x_i|. \tag{7}$$

U svakom konačno-dimenzionom prostoru, pa stoga i u $\mathcal{C}^n$, sve norme su međusobno ekvivalentne, što znači da postoje pozitivne konstante c_1 i c_2 takve da je

$$c_1\|x\|' \le \|x\|'' \le c_2\|x\|',$$

gde su $\|x\|'$ i $\|x\|''$ proizvoljne norme pomenutog prostora.

U vektorski prostor $\mathcal{C}^n$ skalarni proizvod se može uvesti na sledeći način:

$$(\mathbf{x}\,,\,\mathbf{y}) = \mathbf{y}^*\,\mathbf{x} = (\overline{y}_1,\ldots,\overline{y}_n)\cdot\begin{pmatrix} x_1\\ \vdots\\ x_n\end{pmatrix} = \sum_{i=1}^{n} x_i\overline{y}_i. \tag{8}$$

Pri tome je euklidska norma indukovana ovim skalarnim proizvodom, tj.

$$\|\mathbf{x}\|_2 = \sqrt{(\mathbf{x}\,,\,\mathbf{x})}.$$

Za normu matrice $\|A\|$ se kaže da je indukovana normom vektora $\|\mathbf{x}\|$ ako je

$$\|A\| = \sup_{\mathbf{x}\ne 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}. \tag{9}$$

Norme matrica indukovane redom apsolutnom (5), uniformnom (7) i euklidskom (6) normom vektora su:

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|, \tag{10}$$

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|, \tag{11}$$

$$\|A\|_2 = \sqrt{\max_{1\le i\le n} \lambda_i(A^*A)}, \tag{12}$$

gde je $\lambda_i(A^*A)$ i-ta sopstvena vrednost matrice A^*A. Euklidska norma matrice

$$\|A\|_s = \left(\sum_{i,j=1}^{n} |a_{ij}|^2\right)^{\frac{1}{2}}$$

nije indukovana euklidskom normom vektora (6), ali je saglasna sa njom. Norme matrice $\|A\|$ i vektora $\|\mathbf{x}\|$ su saglasne, ako je

$$\|A\mathbf{x}\| \le \|A\|\,\|\mathbf{x}\| \tag{13}$$

za svaku kvadratnu matricu i vektor dimenzije n.

Koristeći definiciju saglasnih normi matrice i vektora i definiciju sopstvenih vrednosti matrice, imamo da je

$$|\lambda| \, \|\mathbf{x}\| = \|\lambda \mathbf{x}\| = \|A\mathbf{x}\| \leq \|A\| \, \|\mathbf{x}\|,$$

odakle sledi

$$|\lambda| \leq \|A\|. \tag{14}$$

Dakle, sve sopstvene vrednosti matrice A su po modulu manje ili jednake od njene proizvoljne norme, saglasne sa nekom normom vektora. Ako je matrica A Hermiteova, a njene sopstvene vrednosti uređene,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n,$$

tada je

$$\lambda_1 = \max_{\mathbf{x} \neq 0} \frac{(A\mathbf{x}\,,\,\mathbf{x})}{(\mathbf{x}\,,\,\mathbf{x})}, \qquad \lambda_n = \min_{\mathbf{x} \neq 0} \frac{(A\mathbf{x}\,,\,\mathbf{x})}{(\mathbf{x}\,,\,\mathbf{x})}. \tag{15}$$

Uslovljenost regularne matrice A je skalar

$$cond(A) = \|A\| \cdot \|A^{-1}\|, \tag{16}$$

gde je sa $\|A\|$ označena norma matrice A indukovana nekom vektorskom normom. Za singularnu matricu je po definiciji $cond(A) = \infty$. Neposredno sledi da je

$$cond(\alpha A) = cond(A),$$

α je proizvoljan skalar. Za jediničnu matricu I je

$$cond(I) = \|I\| \cdot \|I\| = 1,$$

a za proizvoljnu regularnu matricu

$$cond(A) = \|A\| \cdot \|A^{-1}\| \geq \|A \, A^{-1}\| = \|I\| = 1.$$

Dakle, $1 \leq cond(A) \leq \infty$ i matrica je to lošije uslovljena što je $cond(A)$ veće.

PRIMER 1. Primer loše uslovljene matrice je Hilbertova matrica

$$H_n = \begin{pmatrix} 1 & 1/2 & \dots & 1/n \\ 1/2 & 1/3 & \dots & 1/(n+1) \\ \vdots & \vdots & \ddots & \vdots \\ 1/n & 1/(n+1) & \dots & 1/(2n-1) \end{pmatrix}$$

koja je to lošije uslovljena što joj je dimenzija veća. Sledeća tabela pokazuje zavisnost uslovljenosti Hilbertove matrice od njene dimenzije.

n	3	5	6	8
$cond(H_n)$	$5 * 10^2$	$5 * 10^5$	$15 * 10^6$	$15 * 10^9$

5.2 Gaussova metoda eliminacije

Gaussovom metodom eliminacije za rešavanje sistema linearnih jednačina

$$A\mathbf{x} = \mathbf{b}, \tag{17}$$

gde je A regularna kvadratna matrica dimenzije $n \times n$ i $\mathbf{b}$ n-dimenzioni vektor,

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix},$$

transformiše se u konačno mnogo koraka sistem (17) u sistem sa gornje trougaonom matricom

$$U\mathbf{x} = \mathbf{c}, \qquad U = \begin{pmatrix} u_{11} & \dots & u_{1n} \\ & \ddots & \vdots \\ 0 & & u_{nn} \end{pmatrix}, \tag{18}$$

čije rešenje je identično rešenju polaznog sistema (17). Sistem (18) se, pod pretpostavkom da je $u_{ii} \neq 0$, $i = 1, \dots, n$, direktno može rešiti,

$$x_n = \frac{c_n}{u_{nn}}, \qquad x_i = \frac{1}{u_{ii}}\big(c_i - \sum_{j=i+1}^{n} u_{ij}x_j\big), \quad i = n-1, \dots, 1.$$

Prvi korak algoritma Gaussove eliminacije sastoji se u oduzimanju prve jednačine sistema (17), pomnožene odgovarajućim množiteljima, od svih ostalih jednačina. Množitelji se određuju tako da se anulira promenljiva x_1 u svim jednačinama izuzev u prvoj, tako da je pri oduzimanju od i-te jednačine množitelj a_{i1}/a_{11}, $i = 2, \dots, n$. Očigledno je neophodno za realizaciju ovog koraka da bude $a_{11} \neq 0$. Ukoliko taj uslov u sistemu (17) nije zadovoljen, permutovanjem jednačina sistema, tj. stavljanjem na prvo mesto jednačine kod koje je $a_{p1} \neq 0$ ovaj uslov se može ispuniti. Element $a_{p1} \neq 0$ matrice A postoji, jer je po pretpostavci matrica regularna i ne može imati sve nula elemente u prvoj koloni.

Opisane operacije sa jednačinama sistema (17) možemo predstaviti matričnim operacijama na proširenoj matrici sistema

$$(A\,;\,\mathbf{b}) = \begin{pmatrix} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} & b_n \end{pmatrix}. \tag{19}$$

U prvom koraku transformiše se matrica (19) u matricu

$$(A_1\,;\,\mathbf{b}_1) = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} & b_n^{(1)} \end{pmatrix}.$$

Pri tome se eventualno vrši permutacija prve i p-te vrste, što je ekvivalentno množenju matrice (19) matricom permutacije P_1,

$$(\bar{A}\,;\,\bar{\mathbf{b}}) = P_1(A\,;\,\mathbf{b}), \qquad P_1 = \begin{pmatrix} 0 & & & & 1 & & & 0 \\ & 1 & & & & & & \\ & & \ddots & & & & & \\ & & & 1 & & & & \\ 1 & & & & 0 & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ 0 & & & & & & & 1 \end{pmatrix} \leftarrow p \tag{20}$$

i množenje matrice (20) donje trougaonom matricom sa jedinicama na dijagonali

$$(A_1\,;\,\mathbf{b}_1) = L_1(\bar{A}\,;\,\bar{\mathbf{b}}), \qquad L_1 = \begin{pmatrix} 1 & & & 0 \\ -l_{21} & 1 & & \\ & & \ddots & \\ -l_{n1} & 0 & & 1 \end{pmatrix}, \quad l_{i1} = \frac{\bar{a}_{i1}}{\bar{a}_{11}}. \tag{21}$$

Iz (20) i (21) je

$$(A_1\,;\,\mathbf{b}_1) = L_1P_1(A\,;\,\mathbf{b}). \tag{22}$$

S obzirom da su matrice P_1 i L_1 regularne,

$$P_1^{-1} = P_1, \qquad \text{i} \qquad L_1^{-1} = \begin{pmatrix} 1 & & & 0 \\ l_{21} & 1 & & \\ & & \ddots & \\ l_{n1} & 0 & & 1 \end{pmatrix},$$

rešenja sistema $A\mathbf{x} = \mathbf{b}$ i $A_1\mathbf{x} = \mathbf{b}_1$ su identična, jer

$$A\mathbf{x} = \mathbf{b} \quad \Rightarrow \quad L_1P_1A\mathbf{x} = L_1P_1\mathbf{b} \quad \Rightarrow \quad A_1\mathbf{x} = \mathbf{b}_1,$$

i obrnuto,

$$\begin{aligned} A_1\mathbf{x} = \mathbf{b}_1 \quad &\Rightarrow \quad P_1^{-1}L_1^{-1}A_1\mathbf{x} = P_1^{-1}L_1^{-1}\mathbf{b}_1 \\ &\Rightarrow \quad (L_1P_1)^{-1}A_1\mathbf{x} = (L_1P_1)^{-1}\mathbf{b}_1 \quad \Rightarrow \quad A\mathbf{x} = \mathbf{b}. \end{aligned}$$

Element $a_{p1} = \bar{a}_{11}$ naziva se *glavni element* ili *pivot*, a postupak nalaženja glavnog elementa naziva se *izbor glavnog elementa* ili *pivotiranje*. Zbog numeričke stabilnosti algoritma, obično se među svim nenula elementima bira najveći po modulu,

$$|a_{p1}| = \max_{1\le i\le n} |a_{i1}|.$$

Nalaženje pivota među elementima samo jedne kolone (ili vrste) matrice naziva se delimično pivotiranje, a potpuno pivotiranje je nalaženje najvećeg po modulu elementa cele matrice,

$$|a_{pq}| = \max_{1\le i,j\le n} |a_{ij}|.$$

Ako se u prvom koraku algoritma vrši potpuno pivotiranje, onda treba permutovati prvu i p-tu vrstu matrice, kao i prvu i q-tu kolonu.

U svakom slučaju, posle prvog koraka eliminacije dobija se proširena matrica sistema (22)

$$(A_1\,;\,\mathbf{b}_1) = \left(\begin{array}{c|ccc|c} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\ \hline 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right) = \left(\begin{array}{c|c|c} A_{11}^{(1)} & A_{12}^{(1)} & \mathbf{b}_1^{(1)} \\ \hline 0 & A_{22}^{(1)} & \mathbf{b}_2^{(1)} \end{array}\right),$$

gde je $A_{11}^{(1)}$ kvadratna matrica dimenzije jedan čiji je jedini element $a_{11}^{(1)}$, $A_{12}^{(1)}$ je matrica dimenzije $(n-1)\times 1$ čiji su elementi $a_{1j}^{(1)}$, $j = 2,\dots,n$, $A_{22}^{(1)}$ je kvadratna matrica dimenzije $(n-1)$ čiji su elementi $a_{ij}^{(1)}$, $i,j = 2,\dots,n$, $\mathbf{b}_1^{(1)}$ je vektor dimenzije jedan sa koordinatom $b_1^{(1)}$, i $\mathbf{b}_2^{(1)}$ je vektor dimenzije $(n-1)$ sa koordinatama $b_i^{(1)}$, $i = 2,\dots,n$.

Sledeći korak eliminacije se sastoji u primeni opisanog postupka na sistem dimenzije $(n-1)$, tj. na sistem sa proširenom matricom $(A_{22}^{(1)}\,;\,\mathbf{b}_2^{(1)})$. Na taj način se u svakom koraku dimenzija sistema koji se transformiše smanjuje za jedan. U

j-tom koraku proširena matrica sistema ima sledeći oblik
(23)

$$(A_j\,;\,\mathbf{b}_j)=\left(\begin{array}{ccc|ccc|c}
* & \dots & * & * & \dots & * & * \\
\vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\
0 & \dots & * & * & \dots & * & * \\
\hline
0 & \dots & 0 & * & \dots & * & * \\
\vdots & & \vdots & \vdots & & \vdots & \vdots \\
0 & \dots & 0 & * & \dots & * & *
\end{array}\right)=\left(\begin{array}{c|c|c}
A_{11}^{(j)} & A_{12}^{(j)} & \mathbf{b}_1^{(j)} \\
\hline
0 & A_{22}^{(j)} & \mathbf{b}_2^{(j)}
\end{array}\right),$$

gde $*$ označava u opštem slučaju nenula elemente, $A_{11}^{(j)}$ je gornje trougaona matrica dimenzije j, a dalje se transformiše proširena matrica $(A_{22}^{(j)}\,;\,\mathbf{b}_2^{(j)})$ dimenzije $(n-j)\times(n-j+1)$. U $(n-1)$-om koraku elementi matrice $(A_{n-1}\,;\,\mathbf{b}_{n-1})$ su gornje trougaona matrica $A_{11}^{(n-1)}$ dimenzije $(n-1)\times(n-1)$, matrica $A_{12}^{(n-1)}$ dimenzije $1\times(n-1)$, matrica $A_{22}^{(n-1)}$ dimenzije 1×1 i vektori $\mathbf{b}_1^{(n-1)}$ dimenzije $(n-1)$ i $\mathbf{b}_2^{(n-1)}$ dimenzije 1. Dakle, matrica A_{n-1} je tražena gornje trougaona matrica U, a vektor $\mathbf{b}_{n-1}$ vektor $\mathbf{c}$ desne strane sistema (18). Tako smo dobili niz matrica oblika (23)

$$(A\,;\,\mathbf{b})\quad\to\quad(A_1\,;\,\mathbf{b}_1)\quad\to\quad\dots\quad\to\quad(A_{n-1}\,;\,\mathbf{b}_{n-1})\equiv(U\,;\,\mathbf{c}),$$

povezanih međusobno rekurentnim relacijama

$$(A_j\,;\,\mathbf{b}_j)=L_jP_j(A_{j-1}\,;\,\mathbf{b}_{j-1}),\qquad j=1,\dots,n-1,$$

gde je P_j matrica permutacije, a

$$L_j=\begin{pmatrix}
1 & & & & & 0 \\
& \ddots & & & & \\
& & 1 & & & \\
& & -l_{j+1,j} & 1 & & \\
& & & & \ddots & \\
0 & & -l_{n,j} & 0 & & 1
\end{pmatrix},\qquad l_{ij}=\frac{\bar{a}_{ij}^{(j-1)}}{\bar{a}_{jj}^{(j-1)}},\quad i>j.$$

Stoga je

$$(U\,;\,\mathbf{c})=L_{n-1}P_{n-1}L_{n-2}P_{n-2}\cdots L_1P_1(A\,;\,\mathbf{b}).$$

Ako se u toku realizacije algoritma ne vrše permutacije, tj. $P_j\equiv I$, $j=1,\dots,n-1$, onda je

$$(U\,;\,\mathbf{c})=L_{n-1}L_{n-2}\cdots L_1(A\,;\,\mathbf{b}),$$

tj.

$$L_1^{-1}\cdots L_{n-1}^{-1}(U\,;\,\mathbf{c})=(A\,;\,\mathbf{b}).$$

S obzirom da je

$$L_j^{-1} = \begin{pmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ & & l_{j+1,j} & 1 & & \\ & & & & \ddots & \\ 0 & & l_{n,j} & 0 & & 1 \end{pmatrix},$$

to je

$$L \equiv L_1^{-1} \cdots L_{n-1}^{-1} = \begin{pmatrix} 1 & & & 0 \\ l_{21} & 1 & & \\ & & \ddots & \\ l_{n1} & l_{n2} & l_{n,n-1} & 1 \end{pmatrix}, \tag{24}$$

pa je

$$A = L \cdot U.$$

Trougaona dekompozicija matrice. Drugim rečima, Gaussovom metodom eliminacije je izvršena dekompozicija matrice A sistema (17), tzv. *LU dekompozicija*, na proizvod dve trougaone matrice – donje trougaone matrice L sa jedinicama na dijagonali i gornje trougaone matrice U.

Ako se u toku realizacije algoritma vrše permutacije, dobija se dekompozicija matrice $P\,A$,

$$P\,A = L\,U, \tag{25}$$

gde je P proizvod svih korišćenih matrica permutacija. Svakoj regularnoj matrici se može naći permutacija koja dopušta dekompoziciju oblika (25), iako se sama matrica ne može napisati u obliku $A = L\,U$.

PRIMER 2.

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad P\,A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = L\,U, \quad \text{gde je } L = U = I.$$

S obzirom na redosled izračunavanja i strukture matrica L_j i $(A_j\,;\,\mathbf{b}_j)$, obično se elementi matrice L_j zapisuju ispod glavne dijagonale tekuće matrice $(A_j\,;\,\mathbf{b}_j)$.

Označimo tako dobijenu matricu sa T_j,

$$(26)\qquad T_j = \left(\begin{array}{cccc|ccc|c} \underline{u_{11}} & u_{12} & \dots & u_{1j} & u_{1,j+1} & \dots & u_{1n} & c_1 \\ \underline{\lambda_{21}} & \underline{u_{22}} & & u_{2j} & & & & \\ \lambda_{31} & \lambda_{32} & \underline{\ddots} & \vdots & \vdots & & \vdots & \vdots \\ & & \ddots & u_{jj} & u_{j,j+1} & \dots & u_{jn} & c_j \\ \hline \vdots & & & \lambda_{j+1,j} & a^{(j)}_{j+1,j+1} & \dots & a^{(j)}_{j+1,n} & b^{(j)}_{j+1} \\ & & & \vdots & \vdots & & \vdots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \dots & \lambda_{nj} & a^{(j)}_{n,j+1} & \dots & a^{(j)}_{n,n} & b^{(j)}_{n} \end{array}\right).$$

Poslednja matrica ovoga niza matrica T_{n-1} sadrži ispod glavne dijagonale elemente matrice L, a na i iznad glavne dijagonale elemente matrice U.

PRIMER 3. Ilustrujmo opisani algoritam na sledećem sistemu

$$\begin{aligned} 3x_1 + x_2 + 6x_3 &= 2 \\ 2x_1 + x_2 + 3x_3 &= 7 \\ x_1 + x_2 + x_3 &= 4. \end{aligned}$$

Niz matrica T_j je

$$(A\,;\,\mathbf{b}) = \begin{pmatrix} \boxed{3} & 1 & 6 & 2 \\ 2 & 1 & 3 & 7 \\ 1 & 1 & 1 & 4 \end{pmatrix} \xrightarrow{L_1} T_1 = \begin{pmatrix} 3 & 1 & 6 & 2 \\ \frac{2}{3} & \frac{1}{3} & -1 & \frac{17}{3} \\ \frac{1}{3} & \boxed{\frac{2}{3}} & -1 & \frac{10}{3} \end{pmatrix}$$

$$\xrightarrow{P_1} \begin{pmatrix} 3 & 1 & 6 & 2 \\ \frac{1}{3} & \boxed{\frac{2}{3}} & -1 & \frac{10}{3} \\ \frac{2}{3} & \frac{1}{3} & -1 & \frac{17}{3} \end{pmatrix} \xrightarrow{L_2} T_2 = \begin{pmatrix} 3 & 1 & 6 & 2 \\ \frac{1}{3} & \frac{2}{3} & -1 & \frac{10}{3} \\ \frac{2}{3} & \frac{1}{2} & \boxed{-\frac{1}{2}} & 4 \end{pmatrix}.$$

Zaokruženi brojevi predstavljaju delimične pivote. L_j, $j = 1, 2$, su pomenute donje trougaone matrice transformacija čiji su elementi različiti od nula i jedan zapisani ispod glavne dijagonale matrice T_j, a

$$P_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Rešenje polaznog sistema $x_1 = 19, x_2 = -7, x_3 = -8$ se dobija kao rešenje trougaonog sistema

$$\begin{aligned} 3x_1 + x_2 + 6x_3 &= 2 \\ \tfrac{2}{3}x_2 - x_3 &= \tfrac{10}{3} \\ -\tfrac{1}{2}x_3 &= 4. \end{aligned}$$

Trougaone matrice

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{2}{3} & \frac{1}{2} & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} 3 & 1 & 6 \\ 0 & \frac{2}{3} & -1 \\ 0 & 0 & -\frac{1}{2} \end{pmatrix},$$

koje se očitavaju iz matrice T_2, predstavljaju elemente LU dekompozicije matrice

$$P_1 A = \begin{pmatrix} 3 & 1 & 6 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{pmatrix}.$$

Ako je poznata LU dekompozicija neke matrice A, tj. poznate su matrice L, U i P u (25), imamo da je

$$P\,A\mathbf{x} = L\,U\mathbf{x} = P\mathbf{b},$$

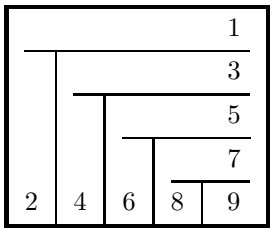
te se rešavanje sistema (17) svodi na rešavanje dva trougaona sistema

$$L\mathbf{y} = P\mathbf{b}, \qquad U\mathbf{x} = \mathbf{y}.$$

LU dekompoziciju (25) možemo izvršiti direktno, ne formirajući niz matrica T_j datih u (26). Radi jednostavnosti, pretpostavimo da nije neophodno vršiti permutacije vrsta i kolona matrice A, tj. da je $P = I$. Tada, iz (25) dobijamo n^2 veza između elemenata matrice A sa jedne strane, i elemenata matrica L i U sa druge strane. Sistem od n^2 jednačina

$$a_{ij} = \sum_{k=1}^{min(i,j)} l_{ik}u_{kj} \qquad (l_{ii} = 1)$$

ima $\frac{(n-1)n}{2}$ nepoznatih l_{ij}, $1 \le j < i \le n$, i $\frac{n(n+1)}{2}$ nepoznatih u_{ij}, $1 \le i \le j \le n$. Poredak izračunavanja l_{ij} i u_{ij} može biti različit. U Croutovoj metodi redosled izračunavanja se može šematski prikazati na sledeći način



gde šema predstavlja pomenutu matricu T_{n-1}, a brojevima je naznačen redosled određivanja njenih vrsta i kolona. Pri tome se računa po formulama

$$\begin{aligned} &u_{1k} = a_{1k}, \qquad u_{ik} = a_{ik} - \sum_{j=1}^{i-1} l_{ij}u_{jk}, \quad k = i, \dots, n \\ &l_{k1} = \frac{a_{k1}}{u_{11}}, \qquad l_{ki} = \frac{1}{u_{ii}}\big(a_{ki} - \sum_{j=1}^{i-1} l_{kj}u_{ji}\big), \quad k = i+1, \dots, n, \end{aligned} \qquad i = 2, \dots, n. \tag{27}$$

Algoritam se znatno usložnjava ako se vrši pivotiranje ($P \neq I$).

Gaussova eliminacija i LU dekompozicija se razlikuju jedino u redosledu operacija. Kako je element $a_{ik}^{(j)}$ matrice A_j, date u (23), ustvari j-ta parcijalna suma prve od formula (27),

$$a_{ik}^{(j)} = a_{ik} - \sum_{s=1}^{j} l_{is} u_{sk},$$

to znači da se u Gaussovoj eliminaciji skalarni proizvod (27) računa postepeno i međurezultati se memorišu, dok se LU dekompozicijom taj skalarni proizvod računa odjednom u celini, što može biti prednost u smislu smanjenja računske greške (ako se međurezultati ne memorišu u dvostrukoj tačnosti).

Potreban broj množenja za realizaciju LU dekompozicije je asimptotski jednak $n^3/3$.

Ovom metodom se može izračunati i determinanta matrice A. Naime, iz (25) je

$$det(P\,A) = \pm det(A) = det(L) \cdot \det(U) = u_{11} \cdots u_{nn},$$

jer je det(L)=1. Znači, determinanta matrice A je do na znak jednaka proizvodu glavnih elemenata.

Ako je $P \equiv I$, glavni elementi u_{ii} su jednaki količnicima determinanti glavnih minora. Naime, ako LU dekompoziciju matrice A predstavimo relacijom

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \cdot \begin{pmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{pmatrix},$$

gde su matrice sa indeksom 11 dimenzije $i \times i$, onda je $A_{11} = L_{11} \cdot U_{11}$, te je

$$det(A_{11}) = det(U_{11}) = u_{11} \cdots u_{ii}.$$

U opštem slučaju, ako sa A_i označimo i-ti glavni minor matrice A, imamo da je

$$u_{11} = det(A_1), \qquad u_{ii} = \frac{det(A_i)}{det(A_{i-1})}, \quad i \geq 2.$$

LU dekompozicija matrice A će se moći izvršiti bez pivotiranja, ukoliko su sve determinante njenih glavnih minora različite od nule, tj. $det(A_i) \neq 0$, $i = 1, \ldots, n$.

Ovom metodom se može odrediti i inverzna matrica matrici A. Ako sa $\mathbf{x}_i$ označimo vektor i-te kolone matrice A^{-1}, a sa $\mathbf{e}_i$ i-ti koordinatni vektor, onda iz $A\,A^{-1} = I$ sledi da je

$$A\mathbf{x}_i = \mathbf{e}_i, \qquad i = 1, \ldots, n,$$

pa je, prema (25),

$$L\,U\mathbf{x}_i = P\,A\mathbf{x}_i = P\mathbf{e}_i, \qquad i = 1, \ldots, n. \tag{28}$$

A^{-1} nalazimo rešavanjem n sistema linearnih jednačina (28) sa istom matricom sistema (čiju LU dekompoziciju jednom odredimo) i desnim stranama određenim koordinatnim vektorima.

Trodijagonalni sistem jednačina. Pri rešavanju različitih problema u numeričkoj matematici (interpolacija splajnovima, diferencijski granični zadaci,...), javlja se potreba za rešavanjem sistema linearnih jednačina sa trodijagonalnom matricom

$$\begin{aligned} c_1x_1 + b_1x_2 &= d_1 \\ a_ix_{i-1} + c_ix_i + b_ix_{i+1} &= d_i, \qquad i = 2, \dots, n-1, \\ a_nx_{n-1} + c_nx_n &= d_n. \end{aligned} \tag{29}$$

Ovakav sistem se efikasno rešava upravo Gaussovom metodom eliminacije, jer je broj računskih operacija koje treba izvršiti asimptotski jednak $8n$. U prvom koraku se iz prve jednačine sistema (29), pod pretpostavkom da je $c_1 \neq 0$, izrazi x_1,

$$x_1 = \alpha_2x_2 + \beta_2, \qquad \alpha_2 = -\frac{b_1}{c_1}, \quad \beta_2 = \frac{d_1}{c_1},$$

i eliminiše promenljiva x_1 u drugoj jednačini sistema (29). Sada ova jednačina sadrži samo promenljive x_2 i x_3, te se na isti način u drugom koraku izrazi x_2 pomoću x_3. U $(i-1)$-vom koraku se dobija veza

$$x_{i-1} = \alpha_ix_i + \beta_i, \tag{30}$$

pomoću koje eliminišemo promenljivu x_{i-1} u i-toj jednačini sistema (29)

$$a_i(\alpha_ix_i + \beta_i) + c_ix_i + b_ix_{i+1} = d_i.$$

Ako napišemo ovu jednačinu u obliku (30),

$$x_i = -\frac{b_i}{c_i + a_i\alpha_i}x_{i+1} + \frac{d_i - a_i\beta_i}{c_i + a_i\alpha_i} = \alpha_{i+1}x_{i+1} + \beta_{i+1},$$

dobijamo rekurentne veze za izračunavanje koeficijenata α_i, β_i,

$$\alpha_{i+1} = -\frac{b_i}{c_i + a_i\alpha_i}, \qquad \beta_{i+1} = \frac{d_i - a_i\beta_i}{c_i + a_i\alpha_i}, \qquad i = 1, \dots, n \quad (a_1 = 0). \tag{31}$$

Posle $(n-1)$ ovakvih koraka, sistem (29) se svodi na sistem

$$\begin{aligned} x_{i-1} &= \alpha_ix_i + \beta_i, \quad i = 2, \dots, n, \\ a_nx_{n-1} + c_nx_n &= d_n, \end{aligned}$$

čije je rešenje neposredno određeno formulama

$$\begin{aligned} x_n &= \frac{d_n - a_n\beta_n}{c_n + a_n\alpha_n} \\ x_i &= \alpha_{i+1}x_{i+1} + \beta_{i+1}, \quad i = n-1, \dots, 1. \end{aligned} \tag{32}$$

Stoga se Gaussova metoda eliminacije za rešavanje linearnog sistema jednačina sa trougaonom matricom sistema svodi na izračunavanje α_i i β_i, $i = 2, \dots, n$, po formulama (31) u direktnom hodu, i nalaženje rešenja sistema (32) u obrnutom hodu.

Gauss–Jordanova metoda. Kod Gaussove metode eliminacije, jednačina iz koje je izabran glavni element se posle tog koraka više ne transformiše. Stoga se posle $(n-1)$-og koraka dobija trougaoni sistem. U Gauss–Jordanovoj metodi, u svakom koraku se transformišu sve jednačine, osim one koja sadrži glavni element tog koraka. Pri izboru glavnog elementa ne uzimaju se u obzir koeficijenti jednačina iz kojih je već biran glavni element. Tako se posle $(n-1)$-og koraka dobija sistem sa dijagonalnom matricom.

I ovom metodom je determinanta matrice A do na znak određena proizvodom glavnih elemenata.

U poređenju sa Gaussovom metodom eliminacije, broj računskih operacija koje je potrebno izvršiti da bi se ovom metodom rešio sistem je veći; ali, ova metoda je efikasnija kada se određuje inverzna matrica, jer se u obrnutom hodu rešava n dijagonalnih umesto n trougaonih sistema.

5.3 Cholesky dekompozicija

Kada je A pozitivno definisana matrica, postoji jedinstveno određena donje trougaona matrica L, koja je realna ako je A realna matrica,

$$L = \begin{pmatrix} l_{11} & & & 0 \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix}, \qquad l_{ii} > 0,\ i = 1, \dots, n,$$

takva da je

$$A = L\, L^*. \tag{33}$$

Iz relacije (33) se dobijaju veze između elemenata matrica A i L,

$$\begin{aligned} a_{ii} &= |l_{i1}|^2 + \dots + |l_{ii}|^2 \\ a_{ij} &= l_{i1}\bar{l}_{j1} + \dots + l_{ij}\bar{l}_{jj}, \quad j < i, \end{aligned} \qquad i = 1, \dots, n, \tag{34}$$

te se elementi matrice L računaju po formulama

$$\begin{aligned} l_{11} &= \sqrt{a_{11}}, & l_{i1} &= \frac{a_{i1}}{l_{11}}, \\ l_{ii} &= \sqrt{a_{ii} - \sum_{k=1}^{i-1} |l_{ik}|^2}, & l_{ij} &= \frac{1}{l_{jj}}\Big(a_{ij} - \sum_{k=1}^{i-1} l_{ik}\bar{l}_{jk}\Big), \end{aligned} \qquad 1 < j < i \le n. \tag{35}$$

S obzirom na pozitivnu definisanost matrice A, potkorene veličine u formulama (35) su pozitivne. Kako iz (34) sledi da je

$$|l_{ij}| \leq \sqrt{a_{ii}},$$

elementi matrice L ne mogu biti suviše veliki, pa je metoda stabilna.

Kada je matrica L određena, rešenje sistema (17) se nalazi rešavanjem dva trougaona sistema

$$L\mathbf{y} = \mathbf{b}, \qquad L^*\mathbf{x} = \mathbf{y}.$$

Iz (33) je $det(A) = det(L) \cdot det(L^*) = \big(det(L)\big)^2$, te je

$$det(A) = (l_{11} \cdots l_{nn})^2.$$

PRIMER 4. Direktno primenjujući formule (35) na matricu sistema

$$\begin{aligned} 3.1x_1 + 1.5x_2 + \quad x_3 &= 10.83 \\ 1.5x_1 + 2.5x_2 + 0.5x_3 &= \ 9.20 \\ x_1 + 0.5x_2 + 4.2x_3 &= 17.10 \end{aligned}$$

dobijamo da je

$$L = \begin{pmatrix} 1.76068 & 0 & 0 \\ 0.85194 & 1.33199 & 0 \\ 0.56796 & 0.01211 & 1.96908 \end{pmatrix}.$$

Rešenje sistema $L\mathbf{y} = \mathbf{b}$, gde je $\mathbf{b}$ vektor desne strane datog sistema, je

$$\mathbf{y} = (6.15103 \quad 2.97276 \quad 6.89178)^T.$$

Rešenje sistema $L^*\mathbf{x} = \mathbf{y}$ je

$$\mathbf{x} = (1.3 \quad 2.2 \quad 3.5)^T,$$

a to predstavlja i rešenje polaznog sistema. Determinanta matrice sistema je

$$det = (1.76068 \cdot 1.33199 \cdot 1.96908)^2 = 21.3250.$$

5.4 Numerička stabilnost

Linearni sistem je stabilan ukoliko malim promenama ulaznih parametara (elementi matrice sistema i vektora slobodnih članova) odgovaraju male promene rešenja. Ako je matrica A singularna, tada za neke vektore $\mathbf{b}$ rešenje sistema (17) ne postoji, a za druge postoji ali nije jedinstveno. Stoga se prirodno nameće zaključak da je rešenje sistema utoliko osetljivije na izmene koeficijenata matrice i desne strane,

ukoliko je matrica sistema "bliža" singularnoj, jer zbog malih promena ulaznih parametara ona može postati singularna.

Determinanta matrice sistema nije pouzdana mera stabilnosti sistema jednačina, jer se množenjem svih jednačina sistema konstantom c determinanta povećava c^n puta, mada se karakteristike sistema bitno ne menjaju. Tačniju meru regularnosti matrice (u smislu stabilnosti) predstavlja njena uslovljenost $cond(A)$, koja je definisana relacijom (16). Pokažimo da je sistem to stabilniji, što je matrica sistema bolje uslovljena (tj. $cond(A)$ manje).

Rešenje sistema (17) se, zbog računske greške, u opštem slučaju određuje samo približno. Ako je $\mathbf{x}$ tačno, a $\mathbf{x}'$ približno rešenje, interesuje nas ocena razlike $\mathbf{x}-\mathbf{x}'$ u funkciji parametara sistema. Približno rešenje $\mathbf{x}'$ je ustvari tačno rešenje nekog drugog sistema

$$A\mathbf{x}' = \mathbf{b}', \qquad \mathbf{b}' \neq \mathbf{b}. \tag{36}$$

Oduzimanjem matrične jednakosti (17) od (36) dobijamo da je

$$A(\mathbf{x}' - \mathbf{x}) = \mathbf{b}' - \mathbf{b}, \qquad \text{tj.} \qquad \mathbf{x}' - \mathbf{x} = A^{-1}(\mathbf{b}' - \mathbf{b}).$$

Na osnovu (13) je

$$\|\mathbf{x}' - \mathbf{x}\| \leq \|A^{-1}\|\,\|\mathbf{b}' - \mathbf{b}\|, \tag{37}$$

a primenjujući istu nejednakost na (17), dobijamo ocenu

$$\|\mathbf{b}\| \leq \|A\|\,\|\mathbf{x}\|. \tag{38}$$

Sada je, na osnovu (37) i (38),

$$\frac{\|\mathbf{x}' - \mathbf{x}\|}{\|A\|\,\|\mathbf{x}\|} \leq \|A^{-1}\|\frac{\|\mathbf{b}' - \mathbf{b}\|}{\|\mathbf{b}\|},$$

što, uzimajući u obzir (16), daje ocenu

$$\frac{\|\mathbf{x}' - \mathbf{x}\|}{\|\mathbf{x}\|} \leq cond(A)\frac{\|\mathbf{b}' - \mathbf{b}\|}{\|\mathbf{b}\|}. \tag{39}$$

Dakle, relativna greška rešenja je srazmerna relativnoj grešci vektora desne strane i raste sa porastom uslovljenosti matrice sistema. U slučaju da su elementi matrice sistema dati približno, u oceni tipa (39) figurišu i relativne greške tih parametara.

Standardne metode (na primer Gaussova eliminacija) su nepogodne za rešavanje sistema linearnih jednačina sa loše uslovljenim matricama, jer male računske greške mogu dovesti do velikih grešaka rezultata.

PRIMER 5. Rešenje sistema linearnih jednačina

$$\begin{aligned} x_1 + 12x_2 + 3x_3 &= 38 \\ 12x_1 - 2x_2 + 9x_3 &= 9 \\ 3x_1 + 9x_2 + 4x_3 &= -49 \end{aligned}$$

je

$$x_1 = -9157, \qquad x_2 = -2166, \qquad x_3 = 11729,$$

a približno rešenje određeno Gaussovom metodom eliminacije, računato u običnoj tačnosti, je

$$x_1' = -9157.064000, \qquad x_2' = -2166.015000, \qquad x_3' = 11729.080000.$$

S obzirom na malu dimenziju sistema, greška je velika i uzrok tome je loša uslovljenost matrice sistema ($cond(A) = 6601$). Kada se sistem rešava Gaussovom eliminacijom u dvostrukoj tačnosti, postiže se tačnost 10^{-6}.

Primer ukazuje da treba računati u dvostrukoj tačnosti, ukoliko se primenjuju standardne metode za rešavanje loše uslovljenih sistema. Ovakve sisteme je, međutim, bolje rešavati iterativnim metodama (poglavlje 7) ili tzv. metodom singularne dekompozicije, o kojoj će biti reči u narednom odeljku.

5.5 Singularna dekompozicija

Analizirajmo sada opštiji oblik sistema linearnih jednačina (17) sa pravougaonom matricom sistema dimenzije $m \times n$, kod koga je $\mathbf{x}$ vektor dimenzije n, a $\mathbf{b}$ vektor dimenzije m. Proizvod $A\mathbf{x}$ je vektor nastao linearnom kombinacijom vektora kolona $\mathbf{a}_i$, $i = 1, \ldots, n$, matrice A,

$$A\mathbf{x} = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \ldots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ \vdots \\ a_{m2} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix},$$

te on pripada prostoru generisanom vektorima kolona matrice A, tzv. *prostoru kolona* $\mathcal{R}(A)$. Dimenzija prostora $\mathcal{R}(A)$, označimo je sa r, jednaka je broju linearno nezavisnih vektora kolona, i naziva se *rang* matrice A. Vektori kolona matrice A dimenzije $m \times n$ pripadaju m-dimenzionom vektorskom prostoru $\mathcal{C}^m$ (koordinate vektora mogu biti kompleksni brojevi), te je $\mathcal{R}(A)$ potprostor prostora $\mathcal{C}^m$ i $r \leq m$. Kako i $\mathbf{b} \in \mathcal{C}^m$, sistem (17) će imati rešenje za svako $\mathbf{b}$ ako je $\mathcal{R}(A) \equiv \mathcal{C}^m$, tj. ako je $r = m$. Vektori kolona matrice A tada predstavljaju jednu bazu prostora $\mathcal{C}^m$, i svaki vektor tog prostora može da se predstavi njihovom linearnom kombinacijom. Koeficijenti x_i, $i = 1, \ldots, n$, u linearnoj kombinaciji vektora kolona kojom je predstavljen vektor $\mathbf{b}$ su rešenja sistema (17). Ako je $\mathcal{R}(A) \subset \mathcal{C}^m$, tj. $r < m$, onda će sistem (17) imati rešenje samo ako $\mathbf{b} \in \mathcal{R}(A)$.

Rešenje sistema (17) kada je $\mathbf{b}$ nula vektor, odnosno rešenje homogenog sistema, pripada tzv. *prostoru nultih vektora* $\mathcal{N}(A)$ matrice A. Ovaj prostor sadrži uvek nula vektor $\mathbf{x} = (0, \ldots, 0)^T$, ali je pitanje da li sadrži i neki drugi vektor. Dimenzija prostora $\mathcal{N}(A)$, tj. broj linearno nezavisnih nultih vektora, jednaka je broju linearno zavisnih vektora kolona matrice A $n-r$. Ako je vektor $\mathbf{x}$ neko rešenje sistema (17), onda je i svaki vektor oblika $\mathbf{x}+\mathbf{x}_\circ$, gde je $\mathbf{x}_\circ \in \mathcal{N}(A)$, takođe rešenje ovog sistema.

Stoga je prostor kolona $\mathcal{R}(A)$ bitan u pitanju egzistencije rešenja, a prostor nultih vektora $\mathcal{N}(A)$ u pitanju jedinosti rešenja sistema (17). Ako $\mathbf{b} \in \mathcal{R}(A)$, rešenje sistema (17) postoji, a ako ne pripada, rešenje ne postoji. To rešenje (ako postoji) je jedinstveno ako prostor $\mathcal{N}(A)$ sadrži samo nula vektor, tj. ako mu je dimenzija nula; tada je rang matrice $r = n$. Ako dimenzija prostora $\mathcal{N}(A)$ nije nula, tada je ma koja kombinacija nađenog rešenja i proizvoljnog vektora iz $\mathcal{N}(A)$ takođe rešenje sistema (17).

Sistem (17) sa loše uslovljenom kvadratnom matricom je blizak sistemu sa singularnom matricom, koji, pak, može da nema rešenja ili da ima više rešenja, u zavisnosti od vektora $\mathbf{b}$. Stoga se ovaj slučaj može obuhvatiti opštim tipom zadatka (17), koji se ne može rešavati LU dekompozicijom ili nekom njoj srodnom metodom, već tzv.*singularnom dekompozicijom*. Ova metoda se zasniva na sledećoj teoremi linearne algebre

TEOREMA 1. *Neka je A proizvoljna matrica dimenzije $m \times n$. Tada postoje unitarna matrica U dimenzije $m \times m$ i unitarna matrica V dimenzije $n \times n$, takve da je*

$$U^* A V = W, \tag{40}$$

gde je W "dijagonalna" matrica dimenzije $m \times n$ sledećeg oblika

$$W = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}, \qquad D = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix}. \tag{41}$$

Vrednosti $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ su singularne vrednosti matrice A različite od nule, a r je rang matrice A.

Singularne vrednosti matrice A su nenegativni brojevi σ_k, takvi da je

$$\sigma_k^2 = \lambda_k(A^*A), \qquad k = 1, \ldots, n,$$

te je, s obzirom na (15),

$$\begin{aligned} \sigma_1 &= \sqrt{\max_{\mathbf{x}\neq 0} \frac{(A^*A\mathbf{x}\,,\,\mathbf{x})}{(\mathbf{x}\,,\,\mathbf{x})}} = \max_{\mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \|A\|_2 \\ \sigma_n &= \sqrt{\min_{\mathbf{x}\neq 0} \frac{(A^*A\mathbf{x}\,,\,\mathbf{x})}{(\mathbf{x}\,,\,\mathbf{x})}} = \min_{\mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \end{aligned} \tag{42}$$

Singularna dekompozicija matrice A sledi iz relacije (40),

$$A = U\,W\,V^*. \tag{43}$$

Dekompozicija (43) je uvek moguća i algoritam je stabilan, jer se koriste unitarne transformacije matrice A koje ne menjaju njenu uslovljenost. U praktičnoj realizaciji dekompozicije (43) dijagonalni elementi matrice W ne moraju biti uređeni,

te ona nije jedinstveno određena u tom smislu da su moguće permutacije vektora kolona matrice U, dijagonalnih elemenata matrice W i vektora kolona matrice V, ili se formiraju linearne kombinacije kolona matrice U, tj. V, koje odgovaraju skoro jednakim elementima matrice W.

Ako pretpostavimo da je $m \geq n$, dimenzija matrice D, date u (41), je u opštem slučaju n; pri tome su jednake nuli one singularne vrednosti σ_i koje odgovaraju linearno zavisnim vektorima kolona matrice A. Preostalih $m-n$ vrsta matrice W su nula vektori, te se, izostavljajući ove vrste i poslednjih $m-n$ kolona matrice U, dekompozicija (43) može prikazati u sledećem obliku

$$A = U_{m\times n}\, D_{n\times n}\, V^*_{n\times n}, \tag{44}$$

pri čemu indeksi ukazuju na dimenzije odgovarajućih matrica. U osnovi većine algoritama za efektivno određivanje dekompozicije (44), tj. nalaženje matrica U, D i V, je Householderova redukcija matrice na skoro dijagonalnu formu i dijagonalizacija QR algoritmom. O ovim metodama će biti reči u poglavlju 6 o sopstvenim vrednostima i vektorima matrica.

Vektori kolona $\mathbf{u}_j$ matrice U, koji odgovaraju singularnim vrednostima $\sigma_j \neq 0$, čine ortonormiranu bazu prostora kolona $\mathcal{R}(A)$. Vektori kolona $\mathbf{v}_k$ matrice V, koji odgovaraju singularnim vrednostima $\sigma_k = 0$, čine ortonormiranu bazu prostora nultih vektora $\mathcal{N}(A)$. Zaista, ma koji vektor $\mathbf{b} \in \mathcal{R}(A)$ može da se prikaže linearnom kombinacijom vektora kolona matrice A, tj. $\mathbf{b} = A\mathbf{x}$, a može da se izrazi, s obzirom na (44) i (8), i na sledeći način:

$$\mathbf{b} = A\mathbf{x} = U\,D\,V^*\mathbf{x} = \begin{pmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_n\end{pmatrix} \begin{pmatrix}\sigma_1 & & 0\\ & \ddots & \\ 0 & & \sigma_n\end{pmatrix} \begin{pmatrix}\mathbf{v}_1^*\\ \vdots \\ \mathbf{v}_n^*\end{pmatrix} \mathbf{x}$$

$$= \begin{pmatrix}\sigma_1\mathbf{u}_1 & \dots & \sigma_n\mathbf{u}_n\end{pmatrix} \begin{pmatrix}(\mathbf{x}\,,\,\mathbf{v}_1)\\ \vdots \\ (\mathbf{x}\,,\,\mathbf{v}_n)\end{pmatrix} = \sigma_1(\mathbf{x}\,,\,\mathbf{v}_1)\mathbf{u}_1 + \dots + \sigma_n(\mathbf{x}\,,\,\mathbf{v}_n)\mathbf{u}_n,$$

tj. linearnom kombinacijom vektora $\mathbf{u}_j$ koji odgovaraju nenula singularnim vrednostima $\sigma_j \neq 0$, a kojih ima r.

Pokažimo sada da bazu prostora $\mathcal{N}(A)$ čine oni vektori $\mathbf{v}_k$ matrice V koji odgovaraju singularnim vrednostima $\sigma_k = 0$, tj. da se svaki vektor $\mathbf{x} \in \mathcal{N}(A)$ može predstaviti u obliku

$$\mathbf{x} = \sum_{\substack{k\\ \sigma_k = 0}} c_k \mathbf{v}_k. \tag{45}$$

Kako je $(\mathbf{v}_i\,,\,\mathbf{v}_j) = \delta_{ij}$ zbog unitarnosti matrice V, to je za svaki vektor $\mathbf{v}_k$

$$A\mathbf{v}_k = U\,D\,V^*\mathbf{v}_k = \begin{pmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_n\end{pmatrix} \begin{pmatrix}\sigma_1\mathbf{v}_1^*\\ \vdots \\ \sigma_n\mathbf{v}_n^*\end{pmatrix} \mathbf{v}_k = \begin{pmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_n\end{pmatrix} \begin{pmatrix}0\\ \vdots \\ \sigma_k \\ \vdots \\ 0\end{pmatrix},$$

te je $A\mathbf{v}_k = 0$ ako je $\sigma_k = 0$, tj. $\mathbf{v}_k \in \mathcal{N}(A)$. Ovih vektora, odnosno singularnih vrednosti $\sigma_k = 0$, ima $n-r$, dakle upravo onoliko kolika je dimenzija prostora $\mathcal{N}(A)$, te oni čine bazu ovog prostora. Stoga se singularnom dekompozicijom konstruišu baze prostora $\mathcal{R}(A)$ i $\mathcal{N}(A)$.

Analizirajmo sada kako se metodom singularne dekompozicije može rešiti sistem sa loše uslovljenom, ili čak singularnom kvadratnom matricom ($m = n$), ili tzv. preodređeni sistem ($m > n$), i šta predstavljaju dobijena rešenja.

Uslovljenost (16) regularne kvadratne matrice je količnik njene najveće i najmanje singularne vrednosti

$$cond(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n},$$

jer je, na osnovu (42) i (9),

$$\frac{1}{\sigma_n} = \max_{\mathbf{x}\neq 0} \frac{\|\mathbf{x}\|_2}{\|A\mathbf{x}\|_2} = \max_{\mathbf{y}\neq 0} \frac{\|A^{-1}\mathbf{y}\|_2}{\|\mathbf{y}\|_2} = \|A^{-1}\|_2.$$

Za singularnu matricu je najmanja singularna vrednost nula i njen je rang $r < n$. Dakle, neki od dijagonalnih elemenata matrice D u dekompoziciji (44) će biti jednak nuli, te se ne može u opštem slučaju rešenje sistema (17) dobiti po formuli

$$\mathbf{x} = A^{-1}\mathbf{b} = V\, D^{-1}\, U^*\mathbf{b}.$$

Neka je $\bar{D}^{-1}$ dijagonalna matrica čiji su elementi

$$s_i = \begin{cases} \frac{1}{\sigma_i}, & \text{ako je } \sigma_i \neq 0 \\ 0, & \text{ako je } \sigma_i = 0, \end{cases} \tag{46}$$

tj. matrica dobijena od matrice D^{-1} tako što su zamenjeni nulama oni dijagonalni elementi matrice D^{-1} koji odgovaraju singularnim vrednostima $\sigma_k = 0$ (a koji bi, ustvari, trebalo da budu beskonačno veliki). Pokazaćemo da vektor $\mathbf{x}$, određen izrazom

$$\mathbf{x} = V\, \bar{D}^{-1}\, U^*\mathbf{b}, \tag{47}$$

predstavlja rešenje sistema (17) sa najmanjom euklidskom normom ako ovaj sistem ima više rešenja (tj. ako $\mathbf{b} \in \mathcal{R}(A)$), ili rešenje u smislu najmanje srednjekvadratne greške, ako sistem nema rešenja (tj. ako $\mathbf{b} \notin \mathcal{R}(A)$).

Neka je prvo $\mathbf{b} \in \mathcal{R}(A)$. Ako je $\mathbf{x}$ rešenje sistema (17), određeno formulom (47), onda je rešenje tog sistema i svaki vektor $\mathbf{x} + \mathbf{x}_\circ$ za proizvoljno $\mathbf{x}_\circ \in \mathcal{N}(A)$. Ocenimo normu vektora $\mathbf{x} + \mathbf{x}_\circ$. Zbog unitarnosti matrice V je $V\,V^* = I$, i

$$\|V\mathbf{z}\|_2^2 = (V\mathbf{z}\,,\, V\mathbf{z}) = (V^*V\mathbf{z}\,,\, \mathbf{z}) = (\mathbf{z}\,,\, \mathbf{z}) = \|z\|_2^2, \tag{48}$$

za proizvoljan vektor $\mathbf{z}$. Stoga je

$$\begin{aligned} \|\mathbf{x} + \mathbf{x}_\circ\|_2 &= \|V\bar{D}^{-1}U^*\mathbf{b} + \mathbf{x}_\circ\|_2 = \|V(\bar{D}^{-1}U^*\mathbf{b} + V^*\mathbf{x}_\circ)\|_2 \\ &= \|\bar{D}^{-1}U^*\mathbf{b} + V^*\mathbf{x}_\circ\|_2. \end{aligned} \tag{49}$$

U vektoru $\bar{D}^{-1}U^*\mathbf{b} + V^*\mathbf{x}_\circ$ prvi sabirak ima različite od nule j-te koordinate koje odgovaraju vrednostima $s_j \neq 0$, datim u (46), tj. vrednostima $\sigma_j \neq 0$. Nasuprot tome, drugi sabirak ima različite od nule samo k-te koordinate koje odgovaraju vrednostima $s_k = 0$, tj. $\sigma_k = 0$, jer $\mathbf{x}_\circ \in \mathcal{N}(A)$ pa je, na osnovu (45),

$$V^*\mathbf{x} = \begin{pmatrix} \mathbf{v}_1^* \\ \vdots \\ \mathbf{v}_n^* \end{pmatrix} \sum_{\substack{k \\ \sigma_k = 0}} c_k \mathbf{v}_k = \begin{pmatrix} \sum_{\substack{k \\ \sigma_k = 0}} c_k(\mathbf{v}_k\,,\,\mathbf{v}_1) \\ \vdots \\ \sum_{\substack{k \\ \sigma_k = 0}} c_k(\mathbf{v}_k\,,\,\mathbf{v}_n). \end{pmatrix}$$

Stoga će vektor $\mathbf{x} + \mathbf{x}_\circ$ imati minimalnu normu (49) ako je $\mathbf{x}_\circ = 0$, odnosno od svih rešenja sistema (17) rešenje određeno izrazom (47) ima minimalnu euklidsku normu.

Ako $\mathbf{b} \notin \mathcal{R}(A)$ sistem (17) nema rešenja, a vektor $\mathbf{x}$, određen izrazom (47), je vektor iz prostora $\mathcal{C}^n$ za koji vektor ostatka

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

ima minimalnu euklidsku normu. Zaista, ako umesto vektora $\mathbf{x}$, određenog izrazom (47), uzmemo neki drugi vektor $\mathbf{x} + \mathbf{x}'$, onda će se ostatak $\mathbf{r}$ promeniti za vektor $\mathbf{b}' = A\mathbf{x}'$ koji pripada prostoru kolona $\mathcal{R}(A)$. Tada je, s obzirom na (44) i (47) i unitarnost matrice U,

$$\begin{aligned} \|A(\mathbf{x}+\mathbf{x}') - \mathbf{b}\|_2 = \|A\mathbf{x} - \mathbf{b} + \mathbf{b}'\|_2 = \|(U\,D\,V^*)(V\,\bar{D}^{-1}\,U^*\mathbf{b}) - \mathbf{b} + \mathbf{b}'\|_2 \\ = \|U\,D\,\bar{D}^{-1}\,U^*\mathbf{b} - U\,U^*\mathbf{b} + U\,U^*\mathbf{b}'\|_2 = \|U\big((D\,\bar{D}^{-1} - I)U^*\mathbf{b} + U^*\mathbf{b}'\big)\|_2 \\ = \|(D\,\bar{D}^{-1} - I)U^*\mathbf{b} + U^*\mathbf{b}'\|_2. \end{aligned}$$

Matrica $D\,\bar{D}^{-1} - I$ je dijagonalna matrica kod koje su samo k-ti dijagonalni elementi, koji odgovaraju singularnim vrednostima $\sigma_k = 0$, različiti od nule. Sa druge strane, vektor $U^*\mathbf{b}'$ ima različite od nule samo j-te koordinate koje odgovaraju vrednostima $\sigma_j \neq 0$, jer $\mathbf{b}' \in \mathcal{R}(A)$. Stoga se minimum euklidske norme vektora ostatka postiže za $\mathbf{b}' = 0$.

PRIMER 6. Matrica sistema

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 1 \\ 2x_1 + 4x_2 + 6x_3 &= 1 \\ 2x_1 + 2x_2 + 3x_3 &= 1 \end{aligned}$$

je singularna, i sistem nema rešenja. Singularnom dekompozicijom ove matrice

dobijaju se matrice

$$U = \begin{pmatrix} -0.4025 & -0.1950 & 0.8944 \\ -0.8050 & -0.3899 & -0.4472 \\ -0.4359 & 0.9000 & 0.0000 \end{pmatrix},$$

$$W = diag\,(9.2869, \quad 0.8681, \quad 0),$$

$$V = \begin{pmatrix} -0.3106 & 0.9505 & 0.0000 \\ -0.5273 & -0.1723 & 0.8320 \\ -0.7909 & -0.2584 & -0.5547 \end{pmatrix},$$

a kao rešenje vektor

$$\mathbf{x} = (0.4 \quad 0.0308 \quad 0.0462)^T.$$

Kod vrlo loše uslovljene matrice neke od singularnih vrednosti će biti male, nekad čak van granica tačnosti računske mašine, te će ovakav sistem biti u metodi singularne dekompozicije tretiran kao sistem sa singularnom matricom.

Preodređeni sistem ($m > n$) se često javlja pri obradi eksperimentalnih podataka. Ako ne postoji linearna zavisnost jednačina sistema , ovom metodom se nalazi rešenje sa najmanjom srednjekvadratnom greškom, kao što je pokazano. U tom slučaju nema potrebe za modifikacijom matrice D^{-1}, jer su sve singularne vrednosti matrice sistema različite od nule.

Ako je $m < n$, za sistem (17) kažemo da je neodređen jer ima više nepoznatih nego jednačina. Stoga, u opštem slučaju, on nema jedinstveno rešenje već postoji $(n - m)$ linearno nezavisnih rešenja. Ovaj prostor rešenja određuje se metodom singularne dekompozicije, tako što se matrica sistema A, dimenzije $m \times n$, dopuni nulama do kvadratne matrice dimenzije n. Vektor desne strane $\mathbf{b}$ se takođe dopuni nulama da se dobije n-dimenzioni vektor. Na taj način smo polazni problem sveli na sistem linearnih jednačina sa singularnom, kvadratnom matricom, čije rešavanje singularnom dekompozicijom je već analizirano. U opštem slučaju, formulom (47) se određuje rešenje sa minimalnom euklidskom normom, a sva ostala rešenja se dobijaju linearnom kombinacijom ovog rešenja i vektora kolona matrice V koji odgovaraju singularnim vrednostima $\sigma_k = 0$, jer ovi čine bazu prostora $\mathcal{N}(A)$.

PRIMER 7. Sistem

$$x_1 + 2x_2 + 3x_3 + 4x_4 = 10$$
$$2x_1 + 3x_2 + 4x_3 + 5x_4 = 14$$

je neodređen i nema jedinstveno rešenje. Singularnom dekompozicijom matrice dobijene od matrice sistema dopunjene nulama do kvadratne matrice, dobijaju se

matrice

$$U = \begin{pmatrix} -0.5969 & -0.8023 & 0 & 0 \\ -0.8023 & 0.5969 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$W = diag\,(9.1521, \quad 0.4886, \quad 0, \quad 0),$$

$$V = \begin{pmatrix} -0.2405 & 0.8013 & 0.5477 & 0 \\ -0.3934 & 0.3811 & -0.7303 & 0.4082 \\ -0.5463 & -0.0392 & -0.1826 & -0.8165 \\ -0.6992 & -0.4595 & 0.3651 & 0.4082 \end{pmatrix}.$$

Rešenje sistema sa minimalnom euklidskom normom je

$$\mathbf{x} = (3.2032 \quad -0.3270 \quad -2.9558 \quad 4.0795)^T.$$

6

Sopstvene vrednosti i sopstveni vektori matrica

Kao što je napomenuto u poglavlju 5, određivanje sopstvenih vrednosti i sopstvenih vektora matrica je jedan od četiri osnovna zadatka linearne algebre. S obzirom da se numeričke metode za rešavanje ovog problema razlikuju od metoda kojima se rešavaju ostali zadaci, one su izdvojene u posebno poglavlje.

Sopstvene vrednosti matrice A su koreni njenog karakterističnog polinoma, koji je predstavljen determinantom (5.3). Teorijski, determinantu je uvek moguće razviti standardnim metodama linearne algebre. Međutim, ako je dimenzija matrice n velika, izračunavanja su vrlo glomazna i po pravilu dovode do nagomilavanja računske greške, pri čemu je dodatna poteškoća i to što se u svakoj vrsti i koloni ove determinante javlja promenljiva λ.

Iz navedenih razloga se za rešavanje problema sopstvenih vrednosti matrica koriste numeričke metode, koje se mogu podeliti u dve osnovne grupe:

(*i*) metode za rešavanje *potpunog problema*, tj. metode kojima se određuju sve sopstvene vrednosti i svi sopstveni vektori, i

(*ii*) metode za rešavanje *delimičnog problema*, tj. metode kojima se određuje jedna sopstvena vrednost (najčešće najveća po modulu) i njoj odgovarajući sopstveni vektor.

6.1 Potpun problem sopstvenih vrednosti

Mogu se izdvojiti dva osnovna pristupa u rešavanju potpunog problema sopstvenih vrednosti. U prvom se sopstvene vrednosti određuju kao nule karakterističnog polinoma (5.3), te se numeričkim metodama nalazi ovaj polinom

$$D(\lambda) = (-1)^n(\lambda^n + p_1\lambda^{n-1} + \cdots + p_n), \tag{1}$$

odnosno njegovi koeficijenti p_i, $i = 1, \ldots, n$.

Metoda interpolacije. Zasniva se na činjenici da je interpolacioni polinom stepena n funkcije koja je i sama polinom stepena n, identičan funkciji. Stoga se tabelira polinom $D(\lambda)$, dat u (5.3), u $(n+1)$-oj tački, i na osnovu date tabele formira interpolacioni polinom. Obično se čvorovi interpolacije μ_i biraju tako da je $|\mu_i| \leq \|A\|$, $i = 0, \ldots, n$; na taj način je, s obzirom na (5.14), najmanja greška upravo u intrvalu u kome se nalaze sopstvene vrednosti matrice A. Vrednosti $D(\mu_i)$ se računaju nekom od metoda iz prethodnog poglavlja.

Metoda Le Verriera. Na osnovu Vieteovih formula, kojima se izražavaju veze koeficijenata i korena polinoma, i veze korena polinoma i traga matrice A^k (trag matrice je suma njenih dijagonalnih elemenata),

$$tr(A^k) \equiv S_k = \lambda_1^k + \cdots + \lambda_n^k, \tag{2}$$

dobijaju se rekurentne formule za nalaženje koeficijenata p_k polinoma (1)

$$p_1 = -S_1, \qquad p_k = -\frac{1}{k}(S_k + p_1 S_{k-1} + p_2 S_{k-2} + \cdots + p_{k-1} S_1), \quad k = 2, \ldots, n.$$

Metoda Krilova. Zasniva se na Cayley-Hamiltonovoj teoremi po kojoj matrica anulira svoj karakteristični polinom,

$$D(A) \equiv (-1)^n (A^n + p_1 A^{n-1} + \cdots + p_n I) = 0. \tag{3}$$

Množenjem jednakosti (3) proizvoljnim vektorom $\mathbf{b}_0$, dobijamo sistem linearnih jednačina po koeficijentima p_k,

$$b_i^{(n-1)} p_1 + b_i^{(n-2)} p_2 + \cdots + b_i^{(0)} p_n = -b_i^{(n)}, \quad i = 1, \ldots, n, \tag{4}$$

gde je $\mathbf{b}_k = A\mathbf{b}_{k-1} = A^k \mathbf{b}_0 = (b_1^{(k)}, b_2^{(k)}, \ldots, b_n^{(k)})^T$. Kada se metodama za nalaženje korena polinoma (poglavlje 7) odrede sopstvene vrednosti, sopstveni vektori se mogu naći rešavanjem homogenog sistema (5.2) kojim su definisani. Pri tome se dobijaju jednostavnije formule, ako se predstavi sopstveni vektor linearnom kombinacijom vektora $\mathbf{b}_k$, $k = 0, \ldots, n-1$, i uzme u obzir linearna zavisnost (4) vektora $\mathbf{b}_n$ i ovih vektora.

Metoda Danilevskog. Transformacijama sličnosti se matrica A transformiše u Frobeniusovu matricu

$$P = \begin{pmatrix} p_1 & \cdots & p_{n-1} & p_n \\ 1 & & 0 & 0 \\ & \ddots & & \\ 0 & & 1 & 0 \end{pmatrix},$$

za koju su koeficijenti karakterističnog polinoma elementi prve vrste uzeti sa promenjenim znakom. Kako slične matrice imaju isti karakteristični polinom, na ovaj

način je određen i karakteristični polinom date matrice A. Između sopstvenih vektora $\mathbf{x}_k$ i $\mathbf{y}_k$ matrica A i P, koji odgovaraju istoj sopstvenoj vrednosti λ_k, postoji relacija

$$\mathbf{x}_k = T\mathbf{y}_k, \qquad k = 1, \ldots, n,$$

gde je T matrica kojom je definisana transformacija A u P, tj. $P = T^{-1}AT$. Sopstveni vektori $\mathbf{y}_k$ matrice P se jednostavno nalaze iz sistema $P\mathbf{y}_k = \lambda_k \mathbf{y}_k$, $k = 1, \ldots, n$, jer je matrica P retka.

Dakle, svim pomenutim metodama nalazi se samo karakteristični polinom. Da bi se problem u potpunosti rešio, treba naći njegove korene. S obzirom da su koeficijenti polinoma po pravilu određeni samo približno, prirodno se postavlja pitanje kako perturbacije koeficijenata polinoma utiču na njegove korene. Bez ulaženja u teorijsku analizu ove greške, navešćemo primer Wilkinsona ([30]) koji ilustruje ovaj uticaj kod polinoma sa loše uslovljenim korenima.

PRIMER 1. Koreni polinoma

$$P(x) = \prod_{k=1}^{20}(x-k) = \sum_{k=0}^{20} c_k x^{20-k}$$

su $\xi_k = k$, $k = 1, \ldots, 20$. Ako umesto koeficijenta $c_1 = -(1+2+\cdots+20) = -210$ imamo $c_1(1+\epsilon)$, greška u korenu ξ_{20} je $0.9 \cdot 10^{10}\epsilon$. Još drastičnija promena nastaje u korenu ξ_{16}, koji se menja za $3.7 \cdot 10^{14}\epsilon$, ako je ϵ relativna greška koeficijenta c_5. To znači da ni računanje u dvostrukoj tačnosti ne bi moglo da da ni jednu sigurnu cifru ovog korena.

Zbog ovakvih poteškoća, navedeni pristup rešavanju potpunog problema sopstvenih vrednosti se sve više napušta.

Drugi pristup je direktno nalaženje sopstvenih vrednosti, bez prethodnog određivanja karakterističnog polinoma. To se obično postiže tako što se polazna matrica transformiše u njoj sličnu dijagonalnu ili trougaonu matricu, čije su sopstvene vrednosti dijagonalni elementi. Lemama 5.6 i 5.7 garantuje se, štaviše, postojanje takve unitarne transformacije, što je od posebnog značaja za stabilnost algoritma. Naime, pri unitarnim transformacijama ne menja se euklidska norma (5.6) vektora, na osnovu (5.48), niti njoj saglasna norma matrice (5.12). Zaista, kako je sferna norma unitarne matrice U, prema (5.9),

$$\|U\|_2 = \sup_{\mathbf{x}\neq 0} \sqrt{\frac{(U\mathbf{x}\,,\, U\mathbf{x})}{(\mathbf{x}\,,\,\mathbf{x})}} = \sup_{\mathbf{x}\neq 0} \sqrt{\frac{\mathbf{x}^* U^* U\mathbf{x}}{\mathbf{x}^*\mathbf{x}}} = 1,$$

to je

$$\|A\|_2 = \|U^*\,U\,A\|_2 \leq \|U^*\|_2 \|U\,A\|_2 = \|U\,A\|_2 \leq \|U\|_2 \|A\|_2 = \|A\|_2,$$

tj. $\|U\,A\|_2 = \|A\|_2$, i, slično, $\|A\,U\|_2 = \|A\|_2$. Analogno je

$$\begin{aligned}\|A^{-1}\|_2 &= \|U\,U^*\,A^{-1}\|_2 \leq \|U\|_2 \|U^{-1}A^{-1}\|_2 \\ &= \|(A\,U)^{-1}\|_2 \leq \|U^*\|_2 \|A^{-1}\|_2 = \|A^{-1}\|_2,\end{aligned}$$

te je $\|(A\,U)^{-1}\|_2 = \|A^{-1}\|_2$, i, slično, $\|(U\,A)^{-1}\|_2 = \|A^{-1}\|_2$. To znači da je uslovljenost (5.16) matrice A

$$cond(A) = \|A\|_2 \cdot \|A^{-1}\|_2 = \|U\,A\|_2 \cdot \|(U\,A)^{-1}\|_2 = cond(U\,A),$$

odnosno da se uslovljenost matrice pri unitarnoj transformaciji ne menja, što garantuje stabilnost.

Problem je u tome što ne postoji efektivni algoritam za nalaženje pomenute unitarne matrice. Ona se konstruiše iterativnim algoritmom kao granica niza međusobno sličnih matrica (metoda Jacobija, LR algoritam, QR algoritam,...). Većina ovih iterativnih metoda se znatno efikasnije primenjuje ukoliko polazna matrica ima skoro trougaonu, ili, još bolje, skoro dijagonalnu formu. Stoga se prethodno posebnim metodama (Givensova metoda rotacije, Householderova metoda redukcije,...) matrica opšte strukture unitarnim transformacijama sličnosti transformiše u tzv. *gornju Hessenbergovu matricu*

$$A \sim \begin{pmatrix} * & \dots & \dots & * \\ * & & & \vdots \\ & \ddots & & \vdots \\ 0 & & * & * \end{pmatrix}, \tag{5}$$

ili, kada je polazna matrica Hermiteova, u trodijagonalnu matricu

$$A \sim \begin{pmatrix} * & * & & & 0 \\ * & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & * \\ 0 & & & * & * \end{pmatrix}. \tag{6}$$

Redukcija matrice na gornju Hessenbergovu, tj. trodijagonalnu, formu može se postići i primenom modifikovane Gaussove metode eliminacije – modifikovane u tom smislu da se uporedo sa formiranjem linearnih kombinacija vrsta matrice, formiraju i iste linearne kombinacije odgovarajućih kolona, kako bi se na svakom koraku dobila matrica slična prethodnoj. Ovaj algoritam je čak efikasniji od pomenutih algoritama koji koriste unitarne transformacije, ali, kao što je pokazano u poglavlju 5, može biti nestabilan.

6.2 Givensova metoda rotacije

Ovo je metoda kojom se pomoću konačno mnogo unitarnih transformacija sličnosti matrica A svodi, u opštem slučaju, na gornju Hessenbergovu matricu (5), ili, kada

je A Hermiteova matrica, na trodijagonalnu matricu (6). Unitarne transformacije se vrše pomoću tzv. matrica rotacije

$$U_{kl} = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & \cos\phi & & -e^{-\imath\psi}\sin\phi & & \leftarrow k \\ & & & \ddots & & & \\ & & e^{\imath\psi}\sin\phi & & \cos\phi & & \leftarrow l \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{pmatrix} \tag{7}$$

gde su ϕ i ψ realni parametri. U realnom slučaju, za $\psi = 0$, matrica U_{kl} je ortogonalna.

Neka je $A = (a_{ij})$, $B = (b_{ij}) = A\,U_{kl}$ i $C = (c_{ij}) = U_{kl}^*\,B = U_{kl}^*\,A\,U_{kl}$. Matrica C je unitarno slična matrici A, a veze među elementima matrica A, B i C su

$$\begin{aligned} b_{ik} &= a_{ik}\alpha + a_{il}\beta & c_{kj} &= b_{kj}\alpha + b_{lj}\bar{\beta} & \\ b_{il} &= -a_{ik}\bar{\beta} + a_{il}\alpha & c_{lj} &= -b_{kj}\beta + b_{lj}\alpha & i,j = 1,\ldots,n, \\ b_{ij} &= a_{ij}, \quad j \neq k,l, & c_{ij} &= b_{ij}, \quad i \neq k,l, & \end{aligned} \tag{8}$$

pri čemu je, radi kratkoće zapisa, korišćena oznaka $\alpha = \cos\phi$ i $\beta = e^{\imath\psi}\sin\phi$. Iz formula (8) je očigledno da se pri ovoj transformaciji menjaju samo elementi k-te i l-te kolone i k-te i l-te vrste. Parametre rotacije ϕ i ψ ćemo odrediti tako da element $c_{l,k-1}$ matrice C bude jednak nuli,

$$c_{l,k-1} = -b_{k,k-1}\beta + b_{l,k-1}\alpha = -a_{k,k-1}\beta + a_{l,k-1}\alpha = 0,$$

a to znači da treba da je $\beta = \frac{a_{l,k-1}}{a_{k,k-1}}\alpha$. Kako je $\alpha^2 + |\beta|^2 = 1$, sledi da će biti $c_{l,k-1} = 0$ ako je

$$\alpha = \frac{|a_{k,k-1}|}{\sqrt{|a_{k,k-1}|^2 + |a_{l,k-1}|^2}}, \qquad \beta = \frac{a_{l,k-1}}{a_{k,k-1}}\alpha. \tag{9}$$

Matricom U_{23}, određenom formulama (9), transformišemo matricu A u njoj sličnu matricu $A_1 = (a_{ij}^{(1)}) = U_{23}^*\,A\,U_{23}$ kod koje je $a_{31}^{(1)} = 0$. Zatim, pomoću matrice U_{24} određujemo matricu kod koje i element sa indeksom $(4,1)$ jednak nuli. Pošto se pri toj transformaciji menjaju samo elementi druge i četvrte vrste i kolone, element sa indeksom $(3,1)$ ostaće nula. Produžavajući postupak na isti način, anuliraćemo redom elemente prve, zatim druge,..., i konačno $(n-2)$-ge kolone, tako da se na kraju dobija gornja Hessenbergova matrica A',

$$A' = U_{n-1,n}^* \cdots U_{24}^* U_{23}^*\,A\,U_{23}U_{24}\cdots U_{n-1,n} \sim A,$$

Ako je A Hermiteova matrica, onda su i sve matrice A_k, dobijene u toku transformacija, takođe Hermiteove. Stoga će se na svakom koraku anulirati ne samo

element $a_{pq}^{(k)}$ matrice A_k, već i njemu simetričan element $a_{qp}^{(k)}$, zbog čega matrica A' ima trodijagonalnu formu. S obzirom da se anuliraju $(n-2)$ elementa prve, $(n-3)$ elementa druge, ..., i jedan element $(n-2)$-ge kolone, ukupan broj transformacija elementarnim matricama rotacije je $(n-2)(n-1)/2$.

PRIMER 2. Ilustrujmo metodu na primeru simetrične matrice

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix}.$$

Trodijagonalna matrica A', slična matrici A, dobija se posle transformacije samo matricom U_{23} (jer je $n = 3$),

$$U_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}, \qquad A' = U_{23}^T A U_{23} = \begin{pmatrix} 2 & \sqrt{2} & 0 \\ \sqrt{2} & \frac{7}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{3}{2} \end{pmatrix}.$$

Givensovom metodom rotacije nije moguće transformisati matricu u njoj sličnu trougaonu, odnosno dijagonalnu, matricu, jer se anuliranjem elemenata sa indeksima $(i+1, i)$ kvare već dobijene nule. Stoga se ovom metodom ne određuju direktno sopstvene vrednosti, već se matrica svodi na jednostavniji oblik; polazeći od njega, iterativnim metodama (obično LR ili QR algoritmom) se nalaze sve sopstvene vrednosti matrice.

6.3 Jacobijeva metoda

Jacobijeva metoda je iterativna metoda kojom se određuju sopstvene vrednosti Hermiteove matrice A. Pošto efektivno nije moguće naći unitarnu matricu U kojom se Hermiteova matrica A transformiše u dijagonalnu formu, a koja postoji na osnovu leme 5.6, ovom metodom se beskonačnim nizom unitarnih transformacija sličnosti anuliraju nedijagonalni elementi matrice A, čime se ona transformiše u matrice koje teže dijagonalnoj matrici. Pri tome se koriste matrice rotacije (7), te se metoda naziva i iterativna metoda rotacije.

Ilustrovaćemo metodu na primeru realne simetrične matrice A dimenzije n. Pomoću realne matrice rotacije (7),

$$U_{kl} = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & \cos\phi & & -\sin\phi & & \\ & & & \ddots & & & \\ & & \sin\phi & & \cos\phi & & \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \leftarrow k \\ \\ \leftarrow l \\ \\ \\ \end{matrix} \qquad (10)$$

nalazimo matricu B sličnu matrici A,

$$B = (b_{ij}) = U_{kl}^T A U_{kl}, \tag{11}$$

čiji su elementi

$$\begin{aligned}
b_{kj} &= a_{kj}\cos\phi + a_{lj}\sin\phi \qquad b_{ik} = a_{ik}\cos\phi + a_{il}\sin\phi \\
b_{lj} &= -a_{kj}\sin\phi + a_{lj}\cos\phi \qquad b_{il} = -a_{ik}\sin\phi + a_{il}\cos\phi \\
b_{ij} &= a_{ij} \qquad i,j \neq k,l \\
b_{kk} &= a_{kk}\cos^2\phi + a_{ll}\sin^2\phi + 2a_{kl}\sin\phi\cos\phi \\
b_{ll} &= a_{kk}\sin^2\phi + a_{ll}\cos^2\phi - 2a_{kl}\sin\phi\cos\phi \\
b_{kl} &= (a_{ll} - a_{kk})\sin\phi\cos\phi + a_{kl}(\cos^2\phi - \sin^2\phi) \\
&= \tfrac{1}{2}(a_{ll} - a_{kk})\sin 2\phi + a_{kl}\cos 2\phi = b_{lk}
\end{aligned} \tag{12}$$

Matricu U_{kl}, tj. ugao rotacije ϕ, odredimo tako da se ovom unitarnom transformacijom anulira element b_{kl}, a zbog simetričnosti i b_{lk}. Iz (12) sledi da je ϕ određeno jednačinom

$$\tfrac{1}{2}(a_{ll} - a_{kk})\sin 2\phi + a_{kl}\cos 2\phi = 0,$$

te je

$$\tan 2\phi = \frac{-a_{kl}}{\frac{1}{2}(a_{ll} - a_{kk})}. \tag{13}$$

S obzirom da imenilac izraza (13) može biti mali broj ili čak nula, korišćenjem trigonometrijskih transformacija iz ovog izraza se izvodi stabilan algoritam za nalaženje elemenata matrice (10)

$$\sin\phi = \frac{\omega}{\sqrt{2(1+\sqrt{1-\omega^2})}}, \qquad \cos\phi = \sqrt{1-\sin^2\phi},$$

gde je

$$\omega = sign(\mu)\frac{\lambda}{\sqrt{\lambda^2+\mu^2}}, \qquad \lambda = -a_{kl}, \qquad \mu = \tfrac{1}{2}(a_{ll} - a_{kk}).$$

Dakle, na odgovarajući način definisanom unitarnom transformacijom određujemo, na osnovu (11), matricu B sličnu matrici A, kod koje je nedijagonalni element b_{kl}, a samim tim i b_{lk}, jednak nuli. Transformacija koju vršimo menja i druge elemente polazne matrice, s obzirom na (12), te će neki nedijagonalni elementi koji su u prethodnim koracima bili svedeni na nulu, biti "pokvareni". Stoga Jacobijev algoritam nije konačan, ali je konvergentan u tom smislu da niz matrica A_m unitarno sličnih matrici A, koje su određene na način opisan za matricu B, teži ka dijagonalnoj matrici D,

$$\lim_{m\to\infty} A_m = D, \qquad A_m = U_m^T \cdots U_1^T A U_1 \cdots U_m. \tag{14}$$

U_j je oznaka za matricu rotacije (10), korišćenu u j-tom koraku.

TEOREMA 1. *Jacobijevom metodom određen niz matrica A_m konvergira ka dijagonalnoj matrici $D = diag\,(\lambda_1, \ldots, \lambda_n)$.*

DOKAZ: Uočimo da je, s obzirom na formule (12),

$$b_{kj}^2 + b_{lj}^2 = a_{kj}^2 + a_{lj}^2, \qquad b_{ik}^2 + b_{il}^2 = a_{ik}^2 + a_{il}^2, \qquad i, j \neq k, l.$$

Kako je još $b_{ij} = a_{ij}$, $i, j \neq k, l$, znači da se suma kvadrata nedijagonalnih elemenata, iz koje su izuzeti elementi a_{kl} i a_{lk}, pri transformaciji (11) ne menja. Takođe, iz (12) i činjenice da je $b_{kl} = b_{lk} = 0$, sledi da je

$$b_{kk}^2 + b_{ll}^2 = b_{kk}^2 + b_{kl}^2 + b_{lk}^2 + b_{ll}^2 = a_{kk}^2 + a_{kl}^2 + a_{lk}^2 + a_{ll}^2. \tag{15}$$

Dakle, suma kvadrata nedijagonalnih elemenata matrice B je za $2a_{kl}^2$ manja od iste sume matrice A, dok je suma kvadrata dijagonalnih elemenata matrice B za $2a_{kl}^2$ veća od sume kvadrata dijagonalnih elemenata matrice A. To znači da se euklidska norma matrice pri ovim transformacijama ne menja, ali i da se suma kvadrata dijagonalnih elemenata povećava na svakom koraku za onoliko za koliko se smanjuje tzv. nedijagonalna norma matrice A,

$$\nu \equiv \Big(\sum_{\substack{i,j=1 \\ i \neq j}}^{n} a_{ij}^2 \Big)^{1/2}. \tag{16}$$

Dokažimo da nedijagonalne norme (16) matrica A_m, definisanih u (14), teže nuli kada $m \to \infty$, što znači da i svi nedijagonalni elementi teže nuli, tj. da $A_m \to D$.

Na prvom koraku unitarnim transformacijama (11) anuliraćemo sve nedijagonalne elemente matrice A za koje je

$$|a_{ij}| \geq \frac{\nu}{\sigma} \equiv c_1, \qquad i \neq j, \tag{17}$$

gde je σ data konstanta. Ako se izabere da je $\sigma \geq n$, postojaće bar jedan element matrice A koji zadovoljava uslov (17), jer bi u protivnom imali

$$\nu^2 = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} a_{ij}^2 < \sum_{\substack{i,j=1 \\ i \neq j}}^{n} c_1^2 = n(n-1)c_1^2 < n^2 c_1^2 \leq \sigma^2 c_1^2 = \nu^2, \tag{18}$$

što je nemoguće. Ustvari, anuliraju se bar dva elementa – a_{kl} i njemu simetričan a_{lk}, te je nedijagonalna norma dobijene matrice, prema (15),

$$\nu_1^2 = \nu^2 - \sum_{|a_{ij}| \geq c_1} a_{ij}^2 \leq \nu^2 - 2c_1^2 = \nu^2\big(1 - \frac{2}{\sigma^2}\big). \tag{19}$$

U sledećem koraku anulirajmo sve nedijagonalne elemente koji su po apsolutnoj vrednosti veći od $c_2 = \nu_1/\sigma$; dokaz da postoje bar dva takva elementa, ako je $\sigma \geq n$, je analogan (18). Anuliranjem ovih elemenata na već poznati način, dobija

se matrica unitarno slična prethodnoj, čija se nedijagonalna norma, na osnovu (19), ocenjuje izrazom

$$\nu_2^2 = \nu_1^2 - \sum_{|b_{ij}|\geq c_2} b_{ij}^2 \leq \nu_1^2 - 2c_2^2 \leq \nu_1^2 - 2\big(\frac{\nu_1}{\sigma}\big)^2 = \nu^2\big(1 - \frac{2}{\sigma^2}\big)^2.$$

Indukcijom zaključujemo da se nedijagonalna norma ν_m matrice A_m dobijene posle m ovakvih koraka, pri čemu je u svakom koraku izvršena bar jedna transformacija, ocenjuje nejednakošću

$$\nu_m^2 \leq \nu^2\big(1 - \frac{2}{\sigma^2}\big)^m, \tag{20}$$

i, stoga, $\nu_m \to 0$ kada $m \to \infty$. ∎

U praksi, ovaj beskonačan iterativni proces zamenjujemo konačnim algoritmom, pri čemu smatramo da je $A_m \approx D$ za dovoljno veliko m. Dijagonalni elementi matrice A_m su približno sopstvene vrednosti matrice A. Kriterijum na osnovu koga određujemo broj iteracija m za koji je greška manja od dozvoljene, se obično iskazuje odnosom nedijagonalnih normi matrica A_m i A,

$$\nu_m \leq \epsilon\nu \tag{21}$$

gde je ϵ željena tačnost. Ako sa c_m označimo konstantu od koje su manji svi nedijagonalni elementi matrice A_m, onda je

$$\nu_m^2 < n(n-1)c_m^2 < n^2c_m^2,$$

te je, na osnovu (21), tačnost postignuta ako su svi nedijagonalni elementi matrice A_m manji od

$$c_m = \frac{\epsilon}{n}\nu.$$

Broj koraka m koje je potrebno izvršiti da bi bio zadovoljen kriterijum (21) je, prema (20),

$$m \geq \frac{2\ln\epsilon}{\ln\big(1 - \frac{2}{\sigma^2}\big)}.$$

Ako je U_k matrica dobijena množenjem svih matrica rotacija oblika (10) korišćenih u k-tom koraku, onda je, s obzirom na (14), unitarna matrica U definisana lemom 5.6 jednaka

$$U = \lim_{m\to\infty} \prod_{k=1}^{m} U_k.$$

Vektori kolona matrice U su sopstveni vektori matrice A, te se aproksimacijom ovih vektora mogu smatrati vektori kolona matrice $\prod_{k=1}^{m} U_k$.

Mada nije naročito brza, metoda se široko primenjuje u praksi jer je numerički stabilna i pogodna za realizaciju na računaru.

PRIMER 3. Jacobijevom metodom se u dva koraka određuju tačno sopstvene vrednosti i sopstveni vektori matrice iz primera 2,

$$\begin{pmatrix} 2 & 1 & 1 \\ \boxed{1} & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix} \xrightarrow{U_{21}} \begin{pmatrix} 3 & 0 & \sqrt{2} \\ 0 & 1 & 0 \\ \boxed{\sqrt{2}} & 0 & 3 \end{pmatrix} \xrightarrow{U_{31}} \begin{pmatrix} 3+\sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3-\sqrt{2} \end{pmatrix}.$$

Zaokruživanjem su naznačeni elementi koji se anuliraju u odgovarajućem koraku, a matrice kojima se to postiže su

$$U_{21} = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad U_{31} = \begin{pmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{pmatrix}.$$

Vektori kolona matrice

$$U = U_{21} \cdot U_{31} = \begin{pmatrix} 1/2 & -1/\sqrt{2} & -1/2 \\ 1/2 & 1/\sqrt{2} & -1/2 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{pmatrix}$$

su jedinični sopstveni vektori,

$$\begin{array}{lll} \lambda_1 = 3+\sqrt{2} & \lambda_2 = 1 & \lambda_3 = 3-\sqrt{2} \\ \mathbf{x}_1 = \begin{pmatrix} 1/2 \\ 1/2 \\ 1/\sqrt{2} \end{pmatrix} & \mathbf{x}_2 = \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix} & \mathbf{x}_3 = \begin{pmatrix} -1/2 \\ -1/2 \\ 1/\sqrt{2} \end{pmatrix}. \end{array}$$

6.4 Householderova metoda

Ovo je metoda, kao i Givensova (§2), kojom se kvadratna matrica A dimenzije n transformiše pomoću konačno mnogo unitarnih transformacija sličnosti u gornju Hessenbergovu matricu (5) ili trodijagonalnu matricu (6), ako je A Hermiteova. Za razliku od Givensove metode u kojoj se svodi na nulu jedan po jedan element matrice, te se ona realizuje u $(n-2)(n-1)/2$ koraka, ovom metodom se u jednom koraku svode na nulu svi potrebni elementi odgovarajućeg vektora kolone tekuće matrice, te se ona realizuje u $(n-2)$ koraka.

Radi nalaženja matrice T_j, $j = 1, \ldots, n-2$, kojom se realizuje pomenuta transformacija u j-tom koraku, pođimo od Householderove matrice

$$T = I - 2\mathbf{w}\,\mathbf{w}^*, \tag{22}$$

gde je I jedinična matrica, a $\mathbf{w} \in \mathcal{C}^n$ i $\|\mathbf{w}\|_2^2 = \mathbf{w}^*\mathbf{w} = 1$. Matrica T je Hermiteova

$$T^* = I^* - (2\mathbf{w}\,\mathbf{w}^*)^* = I - 2\mathbf{w}\,\mathbf{w}^* = T,$$

i, s obzirom na pretpostavku da je $\mathbf{w}$ jedinični vektor, unitarna

$$T^*\,T = (I - 2\mathbf{w}\,\mathbf{w}^*)(I - 2\mathbf{w}\,\mathbf{w}^*) = I - 4\mathbf{w}\,\mathbf{w}^* + 4\mathbf{w}\,\mathbf{w}^*\,\mathbf{w}\,\mathbf{w}^* = I,$$

pa je

$$T^2 = I. \tag{23}$$

Ako je vektor $\mathbf{y}$ dobijen transformacijom vektora $\mathbf{x}$ matricom T,

$$\mathbf{y} = T\mathbf{x} = (I - 2\mathbf{w}\,\mathbf{w}^*)\mathbf{x} = \mathbf{x} - 2\mathbf{w}\,\mathbf{w}^*\,\mathbf{x} = \mathbf{x} - 2(\mathbf{x}\,,\,\mathbf{w})\mathbf{w}, \tag{24}$$

onda je

$$\|\mathbf{y}\|_2^2 = \mathbf{y}^*\mathbf{y} = (T\mathbf{x})^*(T\mathbf{x}) = \mathbf{x}^*T^*\,T\mathbf{x} = \mathbf{x}^*\mathbf{x} = \|\mathbf{x}\|_2^2, \tag{25}$$

$$(\mathbf{y}\,,\,\mathbf{x}) = \mathbf{x}^*\mathbf{y} = \mathbf{x}^*T\mathbf{x} = \mathbf{x}^*T^*\mathbf{x} = (T\mathbf{x})^*(\mathbf{x}^*)^* = (\mathbf{x}^*T\mathbf{x})^*. \tag{26}$$

Dakle, iz (25) sledi da se transformacijom matricom T euklidska norma vektora ne menja, a iz (26) da je skalarni proizvod $(\mathbf{y}\,,\,\mathbf{x})$ realan broj, jer je $\mathbf{x}^*T\mathbf{x} = (\mathbf{x}^*T\mathbf{x})^*$.

Odredimo vektor $\mathbf{w}$, a time i matricu T, tako da se dati vektor $\mathbf{x} = (x_1, \ldots, x_m)^T$ preslika u vektor kolinearan sa koordinatnim vektorom $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$,

$$T\mathbf{x} = k\,\mathbf{e}_1. \tag{27}$$

Konstanta k je, s obzirom na (25),

$$|k|^2 = \|\mathbf{x}\|_2^2 = \mathbf{x}^*\mathbf{x}. \tag{28}$$

Neka je $\sigma = \|\mathbf{x}\|_2$, i neka je prva koordinata vektora $\mathbf{x}$ oblika $x_1 = |x_1|e^{\imath\alpha}$. Zbog (26) je

$$(k\mathbf{e}_1\,,\,\mathbf{x}) = k\mathbf{x}^*\mathbf{e}_1 = k|x_1|e^{-\imath\alpha}$$

realan broj, te k mora imati činilac $e^{\imath\alpha}$. Kako iz (28) sledi da je $|k| = \sigma$, to je

$$k = \mp\sigma e^{\imath\alpha}. \tag{29}$$

Matrica T, tj. vektor $\mathbf{w}$, određeni su relacijama (27) i (24),

$$T\mathbf{x} = \mathbf{x} - 2(\mathbf{x}\,,\,\mathbf{w})\mathbf{w} = k\mathbf{e}_1,$$

odakle je, s obzirom da je $\mathbf{w}$ jedinični vektor,

$$\mathbf{w} = \frac{\mathbf{x} - k\mathbf{e}_1}{\|\mathbf{x} - k\mathbf{e}_1\|_2}. \tag{30}$$

Znak konstante k, date izrazom (29), izaberimo tako da imenilac izraza (30) ne bude blizak nuli, radi numeričke stabilnosti algoritma. Kako je

$$\begin{aligned}\|\mathbf{x}-k\mathbf{e}_1\|_2^2 &= (\mathbf{x}-k\mathbf{e}_1\,,\,\mathbf{x}-k\mathbf{e}_1) = \|\mathbf{x}\|_2^2 - (\mathbf{x}\,,\,k\mathbf{e}_1) - (k\mathbf{e}_1,\mathbf{x}) + \|k\mathbf{e}_1\|_2^2 \\ &= \sigma^2 \pm 2\sigma|x_1| + |k|^2 = 2\sigma(\sigma \pm |x_1|),\end{aligned} \tag{31}$$

a $\sigma > 0$, treba u (29) uzeti znak minus, tj. uzeti da je

$$k = -\sigma e^{\imath\alpha}.$$

Zamenom (30) u (22), uzimajući u obzir (31), dobijamo da je matrica T

$$T = I - \big(\sigma(\sigma+|x_1|)\big)^{-1}(\mathbf{x}-k\mathbf{e}_1)\,(\mathbf{x}-k\mathbf{e}_1)^*.$$

Dakle, Householderova matrica T kojom se proizvoljan m-dimenzioni vektor $\mathbf{x}$ preslikava u vektor kolinearan sa prvim koordinatnim vektorom $\mathbf{e}_1$ je

$$\begin{aligned}&T = I - \beta\mathbf{u}\,\mathbf{u}^*, \qquad \mathbf{u} = \mathbf{x} - k\mathbf{e}_1, \qquad \beta = \big(\sigma(\sigma+|x_1|)\big)^{-1} \\ &\sigma = \sqrt{\sum_{i=1}^{m}|x_i|^2}, \qquad x_1 = |x_1|e^{\imath\alpha}, \qquad k = -\sigma e^{\imath\alpha}.\end{aligned} \tag{32}$$

Unitarnom transformacijom definisanom formulama (32) vektor $\mathbf{x}$ se preslikava u vektor čija je samo prva koordinata različita od nule i čija je euklidska norma jednaka euklidskoj normi vektora $\mathbf{x}$. Stoga se u prvom koraku Householderovog algoritma Householderova matrica $T_1^{(n-1)}$ (gornji indeks ukazuje na njenu dimenziju) definiše vektorom $\mathbf{x} = (a_{21},\dots,a_{n1})^T$, određenim svim elementima prve kolone matrice A izuzev elementa a_{11}, a matrica transformacije T_1 dimenzije n se dobija dopunjavanjem matrice $T_1^{(n-1)}$ koordinatnim vektorima do n-dimenzione matrice. Prema dokazanom, svi elementi prve kolone matrice $T_1\,A$, izuzev prva dva, su nula,

$$T_1\,A = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & T_1^{(n-1)} & \\ 0 & & & \end{array}\right) \cdot \left(\begin{array}{c|ccc} a_{11} & a_{12} & \dots & a_{1n} \\ \hline a_{21} & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & * & \dots & * \end{array}\right)$$

$$= \left(\begin{array}{c|ccc} a_{11} & a_{12} & \dots & a_{1n} \\ \hline k & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \dots & * \end{array}\right).$$

Da bi matrica A_1, dobijena posle prvog koraka Householderovog algoritma, bila slična matrici A, potrebno je još matricu $T_1\,A$ pomnožiti sa desne strane matricom

T_1^{-1}. S obzirom na svojstvo (23) Householderovih matrica i strukturu matrice T_1, neposredno se dokazuje da je $T_1^{-1} = T_1$, te je

$$A_1 = T_1 \, A \, T_1.$$

U drugom koraku algoritma transformacija se vrši matricom

$$T_2 = \left(\begin{array}{cc|ccc} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \hline 0 & 0 & & & \\ \vdots & \vdots & & T_2^{(n-2)} & \\ 0 & 0 & & & \end{array}\right),$$

gde je $T_2^{(n-2)}$ Householderova matrica dimenzije $(n-2)$, formirana pomoću $(n-2)$-dimenzionog vektora određenog vektorom druge kolone matrice A_1, izuzimajući prve dve njegove koordinate. Nastavljajući ovaj postupak, posle $(n-2)$ koraka dobija se gornja Hessenbergova matrica (5), ili trodijagonalna matrica (6) ako je A Hermiteova matrica, koja je slična polaznoj matrici A.

PRIMER 4. Za transformaciju matrice iz primera 2 u trodijagonalnu njoj sličnu matricu A_1 može se koristiti i Householderova metoda

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1/\sqrt{2} & -1/\sqrt{2} \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \quad \longrightarrow \quad A_1 \begin{pmatrix} 2 & -\sqrt{2} & 0 \\ -\sqrt{2} & 7/2 & -1/2 \\ 0 & -1/2 & 3/2 \end{pmatrix}.$$

Householderovom metodom se ne može dobiti trougaona, tj. dijagonalna, matrica slična polaznoj, jer bi se matricama T_j, $j = 1, \dots, n-1$, formiranim Householderovim matricama $T_j^{(n-j+1)}$ pokvarila željena struktura pri množenju inverznim matricama radi očuvanja sličnosti.

6.5 LR metoda

LR algoritam (Rutishauser, 1958) je iterativna metoda kojom se određuju sve sopstvene vrednosti kvadratne matrice A. Formira se, polazeći od date matrice $A_1 \equiv A$, niz matrica A_i na sledeći način. LU dekompozicijom (§5.2) matrica A_i se predstavi u obliku proizvoda donje trougaone matrice L_i, sa jedinicama na dijagonali, i gornje trougaone matrice R_i,

$$A_i = L_i \, R_i, \qquad L_i = \begin{pmatrix} 1 & & 0 \\ \vdots & \ddots & \\ * & \dots & 1 \end{pmatrix}, \quad R_i = \begin{pmatrix} * & \dots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix}. \tag{33}$$

Sledeći član niza matrica, matrica A_{i+1} jednaka je permutovanom proizvodu (33) nađenih trougaonih matrica, i takođe se razlaže LU dekompozicijom,

$$A_{i+1} = R_i\, L_i = L_{i+1}\, R_{i+1}, \qquad i = 1, 2, \ldots .$$

Pod pretpostavkom da se svaka matrica A_i može predstaviti proizvodom (33), bez prethodnog permutovanja vrsta ili kolona, dokažimo neke osobine ovog niza matrica.

TEOREMA 2. *Ako postoje sve dekompozicije $A_i = L_i\, R_i$, tada*

(i) matrica A_{i+1} je slična matrici A_i, tj.

$$A_{i+1} = L_i^{-1}\, A_i\, L_i;$$

(ii) $A_{i+1} = (L_1 \cdots L_i)^{-1}\, A_1\, (L_1 \cdots L_i), \quad i = 1, 2, \ldots;$

(iii) ako označimo sa T_i donje trougaonu matricu $T_i = L_1 \cdots L_i$ i sa U_i gornje trougaonu matricu $U_i = R_i \cdots R_1$, onda je

$$A^i \equiv A_1^i = T_i\, U_i, \qquad i = 1, 2, \ldots . \tag{34}$$

DOKAZ: (*i*) Kako je $A_i = L_i\, R_i$, tvrđenje neposredno sledi

$$L_i^{-1}\, A_i\, L_i = L_i^{-1}\, L_i\, R_i\, L_i = R_i\, L_i = A_{i+1}.$$

(*ii*) Ovo tvrđenje sledi neposredno na osnovu dokazanog pod (*i*),

$$\begin{aligned} A_{i+1} &= L_i^{-1}\, A_i\, L_i = L_i^{-1}(L_{i-1}^{-1}\, A_{i-1}\, L_{i-1})L_i = \cdots \\ &= L_i^{-1} \cdots L_1^{-1}\, A_1\, L_1 \cdots L_i = (L_1 \cdots L_i)^{-1}\, A_1\, (L_1 \cdots L_i). \end{aligned}$$

(*iii*) Iz (*ii*) sledi da je

$$L_1 \cdots L_i\, A_{i+1} = A_1\, L_1 \cdots L_i, \qquad i = 1, 2, \ldots,$$

pa je

$$\begin{aligned} T_i\, U_i &= L_1 \cdots L_{i-1}(L_i\, R_i)R_{i-1} \cdots R_1 = (L_1 \cdots L_{i-1}A_i)\, R_{i-1} \cdots R_1 \\ &= A_1(L_1 \cdots L_{i-1})(R_{i-1} \cdots R_1) = A_1\, T_{i-1}\, U_{i-1}, \end{aligned}$$

i, konačno,

$$T_i\, U_i = A_1\, T_{i-1}\, U_{i-1} = A_1^2\, T_{i-2}\, U_{i-2} = \cdots = A_1^i \equiv A^i.$$

∎

S obzirom da su dijagonalni elementi svih donje trougaonih matrica L_i jednaki jedinici, i dijagonalni elementi matrica T_i su jednaki jedinici. Stoga je teoremom 2 određena trougaona dekompozicija matrica A^i, $i = 1, 2, \ldots$.

Pod određenim pretpostavkama o matrici A može se dokazati da je

$$\lim_{i\to\infty} A_i = \lim_{i\to\infty} R_i = \begin{pmatrix} \lambda_1 & \ldots & * \\ & \ddots & \vdots \\ 0 & & \lambda_n \end{pmatrix}, \qquad \lim_{i\to\infty} L_i = I,$$

gde su λ_i, $i = 1, \ldots, n$, sopstvene vrednosti matrice A. Pokažimo to, uvodeći u toku dokaza neophodne pretpostavke. Kao prvo, neka je

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|. \tag{35}$$

Polazeći od vektora $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$, formirajmo niz vektora određenih matricom A,

$$\mathbf{z}_i = A^i \mathbf{e}_1, \qquad i = 0, 1, \ldots . \tag{36}$$

Da bi analizirali ponašanje vektora $\mathbf{z}_i$ kada $i \to \infty$, predstavimo $\mathbf{e}_1$ linearnom kombinacijom sopstvenih vektora $\mathbf{x}_k$, $k = 1, \ldots, n$, matrice A

$$\mathbf{e}_1 = \alpha_1 \mathbf{x}_1 + \cdots + \alpha_n \mathbf{x}_n, \qquad \text{gde je} \quad A\mathbf{x}_k = \lambda_k \mathbf{x}_k. \tag{37}$$

Tada je

$$\mathbf{z}_i = A^i(\alpha_1 \mathbf{x}_1 + \cdots + \alpha_n \mathbf{x}_n) = \alpha_1 \lambda_1^i \mathbf{x}_1 + \cdots + \alpha_n \lambda_n^i \mathbf{x}_n, \tag{38}$$

pa je , za $\alpha_1 \neq 0$,

$$\lim_{i\to\infty} \frac{1}{\lambda_1^i} \mathbf{z}_i = \alpha_1 \mathbf{x}_1, \tag{39}$$

jer je, s obzirom na (35), $\lim_{i\to\infty} \left(\frac{\lambda_k}{\lambda_1}\right)^i = 0$, $k = 2, \ldots, n$. Sa druge strane, iz (36) i (34) je

$$\mathbf{z}_i = A^i \mathbf{e}_1 = T_i\, U_i \mathbf{e}_1 = \begin{pmatrix} \mathbf{t}_1^{(i)} & \ldots & \mathbf{t}_n^{(i)} \end{pmatrix} \cdot \begin{pmatrix} u_{11}^{(i)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = u_{11}^{(i)} \mathbf{t}_1^{(i)},$$

gde su $\mathbf{t}_j^{(i)}$, $j = 1, \ldots, n$, vektori kolona matrice T_i, a $u_{jk}^{(i)}$ elementi matrice U_i. Kako je $U_i = R_i \cdots R_1$ i kako su sve matrice $R_i = (r_{jk}^{(i)})$ gornje trougaone, sledi da je $u_{11}^{(i)} = r_{11}^{(1)} \cdots r_{11}^{(i)}$, pa je

$$\mathbf{z}_i = r_{11}^{(1)} \cdots r_{11}^{(i)} \mathbf{t}_1^{(i)}. \tag{40}$$

Pri tome je još vektor $\mathbf{t}_1^{(i)}$, za svako $i = 1, 2, \ldots$, oblika

$$\mathbf{t}_1^{(i)} = (1, t_{21}^{(i)}, \ldots, t_{n1}^{(i)})^T, \tag{41}$$

tj. njegova prva koordinata je jedan za svako i, jer je T_i donje trougaona matrica sa jediničnim dijagonalnim elementima. Stoga je prva koordinata vektora $\mathbf{z}_i$ $r_{11}^{(1)} \cdots r_{11}^{(i)}$. Iz (39) sledi da je

$$\mathbf{z}_{i+1} \sim \lambda_1 \mathbf{z}_i,$$

te, ako prva koordinata x_{11} vektora $\mathbf{x}_1 = (x_{11}, \ldots, x_{n1})^T$ nije nula, količnik prvih koordinata vektora $\mathbf{z}_{i+1} = (r_{11}^{(1)} \cdots r_{11}^{(i+1)}, *, \ldots, *)^T$ i $\mathbf{z}_i = (r_{11}^{(1)} \cdots r_{11}^{(i)}, *, \ldots, *)^T$ teži ka λ_1,

$$\lim_{i\to\infty} r_{11}^{(i)} = \lambda_1. \tag{42}$$

Zamenom (40) u (39) i uzimajući u obzir (42), dokazujemo da je granica niza vektora $\mathbf{t}_1^{(i)}$ sopstveni vektor matrice A koji odgovara sopstvenoj vrednosti λ_1,

$$\lim_{i\to\infty} \frac{r_{11}^{(1)} \cdots r_{11}^{(i)}}{\lambda_1^i} \mathbf{t}_1^{(i)} = \mathbf{t}_1 \neq 0, \qquad \text{gde je} \quad A\mathbf{t}_1 = \lambda_1 \mathbf{t}_1.$$

Ako je prva koordinata x_{11} vektora $\mathbf{x}_1$ jednaka nuli, prema (38) je prva koordinata vektora $\mathbf{z}_i$ oblika $\alpha_2 \lambda_2^i x_{12} + \cdots + \alpha_n \lambda_n^i x_{1n}$, te prva koordinata vektora $\frac{1}{\lambda_1^i}\mathbf{z}_i$, s obzirom na (35), teži nuli kada $i \to \infty$. Sa druge strane je

$$\frac{1}{\lambda_1^i}\mathbf{z}_i = \frac{r_{11}^{(1)} \cdots r_{11}^{(i)}}{\lambda_1^i} \mathbf{t}_1^{(i)},$$

gde je vektor $\mathbf{t}_1^{(i)}$ dat u (41), što znači da je

$$\lim_{i\to\infty} \frac{r_{11}^{(1)} \cdots r_{11}^{(i)}}{\lambda_1^i} = 0,$$
$$\lim_{i\to\infty} \frac{r_{11}^{(1)} \cdots r_{11}^{(i)}}{\lambda_1^i} t_{k1}^{(i)} = \alpha_1 x_{k1}, \quad k = 2, \ldots, n.$$

Dakle, niz vektora $\mathbf{t}_1^{(i)}$ divergira,

$$\lim_{i\to\infty} \|\mathbf{t}_1^{(i)}\|_2 = \infty,$$

što znači da i niz matrica $T_i = L_1 \cdots L_i$ divergira.

Ako je u (37) $\alpha_1 = 0$ i $\alpha_2 \neq 0$, umesto (39) imamo

$$\lim_{i\to\infty} \frac{1}{\lambda_2^i}\mathbf{z}_i = \alpha_2 \mathbf{x}_2.$$

Stoga, u slučaju da je prva koordinata x_{12} vektora $\mathbf{x}_2 = (x_{12}, \ldots, x_{n2})^T$ različita od nule, imamo konvergenciju

$$\lim_{i\to\infty} r_{11}^{(i)} = \lambda_2, \qquad \lim_{i\to\infty} \mathbf{t}_1^{(i)} = \mathbf{t}_2, \qquad \text{gde je} \quad A\mathbf{t}_2 = \lambda_2 \mathbf{t}_2,$$

dok u suprotnom niz T_i divergira.

Navedene uslove o konvergenciji procesa možemo iskazati i na sledeći način. Neka je X matrica čiji su vektori kolona $\mathbf{x}_k$ sopstveni vektori matrice A,

$$X = (x_{jk}) = \begin{pmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{pmatrix}, \qquad A\mathbf{x}_k = \lambda_k \mathbf{x}_k, \quad k = 1, \ldots, n.$$

Tada je

$$A\,X = X\,D, \qquad \text{gde je} \quad D = diag\,(\lambda_1, \ldots, \lambda_n),$$

pa je $X^{-1}\,A = D\,X^{-1}$. Ako sa $\mathbf{y}_k^T = (y_{k1}, \ldots, y_{kn})$ označimo vektor k-te vrste matrice $Y = (y_{jk}) = X^{-1}$, iz $Y\,X = I$ i $Y\,A = D\,Y$ je

$$\mathbf{y}_j^T \mathbf{x}_k = \begin{cases} 0, & j \neq k \\ 1, & j = k \end{cases} \qquad \text{i} \qquad \mathbf{y}_j^T A = \lambda_j \mathbf{y}_j^T.$$

Stoga, množenjem vektora (37) sa $\mathbf{y}_j^T$ dobijamo da je

$$y_{j1} = \mathbf{y}_j^T \mathbf{e}_1 = \alpha_j, \qquad j = 1, \ldots, n.$$

S obzirom na prethodnu analizu sledi da, od toga da li je y_{j1}, $j = 1, \ldots, n$, jednako ili različito od nule, zavisi kojoj sopstvenoj vrednosti matrice A niz $r_{11}^{(i)}$ konvergira, ako uopšte konvergira. Da li proces konvergira ili ne zavisi od toga da li x_{1k} nije ili jeste nula. U procesu koji konvergira, ponašanje ostalih kolona matrica T_i i R_i, a time i L_i, R_i i A_i, može se analizirati na sličan način.

Ovi zaključi su objedinjeni u teoremi koja sledi, a kojom se formulišu opšti uslovi konvergencije LR algoritma.

TEOREMA 3. *Neka je $A \equiv A_1$ kvadratna matrica dimenzije n koja ispunjava sledeće pretpostavke:*

(i) LR metoda se može primeniti na matricu A, tj. za svako $i = 1, 2, \ldots,$ dekompozicija $A_i = L_i\,R_i$, a time i matrica $A_{i+1} = R_i\,L_i$, postoji;

(ii) sopstvene vrednosti matrice A zadovoljavaju relaciju

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|;$$

(iii) Za matrice X i $Y = X^{-1}$, takve da je $A = X\,D\,Y$, $D = diag\,(\lambda_1, \ldots, \lambda_n)$, moguća je trougaona dekompozicija

$$X = L_x\,R_x, \qquad Y = L_y\,R_y$$

pri čemu su dijagonalni elementi matrica L_x i L_y jednaki jedinici.

Tada nizovi matrica $\{A_i\}$, $\{R_i\}$ i $\{L_i\}$ konvergiraju, i

$$\lim_{i\to\infty} A_i = \lim_{i\to\infty} R_i = \begin{pmatrix} \lambda_1 & \dots & * \\ & \ddots & \vdots \\ 0 & & \lambda_n \end{pmatrix}, \qquad \lim_{i\to\infty} L_i = I.$$

Hipoteza (*iii*) teoreme, koja se odnosi na matrice $X = (x_{ij})$ i $Y = (y_{ij})$, obezbeđuje da je $x_{11} \neq 0$ i $y_{11} \neq 0$. Postojanje matrica L_x i R_x garantuje konvergenciju metode.

DOKAZ: Pri dokazu ćemo još pretpostaviti da je $\lambda_n \neq 0$.
Prvo ćemo indukcijom dokazati da je

$$A^i = X\, D^i\, Y.$$

Za $i = 1$ iskaz je tačan prema hipotezi (*iii*) teoreme. Pretpostavimo da je tačan za $i = k$, tj. da je $A^k = X\, D^k\, Y$. Tada je, s obzirom da je $Y = X^{-1}$,

$$A^{k+1} = A^k\, A = (X\, D^k\, Y)(X\, D\, Y) = X\, D^k (Y\, X) D\, Y = X\, D^{k+1}\, Y.$$

Dalje, pod pretpostavkom da D^{-1} postoji, je

$$A^i = X\, D^i Y = L_x R_x D^i L_y R_y = L_x R_x (D^i L_y D^{-i}) D^i R_y. \tag{43}$$

Matrica $D^i L_y D^{-i} = (l_{jk}^{(i)})$ je donje trougaona matrica sa jedinicama na dijagonali,

$$D^i L_y D^{-i} = \begin{pmatrix} 1 & & & & 0 \\ \left(\frac{\lambda_2}{\lambda_1}\right)^i l_{21} & 1 & & & \\ \left(\frac{\lambda_3}{\lambda_1}\right)^i l_{31} & \left(\frac{\lambda_3}{\lambda_2}\right)^i l_{32} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \left(\frac{\lambda_n}{\lambda_1}\right)^i l_{n1} & \left(\frac{\lambda_n}{\lambda_2}\right)^i l_{n2} & \left(\frac{\lambda_n}{\lambda_3}\right)^i l_{n3} & \dots & 1 \end{pmatrix},$$

gde je $L_y = (l_{jk})$. Prema hipotezi (*ii*) teoreme je $|\lambda_j| < |\lambda_k|$ za $j > k$, te je

$$\lim_{i\to\infty} l_{jk}^{(i)} = \lim_{i\to\infty} \left(\frac{\lambda_j}{\lambda_k}\right)^i l_{jk} = 0.$$

Stoga se može napisati da je

$$D^i L_y D^{-i} = I + E_i, \qquad \text{gde je} \quad \lim_{i\to\infty} E_i = 0,$$

što zamenom u (43) daje

$$A^i = L_x R_x (I + E_i) D^i R_y = L_x (I + R_x E_i R_x^{-1}) R_x D^i R_y = L_x (I + F_i) R_x D^i R_y,$$

gde je

$$F_i = R_x E_i R_x^{-1}, \qquad \lim_{i\to\infty} F_i = 0. \tag{44}$$

Za $i \geq i_0$ dovoljno veliko, postoji trougaona dekompozicija matrice $I + F_i$,

$$I + F_i = \tilde{L}_i \tilde{R}_i, \qquad \tilde{L}_i = \begin{pmatrix} 1 & & 0 \\ \vdots & \ddots & \\ * & \dots & 1 \end{pmatrix}, \quad \tilde{R}_i = \begin{pmatrix} * & \dots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix},$$

pri čemu je, zbog (44),

$$\lim_{i\to\infty} \tilde{L}_i = \lim_{i\to\infty} \tilde{R}_i = I. \tag{45}$$

Matrica A^i je predstavljena u obliku

$$A^i = (L_x \tilde{L}_i)(\tilde{R}_i R_x D^i R_y),$$

pri čemu je proizvod $L_x \tilde{L}_i$ donje, a proizvod $\tilde{R}_i R_x D^i R_y$ gornje trougaona matrica. Kako je trougaona dekompozicija regularne matrice jednoznačno određena, to je

$$T_i = L_1 \cdots L_i = L_x \tilde{L}_i, \qquad U_i = R_i \cdots R_1 = \tilde{R}_i R_x D^i R_y.$$

Kada $i \to \infty$, s obzirom na (45), je

$$\begin{aligned}
\lim_{i\to\infty} T_i &= L_x, \\
\lim_{i\to\infty} L_i &= \lim_{i\to\infty} T_{i-1}^{-1} T_i = I, \\
\lim_{i\to\infty} R_i &= \lim_{i\to\infty} U_i U_{i-1}^{-1} = \lim_{i\to\infty} (\tilde{R}_i R_x D^i R_y)(\tilde{R}_{i-1} R_x D^{i-1} R_y)^{-1} \\
&= \lim_{i\to\infty} \tilde{R}_i R_x D^i R_y R_y^{-1} D^{-i+1} R_x^{-1} \tilde{R}_{i-1}^{-1} = \lim_{i\to\infty} \tilde{R}_i R_x D\, R_x^{-1} \tilde{R}_{i-1}^{-1} \\
&= R_x D\, R_x^{-1}.
\end{aligned}$$

Međutim, $R_x D\, R_x^{-1}$ je gornje trougaona matrica oblika

$$\begin{pmatrix} \lambda_1 & \dots & * \\ & \ddots & \vdots \\ 0 & & \lambda_n \end{pmatrix},$$

čime je teorema u potpunosti dokazana. ∎

Iz dokaza teoreme je jasno da je konvergencija brža ukoliko je $|\frac{\lambda_j}{\lambda_k}|$, $j > k$, manje, odnosno ako su sopstvene vrednosti više razdvojene. Ako su neke sopstvene vrednosti jednake po modulu,

$$|\lambda_1| > \cdots > |\lambda_r| = |\lambda_{r+1}| > \cdots > |\lambda_n|,$$

konvergencija će postojati, osim u naznačenoj oblasti matrice

$$A_i \xrightarrow{i\to\infty} \left(\begin{array}{ccc|cc|c|cc} \lambda_1 & \cdots & * & * & * & * & \cdots & * \\ & \ddots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ & & \lambda_{r-1} & * & * & \vdots & & \vdots \\ \hline & & & - & - & - & & \vdots \\ & & & * & * & \vdots & & \vdots \\ & & & * & * & \vdots & & \vdots \\ & & & - & - & * & & \vdots \\ & & & & & \lambda_{r+2} & & \vdots \\ & & & & & & \ddots & \vdots \\ 0 & & & & & & & \lambda_n \end{array}\right).$$

Sopstvene vrednosti blok matrice dimenzije dva konvergiraće ka λ_r i λ_{r+1}. Analogan zaključak se može izvesti i ako imamo više od dve jednake po modulu sopstvene vrednosti.

Nedostaci LR metode su:

(*i*) Skupa je, jer je broj množenja na jednom koraku (da se od A_i dobije A_{i+1}) asimptotski jednak $\frac{2}{3}n^3$;

(*ii*) Ne može se primeniti ako nije moguće izvršiti trougaonu dekompoziciju svih matrica A_i, i matrica X i Y;

(*iii*) Sporo konvergira ako je λ_j/λ_k blisko jedinici.

Da bi se ublažili nabrojani nedostaci, obično se čini sledeće. Broj operacija na jednom koraku se smanjuje primenom metode na redukovane matrice, tj. gornje Hessenbergove ili trodijagonalne matrice. Jednostavno se može pokazati da su ovako redukovane matrice invarijantne u odnosu na LR transformacije – ako je A_i gornja Hessenbergova ili trodijagonalna matrica, onda je takva i matrica A_{i+1}.

Konvergencija se može ubrzati tzv. pomeranjem. Za elemente matrice A_i važi ocena ([26])

$$a_{jk}^{(i)} = O\left(\left(\frac{\lambda_j}{\lambda_k}\right)^i\right), \quad j > k. \tag{46}$$

Ako je A gornja Hessenbergova matrica, onda su takve i sve matrice A_i. Stoga je $a^{(i)}_{n,n-1}$ jedini nedijagonalni element u poslednjoj vrsti matrice A_i koji nije nula, i, prema (46), on teži nuli istom brzinom kao i $(\lambda_n/\lambda_{n-1})^i$ kada $i \to \infty$. Da bi ubrzali konvergenciju, primenjujemo LR metodu na matricu

$$\tilde{A} = A - p\,I,$$

gde je p aproksimacija sopstvene vrednosti λ_n. Element $\tilde{a}^{(i)}_{n,n-1}$ matrice $\tilde{A}_i$ se ponaša kao

$$\left(\frac{\lambda_n - p}{\lambda_{n-1} - p}\right)^i,$$

i brže konvergira ka nuli kada $i \to \infty$, jer je $\lambda_n - p$ blisko nuli. Ako je $|\lambda_n| < |\lambda_{n-1}|$, prema dokazanom je $\lim_{i\to\infty} a^{(i)}_{nn} = \lambda_n$, te se kao dobra aproksimacija vrednosti λ_n može uzeti $a^{(i)}_{nn}$, ako je i dovoljno veliko. Konstanta p se menja na svakom koraku i uzima se da je $p_i = a^{(i)}_{nn}$. Ako je λ_n višestruki ili kompleksni koren, onda se ova ocena dobija procenom sopstvenih vrednosti matrice niže dimenzije koja čini blok u matrici $\tilde{A}_i$. Kada je p_i ocenjeno, algoritam i-tog koraka LR metode je

$$A_i - p_i\,I = L_i\,R_i, \qquad A_{i+1} = R_i\,L_i + p_i\,I, \tag{47}$$

čime je obezbeđena sličnost matrica A_i i A_{i+1},

$$A_{i+1} = R_iL_i + p_iI = L_i^{-1}(A_i - p_iI)L_i + p_iI = L_i^{-1}A_iL_i - p_iL_i^{-1}L_i + p_iI = L_i^{-1}A_iL_i.$$

6.6 QR metoda

QR metoda (Francis, 1961) je slična LR metodi. Polazeći od kvadratne matrice $A_1 \equiv A$ reda n, formiraju se matrice A_i, Q_i i R_i na sledeći način:

$$\begin{aligned} &A_i = Q_i\,R_i, \qquad Q_i^*\,Q_i = I, \qquad R_i = \begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix}, \\ &A_{i+1} = R_i\,Q_i. \end{aligned} \tag{48}$$

Za razliku od LR metode koja se ne može uvek primeniti, u ovoj metodi se vrši dekompozicija matrice A_i na unitarnu Q_i i gornje trougaonu matricu R_i, što je uvek moguće realizovati. Druga prednost QR nad LR metodom je njena stabilnost – koriste se unitarne transformacije koje su stabilne, za razliku od LU dekompozicije. QR dekompozicija matrice A_i se može realizovati napr. Householderovom

metodom (§4) tako što se matricama $H_1^{(i)}, \ldots, H_{n-1}^{(i)}$, formiranim od Householderovih matrica (32), matrica A_i transformiše u gornje trougaonu matricu R_i,

$$H_{n-1}^{(i)} \cdots H_1^{(i)} A_i = R_i. \tag{49}$$

Iz (49) sledi da je

$$A_i = \left(H_{n-1}^{(i)} \cdots H_1^{(i)}\right)^{-1} R_i = H_1^{(i)} \cdots H_{n-1}^{(i)} R_i,$$

jer je $\left(H_j^{(i)}\right)^* = \left(H_j^{(i)}\right)^{-1} = H_j^{(i)}$. Stoga je u i-tom koraku

$$Q_i = H_1^{(i)} \cdots H_{n-1}^{(i)} \qquad \text{i} \qquad A_{i+1} = R_i\, Q_i = R_i\, H_1^{(i)} \cdots H_{n-1}^{(i)}.$$

Treba napomenuti da QR dekompozicija matrice nije jednoznačno određena. Naime, ako je S proizvoljna dijagonalna matrica oblika

$$S = \begin{pmatrix} e^{\imath\phi_1} & & 0 \\ & \ddots & \\ 0 & & e^{\imath\phi_n} \end{pmatrix}$$

($\imath$ ovde označava imaginarnu jedinicu), matrica $Q_i\, S$ je unitarna, jer je i S unitarna matrica. Kako je još $S^*\, R_i$ gornje trougaona matrica, to je izrazom

$$A_i = (Q_i\, S)(S^*\, R_i)$$

definisano beskonačno mnogo oblika QR dekompozicije matrice A_i (za različito ϕ).

Osobine matrica A_i, Q_i i R_i su formulisane u teoremi koja sledi, a koja je analogna teoremi 2.

TEOREMA 4. *Matrice A_i, Q_i i R_i QR algoritma (48), i matrice $P_i = Q_1 \cdots Q_i$ i $U_i = R_i \cdots R_1$ imaju sledeća svojstva:*

(i) A_{i+1} *i* A_i *su slične matrice, tj.* $A_{i+1} = Q_i^*\, A_i\, Q_i$;

(ii) $A_{i+1} = (Q_1 \cdots Q_i)^*\, A_1\, (Q_1 \cdots Q_i) = P_i^*\, A_1\, P_i$;

(iii) $A^i = P_i\, U_i, \qquad i = 1, 2, \ldots$.

Dokaz ove teoreme je analogan dokazu teoreme 2, u kome samo treba L_i zameniti sa Q_i.

Analiza konvergencije QR algoritma se može izvršiti onako kako je to urađeno za LR algoritam. Pretpostavimo da za sopstvene vrednosti λ_k matrice A važi (35), i neka su

$$X = (x_{jk}) = \begin{pmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_n \end{pmatrix}, \qquad Y = X^{-1} = (y_{jk}) = \begin{pmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \end{pmatrix},$$

matrice takve da je

$$A = X\,D\,Y, \qquad D = diag\,(\lambda_1, \ldots, \lambda_n).$$

$\mathbf{x}_k$ je desni, a $\mathbf{y}_k^T$ levi sopstveni vektor koji odgovaraju sopstvenoj vrednosti λ_k,

$$\begin{array}{lll} A\,X = X\,D & \Rightarrow & A\mathbf{x}_k = \lambda_k \mathbf{x}_k, \\ Y\,A = D\,Y & \Rightarrow & \mathbf{y}_k^T A = \lambda_k \mathbf{y}_k^T \end{array} \qquad \text{i} \qquad \mathbf{y}_j^T \mathbf{x}_k = \begin{cases} 0, & j \neq k \\ 1, & j = k \end{cases}.$$

Ako je $\alpha_1 = y_{11} \neq 0$ u razvoju koordinatnog vektora $\mathbf{e}_1$ po vektorima $\mathbf{x}_k$,

$$\mathbf{e}_1 = \alpha_1 \mathbf{x}_1 + \cdots + \alpha_n \mathbf{x}_n, \qquad \alpha_j = \mathbf{y}_j^T \mathbf{e}_1 = y_{j1},$$

onda za niz vektora $\mathbf{z}_i$, datih u (38), važi (39). Izrazimo vektor $\mathbf{z}_i$ pomoću matrica $P_i = \begin{pmatrix} \mathbf{p}_1^{(i)} & \ldots & \mathbf{p}_n^{(i)} \end{pmatrix}$ i $U_i = (u_{jk}^{(i)})$, definisanih u teoremi 4,

$$\mathbf{z}_i = A^i \mathbf{e}_1 = P_i\,U_i \mathbf{e}_1 = u_{11}^{(i)} \mathbf{p}_1^{(i)} = r_{11}^{(1)} \cdots r_{11}^{(i)} \mathbf{p}_1^{(i)}, \tag{50}$$

gde je, kao i ranije $R_i = (r_{jk}^{(i)})$. Matrica P_i je unitarna, te su njeni vektori kolona $\mathbf{p}_k^{(i)}$ jedinični u euklidskoj normi. Neka je i sopstveni vektor $\mathbf{x}_1$ normiran,

$$\mathbf{p}_1 = \frac{1}{\|\mathbf{x}_1\|_2} \mathbf{x}_1.$$

Tada, na osnovu (39), postoje fazni činioci $s_k = e^{\imath\phi_k}$ takvi da je

$$\lim_{k\to\infty} s_k \mathbf{p}_1^{(k)} = \mathbf{p}_1. \tag{51}$$

Iz (39) takođe sledi da je $\mathbf{z}_i \sim \lambda_1 \mathbf{z}_{i-1}$, te je na osnovu (50) $r_{11}^{(i)} \mathbf{p}_1^{(i)} \sim \lambda_1 \mathbf{p}_1^{(i-1)}$, što znači da se λ_1 može izračunati kao granična vrednost niza količnika odgovarajućih koordinata vektora $r_{11}^{(i)} \mathbf{p}_1^{(i)}$ i $\mathbf{p}_1^{(i-1)}$. Iz (51) je $\lim_{k\to\infty} \mathbf{p}_1^{(k)} = \lim_{k\to\infty} \frac{\mathbf{p}_1}{s_k}$, te je, konačno,

$$\lim_{k\to\infty} r_{11}^{(k)} \frac{s_{k-1}}{s_k} = \lambda_1.$$

Znači da $\mathbf{p}_1^{(k)}$ i $r_{11}^{(k)}$ ne konvergiraju u uobičajenom smislu ka $\mathbf{p}_1$ i λ_1, već "suštinski", tj. do na fazne činioce. Ova "suštinska" konvergencija postoji bez dodatnih pretpostavki, kakva je $x_{11} \neq 0$ u LR metodi.

Narednom teoremom su objedinjeni prethodni zaključci i dati opšti uslovi konvergencije QR metode.

TEOREMA 5. *Neka je $A \equiv A_1$ kvadratna matrica dimenzije n koja ispunjava sledeće pretpostavke:*

(i) sopstvene vrednosti matrice A su različite po modulu,

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|;$$

(ii) Na matricu Y, takvu da je $A = X\,D\,Y$, gde je $D = diag\,(\lambda_1, \ldots, \lambda_n)$ i $X = Y^{-1}$, je moguće primeniti trougaonu dekompoziciju

$$Y = L_y\,R_y, \qquad L_y = \begin{pmatrix} 1 & & 0 \\ \vdots & \ddots & \\ * & \ldots & 1 \end{pmatrix}, \quad R_y = \begin{pmatrix} * & \ldots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix}.$$

Tada nizovi matrica $\{A_i\}$, $\{R_i\}$ i $\{Q_i\}$, kojima je definisana QR metoda (48), imaju sledeća svojstva: postoje fazne matrice

$$S_k = diag\,(e^{\imath\phi_1^{(k)}}, \ldots, e^{\imath\phi_n^{(k)}})$$

takve da je

$$\lim_{k\to\infty} S_{k-1}^*\,Q_k\,S_k = I, \tag{6.1}$$

$$\lim_{k\to\infty} S_{k-1}^*\,A_k\,S_{k-1} = \lim_{k\to\infty} S_k^*\,R_k\,S_{k-1} = \begin{pmatrix} \lambda_1 & \ldots & * \\ & \ddots & \vdots \\ 0 & & \lambda_n \end{pmatrix}. \tag{6.2}$$

Još je $\lim_{k\to\infty} a_{jj}^{(k)} = \lambda_j$, $j = 1, \ldots, n$, gde je $A_k = (a_{jj}^{(k)})$.

Dokaz ove teoreme je analogan dokazu teoreme 3.

Ako pretpostavka (*ii*) teoreme nije ispunjena, QR metoda će i dalje konvergirati, samo što niz sopstvenih vrednosti λ_i na dijagonali ne mora biti uređen.

Kao i LR metoda, i ova metoda se primenjuje samo na Hessenbergove matrice, Hermiteove trodijagonalne matrice i neke druge forme retkih matrica. Razlog leži isključivo u efikasnosti algoritma. Naime, za punu kvadratnu matricu dimenzije n broj operacija po iteraciji je $O(n^3)$, za Hessenbergovu matricu taj broj je $O(n^2)$, a za trodijagonalnu je $O(n)$.

Kada se metoda primenjuje na Hessenbergovu ili trodijagonalnu matricu, obzirom da je većina elemenata ispod glavne dijagonale nula, bolje je u (49) umesto Householderovih matrica $H_k^{(i)}$ koristiti matrice rotacije (7),

$$G_{n-1,n} \cdots G_{23}\,G_{12}\,A_i = R_i,$$

gde je sa $G_{k-1,k}$ označena matrica rotacije kojom se anulira element $a_{k,k-1}^{(i)}$ matrice A_i. U ovom slučaju je

$$A_i = Q_i\,R_i, \qquad Q_i = G_{12}^* G_{23}^* \cdots G_{n-1,n}^*$$
$$A_{i+1} = R_i\,Q_i = R_i\,G_{12}^* \cdots G_{n-1,n}^*.$$

PRIMER 5. Ilustrujmo jedan korak QR algoritma na matrici A_1, određenoj u primeru 4. Koristićemo Givensove matrice rotacije $G_{k-1,k}$, $k = 2, 3$, za nalaženje matrice R_1.

$$G_{12} = \begin{pmatrix} 0.8165 & -0.5774 & 0 \\ 0.5774 & 0.8165 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad G_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.9713 & -0.2379 \\ 0 & 0.2379 & 0.9713 \end{pmatrix}$$

$$R_1 = G_{23}G_{12}A_1 = \begin{pmatrix} 2.4495 & -3.1754 & 0.2887 \\ 0 & 2.1015 & -0.7534 \\ 0 & 0 & 1.3598 \end{pmatrix}$$

$$Q_1 = G_{12}^T G_{23}^T = \begin{pmatrix} 0.8165 & 0.5608 & 0.1374 \\ -0.5774 & 0.7931 & 0.1943 \\ 0 & -0.2379 & 0.9713 \end{pmatrix}$$

$$A_2 = R_1 Q_1 = \begin{pmatrix} 3.8335 & -1.2134 & 0 \\ -1.2134 & 1.8459 & -0.3235 \\ 0 & -0.3235 & 1.3208 \end{pmatrix}.$$

Kao u LR metodi, pomeranjem (47) se može ubrzati konvergencija ove metode. Za Hermiteove pozitivno definisane matrice je QR metoda dvostruko brža od LR metode, iako je na svakom koraku za realizaciju QR metode potrebno izvršiti dva puta više računskih operacija. U poređenju sa metodom Jacobija, praksa pokazuje da je QR metoda oko četiri puta brža ako se traže i sopstvene vrednosti i sopstveni vektori, a oko deset puta brža ako se traže samo sopstvene vrednosti.

6.7 Delimičan problem sopstvenih vrednosti

U praksi je često potrebno odrediti samo jednu sopstvenu vrednost, najčešće najveću po modulu, i njoj odgovarajući sopstveni vektor, drugim rečima, rešiti delimičan problem sopstvenih vrednosti. Metode navedene u prethodnim odeljcima su nepodesne za rešavanje ovakvih zadataka jer su suviše komplikovane i skupe, a većina rezultata koje njima dobijamo nam nije od koristi. Stoga postoje posebne metode za rešavanje delimičnog problema, koje su po pravilu iterativne i zasnivaju se na lemi 5.4.

Metoda proizvoljnog vektora. Ovom metodom se određuje najveća po modulu sopstvena vrednost λ_1 i odgovarajući sopstveni vektor $\mathbf{x}_1$ matrice $A \in \mathcal{C}^{n\times n}$, za čije sopstvene vrednosti važi

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Ako su njima odgovarajući sopstveni vektori $\mathbf{x}_1, \ldots, \mathbf{x}_n$ linearno nezavisni, oni obrazuju bazu prostora $\mathcal{C}^n$, te se proizvoljan vektor $\mathbf{v}_0$ može izraziti u obliku

$$\mathbf{v}_0 = \alpha_1 \mathbf{x}_1 + \cdots + \alpha_n \mathbf{x}_n.$$

Vektori

$$\mathbf{v}_k = A\mathbf{v}_{k-1} = A^k \mathbf{v}_0, \qquad k = 1, 2, \ldots, \tag{52}$$

s obzirom na lemu 5.4, imaju reprezentaciju

$$\mathbf{v}_k = \lambda_1^k \alpha_1 \mathbf{x}_1 + \cdots + \lambda_n^k \alpha_n \mathbf{x}_n, \qquad k = 0, 1, \ldots. \tag{53}$$

Stoga je

$$\frac{v_j^{(k+1)}}{v_j^{(k)}} = \lambda_1 \frac{\alpha_1 x_j^{(1)} + \left(\frac{\lambda_2}{\lambda_1}\right)^{k+1} \alpha_2 x_j^{(2)} + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^{k+1} \alpha_n x_j^{(n)}}{\alpha_1 x_j^{(1)} + \left(\frac{\lambda_2}{\lambda_1}\right)^{k} \alpha_2 x_j^{(2)} + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^{k} \alpha_n x_j^{(n)}}, \tag{54}$$

gde je

$$\mathbf{x}_i = (x_1^{(i)}, \ldots, x_n^{(i)})^T, \qquad \mathbf{v}_k = (v_1^{(k)}, \ldots, v_n^{(k)})^T.$$

(*i*) Ako je najveća po modulu sopstvena vrednost jednostruka i realna, tj. $|\lambda_1| > |\lambda_2|$, i $\alpha_1 \neq 0$, tada je bar za jedno j, $1 \leq j \leq n$

$$\lim_{k \to \infty} \frac{v_j^{(k+1)}}{v_j^{(k)}} = \lambda_1,$$

a zbog (53)

$$\mathbf{v}_k \sim \lambda_1^k \alpha_1 \mathbf{x}_1, \qquad k \to \infty. \tag{55}$$

Ako je $\alpha_1 = 0$, $\alpha_2 \neq 0$ i $|\lambda_2| > |\lambda_3|$, iz (54) sledi da je

$$\lim_{k \to \infty} \frac{v_j^{(k+1)}}{v_j^{(k)}} = \lambda_2.$$

U praksi se, međutim, usled računske greške, posle nekoliko iteracija pojavljuje komponenta vektora $\mathbf{v}_k$ u pravcu vektora $\mathbf{x}_1$, tako da količnik $v_j^{(k+1)}/v_j^{(k)}$ ipak konvergira ka λ_1, mada znatno sporije.

(*ii*) Ako je najveća po modulu sopstvena vrednost višestruka i realna, tj. $\lambda_1 = \cdots = \lambda_m$ i $|\lambda_m| > |\lambda_{m+1}|$, tada količnik

$$\frac{v_j^{(k+1)}}{v_j^{(k)}} = \lambda_1 \frac{\alpha_1 x_j^{(1)} + \cdots + \alpha_m x_j^{(m)} + \left(\frac{\lambda_{m+1}}{\lambda_1}\right)^{k+1} \alpha_{m+1} x_j^{(m+1)} + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^{k+1} \alpha_n x_j^{(n)}}{\alpha_1 x_j^{(1)} + \cdots + \alpha_m x_j^{(m)} + \left(\frac{\lambda_{m+1}}{\lambda_1}\right)^{k} \alpha_{m+1} x_j^{(m+1)} + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^{k} \alpha_n x_j^{(n)}}$$

teži takođe ka λ_1 kada $k \to \infty$ za ono j za koje je $\alpha_1 x_j^{(1)} + \cdots + \alpha_m x_j^{(m)} \neq 0$.

(*iii*) Ako je $|\lambda_1| = |\lambda_2|$ i $\lambda_1 \neq \lambda_2$, niz $v_j^{(k+1)}/v_j^{(k)}$ ne konvergira. Na primer, ako je $\lambda_2 = -\lambda_1$ i $|\lambda_2| > |\lambda_3|$, tada je

$$\frac{v_j^{(k+1)}}{v_j^{(k)}} = \lambda_1 \frac{\alpha_1 x_j^{(1)} + (-1)^{k+1} \alpha_2 x_j^{(2)} + \left(\frac{\lambda_3}{\lambda_1}\right)^{k+1} \alpha_3 x_j^{(3)} + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^{k+1} \alpha_n x_j^{(n)}}{\alpha_1 x_j^{(1)} + (-1)^{k} \alpha_2 x_j^{(2)} + \left(\frac{\lambda_3}{\lambda_1}\right)^{k} \alpha_3 x_j^{(3)} + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^{k} \alpha_n x_j^{(n)}},$$

i ovaj količnik ne konvergira kada $k \to \infty$. Ali, ako je $\alpha_1 x_j^{(1)} \pm \alpha_2 x_j^{(2)} \neq 0$, količnik

$$\frac{v_j^{(k+2)}}{v_j^{(k)}} = \lambda_1^2 \frac{\alpha_1 x_j^{(1)} + (-1)^{k+2} \alpha_2 x_j^{(2)} + \left(\frac{\lambda_3}{\lambda_1}\right)^{k+2} \alpha_3 x_j^{(3)} + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^{k+2} \alpha_n x_j^{(n)}}{\alpha_1 x_j^{(1)} + (-1)^{k} \alpha_2 x_j^{(2)} + \left(\frac{\lambda_3}{\lambda_1}\right)^{k} \alpha_3 x_j^{(3)} + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^{k} \alpha_n x_j^{(n)}}$$

konvergira ka λ_1^2 kada $k \to \infty$.

PRIMER 6. Nađimo najveću po modulu sopstvenu vrednost matrice iz primera 2. (U primeru 3 je nađeno da je $\lambda_1 = 3 + \sqrt{2} = 4.414$.)

k	$v_1^{(k)}$	$v_2^{(k)}$	$v_3^{(k)}$	$v_1^{(k)}/v_1^{(k-1)}$	$v_2^{(k)}/v_2^{(k-1)}$	$v_3^{(k)}/v_3^{(k-1)}$
0	1	0	0			
1	2	1	1			
2	6	5	6	3	5	6
3	23	22	29	3.833	4.400	4.483
4	97	96	132	4.217	4.364	4.552
9	$1.5910 \cdot 10^5$	$1.5910 \cdot 10^5$	$2.2495 \cdot 10^5$	4.413	4.413	4.415
10	$7.0225 \cdot 10^5$	$7.0225 \cdot 10^5$	$9.9305 \cdot 10^5$	4.414	4.414	4.414

Odgovarajući sopstveni vektor je približno

$$\mathbf{x}_1 = c \begin{pmatrix} 7.0225 \cdot 10^5 \\ 7.0225 \cdot 10^5 \\ 9.9305 \cdot 10^5 \end{pmatrix} = \begin{pmatrix} 0.500 \\ 0.500 \\ 0.707 \end{pmatrix},$$

pri čemu je konstanta c određena tako da $\mathbf{x}_1$ bude jedinični vektor.

Metoda skalarnog proizvoda. Ova metoda predstavlja jednu varijantu metode proizvoljnog vektora. Za nalaženje najveće po modulu sopstvene vrednosti λ_1 koristi se, pored niza vektora $\mathbf{v}_k$ definisanih u (52), i niz

$$\mathbf{w}_k = A^*\mathbf{w}_{k-1} = (A^*)^k\mathbf{w}_0, \qquad k = 1, 2, \ldots,$$

gde je $\mathbf{w}_0$ proizvoljni vektor. Sopstvene vrednosti matrice A^* su, prema lemi 5.5, $\bar{\lambda}_1, \ldots, \bar{\lambda}_n$, a odgovarajući sopstveni vektori $\mathbf{y}_1, \ldots, \mathbf{y}_n$ su linearno nezavisni i mogu se normirati tako da je

$$(\mathbf{x}_i \, , \, \mathbf{y}_j) = \delta_{ij}. \tag{56}$$

Vektor $\mathbf{w}_0$ se može predstaviti linearnom kombinacijom vektora $\mathbf{y}_j$,

$$\mathbf{w}_0 = \beta_1\mathbf{y}_1 + \cdots + \beta_n\mathbf{y}_n,$$

te je

$$\mathbf{w}_k = \bar{\lambda}_1^k\beta_1\mathbf{y}_1 + \cdots + \bar{\lambda}_n^k\beta_n\mathbf{y}_n. \tag{57}$$

Skalarni proizvod vektora (53) i (57), uzimajući u obzir (56), je

$$\begin{aligned}(\mathbf{v}_k \, , \, \mathbf{w}_k) &= \left(\sum_{i=1}^n \lambda_i^k\alpha_i\mathbf{x}_i \, , \, \sum_{j=1}^n \bar{\lambda}_j^k\beta_j\mathbf{y}_j\right) \\ &= \sum_{i=1}^n\sum_{j=1}^n \lambda_i^k\alpha_i\lambda_j^k\bar{\beta}_j(\mathbf{x}_i \, , \, \mathbf{y}_j) = \sum_{i=1}^n \lambda_i^{2k}\alpha_i\bar{\beta}_i,\end{aligned}$$

i, slično,

$$(\mathbf{v}_{k-1} \, , \, \mathbf{w}_k) = \sum_{i=1}^n \lambda_i^{2k-1}\alpha_i\bar{\beta}_i.$$

Stoga količnik

$$\frac{(\mathbf{v}_k \, , \, \mathbf{w}_k)}{(\mathbf{v}_{k-1} \, , \, \mathbf{w}_k)} = \lambda_1\frac{\alpha_1\bar{\beta}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\alpha_2\bar{\beta}_2 + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^{2k}\alpha_n\bar{\beta}_n}{\alpha_1\bar{\beta}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^{2k-1}\alpha_2\bar{\beta}_2 + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^{2k-1}\alpha_n\bar{\beta}_n}$$

konvergira ka λ_1 kada $k \to \infty$ ako je $\alpha_1\bar{\beta}_1 \neq 0$.

Broj množenja potrebnih za izračunavanje jedne iteracije je u ovoj metodi dvostruko veći nego u metodi proizvoljnog vektora. Konvergencija se može ubrzati kada je $A = A^*$ izborom $\mathbf{w}_0 = \mathbf{v}_0$.

Metoda tragova. Ovom metodom se iterativnim algoritmom određuje najveća po modulu sopstvena vrednost λ_1 matrice A pomoću tragova $tr(A^k)$, $k = 1, 2, \ldots$. Naime, ako je

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|,$$

tada je, na osnovu (2),

$$\sqrt[k]{|tr(A^k)|} = |\lambda_1| \sqrt[k]{|1 + \left(\tfrac{\lambda_2}{\lambda_1}\right)^k + \cdots + \left(\tfrac{\lambda_n}{\lambda_1}\right)^k|} \to |\lambda_1|, \qquad k \to \infty.$$

Određujemo matrice A^2, $A^3, \ldots$ i njihove tragove, tj. zbirove njihovih dijagonalnih elemenata, pa nalazeći $\sqrt[k]{|tr(A^k)|}$, $k = 1, 2, \ldots$, dobijamo niz brojeva koji konvergira ka $|\lambda_1|$.

Druga varijanta ove metode je da se λ_1 odredi kao granična vrednost količnika

$$\frac{tr(A^{k+1})}{tr(A^k)} = \lambda_1 \frac{1 + \left(\tfrac{\lambda_2}{\lambda_1}\right)^{k+1} + \cdots + \left(\tfrac{\lambda_n}{\lambda_1}\right)^{k+1}}{1 + \left(\tfrac{\lambda_2}{\lambda_1}\right)^k + \cdots + \left(\tfrac{\lambda_n}{\lambda_1}\right)^k} \to \lambda_1, \qquad k \to \infty.$$

Kao i u prethodnim metodama, aproksimacija sopstvenog vektora je $A^k\mathbf{v}$, gde je $\mathbf{v}$ proizvoljni vektor a k dovoljno veliko, što sledi iz (55).

Metoda je sporija od prethodnih, jer je na svakom koraku potrebno pomnožiti dve matrice, a ne matricu i vektor. U prvoj varijanti konvergencija se može ubrzati ako se umesto niza matrica A, A^2, $A^3, A^4, \ldots$ formira niz A, A^2, A^4, $A^8, \ldots$ i određuje $\sqrt[2^k]{|tr(A^{2^k})|}$.

Metoda iscrpljivanja. Kada je određena najveća po modulu sopstvena vrednost matrice i njoj odgovarajući sopstveni vektor, ovom metodom se može naći sledeća po veličini modula sopstvena vrednost i njoj odgovarajući sopstveni vektor.

Neka sopstvene vrednosti matrice A zadovoljavaju relaciju

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|,$$

i neka su odgovarajući sopstveni vektori $\mathbf{x}_1, \ldots, \mathbf{x}_n$ linearno nezavisni. Pretpostavimo da su λ_1 i $\mathbf{x}_1$ poznati, a takođe da je poznat i sopstveni vektor $\mathbf{y}_1$ matrice A^* koji odgovara sopstvenoj vrednosti $\bar{\lambda}_1$ i koji je normiran tako da važi (56) za $j = 1$. Matrica

$$A_1 = A - \lambda_1 \mathbf{x}_1 \mathbf{y}_1^*$$

ima takođe sopstvene vektore $\mathbf{x}_1, \ldots, \mathbf{x}_n$, a odgovarajuće sopstvene vrednosti su $0, \lambda_2, \ldots, \lambda_n$, jer je

$$\begin{aligned} A_1\mathbf{x}_1 &= A\mathbf{x}_1 - \lambda_1 \mathbf{x}_1 \mathbf{y}_1^* \mathbf{x}_1 = \lambda_1 \mathbf{x}_1 - \lambda_1 \mathbf{x}_1 (\mathbf{x}_1 \,,\, \mathbf{y}_1) = 0 \cdot \mathbf{x}_1, \\ A_1\mathbf{x}_i &= A\mathbf{x}_i - \lambda_1 \mathbf{x}_1 \mathbf{y}_1^* \mathbf{x}_i = \lambda_i \mathbf{x}_i - \lambda_1 \mathbf{x}_1 (\mathbf{x}_i \,,\, \mathbf{y}_1) = \lambda_i \mathbf{x}_i, \quad i = 2, \ldots, n. \end{aligned}$$

Pošto je λ_2 najveća po modulu sopstvena vrednost matrice A_1, ona se može naći nekom od izloženih metoda.

PRIMER 7. Primenom metode iscrpljivanja i metode tragova odredimo drugu po veličini modula sopstvenu vrednost i njoj odgovarajući sopstveni vektor matrice

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix},$$

koristeći rezultat iz primera 6.

S obzirom da je A realna, simetrična matrica, tj. $A = A^*$, biće $\mathbf{x}_1 = \mathbf{y}_1$, te je

$$A_1 = A - \lambda_1 \mathbf{x}_1 \mathbf{x}_1^* = \begin{pmatrix} 0.8965 & -0.1035 & -0.5606 \\ -0.1035 & 0.8965 & -0.5606 \\ -0.5606 & -0.5606 & 0.7929 \end{pmatrix}.$$

Nalazeći tragove matrica A_1^2, A_1^4, A_1^8, A_1^{16} i A_1^{32}, dobijamo niz aproksimacija

$$|\lambda_2| : \quad 1.8747, \quad 1.6451, \quad 1.5907, \quad 1.5858, \quad 1.5858.$$

Množenjem proizvoljnog vektora sa A_1^{32} i normiranjem, dobijamo da je jedinični sopstveni vektor $\mathbf{x}_2 = (0.500, 0.500, -0.707)^T$, a znak sopstvene vrednosti λ_2 određujemo zamenom u sistem $A\mathbf{x}_2 = \lambda_2 \mathbf{x}_2$. (Poređenja radi, tačne vrednosti su date u primeru 3.)

Kada se odrede λ_2 i $\mathbf{x}_2$, ponavljanjem prethodnog postupka možemo odrediti matricu

$$A_2 = A_1 - \lambda_2 \mathbf{x}_2 (\mathbf{y}_2)^*$$

čija je najveća po modulu sopstvena vrednost λ_3, itd.

U praksi λ_1, $\mathbf{x}_1$ i $\mathbf{y}_1$ obično nisu tačno određeni. Stoga će i matrica A_1 biti određena sa izvesnom greškom, koja će uticati na tačnost izračunavanja λ_2 i $\mathbf{x}_2$, itd. Na taj način, svaka sledeća sopstvena vrednost i odgovarajući sopstveni vektor biće izračunati sa većom greškom nego prethodni. Stoga, ako se traži veći broj sopstvenih vrednosti, bolje je koristiti metode za rešavanje potpunog problema.

7

Nelinearne jednačine i sistemi

Opšti oblik sistema jednačina je

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_m) &= 0 \\ f_2(x_1, x_2, \ldots, x_m) &= 0 \\ &\vdots \\ f_m(x_1, x_2, \ldots, x_m) &= 0 \end{aligned} \tag{1}$$

što može kraće da se zapiše u vektorskom obliku

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}, \tag{2}$$

gde je $\mathbf{f} = (f_1, \ldots, f_m)^T$ i $\mathbf{x} = (x_1, \ldots, x_m)^T$. Dopustićemo i mogućnost da je $m = 1$, tako da je formulacijom (1), tj. (2), obuhvaćen i jednodimenzioni problem.

Euklidskom normom vektora $\mathbf{f}$ definisan je funkcional

$$F(\mathbf{x}) = \|\mathbf{f}(\mathbf{x})\|^2 = \mathbf{f}^T(\mathbf{x})\,\mathbf{f}(\mathbf{x}) = \sum_{i=1}^{m} f_i^2(\mathbf{x}), \tag{3}$$

koji je nenegativna funkcija po $\mathbf{x}$, $F(\mathbf{x}) \geq 0$, i jednak je nuli samo za ono $\mathbf{x}$ za koje je $f_i(\mathbf{x}) = 0$, $i = 1, \ldots, m$. Rešenje zadatka (1) je tačka minimuma funkcionala (3), te se za rešavanje sistema nelinearnih jednačina mogu koristiti i numeričke metode minimizacije.

Bilo da se problem (1) rešava direktno bilo minimizacijom odgovarajućeg funkcionala, koriste se numeričke metode koje spadaju u grupu iterativnih metoda. To znači da se, polazeći od proizvoljno izabranog vektora $\mathbf{x}_0$, određuje niz vektora $\mathbf{x}_1, \mathbf{x}_2, \cdots$ rekurentnom formulom

$$\mathbf{x}_{n+1} = G_n(\mathbf{x}_0, \ldots, \mathbf{x}_n), \qquad n = 0, 1, \ldots, \tag{4}$$

koji pod određenim uslovima konvergira ka rešenju zadatka. Ako G_n zavisi od $\mathbf{x}_n, \ldots, \mathbf{x}_{n-k+1}$, a ne zavisi od $\mathbf{x}_{n-k}, \ldots, \mathbf{x}_0$, iterativna metoda definisana formulom (4) je $(k+1)$-slojna metoda. Ako G_n ne zavisi od n, metoda je *stacionarna*, a ako je G_n linearna funkcija svojih argumenata, metoda je *linearna*.

Niz vektora $\mathbf{x}_n$, određenih formulom (4), konvergira ka vektoru $\mathbf{x}^* \in \mathcal{R}^m$ ako za svako $\epsilon > 0$ postoji prirodan broj $N(\epsilon)$ takav da je

$$\|\mathbf{x}_n - \mathbf{x}^*\| < \epsilon, \qquad \text{za svako } n \geq N(\epsilon).$$

Pri tome konvergencija ne zavisi od izbora norme $\|\cdot\|$ prostora $\mathcal{R}^m$, jer su sve one među sobom ekvivalentne (za proizvoljne dve norme $\|\cdot\|_1$ i $\|\cdot\|_2$ prostora $\mathcal{R}^m$ postoje konstante c i C takve da je $c\|\cdot\|_1 \leq \|\cdot\|_2 \leq C\|\cdot\|_1$).

Iterativna metoda (4) je reda p ukoliko je

$$\|\mathbf{x}_{n+1} - \mathbf{x}^*\| \leq C\|\mathbf{x}_n - \mathbf{x}^*\|^p, \qquad n = 0, 1, \ldots, \tag{5}$$

gde je C konstanta, koja u slučaju da je $p = 1$ zadovoljava uslov $C < 1$.

7.1 Teorema o nepokretnoj tački

Konvergencija familije dvoslojnih, stacionarnih iterativnih metoda reda jedan zasniva se na egzistenciji i jedinosti nepokretne tačke operatora kontrakcije.

Ako je G operator koji preslikava normirani prostor $\mathcal{B}$ u samoga sebe, onda je *nepokretna tačka* operatora G svaka tačka $x \in \mathcal{B}$ za koju je

$$x = G(x). \tag{6}$$

Operator G se naziva *operatorom kontrakcije* u zatvorenoj lopti

$$S(x_0, r) = \{x \,|\, \|x - x_0\| \leq r\} \subset \mathcal{B}$$

ako postoji realan broj q, $0 \leq q < 1$, takav da je za proizvoljno $x, y \in S(x_0, r)$

$$\|G(x) - G(y)\| \leq q\|x - y\|. \tag{7}$$

Konstanta q naziva se koeficijentom kontrakcije. Drugim rečima, operator G je operator kontrakcije ako je rastojanje slika manje od rastojanja originala.

TEOREMA 1. *Neka je G operator kontrakcije u $S(x_0, r)$ sa koeficijentom kontrakcije q, i neka je x_0 takvo da je*

$$\frac{1}{1-q}\|G(x_0) - x_0\| \equiv r_0 \leq r. \tag{8}$$

Tada

(*i*) niz $\{x_n\}$ određen rekurentnom formulom

$$x_{n+1} = G(x_n), \qquad n = 0, 1, \ldots \tag{9}$$

konvergira ka nekoj tački $x^* \in S(x_0, r_0)$;

(*ii*) x^* je nepokretna tačka operatora G, tj.

$$x^* = G(x^*); \tag{10}$$

(*iii*) x^* je jedinstvena nepokretna tačka operatora G u $S(x_0, r)$.

DOKAZ: Da bismo dokazali tvrđenje (*i*), dokažimo prvo da sve iteracije x_n, određene formulom (9), pripadaju lopti $S(x_0, r_0)$. Kako je, zbog (8) i uslova $q < 1$,

$$\|x_1 - x_0\| = \|G(x_0) - x_0\| = (1-q)r_0 \leq r_0, \tag{11}$$

to $x_1 \in S(x_0, r_0)$. Pretpostavimo da $x_k \in S(x_0, r_0)$, $k = 0, \ldots, n$. Tada je, na osnovu (7), (9) i (11),

$$\begin{aligned} \|x_{n+1} - x_n\| &= \|G(x_n) - G(x_{n-1})\| \leq q\|x_n - x_{n-1}\| \\ &\leq q^2\|x_{n-1} - x_{n-2}\| \leq \cdots \leq q^n\|x_1 - x_0\| = q^n(1-q)r_0, \end{aligned} \tag{12}$$

te je

$$\begin{aligned} \|x_{n+1} - x_0\| &\leq \|x_{n+1} - x_n\| + \|x_n - x_{n-1}\| + \cdots + \|x_1 - x_0\| \\ &\leq (1-q)(q^n + q^{n-1} + \cdots + 1)r_0 = (1 - q^{n+1})r_0 \leq r_0. \end{aligned}$$

To znači da i $x_{n+1} \in S(x_0, r_0)$, i na osnovu matematičke indukcije sledi da $x_k \in S(x_0, r_0)$ za svako $k = 0, 1, \ldots$, tj. da svi članovi niza određenog rekurentnom relacijom (9) pripadaju lopti $S(x_0, r_0)$.

Za proizvoljno $k > n$, na osnovu (12), je

$$\begin{aligned} \|x_k - x_n\| &\leq \|x_k - x_{k-1}\| + \|x_{k-1} - x_{k-2}\| + \cdots + \|x_{n+1} - x_n\| \\ &\leq q^n(1-q)(q^{k-n-1} + q^{k-n-2} + \cdots + 1)r_0, \end{aligned}$$

što znači da je niz $\{x_n\}$ Cauchyev niz, jer je

$$\|x_k - x_n\| \leq r_0 q^n (1-q) \sum_{j=0}^{k-n-1} q^j \leq r_0 q^n (1-q) \sum_{j=0}^{\infty} q^j = q^n r_0. \tag{13}$$

Stoga niz $\{x_n\}$ konvergira ka nekoj tački $x^* \in S(x_0, r_0)$, čime je tvrđenje (*i*) teoreme dokazano.

Da bismo dokazali tvrđenje (*ii*), tj. da je x^* upravo nepokretna tačka operatora G, ocenimo rastojanje tačaka x^* i $G(x^*)$. Na osnovu (13) je

$$\lim_{k \to \infty} \|x_k - x_n\| = \|x^* - x_n\| \leq q^n r_0, \tag{14}$$

što, zajedno sa nejednakošću (7), daje ocenu

$$\|G(x^*) - x^*\| \leq \|G(x^*) - G(x_n)\| + \|x_{n+1} - x^*\| \leq q\|x_n - x^*\| + \|x_{n+1} - x^*\| \leq 2q^{n+1} r_0.$$

Poslednja nejednakost važi za svako n, pa i kada $n \to \infty$, što je moguće samo ako je

$$\|G(x^*) - x^*\| = 0,$$

jer ova veličina ne zavisi od n. Odavde, na osnovu osobine norme, sledi tvrđenje (10).

Još je potrebno dokazati da je x^* jedina nepokretna tačka operatora G u lopti $S(x_0, r)$. Pretpostavimo suprotno, tj. da je i $\bar{x} \in S(x_0, r)$ nepokretna tačka operatora G. Na osnovu (7) je

$$\|x^* - \bar{x}\| = \|G(x^*) - G(\bar{x})\| \leq q\|x^* - \bar{x}\| < \|x^* - \bar{x}\|,$$

što je nemoguće, čime je teorema u potpunosti dokazana. ∎

PRIMER 1. Preslikavanje $f(x) = \frac{1}{2}x + 2$ jeste kontrakcija na odsečku $[0, 1]$ jer je za svako $x, y \in [0, 1]$

$$|f(x) - f(y)| = \frac{1}{2}|x - y|,$$

i koeficijent kontrakcije je $q = \frac{1}{2}$. Međutim, ovo preslikavanje nema nepokretnu tačku na odsečku $[0, 1]$ – jednačina $x = \frac{1}{2}x + 2$ ima rešenje $x = 4 \notin [0, 1]$. Teorema 1 se u ovom primeru ne može primeniti, jer uslov (8) nije ispunjen,

$$\forall x_0 \in [0, 1] \qquad x_1 = f(x_0) = \frac{1}{2}x_0 + 2 \notin [0, 1].$$

Jednačinu (2) treba transformisati u jednačinu oblika (6),

$$\mathbf{x} = \mathbf{g}(\mathbf{x}), \tag{15}$$

ili, u razvijenom obliku,

$$\begin{aligned} x_1 &= g_1(x_1, \ldots, x_m) \\ &\vdots \\ x_m &= g_m(x_1, \ldots, x_m), \end{aligned}$$

gde je $\mathbf{g} = (g_1, \ldots, g_m)^T$. Za jednačinu (15) metoda iteracije je definisana algoritmom

$$\mathbf{x}_{n+1} = \mathbf{g}(\mathbf{x}_n), \qquad n = 0, 1, \ldots,$$

tj.

$$\begin{aligned} x_1^{(n+1)} &= g_1(x_1^{(n)}, \ldots, x_m^{(n)}) \\ &\vdots \\ x_m^{(n+1)} &= g_m(x_1^{(n)}, \ldots, x_m^{(n)}). \end{aligned}$$

Uslovi konvergencije ove metode određeni su teoremom 1. S obzirom da transformacija jednačine (2) u jednačinu (15) nije jednoznačno određena, treba odabrati onu koja obezbeđuje da je preslikavanje $\mathbf{g}$ kontrakcija u nekoj okolini rešenja. Šta

praktično znači ovaj uslov? Pod pretpostavkom da je funkcija $\mathbf{g}(\mathbf{x})$ diferencijabilna u lopti $S(\mathbf{x}_0, r)$, iz Taylorove teoreme sledi da je za proizvoljne dve tačke $\mathbf{x}, \mathbf{y} \in S(\mathbf{x}_0, r)$

$$|g_i(\mathbf{x}) - g_i(\mathbf{y})| \leq \sum_{j=1}^{m} \Big| \frac{\partial g_i(\mathbf{z}_i)}{\partial x_j} \Big| \, |x_j - y_j|, \qquad i = 1, \ldots, m,$$

gde su $\mathbf{z}_i$, $i = 1, \ldots, m$, tačke iz $S(\mathbf{x}_0, r)$. Ako koristimo uniformnu vektorsku normu (5.7), onda je

$$\begin{aligned} \|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\| &\leq \max_{1 \leq i \leq m} \left(\sum_{j=1}^{m} \Big| \frac{\partial g_i(\mathbf{z}_i)}{\partial x_j} \Big| \, |x_j - y_j| \right) \\ &\leq \max_{1 \leq j \leq m} |x_j - y_j| \max_{1 \leq i \leq m} \left(\sum_{j=1}^{m} \Big| \frac{\partial g_i(\mathbf{z}_i)}{\partial x_j} \Big| \right) \\ &\leq \|\mathbf{x} - \mathbf{y}\| \max_{\mathbf{z} \in S(\mathbf{x}_0, r)} \|J(\mathbf{z})\|, \end{aligned} \tag{16}$$

gde je $J(\mathbf{x})$ Jakobijan preslikavanja $\mathbf{g}(\mathbf{x})$,

$$J(\mathbf{x}) = \begin{pmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_1(\mathbf{x})}{\partial x_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_m(\mathbf{x})}{\partial x_1} & \frac{\partial g_m(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_m(\mathbf{x})}{\partial x_m} \end{pmatrix},$$

a sa $\|J\|$ je označena uniformna norma matrice (5.11). Ocena analogna oceni (16) se može izvesti i korišćenjem drugih vektorskih i njima saglasnih matričnih normi. Dakle, da bi preslikavanje $\mathbf{g}(\mathbf{x})$ bilo kontrakcija, saglasno uslovu (7) dovoljno je da je

$$\max_{\mathbf{z} \in S(\mathbf{x}_0, r)} \|J(\mathbf{z})\| \leq q < 1.$$

Ako je dimenzija problema $m = 1$, Jakobijan preslikavanja je matrica dimenzije jedan sa elementom $g'(x)$, te se uslov da preslikavanje $g(x)$ bude kontrakcija svodi na uslov

$$\max_{x \in [a,b]} |g'(x)| \leq q < 1,$$

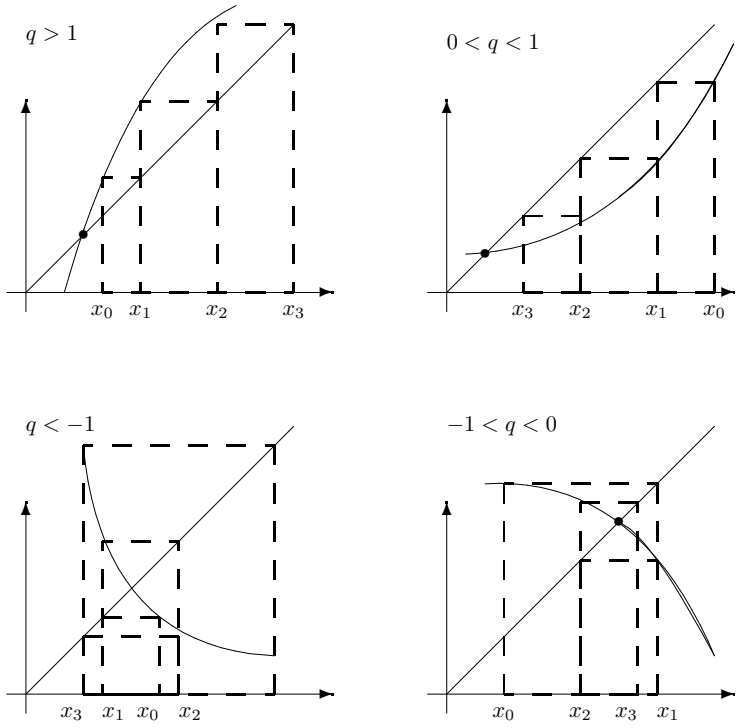
što ima jednostavnu geometrijsku interpretaciju, prikazanu na slici 7.1.

Apriorna ocena greške n-te iteracije se dobija neposredno iz relacija (14) i (8),

$$\|\mathbf{x}^* - \mathbf{x}_n\| \leq q^n r_0 = \frac{q^n}{1-q} \|\mathbf{g}(\mathbf{x}_0) - \mathbf{x}_0\|. \tag{17}$$

Iz ocene (17) se takođe može odrediti broj iteracija koje je potrebno izvršiti da bi se postigla zadata tačnost ϵ,

$$n \geq \frac{\ln \frac{\epsilon(1-q)}{\|\mathbf{g}(\mathbf{x}_0) - \mathbf{x}_0\|}}{\ln q}.$$



Slika 7.1: Geometrijska interpretacija metode iteracije.

Ocena (17) je izvedena korišćenjem većeg broja majoracija, te je dosta gruba u tom smislu da zahteva više izračunavanja nego što je stvarno potrebno da bi se postigla tražena tačnost. U tom smislu, bolja je aposteriorna ocena koja je izražena razlikom dve uzastopne iteracije. Naime, iz

$$\|\mathbf{x}_{n+1} - \mathbf{x}^*\| = \|\mathbf{g}(\mathbf{x}_n) - \mathbf{g}(\mathbf{x}^*)\| \leq q\big(\|\mathbf{x}_{n+1} - \mathbf{x}_n\| + \|\mathbf{x}_{n+1} - \mathbf{x}^*\|\big),$$

sledi da je

$$\|\mathbf{x}_{n+1} - \mathbf{x}^*\| \leq \frac{q}{1-q}\|\mathbf{x}_{n+1} - \mathbf{x}_n\|,$$

te $\mathbf{x}_{n+1}$ aproksimira rešenje jednačine (2) sa tačnošću ϵ ako je

$$\|\mathbf{x}_{n+1} - \mathbf{x}_n\| \leq \frac{1-q}{q}\epsilon.$$

Iz poslednje ocene je očigledno da je konvergencija utoliko sporija ukoliko je koeficijent kontrakcije q bliži jedinici.

PRIMER 2. Radi jednostavnosti, ilustrujmo metodu na jednodimenzionom problemu: naći rešenje jednačine

$$x - \sin x = 0.25$$

na tri sigurne cifre.

Analizom funkcije $f(x) \equiv x - \sin x - 0.25$, zaključujemo da se njen jedini realan koren nalazi u intervalu $[1.1, 1.3]$. Iterativni algoritam definisan formulom

$$x_{n+1} = \sin x_n + 0.25 \equiv g(x_n), \qquad n = 0, 1, \ldots,$$

će biti konvergentan za proizvoljno $x_0 \in [1.1, 1.3]$ jer je

$$\max_{[1.1,1.3]} |g'(x)| = 0.45,$$

pa je $g(x)$ operator kontrakcije sa koeficijentom kontrakcije $q = 0.45$. Kako rešenje pripada intervalu $[1.1, 1.3]$, a treba ga izračunati na tri sigurne cifre, tražena tačnost je $\epsilon = 0.5 \cdot 10^{-2}$. Ako za ocenu tačnosti koristimo rastojanje dve uzastopne iteracije, treba računati dok ne bude ispunjen uslov

$$|x_{n+1} - x_n| \leq 0.006.$$

Polazeći od početne aproksimacije rešenja $x_0 = 1.2$, dobijamo

$$x_0 = 1.2, \qquad x_1 = 1.182, \qquad x_2 = 1.175, \qquad x_3 = 1.173,$$

te je, s obzirom da je $|x_3 - x_2| < 0.006$, traženo rešenje $x^* = 1.17$.

Očigledno su, u smislu nejednakosti (5), metode ovoga tipa u opštem slučaju metode prvog reda, jer je

$$\|\mathbf{x}_{n+1} - \mathbf{x}^*\| = \|\mathbf{g}(\mathbf{x}_n) - \mathbf{g}(\mathbf{x}^*)\| \leq q\|\mathbf{x}_n - \mathbf{x}^*\|, \qquad n = 0, 1, \ldots .$$

U nekim slučajevima konvergencija se može ubrzati primenom tzv. *Gauss-Seidelove metode*

$$\begin{aligned}
x_1^{(n+1)} &= g_1(x_1^{(n)}, x_2^{(n)}, \ldots, x_{m-1}^{(n)}, x_m^{(n)}) \\
x_2^{(n+1)} &= g_2(x_1^{(n+1)}, x_2^{(n)}, \ldots, x_{m-1}^{(n)}, x_m^{(n)}) \\
&\vdots \\
x_m^{(n+1)} &= g_m(x_1^{(n+1)}, x_2^{(n+1)}, \ldots, x_{m-1}^{(n+1)}, x_m^{(n)}).
\end{aligned}$$

Ako je sistem jednačina (1) linearan, $\mathbf{f}(\mathbf{x}) \equiv A\mathbf{x} - \mathbf{b}$ (matrica A i vektor $\mathbf{b}$ ne zavise od $\mathbf{x}$), tada je i funkcija $\mathbf{g}$ u (15) takođe linearna funkcija

$$\mathbf{g}(\mathbf{x}) \equiv B\mathbf{x} + \mathbf{c},$$

gde su B matrica i $\mathbf{c}$ vektor koji ne zavise od $\mathbf{x}$. Jakobijan preslikavanja $\mathbf{g}$ je matrica B, i iterativna metoda

$$\mathbf{x}_{n+1} = B\mathbf{x}_n + \mathbf{c}, \qquad n = 0, 1, \ldots$$

konvergira ukoliko je u ma kojoj matričnoj normi $\|B\| < 1$.

S obzirom da je

$$\mathbf{x}_{n+1} - \mathbf{x}^* = B(\mathbf{x}_n - \mathbf{x}^*) = \cdots = B^{n+1}(\mathbf{x}_0 - \mathbf{x}^*),$$

iterativna metoda će konvergirati ako $B^n \to 0$ kada $n \to \infty$, tj. kada su sve sopstvene vrednosti λ_i matrice B po modulu manje od jedinice. Kako je za proizvoljnu sopstvenu vrednost λ i njoj odgovarajući sopstveni vektor $\mathbf{x}$ matrice B

$$|\lambda| \, \|\mathbf{x}\| = \|\lambda \mathbf{x}\| = \|B\mathbf{x}\| \leq \|B\| \, \|\mathbf{x}\|$$

u ma kojoj normi, to iz $\|B\| < 1$ sledi $|\lambda| < 1$ za svako λ, što znači da je uslov $\|B\| < 1$ dovoljan za konvergenciju iterativne metode. Veličina $\max_i |\lambda_i|$ naziva se *spektralnim radiusom* matrice B, te iterativna metoda konvergira ako je spektralni radius matrice B manji od jedan.

Ako matricu A linearnog sistema predstavimo u obliku sume

$$A = L + D + U,$$

gde je L donje, a U gornje trougaona matrica sa nulama na dijagonali, i D dijagonalna matrica, onda je tzv. *Jacobijeva iterativna metoda* definisana formulom

$$D\mathbf{x}_{n+1} = -(L + U)\mathbf{x}_n + \mathbf{b},$$

tj. formulom $\mathbf{x}_{n+1} = B\mathbf{x}_n + \mathbf{c}$ gde je $B = -D^{-1}(L + U)$ i $\mathbf{c} = D^{-1}\mathbf{b}$. Pomenuta *Gauss-Seidelova iterativna metoda* je data formulom

$$(L + D)\mathbf{x}_{n+1} = -U\mathbf{x}_n + \mathbf{b},$$

što znači da je $B = -(L + D)^{-1}U$ i $\mathbf{c} = (L + D)^{-1}\mathbf{b}$.

7.2 Newton–Raphsonova metoda

Kao što je napomenuto u prethodnom odeljku, jednačina (15) u koju se transformiše jednačina (2) nije jednoznačno određena – postoji više operatora $\mathbf{g}$ čija je nepokretna tačka rešenje jednačine (2). Na primer, operator $\mathbf{g}(\mathbf{x})$ može biti oblika

$$\mathbf{g}(\mathbf{x}) \equiv \mathbf{x} - D(\mathbf{x})\mathbf{f}(\mathbf{x}), \tag{18}$$

gde je $D(\mathbf{x})$ regularna matrica dimenzije $m \times m$. Očigledno je da je nepokretna tačka operatora (18) rešenje jednačine (2) i obrnuto, jer je $D(\mathbf{x})$ po pretpostavci regularna matrica za svako $\mathbf{x}$. Iterativna metoda (9) je u ovom slučaju definisana rekurentnom formulom

$$\mathbf{x}_{n+1} = \mathbf{x}_n - D(\mathbf{x}_n)\mathbf{f}(\mathbf{x}_n). \tag{19}$$

Različitim izborom matrice $D(\mathbf{x})$ dobijaju se različiti iterativni algoritmi.

Ako se uzme da je $D(\mathbf{x})$ matrica inverzna Jakobijanu preslikavanja $\mathbf{f}$,

$$D(\mathbf{x}) = \left(\mathbf{f}'(\mathbf{x})\right)^{-1} = \left(\frac{\partial f_i(\mathbf{x})}{\partial x_j}\right)^{-1},$$

iterativna metoda (19) se naziva Newton-Raphsonova ili, često, samo Newtonova metoda. Na ovu metodu se ne mogu primeniti rezultati iz prethodnog odeljka, jer ona za specijalan slučaj sistema (1) – sistem linearnih jednačina $A\mathbf{x} = \mathbf{b}$, ima oblik

$$\mathbf{x}_{n+1} = \mathbf{x}_n - A^{-1}(A\mathbf{x}_n - \mathbf{b}) = A^{-1}\mathbf{b},$$

što ne definiše iterarivni algoritam.

Stoga ćemo posebno analizirati konvergenciju ove metode, pri čemu ćemo je, kao i ranije, primeniti na opštiji oblik operatorske jednačine

$$F(x) = 0. \tag{20}$$

Neka je F operator koji preslikava linearni normirani prostor $\mathcal{X}$ u linearni normirani prostor $\mathcal{Y}$, pri čemu može biti $\mathcal{Y} \equiv \mathcal{X}$. Linearni operator $P : \mathcal{X} \to \mathcal{Y}$ naziva se izvod Frecheta operatora F u tački x ako je

$$\|F(x+h) - F(x) - Ph\| = o(\|h\|), \qquad \text{kada } \|h\| \to 0, \tag{21}$$

gde su sa $\|\cdot\|$ označene norme u odgovarajućim prostorima. Operator P se obično označava sa $F'(x)$.

PRIMER 3. Neka je $x = (x_1, \ldots, x_m)^T$ i $F = (f_1, \ldots, f_m)^T$. Ako su funkcije $f_i(x)$, $i = 1, \ldots, m$, neprekidno diferencijabilne u okolini tačke x, može se napisati

$$F(x+h) = F(x) + F'(x)h + o(\|h\|),$$

ili, u razvijenom obliku,

$$\begin{pmatrix} f_1(x_1+h_1, \ldots, x_m+h_m) \\ \vdots \\ f_m(x_1+h_1, \ldots, x_m+h_m) \end{pmatrix} = \begin{pmatrix} f_1(x_1, \ldots, x_m) \\ \vdots \\ f_m(x_1, \ldots, x_m) \end{pmatrix}$$
$$+ \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_m} \end{pmatrix} \cdot \begin{pmatrix} h_1 \\ \vdots \\ h_m \end{pmatrix} + o\left(\sqrt{\sum_{i=1}^{m} h_i^2}\right).$$

U ovom slučaju, izvod Frecheta je Jakobijan preslikavanja F.

Neka je x^* tačno, a x_n neko približno rešenje jednačine (20). Na osnovu definicije (21) je

$$\|F(x^*) - F(x_n) - F'(x_n)(x^* - x_n)\| = o(\|x^* - x_n\|),$$

što, pod pretpostavkom da je $\|x^* - x_n\|$ mala veličina, daje približnu jednakost

$$F(x_n) + F'(x_n)(x^* - x_n) \approx F(x^*) = 0.$$

Rešenje x_{n+1} jednačine

$$F(x_n) + F'(x_n)(x_{n+1} - x_n) = 0, \tag{22}$$

se, ako postoji, može uzeti za sledeću aproksimaciju tačke x^*. To rešenje se, ako postoji inverzan operator operatoru F', može zapisati u obliku

$$x_{n+1} = x_n - [F'(x_n)]^{-1} F(x_n).$$

Ovom formulom, za $n = 0, 1, \ldots$, je definisana Newton-Raphsonova metoda za jednačinu (20).

LEMA 1. *Ako postoji izvod $F'(x)$ neprekidnog operatora $F : \mathcal{X} \to \mathcal{Y}$ za svako $x \in C$, gde je C konveksan skup u $\mathcal{X}$, i ako postoji konstanta γ takva da je*

$$\|F'(x) - F'(y)\| \le \gamma \|x - y\| \qquad \text{za svako } x, y \in C, \tag{23}$$

tada za svako $x, y \in C$ važi ocena

$$\|F(x) - F(y) - F'(y)(x - y)\| \le \frac{\gamma}{2} \|x - y\|^2.$$

DOKAZ: S obzirom da je po definiciji skup C konveksan ako je za svako $x, y \in C$ i $0 \le t \le 1$ odsečak $tx + (1 - t)y$ sadržan u C, funkcija

$$\phi(t) = F(y + t(x - y))$$

je definisana za svako $x, y \in C$ i $t \in [0, 1]$, i $\phi : [0, 1] \to Y$. Ona je i diferencijabilna za svako $t \in [0, 1]$,

$$\phi'(t) = F'(y + t(x - y))\,(x - y).$$

Stoga je, zbog (23),

$$\begin{aligned} \|\phi'(t) - \phi'(0)\| &= \|\big(F'(y + t(x - y)) - F'(y)\big)\,(x - y)\| \\ &\le \|F'(y + t(x - y)) - F'(y)\|\,\|x - y\| \le \gamma t \|x - y\|^2. \end{aligned} \tag{24}$$

Kako je

$$F(x) - F(y) - F'(y)(x - y) = \phi(1) - \phi(0) - \phi'(0) = \int_0^1 \big(\phi'(t) - \phi'(0)\big)\,dt,$$

to, koristeći ocenu (24), dobijamo traženu ocenu

$$\begin{aligned} \|F(x) - F(y) - F'(y)(x - y)\| &\le \int_0^1 \|\phi'(t) - \phi'(0)\|\,dt \\ &\le \gamma \|x - y\|^2 \int_0^1 t\,dt = \frac{\gamma}{2} \|x - y\|^2. \end{aligned}$$

∎

Sada možemo dokazati teoremu kojom se utvrđuju uslovi pod kojima Newtonov algoritam konvergira.

TEOREMA 2. *Neka je C konveksna oblast u $\mathcal{X}$ i $x_0 \in C$. $F : \mathcal{X} \to \mathcal{Y}$ je neprekidan operator, takav da postoji $F'(x)$ za svako $x \in C$, koji ima sledeće osobine:*

$$(a) \qquad \|F'(x) - F'(y)\| \le \gamma\|x - y\| \qquad \text{za svako } x, y \in C,$$

$$(b) \qquad [F'(x)]^{-1} \text{ postoji i } \|[F'(x)]^{-1}\| \le \beta \text{ za svako } x \in C,$$

$$(c) \qquad \|[F'(x_0)]^{-1}F(x_0)\| \le \alpha,$$

gde su α, β i γ konstante takve da je

$$h = \frac{\alpha\beta\gamma}{2} < 1. \tag{25}$$

Dalje, neka je

$$S(x_0, r) = \{x \mid \|x - x_0\| < r\} \subseteq C,$$

pri čemu je

$$r = \frac{\alpha}{1 - h}. \tag{26}$$

Tada

(*i*) svi članovi niza određenog formulom

$$x_{n+1} = x_n - [F'(x_n)]^{-1}F(x_n), \qquad n = 0, 1, \ldots \tag{27}$$

sa početnom tačkom x_0, pripadaju lopti $S(x_0, r)$, tj. $x_n \in S(x_0, r)$ za svako $n \ge 0$;

(*ii*) postoji tačka $x^* \in \overline{S(x_0, r)}$ takva da je

$$\lim_{n\to\infty} x_n = x^*, \qquad F(x^*) = 0;$$

(*iii*) za svako $n \ge 0$ je

$$\|x_n - x^*\| \le \alpha\frac{h^{2^n-1}}{1 - h^{2^n}}. \tag{28}$$

DOKAZ: (*i*) Tačka $x_1 \in S(x_0, r)$, jer je na osnovu (27) za $n = 0$, pretpostavke (c) i (26)

$$\|x_1 - x_0\| = \|[F'(x_0)]^{-1}F(x_0)\| \le \alpha < r, \tag{29}$$

pošto je $0 < h < 1$.

Za proizvoljno $n > 0$, zbog pretpostavke (b) sledi da je

$$\begin{aligned}\|x_{n+1} - x_n\| &= \| - [F'(x_n)]^{-1}F(x_n)\| \le \beta\|F(x_n)\| \\ &= \beta\|F(x_n) - F(x_{n-1}) - F'(x_{n-1})(x_n - x_{n-1})\|,\end{aligned}$$

jer je po definiciji niza $\{x_n\}$

$$F'(x_{n-1})(x_n - x_{n-1}) + F(x_{n-1}) = 0.$$

Stoga je, na osnovu pretpostavke (a), leme 1 i (25),

$$\|x_{n+1} - x_n\| \le \beta\frac{\gamma}{2}\|x_n - x_{n-1}\|^2 = \frac{h}{\alpha}\|x_n - x_{n-1}\|^2, \tag{30}$$

što množenjem sa $\frac{h}{\alpha}$ daje

$$\frac{h}{\alpha}\|x_{n+1} - x_n\| \le \left(\frac{h}{\alpha}\|x_n - x_{n-1}\|\right)^2.$$

Primenjujući uzastopno ovu ocenu, imamo da je

$$\begin{aligned}\frac{h}{\alpha}\|x_{n+1} - x_n\| &\le \left(\frac{h}{\alpha}\|x_n - x_{n-1}\|\right)^2 \le \left(\frac{h}{\alpha}\|x_{n-1} - x_{n-2}\|\right)^{2^2} \le \cdots \\ &\le \left(\frac{h}{\alpha}\|x_1 - x_0\|\right)^{2^n},\end{aligned}$$

što, uzimajući u obzir (29), daje

$$\frac{h}{\alpha}\|x_{n+1} - x_n\| \le \left(\frac{h}{\alpha}\|x_1 - x_0\|\right)^{2^n} \le h^{2^n},$$

odnosno

$$\|x_{n+1} - x_n\| \le \alpha h^{2^n-1}, \qquad n = 0, 1, \ldots . \tag{31}$$

Konačno je, na osnovu (26),

$$\begin{aligned}\|x_{n+1} - x_0\| &\le \|x_{n+1} - x_n\| + \cdots + \|x_1 - x_0\| \le \alpha(1 + h + h^3 + \cdots + h^{2^n-1}) \\ &< \alpha\sum_{j=0}^{\infty} h^j = \frac{\alpha}{1-h} = r,\end{aligned}$$

što znači da $x_n \in S(x_0, r)$ za proizvoljno n.

(*ii*) Iz (31) je za proizvoljno $k > n$

$$\begin{aligned}\|x_{k+1} - x_n\| &\le \|x_{k+1} - x_k\| + \cdots + \|x_{n+1} - x_n\| \\ &\le \alpha h^{2^k-1} + \cdots + \alpha h^{2^n-1} \\ &\le \alpha h^{2^n-1}\left(1 + h^{2^n} + (h^{2^n})^2 + \cdots\right) = \alpha\frac{h^{2^n-1}}{1 - h^{2^n}},\end{aligned} \tag{32}$$

što može da se učini proizvoljno malim za dovoljno veliko n, jer je $0 < h < 1$. Stoga je niz $\{x_n\}$ Cauchyev niz, te konvergira ka nekoj tački x^* iz $\overline{S(x_0, r)}$ (tačka pripada lopti jer svi članovi niza, kao što je dokazano, pripadaju unutrašnjosti te lopte),

$$\lim_{n\to\infty} x_n = x^*, \qquad x^* \in \overline{S(x_0, r)}.$$

Dokažimo još da je $F(x^*) = 0$. Iz pretpostavke (a) sledi da je za proizvoljno n

$$\|F'(x_n) - F'(x_0)\| \leq \gamma \|x_n - x_0\| \leq \gamma r,$$

pa je

$$\|F'(x_n)\| \leq \gamma r + \|F'(x_0)\| = C, \qquad C = \text{const.}$$

Stoga se iz (27) dobija ocena

$$\|F(x_n)\| = \| - F'(x_n)(x_{n+1} - x_n)\| \leq C\|x_{n+1} - x_n\|,$$

na osnovu koje je, s obzirom na dokazanu konvergenciju niza $\{x_n\}$,

$$\lim_{n\to\infty} \|F(x_n)\| = 0.$$

Kako je F neprekidni operator, to je

$$\lim_{n\to\infty} \|F(x_n)\| = \|F(\lim_{n\to\infty} x_n)\| = \|F(x^*)\| = 0,$$

odnosno

$$F(x^*) = 0.$$

(*iii*) Ocena (28) sledi neposredno iz (32), kada $k \to \infty$. ∎

NAPOMENA 1. Prema definiciji (5), iz (30) je očigledno da je Newtonova metoda reda dva.

Teoremom 2 se ne garantuje jedinost rešenja jednačine (20), za šta je potrebno da operator F zadovoljava nešto strožije uslove. Potpunu formulaciju uslova pod kojima Newtonova metoda konvergira ka jedinstvenom rešenju jednačine (20) daje teorema Newton-Kantoroviča.

TEOREMA 3 (NEWTON–KANTOROVIČ). *Neka je $F : \mathcal{X} \to \mathcal{Y}$ neprekidan operator za koji na konveksnom skupu $C \subset \mathcal{X}$ postoji neprekidan linearni operator F', i neka su zadovoljeni sledeći uslovi:*

$$(a) \qquad \|F'(x) - F'(y)\| \leq \gamma \|x - y\| \qquad \text{za svako } x, y \in C,$$

$$(b) \qquad \|[F'(x_0)]^{-1}\| \leq \beta,$$

$$(c) \qquad \|[F'(x_0)]^{-1} F(x_0)\| \leq \alpha,$$

za neko $x_0 \in C$ i konstante α, β i γ. Neka je još

$$h = \alpha\beta\gamma, \qquad r_{1/2} = \frac{1 \mp \sqrt{1 - 2h}}{h}\alpha.$$

Ako je $h \leq \frac{1}{2}$ i $\overline{S(x_0, r_1)} \subset C$, niz $\{x_n\}$, određen formulom

$$x_{n+1} = x_n - [F'(x_n)]^{-1} F(x_n), \qquad n = 0, 1, \ldots$$

pripada lopti $S(x_0, r_1)$ *i konvergira ka tački* x^* *koja je jedino rešenje jednačine* $F(x) = 0$ *u oblasti* $C \cap S(x_0, r_2)$. *Pri tome važi ocena greške*

$$\|x_n - x^*\| \leq \frac{(2h)^{2^n}}{\beta\gamma 2^n}, \qquad n = 0, 1, \ldots .$$

Dokaz ove teoreme se može naći u [15] ili [22].

NAPOMENA 2. Konstanta γ u uslovu (a) teoreme 3 (i teoreme 2) se može, u slučaju da je $F(x)$ dva puta neprekidno diferencijabilna funkcija jedne realne promenljive, oceniti sa $\max|F''(x)|$, jer je

$$F'(x) - F'(y) = F''(\xi)(x - y).$$

Izračunavanje $[F'(x_n)]^{-1}$ na svakom koraku je obiman posao, te se često x_{n+1} umesto rekurentnom formulom (27) računa rešavanjem jednačine (22). U slučaju da operatorska jednačina (20) predstavlja sistem (1), jednačina (22) je ustvari sistem linearnih jednačina

$$\mathbf{f}'(\mathbf{x}_n)(\mathbf{x}_{n+1} - \mathbf{x}_n) = -\mathbf{f}(\mathbf{x}_n)$$

po popravkama $x_i^{(n+1)} - x_i^{(n)}$, $i = 1, \ldots, m$.

PRIMER 4. Odredimo sa tačnošću $\epsilon = 0.5 \cdot 10^{-5}$ presečne tačke Descartesovog lista $f_1(x_1, x_2) \equiv x_1^3 + x_2^3 - 3x_1x_2 = 0$ i kruga $f_2(x_1, x_2) = x_1^2 + x_2^2 - 3x_1 - 3x_2 + 3.5 = 0$.

Analizom ove dve krive dolazimo do zaključka da postoje dve presečne tačke, koje su simetrične u odnosu na pravu $x_1 = x_2$, (x_1^*, x_2^*) i (x_2^*, x_1^*), pri čemu $(x_1^*, x_2^*) \in S$, $S = (0.532, 0.546) \times (1.2, 1.25)$. Odredimo prvu od njih Newtonovom metodom, uzimajući za početnu aproksimaciju $x_1^{(0)} = 0.538556$, $x_2^{(0)} = 1.225$ ($x_1^{(0)}$ je određeno kao rešenje kvadratne jednačine $f_2(x_1, x_2^{(0)}) = 0$). Koristeći uniformne norme (5.7) i (5.11), imamo ocene

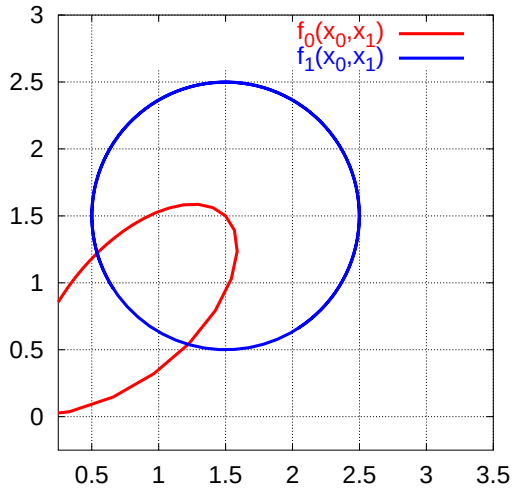
$$\|F(\mathbf{x}_0)\| \leq 0.016, \qquad \|[F'(\mathbf{x}_0)]^{-1}\| \leq 0.667,$$

i

$$\begin{aligned}\|F'(\mathbf{x}) - F'(\mathbf{y})\| &= \max_{\mathbf{x},\mathbf{y}\in S}\{|3(x_1^2 - x_2) - 3(y_1^2 - y_2)| + |3(x_2^2 - x_1) - 3(y_2^2 - y_1)|, \\ &\qquad |2x_1 - 3 - (2y_1 - 3)| + |2x_2 - 3 - (2y_2 - 3)|\} \\ &\leq \max_{\mathbf{x},\mathbf{y}\in S}\{3(|x_1 + y_1| + |x_2 + y_2| + 2),\, 4\}\, \|\mathbf{x} - \mathbf{y}\| \\ &\leq 16.78\, \|\mathbf{x} - \mathbf{y}\|.\end{aligned}$$

Na osnovu teoreme 3, za $\alpha = 0.011$, $\beta = 0.667$, $\gamma = 16.78$, tj. $h = 0.062$, niz aproksimacija određenih Newtonovom metodom konvergira ka traženom rešenju. Rešavajući sisteme linearnih jednačina (22) po priraštajima $x_1^{(n+1)} - x_1^{(n)}$ i $x_2^{(n+1)} - x_2^{(n)}$, $n = 0, 1, \ldots$, dobijamo aproksimacije rešenja

$$x_1^{(1)} = 0.539740, \quad x_2^{(1)} = 1.220858, \quad \text{i} \quad x_1^{(2)} = 0.539754, \quad x_2^{(2)} = 1.220844.$$



Slika 7.2: Primer 4.

Kako je $|x_1^{(3)} - x_1^{(2)}| < 0.5 \cdot 10^{-6}$ i $|x_2^{(3)} - x_2^{(2)}| < 0.5 \cdot 10^{-6}$, to je sa traženom tačnošću

$$x_1^* = 0.53975, \qquad x_2^* = 1.22084.$$

Radi smanjenja obima računanja koristi se i modifikacija ove metode, koja se sastoji u tome da se u više koraka koristi isto $[F'(x_n)]^{-1}$. Unapred se zadaje rastući niz brojeva $n_0 = 0, n_1, n_2, \ldots$, i za $n_k \leq n < n_{k+1}$ iteracije se računaju po formuli

$$x_{n+1} = x_n - [F'(x_{n_k})]^{-1} F(x_n).$$

Konvergencija je nešto sporija, ali je izračunavanje jednostavnije.

Kako u osnovnoj tako i u modifikovanoj Newtonovoj metodi brzina konvergencije umnogome zavisi od dobrog izbora početne aproksimacije rešenja x_0.

7.3 Metode za rešavanje jednačina u $\mathcal{R}^1$

Metoda Newtona. U jednodimenzionom slučaju, Newtonova metoda je znatno jednostavnija, a njena geometrijska interpretacija očiglednija. Neka je $\mathcal{X} \equiv \mathcal{Y} \equiv \mathcal{R}^1$, tj. operator F je realna funkcija jedne promenljive $f(x)$. Iterativni algoritam (27) kojim je definisana Newtonova metoda je

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n = 0, 1, \ldots, \tag{33}$$

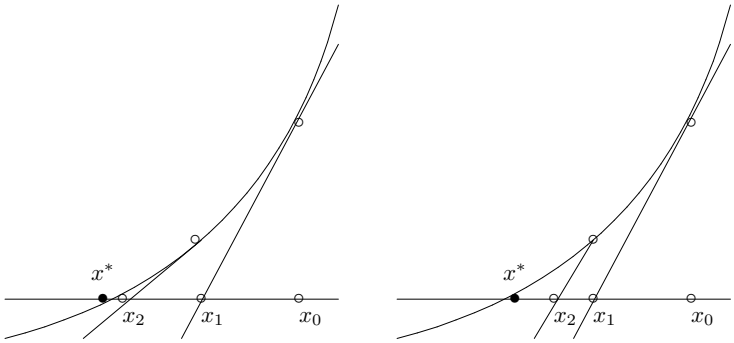
uz uslov da je $f'(x_n) \neq 0$ za svako n. Pošto je jednačina tangente na krivu $f(x)$ u tački x_n

$$y = t_n(x) \equiv f'(x_n)(x - x_n) + f(x_n),$$

očigledno je da je tačka x_{n+1}, određena formulom (33), rešenje jednačine $t_n(x) = 0$. Drugim rečima, na svakom koraku funkcija $f(x)$ se aproksimira svojom tangentom u tački $(x_n, f(x_n))$, i nova aproksimacija rešenja x_{n+1} se određuje kao presečna tačka ove prave sa osom Ox (slika 7.3). Zato se ova metoda u jednodimenzionom slučaju često naziva i metoda tangente. U osnovnoj modifikaciji Newtonove metode

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}, \qquad n = 0, 1, \ldots,$$

u n-toj iteraciji $f(x)$ se aproksimira pravom koja prolazi kroz tačku $(x_n, f(x_n))$, a paralelna je tangenti na krivu u tački $(x_0, f(x_0))$



Slika 7.3: Geometrijska interpretacija metode Newtona i njene modifikacije.

Uslovi konvergencije Newtonove metode i u prostoru $\mathcal{R}^1$ su dati teoremom 2, tj. teoremom 3, ali se mogu i jednostavnije iskazati sledećom teoremom

TEOREMA 4. *Ako funkcija $f : [a, b] \to \mathcal{R}^1$ ima sledeće osobine*

(a) neprekidno je diferencijabilna, $f \in \mathcal{C}^1[a, b]$,

(b) ima različiti znak na krajevima intervala $[a, b]$, $f(a)f(b) < 0$,

(c) za svako $x \in [a, b]$ postoji $f''(x)$,

(d) na intervalu $[a, b]$ $f'(x)$ i $f''(x)$ ne menjaju znak, i $f'(x) \neq 0$ za svako $x \in [a, b]$,

(e) tačka $x_0 \in [a, b]$ je takva da je $f(x_0)f''(x_0) > 0$,

onda niz $\{x_n\}$, sa prvim članom x_0 i određen formulom (33), konvergira ka jedinstvenom rešenju $x^* \in [a, b]$ jednačine $f(x) = 0$.

DOKAZ: Jednačina $f(x) = 0$ ima rešenje na intervalu $[a, b]$ na osnovu pretpostavki (a) i (b) teoreme, a njegova jedinstvenost sledi iz pretpostavke (d) o monotonosti funkcije $f(x)$. Označimo to rešenje sa x^*. Dalje, radi određenosti, pretpostavimo da je $f(a) < 0$, $f(b) > 0$, i $f'(x) > 0$, $f''(x) > 0$ za svako $x \in [a, b]$ (ostali slučajevi se razmatraju analogno). Uzmimo da je $x_0 = b$, što, s obzirom na pretpostavljeno, zadovoljava uslov (e) teoreme.

Indukcijom ćemo dokazati da je $x_n > x^*$ za svako n. Za x_0 smo izabrali desni kraj intervala $[a, b]$, pa je očigledno da je $x_0 > x^*$. Pretpostavimo da je $x_k > x^*$, $k = 1, \ldots, n$. Iz Taylorovog razvoja za funkciju $f(x)$,

$$0 = f(x^*) = f(x_n + (x^* - x_n)) = f(x_n) + f'(x_n)(x^* - x_n) + \frac{1}{2} f''(\xi)(x^* - x_n)^2,$$

gde je $\xi \in (x^*, x_n)$, sledi da je

$$f(x_n) + f'(x_n)(x^* - x_n) < 0$$

jer je po pretpostavci $f''(x) > 0$ za svako x. Kako je po pretpostavci i $f'(x) > 0$ za svako x, to je

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} > x^*,$$

što je i trebalo pokazati.

Funkcija $f(x)$ je monotona, i sve tačke x_n su sa iste strane njene nule x^* kao i tačka $x_0 = b$ u kojoj je $f(x_0) > 0$, pa je $f(x_n) > 0$, $n = 0, 1, \ldots$. Stoga, iz (33) neposredno sledi da je $x_{n+1} < x_n$ za svako n, odnosno da je niz $\{x_n\}$ monotono opadajući. Dakle, niz $\{x_n\}$ je monotono opadajući i ograničen sa donje strane tačkom x^*, te stoga konvergira

$$\lim_{n\to\infty} x_n = \bar{x}, \qquad \bar{x} \in [a, b].$$

Puštajući u formuli (33) da $n \to \infty$, s obzirom na pretpostavku (a) teoreme, imamo da je

$$\lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} x_n - \frac{f(\lim_{n\to\infty} x_n)}{f'(\lim_{n\to\infty} x_n)},$$

tj.

$$\bar{x} = \bar{x} - \frac{f(\bar{x})}{f'(\bar{x})},$$

odakle sledi da je

$$f(\bar{x}) = 0.$$

Kako smo na početku dokaza zaključili da jednačina ima samo jedno rešenje, mora biti $\bar{x} \equiv x^*$, tj. granica niza $\{x_n\}$ je jedino rešenje jednačine $f(x) = 0$ u intervalu $[a, b]$,

$$\lim_{n\to\infty} x_n = x^*, \qquad f(x^*) = 0.$$

∎

Za ocenu greške aproksimacije x_n može se koristiti ocena

$$|x^* - x_n| \le \frac{|f(x_n)|}{m_1}, \qquad m_1 = \min_{x\in[a,b]} |f'(x)|, \tag{34}$$

koja je neposredna posledica teoreme o srednjoj vrednosti. Ako je $|f'(x)| \ge m_1 > 0$ kada $x \in [a, b]$, iz

$$f(x^*) - f(x_n) = f'(\xi)(x^* - x_n),$$

gde je ξ neka tačka iz intervala određenog tačkama x_n i x^*, sledi

$$|f(x_n)| \ge m_1|x^* - x_n|,$$

jer je $f(x^*) = 0$.

Ocena (34) važi nezavisno od načina na koji je aproksimacija x_n određena. Ako je x_n određeno Newtonovom metodom, tj. formulom (33), koristeći ocenu (34) može da se izvede i druga ocena. Iz Taylorovog razvoja

$$\begin{aligned} f(x_n) &= f(x_{n-1} + (x_n - x_{n-1})) \\ &= f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) + \frac{1}{2}f''(\xi)(x_n - x_{n-1})^2, \end{aligned}$$

s obzirom da je prema (33)

$$f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) = 0,$$

sledi da je

$$|f(x_n)| = \frac{1}{2}|f''(\xi)|\,|x_n - x_{n-1}|^2 \le \frac{1}{2}M_2|x_n - x_{n-1}|^2,$$

gde je $M_2 = \max_{x\in[a,b]} |f''(x)|$. Zamenom u (34), dobijamo traženu ocenu

$$\begin{gathered} |x^* - x_n| \le \frac{M_2}{2m_1}|x_n - x_{n-1}|^2, \\ m_1 = \min_{x\in[a,b]} |f'(x)|, \qquad M_2 = \max_{x\in[a,b]} |f''(x)|, \end{gathered}$$

kojom je takođe potvrđena kvadratna brzina konvergencije Newtonove metode.

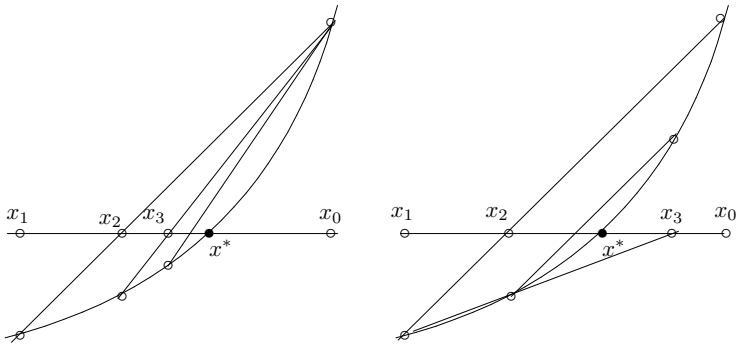
Metoda regula-falsi i metoda sečice. Kao što smo videli, Newtonova metoda ustvari predstavlja linearizaciju jednačine (2), a približno rešenje x_n je rešenje dobijene linearne jednačine. U jednodimenzionom slučaju se ona svodi na aproksimiranje krive njenom tangentom u tački $(x_n, f(x_n))$.

Linearna aproksimacija se može definisati i na drugi način – na primer, kriva se može aproksimirati sečicom zadatom dvema tačkama krive. Ako je jedna tačka koja određuje sečicu fiksirana tačka x_F, a druga tačka poslednja nađena aproksimacija x_n, rekurentna formula kojom se određuje niz približnih rešenja je

$$x_{n+1} = x_n - \frac{f(x_n)}{f(x_F) - f(x_n)}(x_F - x_n), \qquad n = 0, 1, \ldots, \tag{35}$$

i ovako definisana metoda se naziva *metoda regula-falsi* ili metoda lažnog položaja (slika 7.4). Ukoliko je sečica određena tačkama $(x_n, f(x_n))$ i $(x_{n-1}, f(x_{n-1}))$, metoda se naziva *metoda sečice* i definisana je formulom

$$x_{n+1} = x_n - \frac{f(x_n)}{f(x_{n-1}) - f(x_n)}(x_{n-1} - x_n), \qquad n = 0, 1, \ldots .$$



Slika 7.4: Geometrijska interpretacija metode regula-falsi i metode sečice.

Ocenimo tačnost približnog rešenja dobijenog metodom regula-falsi. Iz (35), obzirom da je $f(x^*) = 0$, je

$$x_{n+1} - x^* = x_n - x^* - \frac{f(x_n) - f(x^*)}{f(x_F) - f(x_n)}(x_F - x_n),$$

pa je na osnovu teoreme o srednjoj vrednosti

$$x_{n+1} - x^* = x_n - x^* - \frac{(x_n - x^*)f'(\xi_1)}{(x_F - x_n)f'(\xi_2)}(x_F - x_n) = (x_n - x^*)\left(1 - \frac{f'(\xi_1)}{f'(\xi_2)}\right).$$

Dodajući i oduzimajući x_{n+1} od $x_n - x^*$, i grupišući odgovarajuće sabirke, dobijamo da je

$$(x_{n+1} - x^*)\frac{f'(\xi_1)}{f'(\xi_2)} = (x_n - x_{n+1})\frac{f'(\xi_2) - f'(\xi_1)}{f'(\xi_2)},$$

ili

$$x_{n+1} - x^* = \frac{f'(\xi_1) - f'(\xi_2)}{f'(\xi_1)}(x_{n+1} - x_n).$$

Ako funkcija $f \in \mathcal{C}^1[a, b]$ i monotona je na tom intervalu, postoje konstante m_1 i M_1 takve da je

$$0 < m_1 \leq |f'(x)| \leq M_1 < \infty, \qquad \text{za svako } \; x \in [a, b],$$

te je tražena ocena

$$|x_{n+1} - x^*| \leq \frac{M_1 - m_1}{m_1}|x_{n+1} - x_n|.$$

Ista ocena važi i za metodu sečice. Za ocenu tačnosti može se koristiti i ocena (34).

Metoda sečice se može kombinovati sa Newtonovom metodom, tako da se za definisanje sečice umesto tačke $(x_{n-1}, f(x_{n-1}))$, koristi aproksimacija određena Newtonovom metodom

$$\bar{x}_{n+1} = \bar{x}_n - \frac{f(\bar{x}_n)}{f'(\bar{x}_n)}, \qquad x_{n+1} = x_n - \frac{f(x_n)}{f(\bar{x}_n) - f(x_n)}(\bar{x}_n - x_n), \qquad n = 0, 1, \ldots .$$

PRIMER 5. Kombinovanjem metode sečice i Newtonove metode odredimo sa tačnošću $\epsilon = 10^{-4}$ najmanje pozitivno rešenje jednačine

$$\tan x = x.$$

Ovo rešenje se nalazi u intervalu $(\pi, \frac{3\pi}{2})$, i odredićemo ga kao nulu funkcije $f(x) = \sin x - x\cos x$. Kako je $f(\frac{3\pi}{2})f''(\frac{3\pi}{2}) > 0$, za početnu vrednost Newtonovog algoritma se, na osnovu teoreme 4, može uzeti $\bar{x}_0 = \frac{3\pi}{2}$, a drugi kraj polazne sečice je $x_0 = \pi$. Direktnom primenom prethodnih formula dobijamo dva niza tačaka

$$\begin{aligned} &\bar{x}_1 = 4.50018, \qquad \bar{x}_2 = 4.49342, \qquad \bar{x}_3 = 4.49341, \\ &x_1 = 4.33312, \qquad x_2 = 4.49313, \qquad x_3 = 4.49341. \end{aligned}$$

Pošto je $|\bar{x}_3 - x_3| < \epsilon$, traženi koren je $x^* = 4.4934$.

Uopštenje metode sečice na višedimenzione probleme je dosta složeno, i može se naći u [22].

Metoda polovljenja intervala. Metodu regula-falsi, ako je $f(x_F) \cdot f(x_n) < 0$, možemo interpretirati i kao metodu u kojoj se na svakom koraku interval određen tačkama x_F i x_n deli u razmeri $f(x_F) : f(x_n)$, pri čemu se u sledećem koraku algoritam realizuje na onom od dobijenih podintervala na čijim krajevima funkcija $f(x)$ ima različit znak. Jednostavnija varijanta ove ideje je da se tekući interval podeli na dva jednaka podintervala, što je upravo u osnovi algoritma metode polovljenja intervala, koja se naziva i metoda bisekcije.

Pretpostavimo da je funkcija $f \in \mathcal{C}[a, b]$ i da je $f(a)f(b) < 0$, kako bi u intervalu $[a, b]$ postojalo bar jedno rešenje x^* jednačine $f(x) = 0$. Prvo odredimo sredinu $\frac{a_0+b_0}{2}$ intervala $[a_0, b_0] \equiv [a, b]$. Ukoliko je u toj tački funkcija f jednaka nuli, nađeno je tačno rešenje $x^* = \frac{a_0+b_0}{2}$. Ako to nije slučaj, onda označimo sa $[a_1, b_1]$ onaj od dobijenih podintervala na čijim krajevima funkcija ima različit znak, i ponovimo postupak opisan za interval $[a_0, b_0]$. Ponavljajući ovaj postupak dobićemo niz intervala $[a_n, b_n]$ od kojih je svaki sadržan u svim prethodnim,

$$[a_0, b_0] \supset [a_1, b_1] \supset \cdots \supset [a_n, b_n] \supset \cdots , \tag{36}$$

i takvih da je

$$f(a_n)f(b_n) < 0, \qquad b_n - a_n = \frac{1}{2^n}(b-a), \qquad x^* \in [a_n, b_n], \qquad n = 0, 1, \ldots . \tag{37}$$

Ukoliko se ne dogodi da je neka od deonih tačaka upravo tačka x^*, tj. da je $f(\frac{a_k+b_k}{2}) = 0$ za neko k, niz intervala je beskonačan i nizovi levih i desnih krajeva intervala konvergiraju uravo ka rešenju x^*. Dokažimo to.

TEOREMA 5. *Neka je funkcija $f \in \mathcal{C}[a,b]$ i $f(a)f(b) < 0$. Nizovi $\{a_n\}$ i $\{b_n\}$ levih i desnih krajeva intervala $[a_n, b_n]$, određenih metodom polovljenja intervala, konvergiraju ka tački x^*, koja je rešenje jednačine $f(x) = 0$.*

DOKAZ: S obzirom na (36), levi krajevi intervala obrazuju monotono neopadajući niz ograničen odozgo,

$$a_0 \leq a_1 \leq \cdots \leq a_n \leq \cdots < b_0,$$

a desni krajevi obrazuju monotono nerastući niz ograničen odozdo,

$$b_0 \geq b_1 \geq \cdots \geq b_n \geq \cdots > a_0.$$

Stoga ovi nizovi konvergiraju,

$$\lim_{n\to\infty} a_n = A, \qquad \lim_{n\to\infty} b_n = B,$$

pri čemu je $a_n \leq A \leq B \leq b_n$, $n = 0, 1, \ldots$. Kako je, prema (37),

$$B - A = \lim_{n\to\infty} (b_n - a_n) = \lim_{n\to\infty} \frac{1}{2^n}(b-a) = 0,$$

to je $A = B = x^*$ jedinstvena tačka koja pripada svim intervalima $[a_n, b_n]$, $n = 0, 1, \ldots$. Ova tačka je rešenje jednačine $f(x) = 0$, jer je zbog neprekidnosti funkcije $f(x)$ i (37)

$$\lim_{n\to\infty} f(a_n)f(b_n) = f(\lim_{n\to\infty} a_n)f(\lim_{n\to\infty} b_n) = f(A)f(B) = \big(f(x^*)\big)^2 \leq 0.$$

∎

U praksi, za približno rešenje jednačine x_n sa greškom ne većom od ϵ može se uzeti ma koja tačka iz intervala $[a_n, b_n]$, ukoliko je $b_n - a_n \leq \epsilon$. Obično se uzima da je

$$x_n = \frac{1}{2}(a_n + b_n),$$

pa je greška aproksimacije

$$|x^* - x_n| \leq \frac{1}{2}(b_n - a_n) = \frac{1}{2^{n+1}}(b-a).$$

Iz poslednje ocene je broj iteracija koje je potrebno izračunati da bi se postigla zadata tačnost ϵ

$$n \geq \left[\frac{\ln \frac{b-a}{\epsilon}}{\ln 2} \right].$$

Osnovni nedostaci ove metode su nemogućnost primene na višedimenzione probleme i spora konvergencija.

Njene prednosti su jednostavnost algoritma i direktna ocena greške. Osim toga, ne postavljaju se zahtevi tipa diferencijabilnosti, monotonosti, ..., funkcije f, što omogućava njenu široku primenu. Ako na intervalu $[a, b]$ nije lokalizovano jedno rešenje jednačine $f(x) = 0$, metoda će dati aproksimaciju jednog od rešenja. Takođe se u realizaciji ove metode ne javljaju računski problemi – na primer, deljenje brojevima bliskim ili čak jednakim nuli, što može da se dogodi pri primeni Newtonove metode ili metode regula-falsi. Stoga se ona često kombinuje sa ovim i drugim metodama višeg reda, tako što se jedan korak algoritma uradi ovom metodom kada god nastanu pomenuti računski problemi.

7.4 Metoda Bairstowa za rešavanje algebarskih jednačina

Poseban značaj u praksi imaju jednačine definisane polinomima, tzv. algebarske jednačine. Za njihovo rešavanje se pored već pomenutih opštih metoda koriste i posebne metode – metode za nalaženje korena polinoma. Pri tome treba voditi računa da polinom može biti loše uslovljen, tj. da male promene njegovih koeficijenata dovode do velikih promena korena, što je uzrok nagomilavanju računske greške pri primeni nekog algoritma. Kad god je moguće, polazni problem ne treba svoditi na problem nalaženja korena polinoma, već koristiti posebne metode kojima se ovi direktno određuju (na primer, nalaženje sopstvenih vrednosti matrica).

Metoda Bairstowa, u literaturi nazvana i metoda Hitchcocka, je opšta u tom smislu da se njom mogu odrediti i višestruki i kompleksni koreni polinoma, jer se iterativnim algoritmom nalaze kvadratni faktori polinoma. Polinom drugog stepena $x^2 + px + q$ je faktor polinoma

$$P_m(x) = x^m + a_1 x^{m-1} + \cdots + a_{m-1}x + a_m,$$

ukoliko je ostatak pri deobi ovog polinoma kvadratnim faktorom nula, odnosno ako u reprezentaciji polinoma

$$P_m(x) = (x^2 + px + q)(x^{m-2} + b_1 x^{m-3} + \cdots + b_{m-3}x + b_{m-2}) + rx + s \tag{38}$$

važi da je

$$r(p, q) = 0, \qquad s(p, q) = 0. \tag{39}$$

$r(p, q)$ i $s(p, q)$ su nelinearne funkcije po p i q. Upoređivanjem koeficijenata uz jednake stepene x na levoj i desnoj strani relacije (38), dobijaju se veze koeficijenata polaznog polinoma i polinoma količnika i ostatka,

$$\begin{gathered} a_1 = b_1 + p, \qquad a_2 = b_2 + pb_1 + q, \qquad a_3 = b_3 + pb_2 + qb_1, \\ \ldots \qquad a_k = b_k + pb_{k-1} + qb_{k-2}, \qquad \ldots \qquad a_{m-2} = b_{m-2} + pb_{m-3} + qb_{m-4}, \\ a_{m-1} = r + pb_{m-2} + qb_{m-3}, \qquad a_m = s + qb_{m-2}, \end{gathered}$$

iz kojih sledi rekurentna formula

$$b_k = a_k - pb_{k-1} - qb_{k-2}, \quad k = 1, \ldots, m, \qquad b_{-1} = 0, \quad b_0 = 1, \tag{40}$$

i

$$r = a_{m-1} - pb_{m-2} - qb_{m-3} = b_{m-1}, \qquad s = a_m - qb_{m-2} = b_m + pb_{m-1}.$$

Stoga se sistem (39) svodi na sistem nelinearnih jednačina

$$\begin{aligned} b_{m-1}(p,q) &= 0 \\ b_m(p,q) + pb_{m-1}(p,q) &= 0. \end{aligned} \tag{41}$$

Primenom Newtonove metode (§2) dobijamo sistem linearnih jednačina po popravkama $\Delta p_n = p_{n+1} - p_n$ i $\Delta q_n = q_{n+1} - q_n$,

$$\begin{aligned} \frac{\partial r(p_n,q_n)}{\partial p}\Delta p_n + \frac{\partial r(p_n,q_n)}{\partial q}\Delta q_n + r(p_n,q_n) &= 0 \\ \frac{\partial s(p_n,q_n)}{\partial p}\Delta p_n + \frac{\partial s(p_n,q_n)}{\partial q}\Delta q_n + s(p_n,q_n) &= 0, \end{aligned}$$

koji, obzirom na (41), ima oblik

$$\begin{aligned} \frac{\partial b_{m-1}}{\partial p}\Delta p + \frac{\partial b_{m-1}}{\partial q}\Delta q + b_{m-1} &= 0 \\ \left(\frac{\partial b_m}{\partial p} + p\frac{\partial b_{m-1}}{\partial p} + b_{m-1}\right)\Delta p + \left(\frac{\partial b_m}{\partial q} + p\frac{\partial b_{m-1}}{\partial q}\right)\Delta q + b_m + pb_{m-1} &= 0 \end{aligned}$$

(indeks iteracije n je ovde izostavljen radi jednostavnijeg zapisa). Ovaj sistem može da se uprosti množenjem prve jednačine sa p i oduzimanjem od druge,

$$\begin{aligned} \frac{\partial b_{m-1}}{\partial p}\Delta p + \frac{\partial b_{m-1}}{\partial q}\Delta q + b_{m-1} &= 0 \\ \left(\frac{\partial b_m}{\partial p} + b_{m-1}\right)\Delta p + \frac{\partial b_m}{\partial q}\Delta q + b_m &= 0. \end{aligned} \tag{42}$$

Da bi našli koeficijente sistema, diferencirajmo po p i q koeficijente b_k date u (40),

$$\frac{\partial b_{-1}}{\partial p} = \frac{\partial b_0}{\partial p} = 0, \qquad -\frac{\partial b_k}{\partial p} = b_{k-1} + p\frac{\partial b_{k-1}}{\partial p} + q\frac{\partial b_{k-2}}{\partial p}, \quad k = 1, \ldots, m, \tag{43}$$

$$\frac{\partial b_{-1}}{\partial q} = \frac{\partial b_0}{\partial q} = 0, \qquad -\frac{\partial b_k}{\partial q} = b_{k-2} + p\frac{\partial b_{k-1}}{\partial q} + q\frac{\partial b_{k-2}}{\partial q}, \quad k = 1, \ldots, m, \tag{44}$$

Uvođenjem smene $c_{k-1} = -\frac{\partial b_k}{\partial p}$, $k = -1, \ldots, m$, u (43), imamo da je

$$c_{k-1} = b_{k-1} - pc_{k-2} - qc_{k-3}, \quad k = 1, \ldots, m, \qquad c_{-2} = 0, \quad c_{-1} = 0,$$

ili, pomeranjem indeksa k za 1 i obzirom da je $c_0 = b_0 = 1$,

$$c_k = b_k - pc_{k-1} - qc_{k-2}, \quad k = 1, \ldots, m-1, \qquad c_{-1} = 0, \quad c_0 = 1. \tag{45}$$

Uvođenjem smene $d_{k-2} = -\frac{\partial b_k}{\partial q}$, $k = -1, \ldots, m$, u (44), imamo da je

$$d_{k-2} = b_{k-2} - pd_{k-3} - qd_{k-4}, \quad k = 1, \ldots, m, \qquad d_{-3} = 0, \quad d_{-2} = 0,$$

ili, pomeranjem indeksa k za 2 i obzirom da je $d_{-1} = b_{-1} = 0$ i $d_0 = b_0 = 1$,

$$d_k = b_k - pd_{k-1} - qd_{k-2}, \quad k = 1, \ldots, m-2, \qquad d_{-1} = 0, \quad d_0 = 1. \tag{46}$$

Rekurentne relacije (45) i (46) su identične, te ih možemo objediniti u jednu formulu

$$c_k = b_k - pc_{k-1} - qc_{k-2}, \quad k = 1, \ldots, m-1, \qquad c_{-1} = 0, \quad c_0 = 1, \tag{47}$$

pri čemu je

$$\frac{\partial b_k}{\partial p} = -c_{k-1}, \qquad \frac{\partial b_k}{\partial q} = -c_{k-2}, \qquad k = 1, \ldots, m. \tag{48}$$

Poredeći relacije (40) i (47), vidimo da se koeficijenti c_k izračunavaju pomoću koeficijenata b_k po istoj formuli po kojoj se koeficijenti b_k izračunavaju pomoću koeficijenata a_k, te je

$$\begin{aligned} x^{m-2} + b_1 x^{m-3} + \cdots + b_{m-3}x + b_{m-2} \\ = (x^2 + px + q)(x^{m-4} + c_1 x^{m-5} + \cdots + c_{m-5}x + c_{m-4}) + \bar{r}x + \bar{s}, \end{aligned}$$

gde je

$$\bar{r} = c_{m-3}, \qquad \bar{s} = c_{m-2} + pc_{m-3}.$$

Vratimo se sistemu (42). Njegovi koeficijenti su, obzirom na (48) i (47),

$$\frac{\partial b_{m-1}}{\partial p} = -c_{m-2}, \qquad \frac{\partial b_{m-1}}{\partial q} = -c_{m-3}, \qquad \frac{\partial b_m}{\partial q} = -c_{m-2},$$

$$\frac{\partial b_m}{\partial p} + b_{m-1} = -c_{m-1} + b_{m-1} = pc_{m-2} + qc_{m-3} = -c'_{m-1},$$

i izračunavaju se rekurentnom relacijom (47). Sistem (42) je onda

$$\begin{aligned} c_{m-2}\Delta p + c_{m-3}\Delta q &= b_{m-1} \\ c'_{m-1}\Delta p + c_{m-2}\Delta q &= b_m, \end{aligned} \tag{49}$$

i, ako su odgovarajuće determinante ovoga sistema

$$D = c_{m-2}^2 - c'_{m-1}c_{m-3},$$
$$D_p = b_{m-1}c_{m-2} - b_m c_{m-3}, \qquad D_q = -b_{m-1}c'_{m-1} + b_m c_{m-2},$$

njegova rešenja su

$$\Delta p = \frac{D_p}{D}, \qquad \Delta q = \frac{D_q}{D}.$$

Dakle, polazeći od početnih aproksimacija koeficijenata p i q na svakom koraku iterativnog algoritma se formulom (40) računa niz $\{b_k\}$ i na osnovu njega formulom (47) niz $\{c_k\}$, čija poslednja tri člana određuju koeficijente sistema (49). Rešenja ovog sistema su popravke koje treba dodati poslednje određenim aproksimacijama da bi se dobile nove aproksimacije koeficijenata kvadratnog faktora. Ako su poznate približne vrednosti korena polinoma α_1 i α_2, za početne aproksimacije, zbog ubrzanja konvergencije, treba uzeti

$$p \approx -(\alpha_1 + \alpha_2), \qquad q \approx \alpha_1\alpha_2.$$

Kada je nađen jedan kvadratni faktor, sledeći se određuje primenom ovog algoritma na polinom količnik, čiji koeficijenti su, obzirom na (38), približno jednaki poslednje izračunatim vrednostima b_k, $k = 1, \ldots, m-2$. Na ovaj način se može izvršiti potpuna faktorizacija polinoma $P_n(x)$ na kvadratne i linearne (ako je polinom neparnog stepena) činioce, čiji koreni se neposredno mogu odrediti.

PRIMER 6. Ilustrujmo primenu metode na polinomu

$$P_4(x) = x^4 + 7x^3 + 24x^2 + 25x - 15.$$

Radi preglednijeg zapisa, rezultati izračunavanja su prikazani tabelom

p		0		1.224		1.908		2.000
q		0		-0.625		-0.954		-1.000
k	a	b	c	b	c	b	c	b
0	1	1	1	1	1	1	1	1
1	7	7	7	5.776	4.552	5.092	3.185	5.000
2	24	24	24	17.555	12.609	15.240	10.118	15.000
3	25	25	25	7.123	-5.465	0.786	-15.476	0.001
4	-15	-15		-12.746		-1.969		-0.005
Δp		1.224		0.684		0.092		
Δq		-0.625		-0.328		-0.046		

Polazni polinom se može zapisati u obliku

$$P_4(x) = (x^2 + 2x - 1)(x^2 + 5x + 15),$$

odakle se koreni neposredno izračunavaju.

7.5 Gradijentne metode

U uvodnom delu ovog poglavlja je napomenuto da se rešenje sistema (1), tj. jednačine (2), može odrediti i metodama minimizacije funkcionala, jer je ono tačka minimuma funkcionala F definisanog izrazom (3).

Kao i metode za direktno rešavanje nelinearnih sistema, i metode minimizacije su iterativne i većina od njih se može predstaviti iterativnom formulom

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \lambda_n \mathbf{d}_n, \qquad n = 0, 1, \ldots . \tag{50}$$

$\mathbf{x}_n$ je približno tačka u kojoj funkcional dostiže minimum, nađena u n-tom koraku algoritma, a $\lambda_n \mathbf{d}_n$ je popravka, pri čemu vektor $\mathbf{d}_n = (d_1^{(n)}, \ldots, d_m^{(n)})^T$ definiše pravac, a λ_n veličinu popravke. Metode minimizacije se uglavnom razlikuju po načinu izbora pravca minimizacije $\mathbf{d}_n$. Pošto u većini metoda figuriše neka aproksimacija gradijenta funkcionala

$$\nabla F(\mathbf{x}) = \left(\frac{\partial F(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial F(\mathbf{x})}{\partial x_m} \right), \tag{51}$$

one se obično nazivaju gradijentne metode.

Skalar λ_n se određuje tako da bude

$$F(\mathbf{x}_{n+1}) \leq F(\mathbf{x}_n), \qquad n = 0, 1, \ldots,$$

štaviše, da funkcional F na pravcu $\mathbf{d}_n$ ima minimum u tački $\mathbf{x}_{n+1}$.

Ako je $F(\mathbf{x})$ analitička funkcija u tački $\mathbf{x}_n$, može se razviti u Taylorov red

$$\begin{aligned} F(\mathbf{x}_n + \lambda_n \mathbf{d}_n) &= F(\mathbf{x}_n) + \lambda_n \sum_{i=1}^{m} \frac{\partial F(\mathbf{x}_n)}{\partial x_i} d_i^{(n)} \\ &+ \frac{1}{2} \lambda_n^2 \sum_{i,j=1}^{m} \frac{\partial^2 F(\mathbf{x}_n)}{\partial x_i \partial x_j} d_i^{(n)} d_j^{(n)} + \ldots . \end{aligned} \tag{52}$$

Vektor $\mathbf{d}_n = (d_1^{(n)}, \ldots, d_m^{(n)})^\top$, a simetrična matrica dimenzije $m \times m$

$$H(\mathbf{x}) = \left(\frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_j} \right) \tag{53}$$

naziva se Hessian funkcionala $F(\mathbf{x})$. Koristeći oznake (51) za gradijent i (53) za Hessian, aproksimaciju funkcionala $F(\mathbf{x})$ sa prva tri člana razvoja reda (52) možemo zapisati u obliku

$$F(\mathbf{x}_n + \lambda_n \mathbf{d}_n) \approx F(\mathbf{x}_n) + \lambda_n \nabla F(\mathbf{x}_n) \cdot \mathbf{d}_n + \frac{1}{2} \lambda_n^2 (\mathbf{d}_n)^T H(\mathbf{x}_n)\, \mathbf{d}_n.$$

Pod pretpostavkom da je u okolini tačke $\mathbf{x}_n$ površ $F(\mathbf{x})$ konveksna, za zadati pravac $\mathbf{d}_n$ minimum se određuje iz uslova da je

$$\frac{\partial F(\mathbf{x}_n + \lambda_n \mathbf{d}_n)}{\partial \lambda_n} = 0 \approx \nabla F(\mathbf{x}_n) \cdot \mathbf{d}_n + \lambda_n (\mathbf{d}_n)^T H(\mathbf{x}_n)\, \mathbf{d}_n,$$

tj. približno

$$\lambda_n = -\frac{\nabla F(\mathbf{x}_n) \cdot \mathbf{d}_n}{(\mathbf{d}_n)^T H(\mathbf{x}_n) \mathbf{d}_n} = -\frac{\sum_{i=1}^{m} \frac{\partial F(\mathbf{x}_n)}{\partial x_i} d_i^{(n)}}{\sum_{i=1}^{m} \sum_{j=1}^{m} \frac{\partial^2 F(\mathbf{x}_n)}{\partial x_i \partial x_j} d_i^{(n)} d_j^{(n)}}. \tag{54}$$

Metoda pokoordinatnog spusta. Ovo je jedna od najjednostavnijih gradijentnih metoda. Za vektor $\mathbf{d}_n$, $n = 0, 1, \ldots$, se bira jedan od jediničnih koordinatnih vektora $\mathbf{e}_n^{(k)} = (0, \ldots, 1, \ldots, 0)^T$, sa k-tom koordinatom jednakom jedan. Tada je

$$\nabla F(\mathbf{x}_n) \cdot \mathbf{d}_n = \frac{\partial F(\mathbf{x}_n)}{\partial x_k}, \qquad (\mathbf{d}_n)^T H(\mathbf{x}_n) \mathbf{d}_n = \frac{\partial^2 F(\mathbf{x}_n)}{\partial x_k^2},$$

i zamenom u (54) imamo da je

$$\lambda_n = -\frac{\frac{\partial F(\mathbf{x}_n)}{\partial x_k}}{\frac{\partial^2 F(\mathbf{x}_n)}{\partial x_k^2}}, \tag{55}$$

čime je formulom (50) metoda potpuno određena. Izbor koordinatnog vektora u svakom koraku je proizvoljan. Ako se uoči da spust po nekim koordinatnim pravcima obezbeđuje brže opadanje funkcionala F, mogu se ti pravci češće koristiti.

Metoda najbržeg spusta. Pravac po kome funkcional $F(\mathbf{x})$ najbrže raste je pravac njegovog gradijenta $\nabla F(\mathbf{x})$. Stoga se za vektor $\mathbf{d}_n$ u formuli (50) može uzeti pravac najbržeg opadanja funkcionala $F(\mathbf{x})$,

$$\mathbf{d}_n = -\nabla F(\mathbf{x}_n).$$

Gradijentna metoda u kojoj se ka minimumu krećemo u pravcu gradijenta funkcionala $F(\mathbf{x})$ naziva se metoda najbržeg spusta, i definisana je iterativnom formulom

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \lambda_n \nabla F(\mathbf{x}_n). \tag{56}$$

I u ovoj metodi se skalar λ_n određuje formulom (54), tj. tako da funkcional bude na pravcu $\nabla F(\mathbf{x}_n)$ minimalan u novoj iteraciji $\mathbf{x}_{n+1}$. O konvergenciji ove metode govori sledeća teorema

TEOREMA 6. *Neka je $F(\mathbf{x})$ kvadratni funkcional oblika*

$$F(\mathbf{x}) = c - \mathbf{b}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T H\mathbf{x}, \tag{57}$$

gde je H pozitivno definisana matrica. Neka je $\{\mathbf{x}_n\}$ niz tačaka određenih rekurentnom formulom

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \lambda_n \nabla F(\mathbf{x}_n),$$

gde je λ_n dato formulom (54). Tada

(i) postoji jedinstvena tačka $\mathbf{x}^*$ u kojoj funkcional $F(\mathbf{x})$ dostiže minimum,

(ii) niz $\{\mathbf{x}_n\}$ konvergira ka tački $\mathbf{x}^*$.

Dokaz ove teoreme se može naći u [19].

Ako $F(\mathbf{x})$ nije kvadratni funkcional, iterativni algoritam (56) konvergira ukoliko je $\mathbf{x}_0$ dovoljno blisko $\mathbf{x}^*$ tako da u razvoju (52) glavni doprinos potiče od prva tri člana.

PRIMER 7. Rešenje sistema linearnih jednačina $A\mathbf{x} = \mathbf{b}$, gde je A pozitivno definisana matrica, je tačka minimuma kvadratnog funkcionala

$$F(\mathbf{x}) = (A\mathbf{x}\,,\,\mathbf{x}) - 2(\mathbf{b}\,,\,\mathbf{x}),$$

i obrnuto. Primenjujući metodu najbržeg spusta za nalaženje minimuma ovog funkcionala, dobijamo iterativni algoritam za rešavanje sistema linearnih jednačina

$$\mathbf{r}_n = \mathbf{b} - A\mathbf{x}_n, \qquad \delta_n = \frac{(\mathbf{r}_n\,,\,\mathbf{r}_n)}{(\mathbf{r}_n\,,\,A\mathbf{r}_n)}, \qquad \mathbf{x}_{n+1} = \mathbf{x}_n + \delta_n\mathbf{r}_n.$$

Metoda konjugovanih pravaca. Ovo je metoda u kojoj se za vektore $\mathbf{d}_n$ u formuli (50) biraju elementi konjugovanog bazisa u odnosu na Hessian H funkcionala F. Skalar λ_n se , kao i u ostalim metodama, određuje formulom (54).

Neka je H pozitivno definisana matrica dimenzije $m \times m$. Dva nenula vektora $\mathbf{p}, \mathbf{q} \in \mathcal{R}^m$ su *konjugovana* u odnosu na matricu H, ako je

$$\mathbf{p}^T H\mathbf{q} = 0.$$

Sistem od m vektora $\mathbf{p}_1, \ldots, \mathbf{p}_m$ se naziva *konjugovanim bazisom*, ako su vektori $\mathbf{p}_i$ i $\mathbf{p}_j$ uzajamno konjugovani za sve $i \neq j$. Kako je matrica H pozitivno definisana, može se uvesti skalarni proizvod $(\mathbf{p}\,,\,\mathbf{q})_H = \mathbf{p}^T H\mathbf{q}$. Stoga se može reći da su vektori $\mathbf{p}$ i $\mathbf{q}$ konjugovani u odnosu na matricu H ako i samo ako su *H-ortogonalni*, tj. ortogonalni u smislu uvedenog skalarnog proizvoda.

TEOREMA 7. *Neka je F funkcional zadat formulom*

$$F(\mathbf{x}) = c - \mathbf{b}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T H\mathbf{x},$$

pri čemu je H pozitivno definisana matrica dimenzije $m \times m$, i neka vektori $\mathbf{p}_0, \ldots, \mathbf{p}_{m-1}$ čine konjugovani bazis u odnosu na matricu H. Tada među vektorima $\mathbf{x}_n$, određenim formulama

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \lambda_n \mathbf{p}_n, \tag{58}$$

$$\lambda_n = \frac{(H\mathbf{x}_n - \mathbf{b})^T \mathbf{p}_n}{(H\mathbf{p}_n)^T \mathbf{p}_n}, \tag{59}$$

postoji vektor $\mathbf{x}_k$ takav da je $\mathbf{x}_k = H^{-1}\mathbf{b}$, tj. $\nabla F(\mathbf{x}_k) = -\mathbf{b} + H\mathbf{x}_k = 0$, za neko $k \leq m$.

DOKAZ: Iz (58) i (59) je za svako n i j

$$(H\mathbf{x}_{n+1} - \mathbf{b})^T \mathbf{p}_j = \big(H(\mathbf{x}_n - \lambda_n \mathbf{p}_n) - \mathbf{b}\big)^T \mathbf{p}_j = (H\mathbf{x}_n - \mathbf{b})^T \mathbf{p}_j - \lambda_n (H\mathbf{p}_n)^T \mathbf{p}_j$$
$$= (H\mathbf{x}_n - \mathbf{b})^T \mathbf{p}_j - \frac{(H\mathbf{x}_n - \mathbf{b})^T \mathbf{p}_n}{(H\mathbf{p}_n)^T \mathbf{p}_n} (H\mathbf{p}_n)^T \mathbf{p}_j.$$

Ako je $j \neq n$ drugi sabirak je nula zbog konjugovanosti vektora $\mathbf{p}_n$, a ako je $j = n$ taj sabirak je jednak prvom, te je

$$(H\mathbf{x}_{n+1} - \mathbf{b})^T \mathbf{p}_j = \begin{cases} (H\mathbf{x}_n - \mathbf{b})^T \mathbf{p}_j, & \text{za } j \neq n \\ 0, & \text{za } j = n. \end{cases} \tag{60}$$

Stoga je za $j < m - 1$, na osnovu jednakosti (60) za $j \neq n$,

$$(H\mathbf{x}_m - \mathbf{b})^T \mathbf{p}_j = (H\mathbf{x}_{m-1} - \mathbf{b})^T \mathbf{p}_j = \cdots = (H\mathbf{x}_{j+2} - \mathbf{b})^T \mathbf{p}_j = (H\mathbf{x}_{j+1} - \mathbf{b})^T \mathbf{p}_j,$$

a na osnovu iste jednakosti za $j = n$ je

$$(H\mathbf{x}_{j+1} - \mathbf{b})^T \mathbf{p}_j = 0.$$

Dakle,

$$(H\mathbf{x}_m - \mathbf{b})^T \mathbf{p}_j = 0, \qquad j = 0, \ldots, m-1,$$

pa zbog linearne nezavisnosti vektora $\mathbf{p}_j$ sledi da je

$$H\mathbf{x}_m - \mathbf{b} = 0.$$

Može se dogoditi da je $H\mathbf{x}_k - \mathbf{b} = 0$ za neko $k \leq m$, no tada je, s obzirom na (59), $\lambda_k = 0$, te je $\mathbf{x}_k = \mathbf{x}_{k+1}$, i uopšte $\mathbf{x}_k = \mathbf{x}_{k+1} = \cdots = \mathbf{x}_m$. ∎

Stoga, metoda konjugovanih pravaca u slučaju minimizacije kvadratnog funkcionala daje rešenje u konačnom broju koraka, koji je u najgorem slučaju jednak dimenziji problema. Za funkcional F koji nije kvadratni, u opštem slučaju na kraju ciklusa od m koraka će se dobiti neka aproksimacija tačke minimuma. Uzimamo je

kao početnu aproksimaciju $\mathbf{x}_0$ za sledeći ciklus od m koraka, i na taj način definišemo iterativni postupak koji kvadratno konvergira ka minimumu funkcionala F.

Konjugovane vektore $\mathbf{p}_0, \ldots, \mathbf{p}_{m-1}$, moguće je odrediti na različite načine. Ako se oni određuju pomoću gradijenta funkcionala F,

$$\mathbf{p}_{n+1} = \nabla F(\mathbf{x}_{n+1}) - \mu_n \mathbf{p}_n,$$

pri čemu su skalari μ_n određeni uslovima konjugovanosti $(\mathbf{p}_n)^T H \mathbf{p}_{n+1} = 0$, metoda se naziva *metodom konjugovanih gradijenata*. Formule kojima je definisana su

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \lambda_n \mathbf{p}_n, \qquad \lambda_n = \frac{(H\mathbf{x}_n - \mathbf{b})^T \mathbf{p}_n}{(H\mathbf{p}_n)^T \mathbf{p}_n},$$

$$\mathbf{p}_0 = H\mathbf{x}_0 - \mathbf{b}, \qquad \mathbf{p}_{n+1} = H\mathbf{x}_{n+1} - \mathbf{b} - \mu_n \mathbf{p}_n, \qquad \mu_n = \frac{(H\mathbf{x}_{n+1} - \mathbf{b})^T H \mathbf{p}_n}{(H\mathbf{p}_n)^T \mathbf{p}_n},$$

S obzirom da se ovom metodom minimum kvadratnog funkcionala (57) određuje kao nula njegovog gradijenta $\nabla F(\mathbf{x}) = H\mathbf{x} - \mathbf{b}$, metoda se može koristiti i za rešavanje sistema linearnih jednačina $H\mathbf{x} = \mathbf{b}$.

8

Obične diferencijalne jednačine – Cauchyevi problemi

Zbog značaja diferencijalnih jednačina u praksi, vrlo je važan razvoj algoritama za njihovo rešavanje. U konkretnim problemima mogu se pojaviti diferencijalne jednačine ma kog reda ili sistemi ovih jednačina. Većinom numeričkih metoda se rešava jedna ili sistem jednačina prvog reda. To ne predstavlja ograničavajući faktor, jer se diferencijalna jednačina reda m

$$u^{(m)}(x) = f(x, u, u', \ldots, u^{(m-1)})$$

može smenom, na primer

$$u_k(x) \equiv u^{(k)}(x),$$

svesti na sistem od m diferencijalnih jednačina prvog reda

$$\begin{aligned} u_k'(x) &= u_{k+1}(x), \qquad k = 0, \ldots, m-2, \\ u_{m-1}' &= f(x, u_0, \ldots, u_{m-1}) \qquad (u_0(x) \equiv u(x)). \end{aligned}$$

Stoga će se, po pravilu, u daljem tekstu rešavati sistem jednačina prvog reda

$$u_k'(x) = f_k(x, u_1, \ldots, u_m), \qquad k = 1, \ldots, m,$$

koji se, kratkoće radi, može zapisati u vektorskom obliku

$$\mathbf{u}'(x) = \mathbf{f}(x, \mathbf{u}(x)), \tag{1}$$

gde je $\mathbf{u} = (u_1, \ldots, u_m)^T$, $\mathbf{f} = (f_1, \ldots, f_m)^T$.

Zadavanjem m uslova koje rešenje treba da zadovoljava, iz opšteg rešenja jednačine (1) se određuje tzv. partikularno rešenje. Ako su svi uslovi zadati u jednoj tački, označimo je sa x_0, onda je jednačinom (1), dopunjenom ovim uslovima, definisan problem koji se naziva problem početnih vrednosti ili *Cauchyev problem.*

Dakle, Cauchyev problem se sastoji u određivanju onog rešenja jednačine (1) koje prolazi kroz datu tačku, tj. rešenja koje zadovoljava uslove

$$u_k(x_0) = u_{0k}, \qquad k = 1, \ldots, m,$$

ili, u vektorskom obliku,

$$\mathbf{u}(x_0) = \mathbf{u}_0, \qquad \text{gde je } \mathbf{u}_0 = (u_{01}, \ldots, u_{0m}). \tag{2}$$

Poznato je iz teorije običnih diferencijalnih jednačina da Cauchyev problem (1),(2) ima rešenje ako su funkcije f_k neprekidne i ograničene u nekoj okolini početne tačke $(x_0, u_{01}, \ldots, u_{0m})$. Ako još funkcije f_k zadovoljavaju Lipschitzov uslov po promenljivim u_j,

$$|f_k(x, \bar{u}_1, \ldots, \bar{u}_m) - f_k(x, u_1, \ldots, u_m)| \le L \sum_{j=1}^{m} |\bar{u}_j - u_j|, \qquad k = 1, \ldots, m,$$

rešenje je jedinstveno i neprekidno zavisi od početnih uslova, što znači da je problem korektno postavljen.

Relativno mali broj Cauchyevih problema se može tačno rešiti. Čak i na izgled jednostavna jednačina

$$u'(x) = x^2 + u^2$$

nema rešenje koje se može izraziti elementarnim funkcijama. Sa druge strane, u mnogim slučajevima i kada je moguće naći tačno rešenje problema, ono je u takvom obliku da se opet numeričke metode moraju koristiti za, na primer, sastavljanje tablice vrednosti tog rešenja. Tako, opšte rešenje jednačine

$$u'(x) = \frac{u - x}{u + x} \qquad \text{je} \qquad \frac{1}{2} \ln (x^2 + u^2) + \arctan \frac{u}{x} = const,$$

i da bi se odredile vrednosti u za date vrednosti x treba više puta rešavati transcendentnu jednačinu, što nije nimalo jednostavnije nego numeričkim metodama rešavati direktno diferencijalnu jednačinu.

Numeričkim metodama se najčešće računaju približne, a ponekad i tačne vrednosti traženog rešenja $u(x)$ na nekoj unapred izabranoj mreži tačaka x_n. Pri tome se rešenje dobija u obliku tabele. Ove metode se mogu primeniti samo za rešavanje korektno postavljenih problema. Osim toga, za uspešnu primenu numeričkih metoda je potrebno da je problem i dobro uslovljen, tj. da male promene ulaznih parametara dovode do malih promena rešenja. Ako je problem loše uslovljen, računske greške koje se neminovno javljaju pri realizaciji numeričkog algoritma i koje mogu biti tretirane kao male promene ulaznih parametara, mogu znatno izmeniti približno rešenje.

PRIMER 1. Opšte rešenje jednačine

$$u'(x) = u - x \qquad \text{je} \qquad u(x, c) = 1 + x + ce^x,$$

gde je c proizvoljna konstanta. Partikularno rešenje koje zadovoljava uslov $u(0) = 1$ je $u(x, 0) = 1 + x$ i njegova vrednost $u(100, 0) = 101$. Ako se početni uslov izmeni samo za 10^{-6}, tj. ako je $u(0) = 1.000001$, rešenje Cauchyevog problema je $u(x, 10^{-6}) = 1 + x + 10^{-6}e^x$, te je $u(100, 10^{-6}) = 2.7 * 10^{37}$.

Pored pomenutih, koriste se i aproksimativne metode, mada znatno ređe jer nisu pogodne za realizaciju na računaru. Pomenimo ih ukratko.

8.1 Aproksimativne metode

Ovim metodama se rešenje Cauchyevog problema (1),(2) dobija kao granica niza funkcija $\mathbf{u}_n(x)$, pri čemu se funkcije $\mathbf{u}_n(x)$ izražavaju elementarnim funkcijama i njihovim integralima. Zadržavajući se na nekom konačnom n, dobijamo aproksimativni izraz za rešenje problema $\mathbf{u}(x)$. Obično se ovim metodama određuju približne vrednosti rešenja u nekim tačkama iz neposredne okoline početne tačke koje su neophodne za realizaciju drugih metoda, te se zadovoljavajuća tačnost postiže za malo n.

Metoda uzastopnih aproksimacija. Ova metoda naziva se i metoda Picarda. Integraljenjem jednačine (1) od početne tačke x_0 do proizvoljne tačke x problem (1),(2) svodimo na njemu ekvivalentan problem definisan Volterraovom integralnom jednačinom

$$\mathbf{u}(x) = \mathbf{u}_0 + \int_{x_0}^{x} \mathbf{f}(t, \mathbf{u}(t))\, dt. \tag{3}$$

Integralnom jednačinom (3) definiše se niz funkcija $\mathbf{u}_n(x)$ rekurentnom formulom

$$\begin{aligned} \mathbf{u}_0(x) &= \mathbf{u}_0 \\ \mathbf{u}_{n+1}(x) &= \mathbf{u}_0 + \int_{x_0}^{x} \mathbf{f}(t, \mathbf{u}_n(t))\, dt, \qquad n = 0, 1, \ldots . \end{aligned} \tag{4}$$

Početnu aproksimaciju $\mathbf{u}_0(x)$ je moguće i na drugi način izabrati. Metodu ima smisla koristiti samo ako se integrali u formuli (4) mogu izračunati analitički.

Metoda Taylorovog razvoja. Rešavamo Cauchyev problem definisan diferencijalnom jednačinom prvog reda

$$u'(x) = f(x, u), \qquad u(x_0) = u_0. \tag{5}$$

Pretpostavimo da je funkcija $f(x, u)$ analitička u tački (x_0, u_0). Uzastopnim diferenciranjem jednačine po x imamo da je

$$\begin{aligned} u'' &= f_x + f_u u' \\ u''' &= f_{xx} + 2f_{xu}u' + f_{uu}{u'}^2 + f_u u'' \\ &\vdots \end{aligned}$$

te iz jednačine i dobijenih izraza možemo izračunati vrednosti $u'(x_0)$, $u''(x_0)$, $u'''(x_0), \ldots$. Približno rešenje problema (5) je

$$u_n(x) = \sum_{j=0}^{n} \frac{u^{(j)}(x_0)}{j!}(x - x_0)^j \tag{6}$$

za $|x - x_0| \leq R$, gde je R poluprečnik konvergencije reda

$$\sum_{j=0}^{\infty} \frac{u^{(j)}(x_0)}{j!}(x - x_0)^j.$$

Cauchyev problem definisan diferencijalnom jednačinom višeg reda,

$$u^{(m)}(x) = f(x, u, u', \ldots, u^{(m-1)})$$
$$u(x_0) = u_0, \quad u'(x_0) = u_0', \quad \ldots, \quad u^{(m-1)}(x_0) = u_0^{(m-1)},$$

nema potrebe svoditi na sistem diferencijalnih jednačina prvog reda. Prvih m koeficijenata aproksimacije (6) dato je početnim uslovima problema, a ostalih $n-m+1$ koeficijenata se nalazi uzastopnim diferenciranjem jednačine, na već pomenuti način.

Metoda stepenih redova. Kada je diferencijalna jednačina kojom je definisan Cauchyev problem linearna, rešenje se može tražiti u obliku stepenog reda, ili čak uopštenog stepenog reda (reda sa razlomljenim stepenima od x). Ilustrujmo metodu na linearnoj diferencijalnoj jednačini drugog reda, za koju se ova metoda najčešće primenjuje – na ovaj način se određuje niz specijalnih funkcija. Dakle, rešavamo Cauchyev problem

$$u''(x) + p(x)u'(x) + q(x)u(x) = f(x) \tag{7}$$

$$u(0) = u_0, \qquad u'(0) = u_0', \tag{8}$$

pri čemu pretpostavljamo da je koeficijente jednačine moguće razviti u stepene redove

$$p(x) = \sum_{j=0}^{\infty} p_j x^j, \qquad q(x) = \sum_{j=0}^{\infty} q_j x^j, \qquad f(x) = \sum_{j=0}^{\infty} f_j x^j. \tag{9}$$

Rešenje takođe tražimo u obliku stepenog reda

$$u(x) = \sum_{j=0}^{\infty} c_j x^j, \tag{10}$$

odakle diferenciranjem dobijamo i razvoje za funkcije $u'(x)$ i $u''(x)$,

$$u'(x) = \sum_{j=1}^{\infty} j c_j x^{j-1}, \qquad u''(x) = \sum_{j=2}^{\infty} j(j-1) c_j x^{j-2}. \tag{11}$$

Uvrstimo razvoje (9),(10) i (11) u jednačinu (7), i dobijamo

$$\sum_{j=2}^{\infty} j(j-1)c_j x^{j-2} + \left(\sum_{j=0}^{\infty} p_j x^j\right)\left(\sum_{j=1}^{\infty} j c_j x^{j-1}\right) + \left(\sum_{j=0}^{\infty} q_j x^j\right)\left(\sum_{j=0}^{\infty} c_j x^j\right) = \sum_{j=0}^{\infty} f_j x^j.$$

Pošto rešenje mora da zadovoljava jednačinu (7) za svako x iz intervala definisanosti problema, poslednji izraz mora biti identitet, te koeficijenti uz odgovarajuće stepene x na levoj i desnoj strani moraju biti jednaki. Tako dobijamo rekurentne veze za koeficijente c_j reda (10),

$$\begin{aligned} 2c_2 + p_0c_1 + q_0c_0 &= f_0 \\ 6c_3 + 2p_0c_2 + p_1c_1 + q_0c_1 + q_1c_0 &= f_1 \\ &\vdots \end{aligned}$$

Kada uvrstimo početne vrednosti (8) u (10) i prvi od razvoja (11), dobijamo početne vrednosti ovih rekurentnih formula

$$c_0 = u_0, \qquad c_1 = u_0'.$$

Jasno je da je eksplicitno nalaženje početnih vrednosti moguće samo kada su početni uslovi (8) dati u tački $x_0 = 0$. Ova pretpostavka, međutim, ne umanjuje opštost metode s obzirom da se smenom $x - x_0 = t$ linearan Cauchyev problem sa početnim uslovima zadatim u proizvoljnoj tački x_0 svodi na problem tipa (7),(8).

Približno rešenje $u_n(x)$ je n-ta parcijalna suma reda (10). Poluprečnik konvergencije reda (10) jednak je najmanjem poluprečniku konvergencije redova (9).

8.2 Metode tipa Runge–Kutta

Ovim metodama se na osnovu poznate vrednosti rešenja Cauchyevog problema u tački x, određuje približna vrednost rešenja u tački $x + h$. Radi jednostavnosti, ilustrujmo ih na jednačini prvog reda, u Cauchyevom problemu (1),(2) je $m = 1$. Integralimo jednačinu (1) u granicama od x do $x+h$, pa je tražena vrednost rešenja

$$u(x+h) = u(x) + \int_x^{x+h} f(t, u(t))\, dt. \tag{12}$$

Zamenom integrala na desnoj strani relacije (12) kvadraturnom formulom, dobijamo da je

$$u(x+h) \approx u(x) + h\sum_{i=1}^{n} c_i f(x_i, u(x_i)), \tag{13}$$

gde su c_i koeficijenti i $x_i \in (x, x+h)$ čvorovi kvadraturne formule. Nevolja je što je argument funkcije f rešenje $u(x)$ jednačine (1), a u formuli (13) figurišu vrednosti $u(x_i)$ koje nisu poznate. Ako sukcesivno aproksimiramo ove vrednosti pomoću već određenih aproksimacija rešenja u prethodnim čvorovima, izraz (13) se može napisati u sledećem obliku

$$u(x+h) \approx v(x+h) \equiv u(x) + \sum_{i=1}^{n} c_i k_i(h), \tag{14}$$

gde je za $u \equiv u(x)$ i $x_i = x + \alpha_i h$, $0 = \alpha_1 \le \alpha_2 \le \cdots \le \alpha_n \le 1$,

$$\begin{aligned} k_1(h) &= hf(x,u) \\ k_2(h) &= hf\big(x+\alpha_2 h, u+\beta_{21}k_1(h)\big) \\ &\vdots \\ k_n(h) &= hf\big(x+\alpha_n h, u+\beta_{n1}k_1(h)+\cdots+\beta_{n,n-1}k_{n-1}(h)\big). \end{aligned}$$

Različitim izborom parametara $\alpha_2, \ldots, \alpha_n$, $c_1, \ldots, c_n$ i β_{ij}, $0 < j < i \le n$, formulom (14) su definisane različite metode tipa Runge–Kutta.

Neka je greška metode na jednom koraku

$$\epsilon(h) = u(x+h) - v(x+h). \tag{15}$$

Ako je $f(x,u)$ dovoljno glatka funkcija svojih argumenata, onda su $k_1(h), \ldots, k_n(h)$ i $\epsilon(h)$ glatke funkcije parametra h. Pretpostavimo da je

$$\epsilon(0) = \epsilon'(0) = \cdots = \epsilon^{(p)}(0) = 0$$

za proizvoljnu, dovoljno glatku funkciju $f(x,u)$, a da postoji glatka funkcija $f(x,u)$ za koju je

$$\epsilon^{(p+1)}(0) \neq 0.$$

Na osnovu Taylorove formule je tada greška metode na jednom koraku

$$\epsilon(h) = \sum_{i=0}^{p} \frac{\epsilon^{(i)}(0)}{i!} h^i + \frac{\epsilon^{(p+1)}(\theta h)}{(p+1)!} h^{p+1} = \frac{\epsilon^{(p+1)}(\theta h)}{(p+1)!} h^{p+1}, \qquad 0 < \theta < 1. \tag{16}$$

Broj p je *red greške metode*.

Analizirajmo neke od mogućih izbora parametara formule (14) i greške odgovarajućih metoda.

$\boxed{n=1}$ Prema (14) i (15) je

$$\epsilon(h) = u(x+h) - u(x) - c_1 h f(x,u), \tag{17}$$

te je

$$\epsilon'(h) = u'(x+h) - c_1 f(x,u), \qquad \epsilon''(h) = u''(x+h).$$

Kako je $\epsilon(0) = 0$ i $\epsilon'(0) = (1 - c_1)f(x, u) = 0$ za $c_1 = 1$, za svaku funkciju $f(x, u)$, to za $c_1 = 1$ imamo iz (17) i (16) da je

$$u(x + h) = u(x) + hf(x, u) + \frac{u''(x + \theta h)}{2}h^2.$$

Stoga je formula tipa Runge–Kutta za $n = 1$

$$v(x + h) = u(x) + hf(x, u(x)) \tag{18}$$

i predstavlja formulu Eulera. Eulerovom metodom je rešenje određeno sa greškom $O(h^2)$, te je to metoda reda jedan.

$\boxed{n = 2}$ U ovom slučaju je greška metode

$$\epsilon(h) = u(x + h) - u(x) - c_1hf(x, u) - c_2hf(\bar{x}, \bar{u}), \tag{19}$$

gde je

$$\bar{x} = x + \alpha_2 h, \qquad \bar{u} = u + \beta_{21}hf(x, u).$$

Diferenciranjem izraza (19) po h, dobijamo da je

$$\begin{aligned}
\epsilon'(h) &= u'(x + h) - c_1 f(x, u) - c_2 f(\bar{x}, \bar{u}) - c_2 h\big(\alpha_2 f_x(\bar{x}, \bar{u}) + \beta_{21} f_u(\bar{x}, \bar{u}) f(x, u)\big) \\
\epsilon''(h) &= u''(x + h) - 2c_2\big(\alpha_2 f_x(\bar{x}, \bar{u}) + \beta_{21} f_u(\bar{x}, \bar{u}) f(x, u)\big) \\
&\quad - c_2 h\big(\alpha_2^2 f_{xx}(\bar{x}, \bar{u}) + 2\alpha_2\beta_{21} f_{xu}(\bar{x}, \bar{u}) f(x, u) + \beta_{21}^2 f_{uu}(\bar{x}, \bar{u})(f(x, u))^2\big) \\
\epsilon'''(h) &= u'''(x + h) - 3c_2\big(\alpha_2^2 f_{xx}(\bar{x}, \bar{u}) + 2\alpha_2\beta_{21} f_{xu}(\bar{x}, \bar{u}) f(x, u) \\
&\quad + \beta_{21}^2 f_{uu}(\bar{x}, \bar{u})(f(x, u))^2\big) + O(h)
\end{aligned}$$

Kako je iz diferencijalne jednačine (1)

$$u' = f, \qquad u'' = f_x + f_u f, \qquad u''' = f_{xx} + 2f_{xu}f + f_{uu}f^2 + f_u(f_x + f_u f),$$

to je

$$\begin{aligned}
\epsilon(0) &= 0 \\
\epsilon'(0) &= (1 - c_1 - c_2) f(x, u) \\
\epsilon''(0) &= (1 - 2c_2\alpha_2) f_x(x, u) + (1 - 2c_2\beta_{21}) f_u(x, u) f(x, u) \\
\epsilon'''(0) &= (1 - 3c_2\alpha_2^2) f_{xx}(x, u) + (2 - 6c_2\alpha_2\beta_{21}) f_{xu}(x, u) f(x, u) \\
&\qquad + (1 - 3c_2\beta_{21}^2) f_{uu}(x, u) f^2(x, u) + f_u(x, u) u''(x).
\end{aligned}$$

Za svaku dovoljno glatku funkciju $f(x, u)$ je

$$\epsilon'(0) = 0 \qquad \text{ako je} \qquad 1 - c_1 - c_2 = 0, \tag{20}$$

$$\epsilon''(0) = 0 \qquad \text{ako je} \qquad 1 - 2c_2\alpha_2 = 0 \quad \text{i} \quad 1 - 2c_2\beta_{21} = 0, \tag{21}$$

te je, na osnovu (16) za $p = 2$ i (19),

$$
\begin{aligned}
u(x+h) &= u(x) + c_1 h f(x,u) \\
&\quad + c_2 h f(x+\alpha_2 h, u+\beta_{21} h f(x,u)) + \frac{\epsilon'''(\theta h)}{6} h^3.
\end{aligned} \tag{22}
$$

Uslovi (20) i (21) daju tri veze između četiri parametra, i stoga jedan od njih može biti izabran proizvoljno.

Ako izaberemo da je $c_1 = \frac{1}{2}$, onda je $c_2 = \frac{1}{2}$, $\alpha_2 = 1$, $\beta_{21} = 1$, te zanemarivanjem greške iz (22) dobijamo da je

$$
v(x+h) = u(x) + \frac{h}{2}\Big(f(x,u(x)) + f\big(x+h, u(x) + h f(x,u(x))\big)\Big), \tag{23}
$$

što predstavlja jednu modifikaciju Eulerove metode, čija je greška $O(h^3)$.

Ako izaberemo da je $c_1 = 0$, onda je $c_2 = 1$, $\alpha_2 = \frac{1}{2}$, $\beta_{21} = \frac{1}{2}$, i opet iz (22) dobijamo da je

$$
v(x+h) = u(x) + h f\big(x + \tfrac{h}{2}, u(x) + \tfrac{h}{2} f(x,u(x))\big). \tag{24}
$$

Ovo je druga modifikacija Eulerove metode kojom se povećava njena tačnost, greška joj je takođe $O(h^3)$. Formulom (23) je, ustvari, približno rešenje Cauchyevog problema određeno aproksimacijom integrala u izrazu (12) trapeznim pravilom, a formulom (24) pravilom pravougaonika, pri čemu je u oba slučaja korišćena Eulerova metoda za ocenu nepoznatih vrednosti rešenja u čvorovima.

Slobodni parametar se ne može izabrati tako da za svako $f(x,u)$ bude $\epsilon'''(0) = 0$, tj. tako da bude u (16) $p = 3$, jer je za jednačinu $u'(x) = u(x)$ nezavisno od izbora ovoga parametra $\epsilon'''(0) = u(x)$. To znači da se za $n = 2$ u izrazu (14) ne može dobiti formula reda tri, tj. sa greškom $O(h^4)$ na jednom koraku.

$\boxed{n=3}$ Da bi izrazom (14) bila definisana metoda reda $p = 3$, neophodno je, što se pokazuje analizom sličnom onoj izvršenoj u prethodnim slučajevima, da parametri zadovoljavaju uslove

$$
\alpha_2 = \beta_{21}, \qquad \alpha_3 = \beta_{31} + \beta_{32}, \qquad \alpha_3(\alpha_3 - \alpha_2) - \beta_{32}\alpha_2(2 - 3\alpha_2) = 0,
$$
$$
c_3\beta_{32}\alpha_2 = \frac{1}{6}, \qquad c_2\alpha_2 + c_3\alpha_3 = \frac{1}{2}, \qquad c_1 + c_2 + c_3 = 1.
$$

S obzirom da za osam parametara imamo šest veza, sistem ima beskonačno mnogo rešenja. Najčešće korišćena formula reda tri je

$$
k_1 = h f(x,u), \qquad k_2 = h f(x + \tfrac{h}{2}, u + \tfrac{k_1}{2}), \qquad k_3 = h f(x+h, u - k_1 + 2k_2)
$$
$$
v(x+h) = u(x) + \frac{1}{6}(k_1 + 4k_2 + k_3).
$$

Može se pokazati da se, bez obzira na neodređenost sistema po parametrima, ne može dobiti formula reda četiri.

$\boxed{n=4}$ U ovom slučaju je moguće izvesti formule najviše reda četiri, tj. sa greškom $O(h^5)$ na jednom koraku. Najčešće korišćena od njih je tzv. metoda Runge–Kutta

$$\begin{gathered} k_1 = hf(x,u), \qquad k_2 = hf(x+\tfrac{h}{2}, u+\tfrac{k_1}{2}), \\ k_3 = hf(x+\tfrac{h}{2}, u+\tfrac{k_2}{2}), \qquad k_4 = hf(x+h, u+k_3), \\ v(x+h) = u(x) + \frac{1}{6}(k_1+2k_2+2k_3+k_4). \end{gathered} \tag{25}$$

Metodama tipa Runge–Kutta se mogu rešavati i Cauchyevi problemi za sisteme jednačina prvog reda. I u ovom slučaju metode se mogu zapisati izrazom (14) (Eulerova formulom (18), njene modifikacije formulama (23) i (24) i Runge–Kutta formulom (25)), s tim što su u, v, f i k_i, $i = 1, \ldots, n$, vektorske veličine.

Rungeova ocena greške. Glavni član u grešci približnog rešenja određenog metodom reda p na jednom koraku je, prema (16),

$$\frac{\epsilon^{(p+1)}(0)}{(p+1)!}h^{p+1}.$$

Za malo h tačka $(x+h, v(x+h))$ je bliska tački $(x, u(x))$, pa će greška na sledećem koraku imati isti glavni član. Stoga je, prema (15),

$$u(x+2h) - v_1 \sim 2\frac{\epsilon^{(p+1)}(0)}{(p+1)!}h^{p+1}, \tag{26}$$

gde je sa v_1 označeno približno rešenje u tački $x+2h$ dobijeno tako što je prvo nađeno rešenje u tački $x+h$, a zatim pomoću ovog rešenje u tački $x+2h$. Ako, pak, polazeći od tačke x primenom te metode tipa Runge–Kutta direktno sa korakom $2h$ nađemo približno rešenje v_2 u tački $x+2h$, onda je

$$u(x+2h) - v_2 \sim \frac{\epsilon^{(p+1)}(0)}{(p+1)!}(2h)^{p+1}. \tag{27}$$

Eliminišući $u(x+2h)$ iz (26) i (27), dobijamo da je približno

$$v_1 + 2\frac{\epsilon^{(p+1)}(0)}{(p+1)!}h^{p+1} = v_2 + \frac{\epsilon^{(p+1)}(0)}{(p+1)!}(2h)^{p+1},$$

tj.

$$2\frac{\epsilon^{(p+1)}(0)}{(p+1)!}h^{p+1} = \frac{v_1 - v_2}{2^p - 1}.$$

Kada uvrstimo ovu procenu u (26), dobijamo da je greška rešenja određenog sa korakom h približno jednaka

$$u(x+2h) - v_1 \sim \frac{v_1 - v_2}{2^p - 1}, \tag{28}$$

što predstavlja Rungeov kriterijum za ocenu greške. Popravljena vrednost približnog rešenja je

$$u(x+2h) \approx v_1 + \frac{v_1 - v_2}{2^p - 1}. \tag{29}$$

Kontrolnim računanjem sa dvostrukim korakom imamo mogućnost ocene glavnog člana greške u svakoj drugoj tački. Ako je ova vrednost u zadovoljavajućim granicama, možemo izvršiti popravku po formuli (29). Ako ocenjena greška prelazi dozvoljene granice, treba smanjiti korak h. Obično se korak smanjuje za polovinu, tj. uzima se da je novi korak $h_1 = \frac{h}{2}$, jer se na taj način već izračunate približne vrednosti rešenja sa korakom h mogu koristiti kao vrednosti v_2 u oceni (28) za korak h_1. Ako je i dalje greška suviše velika, korak se polovi sve dok se ne postigne zadovoljavajuća tačnost na osnovu Rungeovog kriterijuma (28).

PRIMER 2. Eulerovom metodom (18) i njenim modifikacijama (23) i (24), kao i metodom Runge–Kutta (25) nađimo približno rešenje Cauchyevog problema

$$u'(x) = x^2 + u^2(x), \qquad u(0) = 0,$$

na intervalu $[0, 1]$, i ocenimo grešku R Rungeovim kriterijumom (28).

Rezultati izračunavanja za različite korake h dati su u tabeli koja sledi.

x_k	v_k			R_k	$\tilde{v}_k$	Met.
	$h=1$	$h=0.5$	$h=0.25$			
0	0	0	0	0	0	
0.25			0.000		0.008	(18)
0.50		0.000	0.016	0.016	0.032	
0.75			0.078		0.134	
1.00	0.000	0.125	0.220	0.095	0.315	
0.50		0.063			0.044	(23)
1.00	0.500	0.386		-0.038	0.348	
0.50		0.031			0.042	(24)
1.00	0.250	0.317		0.022	0.339	
0.50		0.042			0.042	(25)
1.00	0.350	0.351		0.000	0.351	

Poslednja kolona $\tilde{v}_k$ sadrži vrednosti približnog rešenja popravljene po formuli (29). U tačkama u kojima ne postoji ocena greške R, približnom rešenju je dodata aritmetička sredina procena greške R u susednim tačkama.

8.3 Prediktor–korektor metode

Metode tipa Runge–Kutta spadaju u grupu dvoslojnih metoda, jer se njima pomoću vrednosti rešenja u tački $x_{j-1} = x$ određuje približno rešenje u tački $x_j = x + h$.

$(n+1)$-slojne metode su oblika

$$\sum_{i=0}^{n} a_i v_{j-i} - h \sum_{i=0}^{n} b_i f(x_{j-i}, v_{j-i}) = 0, \tag{30}$$

gde su a_i, b_i, $i = 0, \ldots, n$, konstante. Ako je $a_0 \neq 0$ i $b_0 = 0$ metode se nazivaju *ekstrapolacione*, a ako je $a_0 \neq 0$ i $b_0 \neq 0$ one su *interpolacione*.

Prostije, a istovremeno i najčešće korišćene formule oblika (30) dobijaju se takođe pomoću kvadraturnih formula. Naime, integraljenjem jednačine (1) u granicama od x_{j-k} do x_j dobija se da je

$$u(x_j) = u(x_{j-k}) + \int_{x_{j-k}}^{x_j} f(x, u(x))\, dx, \tag{31}$$

pri čemu je radi jednostavnijih oznaka posmatran jednodimenzioni slučaj, mada se veličine u, f i v mogu smatrati i vektorima. Aproksimacijom podintegralne funkcije $f(x, u(x)) \equiv u'(x)$ njenim interpolacionim polinomom određenim čvorovima x_i, dobijamo približno rešenje u tački x_j

$$v_j = v_{j-k} + h \sum_{i=0}^{n} c_i f_{j-i}, \tag{32}$$

gde je $f_{j-i} = f(x_{j-i}, v_{j-i})$.

Ako je $n = 2$ i $k = 2$, i integral aproksimiran Simpsonovom kvadraturnom formulom sa greškom $O(h^5)$, interpolaciona formula je

$$v_j = v_{j-2} + \frac{h}{3}(f_j + 4f_{j-1} + f_{j-2}). \tag{33}$$

Problem sa interpolacionim formulama je taj što je za njihovu primenu potrebno znati vrednost $f_j = f(x_j, v_j)$ koja se računa pomoću nepoznate veličine v_j. Stoga se obično nekom ekstrapolacionom formulom proceni vrednost v_j – označimo tu procenu sa v_j^*, i u formuli (32) umesto f_j koristi $f(x_j, v_j^*)$. Ekstrapolacione formule se takođe izvode aproksimacijom podintegralne funkcije u izrazu (31) njenim interpolacionim polinomom, s tim što je on određen čvorovima $x_{j-1}, x_{j-2}, \ldots$. Na taj način se pomoću ovog polinoma ekstrapoliše funkcija $f \equiv u'$ na interval (x_{j-1}, x_j).

Ekstrapolaciona formula koja se pridružuje formuli (33) dobija se tako što se u (31) stavi da je $k = 4$, a u' aproksimira Newtonovim interpolacionim polinomom trećeg stepena za interpolaciju unapred, napisanim u odnosu na čvor x_{j-4}. Tada

dobijamo, za $q = (x - x_{j-4})/h$,

$$\begin{aligned} v_j &= v_{j-4} + \int_{x_{j-4}}^{x_j} \big(v'_{j-4} + q\Delta v'_{j-4} + \frac{q(q-1)}{2}\Delta^2 v'_{j-4} + \frac{q(q-1)(q-2)}{6}\Delta^3 v'_{j-4}\big)\, dx \\ &= v_{j-4} + h\int_0^4 \big(v'_{j-4} + q\Delta v'_{j-4} + \frac{q(q-1)}{2}\Delta^2 v'_{j-4} + \frac{q(q-1)(q-2)}{6}\Delta^3 v'_{j-4}\big)\, dq \\ &= v_{j-4} + h\big(\frac{8}{3}v'_{j-1} - \frac{4}{3}v'_{j-2} + \frac{8}{3}v'_{j-3}\big), \end{aligned}$$

gde je $v'_i = f(x_i, v_i) \equiv f_i$. Tražena ekstrapolaciona formula je

$$v_j = v_{j-4} + \frac{4h}{3}(2f_{j-1} - f_{j-2} + 2f_{j-3}). \tag{34}$$

Formule (33) i (34) se obično zajedno koriste i definišu tzv. *metodu Milnea*

$$\begin{aligned} v_j^* &= v_{j-4} + \frac{4h}{3}(2f_{j-1} - f_{j-2} + 2f_{j-3}) \\ v_j &= v_{j-2} + \frac{h}{3}(f_j^* + 4f_{j-1} + f_{j-2}). \end{aligned} \tag{35}$$

Obe formule su reda četiri i njihove greške su određene greškama korišćenih kvadraturnih formula. U praksi je jednostavnije koristiti približnu ocenu greške

$$\frac{1}{29}|v - v^*|.$$

Ako u (31) stavimo $k = 1$ dobijamo *formule Adamsa*. Najčešće korišćena formula ovoga tipa je ona koja se dobija kada se podintegralna funkcija aproksimira Newtonovim polinomom trećeg stepena za interpolaciju unazad, napisanim u odnosu na čvor x_j. Za $q = (x - x_j)/h$, formula oblika (32) je

$$\begin{aligned} v_j &= v_{j-1} + \int_{x_{j-1}}^{x_j} \big(v'_j + q\Delta v'_{j-1} + \frac{q(q+1)}{2}\Delta^2 v'_{j-2} + \frac{q(q+1)(q+2)}{6}\Delta^3 v'_{j-3}\big)\, dx \\ &= v_{j-1} + h\int_{-1}^0 \big(v'_j + q\Delta v'_{j-1} + \frac{q(q+1)}{2}\Delta^2 v'_{j-2} + \frac{q(q+1)(q+2)}{6}\Delta^3 v'_{j-3}\big)\, dq \\ &= v_{j-1} + h\big(v'_j - \frac{1}{2}\Delta v'_{j-1} - \frac{1}{12}\Delta^2 v'_{j-2} - \frac{1}{24}\Delta^3 v'_{j-3}\big). \end{aligned}$$

Ovo je interpolaciona formula, te je potrebno prethodno znati neku procenu v_j^* veličine v_j. Ona se izračunava ekstrapolacionom formulom, koja se dobija tako što se u (31) za $k = 1$ podintegralna funkcija takođe aproksimira Newtonovim polinomom trećeg stepena za interpolaciju unazad, ali napisanim u odnosu na čvor x_{j-1} ($q = (x - x_{j-1})/h$),

$$v_j = v_{j-1} + \int_{x_{j-1}}^{x_j} \big(v'_{j-1} + q\Delta v'_{j-2} + \frac{q(q+1)}{2}\Delta^2 v'_{j-3} + \frac{q(q+1)(q+2)}{6}\Delta^3 v'_{j-4}\big)\, dx$$
$$= v_{j-1} + h\int_0^1 \big(v'_{j-1} + q\Delta v'_{j-2} + \frac{q(q+1)}{2}\Delta^2 v'_{j-3} + \frac{q(q+1)(q+2)}{6}\Delta^3 v'_{j-4}\big)\, dq$$
$$= v_{j-1} + h\big(v'_{j-1} + \frac{1}{2}\Delta v'_{j-2} + \frac{5}{12}\Delta^2 v'_{j-3} + \frac{3}{8}\Delta^3 v'_{j-4}\big).$$

Dakle, Adamsove formule reda četiri su

$$\begin{aligned} v_j^* &= v_{j-1} + h(f_{j-1} + \frac{1}{2}\Delta f_{j-2} + \frac{5}{12}\Delta^2 f_{j-3} + \frac{3}{8}\Delta^3 f_{j-4}) \\ v_j &= v_{j-1} + h(f_j^* - \frac{1}{2}\Delta f_{j-1}^* - \frac{1}{12}\Delta^2 f_{j-2}^* - \frac{1}{24}\Delta^3 f_{j-3}^*), \end{aligned} \tag{36}$$

ili, ako se konačne razlike izraze pomoću vrednosti funkcije u čvorovima,

$$v_j^* = v_{j-1} + \frac{h}{24}(55f_{j-1} - 59f_{j-2} + 37f_{j-3} - 9f_{j-4})$$
$$v_j = v_{j-1} + \frac{h}{24}(9f_j^* + 19f_{j-1} - 5f_{j-2} + f_{j-3}).$$

Greške ovih formula se takođe mogu dobiti integraljenjem grešaka korišćenih interpolacionih polinoma, a približno se greška na jednom koraku ocenjuje veličinom

$$\frac{1}{14}|v - v^*|.$$

Ako je u nekoj tački greška veća od dozvoljene, potrebno je od te tačke računati sa umanjenim korakom.

Problem u primeni ovih višeslojnih metoda je u tome što koriste vrednosti rešenja u više čvorova, te nam nije dovoljna samo početna vrednost $u(x_0) = u_0$ za početak izračunavanja. Na primer, prva vrednost koja se može izračunati Milneovom (35) ili Adamsovom (36) metodom je v_4, dok v_1, v_2 i v_3 treba izračunati nekom drugom metodom. Pri tome, metoda koja se koristi za određivanje tog tzv. početnog odsečka mora biti bar isto toliko tačna koliko i metoda koja se zatim koristi. Stoga se za nalaženje početnog odsečka za Milneovu ili Adamsovu metodu obično koristi metoda Runge–Kutta (25) ili neka od aproksimativnih metoda.

PRIMER 3. Milneovom metodom (35) odredimo na odsečku $[0, 1]$ rešenje Cauchyevog problema

$$xu''(x) + u'(x) + xu(x) = 0, \qquad u(0) = 1, \quad u'(0) = 0,$$

sa tačnošću $\epsilon = 3 \cdot 10^{-4}$.

Da bi metoda mogla da se primeni, potrebno je prethodno transformisati jednačinu u sistem jednačina prvog reda. To se postiže smenom $u_1(x) = xu_0'(x),\ u_0(x) \equiv u(x)$,

$$u_0'(x) = \frac{1}{x}u_1(x), \qquad u_0(0) = 1,$$

$$u_1'(x) = -xu_0(x), \qquad u_1(0) = 0.$$

Početni odsečak se može odrediti, na primer, metodom stepenih redova (§1). Primenom ove metode nalazimo razvoj za funkciju $u_0(x)$,

$$u_0(x) \equiv u(x) = 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{[(2k)!!]^2}\, x^{2k},$$

a njegovim diferenciranjem razvoj za funkciju $u_1(x)$,

$$u_1(x) = \sum_{k=1}^{\infty} \frac{(-1)^k 2k}{[(2k)!!]^2}\, x^{2k-1}.$$

Prva tri člana razvoja za funkciju $u_0(x)$ daju sa tačnošću ϵ vrednosti

$$v_0(0.2) = 0.9900, \qquad v_0(0.4) = 0.9604, \qquad v_0(0.6) = 0.9120,$$

a prva dva člana razvoja za $u_1(x)$ daju sa istom tačnošću

$$v_1(0.2) = -0.0199, \qquad v_1(0.4) = -0.0784, \qquad v_1(0.6) = -0.1719.$$

Sada, primenom formula (35) nalazimo

$$\begin{array}{llll} v_0^*(0.8) = 0.8464, & v_1^*(0.8) = -0.2950 & v_0(0.8) = 0.8464, & v_1(0.8) = -0.2951, \\ v_0^*(1.0) = 0.7652, & v_1^*(1.0) = -0.4400, & v_0(1.0) = 0.7652, & \end{array}$$

te je, konačno, sa tačnošću $3 \cdot 10^{-4}$

$$u(0.8) = 0.846, \qquad u(1.0) = 0.765.$$

8.4 Stabilnost numeričkih algoritama

U uvodnom delu ovog poglavlja je ukazano na probleme koji se mogu javiti pri numeričkom rešavanju loše uslovljenih zadataka. Ponekad, međutim, iako je postavljeni problem korektan i dobro uslovljen, numerička greška koja se javlja na svakom koraku može se nagomilavati, tako da numeričko rešenje znatno odstupa od tačnog. Takvi algoritmi se nazivaju nestabilni algoritmi, i da bi greška približnog rešenja ostala u dozvoljenim granicama korak h mora biti dovoljno mali.

Analizirajmo stabilnost nekih od pomenutih metoda, primenjujući ih na jednostavan modelni problem

$$u'(x) = au(x), \qquad u(0) = u_0, \qquad (a = const), \tag{37}$$

čije je rešenje

$$u(x) = u_0 e^{ax}. \tag{38}$$

Eulerovom metodom (18) približno rešenje v_j u tački $x_j = jh$ je određeno diferencijskom jednačinom

$$v_j = v_{j-1} + h(av_{j-1}) = (1 + ah)v_{j-1},$$

te je, s obzirom na zadati početni uslov,

$$v_j = (1 + ah)^j u_0, \qquad j = 0, 1, \ldots . \tag{39}$$

Ako je $a < 0$, rešenje polaznog problema (38) opada, tj. teži nuli kada $x \to \infty$. Međutim, približno rešenje (39) može neograničeno da raste i osciluje u znaku, ako je korak h takav da je

$$1 + ah < -1, \qquad \text{tj.} \qquad h > \frac{2}{|a|}.$$

Dakle, ako je $a \ll 0$, tj. ako rešenje problema (37) brzo opada, da bi numeričko rešenje bilo stabilno neophodno je izabrati korak h dovoljno mali.

Modifikacija (23) Eulerove metode za problem (37) ima u tački x_j oblik

$$v_j = v_{j-1} + \frac{ah}{2}(v_{j-1} + v_j),$$

i numeričko rešenje je

$$v_j = \frac{1 + \frac{ah}{2}}{1 - \frac{ah}{2}} v_{j-1} = \left(\frac{1 + \frac{ah}{2}}{1 - \frac{ah}{2}}\right)^j u_0, \qquad j = 0, 1, \ldots .$$

Kada je $a < 0$ približno rešenje v_j opada i teži nuli kada $j \to \infty$, kao i tačno rešenje (38). Bez obzira na veličinu koraka h približno rešenje se ponaša kao i tačno rešenje, što znači da je ova metoda stabilna.

I vrlo popularna, zbog visoke tačnosti i jednostavnosti, metoda Runge–Kutta (25) pokazuje znake nestabilnosti pri primeni na probleme sa brzo opadajućim rešenjima, ako se ne izabere odgovarajući korak h. Pokažimo to na našem modelnom primeru. Za problem (37) metoda Runge–Kutta je data formulama

$$k_1 = ahv_{j-1}, \qquad k_2 = ah\big(1 + \frac{ah}{2}\big)v_{j-1}, \qquad k_3 = ah\big(1 + \frac{ah}{2} + \frac{a^2h^2}{4}\big)v_{j-1},$$

$$k_4 = ah\big(1 + ah + \frac{a^2h^2}{2} + \frac{a^3h^3}{4}\big)v_{j-1},$$

$$v_j = \big(1 + ah + \frac{a^2h^2}{2} + \frac{a^3h^3}{6} + \frac{a^4h^4}{24}\big)v_{j-1},$$

pa je približno rešenje u tački x_j

$$v_j = \left(1 + ah + \frac{a^2h^2}{2} + \frac{a^3h^3}{6} + \frac{a^4h^4}{24}\right)^j u_0.$$

Ako je $a = -100$, a $h = 0.03$, onda je $v_j = (\frac{11}{8})^j u_0$, te je za $u_0 = 1$ približno rešenje u tački $x = 3$ $v_{100} = 6.8 * 10^{13}$. Ako uzmemo da je korak $h = 0.02$ dobija se za približno rešenje u istoj tački $v_{150} = 2.7 * 10^{-72}$, što odgovara ponašanju tačnog rešenja.

Stoga, konzistentnost metode nije dovoljna garancija da će njom određeno približno rešenje biti zadovoljavajuće tačnosti. Kod tzv. uslovno stabilnih metoda tačnost će biti postignuta samo ako korak zadovoljava određeni uslov.

Zbog pomenute nestabilnosti, posebni problemi se javljaju u rešavanju tzv. *krutih sistema*. Sistem običnih diferencijalnih jednačina je krut sistem ako je količnik najveće i najmanje sopstvene vrednosti jakobijana $\{\frac{\partial f_k}{\partial u_j}\}$ veliki po apsolutnoj vrednosti.

PRIMER 4.

$$\begin{aligned} u_1'(x) &= -u_1(x) + u_2(x), & u_1(0) &= 0 \\ u_2'(x) &= -100u_2(x), & u_2(0) &= 99 \end{aligned}$$

Rešenje ovog sistema je $u_1(x) = e^{-x} - e^{-100x}$, $u_2(x) = 99e^{-100x}$. Iako komponenta rešenja e^{-100x} za $x > 0$ malo utiče na rešenje jer brzo teži nuli, da bi napr. Eulerov algoritam bio stabilan, zbog prisustva ove komponente korak ne sme biti veći od $h = 0.02$. Slično ograničenje bi se pojavilo i pri primeni metode Runge–Kutta.

Da bi se izbegle ove neprijatnost, za rešavanje krutih sistema treba koristiti bezuslovno stabilne metode.

9

Obične diferencijalne jednačine – granični problemi

Granični problem za obične diferencijalne jednačine je problem nalaženja partikularnog rešenja jednačine

$$u^{(m)}(x) = f(x, u, u', \ldots, u^{(m-1)}),$$

koje zadovoljava uslove zadate u više od jedne tačke intervala. Stoga se granični problemi ne mogu definisati za jednačine prvog reda. Prvobitno su to bili problemi kod kojih su uslovi definisani samo na krajevima intervala, te otuda potiče naziv.

PRIMER 1. Rešenjem graničnog problema

$$\begin{gathered} -u''(x) = f(x), \qquad a \le x \le b, \\ u(a) = u(b) = 0 \end{gathered}$$

se može opisati oblik zategnute žice učvršćene na krajevima, kada na nju deluje spoljašnja sila $f(x)$.

Za jednačine ili sisteme jednačina višeg reda, gde je broj zadatih uslova veći pa oni mogu biti zadati i u unutrašnjim tačkama intervala, granični problemi su raznovrsniji.

PRIMER 2. Granični problem

$$\begin{gathered} u^{(4)}(x) = f(x), \qquad a \le x \le b, \\ u(x_i) = 0, \quad i = 1, 2, 3, 4, \qquad a \le x_1 < x_2 < x_3 < x_4 \le b, \end{gathered}$$

opisuje deformaciju grede pod uticajem spoljašnje sile $f(x)$, ako se u četiri tačke x_i greda oslanja na nosače.

Kod Cauchyevih problema integralna kriva je potpuno određena uslovima zadatim u jednoj tački. Stoga, krećući se od te tačke u pravcu određenom jednačinom, približno rekonstruišemo integralnu krivu, tj. računamo približne vrednosti rešenja

u drugim tačkama intervala. Kod graničnih problema, uslovima zadatim u početnoj tački rešenje nije jednoznačno određeno. Iz familije integralnih krivih koje zadovoljavaju dati početni uslov treba odabrati onu koja će proći kroz ostale tačke, odnosno zadovoljiti uslove zadate u ostalim tačkama. Stoga su granični problemi mnogo složeniji od Cauchyevih.

Bez obzira na raznovrsnost graničnih uslova, svi granični problemi se rešavaju u osnovi istim metodama. Možemo ih podeliti u tri osnovne grupe: metode gađanja, metode konačnih razlika i varijacione metode. Pri tome, u poređenju sa Cauchyevim problemima, složenija su pitanja egzistencije i jedinosti rešenja, javlja se potreba za rešavanjem sistema linearnih ili nelinearnih jednačina, itd.

Metode za rešavanje graničnih problema ćemo ilustrovati na primeru prvog i trećeg graničnog problema za linearne diferencijalne jednačine drugog reda. Ovi se uvek, odgovarajućim smenama, mogu svesti na sledeći granični problem sa homogenim graničnim uslovima

$$-u''(x) + q(x)u(x) = f(x), \qquad 0 \le x \le 1, \tag{1}$$

$$\alpha_1 u'(0) + \beta_1 u(0) = 0, \qquad \alpha_2 u'(1) + \beta_2 u(1) = 0, \tag{2}$$

koji može da se zapiše u obliku operatorske jednačine

$$Lu = f. \tag{3}$$

Ako je $\alpha_i = 0$ i $\beta_i \neq 0$, $i = 1, 2$, granični uslovi (2) su prvi ili Dirichletovi granični uslovi

$$u(0) = u(1) = 0, \tag{4}$$

a granični problem (1),(4) se naziva prvi ili *Dirichletov granični problem*.

Ako je $\alpha_i \neq 0$ i $\beta_i = 0$, $i = 1, 2$, granični uslovi (2) su drugi ili Neumannovi granični uslovi

$$u'(0) = u'(1) = 0. \tag{5}$$

Granični problem (1),(5) se naziva drugi ili *Neumannov granični problem*.

Ako je i $\alpha_i \neq 0$ i $\beta_i \neq 0$, $i = 1, 2$, granični uslovi (2) su treći ili mešoviti granični uslovi, i mogu se zapisati u sledećem obliku

$$u'(0) - \sigma_1 u(0) = 0, \qquad u'(1) + \sigma_2 u(1) = 0. \tag{6}$$

Granični problem (1),(6) se naziva treći ili *mešoviti granični problem*.

Iz teorije diferencijalnih jednačina poznati su sledeći rezultati

TEOREMA 1. *Ako su funkcije $q(x)$ i $f(x)$ neprekidne na intervalu $[0, 1]$ i $q(x) \ge 0$, tada prvi granični problem (1),(4) ima jedinstveno rešenje $u(x) \in \mathcal{C}^2[0, 1]$.*

TEOREMA 2. *Ako su funkcije $q(x)$ i $f(x)$ neprekidne na intervalu $[0,1]$, $q(x) \geq 0$ i $\sigma_1 > 0$, $\sigma_2 > 0$, tada treći granični problem (1),(6) ima jedinstveno rešenje $u(x) \in \mathcal{C}^2[0,1]$.*

Što se tiče drugog graničnog problema (1),(5), u važnom specijalnom slučaju kada je $q(x) \equiv 0$, njegovo rešenje nije jedinstveno. Naime, ako je $u(x)$ rešenje tog problema, onda je i svaka funkcija oblika $u(x) + c$, gde je c proizvoljna konstanta, takođe rešenje. Stoga ćemo se baviti samo numeričkim rešavanjem prvog i trećeg graničnog problema.

9.1 Metode gađanja

Metodama ovoga tipa se granični problemi svode na Cauchyeve. Radi jednostavnosti, prikažimo osnovnu ideju ovih metoda na primeru graničnog problema definisanog jednačinom drugog reda

$$u''(x) = f(x, u, u'), \qquad u(a) = A, \quad u(b) = B. \tag{7}$$

Datom jednačinom i graničnim uslovom zadatim u levom kraju $x = a$ definišimo sledeći Cauchyev problem

$$v''(x) = f(x, v, v'), \qquad v(a) = A, \quad v'(a) = s. \tag{8}$$

Njegovo rešenje zavisi od izbora parametra s, $v(x) \equiv v(x; s)$. Da bi rešenja problema (7) i (8) bila identična, treba izabrati $s = \bar{s}$ tako da je $v(b; \bar{s}) = B$, tj. tako da je $\bar{s}$ rešenje jednačine

$$F(s) \equiv v(b; s) - B = 0. \tag{9}$$

Ako se za rešavanje jednačine (9) koristi metoda bisekcije (§7.3), izaberu se početne vrednosti s_0 i s_1 tako da rešenja odgovarajućih Cauchyevih problema (8) zadovoljavaju uslov $F(s_0)F(s_1) < 0$. Zatim se za $s = s_2 = \frac{1}{2}(s_0 + s_1)$ rešava Cauchyev problem (8), i za s_3 uzima sredina onog od intervala $[s_0, s_2]$ ili $[s_2, s_1]$ na čijim krajevima funkcija $F(s)$ ima različiti znak, itd. Metoda bisekcije sporo konvergira, te je potrebno rešavati veliki broj Cauchyevih problema (8), što je značajan nedostatak ove metode. Radi ubrzanja konvergencije, može se koristiti i Newtonova metoda,

$$s_{n+1} = s_n - \frac{F(s_n)}{F'(s_n)},$$

s obzirom da je v, a time i F neprekidna funkcija po s. U svakom slučaju, s se određuje iterativnim algoritmom što zahteva rešavanje velikog broja Cauchyevih problema.

Za linearne granične probleme metoda je mnogo jednostavnija, jer se koristi činjenica da je rešenje nehomogenog linearnog problema jednako zbiru ma kog njegovog partikularnog rešenja i rešenja odgovarajućeg homogenog problema.

Za Dirichletov granični problem (1),(4) rešenje predstavimo u obliku

$$u(x) = u_1(x) + cu_2(x), \tag{10}$$

gde je $u_1(x)$ neko partikularno rešenje jednačine (1), $u_2(x)$ partikularno rešenje homogene jednačine

$$-u''(x) + q(x)u(x) = 0,$$

a c konstanta koju ćemo kasnije pogodno izabrati. Da bi funkcija $u(x)$ zadovoljavala granični uslov u tački $x = 0$ za svako c, treba da je

$$u_1(0) = 0, \qquad u_2(0) = 0. \tag{11}$$

Uslovima (11) pridružimo još dva početna uslova

$$u_1'(0) = A_1, \qquad u_2'(0) = A_2,$$

pri čemu su A_1 i A_2, $A_2 \neq 0$, proizvoljne konstante. Na taj način smo polazni granični problem sveli na dva Cauchyeva problema

$$\begin{aligned} -u_1''(x) + q(x)u_1(x) &= f(x) & \qquad -u_2''(x) + q(x)u_2(x) &= 0 \\ u_1(0) = 0, \quad u_1'(0) &= A_1 & \qquad u_2(0) = 0, \quad u_2'(0) &= A_2. \end{aligned}$$

Funkcija $u(x)$ zadovoljava jednačinu (1) i levi granični uslov za svako c. Ovaj parametar odredimo tako da ona zadovoljava i desni granični uslov,

$$u(1) = u_1(1) + cu_2(1) = 0,$$

što znači da treba da je

$$c = -\frac{u_1(1)}{u_2(1)}.$$

Slično se rešava i mešoviti problem (1),(6). Da bi rešenje oblika (10) zadovoljavalo za svako c granični uslov u levom kraju intervala,

$$u'(0) - \sigma_1 u(0) = u_1'(0) - \sigma_1 u_1(0) + c\big(u_2'(0) - \sigma_1 u_2(0)\big) = 0,$$

treba da je

$$u_1'(0) = \sigma_1 u_1(0), \quad \text{ i } \quad u_2'(0) = \sigma_1 u_2(0).$$

Ako zadamo da je

$$u_1(0) = A_1, \qquad \text{i} \qquad u_2(0) = A_2,$$

gde su A_1 i A_2, $A_2 \neq 0$, proizvoljne konstante, granični problem (1),(6) možemo svesti na dva Cauchyeva problema

$$\begin{aligned} -u_1''(x) + q(x)u_1(x) &= f(x) & -u_2''(x) + q(x)u_2(x) &= 0 \\ u_1(0) = A_1, \quad u_1'(0) &= \sigma_1 A_1, & u_2(0) = A_2, \quad u_2'(0) &= \sigma_1 A_2. \end{aligned}$$

Konstantu c odredimo tako da rešenje (10) graničnog problema zadovolji granični uslov i u desnom kraju intervala,

$$u'(1) + \sigma_2 u(1) = u_1'(1) + \sigma_2 u_1(1) + c\big(u_2'(1) + \sigma_2 u_2(1)\big) = 0,$$

a to znači da je

$$c = -\frac{u_1'(1) + \sigma_2 u_1(1)}{u_2'(1) + \sigma_2 u_2(1)}.$$

Cauchyevi problemi na koje smo metodama gađanja sveli granične probleme, rešavaju se nekom od numeričkih metoda o kojima je bilo reči u poglavlju 8.

9.2 Metode konačnih razlika

Ove metode se zasnivaju na zameni izvoda količnicima konačnih razlika. Prvo se izabere konačno mnogo tačaka intervala $[0, 1]$, i one čine *mrežu*. Izabrane tačke nazivaju se *čvorovi mreže*. Ako su čvorovi ravnomerno raspoređeni, kažemo da je mreža ravnomerna i definisana je *korakom* – rastojanjem između dva susedna čvora,

$$\bar{\omega}_h = \{\, x_i \,|\, x_i = ih,\ i = 0, \ldots, n,\ h = \tfrac{1}{n} \,\}.$$

Ako rastojanje među čvorovima mreže nije konstantno, mreža je neravnomerna. Na ravnomernoj mreži moguće aproksimacije izvoda funkcije u tački x_i su

$$\begin{aligned} u_{x,i} &= \frac{1}{h}\big(u(x_{i+1}) - u(x_i)\big), \qquad u_{\bar{x},i} = \frac{1}{h}\big(u(x_i) - u(x_{i-1})\big), \\ u_{\dot{x},i} &= \frac{1}{2h}\big(u(x_{i+1}) - u(x_{i-1})\big) = \frac{1}{2}(u_{x,i} + u_{\bar{x},i}) \qquad \text{za } u', \\ u_{\bar{x}x,i} &= \frac{1}{h^2}\big(u(x_{i+1}) - 2u(x_i) + u(x_{i-1})\big) = \frac{1}{h}(u_{x,i} - u_{\bar{x},i}), \qquad \text{za } u''. \end{aligned} \tag{12}$$

Pod pretpostavkom da je funkcija $u(x)$ dovoljno glatka, razvojem u Taylorov red možemo oceniti grešku ovih aproksimacija – na primer,

$$\begin{aligned} u'(x_i) - u_{x,i} &= u'(x_i) - \frac{1}{h}\big(u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i + \theta h) - u(x_i)\big) \\ &= -\frac{h}{2}u''(x_i + \theta h), \qquad 0 \leq \theta \leq 1 \end{aligned}$$

Tako dobijamo ocene

$$u'(x_i) = u_{x,i} + O(h) = u_{\bar{x},i} + O(h) = u_{\dot{x},i} + O(h^2)$$
$$u''(x_i) = u_{\bar{x}x,i} + O(h^2). \tag{13}$$

Kada $h \to 0$, tj. kada se mreža zgušnjava, aproksimacije teže vrednostima izvoda funkcije $u(x)$ u čvorovima. Pri tome, konvergencija je brža kod aproksimacija centralnim količnicima konačnih razlika $u_{\dot{x}}$ i $u_{\bar{x}x}$, jer je glavni član greške $O(h^2)$.

Kod neravnomernih mreža izvodi se aproksimiraju podeljenim razlikama.

Zamenom funkcije $u(x)$ i njenih izvoda u graničnom problemu (1),(2) odgovarajućim količnicima konačnih razlika (12) u čvorovima mreže $\bar{\omega}_h$, vršimo diskretizaciju polaznog problema. Kontinualnu veličinu $u(x)$ zamenjujemo vektorom $\mathbf{v} = (v_0, \ldots, v_n)^T$, pri čemu je $v_i \approx u(x_i)$, a granični problem (1),(2) sistemom linearnih jednačina po $\mathbf{v}$,

$$-v_{\bar{x}x,i} + q_i v_i = f_i, \quad i = 1, \ldots, n-1,$$
$$\alpha_1 v_{x,0} + \beta_1 v_0 = 0, \tag{14}$$
$$\alpha_2 v_{\bar{x},n} + \beta_2 v_n = 0,$$

gde je $q_i = q(x_i)$ i $f_i = f(x_i)$. Diskretni problem (14), koji opštije možemo zapisati operatorskom jednačinom

$$L_h v = f_h, \tag{15}$$

naziva se *diferencijskom šemom*. Diskretizacija je dobro izvršena ukoliko, pre svega, diskretni problem (15) ima jedinstveno rešenje – treba, dakle, dokazati egzistenciju i jedinost rešenja problema (15). Dalje, neophodno je da rešenje diskretnog problema konvergira ka rešenju polaznog problema kada $h \to 0$, tj. da greška $\epsilon = u - v$ teži nuli kada $h \to 0$. Ovo će biti ispunjeno ukoliko je šema konzistentna i stabilna. Konzistentnost šeme (15) u odnosu na problem (3) znači da

$$L_h u \to Lu, \qquad f_h \to f \qquad \text{kada} \quad h \to 0,$$

što za šemu (14) sledi iz (13). Stabilnost šeme (15) se svodi na stabilnost sistema linearnih jednačina, koja je detaljnije analizirana u poglavlju 5. Kratko rečeno, šema (15) će biti stabilna ako je L_h^{-1} uniformno ograničeni operator. Dakle, ako je šema (15) konzistentna i stabilna, konvergencija neposredno sledi jer je

$$L_h(u - v) = L_h u - L_h v = L_h u - f_h + f - Lu = (L_h u - Lu) + (f - f_h),$$

te

$$u - v = L_h^{-1}(L_h u - Lu) + L_h^{-1}(f - f_h) \to 0 \qquad \text{kada} \quad h \to 0.$$

Pitanja egzistencije i jedinosti rešenja i konvergencije diferencijske šeme moraju se analizirati posebno za svaku šemu.

Prvi granični problem. U graničnim uslovima (2) je $\alpha_i = 0$, $\beta_i = 1$, $i = 1, 2$, te je diferencijska šema (14)

$$\begin{aligned} -v_{\bar{x}x,i} + q_i v_i &= f_i, \quad i = 1, \ldots, n-1, \\ v_0 &= v_n = 0 \end{aligned} \tag{16}$$

U operatorskom zapisu (15) L_h je linearni operator

$$L_h v_i = \begin{cases} -v_{\bar{x}x,i} + q_i v_i, & \text{za } i = 1, \ldots, n-1 \\ 0, & \text{za } i = 0 \text{ ili } i = n, \end{cases} \tag{17}$$

koji preslikava vektorski prostor $\mathcal{V} = \{\mathbf{v} = (0, v_1, \ldots, v_{n-1}, 0)^T \,|v_i \in \mathcal{R}^1\}$ u samog sebe, i $f_h = (0, f_1, \ldots, f_{n-1}, 0)^T \in \mathcal{V}$. U vektorskom prostoru $\mathcal{V}$ definišimo skalarni proizvod

$$(\mathbf{v}\,,\,\mathbf{w})_h = h \sum_{i=1}^{n-1} v_i w_i$$

i norme

$$\|\mathbf{v}\|_h = (\mathbf{v}\,,\,\mathbf{v})_h^{1/2}, \qquad \|\mathbf{v}\|_{C_h} = \max_{1 \le i \le n-1} |v_i|.$$

U odnosu na ovako definisan skalarni proizvod, linearni operator L_h je samokonjugovan i pozitivno definisan.

Samokonjugovanost sledi iz simetričnosti bilinearne forme

$$(L_h \mathbf{v}\,,\,\mathbf{w})_h = h \sum_{i=1}^{n-1} (-v_{\bar{x}x,i} + q_i v_i) w_i = h \sum_{i=0}^{n-1} v_{x,i} w_{x,i} + h \sum_{i=1}^{n} q_i v_i w_i.$$

Odavde, pošto je $q(x) \ge 0$, takođe sledi da je

$$(L_h \mathbf{v}\,,\,\mathbf{v})_h \ge h \sum_{i=0}^{n-1} v_{x,i}^2. \tag{18}$$

Da bismo dokazali pozitivnu definisanost ovog operatora, ocenimo veličinu v_i. Možemo je, uzimajući u obzir da je $v_0 = v_n = 0$, izraziti na sledeći način:

$$\begin{aligned} v_i^2 &= \left(\sum_{j=1}^{i} v_j - \sum_{j=0}^{i-1} v_j \right)^2 + ih \left[\left(\sum_{j=i+1}^{n} v_j - \sum_{j=i}^{n-1} v_j \right)^2 - \left(\sum_{j=1}^{i} v_j - \sum_{j=0}^{i-1} v_j \right)^2 \right] \\ &= (1 - ih) \left(h \sum_{j=0}^{i-1} v_{x,j} \right)^2 + ih \left(h \sum_{j=i}^{n-1} v_{x,j} \right)^2. \end{aligned}$$

Primenom nejednakosti Cauchy–Schwarza na poslednje dve sume, imamo da je

$$v_i^2 \le (1 - ih) ih^2 \sum_{j=0}^{i-1} v_{x,j}^2 + ih(n - i) h^2 \sum_{j=i}^{n-1} v_{x,j}^2 = ih(1 - ih) \left(h \sum_{j=0}^{n-1} v_{x,j}^2 \right),$$

jer je $(n-i)h = nh - ih = 1 - ih$. Stoga je, na osnovu (18),

$$\|\mathbf{v}\|_{C_h}^2 = \max_{1\le i\le n-1} v_i^2 \le \left(h\sum_{j=0}^{n-1} v_{x,j}^2\right) \max_{1\le i\le n-1} ih(1-ih)$$
$$\le \frac{1}{4}\left(h\sum_{j=0}^{n-1} v_{x,j}^2\right) \le \frac{1}{4}\,(L_h\mathbf{v}\,,\,\mathbf{v})_h.$$

Kako je još

$$\|\mathbf{v}\|_h^2 = h\sum_{i=1}^{n-1} v_i^2 \le h(n-1)\max_{1\le i\le n-1} v_i^2 \le \|\mathbf{v}\|_{C_h}^2,$$

konačno dobijamo da je

$$\|\mathbf{v}\|_h^2 \le \|\mathbf{v}\|_{C_h}^2 \le \frac{1}{4}(L_h\mathbf{v}\,,\,\mathbf{v})_h, \tag{19}$$

što dokazuje pozitivnu definisanost operatora L_h.

Stoga jednačina (15), tj. šema (16), ima rešenje i ono je jedinstveno.

Neposredna posledica nejednakosti (19),

$$4\|\mathbf{v}\|_{C_h}^2 \le (L_h\mathbf{v}\,,\,\mathbf{v})_h \le \|L_h\mathbf{v}\|_h\,\|\mathbf{v}\|_h \le \|L_h\mathbf{v}\|_h\|\mathbf{v}\|_{C_h},$$

su ocene

$$\|\mathbf{v}\|_h \le \|\mathbf{v}\|_{C_h} \le \frac{1}{4}\|L_h\mathbf{v}\|_h. \tag{20}$$

Da bismo dokazali konvergenciju šeme (16), uočimo da greška $\epsilon = u - \mathbf{v}$, koja je kao i $\mathbf{v}$ definisana samo u čvorovima mreže, zadovoljava diferencijsku šemu sa istim operatorom L_h, datim izrazom (17), samo sa promenjenom desnom stranom. Naime, kako je

$$-\epsilon_{\bar{x}x,i} + q_i\epsilon_i = -(u_{\bar{x}x,i} - v_{\bar{x}x,i}) + q_i(u(x_i) - v_i) = -u_{\bar{x}x,i} + q_iu(x_i) - f_i$$
$$= -u_{\bar{x}x,i} + q_iu(x_i) - (-u''(x_i) + q_iu(x_i)) = u''(x_i) - u_{\bar{x}x,i},$$

to je ϵ rešenje diferencijske šeme

$$-\epsilon_{\bar{x}x,i} + q_i\epsilon_i = u''(x_i) - u_{\bar{x}x,i}, \qquad i = 1,\dots,n-1,$$
$$\epsilon_0 = \epsilon_n = 0,$$

pa je, prema (20) i (13),

$$\|\epsilon\|_h \le \|\epsilon\|_{C_h} \le \frac{1}{4}\|u'' - u_{\bar{x}x}\|_h = O(h^2).$$

Prema tome, šema (16) je konvergentna, jer je

$$\|\epsilon\|_h \le \max_{1\le i\le n-1} |u(x_i) - v_i| \to 0, \quad \text{kada} \quad h \to 0.$$

Princip maksimuma. Nesingularnost matrice sistema (16), pa stoga i egzistencija jedinstvenog rešenja ovog sistema, sledi i iz diskretnog principa maksimuma. Ovim principom se dokazuje da homogeni sistem sa trodijagonalnom matricom čiji elementi zadovoljavaju određene pretpostavke, ima samo trivijalno rešenje.

TEOREMA 3. *Neka je*

$$\Lambda v_i = -a_i v_{i-1} + c_i v_i - b_i v_{i+1}, \qquad i = 1, \ldots, n-1,$$
$$a_i > 0, \qquad b_i > 0, \qquad c_i \geq a_i + b_i.$$

Ako je $\Lambda v_i \leq 0$ $(\Lambda v_i \geq 0)$ *za* $i = 1, \ldots, n-1$, *tada funkcija* v_i, *različita od konstante, ne može dostići najveću pozitivnu (najmanju negativnu) vrednost u tačkama* $i = 1, \ldots, n-1$.

DOKAZ: Pretpostavimo da je $\Lambda v_i \leq 0$, $i = 1, \ldots, n-1$, i da u nekoj od unutrašnjih tačaka funkcija v_i dostiže pozitivni maksimum. Tada, pošto je $v_i \neq const$, postoji tačka $i_0 \in \{1, \ldots, n-1\}$ takva da je

$$v_{i_0} = \max_{0 \leq i \leq n} v_i = M_0 > 0,$$

a da u jednoj od njoj susednih tačaka, na primer $i = i_0 - 1$, važi stroga nejednakost $v_{i_0-1} < M_0$. Operator Λ možemo zapisati u sledećem obliku

$$\Lambda v_i = a_i(v_i - v_{i-1}) + b_i(v_i - v_{i+1}) + (c_i - b_i - a_i)v_i.$$

U tački $i = i_0$ su drugi i treći sabirak nenegativni, te je

$$\Lambda v_{i_0} \geq a_{i_0}(v_{i_0} - v_{i_0-1}) > 0,$$

što je suprotno pretpostavci. Dualno tvrđenje se dokazuje analogno. ∎

POSLEDICA 1. Ako je $\Lambda v_i \geq 0$, $i = 1, \ldots, n-1$, $v_0 \geq 0$, $v_n \geq 0$, tada je funkcija v_i nenegativna, $v_i \geq 0$, $i = 1, \ldots, n-1$. Ako je $\Lambda v_i \leq 0$, $i = 1, \ldots, n-1$, $v_0 \leq 0$, $v_n \leq 0$, onda je $v_i \leq 0$, $i = 1, \ldots, n-1$.

POSLEDICA 2. Jedinstveno rešenje homogenog problema

$$\Lambda v_i = 0, \quad i = 1, \ldots, n-1, \qquad v_o = v_n = 0$$

je $v_i = 0$, $i = 0, \ldots, n$, te nehomogeni problem

$$\Lambda v_i = F_i, \quad i = 1, \ldots, n-1, \qquad v_0 = A, \quad v_n = B, \tag{21}$$

ima rešenje za svako F_i, $i = 1, \ldots, n-1$, A i B.

Diferencijska šema (16) se svodi na (21) za $a_i = b_i = \frac{1}{h^2}$, $c_i = \frac{2}{h^2} + q_i$, $F_i = f_i$, $i = 1, \ldots, n-1$ i $A = B = 0$, te, prema posledici 2, ima jedinstveno rešenje.

Konvergencija diferencijske šeme (16) se takođe može dokazati koristeći neke posledice principa maksimuma.

Treći granični problem. Treći granični problem (1),(6)

$$\begin{aligned} -u''(x) + q(x)u(x) &= f(x), \qquad 0 \le x \le 1, \\ u'(0) - \sigma_1 u(0) &= 0, \qquad u'(1) + \sigma_2 u(1) = 0, \end{aligned}$$

s obzirom na (14), aproksimira se diferencijskom šemom

$$\begin{aligned} -v_{\bar{x}x,i} + q_i v_i &= f_i, \quad i = 1, \dots, n-1, \\ v_{x,0} - \sigma_1 v_0 &= 0, \qquad v_{\bar{x},n} + \sigma_2 v_n = 0. \end{aligned}$$

Aproksimacije graničnih uslova, prema (13), imaju grešku $O(h)$, što nepotrebno usporava konvergenciju diferencijske šeme, jer je greška aproksimacije jednačine $O(h^2)$. Granične uslove (6) moguće je aproksimirati takođe sa greškom $O(h^2)$, ukoliko se koristi i jednačina (1). Naime, pođimo od Taylorovog razvoja

$$\begin{aligned} u_{x,0} &= \frac{1}{h}\big(u(h) - u(0)\big) = \frac{1}{h}\big(u(0) + hu'(0) + \frac{h^2}{2}u''(0) + O(h^3) - u(0)\big) \\ &= u'(0) + \frac{h}{2}u''(0) + O(h^2), \end{aligned}$$

u kome $u'(0)$ zamenimo izrazom dobijenim iz prvog graničnog uslova

$$u'(0) = \sigma_1 u(0),$$

a $u''(0)$ izrazom dobijenim iz diferencijalne jednačine

$$u''(0) = q(0)u(0) - f(0).$$

Tako dobijamo da je

$$u_{x,0} = \sigma_1 u(0) + \frac{h}{2}\big(q(0)u(0) - f(0)\big) + O(h^2),$$

pa sa greškom $O(h^2)$ granični uslov u levom kraju intervala aproksimiramo izrazom

$$-v_{x,0} + (\sigma_1 + \frac{h}{2}q_0)v_0 = \frac{h}{2}f_0.$$

Slično se dobija aproksimacija graničnog uslova u desnom kraju intervala

$$v_{\bar{x},n} + (\sigma_2 + \frac{h}{2}q_n)v_n = \frac{h}{2}f_n.$$

Diferencijska šema kojom je sa greškom $O(h^2)$ aproksimiran treći granični problem (1),(6) je

$$\begin{aligned} -v_{\bar{x}x,i} + q_i v_i &= f_i, \quad i = 1, \dots, n-1, \\ -v_{x,0} + (\sigma_1 + \frac{h}{2}q_0)v_0 = \frac{h}{2}f_0 \qquad & v_{\bar{x},n} + (\sigma_2 + \frac{h}{2}q_n)v_n = \frac{h}{2}f_n. \end{aligned} \tag{22}$$

Egzistencija i jedinost rešenja šeme (22) se dokazuju slično kao za prvi granični problem. U operatorskom zapisu (15) šeme (22) linearni operator

$$L_h v_i = \begin{cases} \frac{2}{h}(-v_{x,0} + \sigma_1 v_0) + q_0 v_0, & \text{za } i = 0 \\ -v_{\bar{x}x,i} + q_i v_i, & \text{za } i = 1, \dots, n-1 \\ \frac{2}{h}(v_{\bar{x},n} + \sigma_2 v_n) + q_n v_n, & \text{za } i = n, \end{cases}$$

preslikava vektorski prostor $\mathcal{V} = \{\mathbf{v} = (v_0, \dots, v_n)^T \,|v_i \in \mathcal{R}^1\}$ u samog sebe, i $f_h = (f_0, \dots, f_n)^T \in \mathcal{V}$. U prostoru $\mathcal{V}$ definišimo skalarni proizvod

$$[\mathbf{v}\,,\, \mathbf{w}]_h = h \sum_{i=1}^{n-1} v_i w_i + \frac{h}{2}(v_0 w_0 + v_n w_n)$$

i norme

$$|[\mathbf{v}]|_h = [\mathbf{v}\,,\, \mathbf{v}]_h^{1/2}, \qquad |[\mathbf{v}]|_{C_h} = \max_{0 \le i \le n} |v_i|.$$

Slično dokazu koji je izveden za prvi granični problem, dokazuje se da je linearni operator L_h samokonjugovan i pozitivno definisan, te jednačina (15), tj. šema (22), ima jedinstveno rešenje. Šema je konvergentna, i važi ocena

$$|[u - v]|_{C_h} = \max_{0 \le i \le n} |u(x_i) - v_i| = O(h^2).$$

Što se tiče tehnike rešavanja diferencijskih problema (14), s obzirom da su to sistemi linearnih jednačina sa trodijagonalnim matricama, oni se efikasno rešavaju Gaussovom metodom eliminacije (§5.2).

Diferencijske šeme povišene tačnosti. Da bismo konstruisali šeme povišene tačnosti, neophodno je pre svega odrediti tačniju aproksimaciju za $u''(x)$. Ukoliko rešenje graničnog problema $u(x) \in \mathcal{C}^6[0,1]$, iz Taylorovog razvoja je

$$\begin{aligned} u_{\bar{x}x}(x) &= u''(x) + \frac{h^2}{12} u^{(4)}(x) + O(h^4), \\ u_{\bar{x}x\bar{x}x}(x) &= \frac{1}{h^2}\big(u_{\bar{x}x}(x+h) - 2u_{\bar{x}x}(x) + u_{\bar{x}x}(x-h)\big) = u^{(4)}(x) + O(h^2), \end{aligned} \tag{23}$$

te je u čvorovima x_i, $i = 2, \dots, n-2$,

$$u''(x_i) = u_{\bar{x}x,i} - \frac{h^2}{12}(u_{\bar{x}x\bar{x}x,i} + O(h^2)) + O(h^4) = u_{\bar{x}x,i} - \frac{h^2}{12} u_{\bar{x}x\bar{x}x,i} + O(h^4).$$

Zanemarivanjem $O(h^4)$ dobijamo aproksimaciju četvrtog reda tačnosti za $u''(x)$.

Radi jednostavnosti, pretpostavimo da rešavamo prvi granični problem. Tada se ne javlja problem aproksimiranja graničnih uslova, ali, pošto se šema povišene tačnosti ne može koristiti u čvorovima x_1 i x_{n-1}, potrebno je izvesti posebne aproksimacije iste tačnosti i u ovim čvorovima. Da bi se odredila aproksimacija za $u''(x_1)$, formira se linearna kombinacija $\sum_{i=0}^{5} c_i u(x_i)$, razviju u Taylorov red

oko tačke x_1 veličine $u(x_i)$ i koeficijenti c_i odrede tako da se anuliraju izrazi koji množe h^k, $k = 0, 1, 3, 4, 5$. Sličnim kombinovanjem vrednosti $u(x_i)$, $i = n-5, \ldots, n$, dobija se aproksimacija povišene tačnosti za $u''(x_{n-1})$. Tako je diferencijska šema koja sa greškom $O(h^4)$ aproksimira prvi granični problem

$$\begin{gathered} v_0 = 0 \\ -\frac{1}{12h^2}(10v_0 - 15v_1 - 4v_2 + 14v_3 - 6v_4 + v_5) + q_1 v_1 = f_1 \\ -v_{\bar{x}x,i} + \frac{h^2}{12} v_{\bar{x}x\bar{x}x,i} + q_i v_i = f_i, \quad i = 2, \ldots, n-2 \\ -\frac{1}{12h^2}(10v_n - 15v_{n-1} - 4v_{n-2} + 14v_{n-3} - 6v_{n-4} + v_{n-5}) + q_{n-1} v_{n-1} = f_{n-1} \\ v_n = 0. \end{gathered}$$

Operator L_h u ovom slučaju nije samokonjugovan.

Ako je u jednačini (1) $q(x) \equiv 0$, diferenciranjem jednačine dva puta dobijamo da je $-u^{(4)}(x) = f''(x)$, što zamenom u (23) daje

$$-u_{\bar{x}x}(x) = f(x) + \frac{h^2}{12} f''(x) + O(h^4).$$

U ovom slučaju šema povišene tačnosti dobija znatno jednostavniji oblik,

$$\begin{aligned} v_0 &= 0, \\ -v_{\bar{x}x,i} &= f_i + \frac{h^2}{12} f''_i, \quad i = 1, \ldots, n-1 \\ v_n &= 0. \end{aligned}$$

PRIMER 3. Šemom čija je greška $O(h^4)$ odredimo približno rešenje graničnog problema

$$u''(x) - u(x) = e^x, \qquad u(0) = u(1) = 0.$$

S obzirom da je

$$u^{(4)}(x) = u''(x) + e^x = u(x) + 2e^x,$$

zamenom u (23) se dobija da je

$$u_{\bar{x}x}(x) = u''(x) + \frac{h^2}{12}\big(u(x) + 2e^x\big) + O(h^4),$$

ili

$$u''(x) = u_{\bar{x}x}(x) - \frac{h^2}{12}\big(u(x) + 2e^x\big) + O(h^4).$$

Stoga je aproksimacija jednačine u unutrašnjim čvorovima mreže

$$\frac{1}{h^2}\big(v_{i-1} - 2v_i + v_{i+1}\big) - \big(1 + \frac{h^2}{12}\big)v_i = \big(1 + \frac{h^2}{6}\big)e^{x_i}, \quad i = 1, \ldots, n-1,$$

a diferencijska šema povišene tačnosti za dati problem je

$$\begin{gathered} v_0 = 0 \\ v_{i-1} - \big(2 + h^2 + \frac{h^4}{12}\big)v_i + v_{i+1} = h^2\big(1 + \frac{h^2}{6}\big)e^{x_i}, \quad i = 1, \ldots, n-1, \\ v_n = 0. \end{gathered}$$

Njeno rešenje za $h = 0.2$ je

$$\begin{aligned} &v_0 = 0, \qquad v_1 = -0.11071, \qquad v_2 = -0.17668, \\ &v_3 = -0.18966, \qquad v_4 = -0.13689, \qquad v_5 = 0. \end{aligned}$$

Poređenjem sa tačnim rešenjem, pokazuje se da su sve cifre dobijenih rezultata sigurne cifre.

9.3 Varijacione metode

Posmatrajmo granični problem za funkciju $u(x)$

$$Lu \equiv -\big(p(x)u'(x)\big)' + q(x)u(x) = f(x), \qquad 0 \le x \le 1, \tag{24}$$

gde je $p(x) \ge p_{min} > 0$ i $q(x) \ge 0$, sa Dirichletovim graničnim uslovima

$$u(0) = 0, \qquad u(1) = 0. \tag{25}$$

Ovo je *klasična formulacija* problema – za proizvoljnu funkciju $f \in \mathcal{C}(0,1)$, pri dovoljno glatkim koeficijentima $p(x)$ i $q(x)$, treba odrediti funkciju $u \in \mathcal{C}^2(0,1)$ koja zadovoljava jednačinu (24) i homogene granične uslove (25).

Jednačini (24) pridružimo kvadratni funkcional

$$I(w) = (Lw\,,\, w) - 2(f\,,\, w), \tag{26}$$

gde je sa $(\cdot\,,\,\cdot)$ označen uobičajeni skalarni proizvod u $\mathcal{L}_2(0,1)$,

$$(v\,,\, w) = \int_0^1 v(x)w(x)\,dx. \tag{27}$$

Funkcional (26) se može parcijalnom integracijom, uzimajući u obzir granične uslove (25), predstaviti i u sledećem obliku

$$\begin{aligned} I(w) &= \int_0^1 \big(-(pw')' + qw\big)w\,dx - 2\int_0^1 fw\,dx \\ &= \int_0^1 \big(p(w')^2 + qw^2 - 2fw\big)\,dx. \end{aligned} \tag{28}$$

Veza polaznog problema (24),(25) i funkcionala (26) data je sledećom teoremom:

TEOREMA 4. *Neka je L samokonjugovan i pozitivno definisan linearni operator i $I(w)$ kvadratni funkcional*

$$I(w) = (Lw\,,\, w) - 2(f\,,\, w),$$

a

$$Lu = f \tag{29}$$

granični problem sa homogenim graničnim uslovima. Ako granični problem ima rešenje $u(x)$, onda funkcional $I(w)$ dostiže minimum za $w(x) \equiv u(x)$; i obrnuto, funkcija u kojoj funcional $I(w)$ dostiže minimum, ako postoji, predstavlja rešenje graničnog problema (29).

DOKAZ: Neka je $u(x)$ rešenje graničnog problema (29). Kako je operator L samokonjugovan, to je

$$(Lu\,,\, w) = (u\,,\, Lw) = (Lw\,,\, u),$$

pa je

$$\begin{aligned} I(w) &= (Lw\,,\, w) - 2(Lu\,,\, w) = (Lw\,,\, w) - (Lw\,,\, u) - (Lu\,,\, w) \\ &= (Lw\,,\, w-u) - (Lu\,,\, w-u) - (Lu\,,\, u) = (L(w-u)\,,\, w-u) - (Lu\,,\, u). \end{aligned}$$

$(Lu\,,\, u)$ ne zavisi od w, a $(L(w-u)\,,\, w-u) \geq 0$, pri čemu jednakost važi samo za $w \equiv u$, jer je operator L pozitivno definisan. Stoga funkcional $I(w)$ dostiže svoj minimum za $w \equiv u$,

$$\min_w I(w) = I(u) = -(Lu\,,\, u),$$

čime je prvi deo tvrđenja dokazan.

Neka je sada u funkcija za koju funkcional $I(w)$ dostiže minimum. To znači da je za svaku funkciju

$$w = u + c\eta, \qquad c \in \mathcal{R}^1 \tag{30}$$

iz dopustive klase funkcija

$$I(w) - I(u) \geq 0.$$

Stoga je

$$\begin{aligned} I(w) - I(u) &= (Lw\,,\, w) - 2(Lu\,,\, w) + (Lu\,,\, u) + 2(Lu - f\,,\, w) - 2(Lu - f\,,\, u) \\ &= (L(w-u)\,,\, w-u) + 2(Lu - f\,,\, w-u) \\ &= c^2(L\eta\,,\, \eta) + 2c(Lu - f\,,\, \eta) \geq 0. \end{aligned}$$

Kvadratna funkcija po c će biti nenegativna ako joj diskriminanta nije veća od nule,

$$(Lu - f\,,\, \eta)^2 \leq 0,$$

što je moguće samo ako je

$$(Lu - f\,,\, \eta) = 0.$$

S obzirom da je η proizvoljna funkcija iz klase dopustivih funkcija, ovo je moguće samo ako je

$$Lu - f = 0, \qquad \text{tj.} \qquad Lu = f,$$

što je i trebalo dokazati. Time je teorema u potpunosti dokazana. ∎

Stoga se može, umesto problema (24),(25), rešavati ekvivalentan problem definisan *ekstremalnim principom*

$$\underset{w}{\text{minimizirati}}\ I(w). \tag{31}$$

Jednačina (24) je *Eulerova jednačina* ekstremalnog principa (31), koji predstavlja varijacionu formulaciju jednačine (24). S obzirom na zapis (28) funkcionala $I(w)$, vidimo da je skup dopustivih funkcija proširen, jer su oslabljeni zahtevi za glatkošću. Naime, $w \in \mathcal{V} \subseteq \mathcal{H}^1(0,1)$, gde je

$$\mathcal{H}^1(0,1) = \{w(x) \mid \int_0^1 (w^2 + w'^2)\, dx < \infty\}$$

prostor Soboleva, a $\mathcal{V}$ njegov potprostor funkcija koje zadovoljavaju granične uslove (25),

$$\mathcal{V} = \{w \in \mathcal{H}^1(0,1) \,|\, w(0) = w(1) = 0\}.$$

Rešenje ekstremalnog principa (31) se naziva *slabo rešenje* problema (24),(25). U praksi je česta varijaciona formulacija problema (princip minimuma energije,...).

Iz ekstremalnog principa proizilazi tzv. slaba ili Galerkinova forma problema. Naime, ako za $w = u$ funkcional $I(w)$ dostiže minimum, onda za ma koju drugu dopustivu funkciju, koja može da se predstavi u formi (30), važi da je

$$\begin{aligned} I(u) \le I(u + c\eta) &= (L(u + c\eta)\,,\, u + c\eta) - 2(f\,,\, u + c\eta) \\ &= I(u) + 2c\big((Lu\,,\, \eta) - (f\,,\, \eta)\big) + c^2(L\eta\,,\, \eta). \end{aligned}$$

Pošto c može biti proizvoljnog znaka, da bi ova nejednakost bila zadovoljena mora biti

$$(Lu\,,\, \eta) = (f\,,\, \eta), \qquad \forall \eta \in \mathcal{V}. \tag{32}$$

Jednačina (32) se naziva *slaba* ili *Galerkinova forma* problema (29). Za problem (24),(25) ona se može napisati i u obliku

$$\int_0^1 (pu'w' + quw - fw)\, dx = 0. \tag{33}$$

Granični uslovi Dirichletovog tipa (25) se nazivaju *esencijalni granični uslovi*, i ulaze u definiciju prostora $\mathcal{V}$ u kome tražimo rešenje.

U opštem slučaju, na granici mogu biti zadati i mešoviti granični uslovi,

$$u'(0) - \sigma_1 u(0) = 0, \qquad u'(1) + \sigma_2 u(1) = 0, \qquad \sigma_1\,, \sigma_2 \ge 0. \tag{34}$$

Tada Galerkinova jednačina (32)

$$\int_0^1 \big((-pu')' + qu - f\big)w\, dx = 0, \qquad \forall w \in \mathcal{H}^1(0,1)$$

parcijalnom integracijom prvog sabirka može da se zapiše u obliku

$$\int_0^1 (pu'w' + quw - fw)\,dx - p(1)u'(1)w(1) + p(0)u'(0)w(0) = 0,$$

pri čemu je sada prostor dopustivih funkcija $\mathcal{H}^1(0,1)$. Kako je iz (34) $u'(0) = \sigma_1 u(0)$ i $u'(1) = -\sigma_2 u(1)$, to je konačno Galerkinova jednačina problema (24),(34)

$$\int_0^1 (pu'w' + quw - fw)\,dx + \sigma_1 p(0)u(0)w(0) + \sigma_2 p(1)u(1)w(1) = 0. \tag{35}$$

Mešoviti granični uslovi (34) utiču na oblik Galerkinove jednačine (35), a ne učestvuju u definisanju prostora dopustivih funkcija. Takvi granični uslovi se nazivaju *prirodni granični uslovi*.

Ritzova metoda. Zamenom potprostora $\mathcal{V}$ u varijacionoj formulaciji (32) konačno dimenzionim potprostorom $\mathcal{S}^h \subset \mathcal{V}$, definišemo tzv. Ritzovu varijacionu metodu za rešavanje graničnih problema. Elementi potprostora $\mathcal{S}^h$ nazivaju se *probne funkcije*. Neka je $\phi_1(x), \ldots, \phi_n(x)$ jedan bazis potprostora $\mathcal{S}^h$. Ritzova aproksimacija rešenja graničnog problema (24),(25) se traži u obliku

$$v(x) = \sum_{i=1}^{n} c_i \phi_i(x), \tag{36}$$

pri čemu se konstante c_i određuju tako da aproksimacija (36) zadovoljava slabu formu za svaku funkciju $w(x)$ iz $\mathcal{S}^h$,

$$(Lv\,,\,w) = (f\,,\,w) \qquad \forall w \in \mathcal{S}^h. \tag{37}$$

Dovoljno je da ovaj uslov bude ispunjen za bazisne funkcije $\phi_i(x)$, $i = 1, \ldots, n$, da bi važio i za njihovu proizvoljnu linearnu kombinaciju $w \in \mathcal{S}^h$. Uvrstimo reprezentaciju (36) u (33), i, uzimajući u obzir prethodnu napomenu, dobijamo da je

$$\sum_{i=1}^{n} c_i \int_0^1 \left(p\phi_i'\phi_j' + q\phi_i\phi_j\right) dx = \int_0^1 f\phi_j\,dx, \qquad j = 1, \ldots, n. \tag{38}$$

Dakle, granični problem smo zamenili sistemom linearnih jednačina dimenzije jednake dimenziji potprostora $\mathcal{S}^h$.

U slučaju mešovitih graničnih uslova (34), slaba forma graničnog problema je data izrazom (35), a $\mathcal{S}^h$ je konačno dimenzioni potprostor prostora $\mathcal{H}^1(0,1)$. Opisanim algoritmom se i u ovom slučaju granični problem aproksimira sistemom linearnih jednačina po nepoznatim parametrima c_i.

Ako u polaznom problemu granični uslovi nisu homogeni, oduzimanjem od rešenja $u(x)$ funkcije $\phi_0(x)$ koja zadovoljava date nehomogene granične uslove, dobijamo problem sa homogenim graničnim uslovima za funkciju $u_1(x) = u(x) - \phi_0(x)$

zadat jednačinom $Lu_1 = f - L\phi_0$. Aproksimaciju funkcije $u_1(x)$ određujemo u obliku (36), te je aproksimacija rešenja polaznog problema

$$u(x) \approx \phi_0(x) + \sum_{i=1}^{n} c_i \phi_i(x).$$

Iz (32) i (37) je

$$\big(L(u-v)\,,\, w\big) = 0, \qquad \forall w \in \mathcal{S}^h,$$

što znači da je funkcija $L(u-v)$ ortogonalna na potprostor $\mathcal{S}^h$, tj. da je $v(x)$ ortogonalna projekcija rešenja $u(x)$ na potprostor $\mathcal{S}^h$ u smislu skalarnog proizvoda definisanog operatorom L,

$$(v\,,\, w)_L = (Lv\,,\, w) = \int_0^1 (pv'w' + qvw)\, dx. \tag{39}$$

Ako funkcije $\phi_i(x)$, $i = 1, 2, \ldots$,čine potpun ortogonalni sistem funkcija, onda iz $\big(L(u-v)\,,\, \phi_i\big) = 0$, $i = 1, 2, \ldots$, sledi da je $u = v$ skoro svuda, kada u (36) $n \to \infty$.

Galerkinova metoda. Slabu formu (32) jednačine

$$Lu = f \tag{40}$$

možemo izvesti ne koristeći ekstremalni princip (31), tj. i u slučaju kada L nije samokonjugovan i pozitivno definisan operator. Naime, skalarnim množenjem jednačine (40) tzv. *test funkcijom* $w \in \mathcal{W}$ dobijamo upravo Galerkinovu formu ove jednačine,

$$(Lu\,,\, w) = (f\,,\, w) \qquad \forall w \in \mathcal{W}. \tag{41}$$

Prostor test funkcija $\mathcal{W}$ ne mora biti identičan prostoru $\mathcal{V}$ kome pripada slabo rešenje problema.

Galerkinova metoda predstavlja diskretizaciju slabe forme (41). Kao i u Ritzovoj metodi, potprostor $\mathcal{V}$ se zamenjuje konačno dimenzionim potprostorom $\mathcal{S}^h \subseteq \mathcal{V}$, i aproksimacija traži u obliku linearne kombinacije (36), pri čemu se konstante c_i određuju tako da je

$$(Lv\,,\, w) = (f\,,\, w) \qquad \forall w \in \mathcal{W}. \tag{42}$$

U opštem slučaju, prostori probnih ($\mathcal{S}^h$) i test ($\mathcal{W}$) funkcija nisu identični. Ova metoda se svodi na Ritzovu metodu kada je L samokonjugovan i pozitivno definisan operator, a $\mathcal{W} = \mathcal{S}^h$.

Dovoljno je da jednačina (42) važi za bazisne funkcije $\psi_1, \ldots, \psi_n$ prostora $\mathcal{W}$,

$$(Lv\,,\, \psi_j) = (f\,,\, \psi_j) \qquad j = 1, \ldots, n, \tag{43}$$

što daje sledeći sistem linearnih jednačina po nepoznatim parametrima c_i u reprezentaciji (36)

$$\sum_{i=1}^{n} c_i \int_0^1 L\phi_i \psi_j\, dx = \int_0^1 f\psi_j\, dx, \qquad j = 1, \ldots, n.$$

Dakle, Galerkinovom metodom se približno rešenje određuje tako da funkcija greške

$$R(x; c_1, \ldots, c_n) \equiv Lv - f \tag{44}$$

bude ortogonalna na prostor test funkcija $\mathcal{W}$.

Metoda kolokacije. Diskretizacija slabe forme (41) kada je prostor test funkcija $\mathcal{W}$ generisan sa n δ-funkcija $\delta(x-x_j)$, $j = 1, \ldots, n$, gde su $x_j \in [0,1]$ zadate tačke, naziva se metodom kolokacije. Uslovi (43) za $\psi_j(x) \equiv \delta(x - x_j)$ se svode na

$$Lv(x_j) = f(x_j), \qquad j = 1, \ldots, n. \tag{45}$$

Kada uvrstimo reprezentaciju (36) u izraze (45), i u ovom slučaju dobijamo sistem linearnih jednačina po parametrima c_i,

$$\sum_{i=1}^{n} c_i L\phi_i(x_j) = f(x_j), \qquad j = 1, \ldots, n.$$

Metodom kolokacije se aproksimacija rešenja graničnog problema određuje tako da funkcija greške (44) bude jednaka nuli u n diskretnih tačaka, tzv. *tačaka kolokacije*. Njihov broj je jednak broju nepoznatih parametara c_i u približnom rešenju.

Metoda najmanjih kvadrata. Umesto minimizacije funkcionala (26) minimizira se funkcija greške (44),

$$\begin{aligned} \|R(x; c_1, \ldots, c_n)\|^2 &= (Lv - f\,,\, Lv - f) = (Lv\,,\, Lv) - 2(f\,,\, Lv) + (f\,,\, f) \\ &= (L^*Lv\,,\, v) - 2(L^*f\,,\, v) + (f\,,\, f). \end{aligned}$$

S obzirom da $(f\,,\, f)$ ne zavisi od v, ovom metodom se ustvari minimizira funkcional

$$J(v) = (L^*Lv\,,\, v) - 2(L^*f\,,\, v),$$

što ne predstavlja ništa drugo nego Ritzov funkcional (26) pridružen jednačini

$$L^*Lu = L^*f. \tag{46}$$

Novi problem je, očigledno, definisan samokonjugovanim operatorom bez obzira kakav je operator L, ali dvostruko višeg reda, što predstavlja značajan nedostatak ove metode. Osim što je složeniji račun, povećanim zahtevima za glatkošću približnog rešenja smanjuje se klasa dopustivih funkcija. Naime, diskretan oblik slabe forme jednačine (46) je

$$(L^*Lv\,,\, w) = (L^*f\,,\, w), \qquad \text{tj.} \qquad (Lv\,,\, Lw) = (f\,,\, Lw), \tag{47}$$

što znači da $v, w \in \mathcal{S}^h \subset \mathcal{H}^2(0,1)$, gde je $\mathcal{H}^2(0,1)$ prostor Soboleva

$$\mathcal{H}^2(0,1) = \{w \,|\, \int_0^1 (w^2 + w'^2 + w''^2)\, dx < \infty\}.$$

Kada u jednačinu (47) uvrstimo reprezentaciju (36), i stavimo da je $w(x) \equiv \phi_j(x)$, $j = 1, \ldots, n$, ponovo dobijamo sistem linearnih jednačina po parametrima c_i,

$$\sum_{i=1}^{n} c_i \int_0^1 L\phi_i L\phi_j \, dx = \int_0^1 f L\phi_j \, dx, \qquad j = 1, \ldots, n.$$

PRIMER 4. Galerkinovom metodom, metodom kolokacije i metodom najmanjih kvadrata nađimo aproksimaciju rešenja graničnog zadatka

$$u''(x) + (1 + x^2)u(x) + 1 = 0, \qquad u(-1) = u(1) = 0$$

polinomom četvrtog stepena.

Kako je interval simetričan u odnosu na koordinatni početak, a koeficijenti jednačine parne funkcije, i rešenje problema će biti parna funkcija. Stoga ćemo ga tražiti u obliku

$$v(x) = \sum_{i=1}^{2} c_i \phi_i(x) = c_1(1 - x^2) + c_2 x^2(1 - x^2).$$

Zamenom funkcije $v(x)$ u jednačinu, dobijamo da je funkcija greške

$$R(x; c_1, c_2) = 1 - c_1(1 + x^4) + c_2(2 - 11x^2 - x^6).$$

Galerkinovom metodom, pri izboru istih funkcija za probne i test funkcije, parametri c_i su rešenja sistema linearnih jednačina

$$c_1 \int_{-1}^{1} (1 + x^4)(1 - x^2) \, dx - c_2 \int_{-1}^{1} (2 - 11x^2 - x^6)(1 - x^2) \, dx = \int_{-1}^{1} (1 - x^2) \, dx$$
$$c_1 \int_{-1}^{1} (1 + x^4)x^2(1 - x^2) \, dx - c_2 \int_{-1}^{1} (2 - 11x^2 - x^6)x^2(1 - x^2) \, dx$$
$$= \int_{-1}^{1} x^2(1 - x^2) \, dx,$$

tj. $c_1 = 0.933, \quad c_2 = -0.054.$

Metodom kolokacije sa tačkama kolokacije $x_1 = 0$ i $x_2 = \frac{1}{2}$, rešavanjem sistema

$$R(x_i; c_1, c_2) = 0, \qquad i = 1, 2,$$

dobija se da je $c_1 = 0.957, \quad c_2 = -0.022.$

Metodom najmanjih kvadrata se minimizacijom funkcionala

$$\|R(x; c_1, c_2)\|^2 = \int_{-1}^{1} R^2(x; c_1, c_2) \, dx$$

dobija sistem

$$c_1 \int_{-1}^{1} (1+x^4)^2\,dx - c_2 \int_{-1}^{1} (1+x^4)(2-11x^2-x^6)\,dx = \int_{-1}^{1} (1+x^4)\,dx$$
$$c_1 \int_{-1}^{1} (1+x^4)(2-11x^2-x^6)\,dx - c_2 \int_{-1}^{1} (2-11x^2-x^6)^2\,dx$$
$$= \int_{-1}^{1} (2-11x^2-x^6)\,dx,$$

čije je rešenje $c_1 = 0.933, \quad c_2 = -0.068.$

9.4 Metoda konačnih elemenata

Svakom od navedenih varijacionih metoda se, u slučaju rešavanja graničnog problema definisanog linearnim operatorom, problem svodi na sistem linearnih jednačina po nepoznatim parametrima reprezentacije (36). U opštem slučaju, matrica tog sistema je puna matrica, tj. najveći broj njenih elemenata je različit od nule. Čak i kada bi koristili ortogonalne sisteme funkcija, ne bi dobili sisteme sa retkim matricama, jer to što je $(\phi_i\,,\,\phi_j) = \delta_{ij}$ ne znači i da je $(L\phi_i\,,\,\phi_j) = \delta_{ij}$.

Navedene varijacione metode se koriste i u višedimenzionom slučaju, tj. za rešavanje parcijalnih diferencijalnih jednačina, kada se javljaju dodatne poteškoće pri izboru funkcija ϕ_i koje zadovoljavaju granične uslove. U opštem slučaju granica može biti vrlo složena, te je praktično nemoguće konstruisati funkcije koje će na celoj granici zadovoljiti ovaj zahtev.

Metoda konačnih elemenata je upravo metoda koja koristi Ritz–Galerkinovu tehniku, a navedene probleme prevazilazi na poseban način izabranim probnim funkcijama ϕ_i. Ove funkcije se biraju tako da su identički jednake nuli u najvećem delu oblasti, a na preostalom delu oblasti su opisane deo po deo polinomima. S obzirom da su za najveći broj ovih funkcija oblasti u kojima su one različite od nule disjunktne, jasno je da će matrice pomenutih sistema linearnih jednačina biti retke. Što se tiče graničnih uslova u višedimenzionom slučaju, samo manji broj ovako definisanih bazisnih funkcija mora da zadovoljava granične uslove, i to samo na delu granice gde su različite od nule. Većina ovih funkcija je različita od nule unutar oblasti definisanosti problema, a na celoj granici su jednake nuli.

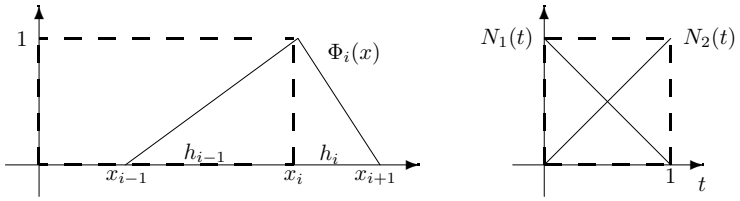
Ilustrujmo metodu na modelnom primeru (24),(25), čije se rešenje aproksimira u prostoru deo po deo linearnih funkcija. Na intervalu $[0,1]$ zadat je skup čvorova

$$\bar{\omega} = \{x_i \,|\, x_{i+1} - x_i = h_i,\ i = 0,\ldots,n,\ \sum_{i=0}^{n} h_i = 1\}.$$

Čvorovima x_i ovaj interval je podeljen na $n+1$ podintervala $[x_i, x_{i+1}]$, $i = 0,\ldots,n$, koji se nazivaju *konačni elementi*. Prostor probnih funkcija $\mathcal{S}^h$ je prostor tzv.

"krov" funkcija – neprekidnih funkcija koje su na dva susedna elementa linearne, a na ostalim identički jednake nuli (slika 9.1),

$$\phi_i(x) = \begin{cases} 1 + \frac{x-x_i}{h_{i-1}}, & x \in [x_{i-1}, x_i] \\ 1 - \frac{x-x_i}{h_i}, & x \in [x_i, x_{i+1}] \\ 0, & x \notin [x_{i-1}, x_{i+1}] \end{cases} . \tag{48}$$



Slika 9.1: Krov funkcije i njihove slike na kanonskom elementu.

Pošto je $\phi_i(x_j) = \delta_{ij}$, iz (36) sledi da je $c_i = v(x_i) \equiv v_i$, $i = 1, \ldots, n$, odnosno da koeficijenti u reprezentaciji (36) predstavljaju približne vrednosti rešenja graničnog problema u unutrašnjim čvorovima mreže $\bar{\omega}$. Stoga se približno rešenje (36) može zapisati i u sledećem obliku

$$v(x) = \sum_{i=1}^{n} v_i \phi_i(x). \tag{49}$$

Koeficijenti v_i se određuju kao rešenja sistema linearnih jednačina (38), u matričnom zapisu

$$K\mathbf{v} = \mathbf{f}, \tag{50}$$

gde je $K = (k_{ij})$ simetrična, kvadratna matrica dimenzije n sa elementima

$$k_{ij} = \int_0^1 (p\phi_i'\phi_j' + q\phi_i\phi_j)\, dx,$$

$\mathbf{v} = (v_1, \ldots, v_n)^T$ nepoznati vektor i $\mathbf{f} = (f_1, \ldots, f_n)^T$,

$$f_j = \int_0^1 f\phi_j\, dx, \qquad j = 1, \ldots, n.$$

S obzirom na (48), biće za $i = 1, \ldots, n$

$$k_{i,j} = 0, \qquad |i-j| > 1,$$

$$k_{i-1,i} = \int_{x_{i-1}}^{x_i} (p\phi_{i-1}'\phi_i' + q\phi_{i-1}\phi_i)\, dx \qquad k_{i,i} = \sum_{j=0}^{1} \int_{x_{i-1+j}}^{x_{i+j}} (p{\phi_i'}^2 + q{\phi_i}^2)\, dx$$

$$f_i = \sum_{j=0}^{1} \int_{x_{i-1+j}}^{x_{i+j}} f\phi_i\, dx.$$

Očigledno je da se parametri matrice K i vektora $\mathbf{f}$ mogu dobiti sabiranjem odgovarajućih integrala po elementima (asembliranjem). To je osnovna prednost metode konačnih elemenata – aproksimacija se određuje lokalno, tj. na svakom konačnom elementu, a zatim se globalna aproksimacija na celoj oblasti dobija asembliranjem ovih lokalnih aproksimacija. Pošto su lokalne aproksimacije na elementima istog tipa i jednoobrazno se određuju, dovoljno je odrediti aproksimaciju rešenja na tzv. kanonskom elementu. U ovom slučaju, kanonski element može biti jedinični interval, i preslikavanje ma kog elementa $[x_i, x_{i+1}]$, $i = 0, \ldots, n$, na njega je zadato izrazom

$$x = x_i + h_i t = x_i(1-t) + x_{i+1}t. \tag{51}$$

Na elementu $[x_i, x_{i+1}]$ probna funkcija je, prema (49) i (48),

$$\begin{aligned} v(x) &= v_i\phi_i(x) + v_{i+1}\phi_{i+1}(x) = v_i\frac{x_{i+1}-x}{h_i} + v_{i+1}\frac{x-x_i}{h_i} \\ &= v_iN_1^i(x) + v_{i+1}N_2^i(x), \end{aligned} \tag{52}$$

tj. predstavljena je linearnom kombinacijom bazisnih funkcija

$$N_1^i(x) = \frac{1}{h_i}(x_{i+1}-x), \qquad N_2^i(x) = \frac{1}{h_i}(x-x_i) \qquad i = 1, \ldots, n. \tag{53}$$

Smenom (51) u izrazu (53) dobijamo da je za svako $i = 1, \ldots, n$,

$$N_1^i(x) \equiv N_1(t) = 1-t, \qquad N_2^i(x) \equiv N_2(t) = t, \qquad t \in [0,1],$$

gde su $N_1(t)$ i $N_2(t)$ bazisne funkcije kanonskog elementa (slika 9.1). Pomoću ovih funkcija preslikavanje (51) može da se zapiše izrazom

$$x(t) = x_iN_1(t) + x_{i+1}N_2(t), \qquad i = 0, \ldots n,$$

a probna funkcija (52) je

$$v(x(t)) = v_iN_1(t) + v_{i+1}N_2(t). \tag{54}$$

Konačni elementi kod kojih je bazisnim funkcijama definisano preslikavanje na kanonski element nazivaju se *izoparametarski elementi*.

Na jednom elementu $[x_i, x_{i+1}]$ sistem (50) se svodi, s obzirom na (54), na sistem od dve jednačine

$$\begin{aligned} &\int_{x_i}^{x_{i+1}} \Big(p\big(v_iN_1^i + v_{i+1}N_2^i\big)'\big(N_j^i\big)' + q\big(v_iN_1^i + v_{i+1}N_2^i\big)N_j^i\Big)\,dx \\ &\qquad = \int_{x_i}^{x_{i+1}} fN_j^i\,dx, \qquad j = 1, 2, \end{aligned} \tag{55}$$

koji, posle smene (51), ima sledeći oblik

$$h_i \int_0^1 \Big(-\tfrac{1}{h^2} p\big(x(t)\big)\big(v_{i+1}-v_i\big) + q\big(x(t)\big)\big(v_i(1-t)+v_{i+1}t\big)(1-t) \Big)\, dt = h_i \int_0^1 f\big(x(t)\big)(1-t)\, dt,$$

$$h_i \int_0^1 \Big(\tfrac{1}{h^2} p\big(x(t)\big)\big(v_{i+1}-v_i\big) + q\big(x(t)\big)\big(v_i(1-t)+v_{i+1}t\big)t \Big)\, dt = h_i \int_0^1 f\big(x(t)\big)t\, dt.$$

Da bismo ilustrovali algoritam do kraja, pretpostavimo da je $p(x) \equiv p = const$ i $q(x) \equiv q = const$. Tada se integrali na levoj strani jednačina poslednjeg sistema mogu izračunati tačno, i sistem postaje

$$\frac{p}{h_i}(v_i - v_{i+1}) + qh_i(\tfrac{1}{3}v_i + \tfrac{1}{6}v_{i+1}) = h_i \int_0^1 f(x(t))(1-t)\, dt$$

$$\frac{p}{h_i}(-v_i + v_{i+1}) + qh_i(\tfrac{1}{6}v_i + \tfrac{1}{3}v_{i+1}) = h_i \int_0^1 f(x(t))t\, dt.$$

Sistem se može zapisati u matričnom obliku

$$k^i \mathbf{v}^i \equiv (k_s^i + k_m^i)\mathbf{v}^i = \mathbf{f}^i, \tag{56}$$

gde je

$$k_s^i = \begin{pmatrix} \frac{p}{h_i} & -\frac{p}{h_i} \\ -\frac{p}{h_i} & \frac{p}{h_i} \end{pmatrix} = \frac{p}{h_i}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

matrica krutosti elementa,

$$k_m^i = \begin{pmatrix} \frac{qh_i}{3} & \frac{qh_i}{6} \\ \frac{qh_i}{6} & \frac{qh_i}{3} \end{pmatrix} = \frac{qh_i}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

matrica mase elementa, i

$$\mathbf{v}^i = \begin{pmatrix} v_i \\ v_{i+1} \end{pmatrix}, \qquad \mathbf{f}^i = \begin{pmatrix} h_i \int_0^1 f(x(t))(1-t)\, dt \\ h_i \int_0^1 f(x(t))t\, dt \end{pmatrix}.$$

U opštem slučaju, vektor $\mathbf{f}^i$ se računa primenom kvadraturnih formula. U slučaju da koeficijenti jednačine $p(x)$ i $q(x)$ nisu konstante, i parametri matrica krutosti i mase bi se računali pomoću kvadraturnih formula.

Sumiranjem jednačina (55) po svim elementima, a to znači asembliranjem svih sistema (56) za $i = 0, \ldots, n$, dobijamo sistem (50). Asembliranje matrica svih sistema (56) znači njihovo spajanje u jednu matricu dimenzije $n \times n$, ali tako da se parametri matrica k^i različitih konačnih elemenata sabiraju ako odgovaraju istom

globalnom čvoru. Naime, svaki unutrašnji čvor je zajednički za po dva elementa, te će doprinosi koji potiču u matricama k^{i-1} i k^i od čvora x_i biti sabrani pri asembliranju. Tako se, u slučaju jednakih konačnih elemenata, tj. za $h_i = h$, $i = 0, \ldots, n$, dobija da je matrica sistema (50)

$$K = K_s + K_m,$$

gde je

$$K_s = \frac{p}{h}\begin{pmatrix} 1+ & 1 & -1 & & & 0 \\ & -1 & 1+1 & -1 & & \\ & & & & \ddots & \\ & 0 & & & & 1+1 \end{pmatrix} = \frac{p}{h}\begin{pmatrix} 2 & -1 & & & & 0 \\ -1 & 2 & -1 & & & \\ & & & & \ddots & \\ 0 & & & & & 2 \end{pmatrix}$$

globalna matrica krutosti, a

$$K_m = \frac{qh}{6}\begin{pmatrix} 2+2 & 1 & & & & 0 \\ 1 & 2+2 & 1 & & & \\ & & & & \ddots & \\ 0 & & & & & 2+2 \end{pmatrix} = \frac{qh}{6}\begin{pmatrix} 4 & 1 & & & & 0 \\ 1 & 4 & 1 & & & \\ & & & & \ddots & \\ 0 & & & & & 4 \end{pmatrix}$$

globalna matrica mase. Zbog homogenih graničnih uslova, prva i poslednja vrsta i kolona (one koje odgovaraju čvorovima x_0 i x_{n+1}) su izostavljene. Vektor desne strane $\mathbf{f}$ ima elemente

$$f_1 = h\int_0^1 f\big(x(t)\big)(1-t)\,dt, \qquad f_n = h\int_0^1 f\big(x(t)\big)t\,dt,$$

$$f_i = h\int_0^1 f\big(x(t)\big)\,dt, \quad i = 2, \ldots, n-1.$$

U razvijenom obliku, sistem (50) je

$$\begin{aligned} (2\frac{p}{h} + \frac{2}{3}qh)v_1 - (\frac{p}{h} - \frac{1}{6}qh)v_2 &= f_1 \\ -(\frac{p}{h} - \frac{1}{6}qh)v_{i-1} + (2\frac{p}{h} + \frac{2}{3}qh)v_i - (\frac{p}{h} - \frac{1}{6}qh)v_{i+1} &= f_i, \qquad i = 2, \ldots, n-1. \\ -(\frac{p}{h} - \frac{1}{6}qh)v_{n-1} + (2\frac{p}{h} + \frac{2}{3}qh)v_n &= f_n \end{aligned}$$

Koristeći uobičajeni diferencijski operator

$$v_{\bar{x}x,i} = \frac{1}{h^2}(v_{i+1} - 2v_i + v_{i-1}),$$

poslednji sistem, posle deljenja sa h, može da se zapiše u vidu diferencijske šeme

$$\begin{aligned} v_0 &= 0 \\ (-p + \frac{qh^2}{6}) v_{\bar{x}x,i} + q v_i &= \frac{1}{h} f_i, \qquad i = 1, \ldots, n. \\ v_{n+1} &= 0 \end{aligned}$$

Greška aproksimacije (49) određene linearnim elementima (48) je ([29])

$$\|u - v\|_L \leq ch\|f\|,$$

gde je sa $\|\cdot\|_L$ označena tzv. energetska norma definisana skalarnim proizvodom (39), a $\|\cdot\|$ je uobičajena L_2-norma (27). Štaviše, u čvorovima se dokazuje tzv. superkonvergencija,

$$\max_{1 \leq i \leq n} |u(x_i) - v_i| = O(h^2).$$

Isti algoritam, samo tehnički znatno složeniji, se primenjuje i u slučaju kada se koriste bazisne funkcije sastavljene deo po deo od polinoma višeg stepena, pri čemu mogu biti postavljeni još dodatni uslovi glatkosti na granicama elemenata.

Metoda je naročito stekla popularnost za rešavanje parcijalnih diferencijalnih jednačina, gde njene prednosti posebno dolaze do izražaja.

9.5 Problem sopstvenih vrednosti

Specijalan slučaj graničnog problema (1),(4) je

$$\begin{aligned} -u''(x) &= \lambda u(x), \qquad 0 < x < 1 \\ u(0) &= u(1) = 0, \end{aligned} \tag{57}$$

koji ima netrivijalno rešenje samo za neke vrednosti parametra λ. Vrednost parametra λ za koju problem (57) ima netrivijalno rešenje naziva se *sopstvena vrednost*, a odgovarajuće rešenje *sopstvena funkcija* graničnog problema (57). U opštem slučaju, problem sopstvenih vrednosti je definisan jednačinom $Lu = \lambda u$, gde je L neki operator.

Da bi prikazali numeričke metode za rešavanje problema sopstvenih vrednosti diferencijalnih operatora, vratimo se modelnom problemu (57), bez obzira što su njegove sopstvene vrednosti i funkcije poznati,

$$\lambda = \lambda_k = k^2\pi^2, \qquad u(x) \equiv u_k(x) = \sin k\pi x, \qquad k = 1, 2, \ldots . \tag{58}$$

S obzirom da problem sopstvenih vrednosti predstavlja poseban oblik graničnog problema, za njegovo numeričko rešavanje se koriste već pomenute metode konačnih razlika i varijacione metode.

Diferencijska šema koja aproksimira problem (57) sa greškom $O(h^2)$ je, prema (16),

$$\begin{aligned} -v_{\bar{x}x,i} &= \lambda_h v_i, \qquad i = 1, \dots, n-1 \\ v_0 &= v_n = 0, \end{aligned} \tag{59}$$

gde je sa λ_h označena aproksimacija sopstvene vrednosti. Ova šema, ustvari, predstavlja homogeni sistem linearnih jednačina sa trodijagonalnom matricom u kojoj figuriše parametar λ_h. Stoga se granični problem (57), uzimajući u obzir granične vrednosti, metodom konačnih razlika svodi na problem sopstvenih vrednosti

$$A\mathbf{v} = \lambda_h \mathbf{v},$$

$(n-1)$-dimenzione kvadratne matrice

$$A = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & \\ & & \ddots & \\ 0 & & & 2 \end{pmatrix}.$$

Sopstveni vektori su diskretizacija prvih $(n-1)$ sopstvenih funkcija (58) problema (57) na čvorovima mreže,

$$\mathbf{v}^k = \begin{pmatrix} \sin k\pi x_1 \\ \vdots \\ \sin k\pi x_{n-1} \end{pmatrix}, \qquad k = 1, \dots, n-1,$$

što se lako proverava zamenom u (59),

$$\begin{aligned} -(\sin k\pi x)_{\bar{x}x,i} &= \frac{1}{h^2}\big(-\sin k\pi(x_i + h) + 2\sin k\pi x_i - \sin k\pi(x_i - h)\big) \\ &= \frac{1}{h^2}(-2\sin k\pi x_i \cos k\pi h + 2\sin k\pi x_i) \\ &= \frac{4}{h^2}\sin^2\frac{k\pi h}{2}\sin k\pi x_i. \end{aligned}$$

Iz poslednjeg izraza neposredno sledi da su aproksimacije sopstvenih vrednosti

$$\lambda_{h,k} = \frac{4}{h^2}\sin^2\frac{k\pi h}{2}, \qquad k = 1, \dots, n-1. \tag{60}$$

Iz asimptotskog razvoja za malo k je

$$\lambda_{h,k} = \frac{4}{h^2}\big(\frac{k\pi h}{2} + O(k^3h^3)\big)^2 = k^2\pi^2 + O(k^4h^2) = \lambda_k + O(k^4h^2),$$

što znači da $\lambda_{h,k} \to \lambda_k$ kada $h \to 0$. Za veće k, na primer za $k = n-1$ je

$$\lambda_{h,n-1} = \frac{4}{h^2}\sin^2\frac{(n-1)\pi h}{2} = \frac{4}{h^2}\sin^2(\frac{\pi}{2} - \frac{\pi h}{2}) = \frac{4}{h^2}\cos^2\frac{\pi h}{2},$$

pa je

$$\frac{\lambda_{h,n-1}}{\lambda_{n-1}} = \frac{\frac{4}{h^2}\cos^2\frac{\pi h}{2}}{(n-1)^2\pi^2} = \frac{4}{\pi^2}\frac{\cos^2\frac{\pi h}{2}}{(1-h)^2} = \frac{4}{\pi^2}\big(1 + 2h + O(h^2)\big) \to \frac{4}{\pi^2}, \quad \text{kada } h \to 0.$$

Možemo da zaključimo sledeće. Kontinualan problem (57) ima prebrojivo mnogo sopstvenih vrednosti i sopstvenih funkcija. Njegova diskretna aproksimacija konačnim razlikama (59) ima ih konačno mnogo, tačnije $(n-1)$, gde je $(n+1)$ broj čvorova mreže diferencijske šeme. Stoga diferencijskom šemom (59) određujemo aproksimacije (60) prvih $(n-1)$ sopstvenih vrednosti i sopstvenih vektora, pri čemu je aproksimacija sopstvene vrednosti to lošija što je k veće. Zgušnjavajem mreže, tj. smanjivanjem koraka i povećavanjem broja čvorova n, dobijaju se aproksimacije većeg broja sopstvenih vrednosti, i veća tačnost aproksimacije onih sa nižim indeksom k.

Slična situacija nastaje i pri korišćenju varijacionih metoda. S obzirom da se svakom od pomenutih metoda problem svodi na sistem linearnih jednačina po koeficijentima c_i reprezentacije (36), čija je desna strana određena funkcijom $f(x)$, očigledno ja da će za problem (57) sistem biti homogen. U sistemu figuriše i parametar λ_h, koji određujemo tako da matrica sistema bude singularna, tj. da sistem ima netrivijalno rešenje. Prema tome, i u ovom slučaju se problem svodi na nalaženje sopstvenih vrednosti i vektora matrice pomenutog sistema. Broju čvorova koji u metodi konačnih razlika određuje dimenziju problema sopstvenih vrednosti, u varijacionoj metodi odgovara broj sabiraka u približnom rešenju (36). Veće n omogućava aproksimaciju većeg broja sopstvenih vrednosti i funkcija.

Numeričke metode za rešavanje problema sostvenih vrednosti i vektora matrica date su u poglavlju 6.

Literatura

[1] Aljančić S., *Uvod u realnu i funkcionalnu analizu*, Građevinska knjiga, Beograd, 1968.

[2] Bahvalov N.S., *Numerical Methods*, Mir Publishers, Moscow, 1977.

[3] Berezin I.S., Židkov N., *Numerička analiza*, Naučna knjiga, Beograd, 1963.

[4] Daubechies I., *Ten lectures on wavelets*, SIAM, Philadelphia, 1992.

[5] Davis P., *Interpolation and Approximation*, Dover Publications, New York, 1975.

[6] Demidovich B.P., Maron I.A., *Computational Mathematics*, Mir Publishers, Moscow, 1987.

[7] Fox L., Parker J.B., *Chebyshev Polynomials in Numerical Analysis*, Oxford University Press, London, 1972.

[8] Graps A., *An introduction to wavelets*, IEEE Comp. Science and Engineering, 2, 1995.

[9] Hageman L.,Young D., *Applied Iterative Methods*, Academic Press, New York, 1981.

[10] Hall G., Watt J.M. (editors), *Modern Numerical Methods for Ordinary Differential Equations*, Clarendon Press, Oxford, 1976.

[11] Hildebrand F.B., *Introduction to Numerical Analysis*, McGraw–Hill, New York, 1974.

[12] Hildebrand F.B., *Advanced Calculus for Applications*, Prentice–Hall, Englewood Cliffs, N.J., 1976.

[13] Jovanović B., Radunović D., *Numerička analiza*, Matematički fakultet, 2003.

[14] Kalitkin N.N., *Numerical methods* (rus.), Nauka, Moskva, 1978.

[15] Kantorovch L.V., Akilov G.P., *Functional analyis* (rus.), Nauka, Moskva, 1977.

[16] Mallat S., *A theory of multiresolution signal decomposition: The wavelet representation*, IEEE Trans. Pattern Anal. and Machine Intel., 11, 674–693, 1989.

[17] Marchuk G.I., *Methods of numerical mathematics* (rus.), Nauka, Moskva, 1980.

[18] Meyer Y., *Wavelets - Algorithms & Applications*, SIAM, Philadelphia, 1993.

[19] Oden J.T., *Finite Elements of Nonlinear Continua*, McGraw–Hill, New York, 1972.

[20] Oden J.T., Reddy J.N., *An Introduction to the Mathematical Theory of Finite Elements*, Wiley, New York, 1976.

[21] Ortega J., *Numerical Analysis – A Second Course*, SIAM, Philadelphia, 1990.

[22] Ortega J., Rheinboldt W., *Iterative Solutions of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[23] Parlett B., *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, N.J., 1980.

[24] Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P., *Numerical Recipes in C*, Cambridge University Press, 1992

[25] Ralston A., *A First Course in Numerical Analysis*, McGraw–Hill, Tokyo, 1965.

[26] Stoer J., Bulirsch R., *Introduction to Numerical Analysis*, Springer–Verlag, New York 1980.

[27] Strang G., *Introduction to Applied Mathematics*, Willesley–Cambridge Press, 1986.

[28] Strang G., *Wavelets and dilation equations: a brief introduction*, SIAM Rew., 31, 614–627, 1989.

[29] Strang G., Fix G.J., *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, N.J., 1973.

[30] Wilkinson J.H., *The evaluation of the zeros of ill-conditioned polynomials, Part I*, Numer. Math. 1, 150–180, 1959.