UNIVERZITET U KRAGUJEVCU
PRIRODNO–MATEMATIČKI FAKULTET

Miodrag Spalević, Miroslav Pranić

NUMERIČKE METODE

KRAGUJEVAC, 2007.

NUMERIČKE METODE

Univerzitetski udžbenik

AUTORI: *dr Miodrag Spalević*, vanredni profesor PMF-a u Kragujevcu
dr Miroslav Pranić, viši asistent PMF-a u Banjaluci

RECENZENTI: *dr Boško Jovanović*,
redovni profesor Matematičkog fakulteta u Beogradu
dr Gradimir Milovanović,
redovni profesor Elektronskog fakulteta u Nišu,
dopisni član SANU

IZDAVAČ: Prirodno–matematički fakultet u Kragujevcu
`www.pmf.kg.ac.yu`

ZA IZDAVAČA: *Prof. dr Radoslav Žikić*, dekan

SLOG: autori

ŠTAMPA: "SKVER", Kragujevac

TIRAŽ: 200 primeraka

ISBN 978–86–81829–84-4

Predgovor

Ovaj udžbenik je namenjen prvenstveno studentima prirodno-matematičkih fakulteta kao uvod u numeričku matematiku. Napisan je u skladu sa programom za jednosemestralni predmet Numerička matematika koji će se se od ove školske godine držati u četvrtom semestru studija matematike i informatike Prirodno-matematičkog fakulteta u Banjaluci. Takav isti predmet je predviđen i za studente matematike Prirodno-matematičkog fakulteta u Kragujevcu, po novim programima u skladu sa novim Zakonom o visokom obrazovanju. Dodavanjem poslednje glave, kao i sekcije o numeričkom diferenciranju, pokrivamo program istoimenog predmeta Numeričke metode koji je od ove godine uveden studentima Mašinskog fakulteta u Beogradu. Prvi autor, koji je prethodne godine držao predmet Numeričke metode na doktorskim studijama Mašinskog fakulteta u Kraljevu, smatra da će ovom knjigom biti pokriven i dobar deo programa odgovarajućeg predmeta. Udžbenik može biti od koristi i na drugim tehničkim fakultetima gde se izučava odgovarajuća problematika. Takođe, udžbenik može jednim svojim delom biti korišćen i za sticanje osnovnih znanja iz numeričke matematike za obdarene učenike odeljenja matematičke gimnazije, za koje ne postoji dovoljno prikladne literature iz ove oblasti.

Celokupan rukopis ove knjige podeljen je u sedam glava, teorijska razmatranja praćena su odgovarajućim primerima. Većina metoda koje se obrađuju u knjizi može se programirati korišćenjem nekog programskog jezika ili po-

moću nekog savremenog programskog sistema. Deo rezultata koji se navode u primerima u ovoj knjizi, dobili smo korišćenjem programskog jezika FORTRAN i/ili programskog sistema MATLAB. Autorima je od posebne pomoći u pisanju ovog udžbenika bila literatura koja se navodi na kraju.

Prva glava je uvodnog karaktera, daje kratak pregled razvoja numeričke matematike, kao i osnovnih pojmova vezanih za ovu oblast matematike. Posebno su obrađene osnove teorije grešaka. Izložene su u kratkim crtama osnove o uslovljenosi kako algoritma, tako i samog problema koji se rešava, a sve u svetlu rešavanja datog numeričkog problema na računaru.

Druga glava se bavi problemima u linearnoj algebri, kako direktnim tako i indirektnim iterativnim metodama za rešavanje sistema linearnih jednačina.

Osnovni problemi interpolacije funkcije pomoću Lagrangeovog, Newtonovih i Hermiteovog interpolacionog polinoma razmatraju se u trećoj glavi. Dodate su i osnove numeričkog diferenciranja.

U četvrtoj glavi se razmatraju nelinearne jednačine i sistemi nelinearnih jednačina i poznate iterativne metode za njihovo približno rešavanje.

U petoj glavi je opisana veoma poznata i često upotrebljavana metoda najmanjih kvadrata.

Šesta glava je posvećena približnom izračunavanju određenog integrala, ili kako se često kaže numeričkoj integraciji.

Konačno, sedma glava je posvećena problemima približnog rešavanja običnih diferencijalnih jednačina. Posebno su tretirane numeričke metode Runge-Kutta i linearni višekoračni metodi.

Autori se nadaju da će knjiga biti od koristi studentima kojima je namenjena, ali i svima onima koji koriste numeričke metode u svojim istraživanjima.

Autori se zahvaljuju recenzentima dr Bošku Jovanoviću (redovni profesor Matematičkog fakulteta u Beogradu) i dr Gradimiru Milovanoviću (redovni

profesor Elektronskog fakulteta u Nišu, dopisni član SANU) na korisnim sugestijama. Njihove primedbe su doprinele poboljšanju prve verzije ovog rukopisa.

Unapred se zahvaljujemo svim čitaocima koji nam ukažu na eventualne propuste u tekstu, kako bismo iste u narednom izdanju ove knjige korigovali.

Kragujevac/Banjaluka, 2007. godine AUTORI

Sadržaj

Glava 1

Elementi teorije grešaka

1.1 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.

Programski realizovane numeričke metode (uz pomoć računara) omogućavaju korisnicima brzo rešavanje problema sa proizvoljnom tačnošću. Kao baza numeričke matematike razvila se oblast pod nazivom *teorija aproksimacija*. Kao posebna oblast numeričke matematike izdvaja se *teorija optimizacija* koja tretira razne optimizacione probleme. Sve ove oblasti su imale snažan razvoj u poslednjih nekoliko decenija, o čemu svedoči veliki broj naučnih rezultata koji se publikuju u specijalizovanim časopisima za primenjenu matematiku i numeričku analizu, kao što su:

Mathematics of Computation (Američko matematičko društvo); *Numerische Mathematik*, *BIT*, *Calcolo*, *Constructive Approximation*, *Numerical Algorithms* (Springer Verlag); *SIAM Journal on Numerical Analysis*, *SIAM Journal on Optimization*, *SIAM Journal on Scientific Computing* (i ostali časopisi SIAM-a, društva za industrijsku i primenjenu matematiku, SAD); *Journal of Computational and Applied Mathematics*, *Applied Mathematics and Computation*, *Journal of Approximation Theory* (Elsevier); *IMA Journal on Numerical Analysis* (Oxford University Press) i mnogi drugi.

Značajan napredak je učinjen i u realizaciji programskih paketa visokokvalitetnog numeričkog softvera. Programski paketi su mahom implementirani na programskom jeziku FORTRAN. U novije vreme postoje implementacije i na jeziku C++. Veliki broj matematičkih softverskih paketa danas se slobodno distribuira (videti: `http://gams.nist.gov`).

Posebnu pažnju treba obratiti na programske sisteme, koji su se pojavili u poslednje vreme,

MATLAB (The MathWorks, Inc., `http://www.mathworks.com`),

MATHEMATICA (Wolfram Research, Inc., `http:www.wolfram.com`),

MAPLE (Waterloo Maple, Inc., `http://www.maplesoft.com`), itd.

Na primer, u MATLAB-u je ugrađen dobar deo visokokvalitetnog softvera kojim se rešavaju problemi iz raznih oblasti numeričke matematike, posebno oni koji se odnose na rešavanje problema u linearnoj algebri. Svi pomenuti programski sistemi omogućavaju korisniku programiranje na jedan veoma jednostavan način.

Na kraju ove uvodne sekcije, posebno bismo ukazali na knjigu jednog od rodonačelnika moderne numeričke analize Waltera Gautschija [3] koja je nedavno objavljena, a odnosi se na izračunavanja i aproksimacije u teoriji ortogonalnih polinoma i kvadraturnih formula. Kao prateći deo ove knjige javlja se skup programskih kodova (za metode obrađene u knjizi) urađenih u programskom sistemu MATLAB koji se slobodno mogu preuzeti sa Web adrese

`http://www.cs.purdue.edu/archives/2002/wxg/codes/` .

1.2 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 zaokruživanja.

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.

1.3 Približni brojevi

Neki brojevi, na primer π, $\sqrt{2}$, e, $\frac{2}{3}$, ..., ne mogu se zapisati 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 1.1. 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

$$a = p \cdot 10^q \quad (a = p \cdot 2^q)\,, \qquad |p| < 1\,,\ \ q \text{ ceo broj}, \tag{1.1}$$

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

Primer 1.2. Ako je $n = 10$, $t = 7$ i $e = 3$, zapisi u pokretnom zarezu

broja iz primera 1.1 mogu biti

$$\boxed{3120700\,|\,002}\,, \qquad \boxed{0312070\,|\,003}\,, \quad \ldots$$

Očigledno je da zadavanjem brojeva t 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 (1.1) je $|p| \geq 10^{-1}$ ($|p| \geq 2^{-1}$ u slučaju binarnog zapisa). U primeru 1.2 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 t 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 $\bar{a}$ njena približna vrednost onda je veličina $|a - \bar{a}|$ apsolutna, a $|a - \bar{a}|/|a|$ relativna greška, i

$$\begin{aligned} &|a - \bar{a}| \leq \Delta(\bar{a}) && \textit{granica apsolutne greške}\,, \\ &\left|\frac{a - \bar{a}}{a}\right| \leq \delta(\bar{a}) && \textit{granica relativne greške}\,. \end{aligned} \tag{1.2}$$

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

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

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

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

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

$$\bar{a} = \pm\left(\alpha_1 10^n + \ldots + \alpha_k 10^{n-k+1} + \ldots + \alpha_m 10^{n-m+1}\right), \tag{1.3}$$

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

PRIMER 1.3. U broju $\bar{a} = 0.03120700$ 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 $\bar{a} = 3.120700 \cdot 10^{-2}$. 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 dekadne jedinice, koja odgovara toj cifri, pomnožene zadatim težinskim faktorom, tj. α_k je sigurna cifra ako je

$$\Delta(\bar{a}) \leq \omega \cdot 10^{n-k+1}, \qquad 0 < \omega \leq 1. \tag{1.4}$$

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 1.4. Ako se zna da je $\Delta(\bar{a}) = 0.5 \cdot 10^{-5}$ apsolutna greška približnog broja $\bar{a} = 0.03120700$, onda su, s obzirom na (1.4), sigurne cifre $3, 1, 2$ i 0. Poslednje tri cifre $(7, 0, 0)$ nisu sigurne, jer u broju a čija je $\bar{a}$ približna vrednost, umesto ovih cifara mogu stajati i ma koje druge. Naime, s obzirom na definiciju (1.2) apsolutne greške, a se nalazi u intervalu

$$\begin{array}{l} 0.03120700 - 0.5 \cdot 10^{-5} \leq \quad a \quad \leq 0.03120700 + 0.5 \cdot 10^{-5} \quad \text{tj.} \\ 0.03120200 \leq \quad a \quad \leq 0.03121200 \end{array}$$

te se poslednje tri cifre brojeva a i $\bar{a}$ 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 $\alpha_{k+1} = 5$, $\alpha_i = 0$ $(i > k+1)$, a α_k parno. U ostalim slučajevima se α_k povećava za jedan. U primeru 1.4, posle odbacivanja cifara koje nisu sigurne, biće $\bar{a} = 0.03121$.

PRIMER 1.5.1. Zaokružimo broj

$$\overline{a} = 72.353,\ \Delta(\overline{a}) = 0.026,$$

tako da mu sve cifre budu sigurne u užem smislu.

Kako je $\Delta(\overline{a}) = 0.026 < 0.05 = \frac{1}{2} \cdot 10^{-1}$ i $n = 1$, iz $n - k + 1 = -1$ dobijamo da je $k = 3$, pa su sigurne cifre broja $\overline{a}$ u užem smislu $7, 2$ i 3. Zaokružimo broj $\overline{a}$ na broj sa tri cifre, tj. $\overline{a}_1 = 72.4$. Greška koje se javlja pri ovom zaokruživanju je $\Delta = 0.047$. Približan broj $\overline{a}_1$ je dat sa greškom

$$\Delta(\overline{a}_1) = \Delta(\overline{a}) + \Delta = 0.026 + 0.047 = 0.073.$$

No, kako je $0.073 > 0.05$ to znači da u broju $\overline{a}_1 = 72.4$ poslednja cifra 4 nije sigurna u užem smislu. Zato, zaokružimo broj $\overline{a} = 72.353$ na broj sa dve cifre, tj. $\overline{a}_2 = 72$. Greška koje se javlja pri ovom zaokruživanju je $\Delta_1 = 0.353$. Približan broj $\overline{a}_2$ je dat sa greškom

$$\Delta(\overline{a}_2) = \Delta(\overline{a}) + \Delta_1 = 0.026 + 0.353 = 0.379 < 0.5,$$

što znači da su 7 i 2 sigurne cifre, tj. u zaokruženom broju $\overline{a}_2 = 72.$ sve cifre su sigurne u užem smislu.

Između broja sigurnih cifara i relativne greške postoji sledeća veza:

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

gde je k broj sigurnih cifara broja $\bar{a}$, a α_1 njegova prva sigurna cifra. Zaista, s obzirom da je cifra α_k poslednja sigurna cifra broja $\bar{a}$, prema (1.4) je

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

Deljenjem ove dvostruke nejednakosti sa $|\bar{a}| \neq 0$ i korišćenjem reprezentacije (1.3), dobijamo

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

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

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

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 1.5.2. U primeru 1.4, u broju 0.03120700 datom sa tačnošću $\Delta(\bar{a}) = 0.5 \cdot 10^{-5}$, 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 $\bar{a} = 0.03121$, tj. broj ima četiri sigurne cifre. Njegova relativna greška je $\delta(\bar{a}) = 1.6 \cdot 10^{-4}$.

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 1.6. 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

$$I_0 = \ln 1.1 \ , \qquad I_n + 10\, I_{n-1} = \frac{1}{n} \ , \quad n = 1, 2, \ldots \ . \tag{1.5}$$

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 (1.5) definisan je nestabilan algoritam, iako nikakva aproksimacija nije vršena.

Sa druge strane, algoritam

$$I_n = \frac{0.1}{n+1} - \frac{0.01}{n+2} + \frac{0.001}{n+3} - \dots\ , \qquad n = 0, 1, \dots\ , \tag{1.6}$$

koji je dobijen razvojem podintegralne funkcije u red

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

je stabilan. Štaviše, alternativni red (1.6) brzo konvergira, te se sa svega nekoliko sabiraka može postići zadovoljavajuća tačnost. Poređenja radi, u sledećoj tabeli je dato nekoliko vrednosti integrala izračunatih formulama (1.5) i (1.6):

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 1.7. Manji koren kvadratne jednačine $x^2 - 140\,x + 1 = 0$ 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 $\bar{x}_2 = 70 - 69.99 = 0.01$. Dakle, rezultat je dobijen sa samo jednom sigurnom cifrom, tj. relativnom greškom $\delta(\bar{x}_2) = 1 = 100\%$, 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 $1.4 \cdot 10^{-4}$.

PRIMER 1.8. Navedimo još jedan primer oduzimanja približnih brojeva. Zaokruživanjem brojeva y_1 i y_2 dobijeni su brojevi $\overline{y}_1 = 2.78493$ i $\overline{y}_2 = 2.78469$. Ocenimo apsolutnu i relativnu grešku njihove razlike $\overline{u} = \overline{y}_1 - \overline{y}_2$ i analizirajmo problem gubitka značajnih cifara. S ozirom da su brojevi $\overline{y}_1 = 2.78493$ i $\overline{y}_2 = 2.78469$ nastali zaokruživanjem brojeva y_1 i y_2 oni aproksimiraju brojeve y_1 i y_2 sa 6 značajnih cifara i za apsolutne greške važi

$$|\overline{y}_i - y_i| \leq 0.5 \cdot 10^{-5}, \quad i = 1, 2,$$

a za relativne greške

$$\frac{|\overline{y}_i - y_i|}{|y_i|} \approx \frac{|\overline{y}_i - y_i|}{|\overline{y}_i|} \leq 0.18 \cdot 10^{-5}, \quad i = 1, 2.$$

Za apsolutnu grešku razlike brojeva sada imamo

$$\begin{aligned} |\overline{u} - u| &= |(\overline{y}_1 - \overline{y}_2) - (y_1 - y_2)| \\ &= |(\overline{y}_1 - y_1) - (\overline{y}_2 - y_2)| \leq |(\overline{y}_1 - y_1)| + |(\overline{y}_2 - y_2)| = 10^{-5}. \end{aligned}$$

S ozirom da je

$$\overline{u} = \overline{y}_1 - \overline{y}_2 = 0.00024,$$

za relativnu grešku razlike imamo

$$\frac{|\overline{u} - u|}{|u|} \approx \frac{|\overline{u} - u|}{|\overline{u}|} \leq \frac{10^{-5}}{24 \cdot 10^{-5}} \approx 0.42 \cdot 10^{-1},$$

pa zaključujemo da $\overline{u} = 0.24 \cdot 10^{-3}$ aproksimira tačnu vrednost $u = y_1 - y_2$ sa dve značajne cifre.

Dakle, pri oduzimanju bliskih brojeva, došlo je do"gubitka" značajnih cifara (operandi su imali po 6 značajnih cifara a rezultat ima samo dve značajne cifre). Naravno, s ozirom da je broj značajnih cifara povezan sa granicom relativne greške to u stvari znači da je došlo do povećanja granice relativne greške (sa 10^{-5} na 10^{-1}). To je i logično s obzirom da je pri oduzimanju

bliskih brojeva rezultat daleko manji od svakog od operanada ponaosob, naravno, posmatrano po modulu.

Može se pokazati da nema gubitka značajnih cifara kod ostalih računskih operacija (sabiranja, množenja i deljenja).

Primerima 1.6, 1.7 i 1.8 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 1.9. Opšte rešenje diferencijalne jednačine $y''(x) = y(x)$ je

$$y(x) = \frac{1}{2}\left(y(0) + y'(0)\right)e^{x} + \frac{1}{2}\left(y(0) - y'(0)\right)e^{-x}, \tag{1.7}$$

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 (1.7) ne anulira, te se u rešenju pojavljuje i član oblika $\varepsilon\, e^{x}$, koji za veće vrednosti x unosi veliku grešku u približno rešenje.

1.3.1 Reprezentacija realnih brojeva u računaru

Realni brojevi se u računaru predstavljaju pomoću binarne reprezentacije,

$$a = \pm\, p \cdot 2^{e}, \quad \text{gde je} \quad 1 \le p < 2. \tag{1.8}$$

Stoga je

$$p = (b_0\,.\,b_1\,b_2\,b_3\,\ldots\,b_{t-1})_2 \quad \text{pri čemu je} \quad b_0 = 1. \tag{1.9}$$

Na primer, broj $111.5 = (1101111.1)_2$ se može napisati kao $(1.1011111)_2 \cdot 2^6$. Vidimo da je u prikazu ovog broja, oblika (1.8), binarna tačka pomaknuta za šest mesta ulevo, a pritom se eksponent povećao za 6. Kako se zahteva da je

$b_0 = 1$, možemo pisati

$$p = (1\,.\,b_1\,b_2\,b_3\ldots b_{t-1})_2.$$

U tom normalizovanom prikazu, binarne cifre desno od binarne tačke čine razlomljeni deo mantise.

Da bismo memorisali normalizovane brojeve u računar, podelimo memorijsku reč (sadržaj jedne ćelije) u tri dela koja zovemo polja. Kod 32-bitnih računara, reč uobičajeno ima 32 bita pa se obično deli na sledeći način (tip `single`): 1 bit za predznak, 8 bita za eksponent e i 23 bita za mantisu. Bit za predznak je 0 (1) ako je broj pozitivan (negativan). Polje za eksponent ima osam bitova pa može reprezentovati eksponent e koji je između granica -128 i 127. Preostala 23 bita za smeštaj mantise koriste se za smeštaj razlomljenog dela mantise, jer je uvek $b_0 = 1$, pa ga ne treba memorisati. Zato se b_0 obično naziva skriveni bit. Realni broj a nazivamo *egzaktno reprezentabilnim* u računaru ili *brojem s pokretnom tačkom*[1] ako se na opisani način može bez greške smestiti u računar. Ako broj nije egzaktno reprezentabilan u računaru, on se mora pre memorisanja u računaru zaokružiti.

Na računaru generalno ne možemo egzaktno izvršavati aritmetičke operacije. Rezultat sabiranja, oduzimanja, množenja ili deljenja dva broja x i y, reprezentabilna u računaru, po definiciji je broj u računaru koji je najbliži tačnom (egzaktnom) zbiru, razlici, proizvodu, odnosno količniku x i y. Pri tome je relativna greška tako izvedenih operacija manja ili jednaka polovini najvećeg relativnog odstojanja dva susedna broja u računaru[2]. Na primer, u standardnoj jednostrukoj preciznosti (32-bitna reprezentacija) je relativno odstojanje susednih brojeva omeđeno s 2^{-23} pa je relativna greška aritmetičkih operacija najviše $\approx 10^{-8}$.

[1] Skup svih ovakvih brojeva označavaćemo sa $\mathbb{R}(t, s)$

[2] Ovaj broj zovemo mašinska preciznost i u ovoj glavi ćemo ga označavati sa eps.

Navedena pravila za izvršavanje elementarnih aritmetičkih operacija lako zapisujemo na sledeći način:

- sabiranje: $x \oplus y = (x+y)(1+\epsilon_1), \quad |\epsilon_1| \leq \text{eps},$
- oduzimanje: $x \ominus y = (x-y)(1+\epsilon_2), \quad |\epsilon_2| \leq \text{eps},$
- množenje: $x \odot y = xy(1+\epsilon_3), \quad |\epsilon_3| \leq \text{eps},$
- deljenje: $x \oslash y = \frac{x}{y}(1+\epsilon_4), \quad |\epsilon_4| \leq \text{eps}, \quad y \neq 0.$

Ove relacije važe ako su rezultati navedenih operacija po apsolutnoj vrednosti u intervalu (μ, M), gde je npr. u 32-bitnoj reprezentaciji $\mu = 2^{-126} \approx 10^{-38}$ najmanji, a $M = (1 + 2^{-1} + \ldots + 2^{-23})2^{127} \approx 10^{38}$ najveći normalizovani mašinski broj. U dvostrukoj tačnosti (64-bitna reprezentacija brojeva) je $\mu \approx 10^{-308}, M \approx 10^{308}$.

1.4 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 $\bar{y}$ približna vrednost za y. *Apsolutna greška* veličine $\bar{y}$ je

$$A(\bar{y}) = \sup_{(a_1,\ldots,a_n)\in G} |y(a_1, \ldots, a_n) - \bar{y}|, \tag{1.10}$$

a *relativna greška* je $\dfrac{A(\bar{y})}{|\bar{y}|}$.

Ako je oblast G n–dimenzioni pravougaonik

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

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

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

$$0 \leq \theta \leq 1.$$

Stoga je, na osnovu (1.10),

$$A(\bar{y}) \leq \sum_{k=1}^{n} \sup_{G} \left| \frac{\partial y}{\partial a_k}(a_1, \ldots, a_n) \right| \Delta(\bar{a}_k) \,. \tag{1.11}$$

U praksi se umesto ocene (1.11) koristi tzv. *linearna ocena* apsolutne greške funkcije

$$\Delta(\bar{y}) = \sum_{k=1}^{n} \left| \frac{\partial y}{\partial a_k}(\bar{a}_1, \ldots, \bar{a}_n) \right| \Delta(\bar{a}_k) \,. \tag{1.12}$$

Pri tome može se pokazti da važi

$$\Delta(\bar{y}) + \varepsilon_1(\rho) \leq A(\bar{y}) \leq \Delta(\bar{y}) + \varepsilon_2(\rho) \,,$$

gde je

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

što znači da je ocena (1.12) zadovoljavajuća za male apsolutne greške argumenata.

Primer 1.10. Odrediti grešku vrednosti funkcije $y = a^{10}$ za $\bar{a} = 1$ i $\Delta(\bar{a}) = 10^{-3}$.

Kako je $\bar{y} = 1$, $\sup_{|a-1| \leq 10^{-3}} \left| \frac{dy}{da}(a) \right| = 10.09$ i $\frac{dy}{da}(1) = 10$, to je, prema (1.10), apsolutna greška funkcije

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

ocena ove greške izrazom (1.11)

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

a linearna ocena greške je

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

U ovom slučaju nema značajnije razlike između navedenih ocena.

Ako je, međutim, $\Delta(\bar{a}) = 10^{-1}$, apsolutna greška je $A(\bar{y}) = 1.5$, ocena ove greške izrazom (1.11) je $A(\bar{y}) \leq 2.3$, a linearna ocena greške je $\Delta(\bar{y}) = 1$. 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

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

gde su $\gamma_k,\ k = 1, \ldots, n$, 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 $\Delta(\bar{y}) = \sum_{k=1}^{n} \Delta(\bar{a}_k)$.

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

$$y = a_1^{e_1} \cdot \ldots \cdot a_n^{e_n} \ , \tag{1.13}$$

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

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

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

$$\delta(\bar{y}) = \frac{\Delta(\bar{y})}{|\bar{y}|} = \sum_{k=1}^{n} |e_k|\,\delta(\bar{a}_k) \, . \tag{1.14}$$

Kada je y proizvod ili količnik $|e_k| = 1$, te je tvrđenje dokazano. Očigledno, ocena (1.14) važi i za opštiji oblik stepene funkcije (1.13).

PRIMER 1.11. Odredimo granicu apsolutne greške i granicu relativne greške približne vrednosti funkcije

$$f(x,y,z)=\frac{x^2+y\sqrt{z}}{x+2y}$$

ako su date približne vrednosti

$$\bar{x}=1.24,\quad \bar{y}=0.66,\quad \bar{z}=1.96,$$

a sve su napisane cifre sigurne u užem smislu.

Pošto su sve napisane cifre sigurne u užem smislu, a u svakom od približnih brojeva $\bar{x},\bar{y},\bar{z}$ poslednja napisana cifra se nalazi na decimalnom mestu koje karakteriše dekadna jedinica 10^{-2}, zaključujemo da su granice grešaka

$$\Delta(\bar{x}),\Delta(\bar{y}),\Delta(\bar{z})$$

takve da je

$$\Delta(\bar{x})\le\frac{1}{2}\,10^{-2},\quad \Delta(\bar{y})\le\frac{1}{2}\,10^{-2},\quad \Delta(\bar{z})\le\frac{1}{2}\,10^{-2}.$$

Dalje, imamo da je

$$\Delta(\bar{f})\le\left|\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial x}\right|\Delta(\bar{x})+\left|\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial y}\right|\Delta(\bar{y})+\left|\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial z}\right|\Delta(\bar{z}),$$

$$\frac{\partial f}{\partial x}=\frac{x^2+4xy-y\sqrt{z}}{(x+2y)^2},\quad \frac{\partial f}{\partial y}=\frac{x(\sqrt{z}-2x)}{(x+2y)^2},\quad \frac{\partial f}{\partial z}=\frac{y}{2\sqrt{z}(x+2y)},$$

$$\bar{f}=0.9615625,\qquad \overline{\overline{f}}=0.962,$$

$$\left|\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial x}\right|\le 0.5932,\quad \left|\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial y}\right|\le 0.2044,\quad \left|\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial z}\right|\le 0.0921,$$

$$\begin{aligned}
\Delta(\bar{f}) &\le \frac{1}{2}\,10^{-2}\cdot 0.8897,\quad \Delta\left(\overline{\overline{f}}\right)\le\Delta(\bar{f})+\frac{1}{2}\,10^{-3},\\
\Delta(\bar{f}) &\le 0.45\cdot 10^{-2},\quad \Delta\left(\overline{\overline{f}}\right)\le 0.5\cdot 10^{-2}=\frac{1}{2}\,10^{-2},\\
\delta(\bar{f}) &\le \frac{0.45\cdot 10^{-2}}{0.9615625},\\
\delta(\bar{f}) &\le 0.4680\cdot 10^{-2},\\
\delta(\bar{f}) &\le 0.47\%\,.
\end{aligned}$$

1.5 Obratan (inverzan) 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

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

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

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

Približno, granica apsolutne greške argumenta je određena relacijom

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

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 $\Delta(\bar{a}_1), \ldots, \Delta(\bar{a}_n)$. Ako je zadata linearna ocena apsolutne greške funkcije (1.12), dodatni uslovi koje apsolutne greške argumenata treba da zadovoljavaju obično se definišu na jedan od sledećih načina:

(*i*) *Princip jednakih uticaja*

$$\left|\frac{\partial y}{\partial a_1}(\bar{a}_1, \ldots, \bar{a}_n)\right| \Delta(\bar{a}_1) = \ldots = \left|\frac{\partial y}{\partial a_n}(\bar{a}_1, \ldots, \bar{a}_n)\right| \Delta(\bar{a}_n)\,.$$

Onda je

$$\Delta(\bar{y}) = n \left|\frac{\partial y}{\partial a_k}(\bar{a}_1, \ldots, \bar{a}_n)\right| \Delta(\bar{a}_k)\,,$$

te je

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

(*ii*) *Princip jednakih apsolutnih grešaka*

$$\Delta(\bar{a}_1) = \ldots = \Delta(\bar{a}_n)\,.$$

Iz (1.12) je

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

odakle sledi da je

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

Napomenimo da se ovaj princip može primeniti samo ako su sve promenljive istorodne (tj. imaju istu fizičku dimenziju).

(*iii*) *Princip jednakih relativnih grešaka*

$$\delta(\bar{a}_1) = \ldots = \delta(\bar{a}_n) .$$

Sada, (1.12) može da se napiše u obliku

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

pa je

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

Primer 1.12. Odredimo s kolikom tačnošću treba naći promenljive x, y, z da bi se veličina

$$f(x, y, z) = \frac{xy + \sqrt{z}}{x + 2z}$$

odredila s tačnošću 10^{-3}, ako su približne vrednosti argumenata

$$\bar{x} = 2.16, \quad \bar{y} = 1.12, \quad \bar{z} = 1.44$$

i ako se usvoji princip jednakih uticaja na grešku.

Kako granica apsolutne greške $\Delta(\bar{f})$ približne vrednosti

$$\bar{f} = \frac{\bar{x}\bar{y} + \sqrt{\bar{z}}}{\bar{x} + 2\bar{z}}$$

mora biti manja od 10^{-3}, što znači, da je, ako se priraštaj funkcije f zameni njenim totalnim diferencijalom,

$$\Delta(\bar{f}) = \left|\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial x}\right| \Delta(\bar{x}) + \left|\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial y}\right| \Delta(\bar{y}) + \left|\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial z}\right| \Delta(\bar{z}) < 10^{-3}. \tag{1.15}$$

Ako se usvoji princip jednakih uticaja na grešku, onda je

$$\left|\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial x}\right| \Delta(\bar{z}) = \left|\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial y}\right| \Delta(\bar{y}) = \left|\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial x}\right| \Delta(\bar{z}). \tag{1.16}$$

Kako je

$$\frac{\partial f}{\partial x} = \frac{2yz - \sqrt{z}}{(x+2z)^2}, \quad \frac{\partial f}{\partial y} = \frac{x}{x+2z}, \quad \frac{\partial f}{\partial z} = \frac{x - 2z - 4xy\sqrt{z}}{2(x+2z)^2\sqrt{z}},$$

imamo

$$\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial x} = \frac{2\cdot 1.12\cdot 1.44 - \sqrt{1.44}}{(2.16 + 2\cdot 1.44)^2} = 0.079743\ldots, \tag{1.17}$$

$$\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial y} = \frac{2.16}{2.16 + 2\cdot 1.44} = 0.4285714\ldots, \tag{1.18}$$

$$\frac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial z} = \frac{2.16 - 2\cdot 1.44 - 4\cdot 2.16\cdot 1.12\cdot\sqrt{1.44}}{2(2.16 + 2\cdot 1.44)^2\sqrt{1.44}} = -0.2022864\ldots. \tag{1.19}$$

Iz (1.15), (1.16) i (1.17) sledi

$$\Delta(\bar{x}) \le \frac{10^{-3}}{3\left|\dfrac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial x}\right|} < \frac{1}{3\cdot 10^3\cdot 0.079743} = 0.00418\ldots.$$

Analogno, iz (1.15), (1.16) i (1.18) imamo

$$\Delta(\bar{y}) \le \frac{10^{-3}}{3\left|\dfrac{\partial f(\bar{x},\bar{y},\bar{z})}{\partial y}\right|} < \frac{1}{3\cdot 10^3\cdot 0.4285714} = 0.00077\ldots,$$

i, konačno, iz (1.15), (1.16) i (1.19) je

$$\Delta(\bar{z}) \le \frac{10^{-3}}{3\left|\dfrac{\partial f(x^*,y^*,z^*)}{\partial z}\right|} < \frac{1}{3\cdot 10^3\cdot 0.2022864} = 0.00164\ldots.$$

Praktično, treba uzeti na primer,

$$\Delta(\bar{x}) \leq 0.004, \quad \Delta(\bar{y}) \leq 0.0007, \quad \Delta(\bar{z}) \leq 0.001.$$

PRIMER 1.13. Moment inercije valjka poluprečnika osnove r i mase m izračunava se po obrascu

$$J = \frac{m\,r^2}{2}.$$

Ako su m i r dati približnim vrednostima $\overline{m} = 500\,g$ i $\overline{r} = 2\,cm$, odredimo sa kakvim granicama apsolutnih grešaka treba da budu određene ove veličine, ako zahtevamo da su odgovarajuće granice relativnih grešaka jednake, da bi moment inercije bio određen s granicom relativne greške od 2% .

Granica apsolutne greške $\Delta(\overline{J})$ približne vrednosti $\overline{J} = \overline{m}\,\overline{r}^2/2$ iznosi

$$\Delta(\overline{J}) = \left|\frac{\partial J(\overline{m},\overline{r})}{\partial m}\right| \Delta(\overline{m}) + \left|\frac{\partial J(\overline{m},\overline{r})}{\partial r}\right| \Delta(\overline{r}),$$

gde je $\partial J/\partial m = r^2/2$ i $\partial J/\partial r = mr$. Granica relativne greške za $\overline{J}$ mora biti manja od 2% , odnosno,

$$\delta(\overline{J}) = \frac{\Delta(\overline{J})}{|\overline{J}|} = \frac{\left|\overline{m}\,\dfrac{\partial J}{\partial m}\right|}{|\overline{J}|}\,\delta(\overline{m}) + \frac{\left|\overline{r}\,\dfrac{\partial J}{\partial r}\right|}{|\overline{J}|}\,\delta(\overline{r}) \leq 0.02, \tag{1.20}$$

gde su

$$\delta(\overline{m}) = \frac{\Delta(\overline{m})}{|\overline{m}|} \quad \text{i} \quad \delta(\overline{r}) = \frac{\Delta(\overline{r})}{|\overline{r}|} \tag{1.21}$$

redom granice relativnih grešaka za $\overline{m}$ i $\overline{r}$. Po uslovu zadatka je

$$\delta(\overline{m}) = \delta(\overline{r}) = \delta\,. \tag{1.22}$$

Iz (1.20) i (1.22) sledi

$$\delta \leq \frac{0.02\,|\overline{J}|}{\left|\overline{m}\,\dfrac{\partial J(\overline{m},\overline{r})}{\partial m}\right| + \left|\overline{r}\,\dfrac{\partial J(\overline{m},\overline{r})}{\partial r}\right|} = \frac{0.02 \cdot \dfrac{1}{2}\,\overline{m}\,\overline{r}^2}{\overline{m} \cdot \dfrac{1}{2}\overline{r}^2 + \overline{r}\,\overline{m}\,\overline{r}} = \frac{0.02}{3} = 0.0066\ldots\,, \tag{1.23}$$

što je ispunjeno ako se uzme $\delta \leq 0.006$. Odavde se, zbog (1.21) i (1.22) dobija

$$\Delta(\overline{m}) \leq |\overline{m}| \cdot 0.006 = 500\,g \cdot 0.006 = 3\,g,$$
$$\Delta(\overline{r}) \leq |\overline{r}| \cdot 0.006 = 2\,cm \cdot 0.006 = 0.012\,cm.$$

1.6 Uslovljenost problema

Neka je $\mathbb{R}^l$ l-dimenzionalni realni vektorski prostor ($l \in \mathbb{N}$). Neka je zadat problem (u vidu preslikavanja f):

$$f : \mathbb{R}^m \mapsto \mathbb{R}^n, \quad y = f(x). \tag{1.24}$$

Analiziraaćemo kompjutersko rešavanje problema i proceniti granicu totalne greške tako dobijenog rešenja.

Problem (1.24) možemo prikazati grafički na sledeći način:

$$x \longrightarrow \boxed{f} \longrightarrow y\,.$$

Dakle, posmatrajmo preslikavanje f, u kojem je ulaz (input) dat u obliku vektora $x \in \mathbb{R}^m$, izlaz (output) je vektor $y \in \mathbb{R}^n$, a okvir (box) f prihvata ulaz x i zatim rešava problem za taj ulaz pri čemu se dobije izlaz y. Pokušajmo da analiziramo kako će se mala promena ulaza (x) odraziti na promenu izlaza (y). Drugim rečima, pokušajmo da ustanovimo osetljivost preslikavanja f u nekoj datoj tački x na male promene x. Stepen te osetljivosti iskazujemo jednim brojem kojeg nazivamo *faktor uslovljenosti* ili *kondicioni broj* preslikavanja f u tački x. Pri tome, za sada, pretpostavljamo da se funkcija f izračunava tačno, sa beskonačnom preciznošću. Dakle, uslovljenost funkcije f je njeno svojstvo koje ne zavisi od algoritma kojim se ona izračunava.

Videli smo da se u izračunavanju na računskoj mašini često na mesto vektora x u izračunavanju pojavljuje njemu "blizak" vektor $\overline{x}$, gde je $\overline{x} =$

$x + \delta$ i štaviše, rastojanje $\|\delta\|$ od x do $\overline{x}$ možemo oceniti pomoću izraza u kome figuriše mašinska preciznost. Ovo, naravno, pri tačnom izračunavanju funkcije f, dovodi, ne do vrednosti y, nego do $\overline{y}$, tj. $\overline{y} = f(\overline{x})$. Ako, pak, znamo kako preslikvanje f reaguje na male promene ulaza, takve kao što je δ, možemo reći nešto o greški $\overline{y} - y$, u rešenju $\overline{y}$, koja je uzrokovana tom promenom.

Faktor uslovljenosti. Startovaćemo sa najprostijim slučajem jedne funkcije jedne promenljive.

Slučaj $m = n = 1$*:* $y = f(x)$. Pretpostavimo najpre da je $x \neq 0, y \neq 0$, i označimo sa Δx male promene od x. Korišćenjem Taylorove formule za odgovarajuću promenu Δy, imamo

$$\Delta y = f(x + \Delta x) - f(x) \approx f'(x)\Delta x, \tag{1.25}$$

pod pretpostavkom da je funkcija f diferencijabilna u tački x. Formulu (1.25), s obzirom da nas interesuju relativne greške, napišimo u formi

$$\frac{\Delta y}{y} \approx \frac{x f'(x)}{f(x)} \cdot \frac{\Delta x}{x}. \tag{1.26}$$

Aproksimaciona jednakost postaje tačna jednakost ako je f linearna funkcija ili u graničnom slučaju kada $\Delta x \to 0$. Ovo sugeriše definisanje uslovljenosti f u x sa

$$(\operatorname{cond} f)(x) := \left| \frac{x f'(x)}{f(x)} \right|. \tag{1.27}$$

Ovaj broj, koji smo nazvali faktor uslovljenosti ili kondicioni broj, nam pokazuje koliko puta je veća relativna promena y u odnosu na relativnu promenu x. Što je ovaj broj veći kažemo da je problem (1.24) slabije uslovljen i, obrnuto, što je on manji to je problem (1.24) bolje uslovljen.

U slučaju kada je $x = 0$ a $y \neq 0$ faktor uslovljenosti definišemo sa $|f'(x)/f(x)|$. Slično, za $y = 0, x \neq 0$ faktor uslovljenosti je $|x f'(x)|$. Ako je $x = y = 0$, korišćenjem (1.25), faktor uslovljenosti bi bio $|f'(x)|$.

Slučaj proizvoljnih m, n. Ovde imamo

$$x = [x_1\ x_2\ \ldots\ x_m]^\top \in \mathbb{R}^m\,, \quad y = [y_1\ y_2\ \ldots\ y_n]^\top \in \mathbb{R}^n\,.$$

Preslikavanje f predstavljamo preko komponenti

$$y_\nu = f_\nu(x_1, x_2, \ldots, x_m), \quad \nu = 1, 2, \ldots, n. \tag{1.28}$$

Pretpostavljamo da svaka funkcija f_ν ima parcijalne izvode u odnosu na m promenljivih u tački x. Ako imamo promenu u komponenti x_μ u funkciji (1.28), a na osnovu (1.27), promena se može okarakterisati vrednostima koje definišemo sa

$$\gamma_{\nu\mu}(x) := (\operatorname{cond} f)(x) := \left| \frac{x_\mu\ \dfrac{\partial f_\nu}{\partial x_\mu}}{f_\nu(x)} \right|. \tag{1.29}$$

To nam daje celokupnu matricu faktora uslovljenosti $\Gamma(x) = [\gamma_{\nu\mu}(x)] \in \mathbb{R}_+^{n\times m}$. Da bismo dobili jedinstven faktor uslovljenosti, možemo uzeti neku pogodnu meru "odstupanja" matrice $\Gamma(x)$ kakva je, na primer, norma matrice definisana nešto kasnije u (1.32)

$$(\operatorname{cond} f)(x) = \|\Gamma(x)\|, \quad \Gamma(x) = [\gamma_{\nu\mu}(x)]\,. \tag{1.30}$$

Uslovljenost tako definisana, naravno, zavisi od norme, ali red odstupanja mogao bi biti manje-više isti za bilo koju razumnu normu.

Ako su komponente od x, ili od y, jednake nuli, (1.29) se modifikuje na isti način kako je to urađeno prethodno, za jednodimenzionalni slučaj.

Nešto grublja analiza, slična onoj za jednodimenzionalni slučaj, može se izvesti definisanjem relativne promene $x \in \mathbb{R}^m$ pomoću

$$\frac{\|\Delta x\|_{\mathbb{R}^m}}{\|x\|_{\mathbb{R}^m}}\,, \quad \Delta x = [\Delta x_1\ \Delta x_2\ \ldots\ \Delta x_m]^\top\,,$$

gde je Δx promena vektora čije komponente Δx_μ su promene komponenti x_μ, i gde je $\|\cdot\|_{\mathbb{R}^m}$ neka norma vektora u $\mathbb{R}^m$. Za promenu Δy prouzrokovanu

sa Δx, slično se definiše relativna promena $\|\Delta y\|_{\mathbb{R}^n}/\|y\|_{\mathbb{R}^n}$, sa podesnom vektorskom normom $\|\cdot\|_{\mathbb{R}^n}$ u $\mathbb{R}^n$. Onda je cilj uporediti relativne promene za y i x.

Da bismo to izveli, potrebno je definisati matričnu normu za matrice $A \in \mathbb{R}^{n\times m}$. Izaberimo takozvanu "operator normu",

$$\|A\|_{\mathbb{R}^{n\times m}} := \max_{0\neq x\in\mathbb{R}^m} \frac{\|Ax\|_{\mathbb{R}^n}}{\|x\|_{\mathbb{R}^m}}\,.$$

Na dalje uzećemo za vektorske norme "uniformnu" (ili beskonačnu) normu,

$$\|x\|_{\mathbb{R}^m} = \max_{1\leq\mu\leq m} |x_\mu| =: \|x\|_\infty, \quad \|y\|_{\mathbb{R}^n} = \max_{1\leq\nu\leq n} |y_\nu| =: \|y\|_\infty\,. \tag{1.31}$$

Onda je lako pokazati da

$$\|A\|_{\mathbb{R}^{n\times m}} := \|A\|_\infty = \max_{1\leq\nu\leq n} \sum_{\mu=1}^{m} |\alpha_{\nu\mu}|, \quad A = [\alpha_{\nu\mu}] \in \mathbb{R}^{n\times m}\,. \tag{1.32}$$

Sada, po analogiji sa (1.25), imamo

$$\Delta y_\nu = f_\nu(x+\Delta x) - f_\nu(x) \approx \sum_{\mu=1}^{m} \frac{\partial f_\nu}{\partial x_\mu}\,\Delta x_\mu\,.$$

Dakle, bar aproksimativno,

$$\begin{aligned} |\Delta y_\nu| &\leq \sum_{\mu=1}^{m} \left|\frac{\partial f_\nu}{\partial x_\mu}\right| |\Delta x_\mu| \leq \max_\mu |\Delta x_\mu| \cdot \sum_{\mu=1}^{m} \left|\frac{\partial f_\nu}{\partial x_\mu}\right| \\ &\leq \max_\mu |\Delta x_\mu| \cdot \max_\nu \sum_{\mu=1}^{m} \left|\frac{\partial f_\nu}{\partial x_\mu}\right|\,. \end{aligned}$$

Pošto ovo važi za svako $\nu = 1, 2, \ldots, n$, to takođe važi za $\max_\nu|\Delta y_\nu|$, dajući, u smislu (1.31) i (1.32),

$$\|\Delta y\|_\infty \leq \|\Delta x\|_\infty \left\|\frac{\partial f}{\partial x}\right\|_\infty\,. \tag{1.33}$$

Ovde je

$$\left\| \frac{\partial f}{\partial x} \right\| = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix} \in \mathbb{R}^{n \times m}$$

Jacobijeva matrica preslikavanja f. (To je analogon prvog izvoda funkcije jedne promenljive, sada za sistem funkcija sa više promenljivih.) Iz (1.33) sada se neposredno dobija za relativne promene

$$\frac{\|\Delta y\|_\infty}{\|y\|_\infty} \leq \frac{\|x\|_\infty \|\partial f/\partial x\|_\infty}{\|f(x)\|_\infty} \cdot \frac{\|\Delta x\|_\infty}{\|x\|_\infty} .$$

Iako je ovo nejednakost, ona je tačna u smislu da jednakost može biti dostignuta za neku podesnu promenu Δx. Tako, možemo definisati globalni faktor uslovljenosti sa

$$(\operatorname{cond} f)(x) := \frac{\|x\|_\infty \|\partial f/\partial x\|_\infty}{\|f(x)\|_\infty} . \tag{1.34}$$

Jasno je da se u slučaju $m = n = 1$, definicija (1.34) redukuje tačno na definiciju (1.27) (kao i na (1.30)) datu ranije. Za veće dimenzije (m i/ili n veće od 1), međutim, faktor uslovljenosti u (1.34) je mnogo grublji nego onaj u (1.30). To možemo objasniti time što norme teže da unište detalje: ako x, na primer, ima komponente sa prilično različitim odstupanjima, onda je $\|x\|_\infty$ naprosto jednaka najvećoj od ovih komponenti uzetih po modulu, dok se sve ostale ignorišu. Zbog toga se zahteva opreznost kod korišćenja (1.34).

Uslovljenost algoritma. Neka je za problem (1.24) dat algoritam A za njegovo rešavanje na računaru, tj. za dati vektor $x \in \mathbb{R}^m(t, s)$ algoritam A daje vektor y_A (u mašinskoj aritmetici) za koji se pretpostavlja da aproksimira $y = f(x)$. Tako, mi sada imamo drugo preslikavanje f_A, koje

opisuje kako je problem f rešen algoritmom A,

$$f_A : \mathbb{R}^m(t,s) \mapsto \mathbb{R}^n(t,s), \quad y_A = f_A(x).$$

Da bismo mogli analizirati f_A u ovim opštim izrazima, moramo formulisati osnovnu pretpostavku, naime,

$$(\forall x \in \mathbb{R}^m(t,s))\,(\exists x_A \in \mathbb{R}^m)\,(f_A(x) = f(x_A)). \tag{1.35}$$

Zapravo, izračunato rešenje koje odgovara nekom ulazu x je tačno rešenje za neki različit ulaz x_A (ne obavezno mašinski vektor i ne obavezno jedinstveno određen) za koji se nadamo da je blizak sa x. Mi, dakle, definišemo faktor uslovljenosti algoritma A pomoću izraza u kome figuriše x_A (najbliži x ako ih ima više od jednog), upoređivanjem njegove relativne greške sa mašinskom preciznošću eps:

$$(\operatorname{cond} A)(x) = \inf_{x_A} \frac{\|x_A - x\|}{\|x\|}/\text{eps}\,. \tag{1.36}$$

Ovde, infimum se uzima preko svih x_A koji zadovoljavaju $f(x) = f(x_A)$. Praktično može se uzeti bilo koje takvo x_A i onda dobiti gornja granica za faktor uslovljenosti:

$$(\operatorname{cond} A)(x) \leq \frac{\|x_A - x\|}{\|x\|}/\text{eps}\,. \tag{1.37}$$

U (1.36), ili (1.37), uzima se ona vektorska norma koja se učini pogodnom za primenu.

Naravno što je x_A bliže x (u smislu odgovarajuće metrike koja proizilazi iz izabrane norme u (1.36)) to će faktor uslovljenosti biti manji, tj. kažemo da je algoritam bolje uslovljen i obrnuto.

Kompjutersko rešenje problema. Totalna greška. Posmatrajmo opet problem (1.24), koji treba da rešimo. To je matematički (idealizovan)

problem, gde su podaci tačni realni brojevi, a rešenje je matematički tačno rešenje.

Kada rešavamo takav problem na računaru, u aritmetici sa pokretnim zarezom sa preciznošću eps, korišćenjem algoritma A, imamo prvo zaokruživanje svih podataka, a zatim primenu na tako zaokružene podatke ne f, već f_A:

$$\begin{aligned} \overline{x} &= \text{zaokruženi podaci}, \qquad \frac{\|\overline{x}-x\|}{\|x\|} = \varepsilon\,, \\ \overline{y}_A &= f_A(\overline{x})\,. \end{aligned}$$

Ovde je ε greška zaokruživanja podataka. (Izvor ove greške može nastati ne samo zaokruživanjem, već, na primer, i prilikom merenja). Totalna greška koju mi želimo da ocenimo je onda

$$\frac{\|\overline{y}_A - y\|}{\|y\|}\,.$$

Korišćenjem osnovne pretpostavke (1.35) nametnute algoritmu A, i biranjem optimalnog $\overline{x}_A$, imamo

$$f_A(\overline{x}) = f(\overline{x}_A), \quad \frac{\|\overline{x}_A - \overline{x}\|}{\|\overline{x}\|} = (\operatorname{cond} A)(\overline{x}) \cdot \operatorname{eps}\,. \tag{1.38}$$

Neka $\overline{y} = f(\overline{x})$. Onda, korišćenjem nejednakosti trougla, imamo

$$\frac{\|\overline{y}_A - y\|}{\|y\|} \le \frac{\|\overline{y}_A - \overline{y}\|}{\|y\|} + \frac{\|\overline{y} - y\|}{\|y\|} \approx \frac{\|\overline{y}_A - \overline{y}\|}{\|\overline{y}\|} + \frac{\|\overline{y} - y\|}{\|y\|}\,,$$

gde smo iskoristili aproksimaciju $\|y\| \approx \|\overline{y}\|$. Korišćenjem (1.38), imamo za prvi izraz na desnoj strani u prethodnoj nejednakosti

$$\begin{aligned} \frac{\|\overline{y}_A - \overline{y}\|}{\|\overline{y}\|} &= \frac{\|f_A(\overline{x}) - f(\overline{x})\|}{\|f(\overline{x})\|} = \frac{\|f(\overline{x}_A) - f(\overline{x})\|}{\|f(\overline{x})\|} \\ &\le (\operatorname{cond} f)(\overline{x}) \cdot \frac{\|\overline{x}_A - \overline{x}\|}{\|\overline{x}\|} \\ &= (\operatorname{cond} f)(\overline{x}) \cdot (\operatorname{cond} A)(\overline{x}) \cdot \operatorname{eps}\,. \end{aligned}$$

Za drugi izraz imamo

$$\frac{\|\overline{y}-y\|}{\|y\|}=\frac{\|f(\overline{x})-f(x)\|}{\|f(x)\|}\leq(\operatorname{cond}f)(x)\cdot\frac{\|\overline{x}-x\|}{\|x\|}=(\operatorname{cond}f)(x)\cdot\varepsilon\,.$$

Pretpostavljajući, konačno, da $(\operatorname{cond}f)(\overline{x})\approx(\operatorname{cond}f)(x)$, dobijamo

$$\frac{\|\overline{y}_A-y\|}{\|y\|}\leq(\operatorname{cond}f)(x)[\varepsilon+(\operatorname{cond}A)(\overline{x})\cdot\operatorname{eps}]\,. \tag{1.39}$$

Formula (1.39) pokazuje koliko greške u ulaznim podacima (ε) i mašinska preciznost (eps) doprinose totalnoj greški: obe su uvećane uslovljenošću problema, dok je druga uvećana i uslovljenošću algoritma.

Primer 1.14. Data je algebarska jednačina

$$x^n+x^{n-1}-a=0,\quad a>0,\quad n\geq 2. \tag{1.40}$$

a) Pokazaćemo da postoji tačno jedan pozitivan koren $\xi(a)$ jednačine (1.40). Neka je $p(x)=x^n+x^{n-1}-a$. Tada je $p'(x)=nx^{n-1}+(n-1)x^{n-2}=x^{n-2}(nx+n-1)>0$ za $x>0$. S obzirom da je je $p(0)=-a<0$, $p(+\infty)>0$, postoji tačno jedan pozitivan koren jednačine (1.40).

b) Pokažimo sada da je koren $\xi(a)$ dobro–uslovljen kao funkcija od a. Mi imamo

$$[\xi(a)]^n+[\xi(a)]^{n-1}-a\equiv 0.$$

Diferenciranjem dobijamo

$$n[\xi(a)]^{n-1}\xi'(a)+(n-1)[\xi(a)]^{n-2}\xi'(a)-1=0,$$

gde je

$$\begin{aligned}\xi'(a) &= \frac{1}{n[\xi(a)]^{n-1}+(n-1)[\xi(a)]^{n-2}}=\frac{\xi(a)}{n[\xi(a)]^n+(n-1)[\xi(a)]^{n-1}}\\ &= \frac{\xi(a)}{n[\xi(a)]^n+(n-1)\left(a-[\xi(a)]^n\right)}\\ &= \frac{\xi(a)}{(n-1)a+[\xi(a)]^n}\,.\end{aligned}$$

Dakle (videti (1.27)),

$$\begin{aligned}(\operatorname{cond}\xi)(a) &= \left|\frac{a\xi'(a)}{\xi(a)}\right| = \frac{a}{(n-1)a+[\xi(a)]^n} \\ &= \frac{1}{n-1+\dfrac{[\xi(a)]^n}{a}} < \frac{1}{n-1} \le 1\,.\end{aligned}$$

PRIMER 1.15. U teoriji Fourierovih redova brojevi

$$\lambda_n = \frac{1}{2n+1} + \frac{2}{\pi}\sum_{k=1}^{n}\frac{1}{k}\tan\frac{k\pi}{2n+1}, \quad n = 1,2,3,\dots\ , \tag{1.41}$$

su poznati kao Lebesgueove konstante.

a) Pokažimo da izrazi pod sumom monotono rastu po k i ispitajmo kako se ti izrazi ponašaju kada je k blisko n, a n je veliko.

Neka je $x = k\pi/(2n+1)$, tako da je $0 < x < \pi/2$ za $1 \le k \le n$. Onda je, do na konstantni faktor, opšti član sume

$$f(x) = \frac{1}{x}\tan x.$$

Pokažimo da f monotono raste. Imamo

$$[xf(x)]' = \frac{1}{\cos x^2}\,,$$

pa je

$$\begin{aligned}xf'(x) &= \frac{1}{\cos^2 x} - f(x) = \frac{1}{\cos^2 x} - \frac{\sin x}{x\cos x} \\ &= \frac{1}{\cos^2 x}\left(1 - \frac{1}{x}\cdot\sin x\cos x\right) \\ &= \frac{1}{\cos^2 x}\left(-\frac{\sin 2x}{2x}\right) > 0\,.\end{aligned}$$

Dakle, izraz pod sumom monotono raste. Za n vrlo veliko, na primer $n = 10^5$, najveći broj sabiraka sume je zanemarljivo mali, izuzev nekoliko njih kod kojih se indeks sume (k) približava vrednosti n, pa oni naglo rastu ka

maksimalnoj vrednosti $\approx 4/\pi$. To može biti pokazano stavljanjem $k = n - r$ za neki fiksirani (mali) prirodan broj r i veliko n. Tada imamo

$$\frac{n-r}{2n+1} = \frac{1}{2} - \frac{2r+1}{2(2n+1)},$$

i, kada $n \to \infty$,

$$\tan\frac{(n-r)\pi}{2n+1} = \tan\left(\frac{\pi}{2} - \frac{\pi}{2}\frac{2r+1}{2n+1}\right) = \frac{\cos\left(\frac{\pi}{2}\frac{2r+1}{2n+1}\right)}{\sin\left(\frac{\pi}{2}\frac{2r+1}{2n+1}\right)} \sim \frac{4}{\pi}\frac{n}{2r+1}.$$

Dakle,

$$\frac{1}{n-r}\tan\frac{(n-r)\pi}{2n+1} \sim \frac{4}{\pi}\frac{1}{2r+1}, \quad \text{kada} \quad n \to \infty.$$

b) Izračunajmo λ_n za $n = 1, 10, 10^2, \ldots, 10^5$ u FORTRAN-u u "single" i "double precision" aritmetici i uporedimo rezultate. Objasnimo šta se primećuje.

Dobijeni su sledeći rezultati:

n	Lebesgue	Lebesgue double
1	$0.1435991 \cdot 10$	$0.1435991124 \cdot 10$
10	$0.2223358 \cdot 10$	$0.2223356924 \cdot 10$
100	$0.3138774 \cdot 10$	$0.3138780093 \cdot 10$
1000	$0.4070242 \cdot 10$	$0.4070163604 \cdot 10$
10000	$0.5003402 \cdot 10$	$0.5003183862 \cdot 10$
100000	$0.5939833 \cdot 10$	$0.5936368213 \cdot 10$

Zbog ponašanja izraza pod sumom, kada je n veliko, tačnost sume je uveliko određena tačnošću sabiraka u kojima je k veoma blisko n. No, u tim slučajevima, argument tangensa je vrlo blizak $\pi/2$. S obzirom da je (videti (1.27))

$$(\text{cond } \tan)(x) = \frac{x(1+\tan^2 x)}{\tan x}, \quad 0 < x < \pi/2,$$

to je tangens veoma slabo uslovljen za x blisko $\pi/2$. Zaista, ako je $\varepsilon > 0$ veoma malo, tada je

$$(\text{cond tan})\left(\frac{\pi}{2}-\varepsilon\right) \sim \frac{\pi}{2}\tan\left(\frac{\pi}{2}-\varepsilon\right) = \frac{\pi}{2}\frac{\cos\varepsilon}{\sin\varepsilon} \sim \frac{\pi}{2\varepsilon}\,.$$

S obzirom da $k = n$ odgovara $\varepsilon = \dfrac{\pi}{2(2n+1)} \sim \dfrac{\pi}{4n}$, važi

$$(\text{cond tan})\left(\frac{\pi}{2}-\varepsilon\right) \sim \frac{\pi}{2\pi/(4n)} = 2n, \quad n \to \infty\,.$$

Tako, na primer, za $n = 10^5$, možemo očekivati gubitak od oko pet decimalnih cifara. To je potvrđeno dobijenim numeričkim rezultatima prethodno prikazanim.

Uočena netačnost ne može biti pripisana samo velikoj količini izračunavanja, tj. nagomilavanju greške zaokruživanja međurezultata u procesu izračunavanja na računskoj mašini. Zapravo, da smo sumu iz (1.41) uzeli, na primer, od $k = 1$ do $k = [n/2]$ izbegli bismo slabu–uslovljenost tangensa i, čak za $n = 10^5$, dobili tačnije rezultate u običnoj ("single") tačnosti. To pokazuju sledeći numerički rezultati (suma iz (1.41) uzeta od $k = 1$ do $k = [n/2]$):

n	Lebesgue	Lebesgue double
1	0.3333333	0.3333333333
10	0.5706023	0.5706023118
100	0.5436349	0.5436349731
1000	0.5407878	0.5407873971
10000	0.5405016	0.5405010908
100000	0.5404736	0.5404724446

Primer 1.16. Izračunajmo

$$I_n = \int_0^1 \frac{t^n}{t+5}\,dt$$

za neki fiksirani prirodan broj n.

Za $n = 0$ imamo

$$I_0 = \int_0^1 \frac{dt}{t+5} = \log(t+5)|_0^1 = \log\frac{6}{5}\,. \tag{1.42}$$

Da bi smo našli rekurzionu formulu za određivanje traženog integrala, uočimo da

$$\frac{t}{t+5} = 1 - \frac{5}{t+5}\,.$$

Množenjem obe strane sa t^{k-1} i integracijom od 0 do 1 dobijamo

$$I_k = -5I_{k-1} + \frac{1}{k}, \quad k = 1, \ldots, n. \tag{1.43}$$

Dakle shema za izračunavanje I_n bi se mogla ovako definisati: Startujući sa I_0 koje je dato sa (1.42), a onda sukcesivno primenjujući (1.43) za $k = 1, 2, \ldots, n$, dobijamo I_n.

Rekurzija (1.43), za bilo koju početnu vrednost I_0, definiše funkciju

$$I_n = f_n(I_0). \tag{1.44}$$

Tako smo dobili problem $f_n : \mathbb{R} \mapsto \mathbb{R}$ (n je parametar) koji možemo prikazati grafički na sledeći način:

$$I_0 \longrightarrow \boxed{f_n} \longrightarrow I_n\,.$$

Sl. 1.1.

Mi smo zainteresovani za uslovljenost preslikavanja f_n u tački I_0. Zaista, s obzirom da broj I_0 iz (1.42) nije mašinski reprezentabilan, to mora biti zaokružen na $\overline{I}_0$ pre startovanja rekurzionog procesa (1.43). Čak i kada ne bi bilo unošenja novih grešaka tokom rekurzije (1.43), konačni rezultat neće biti tačno I_n već neka aproksimacija $\overline{I}_n = f_n(\overline{I}_0)$, i mi imamo

$$\left|\frac{\overline{I}_n - I_n}{I_n}\right| = (\operatorname{cond} f_n)(I_0)\,\left|\frac{\overline{I}_0 - I_0}{I_0}\right|\,. \tag{1.45}$$

Ovde važi jednakost s obzirom na linearnost funkcije f_n po I_0, kako je to napomenuto posle (1.26). Zaista, ako je $n = 1$, onda

$$I_1 = f_1(I_0) = -5I_0 + 1\,.$$

Ako je $n = 2$, onda

$$I_2 = f_2(I_0) = -5I_1 + \frac{1}{2} = (-5)^2 I_0 - 5 + \frac{1}{2},$$

itd. Uopšte, imamo

$$I_n = f_n(I_0) = (-5)^n I_0 + p_n,$$

gde je p_n neki broj (nezavisan od I_0). Sledi da

$$(\mathrm{cond}\, f_n)(I_0) = \left|\frac{I_0 f_n'(I_0)}{I_n}\right| = \left|\frac{I_0(-5)^n}{I_n}\right| = \frac{I_0 \cdot 5^n}{I_n}. \tag{1.46}$$

Iz definicije I_n kao integrala jasno je da I_n opada monotono po n (zapravo konvergira monotono ka 0 kada $n \to \infty$), pa dakle, vidimo da je $f_n(I_0)$ slabo–uslovljeno u I_0 i to sve više što je n veće.

Uočavamo da do stalnog uvećavanja greške u procecu izračunavanja, pomoću rekurzione formule (1.43), dolazi usled množenja sa (-5) u svakom koraku izračunavanja.

Kako možemo izbeći ovu slabu–uslovljenost? Rešenje nalazimo u zapažanju da umesto da množimo velikim brojem bolje bi bilo da delimo velikim brojem, pogotovu ako dobijamo veće rezultate u isto vreme. To se izvodi izračunavanjem unazad u formuli (1.43), tj. biranjem nekog $\nu > n$ i izračunavanjem po formuli

$$I_{k-1} = \frac{1}{5}\left(\frac{1}{k} - I_k\right), \quad k = \nu, \nu - 1, \ldots, n + 1.$$

Problem je tada, naravno, kako izračunati početnu vrednost I_ν.

Pre nego se pozabavimo sa tim, primetimo da sada imamo novi problem $g_n : \mathbb{R} \mapsto \mathbb{R}$ (n je parametar $< \nu$) koji možemo prikazati grafički na sledeći način:

$$I_\nu \longrightarrow \boxed{g_n} \longrightarrow I_n \,.$$

Sl. 1.2.

Kao i gore, razmatramo funkciju g_n kao linearnu funkciju od I_ν, i na sličan način kako smo došli do (1.46), zaključujemo

$$(\operatorname{cond} g_n)(I_\nu) = \left| \frac{I_\nu \left(-\frac{1}{5}\right)^{\nu-n}}{I_n} \right| , \qquad \nu > n.$$

Opet, na osnovu monotonosti za I_n, dobijamo

$$(\operatorname{cond} g_n)(I_\nu) < \left(\frac{1}{5}\right)^{\nu-n} , \quad \nu > n.$$

Po analogiji sa (1.45), sada imamo

$$\left| \frac{\overline{I}_n - I_n}{I_n} \right| = (\operatorname{cond} g_n)(I_\nu) \left| \frac{\overline{I}_\nu - I_\nu}{I_\nu} \right| < \left(\frac{1}{5}\right)^{\nu-n} \left| \frac{\overline{I}_\nu - I_\nu}{I_\nu} \right| , \tag{1.47}$$

gde je $\overline{I}_\nu$ neka aproksimacija od I_ν. Zapravo, $\overline{I}_\nu$ čak ne mora biti blizu I_ν da bi važilo (1.47), s obzirom da je funkcija g_n linearna po I_ν. Tako, mi možemo uzeti početnu vrednost sa 100% relativnom greškom, tj. $\overline{I}_\nu = 0$, da bismo dobili $\overline{I}_n$ sa relativnom greškom

$$\left| \frac{\overline{I}_n - I_n}{I_n} \right| < \left(\frac{1}{5}\right)^{\nu-n} , \quad \nu > n.$$

Granica sa desne strane može da se učini proizvoljno malom, na primer, $\leq \varepsilon$, ako izaberemo ν dovoljno velikim, tj.

$$\nu \geq n + \frac{\log \dfrac{1}{\varepsilon}}{\log 5} \,. \tag{1.48}$$

Dakle, konačna procedura je:

Dati opisanu relativnu tačnost ε. Izabrati ν tako da bude najmanji prirodan broj koji zadovoljava (1.48), a onda računati

$$\begin{aligned} \overline{I}_{\nu} &= 0, \\ \overline{I}_{k-1} &= \frac{1}{5}\left(\frac{1}{k} - \overline{I}_k\right), \quad k = \nu, \nu-1, \ldots, n+1. \end{aligned} \tag{1.49}$$

To će biti procedura za određivanje $\overline{I}_n$ koje dovoljno tačno aproksimira I_n, čak će prisutne greške zaokruživanja tokom izvršavanja (1.49) biti stalno smanjivane.

Primer 1.17. Ispitatajmo uslovljenost algoritma za množenje n realnih brojeva koji su zadati tačno i mašinski su reprezentabilni, na računskoj mašini (računaru).

Neka su x_i $(i = 1, \ldots, n)$, brojevi koje treba pomnožiti. Označimo $x = [x_1 \quad x_2 \quad \ldots \quad x_n]^\top \in \mathbb{R}^n$.

Matematički posmatrano (sva izračunavanja se izvode apsolutno tačno), imamo problem koji bi se mogao interpretirati kao preslikavanje

$$f : \mathbb{R}^n \mapsto \mathbb{R}, \quad y = f(x) = x_1 x_2 \cdots x_n. \tag{1.50}$$

Ono bi se moglo, na primer, ovako realizovati

$$\begin{aligned} A: \quad & p_1 = x_1, \\ & p_k = x_k p_{k-1}, \quad k = 2, 3, \ldots, n, \\ & y = p_n\,. \end{aligned} \tag{1.51}$$

Pri izračunavanju na računaru, po istom algoritmu (1.51), situacija je nešto drugačija. Prema uslovu u zadatku, brojevi x_i $(i = 1, \ldots, n)$ su mašinski reprezentabilni brojevi, tj. $x_i \in \mathbb{R}(t, s) \quad (i = 1, \ldots, n)$. Međutim, s obzirom na konačnost broja (t) cifara mantise svakog broja u računaru, posle svake operacije množenja javlja se odgovarajuća mašinska greška (kao posledice zaokruživanja rezultata na t cifara mantise). Ove mašinske greške označimo sa r_i $(i = 2, \ldots, n)$ i neka je $|r_i| \leq \text{eps}$.

Dakle, korišćenjem istog algoritma (1.51), nećemo imati preslikavanje f već preslikavanje f_A, koje je definisano sa

$$f_A : \mathbb{R}^n(t,s) \mapsto \mathbb{R}(t,s), \quad y_A = f_A(x)$$

tj. primenom algoritma (1.51), mi umesto p_i $(i = 2, \ldots, n)$, dobijamo $\overline{p}_i$ $(i = 2, \ldots, n)$, a umesto y dobijamo y_A, pri čemu je (ovde koristimo oznaku $\odot$ za množenje kako bismo naznačili da se ono izvršava na računaru, posle čega imamo pojavu mašinske greške):

$$\begin{aligned}
p_1 &= x_1, \\
\overline{p}_2 &= x_2 \odot p_1 = x_2 p_1 (1 + r_2) = x_2 x_1 (1 + r_2), \\
\overline{p}_3 &= x_3 \odot \overline{p}_2 = x_3 \overline{p}_2 (1 + r_3) = x_3 x_2 x_1 (1 + r_2)(1 + r_3), \\
&\vdots \\
\overline{p}_n &= x_n \odot \overline{p}_{n-1} = x_n \overline{p}_{n-1} (1 + r_n) \\
&= x_n x_{n-1} \cdots x_1 (1 + r_2)(1 + r_3) \cdots (1 + r_n), \\
y_A &= \overline{p}_n.
\end{aligned}$$

U smislu (1.35) možemo uzeti, na primer (ne postoji jedinstvenost), da je

$$x_A = \left[x_1 \quad x_2(1 + r_2) \quad \ldots \quad x_n(1 + r_n) \right]^\top,$$

pri čemu je $f_A(x) = f(x_A)$.

Korišćenjem $\| \cdot \|_\infty$ norme, imamo

$$\frac{\|x_A - x\|_\infty}{\|x\|_\infty \cdot \text{eps}} = \frac{\|[0 \; x_2 r_2 \; \ldots \; x_n r_n]^\top\|_\infty}{\|x\|_\infty \cdot \text{eps}} \leq \frac{\|x\|_\infty \cdot \text{eps}}{\|x\|_\infty \cdot \text{eps}} = 1\,,$$

i tako, pomoću (1.37), $(\text{cond}\, A)(x) \leq 1$ za svako $x \in \mathbb{R}^n(t,s)$. Dakle, kako se i očekivalo, ovaj algoritam je perfektno dobro–uslovljen.

Primer 1.18. Ispitajmo uslovljenost algoritma za računanje skalarnog proizvoda dva vektora $x, y \in \mathbb{R}^m$, čiji su elementi mašinski reprezentabilni, na računaru.

Ako vrednost izraza

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

računamo kao

$$\begin{aligned}
&\bar{s} = x_1 \odot y_1; \\
&\bar{s} = \bar{s} \oplus x_i \odot y_i; \quad i = 2, 3, \ldots, m\,,
\end{aligned}$$

dobijamo, npr. za $m = 4$, izraz oblika

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

Primenom osnovnih osobina aritmetike na računaru, lako se proveri da važi

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

gde su vrednosti ϵ_i, ξ_i po modulu manje od eps. Sada je jasno kako bi izgledala formula za proizvoljan broj od m sumanada. Primetimo da $1+\zeta_k$ možemo oceniti sa

$$1 - m\,\mathrm{eps} \le 1+\zeta_k \le \frac{1}{1-m\,\mathrm{eps}}, \quad \text{tj. važi} \quad |\zeta_k| \le \frac{m\,\mathrm{eps}}{1-m\,\mathrm{eps}} \quad (k = 1, \ldots, m).$$

Izračunati skalarni proizvod $\bar{s}$ je tačan skalaran proizvod malo perturbovanih vektora x i y. Kao perturbovane vektore možemo uzeti, npr.

$$x + \delta x = [x_1 + \zeta_1, x_2 + \zeta_2, \ldots, x_m + \zeta_m]^T \text{ i } y\,.$$

Zaključujemo da je posmatrani algoritam dobro uslovljen, za razliku od problema koji rešava. Naime, iz

$$|\bar{s}_m - s_m| \leq \frac{m\,\mathrm{eps}}{1 - m\,\mathrm{eps}}|x|^T|y|\,,$$

vidimo da relativna greška izračunatog $\bar{s}_m$ može biti veoma velika ako je $|x^T y| << |x|^T|y|$.

Jasno je da iste zaključke možemo preneti i na množenje matrica.

Glava 2

Sistemi linearnih jednačina

Neka su zadani matrica $A \in \mathbb{R}^{m\times n}$ i vektor $b \in \mathbb{R}^m$. Odgovor na pitanje kad linearni sistem $Ax = b$ ima rešenje $x \in \mathbb{R}^n$ i kad je ono jedinstveno daje Kronecker-Capellijeva teorema. Posebna pažnja posvećuje se sistemima kad je matrica A regularna i kvadratna, s obzirom da se oni najčešće javljaju u praksi.

U ovom poglavlju istraživaćemo metode za rešavanje regularnih, kvadratnih sistema

$$\begin{array}{ccccccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1j}x_j & + & \dots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2j}x_j & + & \dots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & & & & \vdots & & \vdots \\ a_{i1}x_1 & + & a_{i2}x_2 & + & \dots & + & a_{ij}x_j & + & \dots & + & a_{in}x_n & = & b_i \\ \vdots & & \vdots & & & & \vdots & & & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \dots & + & a_{nj}x_j & + & \dots & + & a_{nn}x_n & = & b_n. \end{array} \tag{2.1}$$

Matrica $A = [a_{ij}]_{i,j=1}^n \in \mathbb{R}^{n\times n}$ je matrica sistema, njeni elementi su koeficijenti sistemi, a vektor $b = [b_i]_{i=1}^n \in \mathbb{R}^n$ je vektor desne strane sistema. Rešiti sistem znači odrediti vektor nepoznatih $x = [x_i]_{i=1}^n \in \mathbb{R}^n$ tako da važi $Ax = b$.

Za teorijsku matematiku je rešavanje ovog sistema veoma jednostavan problem. Rešenje x je dato formulom $x = A^{-1}b$, gde je A^{-1} inverzna matrica od A. Navedenom formulom je data jedna metoda za rešavanje sistema (2.1) koja nije praktična jer je zadani problem zamenjen novim, težim proble-

mom računanja inverzne matrice. Naime, j-ta kolona matrice A^{-1} je rešenje sistema $Ax = e_j$.

Do rešenja sistema se može doći i upotrebom Cramerovih formula gde je potrebno izvršiti $O(n^2 n!)$ elementarnih operacija ili Gaussovom metodom u $O(n^3)$ elementarnih operacija. Vidimo da je upotreba Cramerovih formula dosta zahtevnija.

U primenjenoj matematici je situacija puno komplikovanija. Rešiti problem znači biti u stanju u konkretnoj situaciji doći do dovoljno dobre numeričke aproksimacije rešenja koristeći računar. Zato se mora obratiti posebna pažnja na mogućnosti računara, bez obzira na brzinu kojom se menjaju. Iako današnji računari imaju fascinantnu brzinu (dobar jednoprocesorski računar može izvesti 10^9 operacija u sekundi), ona ne može odgovoriti problemima sve većih dimenzija koje nameće moderna nauka. Na primer, broj operacija potrebnih za rešavanja sistem (2.1) Gaussovom metodom za $n = 10^5$ je reda veličine 10^{15}, pa bi brzinom od 10^9 operacija u sekundi bilo potrebno 10^6 sekundi za rešavanje problema, što je više od 11 dana.[1] Kad još uzmemo u obzir mogućnost da je svaka od $O(n^3)$ operacija u Gaussovom algoritmu izvršena s greškom zaokruživanja, postavlja se pitanje koliko je tačno dobijeno rešenje.

2.1 Gaussove eliminacije

Metoda Gaussovih eliminacija je najstariji, najjednostavniji i najpoznatiji algoritam za rešavanje linearnog sistema $Ax = b$. Sastoji se od sukcesivne primene elementarnih transformacija (zamena poretka jednačina, množenje jednačine brojem različitim od 0 i sabiranje jednačina), s ciljem svođenja

[1] Situacija je još drastičnija ako se koriste Cramerove formule, pa se one nikad ne koriste kao metoda numeričkog rešavanja.

sistema na trougaoni oblik koji se jednostavno rešava. Algoritam ima veoma zanimljivu strukturu i može se interpretirati kao faktorizacija matrice A na proizvod trougaonih matrica.

Krenimo od sistema (2.1) i pretpostavimo da je $a_{11} \neq 0$. Ako prvu jednačinu pomnožimo sa $m_{i1} = -\frac{a_{i1}}{a_{11}}$ i dodamo i-toj jednačini, $i = 2, 3, \ldots, n$, dobijamo sistem

$$\begin{array}{ccccccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1j}x_j & + & \ldots & + & a_{1n}x_n & = & b_1 \\ & & a_{22}^{(1)}x_2 & + & \ldots & + & a_{2j}^{(1)}x_j & + & \ldots & + & a_{2n}^{(1)}x_n & = & b_2^{(1)} \\ & & \vdots & & & & \vdots & & & & \vdots & & \vdots \\ & & a_{i2}^{(1)}x_2 & + & \ldots & + & a_{ij}^{(1)}x_j & + & \ldots & + & a_{in}^{(1)}x_n & = & b_i^{(1)} \\ & & \vdots & & & & \vdots & & & & \vdots & & \vdots \\ & & a_{n2}^{(1)}x_2 & + & \ldots & + & a_{nj}^{(1)}x_j & + & \ldots & + & a_{nn}^{(1)}x_n & = & b_n^{(1)}, \end{array} \tag{2.2}$$

gde je $a_{ij}^{(1)} = a_{ij} - m_{i1}a_{1j}$ i $b_i^{(1)} = b_i - m_{i1}b_1$. Sistem (2.2) se može zapisati u obliku

$$M_1\, A\, x = M_1\, b,$$

gde je

$$M_1 = I - m^{(1)}e_j^T = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ -m_{21} & 1 & 0 & 0 & \ldots & 0 \\ -m_{31} & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ -m_{n1} & 0 & 0 & 0 & \ldots & 1 \end{bmatrix}$$

i $m^{(1)} = [0, m_{21}, m_{31}, \ldots, m_{n1}]^T$. Nakom prvog koraka sistem sadrži nepoznatu x_1 samo u prvoj jednačini.

U drugom koraku se vrši isključivanje nepoznate x_2 iz svih jednačina osim prve i druge. Ako je $a_{22}^{(1)} \neq 0$, onda drugu jednačinu množimo sa $m_{i2} = -\frac{a_{i2}^{(1)}}{a_{22}^{(1)}}$

i dodamo i-toj jednačini, $i = 3, 4, \ldots, n$ i dobijamo sistem

$$\begin{array}{ccccccccccccc}
a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & a_{14}x_j & + & \ldots & + & a_{1n}x_n & = & b_1 \\
 & & a_{22}^{(1)}x_2 & + & a_{23}^{(1)}x_3 & + & a_{24}^{(1)}x_j & + & \ldots & + & a_{2n}^{(1)}x_n & = & b_2^{(1)} \\
 & & & & a_{33}^{(2)}x_3 & + & a_{34}^{(2)}x_j & + & \ldots & + & a_{3n}^{(2)}x_n & = & b_2^{(2)} \\
 & & & & \vdots & & \vdots & & & & \vdots & & \vdots \\
 & & & & a_{i3}^{(2)}x_3 & + & a_{i4}^{(2)}x_j & + & \ldots & + & a_{in}^{(2)}x_n & = & b_i^{(2)} \\
 & & & & \vdots & & \vdots & & & & \vdots & & \vdots \\
 & & & & a_{n3}^{(2)}x_3 & + & a_{n4}^{(2)}x_j & + & \ldots & + & a_{nn}^{(2)}x_n & = & b_n^{(2)},
\end{array}$$

gde je $a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i2}a_{2j}^{(1)}$ i $b_i^{(2)} = b_i^{(1)} - m_{i2}b_2^{(1)}$. Novi sistem se može zapisati u obliku

$$M_2\, M_1\, A\, x = M_2\, M_1\, b,$$

gde je

$$M_2 = I - m^{(2)}e_2^T = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & -m_{32} & 1 & 0 & \ldots & 0 \\ 0 & -m_{42} & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & -m_{n2} & 0 & 0 & \ldots & 1 \end{bmatrix}$$

i $m^{(2)} = [0, 0, m_{32}, m_{42}, \ldots, m_{n2}]^T$.

Nastavljajući ovaj postupak, ako su svi $a_{ii}^{(i-1)} \neq 0$, $i = 1, 2, \ldots, n-1$, dobijamo trougaoni sistem koji je ekvivalentan polaznom

$$\begin{array}{ccccccccc}
a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & b_1 \\
 & & a_{22}^{(1)}x_2 & + & \ldots & + & a_{2n}^{(1)}x_n & = & b_2^{(1)} \\
 & & & & \ddots & & \vdots & & \vdots \\
 & & & & & & a_{nn}^{(n-1)}x_n & = & b_n^{(n-1)},
\end{array} \tag{2.3}$$

a u matričnom obliku

$$M_{n-1}\, M_{n-2} \ldots M_2\, M_1\, A\, x = M_{n-1}\, M_{n-2} \ldots M_2\, M_1\, b\,,$$

gde je, za $i = k+1, \ldots, n$,

$$M_k = I - m^{(k)}e_k^T\,, \quad m^{(k)} = [0, \ldots, 0, m_{k+1,k}, \ldots, m_{nk}]^T\,, \quad m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}\,.$$

Označimo matricu sistema (2.3) sa U, tj. $U = (M_{n-1}\, M_{n-2} \ldots M_2\, M_1)\, A$. S obzirom na strukturu matrica M_k važi

$$\begin{aligned} L &:= (M_{n-1}\, M_{n-2} \ldots M_2\, M_1)^{-1} = M_1^{-1}\, M_2^{-1} \ldots M_{n-1}^{-1} \\ &= (I + m^{(1)} e_1^T)(I + m^{(2)} e_2^T) \cdots (I + m^{(n-1)} e_{n-1}^T) = I + \sum_{i=1}^{n-1} m^{(i)} e_i^T \,, \end{aligned}$$

odakle zaključujemo da je matrica L donja trougaona sa jedinicama na dijagonali. Dakle, $A = L\,U$, tj. matrica A se može predstaviti kao proizvod jedne donje i jedne gornje trougaone matrice.

Primer 2.1. Data je matrica

$$A = \begin{bmatrix} 1 & 2 & -1 & 1 \\ 2 & 5 & -1 & 2 \\ 3 & -1 & -2 & 1 \\ 1 & -1 & 3 & -5 \end{bmatrix}.$$

U prvom koraku odredimo matricu M_1 i pomnožimo je sa A

$$M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -3 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}, \quad M_1\, A = \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & -7 & 1 & -2 \\ 0 & -3 & 4 & -6 \end{bmatrix}.$$

Sada biramo M_2 i pravimo proizvod $M_2\, M_1\, A$

$$M_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 7 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}, \quad M_2\, M_1\, A = \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 8 & -2 \\ 0 & 0 & 7 & -6 \end{bmatrix}.$$

U zadnjem koraku je

$$M_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -7/8 & 1 \end{bmatrix}, \quad M_3\, M_2\, M_1\, A = \begin{bmatrix} 1 & 2 & -1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 8 & -2 \\ 0 & 0 & 0 & -17/4 \end{bmatrix} = U\,.$$

Da bi se dobile matrice M_i^{-1} potrebno je samo promeniti predznake elementima ispod dijagonale

$$M_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \; M_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -7 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{bmatrix}, \; M_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 7/8 & 1 \end{bmatrix},$$

a da bi se dobio njihov proizvod treba samo prepisati odgovarajuće kolone

$$M_1^{-1}\, M_2^{-1}\, M_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & -7 & 1 & 0 \\ 1 & -3 & 7/8 & 1 \end{bmatrix} = L.$$

Nameće se pitanje moraju li $a_{ii}^{(i-1)}$, $i = 1, \ldots, n-1$, biti različiti od nule ako je matrica A kvadratna i regularna. Jasno je da ne moraju. Matrica sistema

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

je regularna, ali ga ipak ne možemo rešiti Gaussovim eliminacijama bez menjanja poretka jednačina. U svakom k-tom koraku eliminacija u matrici $M_{k-1} \ldots M_1 A$ jedan od elemenata $a_{rk}^{(k-1)}$, $r \geq k$, mora biti različit od nule. Lako se pokazuje da je u suprotnom matrica A singularna. Ukoliko je $a_{kk}^{(k-1)} = 0$ zamenom k-te i r-te jednačine na njegovo mjesto dolazi broj različit od nule i možemo nastaviti sa eliminacijama. Takve nenula elemente koje zamenom vrsta dovodimo na dijagonalu zovemo pivoti. Biranje pivota tako da je on različit od nule je dovoljno da se postupak eliminacije provede do kraja. Međutim, u obzir treba uzeti i druge činjenice.

Strategiju kojom se za pivot u k-tom koraku bira element

$$|a_{rk}^{(k-1)}| = \max_{k \leq i \leq n} |a_{ik}^{(k-1)}|$$

zovemo parcijalno pivotiranje. Motivacija za nju je jednostavna. Elementi koji se računaju u k-tom koraku su

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik} a_{kj}^{(k-1)}, \quad b_i^{(k)} = b_i^{(k-1)} - m_{ik} b_k^{(k-1)}, \quad m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}},$$

za $i, j = k+1, \ldots, n$. Ako je faktor m_{ik} velik, u aritmetici pokretnog zareza može doći do kraćenja najmanje značajnih cifara broja $a_{ij}^{(k-1)}$, pa izračunati $a_{ij}^{(k)}$ može imati veliku relativnu grešku. Da bi se to izbeglo, brojevi m_{ik}

trebaju biti što manji, a to se postiže biranjem što je moguće većeg (po modulu) imenioca, a to je upravo pivot. Kod parcijalnog pivotiranja je $|m_{ik}| \leq 1$, $i = k+1, \ldots, n$. U praksi ono funkcioniše veoma dobro, mada je moguće konstruisati i primere za suprotno.

Pored parcijalnog, može se koristiti i potpuno pivotiranje, kojim se za pivota u k-tom koraku bira element

$$|a_{rs}^{(k-1)}| = \max_{k \leq i,j \leq n} |a_{ij}^{(k-1)}| .$$

Dakle, umesto jednodimenzionalnog, sada koristimo dvodimenzionalno pretraživanje. Da bi izabrani $|a_{rs}^{(k-1)}|$ došao na dijagonalu, osim vrsta r i k, treba zameniti i kolone s i k. Tu treba biti oprezan jer zamena kolona s i k znači zamenu uloga nepoznatih x_s i x_k. Praksa je pokazala da se ovom strategijom u odnosu na prethodnu ne dobija mnogo, pa se s obzirom na plaćenu cenu, prednost daje parcijalnom pivotiranju. Takođe, treba naglasiti da Gaussove eliminacije sa pivotiranjem ne dovode do LU faktorizacije matrice A.

2.2 LU faktorizacija

U praksi se linearni sistemi najčešće rešavaju metodom koja se sastoji od tri koraka:

1. Napisati matricu A u obliku $A = LU$, gde je L donja trougaona, a U gornja trougaona matrica.

2. Rešavanjem donjeg trougaonog sistema $Ly = b$ odrediti vektor y.

3. Rešavanjem gornjeg trougaonog sistema $Ux = y$ odrediti vektor x.

Gaussove eliminacije predstavljaju samo jedan od načina (verovatno najjednostavniji) da se izračuna LU faktorizacija[2]. Preciznije, transformacijom

[2]U ostatku glave ćemo podrazumevati da L ima jediničnu dijagonalu.

sistema (2.1) u trougaoni oblik (2.3) su izvedeni koraci 1 i 2. Međutim, lako ih je razdvojiti: Za računanje matrica M_i, $i = 1, \ldots, n-1$, tačnije njihove i-te kolone, nije bio potreban vektor b. "Ubacivanjem"brojeva $-m_{i,j}$, $j > i$, $i = 1, \ldots, n$ u jediničnu matricu na poziciju (i, j) dobijamo matricu L i prelazimo na korak 2. Ovim se ništa ne dobija, ali ni ne gubi, s obzirom da je broj operacija potrebnih za rešavanje sistema $L\,y = b$ jednak boju operacija potrebnih za izračunavanje proizvoda $L^{-1}\,b$.

Prebrojimo sve elementarne operacije ovog algoritma. U k-tom koraku, $k = 1, \ldots, n-1$ se obavlja:

- $n - k$ deljenja - računanje brojeva $m_{i,k}$
- $(n-k+1)(n-k)$ množenja - po $n-k$ množenja za računanje elemenata matrice $M_k \ldots, M_1\,A$ i jedno množenje za računanje elementa vektora $M_k \ldots, M_1\,b$ u svakoj od $n - k$ vrsta
- $(n-k+1)(n-k)$ oduzimanja (sabiranja) - javljaju se na istim mestima gde i prethodna množenja.

Ukupan broj operacija (za prvi i drugi korak) je

$$\sum_{k=1}^{n-1}[2(n-k)^2 + 3(n-k)] = \sum_{k=1}^{n-1}(2k^2 + 3k) = \frac{1}{6}(4n^3 + 3n^2 - 7n).$$

Trougaoni sistemi se lako rešavaju: donji - supstitucijom unapred,

$$\begin{aligned} y_1 &= b_1 \\ y_i &= b_i - \sum_{j=1}^{i-1} \ell_{ij} y_j, \quad i = 2, \ldots, n\,, \end{aligned}$$

a gornji - supstitucijom unazad

$$\begin{aligned} x_n &= \frac{y_n}{r_{nn}} \\ x_i &= \frac{1}{r_{ii}}\left(y_i - \sum_{j=i+1}^{n} r_{ij} x_j\right), \quad i = n-1, \ldots, 1\,. \end{aligned}$$

Ako je matrica A regularna elementi $r_{11}, r_{22}, \ldots, r_{nn}$ su različiti od nule, pa nema problema sa supstitucijom unazad. U poslednjem algoritmu potrebno je izvršiti n deljenja, $1+2+\ldots+(n-1) = n(n-1)/2$ množenja i $n(n-1)/2$ sabiranja (oduzimanja), što je ukupno n^2 elementarnih operacija.

Dakle, da bi se rešio sistem upotrebom Gaussovih eliminacija potrebno je izvršiti $(4n^3 + 9n^2 - 7n)/6$ elementarnih operacija, što je približno $2n^3/3$, za veće n.

Dalje je cilj odgovoriti na pitanje koje osobine mora imati proizvoljna kvadratna matrica A da bi imala LU faktorizaciju. Izvedivost operacija koje su dovele do svođenja sistema (2.1) na oblik (2.3), odnosno do faktorizacije $A = LU$ je zavisila od uslova

$$a_{11} \neq 0, \quad a_{22}^{(1)} \neq 0, \quad a_{33}^{(2)} \neq 0, \quad \ldots \quad a_{n-1,n-1}^{(n-2)} \neq 0. \tag{2.4}$$

Videli smo da regularnost matrice A nije dovoljna za ove uslove (npr. ako je $a_{11} = 0$).

Uslovi (2.4) su ekvivalentni uslovu da je prvih $n-1$ glavnih minora matrice A različito od nule. Naime, nakon $(k-1)$-og koraka vredi

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1k} \\ a_{21} & a_{22} & \ldots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \ldots & a_{kk} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ m_{21} & 1 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ m_{k1} & m_{k2} & \ldots & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1k} \\ 0 & a_{22}^{(1)} & \ldots & a_{2k}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & a_{kk}^{(k-1)} \end{bmatrix}$$

i $a_{11} \neq 0,\ a_{22}^{(1)} \neq 0,\ \ldots\ a_{k-1,k-1}^{(k-2)} \neq 0$, pa je očigledno k-ti glavni minor matrice A jednak proizvodu $a_{11} \cdot a_{22}^{(1)} \cdots a_{k-1,k-1}^{(k-2)} \cdot a_{kk}^{(k-1)}$. Podmatricu matrice A sa leve strane prethodne jednakosti[3] označavaćemo sa A_k.

Teorema 2.2.1 *Neka je $A \in \mathbb{R}^{n\times n}$ i neka su svi njeni glavni minori osim $\det(A)$ različiti od nula. Tada postoji donja trougaona matrica sa jedinicama*

[3]Isto važi i za proizvoljnu matricu.

na dijagonali L i gornja trouaona matrica U, tako da vredi $A = LU$. Faktorizacija $A = LU$ je jedinstvena ako i samo ako još i $det(A) \neq 0$.

Dokaz. Egzistenciju LU faktorizacije pod navedenim uslovima je već dokazana. Pokazaćemo njenu jedinstvenost ako je A još i regularna. Pretpostavimo da postoje dve takve faktorizacije,

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

Ako je A regularna, onda su i L, L', U i U' takođe regularne i vredi

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

Gornja relacija predstavlja jednakost donje trougaone i gornje trougaone matrice, a to je moguće samo ako su obe dijagonalne. Matrice L i L' imaju jedinice na dijagonali, pa isto važi i za matricu $L^{-1}L'$. Dakle, $L^{-1}L' = I$, tj. $L = L'$. Tada je i $U = U'$.

S druge strane, pretpostavimo da je A regularna i da ima LU faktorizaciju. Tada je $A_k = L_k U_k$, $k = 1, \ldots, n$, pa je $\det(A_k) = u_{11}u_{22}\cdots u_{kk}$. Pošto je $\det(A) = u_{11}u_{22}\cdots u_{nn} \neq 0$, sledi $\det(A_k) \neq 0$. □

Regularne matrice kojima je neki od glavnih minora jednak nuli mogu imati pomenutu faktorizaciju, ali ona nije jedinstvena. Isto važi i za singularne matrice, što pokazuje sledeći primer

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

Na osnovu prethodne teoreme vidimo da svaka singularna matrica čijih je prvih $n-1$ glavnih minora različito od nule[4] ima LU faktorizaciju.

U prethodnom poglavlju smo videli da se Gaussovim eliminacijama sa parcijalnim (potpunim) pivotiranjem svaki kvadratni sistem sa regularnom

[4]Ovo možemo lako uopštiti na matrice kojima je broj prvih glavnih minora različitih od nule jednak rangu matrice.

matricom A može dovesti na trougaoni oblik. Pokazaćemo sada da to važi za svaki kvadratni sistem, s tim što će sada bar jedan element na dijagonali matrice U biti nula ako je A singularna. Nakon $(k-1)$-og koraka imamo matricu

$$M_{k-1}\cdots M_1 A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & a_{2k}^{(1)} & \dots & a_{2n}^{(1)} \\ \vdots & 0 & \ddots & \vdots & \dots & \vdots \\ & \vdots & & a_{kk}^{(k-1)} & \dots & a_{kn}^{(k-1)} \\ & & & a_{k+1,k}^{(k-1)} & \dots & a_{k+1,n}^{(k-1)} \\ & & & \vdots & \dots & \vdots \\ 0 & 0 & & a_{n,k}^{(k-1)} & \dots & a_{nn}^{(k-1)} \end{bmatrix}$$

i određujemo M_k s ciljem da ona poništi elemente k-te kolone prethodne matrice ispod dijagonale. Videli smo šta se dešava ako je $a_{kk}^{(k-1)} \neq 0$. U suprotnom radimo sledeće. Ako su svi elementi k-te kolone ispod dijagonale jednaki nuli, onda nastavljamo dalje, tj. $M_k = I$, jer nam je ionako bio cilj da njih poništimo. Ako je bar jedan od njih različit od nule, biramo najveći (po modulu) i zamenjujemo vrstu r u kojoj se on nalazi sa k-tom vrstom. To postižemo množeći gornju matricu permutacionom matricom $P^{(k,r)}$ (dobijamo je kad u jediničnoj matrici I zamenimo vrste k i r). Nakon toga biramo M_k na poznati način. Na kraju dobijamo

$$U = M_{n-1}\,P_{n-1}\,M_{n-2}\,P_{n-2}\cdots M_2\,P_2\,M_1\,P_1\,A\,, \tag{2.5}$$

gde su P_k neke permutacione matrice. U odnosu na standardni slučaj imamo permutacione matrice između matrica M_j i M_{j+1}, ali to ne smeta. Naime, lako je pokazati da je $P_i\,M_k = M_k^{(1)}\,P_i$, gde $M_k^{(1)}$ ima istu strukturu kao i M_k. Preciznije, $M_k^{(1)}$ se dobije tako što neka dva elementa (koja, zavisi od matrice P_i) u k-toj koloni matrice M_k zamene mesta. Iz opisa procedure se vidi da će to uvek biti dva elementa ispod dijagonale. Sledi, da se svaka od matrica P_i (krećući od P_2, pa redom do P_{n-1}) u (2.5) može pomeriti desno

do pred A, menjajući pri tom svaku matricu M s kojom zameni mesto ali joj ne menjajući strukturu, tj. važi

$$U = M_{n-1}\, M_{n-2}^{(1)} \cdots M_2^{(n-3)}\, M_1^{(n-2)}\; P_{n-1} \cdots P_2\, P_1\, A\,.$$

Proizvod permutacionih matrica je takođe permutaciona matrica, pa, slično kao u prethodnom poglavlju, dobijamo $L\,U = P\,A$, gdje matrica P pokazuje sva učinjena pivotiranja. Demonstrirajmo ovo na sledećem primeru.

Primer 2.2. Najveći (po modulu) element prve kolone matrice A je na poziciji $(3,1)$, pa je prvo množimo sa $P_1 = P^{(1,3)}$

$$A = \begin{bmatrix} -3 & 5 & -11 & -13 \\ 2 & -1 & 4 & 7 \\ 6 & -6 & 12 & 24 \\ 3 & 1 & -2 & -8 \end{bmatrix}, \quad P_1\,A = \begin{bmatrix} 6 & -6 & 12 & 24 \\ 2 & -1 & 4 & 7 \\ -3 & 5 & -11 & -13 \\ 3 & 1 & -2 & -8 \end{bmatrix}.$$

Sada je

$$M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1/3 & 1 & 0 & 0 \\ 1/2 & 0 & 1 & 0 \\ -1/2 & 0 & 0 & 1 \end{bmatrix}, \quad M_1\,P_1\,A = \begin{bmatrix} 6 & -6 & 12 & 24 \\ 0 & 1 & 0 & -1 \\ 0 & 2 & -5 & -1 \\ 0 & 4 & -8 & -4 \end{bmatrix}.$$

U drugoj koloni, od dijagonalnog elementa pa nadole, najveći element je na poziciji $(4,2)$, pa je $P_2 = P^{(2,4)}$. Sada je

$$P_2\,M_1\,P_1\,A = \begin{bmatrix} 6 & -6 & 12 & 24 \\ 0 & 4 & -8 & -4 \\ 0 & 2 & -5 & -1 \\ 0 & 1 & 0 & -1 \end{bmatrix},$$

a zatim

$$M_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1/2 & 1 & 0 \\ 0 & -1/4 & 0 & 1 \end{bmatrix}, \quad M_2\,P_2\,M_1\,P_1\,A = \begin{bmatrix} 6 & -6 & 12 & 24 \\ 0 & 4 & -8 & 4 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 2 & 0 \end{bmatrix}.$$

U trećoj koloni, od dijagonalnog elementa pa nadole, najveći element je na poziciji $(4,3)$, pa je $P_3 = P^{(3,4)}$. Sada je

$$P_3\, M_2\, P_2\, M_1\, P_1\, A = \begin{bmatrix} 6 & -6 & 12 & 24 \\ 0 & 4 & -8 & -4 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix},$$

a zatim

$$M_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1/2 & 1 \end{bmatrix},\ M_3\, P_3\, M_2\, P_2\, M_1\, P_1\, A = \begin{bmatrix} 6 & -6 & 12 & 24 \\ 0 & 4 & -8 & 4 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = U\,.$$

Dalje važi $P_2\, M_1 = M_1^{(1)}\, P_2$, gde $M_1^{(1)}$ dobijamo iz M_1 zamenom elemenata na pozicijama $(2,1)$ i $(4,1)$ jer je $P_2 = P^{(2,4)}$. Dakle,

$$M_3\, P_3\, M_2\, P_2\, M_1\, P_1\, A = M_3\, P_3\, M_2\, M_1^{(1)}\, P_2\, P_1\, A\,.$$

Ostaje još da se na desnoj strani prethodne jednakosti $P_3 = P^{(3,4)}$ pomeri do pred P_2. Pri tome će se promeniti M_2 u $M_2^{(1)}$ (zamenom elemenata na pozicijama $(3,2)$ i $(4,2)$) i $M_1^{(1)}$ u $M_1^{(2)}$ (zamenom elemenata na pozicijama $(3,1)$ i $(4,1)$). Sada je

$$U = M_3\, M_2^{(1)}\, M_1^{(2)}\, P_3\, P_2\, P_1\, A\,,$$

gde je

$$M_2^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1/4 & 1 & 0 \\ 0 & -1/2 & 0 & 1 \end{bmatrix}, \quad M_1^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1/2 & 1 & 0 & 0 \\ -1/3 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 1 \end{bmatrix}.$$

Na kraju dobijamo

$$L = (M_3\, M_2^{(1)}\, M_1^{(2)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 1/3 & 1/4 & 1 & 0 \\ -1/2 & 1/2 & -1/2 & 0 \end{bmatrix},$$

$$P = P_3\, P_2\, P_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

2.3 Numerička svojstva Gaussovih eliminacija

U prethodnim poglavljima, osim na par mesta, nismo izlazili iz okvira linearne algebre. Sada ćemo razmotriti probleme koji se javljaju prilikom realizacije pomenutih algoritama na računaru.

Krenimo od $L\,U$ faktorizacije (bez pivotiranja). Elementi u_{ij} računaju se prema formuli

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

koja je u stvari specijalni slučaj skalarnog proizvoda vektora sa i koordinata. Na osnovu primera 1.18. zaključujemo da postoje ξ_{ij}, ζ_{ijm}, svi manji od $n\,\varepsilon/(1 - n\,\varepsilon)$, takvi da je

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

Prethodnu relaciju možemo zapisati i u sledećem obliku

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

Time smo dokazali sledeću teoremu.

Teorema 2.3.1 *Ako su $\bar{L}$ i $\bar{U}$ trougaoni faktori matrice A dobijeni Gaussovim eliminacijama (bez pivotiranja) primenom aritmetike računara preciznosti ε, onda je*

$$\bar{L}\,\bar{U} = A + \Delta A, \quad |\Delta A| \le \frac{n\varepsilon}{1 - n\varepsilon}(|A| + |\bar{L}||\bar{U}|) \le \frac{2n\varepsilon}{1 - 2n\varepsilon}|\bar{L}||\bar{U}|\,.^{5}$$

[5]Koristimo sledeće oznake: $|A| = [\,|a_{ij}|\,]$, $|A| \le |B| \Leftrightarrow (\forall\, i, j)\,|a_{ij}| \le |b_{ij}|$.

Prva nejednakost vredi za $n\varepsilon < 1$*, a druga za* $2n\varepsilon < 1$.[6]

Prethodna analiza ne daje odgovor na pitanje koliko se $\bar{L}$ i $\bar{U}$ razlikuju od tačnih L i U. Ona sugeriše da $\bar{L}$ i $\bar{U}$ čine egzaktnu LU faktorizaciju matrice $A+\Delta A$. Vidimo da je matrica ΔA zadovoljavajuće mala ako proizvod $|\bar{L}||\bar{U}|$ nije prevelik u odnosu na $|A|$. Sledeći primer pokazuje da to nije osigurano u LU faktorizaciji bez pivotiranja, tj. ona je nestabilan algoritam.

Primer 2.3. Neka je α mali parametar, $|\alpha| << 1$ i matrica

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

U egzaktnom računu je

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

Ako ovaj račun provodimo u aritmetici s 8 decimalnih cifara, tj. tačnosti $\varepsilon \approx 10^{-8}$, i ako je $|\alpha| < \varepsilon$, vredi

$$\bar{\ell}_{21} = \ell_{21}(1+\epsilon_1),\ \bar{u}_{11} = u_{11},\ \bar{u}_{12} = u_{12},\ \bar{u}_{22} = 1 \ominus 1 \oslash \alpha = -\frac{1}{\alpha}(1+\epsilon_1),$$

za neko $|\epsilon_1| \leq \varepsilon$. Sad se lako proverava da je

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

gde je $\alpha(1 \oslash \alpha) = 1+\epsilon$, $|\epsilon| \leq \varepsilon$. Kako je element na poziciji $(2,2)$ reda veličine $1/|\alpha| > 1/\varepsilon$, algoritam nam ne može garantovati zadovoljavajuće malo ΔA.

Cilj numeričke analize algoritma je da otkrije moguće uzroke nestabilnosti algoritma i ponudi rešenje za njihovo uklanjanje. U našem slučaju je to mogućnost proizvoljnog rasta elemenata u toku faktorizacije, tj. mogućnost da $|\bar{L}||\bar{U}|$ ima proizvoljno velike elemente. Jedan od načina za njeno uklanjanje je uvođenje parcijalnog pivotiranja.

[6]Primetimo da su ova ograničenja skoro beznačajna jer n može biti veličina reda 10^8, a videli smo da problemi nastaju i za $n = 10^5$.

U analizi grešaka zaokruživanja pivotiranje ne predstavlja dodatnu poteškoću, pa je očigledno da važi sledeća teorema.

Teorema 2.3.2 *Ako su $\bar{L}$ i $\bar{U}$ trougaoni faktori matrice A dobijeni Gaussovim eliminacijama sa parcijalnim pivotiranjem primenom aritmetike računara preciznosti ε, onda je*

$$\bar{L}\,\bar{U} = P(A + \Delta A), \quad |\Delta A| \leq \frac{2n\varepsilon}{1 - 2n\varepsilon} P^T |\bar{L}||\bar{U}|\,,$$

uz pretpostavku da je $2n\varepsilon < 1$.

Na osnovu razmatranja iz poglavlja 2.1, parcijalno pivotiranje osigurava da su svi elementi matrice $|\bar{L}|$ manji od jedan. Sada veličina $|\bar{L}||\bar{U}|$ bitno zavisi samo od elemenata matrice $\bar{U}$. Pošto se oni dobijaju od elemenata $\bar{a}_{ij}^{(k)}$, broj

$$\rho = \frac{\max\limits_{ijk} \bar{a}_{ij}^{(k)}}{\max\limits_{ij} a_{ij}}$$

je dobra mera za relativni rast elemenata u proizvodu $|\bar{L}||\bar{U}|$ u odnosu na A. Broj ρ zovemo faktor rasta elemenata u LU faktorizaciji, i definisan je bez obzira koristimo li pivotiranje. Sledeća teorema govori da on sada nije neograničen.

Teorema 2.3.3 *Ako LU faktorizaciju sa parcijalnim pivotiranjem računamo u aritmetici računara preciznosti ε, onda je*

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

Ako je računanje egzaktno, onda je $\rho \leq 2^{n-1}$.

Dokaz. U k-tom koraku algoritma novi elementi matrice $A^{(k)}$ se računaju formulom

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik} a_{kj}^{(k-1)},$$

pa je, s obzirom da je $|m_{ik}| \leq 1$, u egzaktnom računanju

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

Dakle najveći element matrice $A^{(k-1)}$ se u računanju $A^{(k)}$ može najviše udvostručiti. Ako koristimo aritmetiku pokretnog zareza, zbog dva množenja dobijamo

$$|a_{ij}^{(k)}| \leq 2 \max_{ij} |a_{ij}^{(k-1)}|(1+\varepsilon)^2.$$

Nakon izvedenih $n-1$ koraka dobijamo tvrđenje teoreme. □

Gornja granica za ρ je reda veličine 2^{n-1}, što brzo raste kao funkcija od n. Postoje primeri na kojima se ta gornja granica i dostiže. Jedan od njih su matrice

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 1 \\ -1 & 1 & 0 & & 0 & 1 \\ -1 & -1 & 1 & \ddots & \vdots & 1 \\ \vdots & \vdots & & \ddots & 0 & \vdots \\ -1 & -1 & & & 1 & 1 \\ -1 & -1 & & \dots & -1 & 1 \end{bmatrix}.$$

Međutim, iskustvo iz prakse govori da su takvi primeri retki, i da je LU faktorizacija sa parcijalnim pivotiranjem dobar algoritam za rešavanje linearnih sistema. Statistička analiza za razne vrste slučajnih matrica pokazuje da je faktor rasta u proseku oko $n^{2/3}$.

Preostaje nam još analizirati numeričko rešenje trougaonog sistema. Sledeća teorema pokazuje da supstitucije unapred (unazad) pored jednostavnosti imaju i numeričku stabilnost.

Teorema 2.3.4 *Neka je T donja (gornja) trougaona matrica reda n i neka je sistem $Tx = d$ rešen supstitucijama unapred (unazad). Ako je $\bar{x}$ rešenje dobijeno primenom aritmetike računara preciznosti ε, onda postoji donja (gornja) trougaona matrica ΔT takva da vredi*

$$(T + \Delta T)\bar{x} = d, \quad |\Delta T| \leq \frac{n\varepsilon}{1 - n\varepsilon}.$$

Dokaz. Dokaz izvodimo na sličan način kao i za teoremu 2.3.1. U supstitucijama unapred, na osnovu primera 1.18. važi

$$\bar{x}_i = \frac{d_i - \sum_{j=1}^{i-1} t_{ij}\bar{x}_j(1+\xi_j)}{\frac{t_{ii}}{(1+\epsilon_{1,i})(1+\epsilon_{2,i})}}, \quad |\xi_j| \leq \frac{(i-1)\varepsilon}{1-(i-1)\varepsilon}, \quad |\epsilon_{1,i}| \leq \varepsilon, \quad |\epsilon_{2,i}| \leq \varepsilon.$$

Prethodna jednakost se može napisati i u obliku

$$d_i = \sum_{j=1}^{i} t_{ij}\bar{x}_j + \delta d_i\,, \quad \delta d_i = \sum_{j=1}^{i} t_{ij}\bar{x}_j\xi_j\,, \quad \xi_i = \frac{1}{(1+\epsilon_{1,i})(1+\epsilon_{2,i})} - 1\,.$$

Lako se pokazuje da je $|\xi_i| \leq n\varepsilon/(1-n\varepsilon)$.

Slično vredi i za supstitucije unazad. □

Kompozicijom dobijenih rezultata ćemo oceniti koliko tačno možemo rešiti linearni sistem $A\,x = b$.

Teorema 2.3.5 *Neka je $\bar{x}$ rešenje regularnog $n \times n$ sistema jednačina $A\,x = b$ dobijeno Gaussovim eliminacijama sa parcijalnim pivotiranjem primenom aritmetike računara preciznosti ε. Tada postoji matrica ΔA takva da vredi*

$$(A + \Delta A)\bar{x} = b, \quad |\Delta A| \leq \frac{5n\varepsilon}{1-2n\varepsilon} P^T|\bar{L}||\bar{U}|,$$

uz pretpostavku $2n\varepsilon < 1$.

Dokaz. Teoretski, permutacija P bi se mogla primeniti odmah na početku, pa numeričku analizu možemo provesti bez pivotiranja. Na osnovu teorema 2.3.1 i 2.3.4 sledi da izračunato rešenje $\bar{x}$ zadovoljava sistem

$$(A + \Delta_1 A + \bar{L}\,\Delta\bar{U} + \Delta\bar{L}\,\bar{U} + \Delta\bar{L}\,\Delta\bar{U})\,\bar{x} = b,$$

gde je

$$|\Delta_1 A| \leq \frac{2n\varepsilon}{1-2n\varepsilon}|\bar{L}||\bar{U}|, \quad |\Delta\bar{U}| \leq \frac{n\varepsilon}{1-n\varepsilon}|\bar{U}|, \quad |\Delta\bar{L}| \leq \frac{n\varepsilon}{1-n\varepsilon}|\bar{L}|\,.$$

Odavde, uz očigledne nejednakosti

$$\frac{n\varepsilon}{1-n\varepsilon} \leq \frac{n\varepsilon}{1-2n\varepsilon}\,, \quad \left(\frac{n\varepsilon}{1-n\varepsilon}\right)^2 \leq \frac{n\varepsilon}{1-2n\varepsilon}\,,$$

za $2n\varepsilon < 1$, sledi tvrđenje teoreme. □

2.4 Teorija perturbacije linearnog sistema

Teorija perturbacije treba da odgovori na pitanje koliko se (po normi) promeni rešenje linearnog sistema $A\,x = b$ ako se (po normi) malo promene A, b ili oba.

Razmotrimo prvo slučaj kada je perturbovana samo matrica sistema. Umesto sistema $A\,x = b$, egzaktno rešavamo sistem

$$(A + \Delta A)(x + \Delta x) = b\,. \tag{2.6}$$

Pretpostavljamo da je $||\Delta A|| \leq \xi\,||A||$, tj. norma matrice ΔA je mala u odnosu na normu polazne matrice. Iz (2.6) dobijamo

$$A\,\Delta x + \Delta A\,(x + \Delta x) = 0\,,$$

a nakon množenja sleva s A^{-1}

$$\Delta x = -A^{-1}\,\Delta A\,(x + \Delta x)\,.$$

Dalje koristimo osobine norme i dobijamo

$$\begin{aligned} ||\Delta x|| &\leq ||A^{-1}||\,||\Delta A||\,||x + \Delta x|| \leq \xi\,||A^{-1}||\,||A||\,||x + \Delta x|| \\ &\leq \xi\,k(A)\,(||x|| + ||\Delta x||)\,, \end{aligned}$$

gde je $k(A) = ||A||\,||A^{-1}||$ standardna oznaka za uslovljenost matrice A. Ako je $\xi\,k(A) < 1$ (a to znači da je i $||\Delta A||\,||A^{-1}|| < 1$), onda je

$$||\Delta x|| \leq \frac{\xi\,k(A)}{1 - \xi\,k(A)}\,||x||\,. \tag{2.7}$$

Vidimo da je greška rešenja približno proporcionalna uslovljenosti matrice A.

Pretpostavimo sada da umesto sistema $A\,x = b$, egzaktno rešavamo sistem

$$A\,(x + \Delta x) = b + \Delta b\,, \tag{2.8}$$

tj. samo je desna strana sistema malo perturbovana. Neka je $||\Delta b|| \leq \xi\,||b||$. Slično kao u prethodnom slučaju iz (2.8) dobijamo

$$\Delta x = A^{-1}\,\Delta b\,,$$

i dalje

$$\begin{aligned} ||\Delta x|| &\leq ||A^{-1}||\,||\Delta b|| \leq \xi\,||A^{-1}||\,||b|| \leq \xi\,||A^{-1}||\,||A\,x|| \\ &\leq \xi\,||A^{-1}||\,||A||\,||x|| \leq \xi\,k(A)\,||x||\,. \end{aligned}$$

Opet je greška rešenja proporcionalna uslovljenosti matrice A.

Na isti način dokazujemo i sledeću teoremu.

Teorema 2.4.1 *Neka je $A\,x = b$ i*

$$(A + \Delta A)(x + \Delta x) = b + \Delta b \tag{2.9}$$

gde je $||\Delta A|| \leq \xi\,||E||$, $||\Delta b|| \leq \xi\,||f||$, i neka je $\xi\,||A^{-1}||\,||E|| < 1$. Tada za $x \neq 0$ vredi

$$\frac{||\Delta x||}{||x||} \leq \frac{\xi}{1 - \xi\,||A^{-1}||\,||E||}\left(\frac{||A^{-1}||\,||f||}{||x||} + ||A^{-1}||\,||E||\right). \tag{2.10}$$

Ocena u (2.10) je optimalna jer se skoro dostiže za $\Delta A = \xi\,||E||\,||x||\,wv^T$ i $\Delta b = -\xi\,||f||\,w$, gde je $||w|| = 1$, $||A^{-1}\,w|| = ||A^{-1}||$ i $v^T x = 1$.

U teoremi je korišten opšti oblik ocene za normu perturbacija polaznih podataka. Ako koristimo prirodniji "relativni"oblik, tj. ako uzmemo da je $E = A$ i $f = b$, onda se ocena (2.10) može prikazati u obliku

$$\frac{||\Delta x||}{||x||} \leq \frac{\xi}{1 - \xi\,k(A)}\left(\frac{||A^{-1}||\,||b||}{||x||} + k(A)\right).$$

Izraz u zagradi je očigledno manji od $2\,k(A)$, pa možemo dobiti i jednostavniju, ali i nešto lošiju ocenu

$$\frac{||\Delta x||}{||x||} \leq \frac{2\,k(A)\,\xi}{1-\xi\,k(A)}\,.$$

Ova ocena je slabija za faktor 2 od ocene (2.7), ali ne treba zaboraviti da ona uključuje perturbacije od A i b, a ne samo od A.

Videli smo da je broj $k(A)$ pokazatelj osetljivosti linearnog sistema. Ako je $k(A)$ veliko, sistem $A\,x = b$ (ili samo matricu A) zovemo slabo uslovljenim. Primetimo da je to osobina koja zavisi od izbora matrične norme. Međutim, ako je sistem loše uslovljen u jednoj normi, onda je on loše uslovljen i u svakoj, njoj ekvivalentnoj normi. S obzirom da se najčešće koriste p-norme, a one su međusobno ekvivalentne, izbor norme i nije toliko bitan. U svakoj p-normi je $k(A) \geq 1$. Matrice sa malim $k(A)$ zovemo dobro uslovljenim. Ortogonalne matrice su savršeno uslovljene u 2-normi jer je $k_2(A) = 1$.

2.5 Faktorizacija Choleskog

U ovom poglavlju nas interesuje kako na $L\,U$ faktorizaciju utiču neke osobine matrice A, pre svega simetričnost i pozitivna definitnost.

Teorema 2.5.1 *Ako je kvadratna matrica A pozitivno definitna, tada postoje donje trougaone matrice sa jedinicama na dijagonali L i M, i dijagonalna matrica sa pozitivnim elementima na dijagonali, takve da je $A = L\,D\,M^T$. Ova faktorizacija je jedinstvena.*

Dokaz. Svi glavni minori matrice A su pozitivni pa važi $A = L\,U$, gde L na dijagonali ima jedinice, a U pozitivne brojeve. Matrica U se očigledno može napisati u obliku $U = D\,M^T$, gde je $D = \mathrm{diag}(u_{11}, \ldots, u_{nn})$, a $M^T = D^{-1}\,U$. Jedinstvenost sledi iz jedinstvenosti faktorizacija $L\,U$ i $D\,M^T$ □

Primer 2.4.

$$\begin{bmatrix} 10 & 10 & 20 \\ 20 & 25 & 40 \\ 30 & 50 & 61 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{bmatrix} \begin{bmatrix} 10 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Ako je matrica A još i simetrična, iz $A^T = A$ sledi $L\,D\,M^T = M\,D\,L^T$, i dalje

$$D\,M^T\,L^{-T} = L^{-1}\,M\,D\,.$$

Na levoj strani se nalazi proizvod gornjih trougaonih, a na desnoj donjih trougaonih matrica, pa su ti proizvodi dijagonalne matrice. Ta dijagonalna matrica je baš D jer M i L imaju na dijagonali jedinice. Dakle, $L^{-1}\,M$ je jedinična matrica, tj. $L = M$.

Ako sa R označimo sledeću gornju trougaonu matricu

$$R = \mathrm{diag}(\sqrt{d_{11}}, \ldots, \sqrt{d_{nn}})\,L^T\,,$$

onda je $A = R^T R$. Ovim smo dokazali sledeću teoremu.

Teorema 2.5.2 *Ako je matrica A simetrična i pozitivno definitna, tada postoji tačno jedna gornja trougaona matrica R sa pozitivnim elementima na dijagonali, takva da je $A = R^T R$.*

Primer 2.5.

$$\begin{aligned} \begin{bmatrix} 2 & -2 \\ -2 & 5 \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \\ &= \left(\begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{3} \end{bmatrix} \right) \left(\begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{3} \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \right) \\ &= \begin{bmatrix} \sqrt{2} & 0 \\ -\sqrt{2} & \sqrt{3} \end{bmatrix} \begin{bmatrix} \sqrt{2} & -\sqrt{2} \\ 0 & \sqrt{3} \end{bmatrix} \end{aligned}$$

Faktorizacija $A = R^T R$ je poznata kao faktorizacija Choleskog. Videli smo kako doći do nje polazeći od $L\,U$ faktorizacije. Efikasnija metoda za

njeno izračunavanje se može izvesti poredeći elemente u jednakosti

$$A = \begin{bmatrix} r_{11} & 0 & \dots & 0 \\ r_{12} & r_{22} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ r_{1n} & r_{2n} & \dots & r_{nn} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & r_{nn} \end{bmatrix}.$$

Tako dobijamo sledeće formule:

$$\begin{aligned} r_{kk} &= \left(a_{kk} - \sum_{p=1}^{k-1} r_{pk}^2 \right)^{1/2}, \quad k = 1, \dots, n \\ r_{ik} &= \frac{1}{r_{ii}} \left(a_{ik} - \sum_{p=1}^{k-1} r_{pi} r_{pk} \right), \quad i = 1, 2, \dots, k-1\,. \end{aligned}$$

U ovim formula potrebno je izvesti približno $n^3/3$ elementarnih operacija (i n korenovanja, ali je to zanemarivo malo), što je duplo manje nego u Gaussovim eliminacijama. Videli smo da je razlog nestabilnosti Gaussovih eliminacija mogućnost proizvoljnog rasta elemanata u matricama L i U. Ovde takva mogućnost ne postoji, što pokazuje nejednakost

$$r_{ji}^2 \le \sum_{p=1}^{i} r_{pi}^2 = a_{ii}\,,$$

pa zaključujemo da je posmatrani algoritam stabilan. Međutim, mogu se javiti problemi zbog korenovanja, jer je moguće da zbog gomilanja greške zaokruživanja broj koji treba korenovati bude negativan ili nula.

Primer 2.6. Odredimo faktorizaciju Choleskog matrice

$$A = \begin{bmatrix} 100 & 15 & 0.01 \\ 15 & 2.2 & 0.01 \\ 0.01 & 0.01 & 1.00 \end{bmatrix},$$

zaokružujući sve rezultate na dve značajne cifre (aritmetika pokretnog zareza sa osnovom 10, bez ograničenja sa karakteristiku i $t = 2$). Redom dobijamo $\bar{r}_{11} = 10$, $\bar{r}_{21} = 1.5$, $\bar{r}_{31} = 0.001$. Ako prilikom zokruživanja koristimo pravilo parne cifre dobijamo $\bar{r}_{22} = 0$.

Detaljna analiza ovog problema pokazuje da se ovo dešava samo kod veoma loše uslovljenih matrica, preciznije kod matrica kod kojih je $k_2(A) \approx \varepsilon^{-1}$.

Kad imamo faktorizaciju Choleskog, rešavanje sistema $A\,x = b$ se svodi na rešavanje dva trougaona sistema $R^T\,y = b$ i $R\,x = y$, gde u obe supstitucije imamo deljenja, za razliku od Gaussove metode.

Zato se za rešavanje linearnih sistema dosta često koristi $L\,D\,L^T$ faktorizacija. Računamo je na sličan način kao i faktorizaciju Choleskog. U tom algoritmu nema n korenovanja, jer su brojevi r_{ii}^2 elementi matrice D. Rešenje sistema $A\,x = b$ dobijamo rešavanjem tri linearna sistema

$$L\,z = b\,, \quad D\,y = z\,, \quad L^T\,x = y\,.$$

Prvi i treći sistem su trougaoni sistemi sa jediničnom dijagonalom, pa u supstitucijama nema deljenja. Srednji sistem je dijagonalan i trivijalno se rešava sa n deljenja. Ukupno, imamo n deljenja manje nego u sistemima iz faktorizacije Choleskog, što i nije neka ušteda jer, u opštem slučaju, imamo oko n^2 operacija po supstituciji. Međutim, postoje situacije kad je ova ušteda itekako značajna. Na primer, ako je matrica A još i trodijagonalna, za supstitucije iz faktorizacije Choleskog treba oko $6n$ operacija, a ovde samo oko $5n$ operacija.

Dodatnu potvrdu stabilnosti za simetrične pozitivno definitne matrice dobijamo posmatranjem njihove $L\,U$ faktorizacije koju možemo izvesti Gaussovim eliminacijama bez pivotiranja. Nije teško pokazati da je faktor rasta ρ jednak jedinici u egzaktnom računu, ili $(1+\varepsilon)^{2(n-1)}$ u aritmetici računara preciznosti ε.

2.6 QR faktorizacija

U mnogim primenama matrica A je zadana u obliku

$$A = G^T G\,,$$

gde je $G \in \mathbb{R}^{m\times n}$. Matrica A je očigledno simetrična i pozitivno semidefinitna, jer je

$$x^T A\, x = (x^T G^T)(Gx) = (Gx)^T(Gx) \geq 0\,.$$

Ako matrica G ima osobine $m \geq n$ i $\text{rang}(G) = n$, A je pozitivno definitna. Naime, ako je $m < n$, onda je $\text{rang}(G) \leq m$, odakle sledi da je i $\text{rang}(A) \leq m < n$, tj. A je singularna. Ako je $\text{rang}(G) = n$, onda $\forall x \neq 0$ vredi $Gx \neq 0$, pa je i $x^T Ax > 0$.

Pod navedenim uslovima za G, A ima faktorizaciju Choleskog $A = R^T R$. Prva ideja kako da odredimo R je da eksplicitno odredimo matricu A, a zatim primenimo algoritam iz prethodnog poglavlja. Postavlja se pitanje možemo li dobiti R direktno iz G, i tako izbeći množenje matrica koje može dovesti do velike relativne greške u proizvodu (videti primer 1.18).

To je moguće izvesti ako postoji ortogonalna matrica Q reda m, takva da je

$$G = Q\, R = Q \begin{bmatrix} R_0 \\ 0 \end{bmatrix}\,, \tag{2.11}$$

gde je R_0 gornja trougaona matrica reda n sa pozitivnim elementima na dijagonali. Naime, u tom slučaju je

$$A = G^T G = R^T Q^T Q\, R = R^T R = R_0^T R_0\,,$$

pa je R_0 traženi faktor iz faktorizacije Choleskog. Ako faktorizacija (2.11) postoji, očigledno je jedinstvena. U ostatku ovog poglavlja dajemo konstruktivni dokaz njene egzistencije.

Ideja je da se u Gaussovim eliminacijama matrice M_i, $i = 1, \ldots, n-1$, zamene ortogonalnim matricama P_i koje postižu isti efekat poništavanja elemenata ispod dijagonale u i-toj koloni, i pri tome, u prethodnim kolonama zadržavaju sve nule ispod dijagonale. Činjenica da sada polazna matrica ima više vrsta nego kolona ne igra nikakvu ulogu. Šablonski, za matricu tipa 5×3, postupak izgleda ovako

$$\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \end{bmatrix} \rightarrow \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & x & x \\ 0 & x & x \\ 0 & x & x \end{bmatrix} \rightarrow \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & x \\ 0 & 0 & x \\ 0 & 0 & x \end{bmatrix} \rightarrow \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & x \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} .$$

Prva matrica je A, a zatim su redom $P_1\,A$, $P_2\,P_1\,A$, $P_3\,P_2\,P_1\,A$. Jasno, u opštem slučaju imamo

$$P_{n-1}\,P_{n-2}\ldots\,P_2\,P_1\,A = R\,,$$

gde je R gornja trougaona matrica. Proizvod

$$Q^T = P_{n-1}\,P_{n-2}\ldots\,P_2\,P_1$$

je ortogonalna matrica, pa sledi da je $A = QR$. Uz sve ovo matrice P_i moraju biti takve da na kraju postupka R na dijagonali ima pozitivne elemente. Ukoliko je to nemoguće postići samom matricom P_i, možemo izvesti još jedno množenje matricom koju dobijamo tako što u I promenimo jednu jedinicu u -1.

Demonstriraćemo dve mogućnosti za određivanje matrica P_i.

2.6.1 Householderova refleksija

Householderova refleksija je matrica H defnisana sa

$$H = I - 2vv^T\,, \quad ||v||_2 = 1\,.$$

Lako se proverava da je matrica H simetrična i ortogonalna. Vektor Hx predstavlja refleksiju vektora x u odnosu na hiperravan koja je ortogonalna na vektor v.

Za dati nenula vektor x može se odrediti v, tako da je

$$Hx = \begin{bmatrix} c \\ 0 \\ \vdots \\ 0 \end{bmatrix} = c \cdot e_1 \,. \tag{2.12}$$

Iz prethodne jednakosti direktno sledi

$$v = \frac{1}{2\,(v^T x)}(x - c\,e_1)\,,$$

tj. v možemo napisati u obliku

$$v = \alpha\,(x - c\,e_1)\,.$$

Iz ortogonalnosti matrice H i (2.12) sledi

$$||Hx||_2 = ||x||_2 = |c|\,,$$

pa važi

$$v = \frac{\hat{v}}{||\hat{v}||_2}\,, \quad \hat{v} = x \pm ||x||_2\, e_1\,.$$

Ukoliko je prva koordinata vektora x značajno veća od ostalih, tj. ako je $||x||_2 \approx x_1$, pri računanju elementa $\hat{v}_1$ po formuli $\hat{v} = x - \text{sign}(x_1)\,||x||_2\, e_1$ dolazi do oduzimanja približno istih brojeva, što može narušiti stabilnost izračunavanja. Zato se koristi

$$\hat{v} = x + \text{sign}(x_1)\,||x||_2\, e_1\,.$$

Primer 2.7. Neka je $x = [3, 1, 5, 1]^T$. Tada je $||x||_2 = 6$, pa je $\hat{v} = [9, 1, 5, 1]^T$, $Hx = [-6, 0, 0, 0]^T$ i

$$H = I - \frac{2}{\hat{v}^T\,\hat{v}}\,\hat{v}\,\hat{v}^T = \frac{1}{54}\begin{bmatrix} -27 & -9 & -45 & -9 \\ -9 & 53 & -5 & -1 \\ -45 & -5 & 29 & -5 \\ -9 & -1 & -5 & 53 \end{bmatrix}.$$

Householderova refleksija može i selektivnije poništiti delove vektora x, na sledeći način. Neka je $1 \leq k \leq j \leq n$ i

$$\hat{v} = [0, \ldots, 0, x_k + \text{sign}(x_k)\alpha, x_{k+1}, \ldots, x_j, 0, \ldots, 0]^T ,$$

gde je $\alpha = x_k^2 + \ldots + x_j^2$. Tada matrica $H = I - 2vv^T$ ima osobinu da je

$$Hx = [x_1, \ldots, x_{k-1}, -\text{sign}(x_k)\alpha, 0, \ldots, 0, x_{j+1}, \ldots, x_n]^T .$$

Prilikom računanja proizvoda HA možemo iskoristiti strukturu matrice H

$$H\,A = (I - 2vv^T)\,A = A - 2v(A^T v)^T .$$

Vidimo da ne moramo eksplicitno formirati H, dovoljno je znati vektor v. Nakon množenja HA, u matrici A se menjaju vrste $k, k+1, \ldots, j$.

2.6.2 Givensova rotacija

Householderovom refleksijom možemo poništiti sve elemente vektora x (ili jednog njegovog dela) osim prvog elementa. Međutim, često je potrebno selektivnije poništavanje, koje se postiže Givensovom rotacijom. To je ortogonalna matrica oblika

$$J(i,k,\theta) = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & \cos\theta & & & & -\sin\theta & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & \sin\theta & & & & \cos\theta & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix} .$$

U poređenju sa jediničnom matricom, razlikuju se samo četiri elementa na pozicijama $(\pm i, \pm k)$. Vektor $J(i,k,\theta)x$ se dobija rotacijom vektora x u (i,k) ravni za ugao θ.

Ugao θ se može izbrati tako da se u matrici $J(i,k,\theta)\,x$ poništi k-ta koordinata. Pri tome se promeni još samo i-ta koordinata. Iz jednakosti

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_i \\ x_k \end{bmatrix} = \begin{bmatrix} x_i' \\ 0 \end{bmatrix},$$

dobijamo

$$\sin\theta\, x_i + \cos\theta\, x_k = 0\,,$$

a to važi za

$$\sin\theta = -\frac{x_k}{\sqrt{x_i^2 + x_k^2}}\,, \quad \cos\theta = \frac{x_i}{\sqrt{x_i^2 + x_k^2}}\,.$$

Tada je

$$x_i' = \frac{x_i^2 + x_k^2}{\sqrt{x_i^2 + x_k^2}} = \sqrt{x_i^2 + x_k^2} > 0\,.$$

PRIMER 2.8. Ako je $x = [1,2,3,4]^T$, uzmimo da je $\sin\theta = -1/\sqrt{5}$ i $\cos\theta = 2/\sqrt{5}$. Tada je $J(2,4,\theta)x = [1, 2\sqrt{5}, 3, 0]^T$ i

$$J(2,4,\theta) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2/\sqrt{5} & 0 & 1/\sqrt{5} \\ 0 & 0 & 1 & 0 \\ 0 & -1/\sqrt{5} & 0 & 2/\sqrt{5} \end{bmatrix}.$$

Množenjem $J(i,k,\theta)\,A$ u matrici A menjaju se samo vrste i i k.

Da bismo u vektoru $x \in \mathbb{R}^n$ poništili sve koordinate osim prve potrebno ga je $n-1$ puta pomnožiti sa odgovarajućim Givensvim rotacijama, a samo jednom odgovarajućom Householderovom refleksijom. Međutim, ukupan broj operacija je neznatno veći u drugom slučaja jer za računanje proizvoda Jx treba manje operacija nego za računanje proizvoda Hx. S druge strane moguće je odrediti J tako da broj koji se ne poništava[7] bude pozitivan, dok na njegov predznak u proizvodu Hx ne možemo uticati.

Napomena. Sistem linearnih jednačina $A\,x = b$, $A \in \mathbb{R}^{m\times n}$, $m > n$, uglavnom nema rešenje, pa ima smisla tražiti vektor $x \in \mathbb{R}^n$, takav da je

[7] To je broj koji će biti na dijagonali matrice R.

vektor $Ax \in \mathbb{R}^m$ najbliži vektoru b. Preciznije, treba odrediti vektor x koji minimizira $||A\,x - b||_p$. Najpogodniji izbor norme je $p = 2$, jer je $f(x) = ||A\,x - b||_2$ neprekidno diferencijabilna funkcija. Ako je $\text{rang}(A) = n$, tada QR faktorizacijom matrice A možemo jednostavno rešiti ovaj problem. Iz

$$Q^T A = R = \begin{bmatrix} R_0 \\ 0 \end{bmatrix}, \quad Q^T b = \begin{bmatrix} c \\ d \end{bmatrix},$$

sledi

$$||A\,x - b||_2^2 = ||Q^T A\,x - Q^T b||_2^2 = ||R_0\,x - c||_2^2 + ||d||_2^2\,.$$

Zaključujemo da je rešenje posmatranog problema jedinstvano i dobijamo ga jednostavno, rešavajući gornji trougaoni sistem $R_0 x = c$.

2.7 Iterativne metode

Do sada smo pomenuli nekoliko metoda koje se koriste za rešavanje regularnog linearnog sistema $A\,x = b$. Njihova zajednička karakteristika je da one nakon konačno mnogo izvedenih operacija dovode do tačnog rešenja ako se sve operacije izvode egzaktno. Ovakve metode zovemo direktnim.

Videli smo i da rešenje sistema $A\,x = b$ uglavnom ne možemo izračunati apsolutno tačno. Računar nije ograničen samo po pitanju numeričke tačnosti, već i memorijskog prostora i brzine izvođenja operacija. Naveli smo ranije primere na osnovu kojih zaključujemo da su direktne metode nepraktične ako je matrica A velikih dimenzija.

U velikom broju primena (naročito u numeričkom rešavanju parcijalnih diferencijalnih jednačina) matrica A je retko popunjena, u smislu da je veliki broj njenih elemenata jednak nuli, dok su nenula elementi uglavnom pravilno raspoređeni. Direktne metode se mogu prilagoditi tako da iskoriste strukturu takvih matrica kako bi se smanjio broj potrebnih operacija. Međutim, takvu

strukturu je mnogo lakše iskoristiti ako se matrica A u algoritmu koristi samo kao faktor.

Pored toga, u praksi se skoro nikada ne radi s egzaktnim podacima (matrica A može biti rezultat merenja ili prethodnih proračuna, dakle netačna), pa nema smisla tražiti tačno rešenje $x = A^{-1}\, b$.

Prethodna diskusija nas motiviše da potražimo drugačije metode za rešavanje linearnih sistema. Cilj je konstruisati niz $x^{(0)}, x^{(1)} \ldots, x^{(n)}, \ldots$ vektora iz $\mathbb{R}^n$, koji konvergira (po normi) ka tačnom rešenju $x = A^{-1}\, b$, i u čijem formiranju se matrica A (ili neka druga dobijena iz A) koristi samo kao faktor. Ovakve metode zovemo iterativnim. U praksi se gotovo isključivo koriste iterativne metode prvog reda, kojima se iteracija x^{m+1} dobija samo iz prethodne x^m. Svako rastavljanje matrice A u obliku $A = M - K$, gde je matrica M regularna, bi moglo generisati jednu iterativnu metodu na sledeći način:

$$(M - K)\, x = b \Longrightarrow x = M^{-1}\, K\, x + M^{-1}\, b\,.$$

Uvođenjem oznaka $T = M^{-1}\, K$, $c = M^{-1}\, b$ prethodna relacija postaje

$$x = T\, x + c\,,$$

tj. pravo rešenje x je fiksna tačka preslikavanja $f(x) = T\, x + c$. Iterativnu metodu definišemo sa

$$x^{(m+1)} = T\, x^{(m)} + c, \quad m \in \mathbb{N}_0\,. \tag{2.13}$$

Dovoljan uslov za konvergenciju ovakvih metoda daje sledeća teorema.

Teorema 2.7.1 *Niz $\{x^{(m)}\}$, $m \in \mathbb{N}_0$, definisan relacijom (2.13) konvergira ka rešenju linearnog sistema $A\, x = b$ za sve početne vektore $x^{(0)}$ i sve desne strane b, ako je*

$$||T|| < 1\,,$$

pri čemu je $||\ \ ||$ *proizvoljna operatorska norma*[8].

Dokaz. Oduzimanjem $x = T\,x + c$ od (2.13) dobijamo

$$x^{(m+1)} - x = T\,(x^{(m)} - x)\,.$$

i dalje, uzimanjem norme

$$||x^{(m+1)} - x|| \leq ||T||\,||x^{(m)} - x|| \leq ||T||^{m+1}\,||x^{(0)} - x||\,.$$

Iz $||T|| < 1$ sledi $||T||^{m+1} \to 0$ kad $m \to \infty$, a to znači $||x^{(m+1)} - x|| \to 0$. Iz neprekidnosti norme sledi $x^{(m+1)} \to x$ za svako $x^{(0)}$. □

Korištenjem veze spektralnog radijusa ρ i operatorske norme matrice, možemo dobiti i potreban uslov za konvergenciju iterativne metode.

Lema 2.7.1 *Za sve operatorske norme* $||\ \ ||$ *vredi*

$$\rho(T) \leq ||T||\,.$$

Za svaku kvadratnu matricu T *i svako* $\varepsilon > 0$ *postoji operatorska norma* $||\ \ ||_*$, *takva da je*

$$||T||_* \leq \rho(T) + \varepsilon\,.$$

Dokaz. Neka je $\rho(T) = |\lambda|$, a z karakteristični vektor koji odgovara λ. Tada vredi

$$||T|| = \max_{y \neq 0} \frac{||Ty||}{||y||} \geq \frac{||Tz||}{||z||} = \frac{||\lambda z||}{||z||} = \frac{|\lambda|\,||z||}{||z||} = \rho(T)\,.$$

Neka je $\mathcal{J}$ Jordanova forma matrice T, $\mathcal{J} = S^{-1}\,T\,S$ i neka je $D_\varepsilon = \mathrm{diag}(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{n-1})$. Matrica $D_\varepsilon^{-1}\,\mathcal{J}\,D_\varepsilon = (S\,D_\varepsilon)^{-1} T (S\,D_\varepsilon)$ se razlikuje od $\mathcal{J}$ samo po tome što na dijagonali umesto jedinica ima ε. Matrice S i D_ε su regularne, pa možemo definisati sledeću vektorsku normu

$$||x||_* = ||(S\,D_\varepsilon)^{-1}\,x||_\infty\,.$$

[8]Matrična norma indukovana vektorskom normom.

Pokazaćemo da ona indukuje traženu operatorsku normu. U tako dobijenoj operatorskoj normi vredi

$$\begin{aligned} ||T||_* &= \max_{x\neq 0} \frac{||Tx||_*}{||x||_*} = \max_{x\neq 0} \frac{||(S\,D_\varepsilon)^{-1}Tx||_\infty}{||(S\,D_\varepsilon)^{-1}x||_\infty} = \max_{y\neq 0} \frac{||(S\,D_\varepsilon)^{-1}T(S\,D_\varepsilon)y||_\infty}{||y||_\infty} \\ &= ||(S\,D_\varepsilon)^{-1}T(S\,D_\varepsilon)||_\infty \leq \max_i |\lambda_i| + \varepsilon = \rho(T) + \varepsilon\,. \end{aligned}$$

□

Teorema 2.7.2 *Niz* $\{x^{(m)}\}$, $m \in \mathbb{N}_0$, *definisan relacijom (2.13) konvergira ka rešenju linearnog sistema* $A\,x = b$ *za sve početne vektore* $x^{(0)}$ *i sve desne strane* b, *ako i samo ako je*

$$\rho(T) < 1\,,$$

pri čemu je $\rho(T)$ *spektralni radijus matrice* $T = M^{-1}\,K$.

Dokaz. Ako je $\rho(T) \geq 1$, izaberimo početni vektor $x^{(0)}$, tako da je $x^{(0)} - x$ karakteristični vektor koji odgovara karakterističnoj vrednosti λ, takvoj da je $|\lambda| = \rho(T)$. Tada vredi

$$x^{(m+1)} - x = T\,(x^{(m)} - x) = \ldots = T^{m+1}\,(x^{(0)} - x) = \lambda^{m+1}\,(x^{(0)} - x)\,,$$

što očigledno ne teži ka nuli.

S druge strane, ako je $\rho(T) < 1$, onda možemo izabrati $\varepsilon > 0$, takvo da je $\rho(T) + \varepsilon < 1$, a zatim i (lema 2.7.1) operatorsku normu $||\ \ ||_*$, takvu da je

$$||T||_* \leq \rho(T) + \varepsilon < 1\,, \tag{2.14}$$

pa primenom teoreme 2.7.1 dobijamo traženi rezultat. □

Definicija 2.7.1 *Red konvergencije niza vektora* $x^{(m)}$, $m \in \mathbb{N}_0$, *koji konvergira ka* x *je* p, *ako je*

$$||x^{(m+1)} - x|| \leq c\,||x^{(m)} - x||^p, \quad c \in \mathbb{R}_0^+\,.$$

Ako je $p = 1$, *mora biti* $c < 1$ *(tzv. geometrijska konvergencija s faktorom* c*).*

Red konvergencije ne možemo koristiti za upoređivanje brzina konvergencije različitih metoda jer je konvergencija posmatranih iterativnih metoda linearna, a faktor je $\rho(T)$, tj.

$$||x^{(m+1)} - x||_* \leq \rho(T)\, ||x^{(m)} - x||_* .$$

Logaritmiranjem prethodne nejednakosti dobijamo

$$-\log \rho(T) \leq \log ||x^{(m)} - x||_* - \log ||x^{(m+1)} - x||_* ,$$

pa broj

$$r(T) = -\log \rho(T)$$

možemo definisati kao brzinu konvergencije. Što je $\rho(T)$ manji, to je veća brzina konvergencije.

Videli smo da je centralno pitanje konstrukcije iterativnih metoda kako rastaviti matricu A u obliku $A = M - K$, tako da se $Tx = M^{-1}Kx$ i $c = M^{-1}b$ lako računaju i da je $\rho(T)$ mali.

Da bismo odredili rešenje sistema sa zadanom tačnošću ε potrebna nam je granica apsolutne greške svake iteracije. Iz (2.13) sledi

$$||x^{(n+1)} - x^{(n)}|| \leq ||T||\, ||x^{(n)} - x^{(n-1)}|| \leq ||T||^2\, ||x^{(n-1)} - x^{(n-2)}|| \leq \dots ,$$

odakle dobijamo

$$\begin{aligned}
||x^{(n+m)} - x^{(n)}|| &\leq ||x^{(n+m)} - x^{(n+m-1)} + x^{(n+m-1)} - x^{(n+m-2)} + \dots - x^{(n)}|| \\
&\leq ||x^{(n+m)} - x^{(n+m-1)}|| + \dots + ||x^{(n+1)} - x^{(n)}|| \\
&\leq ||T||^m\, ||x^{(n)} - x^{(n-1)}|| + \dots + ||T||\, ||x^{(n)} - x^{(n-1)}|| \\
&= (||T||^m + \dots + ||T||)\, ||x^{(n)} - x^{(n-1)}|| \\
&= ||T||\, \frac{1 - ||T||^m}{1 - ||T||}\, ||x^{(n)} - x^{(n-1)}|| .
\end{aligned}$$

Stavimo li da $m \to \infty$ dobijamo

$$||x - x^{(n)}|| \leq \frac{||T||}{1 - ||T||} \, ||x^{(n)} - x^{(n-1)}|| \,, \tag{2.15}$$

odakle sledi i

$$||x - x^{(n)}|| \leq \frac{||T||^n}{1 - ||T||} \, ||x^{(1)} - x^{(0)}|| \,, \tag{2.16}$$

Nejednakost (2.16) nam omogućava da odredimo granicu apsolutne greške n-te iteracije već nakon što izračunamo prvu iteraciju, ali je ona dosta grublja od nejednakosti (2.15).

2.7.1 Jacobijeva metoda

Uvedimo sledeće oznake: $D = \operatorname{diag}(a_{11}, \ldots, a_{nn})$,

$$\tilde{L} = \begin{bmatrix} 0 & \ldots & & & 0 \\ a_{21} & 0 & \ldots & & 0 \\ \vdots & a_{32} & 0 & & \vdots \\ & \vdots & & \ddots & \\ a_{n1} & a_{n2} & & a_{n,n-1} & 0 \end{bmatrix}, \; \tilde{U} = \begin{bmatrix} 0 & a_{12} & \ldots & & a_{1n} \\ \vdots & 0 & a_{23} & \ldots & a_{2n} \\ & \vdots & 0 & & \\ & & & \ddots & a_{n-1,n} \\ 0 & 0 & \ldots & & 0 \end{bmatrix}.$$

Jacobijevu metodu dobijamo uzimajući da je $M = D$ i $K = -(\tilde{L} + \tilde{U})$. Ovo ima smisla samo ako su svi a_{ii}, $i = 1, \ldots, n$, različiti od nule. Dakle, $x^{(m+1)}$ računamo formulama

$$x_j^{(m+1)} = \frac{1}{a_{jj}} \left(b_j - \sum_{k \neq j} a_{jk} \, x_k^{(m)} \right), \quad j = 1, \ldots, n.$$

Uočimo da ovu petlju možemo izvesti bilo kojim redom po j jer koordinate novog vektora $x^{(m+1)}$ zavise samo od koordinata starog vektora. Zbog toga je Jacobijeva metoda pogodna za paralelno računanje.

Definicija 2.7.2 *Matrica A je dijagonalno dominantna ako je*

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

Teorema 2.7.3 *Ako je matrica A dijagonalno dominantna, tada Jacobijeva iterativna metoda za rešavanje sistema $A\,x = b$ konvergira ka $x = A^{-1}b$, za svako $x^{(0)}$.*

Dokaz.

$$\rho(T) \leq ||D^{-1}(\tilde{L} + \tilde{U})||_\infty = \max_i \sum_{j \neq i} |a_{ij}/a_{ii}| < 1\,.$$

□

2.7.2 Gauss-Seidelova metoda

Ako koordinate novog vektora u Jacobijevoj metodi računamo redom, od prve ka posljednjoj, odmah se nameće ideja za poboljšanje. Prilikom računanja j-te koordinate u $x^{(m+1)}$ koristimo sve koordinate iz prethodne iteracije, iako već imamo poboljšane koordinate $x_i^{(m+1)}$, za $i < j$, iz novog koraka. Koristeći njih umesto starih dobijamo

$$x_j^{(m+1)} = \frac{1}{a_{jj}} \left(b_j - \sum_{k=1}^{j-1} a_{jk}\, x_k^{(m+1)} - \sum_{k=j+1}^{n} a_{jk}\, x_k^{(m)} \right), \quad j = 1, \ldots, n.$$

Nove koordinate množimo, pored dijagonalnih, i sa elementima ispod dijagonale matrice A, dok stare koordinate množimo samo sa elementima iznad dijagonale. Sada je $M = D + \tilde{L}$, $K = -\tilde{U}$.

Teorema 2.7.4 *Ako je matrica A dijagonalno dominantna, tada Gauss-Seidelova iterativna metoda za rešavanje sistema $A\,x = b$ konvergira ka $x = A^{-1}b$, za svako $x^{(0)}$.*

Dokaz. Neka je λ karakteristična vrednost matrice $T = -(D + \tilde{L})^{-1}\,\tilde{U}$, a x odgovarajući karakteristični vektor. Iz $x \neq 0$ sledi da postoji m, takvo da je $|x_m| = ||x||_\infty > 0$. Iz $Tx = \lambda x$ sledi $-\tilde{U}x = \lambda(D + \tilde{L})x$, odnosno

$$-\sum_{j=m+1}^{n} a_{mj}\, x_j = \lambda\, a_{mm}\, x_m + \lambda \sum_{j=1}^{m-1} a_{mj}\, x_j\,.$$

Pošto je $a_{mm} \neq 0$, važi

$$-\lambda = \lambda \sum_{j=1}^{m-1} \frac{a_{mj}}{a_{mm}} \frac{x_j}{x_m} + \sum_{j=m+1}^{n} \frac{a_{mj}}{a_{mm}} \frac{x_j}{x_m},$$

odakle sledi

$$|\lambda| \leq |\lambda| \sum_{j=1}^{m-1} \left| \frac{a_{mj}}{a_{mm}} \right| + \sum_{j=m+1}^{n} \left| \frac{a_{mj}}{a_{mm}} \right|.$$

Ako bi bilo $|\lambda| \geq 1$, iz poslednje nejednakosti bismo dobili

$$|\lambda| \leq |\lambda| \sum_{j \neq m} \left| \frac{a_{mj}}{a_{mm}} \right| < |\lambda|.$$

Dakle, $\rho(T) < 1$. □

Teorema 2.7.5 *Ako je matrica A simetrična i pozitivno definitna, tada Gauss-Seidelova iterativna metoda za rešavanje sistema $A\,x = b$ konvergira ka $x = A^{-1}b$, za svako $x^{(0)}$.*

Dokaz. Matrica A je simetrična i pozitivno definitna, pa je sada $\tilde{L} = \tilde{U}^T$ i $a_{ii} > 0$, $i = 1, \ldots, n$. Pokazaćemo da su sve karakteristične vrednosti matrice $T = -(D + \tilde{L})^{-1}\,\tilde{L}^T$ po modulu manje od jedan. Matrica

$$T_1 = D^{1/2}\,T\,D^{-1/2}$$

ima iste karakteristične vrednosti kao i T. Ako stavimo $L_1 = D^{-1/2}\,\tilde{L}\,D^{-1/2}$, važi

$$T_1 = -(I + L_1)^{-1}\,L_1^T.$$

Ako je $T_1\,x = \lambda\,x$ i $x^H x = 1$, onda je

$$-L_1^T\,x = \lambda\,(I + L_1)\,x,$$

odnosno

$$-x^H\,L_1^T\,x = \lambda\,(I + x^H\,L_1\,x).$$

Proizvod $x^H\, L_1\, x$ je neki kompleksan broj $a + bi$. Sada je $x^H\, L_1^T\, x = a - bi$, pa imamo

$$|\lambda|^2 = \left|\frac{-a+bi}{1+a+bi}\right|^2 = \frac{a^2+b^2}{1+2a+a^2+b^2}\,.$$

Dokaz završavamo pokazujući da je $2a+1 > 0$. Matrica $W = D^{-1/2}\, A\, D^{-1/2} = I + L_1 + L_1^T$ je pozitivno definitna, pa je

$$0 < x^H\, W\, x = 1 + x^H\, L_1\, x + x^H\, L_1^T\, x = 1 + 2a\,.$$

□

Jasno je da će u slučaju da konvergiraju i Jacobijeva i Gauss-Seidelova metoda, ova druga konvergirati brže. Međutim, postoje slučajevi kad konvergira jedna, a ne konvergira druga, i obratno.

Primer 2.9. Rešimo sledeći sistem Jacobijevom i Gauss-Seidelovom metodom, zokružujući rezultate na četiri decimale

$$\begin{aligned} 1.02\,x_1 - 0.05\,x_2 - 0.10\,x_3 &= 0.795 \\ -0.11\,x_1 + 1.03\,x_2 - 0.05\,x_3 &= 0.849 \\ -0.11\,x_1 - 0.12\,x_2 + 1.04\,x_3 &= 1.398\,. \end{aligned}$$

Jacobijevom metodom iteracije računamo sledećim formulama

$$\begin{aligned} x_1^{(n+1)} &= \frac{1}{1.02}\left[0.05\,x_2^{(n)} + 0.10\,x_3^{(n)} + 0.795\right] \\ x_2^{(n+1)} &= \frac{1}{1.03}\left[0.11\,x_1^{(n)} + 0.05\,x_3^{(n)} + 0.849\right] \\ x_3^{(n+1)} &= \frac{1}{1.04}\left[0.11\,x_1^{(n)} + 0.12\,x_2^{(n)} + 1.398\right]\,. \end{aligned}$$

i polazeći od $x^{(0)} = [0.795, 0.849, 1.398]^T$ dobijamo

$$x^{(1)} = [0.9581, 0.9770, 1.5263]^T\,,$$

$$x^{(2)} = [0.9769, 1.0007, 1.5263]^T\,,$$

$$x^{(3)} = [0.9812, 1.0043, 1.5630]^T ,$$

$$x^{(4)} = [0.9819, 1.0049, 1.5639]^T ,$$

$$x^{(5)} = [0.9820, 1.0050, 1.5640]^T ,$$

$$x^{(6)} = [0.9820, 1.0051, 1, 5641]^T = x^{(7)} .$$

Gauss-Seidelovom metodom koristimo formule

$$\begin{aligned} x_1^{(n+1)} &= \frac{1}{1.02}\left[0.05\, x_2^{(n)} + 0.10\, x_3^{(n)} + 0.795\right] \\ x_2^{(n+1)} &= \frac{1}{1.03}\left[0.11\, x_1^{(n+1)} + 0.05\, x_3^{(n)} + 0.849\right] \\ x_3^{(n+1)} &= \frac{1}{1.04}\left[0.11\, x_1^{(n+1)} + 0.12\, x_2^{(n+1)} + 1.398\right] , \end{aligned}$$

i polazeći od iste početne iteracije dobijamo

$$x^{(1)} = [0.9581, 0.9945, 1.5603]^T ,$$

$$x^{(2)} = [0.9811, 1.0048, 1.5639]^T$$

$$x^{(3)} = [0.9820, 1.0051, 1.5641]^T$$

$$x^{(4)} = [0.9821, 1.0051, 1.5641]^T = x^{(5)} .$$

Kao što smo i očekivali, Gauss-Seidelovom metodom nam je trebalo manje iteracija da bismo došli do rešenja. Razlika u x_1 je posledica zokruživanja.

Glava 3

Interpolacija funkcija i numeričko diferenciranje

Jedan od najvažnijih problema teorije aproksimacija funkcija je interpolacija koja je od velikog teorijskog značaja i u numeričkom diferenciranju, numeričkoj integraciji i drugim oblastima numeričke matematike.

Neka je realna funkcija f zadata svojim vrednostima $f_k = f(x_k)$ $(k = 0, 1, \ldots, n)$ u tačkama $x_k \in [a, b]$. Izvršimo aproksimaciju funkcije f pomoću aproksimacione funkcije

$$\varphi(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + \ldots + a_n\varphi_n(x), \tag{3.1}$$

tako da je uslov za određivanje parametara a_i $(i = 0, 1, \ldots, n)$ dat sa

$$\varphi(x_k) = f_k \quad (k = 0, 1, \ldots, n). \tag{3.2}$$

Od sistema poznatih funkcija $\{\varphi_k\}$ $(k = 0, 1, \ldots, n)$ zahteva se da ispunjava određene osobine. Problem (3.2) naziva se problemom interpolacije funkcije, funkcija φ je interpolaciona funkcija a tačke x_k $(k = 0, 1, \ldots, n)$ su interpolacioni čvorovi, za koje se često pretpostavlja da ispunjavaju uslov

$$a \leq x_0 < x_1 < x_2 < \ldots < x_n \leq b. \tag{3.3}$$

Sistem jednačina (3.2) u ovom slučaju svodi se na sistem linearnih jednačina po parametrima a_i $(i = 0, 1, \ldots, n)$

$$a_0\varphi_0(x_k) + a_1\varphi_1(x_k) + \ldots + a_n\varphi_n(x_k) = f_k \quad (k = 0, 1, \ldots, n),$$

tj.

$$\begin{bmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \ldots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_n(x_1) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_n(x_1) \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{bmatrix}. \tag{3.4}$$

Da bi navedeni interpolacioni problem imao jedinstveno rešenje potrebno je da matrica sistema (3.4) bude regularna. Dakle, na osnovu osobina determinanti zaključujemo da sistemu linearno nezavisnih funkcija $\{\varphi_k\}$ $(k = 0, 1, \ldots, n)$ treba nametnuti takve uslove, pod kojima ne postoji linearna kombinacija

$$a_0\varphi_0(x) + a_1\varphi_1(x) + \ldots + a_n\varphi_n(x)$$

koja ima $n + 1$ različitih nula na $[a, b]$. Sistemi funkcija sa ovom osobinom nazivaju se T-sistemi ili Čebiševljevi sistemi. Za ispitivanje da li je neki sistem Čebiševljev od koristi je sledeća teorema, koju navodimo bez dokaza.

Teorema 3.0.6 *Ako su funkcije $\varphi_k : [a, b] \mapsto \mathbb{R}$ $(k = 0, 1, \ldots, n)$ $n + 1$ puta diferencijabilne i ako je za svako $k = 0, 1, \ldots, n$ determinanta Wronskog W_k različita od nule, tj.*

$$W_k = \begin{vmatrix} \varphi_0(x) & \varphi_1(x) & \ldots & \varphi_k(x) \\ \varphi_0'(x) & \varphi_1'(x) & \ldots & \varphi_k'(x) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_0^{(k)}(x) & \varphi_1^{(k)}(x) & \ldots & \varphi_k^{(k)}(x) \end{vmatrix} \neq 0,$$

sistem funkcija $\{\varphi_k\}$ $(k = 0, 1, \ldots, n)$ je T-sistem.

3.1 Lagrangeova interpolacija

Neka je funkcija f data svojim vrednostima $f_k = f(x_k)$ u tačkama x_k $(k = 0, 1, \ldots, n)$. Pretpostavljamo da čvorovi x_k ispunjavaju uslov (3.3). Ako stavimo da su $\varphi(x) = x^k$ $(k = 0, 1, \ldots, n)$ imamo problem interpolacije funkcije f algebarskim polinomom. Označimo ovaj polinom sa π_n, tj.

$$\pi_n(x) = \varphi(x) = a_0 + a_1 x + \ldots + a_n x^n.$$

Teorema 3.1.1 *Polinom π_n je jedinstven i može se predstaviti u obliku*

$$\pi_n(x) = \sum_{k=0}^{n} f(x_k)\ell_k(x), \tag{3.5}$$

gde je

$$\ell_k(x) = \frac{\omega(x)}{(x - x_k)\omega'(x_k)},$$

a čvorna funkcija

$$\omega(x) = (x - x_0)(x - x_1)\cdots(x - x_n).$$

Dokaz. Pošto je u ovom slučaju determinanta matrice sistema (3.4) Vandermondova, tj.

$$\begin{vmatrix} 1 & x_0 & \ldots & x_0^n \\ 1 & x_1 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & \ldots & x_n^n \end{vmatrix} = \prod_{i>j}(x_i - x_j),$$

i s obzirom na pretpostavku (3.3), zaključujemo da je polinom π_n jedinstven. Jedinstvenost sledi i na osnovu toga što je sistem funkcija $\{1, x, x^2, \ldots, x^n\}$ T-sistem, pošto je determinanta Wronskog $W_k = 0!1!2!\cdots k! \neq 0$ $(k = 0, 1, \ldots, n)$.

Dokaz formule (3.5) sledi na osnovu činjenica:

- Svi polinomi ℓ_k su ne višeg stepena od n, tj. $\deg(\ell_k) \leq n$ za svako k,
- $\ell_k(x_i) = \delta_{ik}$, gde je δ_{ik} Kroneckerovo delta, tj.

$$\delta_{ik} = \begin{cases} 1, & \text{ako je } i = k, \\ 0, & \text{ako je } i \neq k, \end{cases}$$

- $\pi_n(x_k) = f(x_k)$ za svako k,
- $\deg(\pi_n) \leq n$.

□

Primetimo da se $\ell_k(x)$ može zapisati i na sledeći način

$$\ell_k(x) = \frac{(x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)}.$$

Teorema 3.1.2 *Neka je $f \in C^{n+1}[a, b]$ i $x_i \in [a, b]$ $(i = 0, 1, \ldots, n)$. Tada postoji $\xi \in (a, b)$ takvo da je greška Lagrangeovog interpolacionog polinoma*

$$r_n(f; x) = f(x) - \pi_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\, \omega(x). \tag{3.6}$$

Dokaz. Posmatrajmo funkciju g, definisanu sa

$$g(x) = f(x) - \pi_n(x) - \kappa_n\, \omega(x), \tag{3.7}$$

koja se anulira u tačkama $x_0, x_1, \ldots, x_n$. Neka je $\bar{x}$ proizvoljna tačka iz $[a, b]$ takva da je $\bar{x} \neq x_i$ $(i = 0, 1, \ldots, n)$. Odredimo konstantu κ_n u (3.7) tako da bude $g(\bar{x}) = 0$. Kako je $\bar{x} \neq x_i$ $(i = 0, 1, \ldots, n)$, ovakva vrednost za κ_n postoji. Naime,

$$\kappa_n = \frac{f(\bar{x}) - \pi_n(\bar{x})}{\omega(\bar{x})}.$$

Kako funkcija g na $[a, b]$ ima bar $n+2$ različite nule, to na osnovu Rolleove teoreme sukcesivno zaključujemo da u (a, b) $g'(x)$ ima bar $n+1$ različitih nula, itd., $g^{(n+1)}(x)$ ima bar jednu nulu.

Neka je $\xi \in (a, b)$ nula funkcije $g^{(n+1)}$ za koju smo upravo pokazali da postoji. Tada iz (3.7) sledi da

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

odakle je

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

Na osnovu prethodnog zaključujemo da je

$$r_n(f;\bar{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\omega(\bar{x}).$$

Kako je $\bar{x}$ proizvoljna tačaka iz $[a,b]$ dokaz je završen. □

Iz (3.6) dobijamo da je jedna gornja ocena ostatka u Lagrangeovoj interpolaciji

$$|r_n(f;x)| = |f(x) - \pi_n(x)| \leq \frac{M_{n+1}}{(n+1)!} \max_{x\in[a,b]} |\omega(x)|,$$

gde je

$$M_k = \max_{x\in[a,b]} \left|f^{(k)}(x)\right| \quad (k \in \mathbb{N}).$$

Teorema 3.1.3 *Neka je* $f \in C^{n+1}[a,b]$, $x_k - x_{k-1} = h = \text{const}$ $(k = 1,\ldots,n)$, $x_0 = a, x_n = b$. *Tada je*

$$|r_n(f;x)| = |f(x) - \pi_n(x)| < \frac{h^{n+1}}{4(n+1)} M_{n+1}.$$

Dokaz. Ako stavimo $x = x_0 + ph$ $(p \in [0,n])$ imamo

$$\omega(x) = p(p-1)\cdots(p-n)h^{n+1}.$$

Kako funkcija $F_n(p) = |p(p-1)\cdots(p-n)|$ na intervalu $[0,n]$ dostiže maksimalnu vrednost kada je $p \in (0,1)$ (ili $p \in (n-1,n)$), važi

$$\max_{p\in[0,n]} F_n(p) < \max_{p\in(0,1)} |p(p-1)| \max_{p\in(0,1)} |(p-2)\cdots(p-n)| < \frac{1}{4}\,n!\,,$$

odakle neposredno sledi tvrđenje teoreme. □

Osobina $x_k - x_{k-1} = h = \text{const}$ $(k = 1,\ldots,n)$ ukazuje da se čvorovi nalaze na istom odstojanju, kažemo da su ekvidistantni. U ovom slučaju nije teško primetiti da za funkciju $\omega(x)$ važi

$$\omega\left(\frac{n}{2} + x\right) = (-1)^{n+1}\,\omega\left(\frac{n}{2} - x\right).$$

PRIMER 4.1. Odredimo sa kojom se tačnošću može izračunati vrednost $\log 11.5$ korišćenjem Lagrangeovog interpolacionog polinoma π_3 ako su poznate vrednosti

$$\log 10, \quad \log 11, \quad \log 12, \quad \log 13.$$

S obzirom da je

$$M_4 = \max_{x\in[10,13]} \left|\frac{-6}{x^4}\right| = 6 \cdot 10^{-4},$$

imamo

$$\begin{aligned} |\log 11.5 - \pi_3(11.5)| &\leq \frac{M_4}{4!} |(11.5-10)(11.5-11)(11.5-12)(11.5-13)| \\ &= 1.41 \cdot 10^{-5}. \end{aligned}$$

Inače za proizvoljno $x \in (10, 13)$ važi ocena

$$|\log x - \pi_3(x)| \leq \frac{1}{24} \cdot 6 \cdot 10^{-4} = 2.5 \cdot 10^{-5}.$$

Kada nije potreban opšti izraz za interpolacioni polinom, već samo vrednost za neko konkretno x, koristi se Aitkenova shema, koja se sastoji u sukcesivnoj primeni sledećih izraza:

$$\begin{aligned} A_k &= f(x_k) \quad (k = 0, 1, \ldots, n); \\ A_{k-1,k} &= \frac{1}{x_k - x_{k-1}} \begin{vmatrix} A_{k-1} & x_{k-1} - x \\ A_k & x_k - x \end{vmatrix} \quad (k = 1, 2, \ldots, n); \\ &\vdots \\ A_{0,1,\ldots,n} &= \frac{1}{x_n - x_0} \begin{vmatrix} A_{0,1,\ldots,n-1} & x_0 - x \\ A_{1,2,\ldots,n} & x_n - x \end{vmatrix}; \end{aligned}$$

pri čemu je

$$\pi_n(x) = A_{0,1,\ldots,n}.$$

U praksi se često javlja zadatak određivanja vrednosti argumenta na osnovu zadate vrednosti funkcije. Ovaj se zadatak rešava metodom inverzne interpolacije. Ako je data funkcija monotona, zadatak inverzne interpolacije najjednostavnije se rešava međusobnom zamenom vrednosti funkcije i argumenta, a zatim konstrukcijom interpolacionog polinoma. U problemima inverzne interpolacije pogodno je koristiti se Aitkenovom shemom.

PRIMER 4.2. Data je funkcija $f(x)$ skupom podataka

x	14	17	31	35
$f(x)$	68.7	64.0	44.0	39.1

Bez konstrukcije interpolacionog polinoma odredićemo približno $f^{-1}(54.0)$.

Tablica za inverznu funkciju je

y	68.7	64.0	44.0	39.1
$f^{-1}(y)$	14	17	31	35

Primenimo Aitkenovu shemu.

Kako je $A_k = f^{-1}(y_k)$ $(k = 0, 1, 2, 3)$, to je

$$A_0 = 14, \quad A_1 = 17, \quad A_2 = 31, \quad A_3 = 35.$$

Kako je

$$A_{k-1,k} = \frac{1}{y_k - y_{k-1}} \begin{vmatrix} A_{k-1} & y_{k-1} - y \\ A_k & y_k - y \end{vmatrix} \quad (k = 1, 2, 3),$$

to je $(y = 54)$

$$A_{0,1} = 23.383, \quad A_{1,2} = 24., \quad A_{2,3} = 22.837.$$

Kako je

$$A_{k-1,k,k+1} = \frac{1}{y_{k+1} - y_{k-1}} \begin{vmatrix} A_{k-1,k} & y_{k-1} - y \\ A_{k,k+1} & y_{k+1} - y \end{vmatrix} \quad (k = 1, 2),$$

to je

$$A_{0,1,2} = 23.75, \quad A_{1,2,3} = 23.533.$$

Konačno, imamo

$$A_{0,1,2,3} = \frac{1}{y_3 - y_0} \begin{vmatrix} A_{0,1,2} & y_0 - y \\ A_{1,2,3} & y_3 - y \end{vmatrix} = 23.642.$$

Dakle, $f^{-1}(54.0) \approx 23.6$.

3.1.1 Optimalni izbor interpolacionih čvorova

Ovu podsekciju počinjemo definisanjem Čebiševljevih polinoma T_n,

$$T_n(x) = \cos(n \arccos x), \quad |x| \leq 1, \ n \in \mathbb{N}_0.$$

Za $n = 0$, na osnovu prethodne formule, imamo da je

$$T_0(x) = 1.$$

Za $n = 1$ imamo da je

$$T_1(x) = \cos(\arccos x) = x.$$

Za $n = 2$ imamo da je

$$\begin{aligned} T_2(x) &= \cos(2 \arccos x) = \cos^2(\arccos x) - \sin^2(\arccos x) \\ &= x^2 - (1 - \cos^2(\arccos x)) \\ &= 2x^2 - 1. \end{aligned}$$

Kako važi

$$\begin{aligned}\cos(n+1)\theta &= \cos(n\theta)\cos\theta - \sin(n\theta)\sin\theta \\ &= \cos(n\theta)\cos\theta + \frac{1}{2}[\cos(n+1)\theta - \cos(n-1)\theta] \\ &= \cos(n\theta)\cos\theta + \frac{1}{2}[\cos(n+1)\theta + \cos(n-1)\theta] - \cos(n-1)\theta \\ &= 2\cos(n\theta)\cdot\cos\theta - \cos(n-1)\theta,\end{aligned}$$

supstitucijom $\theta = \arccos x$ dobijamo

$$\cos[(n+1)\arccos x] = 2x\cdot\cos(n\arccos x) - \cos[(n-1)\arccos x],$$

tj. dobijamo rekurentnu relaciju kojom su povezani članovi niza Čebiševljevih polinoma

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)\,.$$

Poslednja rekurentna relacija služi nam za dalje određivanje niza Čebiševljevih polinoma:

$$T_3(x) = 4x^3 - 3x, \quad T_4(x) = 8x^4 - 8x^2 + 1, \quad \text{itd.}$$

Zaključujemo da je Čebiševljev polinom n-tog stepena

$$T_n(x) = 2^{n-1}x^n + \dots\,,$$

te da je njegov monični (vodeći koeficijent mu je 1) polinom dat sa

$$\overline{T}_n = \frac{1}{2^{n-1}}\,T_n(x).$$

Nule Čebiševljevog polinoma n-tog stepena T_n određujemo rešavanjem jednačine

$$\cos(n\arccos x) = 0.$$

Nije teško zaključiti da su rešenja (koreni) ove jednačine

$$\overline{x}_k = \cos\frac{(2k+1)\pi}{2n} \quad (k = 0, 1, \dots, n-1).$$

Dalje, očigledno je $|T_n(x)| \leq 1,\ x \in [-1,1]$. T_n dostiže ekstremne vrednosti za

$$\cos(n \arccos x) = \mp 1\,,$$

odnosno u tačkama

$$x_k = \cos\frac{k\pi}{n} \quad (k = 0, 1, \ldots, n),$$

i važi

$$T_n(x_k) = (-1)^k \quad (k = 0, 1, \ldots, n).$$

Podsetimo se da smo čvornu, odnosno funkciju greške definisali sa $\omega(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$. Dokazaćemo sada da važi sledeći rezultat.

Teorema 3.1.4 *Veličina*

$$\max_{x\in[-1,1]} |\omega(x)|$$

dostiže najmanju vrednost ako se za čvorove interpolacije izaberu koreni Čebiševljevog polinoma T_{n+1}, tj.

$$\max_{x\in[-1,1]} |\omega(x)| \geq \frac{1}{2^n} = \max_{x\in[-1,1]} |\overline{T}_{n+1}(x)|.$$

Dokaz. Ako pretpostavimo suprotno, tj. da postoji neka od funkcija greške $\widetilde{\omega}$ za koju je

$$\max_{x\in[-1,1]} |\widetilde{\omega}(x)| < \frac{1}{2^n},$$

tada polinom

$$Q_n(x) = \overline{T}_{n+1}(x) - \widetilde{\omega}(x),$$

stepena $\deg(Q_n) \leq n$, menja znak u svakoj tački

$$x_k = \cos\frac{k\pi}{n+1} \quad (k = 0, 1, \ldots, n+1),$$

što znači da polinom Q_n ima $(n+1)$ nula, što je nemoguće. □

Neka je interval $[a, b]$ na kome interpoliramo funkciju f ovde $[-1, 1]$, čime ne umanjujemo opštost razmatranja jer se lako može odrediti linearna funkcija koja preslikava jedan interval na drugi i obrnuto. Razmotrimo problem optimalnog izbora interpolacionih čvorova kod interpolacije funkcije iz klase

$$CM = \{f|\ f \in C^{n+1}[-1,1] \wedge \left|f^{(n+1)}\right| \leq M\ (\forall x \in [-1,1])\}.$$

Naime, odredićemo interpolacione čvorove $x_0, x_1, \ldots, x_n$ tako da veličina

$$\sup_{f \in CM} \max_{x \in [-1,1]} |r_n(f;x)| \tag{3.8}$$

ima najmanju vrednost. Tada je na osnovu (3.6)

$$\max_{x \in [-1,1]} |r_n(f;x)| \leq \frac{M}{(n+1)!} \max_{x \in [-1,1]} |\omega(x)|. \tag{3.9}$$

Nije teško uočiti da u poslednjoj nejednakosti nastupa jednakost ako je f polinom

$$f(x) = \frac{M}{(n+1)!}\, x^{n+1} + d_0 x^n + d_1 x^{n-1} + \ldots + d_n,$$

gde su d_i $(i = 0, 1, \ldots, n)$ proizvoljni realni koeficijenti.

Kako desna strana u nejednakosti (3.9) ne zavisi od f, to je

$$\sup_{f \in CM} \max_{x \in [-1,1]} |r_n(f;x)| \leq \frac{M}{(n+1)!} \max_{x \in [-1,1]} |\omega(x)|.$$

Postavljeni problem očigledno se svodi na određivanje polinoma

$$\omega(x) = (x - x_0)(x - x_1) \cdots (x - x_n),$$

čije nule leže u $[-1, 1]$ i za koji je $\max_{x \in [-1,1]} |\omega(x)|$ minimalan.

Rešenje poslednjeg problema je (Teorema 3.1.4)

$$\omega(x) = \frac{1}{2^n} T_{n+1}(x) = \frac{1}{2^n} \cos[(n+1) \arccos x].$$

Dakle, interpolacioni čvorovi pri kojima veličina (3.8) ima najmanju vrednost su nule Čebiševljevog polinoma $(n+1)$-og stepena T_{n+1}, tj.

$$x_k = \cos \frac{2k+1}{2(n+1)}\pi \quad (k = 0, 1, \ldots, n).$$

Interpolacija kod koje su ovako uzeti interpolacioni čvorovi naziva se Čebiševljevom interpolacijom. Za ostatak u opštem slučaju ($f \in CM$) važi

$$|r_n(f; h)| \leq \frac{M}{2^n \cdot (n+1)!}\,,$$

odnosno, možemo dati strožiju ocenu u konkretnom slučaju, ako znamo M_{n+1},

$$|r_n(f; h)| \leq \frac{M_{n+1}}{2^n \cdot (n+1)!}\,.$$

3.2 Newtonova interpolacija sa podeljenim razlikama

Za funkciju f zadatu vrednostima $f_k = f(x_k)$ u tačkama x_k $(k = 0, 1, \ldots, n)$, definisaćemo najpre podeljene razlike.

Količnik

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

nazivamo podeljenom razlikom prvog reda (funkcije f u tačkama x_0, x_1) i označavamo sa $[x_0, x_1; f]$.

Podeljena razlika reda r definiše se rekurzivno pomoću

$$[x_0, x_1, \ldots, x_r; f] = \frac{[x_1, \ldots, x_r; f] - [x_0, \ldots, x_{r-1}; f]}{x_r - x_0}, \tag{3.10}$$

pri čemu je $[x; f] \equiv f(x)$.

Relacija (3.10) omogućava konstrukciju tablice podeljenih razlika

$$\begin{array}{llllll}
x_0 & f_0 & & & & \\
 & & [x_0,x_1;f] & & & \\
x_1 & f_1 & & [x_0,x_1,x_2;f] & & \\
 & & [x_1,x_2;f] & & [x_0,x_1,x_2,x_3;f] & \\
x_2 & f_2 & & [x_1,x_2,x_3;f] & & \\
 & & [x_2,x_3;f] & & & \\
x_3 & f_3 & & & & \\
\vdots & \vdots & & & &
\end{array}$$

Može se pokazati da podeljena razlika reda r ima osobinu linearnosti, tj. da je

$$[x_0, x_1, \ldots, x_r; c_1 f + c_2 g] = c_1[x_0, \ldots, x_r; f] + c_2[x_0, \ldots, x_r; g],$$

gde su c_1, c_2 proizvoljne konstante, a f, g funkcije.

Na osnovu definicije podeljene razlike prvog reda važi

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

Matematičkom indukcijom se lako pokazuje da važi formula

$$[x_0, x_1, \ldots, x_r; f] = \sum_{k=0}^{r} \frac{f(x_k)}{\omega'(x_k)},$$

gde je $\omega(x) = (x - x_0)(x - x_1) \cdots (x - x_r)$.

Indukcijom se može dokazati i sledeći rezultat.

Teorema 3.2.1 *Neka je $f \in C^n[a, b]$ i neka je ispunjen uslov* (3.3)*. Tada, za svako $r \leq n$ važi formula*

$$[x_0, x_1, \ldots, x_r; f] = \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{r-1}} f^{(r)} \left(x_0 + \sum_{i=1}^{r} (x_i - x_{i-1}) t_i \right) dt_1 \cdots dt_r \,.$$

Na osnovu teoreme o srednjoj vrednosti integrala, iz poslednje jednakosti sledi

$$\begin{aligned}
[x_0, x_1, \ldots, x_r; f] &= f^{(r)}(\xi) \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{r-1}} dt_1 dt_2 \cdots dt_r \\
&= \frac{1}{r!} f^{(r)}(\xi) \quad (a < \xi < b).
\end{aligned}$$

Uzimajući u poslednjoj jednakosti da $x_i \to x_0$ $(i = 1, \ldots, r)$ zaključujemo da

$$[x_0, x_1, \ldots, x_r; f] \to \frac{1}{r!} f^{(r)}(x_0). \tag{3.11}$$

Izrazimo sada vrednost funkcije $f(x_r)$ $(r \leq n)$ pomoću podeljenih razlika $[x_0, \ldots, x_i; f]$ $(i = 0, 1, \ldots, r)$.

Za $r = 1$, na osnovu definicije podeljene razlike prvog reda, imamo

$$f(x_1) = f(x_0) + (x_1 - x_0)[x_0, x_1; f].$$

Analogno, za $r = 2$,

$$\begin{aligned} f(x_2) &= f(x_1) + (x_2 - x_1)[x_1, x_2; f] \\ &= (f(x_0) + (x_1 - x_0)[x_0, x_1; f]) \\ &+ (x_2 - x_1)\left([x_0, x_1; f] + (x_2 - x_0)[x_0, x_1, x_2; f]\right), \end{aligned}$$

tj.

$$f(x_2) = f(x_0) + (x_2 - x_0)[x_0, x_1; f] + (x_2 - x_0)(x_2 - x_1)[x_0, x_1, x_2; f].$$

U opštem slučaju važi

$$\begin{aligned} f(x_r) &= f(x_0) + (x_r - x_0)[x_0, x_1; f] + (x_r - x_0)(x_r - x_1)[x_0, x_1, x_2; f] \\ &+ \ldots + (x_r - x_0)(x_r - x_1) \cdots (x_r - x_{r-1})[x_0, x_1, \ldots, x_r; f]. \end{aligned} \tag{3.12}$$

Korišćenjem podeljenih razlika može se konstruisati interpolacioni polinom za skup podataka $\{(x_k, f_k)\}$ $(k = 0, 1, \ldots, n)$. Ovaj polinom ima oblik

$$\begin{aligned} \pi_n(x) &= f(x_0) + (x - x_0)[x_0, x_1; f] + (x - x_0)(x - x_1)[x_0, x_1, x_2; f] \\ &+ \ldots + (x - x_0)(x - x_1) \cdots (x - x_{n-1})[x_0, x_1, \ldots, x_n; f] \end{aligned}$$

i naziva se Newtonov interpolacioni polinom (sa podeljenim razlikama).

Da bismo dokazali poslednju formulu dovoljno je primetiti da je $\pi_n(x_r) = f(x_r)$ $(r = 0, 1, \ldots, n)$ i da je $\deg(\pi_n(x)) \leq n$. Poslednje tvrđenje sada sledi iz (3.12).

S obzirom na jedinstvenost algebarskog interpolacionog polinoma zaključujemo da je Newtonov interpolacioni polinom ekvivalentan sa Lagrangeovim. Konstrukcija Newtonovog interpolacionog polinoma zahteva prethodno formiranje tablice podeljenih razlika, što nije bio slučaj kod Lagrangeove interpolacije. S druge strane, kada hoćemo da smanjimo grešku u interpolaciji uvođenjem novog interpolacionog čvora, Newtonov polinom je znatno pogodniji od Lagrangeovog, jer ne zahteva ponavljanje celog računskog postupka, pošto kod Newtonove interpolacije imamo

$$\pi_{n+1}(x) = \pi_n(x) + (x - x_0)(x - x_1) \cdots (x - x_n)[x_0, x_1, \ldots, x_n, x_{n+1}; f].$$

Teorema 3.2.2 *Ako funkcija f ima konačne vrednosti u tačkama*

$$x_0, x_1, \ldots, x_n, x,$$

onda važi formula

$$r_n(f; x) = f(x) - \pi_n(x) = \omega(x)\,[x_0, x_1, \ldots, x_n, x; f], \tag{3.13}$$

gde je $\omega(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$.

Dokaz. Stavimo $r = n + 1$ i $x_r = x$. Tada iz (3.12) sledi da je

$$f(x) = \pi_n(x) + (x - x_0)(x - x_1) \cdots (x - x_n)\,[x_0, x_1, \ldots, x_n, x; f],$$

čime je dokaz završen. □

Kako je

$$[x_0, x_1, \ldots, x_n, x; f] = \frac{1}{(n+1)!}\, f^{(n+1)}(\xi) \quad (\xi \in (a, b)),$$

zaključujemo da se ostatak u (3.13) može svesti na Lagrangeov oblik.

Ako u Newtonovom interpolacionom polinomu π_n uzmemo da $x_i \to x_0$ $(i = 1, \ldots, n)$, na osnovu (3.11) zaključujemo da se on svodi na Taylorov polinom.

PRIMER 4.3. Aproksimirajmo funkciju $f(x) = e^x$, na segmentu [0, 0.5], interpolacionim polinomom.

Funkciju $f(x) = e^x$ aproksimiraćemo Lagrangeovim i Newtonovim interpolacionim polinomom, na osnovu sledećih podataka

k	0	1	2
x_k	0.0	0.2	0.5
$f(x_k)$	1.000000	1.221403	1.648721

Lagrangeov interpolacioni polinom za ovaj skup podataka glasi

$$\begin{aligned} \pi_2(x) &= 1\,\frac{(x-0.2)(x-0.5)}{(0-0.2)(0-0.5)} + 1.221403\,\frac{(x-0.2)(x-0.5)}{(0.2-0)(0.2-0.5)} \\ &\quad +1.648721\,\frac{(x-0)(x-0.2)}{(0.5-0)(0.5-0.2)} \\ &= 0.634757\,x^2 + 0.980064\,x + 1, \end{aligned} \tag{3.14}$$

pri čemu su svi rezultati zaokruženi na šest decimala.

U cilju konstruisanja Newtonovog interpolacionog polinoma najpre formiramo, na osnovu prethodne tabele, tablicu podeljenih razlika

k	$[x_k; f]$	$[x_k, x_{k+1}; f]$	$[x_k, x_{k+1}, x_{k+2}; f]$
0	1.000000		
		1.107015	
1	1.221403		0.634756
		1.424393	
2	1.648721		

Dakle, Newtonov interpolacioni polinom je

$$\begin{aligned} \pi_2(x) &= 1 + 1.107015\,(x-0) + 0.634756\,(x-0)(x-0.2) \\ &= 0.634756\,x^2 + 0.980064\,x + 1\,, \end{aligned} \tag{3.15}$$

pri čemu su svi rezultati zaokruženi na šest decimala.

Kao što je rečeno, teorijski, interpolacioni polinom je jedinstven. Prema tome, Lagrangeov interpolacioni polinom i Newtonov bi trebalo da budu identički jednaki. Međutim, upoređivanjem odgovarajućih izraza uočavamo da se koeficijenti uz x^2 razlikuju za 10^{-6}. To je posledica grešaka zaokruživanja koje se neminovno javljaju u procesu izračunavanja na računskim mašinama. Zbog toga se, zavisno od svrhe, često daje prednost interpolacionom polinomu dobijenom na jedan način u odnosu na interpolacioni polinom dobijen na neki drugi način.

Primetimo još jednom da konstrukcija Newtonovog interpolacionog polinoma zahteva prethodno formiranje tablice podeljenih razlika, što nije bio slučaj kod Lagrangeove interpolacije.

S obzirom da je $f^{(k)}(x) = (e^x)^{(k)} = e^x$ $(k = 1, 2, \ldots)$, na osnovu (3.6) imamo

$$|f(x) - \pi_2(x)| \le \frac{M_3}{3!}\,|\,x\,(x-0.2)\,(x-0.5)\,| \qquad (0 \le x \le 0.5)\,,$$

gde je

$$M_3 = \max_{x\in[0,0.5]} |e^x| = e^{0.5} \approx 1.648721\,.$$

Ako hoćemo da smanjimo grešku interpolacionog polinoma, to najjednostavnije možemo učiniti uvođenjem novog interpolacionog čvora. Izaberimo, na primer, $x_3 = 0.4$, pa je $f(x_3) = 1.491825$. Za tu svrhu, konstatujmo opet, Newtonov interpolacioni polinom je znatno pogodniji od Lagrangeovog, jer ne zahteva ponavljanje celog računskog postupka. Naime korišćenjem Newtonove interpolacije, imamo

$$\pi_3(x) = \pi_2(x) + (x-x_0)(x-x_1)(x-x_2)[x_0, x_1, x_2, x_3; f]\,.$$

Dakle, dopunimo poslednju tablicu konačnih razlika novouvedenim interpolacionim čvorom x_3:

k	$[x_k; f]$	$[x_k, x_{k+1}; f]$	$[x_k, x_{k+1}, x_{k+2}; f]$	$[x_k, x_{k+1}, x_{k+2}, x_{k+3}; f]$
0	1.000000			
		1.107015		
1	1.221403		0.634756	
		1.424393		0.220198
2	1.648721		0.722835	
		1.568960		
3	1.491825			

Odavde je

$$\begin{aligned} \pi_3(x) &= \pi_2(x) + 0.220198\, x\,(x-0.2)\,(x-0.5) \\ &= 0.220198\, x^3 + 0.480618\, x^2 + 1.002084\, x + 1\,. \end{aligned}$$

Na osnovu (3.6) imamo

$$|f(x) - \pi_3(x)| \le \frac{M_4}{4!}\,|\,x\,(x-0.2)\,(x-0.5)\,(x-0.4)\,| \qquad (0 \le x \le 0.5).$$

Na primer, za $x = 0.3$ je

$$|f(0.3) - \pi_2(0.3)| = 0.001288$$

i

$$|f(0.3) - \pi_3(0.3)| = 0.000033\,.$$

3.3 Njutnove interpolacione formule sa konačnim razlikama

Uvedimo nekoliko operatora konačnih razlika ili diferencnih operatora (definisanih recimo na skupu neprekidnih funkcija) koji će nam biti od pomoći

u konstrukciji odgovarajućih interpolacionih formula. Sa ovim operatorima sprovodi se formalni račun zasnovan na pravilima algebre i analize, i kao takav se često pokazuje kao elegantno sredstvo za formiranje raznih aproksimacionih formula. Taj se račun naziva računom konačnih razlika.

Najpre definišemo operator prednje razlike sa

$$\Delta f(x) = f(x+h) - f(x) \quad (h = \text{const} > 0),$$

i njegov iterirani operator, tj. stepen operatora Δ,

$$\begin{aligned} \Delta^k f(x) &= \Delta^{k-1} f(x+h) - \Delta^{k-1} f(x) \quad (k \in \mathbb{N}), \\ \Delta^0 f(x) &= f(x). \end{aligned}$$

Matematičkom indukcijom lako se mogu dokazati formule

$$\Delta^k f(x) = \sum_{i=0}^{k} (-1)^i \binom{k}{i} f(x + (k-i)h), \tag{3.16}$$

$$f(x+kh) = \sum_{i=0}^{k} \binom{k}{i} \Delta^i f(x). \tag{3.17}$$

Na potpuno analogan način se definiše i operator zadnje razlike ∇,

$$\nabla f(x) = f(x) - f(x-h)$$

i njegov stepen.

Pored ova dva uvodimo identički operator 1,

$$1 f(x) = f(x),$$

i operator pomeranja E,

$$E f(x) = f(x+h).$$

Kako je $E^k f(x) = E(E^{k-1} f(x)) = f(x+kh)$ $(k \in \mathbb{N})$, tj. kako iterirani operator E^k primenjen na $f(x)$ daje $f(x+kh)$, logično se nameće definisanje stepena operatora za proizvoljni izložilac sledećom jednakošću

$$E^p f(x) = f(x+ph) \quad (p \in \mathbb{R}).$$

Kako je

$$\Delta f(x) = f(x+h) - f(x) = Ef(x) - 1f(x) = (E-1)f(x),$$

imamo formalnu vezu

$$\Delta = E - 1 \qquad \text{ili} \qquad E = 1 + \Delta.$$

Primenom binomne formule na desne strane u poslednjim jednakostima dobijamo

$$\Delta^k = (E-1)^k = \sum_{i=0}^{k} (-1)^i \binom{k}{i} E^{k-i}, \quad E^k = (1+\Delta)^k = \sum_{i=0}^{k} \binom{k}{i} \Delta^i,$$

odakle formalno slede formule (3.16), (3.17).

Primenom računa konačnih razlika mogu se izvesti razne interpolacione formule sa ekvidistantnim čvorovima. Sada ćemo najpre izvesti prvu Newtonovu interpolacionu formulu.

Neka je funkcija f data na $[a,b]$ parovima vrednosti (x_k, f_k), gde je $f_k = f(x_k)$ i $x_k = x_0 + kh$ $(k = 0, 1, \ldots, n)$.

Za dati skup može se formirati tablica konačnih razlika, konkretno ovde tablica prednjih razlika.

$$\begin{array}{llllll}
x_0 & f_0 & & & & \\
 & & \Delta f_0 & & & \\
x_1 & f_1 & & \Delta^2 f_0 & & \\
 & & \Delta f_1 & & \Delta^3 f_0 & \\
x_2 & f_2 & & \Delta^2 f_1 & & \Delta^4 f_0 \\
 & & \Delta f_2 & & \Delta^3 f_1 & \\
x_3 & f_3 & & \Delta^2 f_2 & & \\
 & & \Delta f_3 & & & \\
x_4 & f_4 & & & & \\
\vdots & \vdots & & & &
\end{array}$$

Neka je

$$x = x_0 + ph \quad (0 \le p \le n), \text{ tj.} \quad p = \frac{x - x_0}{h}.$$

Kako je

$$E^p = (1 + \Delta)^p = \sum_{k=0}^{\infty} \binom{p}{k} \Delta^k,$$

imamo

$$f(x) = (1 + \Delta)^p f_0 = E^p f_0 = \sum_{k=0}^{\infty} \binom{p}{k} \Delta^k f_0 = \sum_{k=0}^{n} \binom{p}{k} \Delta^k f_0 + r_n(f; x),$$

tj.

$$f(x) = f(x_0 + ph) = \sum_{k=0}^{n} \binom{p}{k} \Delta^k f_0 + r_n(f; x), \tag{3.18}$$

gde je ostatak $r_n(f; x)$, s obzirom na jedinstvenost interpolacionog polinoma, isti kao kod Lagrangeove interpolacione formule, tj.

$$r_n(f; x) = \frac{h^{n+1}}{(n+1)!} p(p-1) \cdots (p-n) f^{(n+1)}(\xi),$$

gde je ξ tačka iz intervala (x_0, x_n).

Polinom

$$\pi_n(x) = \sum_{k=0}^{n} \binom{p}{k} \Delta^k f_0 \quad (ph = x - x_0), \tag{3.19}$$

dobijen na ovaj način, naziva se prvi Newtonov interpolacioni polinom. Ovaj polinom može se definisati i rekurzivno pomoću

$$\pi_k(x) = \pi_{k-1}(x) + \binom{p}{k}\Delta^k f_0 \quad (k = 1, \ldots, n),$$

polazeći od $\pi_0(x) = f_0$.

Polinom (3.19) se može predstaviti u razvijenom obliku

$$\pi_n(x) = f_0 + p\Delta f_0 + \frac{p(p-1)}{2!}\,\Delta^2 f_0 + \ldots + \frac{p(p-1)\cdots(p-n+1)}{n!}\,\Delta^n f_0,$$

tj.

$$\begin{aligned}\pi_n(x) &= f_0 + \frac{\Delta f_0}{h}(x - x_0) + \frac{\Delta^2 f_0}{2!h^2}(x - x_0)(x - x_1) + \ldots \\ &+ \frac{\Delta^n f_0}{n!h^n}(x - x_0)(x - x_1)\cdots(x - x_{n-1}).\end{aligned}$$

Prvi Newtonov interpolacioni polinom (sa konačnim razlikama) se koristi u slučajevima kada se interpolacija izvodi na početku intervala, tj. u okolini tačke x_0. Ako se on koristi za približno izračunavanje vrednosti funkcije f za $x < x_0$, kažemo da se radi o ekstrapolaciji funkcije.

Kako je

$$\nabla f(x) = f(x) - f(x-h) = 1f(x) - E^{-1}f(x) = (1 - E^{-1})f(x),$$

imamo formalnu vezu

$$\nabla = 1 - E^{-1}, \text{ tj.} \quad E = (1 - \nabla)^{-1}. \tag{3.20}$$

Koristeći se operatorom zadnje razlike ∇ može se formirati tablica konačnih razlika.

$$
\begin{array}{llllll}
\vdots & \vdots & & & & \\
x_{n-4} & f_{n-4} & & & & \\
 & & \nabla f_{n-3} & & & \\
x_{n-3} & f_{n-3} & & \nabla^2 f_{n-2} & & \\
 & & \nabla f_{n-2} & & \nabla^3 f_{n-1} & \\
x_{n-2} & f_{n-2} & & \nabla^2 f_{n-1} & & \nabla^4 f_n \\
 & & \nabla f_{n-1} & & \nabla^3 f_n & \\
x_{n-1} & f_{n-1} & & \nabla^2 f_n & & \\
 & & \nabla f_n & & & \\
x_n & f_n & & & &
\end{array}
$$

Neka je

$$x = x_n + ph \quad (0 \leq -p \leq n), \text{ tj.} \quad p = \frac{x - x_n}{h}.$$

Kako je, na osnovu (3.20),

$$E^p = (1 - \nabla)^{-p} = \sum_{k=0}^{\infty} (-1)^k \binom{-p}{k} \nabla^k,$$

imamo

$$\begin{aligned} f(x) = f(x_n + ph) &= \sum_{k=0}^{\infty} (-1)^k \binom{-p}{k} \nabla^k f_n \\ &= \sum_{k=0}^{\infty} \frac{p(p+1)\cdots(p+k-1)}{k!} \nabla^k f_n \,. \end{aligned}$$

Za $f(x_n + ph)$ često se koristi oznaka f_{n+p}.

Koristeći se razlikama zaključno sa redom n, na osnovu poslednje jednakosti dobijamo drugu Newtonovu interpolacionu formulu

$$\pi_n(x) = f_n + p\nabla f_n + \frac{p(p+1)}{2!} \nabla^2 f_n + \ldots + \frac{p(p+1)\cdots(p+n-1)}{n!} \nabla^n f_n,$$

tj.

$$\begin{aligned} \pi_n(x) &= f_n + \frac{\nabla f_n}{h}(x - x_n) + \frac{\nabla^2 f_n}{2!h^2}(x - x_n)(x - x_{n-1}) + \ldots \\ &+ \frac{\nabla^n f_n}{n!h^n}(x - x_n)(x - x_{n-1})\cdots(x - x_1), \end{aligned}$$

pri čemu se ostatak može izraziti u obliku

$$r_n(f;x) = \frac{h^{n+1}}{(n+1)!}\, p(p+1)\cdots(p+n) f^{(n+1)}(\xi) \quad (x_0 < \xi < x_n).$$

3.3.1 Prostiranje greške u tablici konačnih razlika

Proučićemo ovde kako se slučajna greška ε u vrednosti funkcije u nekom od ekvidistantnih interpolacionih čvorova, manifestuje u tablici konačnih razlika. Prikazujemo tablicu konačnih razlika, sa greškom ε u vrednosti f_n.

Na osnovu tablice može se zaključiti sledeće:

1. Ako vrednost f_n sadrži grešku, biće pogrešne sledeće razlike:

$$\begin{array}{llll} \Delta f_{n-1}, & \Delta f_n; & & \\ \Delta^2 f_{n-2}, & \Delta^2 f_{n-1}, & \Delta^2 f_n; & \\ \Delta^3 f_{n-3}, & \Delta^3 f_{n-2}, & \Delta^3 f_{n-1}, & \Delta^3 f_n; \text{ itd.} \end{array}$$

2. Kod k–te konačne razlike, greška učestvuje po zakonu binomnih koeficijenata uz alternativnu promenu znaka, tj.

$$\binom{k}{0}\varepsilon\,, \quad -\binom{k}{1}\varepsilon\,, \quad \binom{k}{2}\varepsilon\,, \ldots,\ (-1)^k\binom{k}{k}\varepsilon\,.$$

Takođe, apsolutna vrednost maksimalne greške u k–toj konačnoj razlici $\binom{k}{[k/2]}|\varepsilon|$ vrlo brzo raste sa redom razlike.

3. Za svaku konačnu razliku Δ^k važe jednakosti:

$$\binom{k}{0}\varepsilon - \binom{k}{1}\varepsilon + \binom{k}{2}\varepsilon - \ldots + (-1)^k\binom{k}{k}\varepsilon = (1-1)^k\,\varepsilon = 0$$

i

$$\binom{k}{0}|\varepsilon| + \binom{k}{1}|\varepsilon| + \binom{k}{2}|\varepsilon| + \ldots + \binom{k}{k}|\varepsilon| = (1+1)^k\,|\varepsilon| = 2^k\,|\varepsilon|\,.$$

x	f	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
x_{n-4}	f_{n-4}				
		Δf_{n-4}			
x_{n-3}	f_{n-3}		$\Delta^2 f_{n-4}$		
		Δf_{n-3}		$\Delta^3 f_{n-4}$	
x_{n-2}	f_{n-2}		$\Delta^2 f_{n-3}$		$\Delta^4 f_{n-4}+\varepsilon$
		Δf_{n-2}		$\Delta^3 f_{n-3}+\varepsilon$	
x_{n-1}	f_{n-1}		$\Delta^2 f_{n-2}+\varepsilon$		$\Delta^4 f_{n-3}-4\varepsilon$
		$\Delta f_{n-1}+\varepsilon$		$\Delta^3 f_{n-2}-3\varepsilon$	
x_n	$f_n+\varepsilon$		$\Delta^2 f_{n-1}-2\varepsilon$		$\Delta^4 f_{n-2}+6\varepsilon$
		$\Delta f_n-\varepsilon$		$\Delta^3 f_{n-1}+3\varepsilon$	
x_{n+1}	f_{n+1}		$\Delta^2 f_n+\varepsilon$		$\Delta^4 f_{n-1}-4\varepsilon$
		Δf_{n+1}		$\Delta^3 f_n-\varepsilon$	
x_{n+2}	f_{n+2}		$\Delta^2 f_{n+1}$		$\Delta^4 f_n+\varepsilon$
		Δf_{n+2}		$\Delta^3 f_{n+1}$	
x_{n+3}	f_{n+3}		$\Delta^2 f_{n+2}$		
		Δf_{n+3}			
x_{n+4}	f_{n+4}				

U tablici konačnih razlika figurišu vrednosti funkcije f sa određenim, fiksiranim, brojem decimalnih mesta. Ako se funkcija f nad skupom vrednosti $\left\{(x_k, f_k)_{k=\overline{0,m}}\right\}$ iz tablice ponaša kao polinom stepena $r\,(<m)$, tada će konačne razlike reda r biti konstantne, a konačne razlike reda $r+1,\ r+2,$ $\ldots, m$ će biti jednake nuli (ili će biti približno jednake nuli s obzirom da su vrednosti funkcije koje su ušle u tablicu eventualno zaokružene). Primetimo da funkcija f ne mora biti polinom a da iskaže opisano ponašanje. Na primer, ako za funkciju f postoji Taylorov polinom pri čemu je odgovarajući ostatak za svako x_k iz tablice toliko mali da ne utiče na decimale koje figurišu u tablici, tada je funkcija f praktično tabelirana vrednostima iz Taylorovog polinoma.

Svakako, ako postoji greška u vrednosti funkcije u nekom od interpolacionih čvorova, prethodni princip će biti narušen u polju prostiranja greške, kako smo prethodno videli, što nam predstavlja indikaciju o postojanju greške.

Zakon prostiranja greške u tablici konačnih razlika, koji je razmatran, daje mogućnost da se u nekim slučajevima pronađe izvor greške i otkloni.

Primer 4.4. Ispravimo grešku u vrednosti funkcije u jednom od interpolacionih čvorova, ako je dato

1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4
−1.020	−0.692	−0.076	0.872	2.212	3.980	6.228	9.004

2.6	2.8
12.356	16.332

Formirajmo tablicu konačnih razlika, na osnovu datih podataka.

k	x_k	f_k	Δf_k	$\Delta^2 f_k$	$\Delta^3 f_k$	$\Delta^4 f_k$	$\Delta^5 f_k$
0	1.0	−1.020					
			0.328				
1	1.2	−0.692		0.288			
			0.616		0.044		
2	1.4	−0.076		0.332		0.16	
			0.948		0.060		−0.040
3	1.6	0.872		0.392		−0.024	
			1.340		0.036		0.040
4	1.8	2.212		0.428		0.016	
			1.768		0.052		−0.020
5	2.0	3.980		0.480		−0.004	
			2.248		0.048		0.004
6	2.2	6.228		0.528		0.	
			2.776		0.048		0.
7	2.4	9.004		0.576		0.	
			3.352		0.048		
8	2.6	12.356		0.624			
			3.976				
9	2.8	16.332					

Iz tablice uočavamo sledeće: Razlike $\Delta^4 f_4$, $\Delta^4 f_5$ i $\Delta^5 f_4$, su jednake nuli, dok su preostale razlike četvrtog i petog reda različite od nule, s tim što su još i razlike petog reda, po modulu, uvećane u odnosu na odgovarajuće razlike četvrtog reda. Ovo nesumnjivo govori o postojanju greške u nekoj vrednosti funkcije f_k.

Dakle, možemo zaključiti da sve razlike četvrtog i petog reda koje su različite od nule, pripadaju polju prostiranja greške ε u vrednosti funkcije f_k. Na osnovu prethodne analize, u razlikama četvrtog reda postoji pet pogrešnih razlika te s obzirom na njihov raspored zaključujemo da je pogrešna vrednost funkcije za $x = 1.6$ $(k = 3)$.

Odredimo grešku ε.

S obzirom da bi konačne razlike četvrtog reda trebalo da budu jednake nuli, to je, na osnovu tablice, $\Delta^4 f_3 + \varepsilon = \varepsilon = -4 \cdot 10^{-3}$.

Ili, na osnovu trećih razlika, koje bi trebalo da budu konstantne (s obzirom da bi četvrte razlike trebalo da budu jednake nuli), nalazimo

$$\begin{aligned} \Delta^3 f_3 &= \frac{1}{4}((\Delta^3 f_3 - \varepsilon) + (\Delta^3 f_2 + 3\varepsilon) + (\Delta^3 f_1 - 3\varepsilon) + (\Delta^3 f_0 + \varepsilon)) = \\ &= \frac{1}{4}(52 + 36 + 60 + 44) \cdot 10^{-3} = 48 \cdot 10^{-3}\,, \end{aligned}$$

ili direktno, na osnovu polja prostiranja greške ε, očitavamo na osnovu "neporemećenih" trećih razlika $\Delta^3 f_4 = \Delta^3 f_3 = 48 \cdot 10^{-3}$, a dalje, s obzirom na $\Delta^3 f_3 - \varepsilon = 52 \cdot 10^{-3}$, nalazimo $\varepsilon = -4 \cdot 10^{-3}$.

Grešku ε možemo naći u ovom slučaju i na osnovu drugih razlika koje bi u tačnoj tablici morale obrazovati aritmetičku progresiju (s obzirom da bi treće razlike trebalo da budu konstantne). Dakle, tačna vrednost $\Delta^2 f_2$ je

$$\begin{aligned} \Delta^2 f_2 &= \frac{1}{3}((\Delta^2 f_1 + \varepsilon) + (\Delta^2 f_2 - 2\varepsilon) + (\Delta^2 f_3 + \varepsilon)) \\ &= \frac{1}{3}(332 + 392 + 428) \cdot 10^{-3} = 384 \cdot 10^{-3}\,, \end{aligned}$$

pa ε nalazimo na osnovu

$$\varepsilon = \frac{1}{2}(\Delta^2 f_2 - (\Delta^2 f_2 - 2\varepsilon)) = \frac{1}{2}(384 - 392) \cdot 10^{-3} = -4 \cdot 10^{-3}\,.$$

Najzad, ispravljena vrednost f za $x = 1.6$, biće

$$f_3 = (f_3 + \varepsilon) - \varepsilon = 0.872 - (-0.004) = 0.876\,.$$

3.4 Hermiteova interpolacija

Ovde obrađujemo jedan opštiji interpolacioni problem. Neka su za funkciju $f : [a, b] \mapsto \mathbb{R}$ u interpolacionim čvorovima x_i $(i = 0, 1, \ldots, m)$ poznate sledeće vrednosti:

$$f(x_i), f'(x_i), \ldots, f^{(k_i-1)}(x_i) \quad (i = 0, 1, \ldots, m).$$

Neka je broj podataka o funkciji f jednak $n+1$, tj. $n+1 = k_0+k_1+\ldots+k_m$. Broj k_i se naziva višestrukost interpolacionog čvora x_i.

Posmatraćemo sada samo interpolaciju algebarskim polinomima, koja je poznata kao Hermiteova interpolacija. Na osnovu datih podataka o funkciji f, Hermiteov interpolacioni polinom će u opštem slučaju biti

$$H_n(x) = a_0 + a_1x + \ldots + a_nx^n.$$

Koeficijente a_i $(i = 0, 1, \ldots, n)$ određujemo rešavanjem sistema linearnih jednačina

$$H_n^{(j)}(x_i) = f^{(j)}(x_i) \quad (i = 0, 1, \ldots, m;\ j = 0, 1, \ldots, k_i - 1). \tag{3.21}$$

Teorema 3.4.1 *Sistem jednačina* (3.21) *ima jedinstveno rešenje.*

Dokaz. Za dokaz ovog tvrđenja dovoljno je pokazati da homogeni sistem jednačina

$$H_n^{(j)}(x_i) = 0 \quad (i = 0, 1, \ldots, m;\ j = 0, 1, \ldots, k_i - 1)$$

ima samo trivijalno rešenje $a_i = 0$ $(i = 0, 1, \ldots, n)$, tj. $H_n(x) \equiv 0$.

U posmatranom slučaju sledi da je x_i nula polinoma H_n višestrukosti najmanje k_i, što znači da zbir višestrukosti svih nula polinoma H_n nije manji od $n + 1$ $(= k_0 + k_1 + \ldots + k_m)$. No, H_n je polinom stepena n, zaključujemo da on mora biti identički jednak nuli. Dokaz je završen. □

Kako rešavanje sistema linearnih jednačina (3.21) može biti veoma komplikovano to se konstrukcija Hermiteovog interpolacionog polinoma najčešće izvodi korišćenjem Lagrangeove interpolacije, uključivanjem uslova (3.21). Naime, polinom H_n tražimo u obliku

$$H_n(x) = \pi_m(x) + (x - x_0)(x - x_1)\cdots(x - x_m)H_k(x), \tag{3.22}$$

gde je π_m Lagrangeov interpolacioni polinom dobijen samo na osnovu datih podataka o vrednnosti funkcije f u interpolacionim čvorovima, tj. na osnovu parova $(x_i, f(x_i))$ $(i = 0, 1, \ldots, m)$, a H_k za sada nepoznati polinom čiji je stepen takav da je $k + m + 1 = n$. Primetimo da ovakva reprezentacija polinoma H_n zahteva određivanje polinoma H_k. Kako smo iskoristili informacije o vrednosti funkcije (za određivanje π_m) ostaju nam na raspolaganju vrednosti izvoda funkcije. Imajući u vidu interpolacioni zahtev (3.21), to diferenciranjem (3.22) dobijamo potrebne uslove za određivanje polinoma H_k koji su opet tipa (3.21), ali sada za polinom H_k. To znači da se metoda konstrukcije Hermiteovog polinoma može da zasnuje rekurzivno.

Na osnovu prethodne teoreme polinom koji se na ovaj način dobija predstavlja Hermiteov interpolacioni polinom. Evo primera kojim ilustrujemo navedeno.

Primer 4.5. Konstruišimo Hermiteov interpolacioni polinom na osnovu sledećih podataka

x	-1	0	2
$f(x)$	0	-7	3
$f'(x)$	-8	-5	55
$f''(x)$		10	

Kako je dato sedam podataka, interpolacioni polinom će biti stepena ne

više od šestog. Potražimo ga u obliku

$$H_6(x) = \pi_2(x) + (x+1)(x-2)x\, H_3(x)\,, \tag{3.23}$$

gde je $\pi_2(x)$ Lagrangeov interpolacioni polinom formiran na osnovu vrednosti funkcije f u tačkama $x=-1,\ x=0,\ x=2$, tj.

$$\pi_2(x) = -7\,\frac{(x+1)(x-2)}{(0+1)(0-2)} + 3\,\frac{(x+1)(x-0)}{(2+1)(2-0)} = 4x^2 - 3x - 7\,,$$

a $H_3(x)$ za sada nepoznat polinom stepena ne višeg od tri.

Diferenciranjem (3.23) dobijamo

$$H_6'(x) = 8x - 3 + (3x^2 - 2x - 2)H_3(x) + (x+1)(x-2)x\, H_3'(x)\,,$$

odakle, s obzirom na interpolacioni zahtev

$$H_6'(-1) = f'(-1) = -8,\ H_6'(0) = f'(0) = -5 \text{ i } H_6'(2) = f'(2) = 55,$$

sledi

$$H_3(-1) = 1\,, \quad H_3(0) = 1\,, \quad H_3(2) = 7\,. \tag{3.24}$$

Kako je dalje

$$H_6''(x) = 8 + (6x-2)\,H_3(x) + (6x - 4x - 4)\,H_3'(x) + (x+1)(x-2)x\, H_3''(x)$$

i $H_6''(0) = f''(0) = 10$, dobijamo

$$H_3'(0) = -1\,. \tag{3.25}$$

Primenimo sada isti postupak na određivanje polinoma H_3, na osnovu podataka (3.24) i (3.25). Dakle, imamo

$$H_3(x) = \pi_2^*(x) + (x+1)(x-2)x\, a \qquad (a = H_0(x))\,,$$

gde je

$$\pi_2^*(x) = 1\,\frac{(x-0)(x-2)}{(-1-0)(-1-2)} + 1\,\frac{(x+1)(x-2)}{(0+1)(0-2)} + 7\,\frac{(x+1)(x-0)}{(2+1)(2-0)} = x^2 + x + 1\,.$$

Dalje, kako je

$$H_3'(x) = 2x + 1 + (3x^2 - 2x - 2)a$$

i $H_3'(0) = -1$, dobijamo $a = 1$, pa je

$$H_3(x) = x^3 - x + 1\,.$$

Najzad, na osnovu (3.23), dobijamo

$$H_6(x) = x^6 - x^5 - 3x^4 + 2x^3 + 5x^2 - 5x - 7\,.$$

Analogno kao za ostatak Lagrangeove interpolacije, za ostatak Hermiteove interpolacije važi sledeća teorema.

Teorema 3.4.2 *Neka je $f \in C^{n+1}[a,b]$ i $x_i \in [a,b]$ $(i = 0, 1, \ldots, m)$. Tada postoji $\xi \in (a,b)$ takvo da je greška Hermiteovog interpolacionog polinoma*

$$r_n(f;x) = f(x) - H_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\Omega_n(x), \tag{3.26}$$

gde je

$$\Omega_n(x) = (x - x_0)^{k_0}(x - x_1)^{k_1} \cdots (x - x_m)^{k_m}.$$

Dokaz. Uvedimo pomoćnu funkciju g, definisanu sa

$$g(x) = f(x) - H_n(x) - \kappa_n\,\Omega_n(x).$$

Neka je $\bar{x}$ proizvoljna tačka iz $[a,b]$ takva da je $\bar{x} \neq x_i$ $(i = 0, 1, \ldots, m)$.

Pomoćna funkcija g ima nule u tačkama $x_0, x_1, \ldots, x_m$, čiji je red višestrukosti najmanje $k_0, k_1, \ldots, k_m$, respektivno. Odredimo konstantu κ_n tako da bude $g(\bar{x}) = 0$. Kako je $\bar{x} \neq x_i$ $(i = 0, 1, \ldots, m)$, ovakva vrednost za κ_n postoji. Naime,

$$\kappa_n = \frac{f(\bar{x}) - H_n(\bar{x})}{\Omega_n(\bar{x})}. \tag{3.27}$$

Na osnovu Rolleove teoreme, izvod g' ima bar $m+1$ nula koje se nalaze između nula funkcije g. S druge strane, funkcija g' ima nule i u tačkama $x_0, x_1, \ldots, x_m$, čiji je red višestrukosti najmanje $k_0 - 1, k_1 - 1, \ldots, k_m - 1$, respektivno. Dakle, broj nula funkcije g' na $[a, b]$ je najmanje $n+1$ $(= (k_0 - 1) + (k_1 - 1) + \ldots + (k_m - 1) + m + 1)$, pri čemu se višestrukost nula uzima onoliko puta koliki je njen red višestrukosti.

Kako $H_n^{(n+1)}(x) = 0$ i $\Omega_n^{(n+1)}(x) = (n+1)!$, iz $g^{(n+1)}(\xi) = 0$ sledi

$$\kappa_n = \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

što zajedno sa (3.27) daje

$$f(\bar{x}) = H_n(\bar{x}) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\Omega_n(\bar{x}).$$

Kako je $\bar{x}$ proizvoljna tačaka iz $[a, b]$, sledi (3.26), tj. dokaz je završen. □

Ovu sekciju završavamo tako što ćemo odrediti opšti oblik Hermite–ovog interpolacionog polinoma za realnu funkciju $y = f(x)$ na intervalu $[a, b],\ a, b \in \mathbb{R}$.

Neka je zadat bazni sistem interpolacionih funkcija

$$\varphi_0(x), \varphi_1(x),\ \ldots,\ \varphi_n(x),\ \ldots$$

na $[a, b]$. Odredimo takvu linearnu kombinaciju ovih funkcija

$$\varphi(x) = \sum_{i=0}^{m} c_i \varphi_i(x) \tag{3.28}$$

koja zadovoljava uslove

$$\begin{aligned}
\varphi(x_0) &= y_0, \quad \varphi'(x_0) = y_0',\ \ldots\ , \varphi^{(\alpha_0-1)}(x_0) = y_0^{(\alpha_0-1)},\\
\varphi(x_1) &= y_1, \quad \varphi'(x_1) = y_1',\ \ldots\ , \varphi^{(\alpha_1-1)}(x_1) = y_1^{(\alpha_1-1)},\\
&\vdots\\
\varphi(x_n) &= y_n, \quad \varphi'(x_n) = y_n',\ \ldots\ , \varphi^{(\alpha_n-1)}(x_n) = y_n^{(\alpha_n-1)},
\end{aligned}$$

gde su $y_i^{(j)}$ poznate vrednosti, a $x_i \in [a,b]$ $(i = 0,1,2,\ \dots\ ,n;\ x_i \neq x_j$ pri $i \neq j)$. Pošto je broj uslova koje namećemo funkciji $\varphi(x)$ jednak

$$\alpha_0 + \alpha_1 + \ \dots\ + \alpha_n$$

da bi naš zadatak imao jedinstveno rešenje potrebno je da bude

$$m = \alpha_0 + \alpha_1 + \ \dots\ + \alpha_n - 1$$

i

$$\begin{vmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \dots & \varphi_m(x_0) \\ \varphi_0'(x_0) & \varphi_1'(x_0) & \dots & \varphi_m'(x_0) \\ \vdots & \vdots & & \vdots \\ \varphi_0^{(\alpha_0-1)}(x_0) & \varphi_1^{(\alpha_0-1)}(x_0) & \dots & \varphi_m^{(\alpha_0-1)}(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \dots & \varphi_m(x_1) \\ \vdots & \vdots & & \vdots \\ \varphi_0^{(\alpha_n-1)}(x_n) & \varphi_1^{(\alpha_n-1)}(x_n) & \dots & \varphi_m^{(\alpha_n-1)}(x_n) \end{vmatrix} \neq 0.$$

Ako se ograničimo na slučaj kada je $\varphi_i(x) = x^i$ i $y_i^{(j)} = f^{(j)}(x_i)$ $(i = 0,1,\dots,n; j = 0,1,\dots,\alpha_i - 1)$ onda polinom (3.28) predstavlja algebarski interpolacioni Hermiteov polinom za funkciju $y = f(x)$ na intervalu $[a,b]$.

Odredimo sada opšti oblik Hermiteovog interpolacionog polinoma. U tu svrhu uvedimo polinome $H_{ij}(x)$ stepena ne višeg od m, koji zadovoljavaju sledeće uslove:

$$\begin{aligned} H_{ij}(x_k) &= H_{ij}'(x_k) = \ \dots\ = H_{ij}^{(\alpha_k-1)}(x_k) = 0, \quad i \neq k, \\ H_{ij}(x_i) &= H_{ij}'(x_i) = \ \dots\ = H_{ij}^{(j-1)}(x_i) = H_{ij}^{(j+1)}(x_i) = H_{ij}^{(\alpha_i-1)}(x_i) = 0, \\ H_{ij}^{(j)}(x_i) &= 1 \quad (i = 0,1,2,\ \dots\ ,n;\ j = 0,1,2,\ \dots\ ,\alpha_i - 1). \end{aligned}$$

Kako H_{ij} ima nule

$$x_0, x_1,\ \dots\ , x_{i-1}, x_{i+1},\ \dots\ , x_n,$$

višestrukosti

$$\alpha_0, \alpha_1, \ \dots \ , \alpha_{i-1}, \alpha_{i+1}, \ \dots \ , \alpha_n,$$

respektivno, a u tački x_i nulu viěstrukosti j, to je

$$H_{ij}(x) = (x - x_0)^{\alpha_0}(x - x_1)^{\alpha_1} \ \cdots \ (x - x_{i-1})^{\alpha_{i-1}}(x - x_i)^j \times$$

$$\times (x - x_{i+1})^{\alpha_{i+1}} \ \cdots \ (x - x_n)^{\alpha_n} \overline{H}_{ij}(x),$$

gde je $\overline{H}_{ij}(x)$ polinom stepena $\alpha_i - j - 1$, različit od nule za $x = x_i$. Predstavimo ga, zato, u obliku

$$\overline{H}_{ij}(x) = A_{ij}^{(0)} + A_{ij}^{(1)}(x - x_i) + \ \dots \ + A_{ij}^{(\alpha_i - j - 1)}(x - x_i)^{\alpha_i - j - 1}.$$

Neka je

$$\Omega(x) = (x - x_0)^{\alpha_0}(x - x_1)^{\alpha_1} \ \cdots \ (x - x_n)^{\alpha_n},$$

tada

$$A_{ij}^{(0)} + A_{ij}^{(1)}(x - x_i) + \ \dots \ + A_{ij}^{(\alpha_i - j - 1)}(x - x_i)^{\alpha_i - j - 1} = \frac{(x - x_i)^{\alpha_i - j}}{\Omega(x)} H_{ij}(x).$$

Ako pustimo da $x \to x_i$, dobijamo:

$$A_{ij}^{(0)} = \lim_{x \to x_i} \left[\frac{(x - x_i)^{\alpha_i - j}}{\Omega(x)} \frac{H_{ij}(x)}{(x - x_i)^j} \right].$$

Prvi činilac na desnoj strani u prethodnoj jednakosti, pod limesom, je neprekidan za $x = x_i$. Dakle,

$$\lim_{x \to x_i} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right] = \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]_{x = x_i}.$$

Limes drugog člana nalazimo po L'Hôspitalovom pravilu:

$$\lim_{x \to x_i} \left[\frac{H_{ij}(x)}{(x - x_i)^j} \right] = \lim_{x \to x_i} \frac{H_{ij}^{(j)}(x)}{j!} = \frac{1}{j!}.$$

Dakle,

$$A_{ij}^{(0)} = \frac{1}{j!} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]_{x = x_i}.$$

Analogno nalazimo koeficijente $A_{ij}^{(k)}$:

$$A_{ij}^{(k)} = \frac{1}{k!} \lim_{x \to x_i} \frac{d^k}{d\,x^k} \left[\frac{(x-x_i)^{\alpha_i}}{\Omega(x)} \frac{H_{ij}(x)}{(x-x_i)^j} \right] .$$

Primenom Leibnitzovog pravila za diferenciranje proizvoda imamo

$$\frac{d^k}{d\,x^k} \left[\frac{(x-x_i)^{\alpha_i}}{\Omega(x)} \frac{H_{ij}(x)}{(x-x_i)^j} \right] = \sum_{p=0}^{k} \binom{k}{p} \left[\frac{(x-x_i)^{\alpha_i}}{\Omega(x)} \right]^{(p)} \left[\frac{H_{ij}(x)}{(x-x_i)^j} \right]^{(k-p)} .$$

Izvod

$$\left[\frac{(x-x_i)^{\alpha_i}}{\Omega(x)} \right]^{(p)}$$

je neprekidan u tački $x = x_i$. Dakle,

$$\lim_{x \to x_i} \left[\frac{(x-x_i)^{\alpha_i}}{\Omega(x)} \right]^{(p)} = \left[\frac{(x-x_i)^{\alpha_i}}{\Omega(x)} \right]^{(p)}_{x=x_i} .$$

Za nalaženje limesa

$$\lim_{x \to x_i} \left[\frac{H_{ij}(x)}{(x-x_i)^j} \right]^{(k-p)}$$

postupamo na sledeći način.

Polinom $H_{ij}(x)$ je stepena ne višeg od m. On je deljiv sa $(x-x_i)^j$, stoga ga možemo zapisati u obliku

$$H_{ij}(x) = B_{ij}^{(0)}(x-x_i)^j + B_{ij}^{(1)}(x-x_i)^{j+1} + \ \ldots \ + B_{ij}^{(m-j)}(x-x_i)^m,$$

ili

$$\frac{H_{ij}(x)}{(x-x_i)^j} = B_{ij}^{(0)} + B_{ij}^{(1)}(x-x_i) + \ \ldots \ + B_{ij}^{(m-j)}(x-x_i)^{m-j}.$$

Dakle,

$$\lim_{x \to x_i} \left[\frac{H_{ij}(x)}{(x-x_i)^j} \right]^{(k-p)} = (k-p)!\, B_{ij}^{(k-p)}.$$

No, $B_{ij}^{(k-p)}$ kao koeficijente razlaganja $H_{ij}(x)$ po stepenima $(x-x_i)$ možemo zapisati u obliku

$$B_{ij}^{(k-p)} = \frac{H_{ij}^{(j+k-p)}(x_i)}{(j+k-p)\,!} .$$

U našem slučaju je

$$j + k - p \leq j + k \leq j + \alpha_i - j - 1 = \alpha_i - 1.$$

Uočimo da je $B_{ij}^{(k-p)}$ $(k - p = 0, 1, \ldots, \alpha_i - j - 1)$ različito od nule samo za $p = k$, i u tom slučaju je

$$B_{ij}^{(0)} = \frac{1}{j\,!}\,.$$

Dakle,

$$\begin{aligned} A_{ij}^{(k)} &= \frac{1}{k!} \lim_{x \to x_i} \frac{d^k}{dx^k} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \frac{H_{ij}(x)}{(x - x_i)^j} \right] \\ &= \frac{1}{k\,!j\,!} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(k)}_{x=x_i} \end{aligned}$$

i

$$H_{ij}(x) = \frac{1}{j\,!} \frac{\Omega(x)}{(x - x_i)^{\alpha_i - j}} \sum_{k=0}^{\alpha_i - j - 1} \frac{1}{k!} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(k)}_{x=x_i} (x - x_i)^k\,.$$

Na osnovu svojstava funkcija $H_{ij}(x)$ nije teško uočiti da je

$$\varphi(x) \equiv H_m(x) = \sum_{i=0}^{n} \sum_{j=0}^{\alpha_i - 1} y_i^{(j)} H_{ij}(x)\,,$$

odnosno

$$H_m(x) = \sum_{i=0}^{n} \sum_{j=0}^{\alpha_i - 1} \sum_{k=0}^{\alpha_i - j - 1} y_i^{(j)} \frac{1}{k\,!} \frac{1}{j\,!} \left[\frac{(x - x_i)^{\alpha_i}}{\Omega(x)} \right]^{(k)}_{x=x_i} \frac{\Omega(x)}{(x - x_i)^{\alpha_i - j - k}}\,.$$

3.5 Numeričko diferenciranje

Posmatramo realnu funkciju f definisanu na intervalu $[a, b]$. Često imamo potrebu za numeričkim diferenciranjem funkcije f, npr. kada je funkcija f data tabelarno, tj. kada znamo samo vrednosti funkcije na nekom diskretnom podskupu intervala $[a, b]$, kada je analitički izraz za f komlikovan, itd.

Numeričko diferenciranje se uglavnom izvodi tako što aproksimiramo funkciju f nekom funkcijom φ na $[a,b]$ ($f(x) \sim \varphi(x),\ x \in [a,b]$) koju onda diferenciramo određeni broj puta. Dakle,

$$f^{(k)}(x) \sim \varphi^{(k)}(x), \quad (a \leq x \leq b;\ k = 1, 2, \ldots).$$

Naravno, od funkcije φ se zahteva da može jednostavno da se diferencira, zbog čega se najčešće za φ uzimaju algebarski interpolacioni polinomi.

Neka je φ interpolacioni polinom stepena n, tj.

$$\varphi(x) = \pi_n(x).$$

Iz formule za grešku interpolacionim polinomom

$$r_n(f;x) = f(x) - \pi_n(x) \quad (x \in [a,b])$$

dobijamo formulu za grešku diferenciranja, naime

$$r_n^{(k)}(f;x) = f^{(k)}(x) - \pi_n^{(k)}(x) \quad (x \in [a,b]).$$

Za red izvoda ima smisla uzeti samo $k < n$. Primećujemo da je kod interpolacije greška u čvorovima jednaka nuli, dok kod numeričkog diferenciranja to ne mora biti. Dakle u opštem slučaju numeričko diferenciranje ima manju tačnost od interpolacije.

Neka su poznate vrednosti funkcije na skupu ekvidistantnih tačaka

$$\{x_0, x_1, \ldots, x_m\} \subset [a,b]\,,$$

sa korakom h. Dakle, neka je

$$f_k = f(x_k) = f(x_0 + kh) \quad (k = 0, 1, \ldots, m).$$

Na skupu

$$\{x_i, x_{i+1}, \ldots, x_{i+n}\} \quad (0 \leq i \leq m - n)$$

konstruišimo prvi Newtonov interpolacioni polinom

$$\begin{aligned}\pi_n(x) &= f_i + p\Delta f_i + \frac{p(p-1)}{2!}\Delta^2 f_i + \frac{p(p-1)(p-2)}{3!}\Delta^3 f_i \\ &\quad + \ldots + \frac{p(p-1)\cdots(p-n+1)}{n!}\Delta^n f_i,\end{aligned}$$

tj.

$$\begin{aligned}\pi_n(x) = \; & f_i + p\Delta f_i + \frac{p^2-p}{2}\Delta^2 f_i + \frac{p^3-3p^2+2p}{6}\Delta^3 f_i \\ & + \ldots + \frac{p^n - \frac{1}{2}n(n-1)p^{n-1} + \ldots}{n!}\Delta^n f_i,\end{aligned} \tag{3.29}$$

gde je $p = (x - x_i)/h$.

Pošto je

$$\pi_n'(x) = \frac{1}{h}\cdot\frac{d\,\pi_n(x)}{dp},$$

diferenciranjem (3.29) dobijamo

$$\pi_n'(x) = \frac{1}{h}\left(\Delta f_i + \frac{2p-1}{2}\Delta^2 f_i + \frac{3p^2-6p+2}{6}\Delta^3 f_i + \ldots\right). \tag{3.30}$$

Daljim diferenciranjem (3.30), dobijamo π_n'', π_n''', itd.

Na primer

$$\pi_n''(x) = \frac{1}{h^2}\left(\Delta^2 f_i + (p-1)\Delta^3 f_i + \ldots\right). \tag{3.31}$$

Za $x = x_i$, tj. $p = 0$, formule (3.30) i (3.31) se svode na

$$\pi_n'(x_i) = \frac{1}{h}\left(\Delta f_i - \frac{1}{2}\Delta^2 f_i + \frac{1}{3}\Delta^3 f_i - \ldots + \frac{(-1)^{n-1}}{n}\Delta^n f_i\right)$$

i

$$\pi_n''(x_i) = \frac{1}{h^2}\left(\Delta^2 f_i - \Delta^3 f_i + \ldots\right).$$

Dakle, imamo

$$f'(x_i) = \frac{1}{h}\left(\Delta f_i - \frac{1}{2}\Delta^2 f_i + \frac{1}{3}\Delta^3 f_i - \ldots + \frac{(-1)^{n-1}}{n}\Delta^n f_i\right) + r_n'(f; x_i). \tag{3.32}$$

Ako je $f \in C^{n+1}[a,b]$, greška u poslednjoj formuli može se predstaviti u obliku

$$r'_n(f;x_i) = \frac{(-1)^n}{n+1}\, h^n f^{(n+1)}(\xi_n) \quad (\xi_n \in (x_i, x_{n+i})). \tag{3.33}$$

Poslednja formula pokazuje da greška u interpolacionom čvoru ne mora biti jednaka nuli, iako je $r_n(f;x_i) = 0$.

Na osnovu (3.32) i (3.33), za $n = 1,2,3$ dobijamo respektivno

$$\begin{aligned} f'(x_i) &= \frac{1}{h}\,(f_{i+1} - f_i) - \frac{h}{2} f''(\xi_1), \\ f'(x_i) &= \frac{1}{2h}\,(-3f_i + 4f_{i+1} - f_{i+2}) + \frac{h^2}{3} f'''(\xi_2), \\ f'(x_i) &= \frac{1}{6h}\,(-11f_i + 18f_{i+1} - 9f_{i+2} + 2f_{i+3}) - \frac{h^3}{4} f^{(4)}(\xi_3), \end{aligned}$$

gde je ξ_n, u izrazima za ostatak u prethodnim formulama, takvo da $\xi_n \in (x_i, x_{n+i})$, za $n = 1,2,3$.

Korišćenjem operatora zadnje razlike ∇, slično formuli (3.32), dobijamo

$$f'(x_i) = \frac{1}{h}\left(\nabla f_i + \frac{1}{2}\nabla^2 f_i + \frac{1}{3}\nabla^3 f_i + \ldots + \frac{1}{n}\nabla^n f_i\right) + r'_n(f;x_i).$$

Ako je $f \in C^{n+1}[a,b]$, greška u poslednjoj formuli može se predstaviti u obliku

$$r'_n(f;x_i) = \frac{1}{n+1}\, h^n f^{(n+1)}(\eta_n) \quad (\eta_n \in (x_{i-n}, x_i)).$$

Za $n = 1,2,3$ dobijamo respektivno

$$\begin{aligned} f'(x_i) &= \frac{1}{h}\,(f_i - f_{i-1}) + \frac{h}{2} f''(\eta_1), \\ f'(x_i) &= \frac{1}{2h}\,(3f_i - 4f_{i-1} + f_{i-2}) + \frac{h^2}{3} f'''(\eta_2), \\ f'(x_i) &= \frac{1}{6h}\,(11f_i - 18f_{i-1} + 9f_{i-2} - 2f_{i-3}) + \frac{h^3}{4} f^{(4)}(\eta_3), \end{aligned}$$

gde je η_n, u izrazima za ostatak u prethodnim formulama, takvo da $\eta_n \in (x_{i-n}, x_i)$, za $n = 1,2,3$.

Prethodne formule za prvi izvod u čvoru x_i su nesimetrične. One se uspešno primenjuju kada se određuje izvod na krajevima intervala $[a, b]$. Tipična primena ovih formula je kod aproksimacije diferencijalnih konturnih uslova u konturnim problemima kod diferencijalnih jednačina.

Za čvorove unutar intervala $[a, b]$ bolje je koristiti simetrične formule za diferenciranje, tj. one koje se dobijaju primenom operatora centralne razlike, koji nisu predviđeni za obrađivanje ovom knjigom. Kao korisne, ovde navodimo formule za određivanje prvog i drugog izvoda:

$$f'(x_i) = \frac{1}{2h}(f_{i+1} - f_{i-1}) - \frac{1}{6}\,h^2 f'''(\zeta_1) \quad (\zeta_1 \in (x_{i-1}, x_{i+1})), \tag{3.34}$$

$$f''(x_i) = \frac{1}{h^2}(f_{i+1} - 2f_i + f_{i-1}) - \frac{1}{12}\,h^2 f^{(4)}(\zeta_2) \quad (\zeta_2 \in (x_{i-1}, x_{i+1})). \tag{3.35}$$

Primer 4.6. Data je funkcija

$$y(x) = \varepsilon \sin \frac{x}{\varepsilon^2}.$$

Aproksimiraćemo datu funkciju polinomom

$$p(x) \equiv 0$$

i analizirati grešku kod numeričkog diferenciranja.

Ako je ε dovoljno malo, važi:

$$|y(x) - p(x)| < \varepsilon_1,$$

gde je ε_1 pozitivna veličina bliska nuli.

S druge strane,

$$y(x) = 0, \quad \text{za} \quad x_k = k\pi\varepsilon^2,$$

pa ako za čvorove interpolacije uzmemo baš tačke x_k onda je interpolacioni polinom $p(x) \equiv 0$.

Dakle, $y(x) \approx 0$, dok

$$\max |y'(x)| = \frac{1}{\varepsilon} \longrightarrow \infty, \quad \varepsilon \longrightarrow 0+ \ .$$

PRIMER 4.7. Pretpostavimo da je $f \in C^3[a,b]$ i neka $x_0, x_1 \in (a,b)$. Dokažimo najpre da postoji tačka $\xi \in (a,b)$ tako da je za $h = x_1 - x_0$

$$f'\left(x_0 + \frac{h}{2}\right) = \frac{\Delta f_0}{h} - \frac{f'''(\xi)}{24}\, h^2. \tag{3.36}$$

Formula (3.36) važi, jer

$$\begin{aligned} & f'\left(x_0 + \frac{h}{2}\right) - \frac{1}{h}(f(x_1) - f(x_0)) = f'\left(x_0 + \frac{h}{2}\right) \\ - \ & \frac{1}{h}\left[f\left(x_0 + \frac{h}{2}\right) + \frac{h}{2} f'\left(x_0 + \frac{h}{2}\right) + \frac{h^2}{4} \cdot \frac{1}{2!} f''\left(x_0 + \frac{h}{2}\right) + \frac{h^3}{8}\frac{1}{3!} f'''(\eta_1)\right. \\ - \ & \left. f\left(x_0 + \frac{h}{2}\right) + \frac{h}{2} f'\left(x_0 + \frac{h}{2}\right) - \frac{h^2}{4} \cdot \frac{1}{2!} f''\left(x_0 + \frac{h}{2}\right) + \frac{h^3}{8}\frac{1}{3!} f'''(\eta_2)\right] \\ & = -\frac{h^2}{24} f'''(\xi), \quad \xi \in (a,b). \end{aligned}$$

Dokažimo sada

$$\lim_{\varepsilon \to 0} [x_0, x_1, x, x+\varepsilon; f] = [x_0, x_1, x, x; f] = \frac{f'''(\xi)}{3!}\,,$$

gde je $x = x_0 + h/2, \quad \xi \in (x_0, x_1)$.

Označimo interpolacioni polinom prvog stepena, dobijen na osnovu vrednosti funkcije f u čvorovima x_0, x_1, sa π_1. Imamo

$$\pi_1(x) = f(x_0) + [x_0, x_1; f](x - x_0).$$

Kako je na osnovu na osnovu teoreme o ostatku u interpolaciji Newtonovim polinomom sa podeljenim razlikama

$$f(x) - \pi_1(x) = [x_0, x_1, x; f]\,(x - x_0)(x - x_1),$$

imamo

$$f'(x)-\pi_1'(x) = \frac{d}{dx}\{[x_0,x_1,x;f]\}(x-x_0)(x-x_1)+[x_0,x_1,x;f](x-x_1+x-x_0),$$

pa je

$$f'\left(x_0+\frac{h}{2}\right)-\pi_1'\left(x_0+\frac{h}{2}\right) = -\frac{h^2}{4}\frac{d}{dx}\{[x_0,x_1,x;f]\}_{x=x_0+h/2},$$

jer je

$$[x_0,x_1,x;f]\,(x-x_1+x-x_0)_{x=x_0+h/2}=0.$$

S druge strane, na osnovu (3.36), imamo

$$f'\left(x_0+\frac{h}{2}\right)-\pi_1'\left(x_0+\frac{h}{2}\right) = -\frac{h^2}{24}f'''(\xi).$$

Upoređivanjem poslednja dva izraza dobijamo da je

$$-\frac{h^2}{4}\frac{d}{dx}\{[x_0,x_1,x;f]\}_{x=x_0+h/2} = -\frac{h^2}{24}f'''(\xi),$$

tj.

$$\frac{d}{dx}\{[x_0,x_1,x;f]\}_{x=x_0+h/2} = \frac{1}{6}f'''(\xi).$$

Dakle, važi

$$\begin{aligned}\frac{d}{dx}\{[x_0,x_1,x;f]\} &= \lim_{\varepsilon\to 0}\frac{[x_0,x_1,x+\varepsilon;f]-[x_0,x_1,x;f]}{\varepsilon}\\ &= \lim_{\varepsilon\to 0}[x_0,x_1,x,x+\varepsilon;f] = [x_0,x_1,x,x;f]\\ &= \frac{1}{6}f'''(\xi).\end{aligned}$$

PRIMER 4.8. Data je funkcija tablicom

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0.50	0.3521				
		-0.0510			
0.75	0.3011		-0.0081		
		-0.0591		0.0079	
1.00	0.2420		-0.0002		-0.0016
		-0.0593		0.0063	
1.25	0.1827		0.0061		
		-0.0532			
1.50	0.1295				

Ispitajmo da li data funkcija ima tačku prevoja na intervalu interpolacije.

Za izračunavanje drugog izvoda u čvorovima $0.75, 1.00, 1.25$, koristimo formulu (videti (3.35))

$$y'' = \frac{y_{-1} - 2y_0 + y_1}{h^2} + O(h^2).$$

Za izračunavanje izvoda u tački $x = 0.50$ koristimo prvi Newtonov interpolacioni polinom

$$\begin{aligned} y(x) &\approx y_0 + p\Delta y_0 + \frac{p(p-1)}{2!}\Delta^2 y_0 + \frac{p(p-1)(p-2)}{3!}\Delta^3 y_0 \\ &+ \frac{p(p-1)(p-2)(p-3)}{4!}\Delta^4 y_0, \end{aligned}$$

gde je $p = (x - x_0)/h$. Dakle,

$$y''(0.5) \approx \frac{1}{h^2}\left[\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0\right] = -0.2795.$$

Za izračunavanje izvoda u tački $x = 1.50$ koristimo drugi Newtonov interpolacioni polinom

$$\begin{aligned} y(x) &\approx y_4 + p\Delta y_3 + \frac{p(p+1)}{2!}\Delta^2 y_2 + \frac{p(p+1)(p+2)}{3!}\Delta^3 y_1 \\ &+ \frac{p(p+1)(p+2)(p+3)}{4!}\Delta^4 y_0, \end{aligned}$$

gde je $p = (x - x_4)/h$. Dakle,

$$y''(1.5) \approx \frac{1}{h^2}\left[\Delta^2 y_2 - \Delta^3 y_1 + \frac{11}{12}\Delta^4 y_0\right] = 0.1749.$$

Dobili smo tabelu približnih vrednosti drugog izvoda tabelirane funkcije.

k	x_k	$y''(x_k) = y''_k$
0	0.50	-0.2795
1	0.75	-0.1296
2	1.00	-0.0032
3	1.25	0.0976
4	1.50	0.1749

Očigledno je $y'' = 0$ za $x \in (1.00, 1.25)$. Da bismo odredili približno tačku prevoja x^* koristimo podatke iz prethodne tabele i primenjujemo inverznu Lagrangeovu interpolaciju. Dakle,

$$x^* = \pi_4(y'' = 0) = 1.03.$$

Glava 4

Nelinearne jednačine i sistemi

Ova glava je prvenstveno posvećena iterativnim metodama za rešavanje jednačina oblika

$$f(x) = 0\,,$$

gde je $f(x)$ nelinearna funkcija definisana na intervalu $[a, b]$. Pretpostavljaćemo da je f neprekidna na $[a, b]$ i da u njemu ima samo jednu nulu α (ne računajući višestrukosti).[1]

Osnovne karakteristike ovih metoda su sigurnost[2] i brzina konvergencija. Obično je slučaj da brze metode nemaju sigurnu konvergenciju, dok je sporije metode imaju. Brzinu konvergencije izražavamo preko reda konvergencije metode.

Definicija 4.0.1 *Red konvergencije niza brojeva x_n, $n \in \mathbb{N}_0$, koji konvergira ka α je p, ako je*

$$|x_n - \alpha| \leq c\,|x_{n-1} - \alpha|^p, \quad c \in \mathbb{R}_0^+\,. \tag{4.1}$$

Broj c se naziva faktor konvergencije. Ako je $p = 1$, mora biti $c < 1$.

[1] Ovo ćemo uvek podrazumevati i kad posebno ne naglasimo.
[2] U smislu da li uvek konvergiraju ili ne.

Ako je $p = 1$ i $c < 1$, iz (4.1) se lako dobija

$$|x_n - x| \leq c^n \, |x_0 - x|, \quad c \in \mathbb{R}_0^+ \, . \tag{4.2}$$

Ponekad je mnogo lakše pokazati (4.2) nego (4.1). I u ovom slučajemo kažemo da niz iteracija konvergira linearno sa faktorom c.

Za primenu ovih metoda potrebno je imati ocenu za udaljenost iteracije od α, da bismo znali kad prekinuti sa računanjem.

Teorema 4.0.1 *Ako je funkcija f diferencijabilna na $[a, b]$, $|f'(x)| \geq m_1 > 0$ za sve $x \in [a, b]$ i $f(\alpha) = 0$, tada je*

$$|x - \alpha| \leq \frac{|f(x)|}{m_1} \, .$$

Dokaz. Na osnovu teoreme o srednjoj vrednosti postoji $\xi \in (a, b)$ takvo da je

$$f(x) - f(\alpha) = f'(\xi) \, (x - \alpha) \, ,$$

odnosno, s obzirom na pretpostavke,

$$|f(x)| = |f'(\xi)| \, |x - \alpha| \geq m_1 \, |x - \alpha| \, .$$

□

Za upotrebu ove teoreme moramo nakon svake izračunate iteracije, izračunati i vrednost funkcije u njoj, što može zahtevati izvođenje velikog broja operacija.

4.1 Metoda polovljenja intervala

Metoda polovljenja intervala je najjednostavnija metoda za nalaženje nula funkcije f. Jedina pretpostavka za njenu primenu je

$$f(a) \cdot f(b) < 0 \, .$$

Označimo sa α pravu nulu funkcije f, i zatim

$$a_0 = a\,, \quad b_0 = b\,, \quad x_0 = \frac{a_0 + b_0}{2}\,.$$

Algoritam se sastoji od koraka sledećeg oblika. U n-tom koraku se konstruiše interval $[a_n, b_n]$, tako da bude duplo kraći od prethodnog intervala i da α ostane unutar $[a_n, b_n]$. Interval $[a_n, b_n]$ se konstruiše polovljenjem intervala $[a_{n-1}, b_{n-1}]$ tačkom x_{n-1}, i to tako da je

$$\begin{aligned} a_n &= x_{n-1},\ b_n = b_{n-1}, \qquad \text{ako je} \qquad f(a_{n-1}) \cdot f(x_{n-1}) > 0\,, \\ a_n &= a_{n-1},\ b_n = x_{n-1}, \qquad \text{ako je} \qquad f(a_{n-1}) \cdot f(x_{n-1}) < 0\,. \end{aligned}$$

Postupak zaustavljamo kad je $b_n - x_n \leq \varepsilon$.

Grešku n-te iteracije možemo odrediti na sledeći način,

$$|x_n - \alpha| \leq b_n - x_n = \frac{1}{2}\,(b_n - a_n) = \frac{1}{2^2}\,(b_{n-1} - a_{n-1}) = \ldots = \frac{1}{2^{n+1}}\,(b - a).$$

Kako je $(b-a)/2 = b - x_0$, prethodnu ocenu možemo pisati kao

$$|x_n - \alpha| \leq \frac{1}{2^n}\,(b - x_0)\,. \tag{4.3}$$

Ova ocena liči na (4.2), ali na desnoj strani, umesto $|x_0 - \alpha|$, imamo nešto veći broj. Ipak, desna strana daje nam da naslutimo da će konvergencija biti dosta spora.

Odredićemo još i broj koraka potrebnih da bi se odredilo rešenje s tačnošću ε. Da bismo postigli da je $|x_n - \alpha| \leq \varepsilon$, dovoljno je zahtevati

$$\frac{1}{2^{n+1}}\,(b - a) \leq \varepsilon\,,$$

a to je ekvivalentno sa

$$n \geq \frac{\ln(b-a) - \ln\varepsilon}{\ln 2} - 1, \quad n \in \mathbb{N}_0\,.$$

4.2 Metoda regula falsi

Videli smo da metoda polovljenja intervala ima sigurnu konvergenciju, ali je ona dosta spora. Prirodan pokušaj da se ona ubrza je metoda regula falsi (ili metoda pogrešnog položaja). Opet je cilj u svakom narednom koraku konstriusati kraći interval u kome leži traženo α, s tom razlikom što sada iteracije x_n računamo na drugi način.

Funkciju f aproksimiramo pravom koja prolazi kroz tačke $(a_n, f(a_n))$ i $(b_n, f(b_n))$. Njena jednačina je

$$y - f(b_n) = \frac{f(a_n) - f(b_n)}{a_n - b_n}\,(x - b_n), \text{ ili } y - f(a_n) = \frac{f(b_n) - f(a_n)}{b_n - a_n}\,(x - a_n)\,.$$

Presečnu tačku ove prave sa x osom označimo sa x_n. Ona leži unutar intervala $[a_n, b_n]$ jer su tačke $(a_n, f(a_n))$ i $(b_n, f(b_n))$ sa različitih strana x ose. Zatim, kao u metodi polovljenja intervala, pomeramo ili tačku a_n ili tačku b_n u x_n. Tačku x_n je lako izračunati

$$x_n = b_n - f(b_n)\,\frac{b_n - a_n}{f(b_n) - f(a_n)} = a_n - f(a_n)\,\frac{a_n - b_n}{f(a_n) - f(b_n)}\,. \tag{4.4}$$

Pod predpostavkom $f(a) \cdot f(b) < 0$, metoda regula falsi uvek konvergira ka α. Međutim, pitanje je koliko smo uspeli u pokušaju da ubrzamo konvergenciju metode polovljenja intervala.

Uz dodatne pretpostavke za f, pokazaćemo da je konvergencija ove metode linearna. Ako jednakost (4.4) za $n = 0$ pomnožimo sa -1 i dodamo α s obe strane, dobijamo

$$\begin{aligned}
\alpha - x_0 &= \alpha - b + \frac{f(b)}{[a, b\,; f]} = (\alpha - b)\left(1 + \frac{f(b)}{(\alpha - b)\,[a, b\,; f]}\right) \\
&= (\alpha - b)\left(1 + \frac{f(b) - f(\alpha)}{(\alpha - b)\,[a, b\,; f]}\right) \\
&= (\alpha - b)\left(1 + (b - \alpha)\,\frac{[b, \alpha\,; f]}{(\alpha - b)\,[a, b\,; f]}\right)
\end{aligned}$$

$$
\begin{aligned}
&= (\alpha - b)\left(1 - \frac{[b, \alpha\,; f]}{[a, b\,; f]}\right) = (\alpha - b)\left(\frac{[a, b\,; f] - [b, \alpha\,; f]}{[a, b\,; f]}\right) \\
&= -(\alpha - b)(\alpha - a)\frac{[a, b, \alpha\,; f]}{[a, b\,; f]}\,.
\end{aligned}
$$

Uvedimo pretpostavku $f \in C^2[a, b]$. Na osnovu teoreme o srednjoj vrednosti, postoje ξ i ζ iz $[a, b]$, takvi da je

$$
[a, b\,; f] = f'(\xi)\,, \quad [a, b, \alpha\,; f] = \frac{1}{2} f''(\zeta)\,.
$$

Sada je

$$
\alpha - x_0 = -(\alpha - b)(\alpha - a)\frac{f''(\zeta)}{2\, f'(\xi)}\,. \tag{4.5}
$$

Neka su još prvi i drugi izvod funkcije f stalnog znaka na $[a, b]$. Zbog međusobne sličnosti, razmotrićemo samo jedan od moguća četiri slučaja, npr. $f'(x) > 0$, $f''(x) > 0$, $\forall x \in [a, b]$.

Tada je f rastuća i konveksna, pa se duž koja spaja tačke $(a, f(a))$ i $(b, f(b))$ uvek nalazi iznad funkcije f. Desna strana u (4.5) je pozitivna, pa je $\alpha > x_0$, što znači da se u sledećem koraku pomera levi kraj intervala. Isto se događa u svim narednim koracima, tj. desni kraj intervala b je fiksiran, a α uvek ostaje desno od iteracija x_n. Na osnovu (4.5) zaključujemo da je

$$
\alpha - x_n = -(\alpha - b)(\alpha - a_n)\frac{f''(\zeta_n)}{2\, f'(\xi_n)}\,,
$$

tj. konvergencija metode regula falsi je linearna. Poredeći prethodnu relaciju sa (4.3), vidimo da postoje slučajevi kad metoda polovljenja intervala brže konvergira od regula falsi. Prilikom računanja aproksimacija ne moramo proveravati u kom intervali je α, a formule (4.4) su sada

$$
x_{-1} = a, \quad x_n = x_{n-1} - f(x_{n-1})\frac{x_{n-1} - b}{f(x_{n-1}) - f(b)}, \quad n \in \mathbb{N}_0\,. \tag{4.6}
$$

Isto važi i kad su f' i f'' negativni na $[a, b]$.

Kad su f' i f'' različitog znaka na $[a, b]$, fiksiran je desni kraj intervala, a uvek se pomiče b, tj. iteracije x_n su uvek desno od α. Formule (4.4) su sada

$$x_{-1} = b, \quad x_n = x_{n-1} - f(x_{n-1}) \frac{x_{n-1} - a}{f(x_{n-1}) - f(a)}, \quad n \in \mathbb{N}_0 .$$

4.3 Metoda sečica

Ako u metodi regula falsi prave linije ne povlačimo kroz tačke $(a_n, f(a_n))$ i $(b_n, (f(b_n)))$, već kroz $(x_{n-1}, f(x_{n-1}))$ i $(x_n, f(x_n))$, dobijamo metodu sečica. Ovim smo izgubili svojstvo sigurne konvergencije, a pokušavamo povećati brzinu konvergencije.

Formule (4.4) postaju

$$x_{n+1} = x_n - f(x_n) \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, \quad n \in \mathbb{N} .$$

Sad se postavljaju pitanja kako odrediti početne aproksimacije x_0 i x_1, tako da niz x_n konvergira ka α, i kog je reda ta konvergencija.

Pretpostavljaćemo da je α jednostruka nula funkcije f, tj. postoji interval $I = [\alpha - \varepsilon, \alpha + \varepsilon]$, $\varepsilon > 0$, takav da je $f'(x) \neq 0$, $\forall x \in I$. Pretpostavljaćemo još da $f \in C^2(I)$. Ako su početne aproksimacije x_0 i x_1 dovoljno blizu α, niz x_n konvergira ka α sa redom p, gde je

$$p = \frac{1 + \sqrt{5}}{2} \approx 1,618 .$$

Dokažimo ovo tvrđenje. Relacija (4.5) je sada

$$\alpha - x_{n+1} = -(\alpha - x_n)(\alpha - x_{n-1}) \frac{f''(\zeta_n)}{2 f'(\xi_n)} . \tag{4.7}$$

Definišimo broj M,

$$M = \frac{\max\limits_{x \in I} |f''(x)|}{2 \min\limits_{x \in I} |f'(x)|} . \tag{4.8}$$

Na osnovu (4.7), za svako x_0 i x_1 iz intervala I važi

$$|\alpha - x_2| \leq |\alpha - x_1|\,|\alpha - x_0|\,M.$$

Da bismo skratili zapis, označimo grešku n-te aproksimacije sa e_n, tj. $e_n = \alpha - x_n$. Množenjem prethodne nejednakosti sa M dobijamo

$$M\,|e_2| \leq M\,|e_1|\,M\,|e_0|\,.$$

Konačno, pretpostavimo da su x_0 i x_1 toliko blizu α da vredi

$$\delta = \max\{M|e_0|, M|e_1|\} < 1\,.$$

Sada je $M|e_2| \leq \delta^2 < \delta$, pa je

$$|e_2| < \frac{\delta}{M} = \max\{|e_0|, |e_1|\} \leq \varepsilon\,.$$

Dakle, uz sve navedene pretpostavke važi $x_2 \in I$. Nastavimo li ovaj postupak dobijamo $|e_3| \leq \delta^3/M$, $e_4 \leq \delta^5/M$, ... Uopšteno, ako je

$$M\,|e_{n-1}| \leq \delta^{q_{n-1}}\,, \quad M\,|e_n| \leq \delta^{q_n}\,,$$

onda je

$$M\,|e_{n+1}| \leq M\,|e_n|\,M\,|e_{n-1}| \leq \delta^{q_n+q_{n-1}} = \delta^{q_{n+1}}\,.$$

Vidimo da je niz q_n definisan rekurzijom

$$q_{n+1} = q_n + q_{n-1}\,, \quad q_0 = q_1 = 1\,.$$

Radi se o poznatom Fibonaccijevom nizu, koji očigledno teži ka ∞. Kako je $\delta < 1$ zaključujemo da $e_n \to 0$, tj. $x_n \to \alpha$.

Neka je ova konvergencija reda r. To znači da je r najveći broj za koji važi

$$|e_{n+1}| \leq c_2\,|e_{n-1}|^{r^2}\,, \quad n = 2, 3, \ldots\,. \tag{4.9}$$

Iz relacije (4.7) dobijamo

$$|e_{n+1}| \leq M\,|e_n|\,|e_{n-1}| \leq M\,c\,|e_{n-1}|^{r+1}\,, \quad n = 2, 3, \ldots\,, \tag{4.10}$$

a r je najveći broj za koji to vredi. Iz (4.9) i (4.10) zaključujemo da je r pozitivan koren jednačine

$$r^2 = r + 1 \Rightarrow r = \frac{1+\sqrt{5}}{2}\,.$$

Uspeli smo u nameri da ubrzamo konvergenciju metode regula falsi, ali ostaje problem kako odrediti x_0 i x_1.

PRIMER 4.1. Jednačina $x^2 - e^x + 2 = 0$ ima jedinstven prost koren na intervalu $[1, 2]$. Lako se proverava da su prvi i drugi izvod funkcije $f(x) = x^2 - e^x + 2$ negativni na $[1, 2]$. Ako primenimo metodu regula falsi, iteracije se računaju po formuli (4.6), i one sa leve strane teže ka rešenju α. Tako dobijamo

n	x_n	n	x_n
0	1.16861	8	1.31876
1	1.24873	9	1.31893
2	1.28644	10	1.31901
3	1.30400	11	1.31904
4	1.31213	12	1.31906
5	1.31588	13	1.31907
6	1.31760	14	1.31907
7	1.31840		

Računajući iteracije metodom sečica, polazeći od $x_0 = 1$ i $x_1 = 2$ dobijamo

n	x_n
2	1.16861
3	1.24873
4	1.32745
5	1.31867
6	1.31907
7	1.31907

Očekivano, druga metoda konvergira znatno brže, ali generalno, ne možemo biti sigurni u njenu konvergenciju, pogotovo ne sa ovakvim izborom početnih iteracija.

4.4 Hibridna Brent-Dekkerova metoda

U metodi regula falsi i metodi sečica smo funkciju f aproksimirali interpolacionim polinomom prvog stepena. Možda bi se mogle dobiti brže metode ako funkciju f aproksimiramo interpolacionim polinomom većeg stepena? Za takvu interpolaciju nam treba više tačaka, pa se kvadratna interpolacija nameće kao prvi izbor.

Brent-Dekkerova metoda je smišljena kao metoda koja bi trebala konvergirati brže od metode sečica (u najboljem slučaju kvadratno) i uz to imati sigurnu konvergenciju. Sastoji se iz uzastopnih ponavljanja koraka od kojih svaki može obuhvatati delove koje grubo možemo opisati kao inverznu kvadratnu interpolaciju, metodu polovljenja intervala i metodu sečice (preciznije jedan njihov korak).

Korak počinje tako što se metodom sečica odredi treća tačka. Prema nekom kriterijumu se procenjuje da li je ta tačka dobra. Ako jeste formira se kvadratni interpolacioni polinom P_2 kroz poslednje tri tačke, čija nula se uzima za četvrtu tačku. Odnosno, formira se P_2^{-1}, a četvrta tačka je $P_2^{-1}(0)$. Ako je treća tačka odbačena kao loša, radi se jedan korak metode polovljenja intervala.

Što se manje puta bude morala koristiti metoda polovljenja intervala, to će konvergencija biti brža. Precizni kriterijumi kako se procenjuje da li je neka tačka dobra ili loša su dosta složeni, pa ih izostavljamo. Metoda je sastavni velikih numeričkih biblioteka programa.

4.5 Metoda tangenti

Ako u metodi sečica $[x_n, x_{n-1}\,; f]$ zamenimo sa $f'(x_n)$ (uz pretpostavku $f'(x_n) \neq 0$) dobijamo sledeću formulu za generisanje niza x_n,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\,. \tag{4.11}$$

Motivacija za ovo je očigledna, naročito kad su x_n i x_{n-1} veoma blizu. Nova iteracija x_{n+1} je nula tangente funkcije f u tački $(x_n, f(x_n))$. Ovu metodu zovemo još i Newtonova metoda.

Do nje možemo doći i na drugi način. Pretpostavimo da funkciju f možemo razviti u Taylorov red oko x_n zaključno sa prvim članom, tj.

$$f(x) = f(x_n) + f'(x_n)\,(x - x_n) + \frac{f''(\xi_n)}{2}\,(x - x_n)^2\,,$$

pri čemu je ξ_n između x i x_n. Uvrštavanjem $x = \alpha$ dobijamo

$$0 = f(\alpha) = f(x_n) + f'(x_n)\,(\alpha - x_n) + \frac{f''(\xi_n)}{2}\,(\alpha - x_n)^2\,, \tag{4.12}$$

a zatim i

$$\alpha = x_n - \frac{f(x_n)}{f'(x_n)} - (\alpha - x_n)^2 \frac{f''(\xi_n)}{2f'(x_n)}\,.$$

Ako x_{n+1} definišemo pomoću (4.11), dobijamo

$$\alpha - x_{n+1} = -(\alpha - x_n)^2 \frac{f''(\xi_n)}{2f'(x_n)}\,. \tag{4.13}$$

Vidimo da će, ako konvergira, metoda tangenti konvergirati sa redom dva. Zaključak važi samo ako je $f'(x_n) \neq 0$, za svako n.

Nametnimo, za početak, funkciji f iste uslove kao u metodi sečica.

Teorema 4.5.1 *Neka je $f(\alpha) = 0$, f klase C^2 na nekoj okolini od α i neka je $f'(\alpha) \neq 0$. Ako je početna iteracija x_0 dovoljno blizu α, niz x_n konvergira ka α.*

Dokaz. Kako je $f'(\alpha) \neq 0$, možemo izabrati $I = [\alpha - \varepsilon, \alpha + \varepsilon]$, takav da $f'(x) \neq 0$, $\forall x \in I$. Broj M je dobro definisan pomoću (4.8). Na osnovu (4.13), za svako $x_0 \in I$ važi

$$|\alpha - x_1| \leq M\,|\alpha - x_0|^2\,.$$

Izaberimo x_0 tako da, osim što leži u I, vredi i $M|\alpha - x_0| < 1$. Sada je

$$|\alpha - x_1| \leq |\alpha - x_0| \leq \varepsilon\,, \quad M\,|\alpha - x_1| < 1\,.$$

Indukcijom se lako pokazuje da isto važi i za svako x_n, $n \geq 1$. Dalje imamo

$$M\,|\alpha - x_n| \leq (M\,|\alpha - x_{n-1}|)^2 \leq \ldots \leq (M\,|\alpha - x_0|)^{2^n}\,.$$

Iz $M|\alpha - x_0| < 1$, sledi $x_n \to \alpha$, kad $n \to \infty$. □

Pošto ξ_n iz (4.13) leži između x_n i α, važi i $\xi_n \to \alpha$, kad $n \to \infty$, pa je

$$\lim_{n\to\infty} \frac{\alpha - x_{n+1}}{(\alpha - x_n)^2} = -\lim_{n\to\infty} \frac{f''(\xi_n)}{2f'(x_n)} = -\frac{f''(\alpha)}{2f'(\alpha)}\,.$$

Ako funkciji f nametnemo još neke uslove, niz x_n će konvergirati ka α pod puno slabijim uslovima za x_0.

Teorema 4.5.2 *Neka je $f \in C^2[a,b]$, $f(a)\cdot f(b) < 0$ i neka su f' i f'' stalnog znaka na $[a,b]$. Ako polazna iteracija x_0 iz $[a,b]$ ispunjava uslov*

$$f(x_0)\cdot f''(x_0) > 0\,,$$

onda niz iteracija dobijen Newtonovom metodom konvergira prema (prostoj) nuli funkcije f.

Dokaz. Pretpostavimo da su f' i f'' pozitivni na $[a,b]$. Tada je f rastuća funkcija, pa je $f(a) < 0$ i $f(b) > 0$. Za x_0 možemo uzeti bilo koji broj iz $[a,b]$, takav da je $f(x_0) > 0$. Indukcijom pokazujemo da je $\alpha < x_n \leq x_0$,

$\forall n \in \mathbb{N}_0$. Tvrđenje je očigledno za $n = 0$. Pretpostavimo da je $\alpha < x_n \leq x_0$. Tada je $f(x_n) > 0$ i $f'(x_n) > 0$, pa je

$$x_{n+1} < x_n \leq x_0 \,,$$

tj. niz x_n je monotono opadajući. Nejednakosti $x_{n+1} > \alpha$ sledi iz (4.12) jer je $f''(\xi_n) > 0$.

Usput je pokazana i konvergencija niza x_n jer je pokazano da je on monotono opadajući i ograničen odozdo. Neka je

$$\lim_{n\to\infty} x_n = \alpha' \,, \quad \alpha' \in [\alpha, x_0] \,.$$

Prelaskom na limese u (4.11) dobijamo

$$\alpha' = \alpha' - \frac{f(\alpha')}{f'(\alpha')} \,,$$

odakle sledi $f(\alpha') = 0$. Kako je α jedina nula od f iz intervala $[a, b]$, sledi $\alpha = \alpha'$.

Preostala tri slučaja se dokazuju analogno. □

Uslov $f(x_0) \cdot f''(x_0) > 0$ ima jednostavnu geometrijsku interpretaciju. Početnu iteraciju treba izabrati na "strmijoj" strani grafika funkcije.

Izvešćemo još ocenu greške pogodnu za praktičnu upotrebu. Za dve susedne iteracije vredi

$$f(x_n) = f(x_{n-1}) + f'(x_{n-1})\,(x_n - x_{n-1}) + \frac{f''(\xi_{n-1})}{2}\,(x_n - x_{n-1})^2 \,,$$

pri čemi je ξ_{n-1} između x_{n-1} i x_n. Po definiciji iteracija vredi i

$$f(x_{n-1}) + f'(x_{n-1})\,(x_n - x_{n-1}) = 0 \,,$$

pa je

$$f(x_n) = \frac{f''(\xi_{n-1})}{2}\,(x_n - x_{n-1})^2 \,,$$

odnosno

$$|f(x_n)| \leq \frac{M_2}{2} (x_n - x_{n-1})^2 .$$

Prema teoremi 4.0.1 je

$$|\alpha - x_n| \leq \frac{|f(x_n)|}{m_1} .$$

Iz prethodne dve nejednakosti dobijamo

$$|\alpha - x_n| \leq \frac{M_2}{2m_1} (x_n - x_{n-1})^2 .$$

Ako je cilj dobiti rešenje sa tačnošću ε, proces računanja iteracija zaustavljamo kad bude ispunjen uslov

$$|x_n - x_{n-1}| \leq \sqrt{\frac{2m_1\varepsilon}{M_2}} .$$

Jasno, nejednakost vredi do na grešku zaokruživanja.

Primer 4.2. Ovaj primer služi za ilustraciju razlike u brzini kvadratne i linearne konvergencije. Jednačina $x^3 - 1.5 = 0$ ima jedinstven prost koren na intervalu $[1, 1.5]$.

Da bismo rešili ovu jednačinu s tačnošću 10^{-8} metodom polovljenja intervala, moramo izračunati 27 iteracija (računajući i x_0), a približno rešenje je $x_{26} = 1.144714239984751$.

Newtonovom metodom (polazeći od $x_0 = 1.5$) u samo sedam iteracija dobijamo rešenje sa greškom manjom od 10^{-15}. Približno rešenje je $x_7 = 1.144714242553332$.

Prilikom primene Newtonove metode javljaju se sledeći problemi. U svakom koraku mora se izračunati i vrednost funkcije i vrednost njenog izvoda u nekoj tački, što u slučaju komplikovanih izvoda (vrednost funkcije moramo računati u svakoj metodi) zahteva izvođenje velikog broja operacija. To može pokvariti činjenicu da je ova metoda najbrža od do sada posmatranih. U

praksi često imamo samo interval $[a, b]$ u kojem smo izolovali funkciju f, a nemamo dodatne informacije o funkciji f na osnovu kojih bismo zaključili da metoda (ili neka druga koja spada u grupu bržih) konvergira (npr. ne znamo izvode).

Prvi problem se ponekad rešava upotrebom sledeće modifikacije Newtonove metode

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}\,.$$

Tangentu krive f u tački $(x_n, f(x_n))$ zamenjujemo pravom koja prolazi kroz istu tačku i koja je paralelna tangenti u $(x_0, f(x_0))$. Smanjenje broja operacija potrebnih u jednom koraku, plaćeno je smanjenjem brzine konvergencije. Naime, red konvergencije modifikovane Newtonove metode je jedan.

Drugi problem možemo rešiti kombinovanjem Newtonove metode sa metodom polovljenja intervala, na sledeći način. Novu iteraciju prvo izračunamo po Newtonovoj metodi i ako je ona u trenutnom intervalu, koristimo je za skraćivanje intervala. U suprotnom, interval skraćujemo jednim korakom metode polovljenja intervala.

Razmotrimo još i slučaj kad je α višestruka nula funkcije f, tj. neka f ima neprekidnih prvih p izvoda i neka je

$$f(\alpha) = f'(\alpha) = \ldots = f^{(p-1)}(\alpha) = 0, \quad f^{(p)}(\alpha) \neq 0\,.$$

Funkcija f se tada može zapisati u obliku

$$f(x) = (x-\alpha)^p\, h(x), \quad h(\alpha) \neq 0\,.$$

Funkcija

$$u(x) = \frac{f(x)}{f'(x)} = \frac{(x-\alpha)^p\, h(x)}{(x-\alpha)^{p-1}\,[p\, h(x) + (x-\alpha)\, h'(x)]} = \frac{(x-\alpha)\, h(x)}{p\, h(x) + (x-\alpha)\, h'(x)}\,,$$

ima jednostruku nulu u α, pa Newtonovu metodu, umesto na f, možemo

primeniti na u. Niz x_n generišemo pomoću

$$x_{n+1} = x_n - \frac{u(x_n)}{u'(x_n)}\,, \tag{4.14}$$

gde je

$$u'(x) = \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2} = 1 - \frac{f''(x)}{f'(x)}\,u(x)\,.$$

Dobra strane ovog postupka je što ne moramo znati red nule funkcije, a loša što, pored f i f', moramo računati i f''. Slučaj kad znamo p razmotrićemo u narednom poglavlju.

4.6 Metoda proste iteracije

Jednačinu oblika $f(x) = 0$ možemo na beskonačno mnogo načina predstaviti u ekvivalentnom obliku

$$x = g(x)\,.$$

Rešenja ove jednačine su fiksne tačke funkcije g, koju koristimo za generisanje niza x_n

$$x_{n+1} = g(x_n)\,, \quad n \in \mathbb{N}_0\,.$$

Ovakve iterativne metode zovemo prostim, jer se u izračunavanju nove iteracije od već izračunatih iteracija koristi samo prethodna. Sledeća tvrđenja daju odgovor na pitanje kada niz x_n konvergira ka α.

Lema 4.6.1 *Neka je $g \in C[a,b]$ i neka je $g([a,b]) \subseteq [a,b]$. Tada jednačina $g(x) = x$ ima bar jedno rešenje u $[a,b]$.*

Dokaz. Pod uslovima leme, neprekidna funkcija $g(x) - x$ na $[a,b]$ ima sledeće osobine

$$g(a) - a \geq 0\,, \quad g(b) - b \leq 0\,.$$

Dakle, ili je funkcija $g(x) - x$ je promenila znak na $[a,b]$ ili je jedan od krajeva intervala njena nula, što znači da ima bar jednu nulu u $[a,b]$. □

Teorema 4.6.1 *Neka je* $g \in C^1[a,b]$, $g([a,b]) \subseteq [a,b]$ *i neka je*

$$\max_{x\in[a,b]} |g'(x)| = \lambda < 1\,. \tag{4.15}$$

Tada vredi:

1. *Jednačina* $x = g(x)$ *ima tačno jedno rešenje* α *na intervalu* $[a,b]$,
2. *Za proizvoljno* $x_0 \in [a,b]$ *i niz generisan pomoću* $x_{n+1} = g(x_n)$, $n \geq 0$, *vredi*
   (a) $\lim_{n\to\infty} x_n = \alpha\,,$
   (b) $|\alpha - x_n| \leq \lambda^n\,|\alpha - x_0|\,,$
   (c) $\lim_{n\to\infty} \dfrac{\alpha - x_{n+1}}{\alpha - x_n} = g'(\alpha)\,.$

Dokaz. Prema prethodnoj lemi postoji bar jedno rešenje posmatrane jednačine. Pretpostavimo da jednačina ima još jedno rešenje $\beta \in [a,b]$. Na osnovu teoreme o srednjoj vrednosti je

$$|\alpha - \beta| = |g(\alpha) - g(\beta)| = |g'(\xi)|\,|\alpha - \beta|\,, \quad \xi \in (\alpha,\beta)\,,$$

odakle, koristeći (4.15), dobijamo

$$|\alpha - \beta| \leq \lambda\,|\alpha - \beta|\,,$$

a ovo je moguće samo kad je $\alpha = \beta$. Ovim je dokazan prvi deo teoreme.

Opet koristimo teoremu o srednjoj vrednosti

$$\alpha - x_{n+1} = g(\alpha) - g(x_n) = g'(\xi_n)\,(\alpha - x_n)\,,$$

gde je ξ_n neki broj između α i x_n. Iz uslova (4.15) sledi

$$|\alpha - x_{n+1}| \leq \lambda\,|\alpha - x_n|\,,$$

na osnovu čega, indukcijom po n, dobijamo

$$|\alpha - x_n| \leq \lambda^n |\alpha - x_0| .$$

Ako pustimo $n \to \infty$, onda $\lambda^n \to 0$, pa važi $x_n \to \alpha$. Kako je ξ_n između α i x_n, i ono teži ka α, a onda je i

$$\lim_{n\to\infty} \frac{\alpha - x_{n+1}}{\alpha - x_n} = \lim_{n\to\infty} g'(\xi_n) = g'(\alpha) .$$

□

Iz prethodne relacije vidimo da red konvergencije može biti veći od jedan samo ako je $g'(\alpha) = 0$. Prethodna teorema ukazuje na šta moramo obratiti pažnju prilikom prelaska sa oblika $f(x) = 0$ na $x = g(x)$. Uglavnom, problemi nastaju oko uslova (4.15). Ako je on ispunjen, da bi važilo $g([a, b]) \subseteq [a, b]$ dovoljno je da $[a, b]$ obuhvata interval $[\alpha - x_0, \alpha + x_0]$. Dakle, za konvergenciju metode je dovoljno da je $|g'(\alpha) < 1|$ i da je x_0 dovoljno blizu α. Jedna od mogućnosti da uslov (4.15) bude ispunjen je da stavimo

$$g(x) = x - \frac{\operatorname{sign}(f')}{M_1} f(x) , \quad M_1 = \max_{x\in[a,b]} |f'(x)| .$$

Jedan od neophodnih uslova za konvergenciju metode je $g'(\alpha) \leq 1$. Naime, ako pretpostavimo da je $g'(\alpha) > 1$ iz

$$\alpha - x_{n+1} = g(\alpha) - g(x_n) = g'(\xi_n)\,(\alpha - x_n)$$

i za x_n dovoljno blizu α da važi $|g'(\xi_n)| > 1$, sledi $|\alpha - x_{n+1}| \geq |\alpha - x_n|$, pa konvergencija metode nije moguća.

Primer 4.3 Nađimo pozitivno rešenje jednačine

$$x^2 - a = 0 , \quad a > 0 .$$

1. Ako jednačinu napišemo obliku $x = x^2 + x - a$, tj. $g(x) = x^2 + x - a$, dobijamo

$$g'(\sqrt{a}) = 2\sqrt{a} + 1 > 1 ,$$

pa odgovarajuća metoda proste iteracije neće konvergirati.

2. Ako stavimo da je $g(x) = a/x$, dobijamo $g'(\sqrt{a}) = -1$, pa ne možemo biti sigurni da će metoda konvergirati.

3. Ako je $g(x) = 0.5(x + a/x)$, onda je $g'(\sqrt{a}) = 0$, pa će konvergencija metode[3] biti bar kvadratna.

4. Odgovarajuću funkciju g možemo odrediti na sledeći način. Stavimo da je $g(x) = x + c\, f(x)$ i odredimo konstantu c tako da je $-1 < g'(\sqrt{a}) < 1$. Iz $g'(x) = 1 + 2cx$ dobijamo

$$g'(\sqrt{a}) = 1 + 2\, c\sqrt{a} \quad \Longrightarrow \quad -\frac{1}{\sqrt{a}} < c < 0\,.$$

Teorema 4.6.2 *Neka je $\alpha \in [a, b]$ rešenje jednačine $x = g(x)$, $g \in C^p[a, b]$, $p \geq 2$ i neka je*

$$g'(\alpha) = \ldots = g^{(p-1)}(\alpha) = 0\,.$$

Ako metoda proste iteracije $x_{n+1} = g(x_n)$ konvergira ka α, njen red konvergencije je p i važi

$$\lim_{n\to\infty} \frac{x_{n+1} - \alpha}{(x_n - \alpha)^p} = \frac{g^{(p)}(\alpha)}{p\,!}\,.$$

Dokaz Ako razvijemo $g(x)$ u Taylorov red zaključno sa $(p-1)$-im stepenom i uvrstimo $x = x_n$, dobijamo

$$x_{n+1} = g(x_n) = g(\alpha) + \ldots + \frac{g^{(p-1)}(\alpha)}{(p-1)\,!}\,(x_n - \alpha)^{p-1} + \frac{g^{(p)}(\xi_n)}{p\,!}\,(x_n - \alpha)^p\,,$$

za neko ξ_n između x_n i α. Iskoristimo li pretpostavke teoreme, dobijamo

$$x_{n+1} = \alpha + \frac{g^{(p)}(\xi_n)}{p\,!}\,(x_n - \alpha)^p\,.$$

Iz $x_n \to \alpha$ sledi $\xi_n \to \alpha$, pa iz prethodne jednakosti lako dobijamo traženu relaciju. □

[3]Naravno, uz ispunjenost ostalih uslova.

Primenimo rezultate poslednje teoreme na Newtonovu metodu, koja očigledno pripada posmatranoj klasi metoda. Za nju već znamo da ima kvadratnu konvergenciju uz pretpostavku $f'(x) \neq 0$, što znači da mora biti $g'(\alpha) = 0$. Zaista, kako je

$$g'(x) = \left(x - \frac{f(x)}{f'(x)} \right)' = \frac{f(x)\, f''(x)}{[f'(x)]^2} \,,$$

sledi $g'(\alpha) = 0$. Prethodna teorema sugeriše da bi red konvergencije Newtonove metode u nekim slučajevima mogao biti veći od dva. Iz

$$g''(\alpha) = \frac{f''(\alpha)}{f'(\alpha)}$$

zaključujemo da će red konvergencije biti bar tri ako je $f''(\alpha) = 0$.

Razmotrimo još jednom slučaj kad je α nula funkcije f reda $p > 1$, s tim što ćemo sada pretpostavljati da je p poznato. Lako se može pokazati da je tada

$$g'(\alpha) = 1 - \frac{1}{p} \neq 0 \,.$$

Vidimo da uslov $f'(\alpha) \neq 0$ nije neophodan za konvergenciju, ali je ona sada linearna. Za $p = 2$, metoda će biti brza kao metoda polovljenja intervala, a još sporija za $p \geq 3$. Do brže metode možemo doći na sledeći način

$$g(x) = x - p\, \frac{f(x)}{f'(x)} \,. \tag{4.16}$$

Tada je

$$g'(x) = 1 - p + p\, \frac{f(x)\, f''(x)}{[f'(x)]^2} \,,$$

pa sledi da je

$$\lim_{x \to \alpha} g'(x) = 0 \,,$$

što znači da je konvergencija bar kvadratna.

Primer 4.4. Da bismo Newtonovom metodom odredili nulu funkcije

$$f(x) = x^3 - 5.56\, x^2 + 9.1389\, x - 4.68999 = 0$$

s tačnošću 10^{-9}, potrebno nam je trideset iteracija, što je, s obzirom na primer 4.2, neočekivano puno. Ovo nam signalizira da tražena nula nije prosta. Dobijeno približno rešenje je $x_{30} = 1.230000000463810$.

Možemo proveriti da je 1.23 dvostruka nula funkcije f, pa je konvergencija Newtonove metode bila linearna. Ako iteracije računamo po formuli (4.16) za $p = 2$, rešenje sa istom tačnošću dobijamo u svega sedam iteracija. Približno rešenje je $x_7 = 1.229999999995655$, a konvergencija je bila kvadratna.

Kvadratnu konvergenciju možemo postići i primenom formula (4.14).

4.7 Newtonova metoda za nelinearne sisteme

Newtonovu metodu za rešavanje nelinearnih jednačina je moguće uopštiti i dobiti metodu za rešavanje sistema nelinearnih jednačina, oblika

$$f_i(x_1, \ldots, x_n) = 0\,, \quad i = 1, \ldots, n\,,$$

gde su $f_i : \mathbb{R}^n \to \mathbb{R}$ zadane nelinearne funkcije. Sistem možemo zapisati i u obliku

$$F(x) = \mathbf{0}\,, \tag{4.17}$$

gde je $F = (f_1, \ldots, f_n) : \mathbb{R}^n \to \mathbb{R}^n$, $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ i $\mathbf{0}$ nula vektor iz $\mathbb{R}^n$. Jacobijevu matricu funkcije F ćemo označavati sa

$$W = \left[\frac{\partial f_i}{\partial x_j}\right]_{ij}\,,$$

a njenu determinantu sa J_F.

Interesuje nas kada će niz aproksimacija $x^{(n)}$ generisan pomoću

$$x^{(n+1)} = x^{(n)} - \left[W(x^{(n)})\right]^{-1} F(x^{(n)})\,, \quad n \in \mathbb{N}_0\,, \tag{4.18}$$

konvergirati ka rešenju sistema (4.17).

Teorema 4.7.1 *Neka je $F(\alpha) = 0$, F klase C^2 na nekoj okolini od α i neka je $J_F(\alpha) \neq 0$. Ako je početna iteracija $x^{(0)}$ dovoljno blizu α, niz $x^{(n)}$ konvergira kvadratno ka α.*

Dokaz. Uopštavamo dokaz teoreme 4.5.1 na višedimenzionalni slučaj. Na osnovu pretpostavki teoreme postoji okolina $U = \{x : ||x - \alpha|| < \varepsilon\}$ tačke α, takva da je $J_F(x) \neq 0$, za svako $x \in U$.

Funkciju F možemo razviti u Taylorov red oko α zaključno sa prvim članom,

$$\mathbf{0} = F(\alpha) = F(x^{(n)}) + W(x^{(n)}) \cdot (\alpha - x^{(n)}) + R_2(f, \alpha - x^{(n)}), \qquad (4.19)$$

gde je

$$||R_2(f, \alpha - x^{(n)})|| \leq M_1 \, ||\alpha - x^{(n)}||^2 .$$

Množenjem (4.19) sa matricom[4] $\left[W(x^{(n)})\right]^{-1}$ i uvrštavanjem (4.18), zaključujemo da postoji konstanta M takva da je

$$||\alpha - x^{(n+1)}|| \leq M \, ||\alpha - x^{(n)}||^2.$$

Vidimo da ako niz $x^{(n)}$ konvergira ka α, konvergencija je kvadratna.

Da bi niz $x^{(n)}$ konverigirao ka α, dovoljno je da $x^{(0)}$ izaberemo tako da je $||\alpha - x_0|| < \varepsilon$ i $M||\alpha - x_0|| < 1$. □

Izbor početne iteracije $x^{(0)}$ može predstavljati dosta težak problem. Ovaj postupak se još zove i metoda Newton-Kantoroviča.

Kao i kod klasične Newtonove metode, broj operacija potrebnih za dobijanje jedne iteracije možemo smanjiti sledećom modifikacijom

$$x^{(n+1)} = x^{(n)} - \left[W(x^{(0)})\right]^{-1} F(x^{(n)}) . \qquad (4.20)$$

Smanjenje broja operacija potrebnih u jednom koraku, plaćeno je smanjenjem brzine konvergencije.

[4]Matrica je regularna ako $x^{(n)} \in U$.

PRIMER 4.5. Ilustrujmo prethodno na sledećem sistemu nelinearnih jednačina, koji prvo rešavamo metodom Newton-Kantoroviča

$$\begin{array}{rcrcrcl} x^2 & + & y^2 & + & z^2 & = & 1 \\ 2x^2 & + & y^2 & - & 4z & = & 0 \\ 3x^2 & - & 4y & + & z^2 & = & 0 \end{array}$$

uzimajući početne vrednosti $x^{(0)} = [0.5, 0.5, 0.5]^T$.

Kako je

$$F = \begin{bmatrix} x^2 + y^2 + z^2 - 1 \\ 2x^2 + y^2 - 4z \\ 3x^2 - 4y + z^2 \end{bmatrix}, \quad W_F(x) = \begin{bmatrix} 2x & 2y & 2z \\ 4x & 2y & -4 \\ 6x & -4 & 2z \end{bmatrix},$$

dobijamo

$$F(x^{(0)}) = \begin{bmatrix} -0.25 \\ -1.25 \\ -1.00 \end{bmatrix}, \quad W_F(x^{(0)}) = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & -4 \\ 3 & -4 & 1 \end{bmatrix},$$

i

$$[W_F(x^{(0)})]^{-1} = -\frac{1}{40} \begin{bmatrix} -15 & -5 & -5 \\ -14 & -2 & 6 \\ -11 & 7 & -1 \end{bmatrix}.$$

Prva iteracija je

$$\begin{aligned} x^{(1)} &= x^{(0)} - [W_F(x^{(0)})]^{-1} F(x^{(0)}) \\ &= \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix} + \frac{1}{40} \begin{bmatrix} -15 & -5 & -5 \\ -14 & -2 & 6 \\ -11 & 7 & -1 \end{bmatrix} \begin{bmatrix} -0.25 \\ -1.25 \\ -1.00 \end{bmatrix} \\ &= \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix} + \begin{bmatrix} 0.375 \\ 0.000 \\ -0.125 \end{bmatrix} = \begin{bmatrix} 0.875 \\ 0.500 \\ 0.375 \end{bmatrix}. \end{aligned}$$

Nastavljajući dalje, dobijamo

$$x^{(2)} = \begin{bmatrix} 0.78981 \\ 0.49662 \\ 0.36993 \end{bmatrix}, \quad x^{(3)} = \begin{bmatrix} 0.78521 \\ 0.49662 \\ 0.36992 \end{bmatrix}, \quad \ldots$$

Ako se zaustavimo na trećem koraku imamo

$$x \approx \begin{bmatrix} 0.7852 \\ 0.4966 \\ 0.3699 \end{bmatrix}, \quad F(x) = \begin{bmatrix} 0.00003 \\ 0.00006 \\ 0.00003 \end{bmatrix}.$$

Ako iteracije računamo po formulama (4.20), dobijamo

$$x^{(1)} = \begin{bmatrix} 0.87500 \\ 0.50000 \\ 0.37500 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} 0.72656 \\ 0.49688 \\ 0.37031 \end{bmatrix}, \quad x^{(3)} = \begin{bmatrix} 0.81526 \\ 0.49663 \\ 0.36995 \end{bmatrix}, \quad \dots$$

Sada je

$$F(x^{(3)}) = \begin{bmatrix} 0.04815 \\ 0.09614 \\ 0.14429 \end{bmatrix},$$

što pokazuje da je konvergencija druge metode dosta sporija.

Glava 5

Metoda najmanjih kvadrata

Kod primene interpolacije greška se menja od tačke do tačke. U čvorovima interpolacije nema nikakve greške, ali u ostalim tačkama greške mogu biti velike. Zato se postavlja zadatak nalaženja neke funkcije φ (koja ne mora obavezno biti polinom) koja će "dobro" aproksimirati datu funkciju f na celom intervalu $[a, b]$. Funkcija f je zadata najčešće na diskretnom skupu podataka, tj. tabelarno. Na primer, Taylorov polinom na određenom intervalu može dobro da aproksimira datu funkciju f. Izložićemo još jednu takvu metodu ovde.

Neka je na osnovu eksperimenta dobijena tablica vrednosti funkcije f na diskretnom skupu podataka,

$$\{(x_k, f_k)\} \quad (k = 0, 1, \ldots, m),$$

tj.

x	x_0	x_1	$\ldots$	x_m
$y = f(x)$	f_0	f_1	$\ldots$	f_m

Treba naći funkciju $y = \varphi(x)$ koja će najbolje odgovarati vrednostima zadatim tabelarno i ostalim vrednostima koje nisu izmerene, ali naslućujemo u kojim okvirima će se kretati. Oblik zavisnosti vrednosti y od x, tj. oblik

funkcije φ, može se ustanoviti ili teorijskim razmatranjem posmatranog problema ili na osnovu rasporeda tačaka čije su koordinate merene. Na primer, ako su eksperimentom dobijene tačke raspoređene oko neke prave linije, onda ćemo funkciju φ tražiti kao neku linearnu funkciju, tj.

$$\varphi(x) = a_0 + a_1 x,$$

gde za sada proizvoljne parametre treba odrediti tako da su sve date tačke što bliže posmatranoj pravoj φ.

5.1 Preodređeni i normalni sistem jednačina

Razmotrimo opšti slučaj. Pretpostavimo da aproksimacionu funkciju φ tražimo u obliku

$$y = \varphi(x; a_0, a_1, \ldots, a_n),$$

gde je x nezavisno promenljiva, a $a_0, a_1, \ldots, a_n$ parametri koje treba odrediti.

Određivanje parametara a_i $(i = 0, 1, \ldots, n)$ je sa stanovišta teorije aproksimacija mogućno samo ako je $m \geq n$. Kada je $m = n$ imamo slučaj interpolacije, koji je već analiziran. Naravno da je od interesa imati što više podataka u prethodnoj tabeli, tako dobijamo bolju aproksimaciju funkcije f. Dakle, razmotrimo šta se dešava u slučaju kada je $m > n$.

Ako bismo postupili kao u slučaju interpolacije, dobili bismo tzv. *preodređeni sistem* jednačina,

$$\varphi(x_j; a_0, a_1, \ldots, a_n) = f_j \ (\equiv y_j) \quad (j = 0, 1, \ldots, m). \tag{5.1}$$

Označimo sa

$$\varepsilon_j = f_j - \varphi(x_j; a_0, a_1, \ldots, a_n) \quad (j = 0, 1, \ldots, m)$$

odstupanja izmerenih vrednosti od vrednosti funkcije φ u tačkama x_j.

Sada se parametri $a_0, a_1, \ldots, a_n$ određuju tako da zbir kvadrata odstupanja

$$\varepsilon_0^2 + \varepsilon_1^2 + \ldots + \varepsilon_m^2$$

bude minimalan. Dakle, imamo problem minimizacije funkcije F,

$$F(a_0, a_1, \ldots, a_n) = \sum_{j=0}^{m} [f_j - \varphi(x_j; a_0, a_1, \ldots, a_n)]^2 ,$$

ili, u opštijem slučaju,

$$F(a_0, a_1, \ldots, a_n) = \sum_{j=0}^{m} w_j \, [f_j - \varphi(x_j; a_0, a_1, \ldots, a_n)]^2 , \tag{5.2}$$

gde su uključene takozvane težine $w_j = w(x_j)$ $(j = 0, 1, \ldots, m)$[1], po parametrima $a_0, a_1, \ldots, a_n$.

Ako je, pak, u pitanju funkcionalna veza koja zavisi od više nezavisno promenljivih, na primer,

$$z = \varphi(x, y; a_0, a_1, \ldots, a_n),$$

za određivanje aproksimacionih parametara minimizira se veličina

$$F(a_0, a_1, \ldots, a_n) = \sum_{j=0}^{m} w_j \, [z_j - \varphi(x_j, y_j; a_0, a_1, \ldots, a_n)]^2 .$$

Ako je φ linearna aproksimaciona funkcija (po parametrima $a_0, a_1, \ldots, a_n$), tj. oblika

$$\varphi(x) = \sum_{i=0}^{n} a_i \, \varphi_i(x) \quad (n < m),$$

[1]Vrednostima funkcije $y_j = f_j$ sa većom tačnošću dodeljuju se veće težine w_j. Ovo je posebno važno kod aproksimacije eksperimentalnih podataka koji su prilikom merenja dobijeni sa različitom tačnošću. Na primer, ako su merenja izvršena sa različitim disperzijama čiji je odnos poznat, to se težine w_j biraju obrnuto proporcionalno disperzijama, tj. tako da je

$$w_0 : w_1 : \ \cdots \ : w_m = \frac{1}{\sigma_0^2} : \frac{1}{\sigma_1^2} : \ \cdots \ : \frac{1}{\sigma_m^2} .$$

Ako su, pak, merenja izvedena sa istom tačnošću, ali je pri svakoj vrednosti argumenta x_j izvedena serija od m_j merenja, i za $y_j = f_j$ uzeta aritmetička sredina dobijenih rezultata u seriji, to se za težine uzima broj merenja u seriji, tj. $w_j = m_j$ $(j = 0, 1, \ldots, m)$. Najčešće se, međutim, uzima da su sve težine jednake međusobno.

minimizacioni problem za (5.2) se svodi na određivanje minimuma funkcije F,

$$F(a_0, a_1, \ldots, a_n) = \sum_{j=0}^{m} w_j \, \delta_n(x_j)^2 = \overrightarrow{v}^T \, W \, \overrightarrow{v} \tag{5.3}$$

po parametrima $a_0, a_1, \ldots, a_n$, gde smo uveli

$$\delta_n(x) = f(x) - \sum_{i=0}^{n} a_i \, \varphi_i(x) \, ,$$

$$\overrightarrow{f} = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_m \end{bmatrix}, \quad \overrightarrow{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}, \quad X = \begin{bmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \ldots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_n(x_1) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_0(x_m) & \varphi_1(x_m) & \ldots & \varphi_n(x_m) \end{bmatrix},$$

$$W = \operatorname{diag}(w_0, w_1, \ldots, w_m), \quad \overrightarrow{v} = \overrightarrow{f} - X\overrightarrow{a}.$$

Odredimo minimum funkcije F date pomoću (5.3). Imamo

$$\frac{\partial F}{\partial a_i} = 2 \sum_{j=0}^{m} w_j \delta_n(x_j) \frac{\partial \delta_n(x_j)}{\partial a_i} = 0 \quad (i = 0, 1, \ldots, n),$$

odakle dobijamo takozvani *normalni sistem* jednačina

$$\sum_{j=0}^{m} w_j \delta_n(x_j) \varphi_i(x_j) = 0 \quad (i = 0, 1, \ldots, n), \tag{5.4}$$

za određivanje parametara a_i. Shodno gornjim oznakama, sistem jednačina (5.4) se može predstaviti u matričnom obliku

$$X^T W \overrightarrow{v} = \overrightarrow{0},$$

tj.

$$X^T W X \overrightarrow{a} = X^T W \overrightarrow{f}. \tag{5.5}$$

Vektor traženih koeficijenata $\overrightarrow{a}$ određujemo iz (5.5), tako da je

$$\overrightarrow{a} = \left(X^T W X\right)^{-1} X^T W \overrightarrow{f}.$$

U ovom slučaju preodređeni sistem jednačina (5.1) (podsetimo da je aproksimaciona funkcija φ linearna) ima matričnu reprezentaciju

$$X\overrightarrow{a} = \overrightarrow{f}.$$

Primetimo da se normalni sistem jednačina (5.5) dobija iz preodređenog sistema jednačina jednostavnim množenjem matricom $X^T W$ sa leve strane.

U najčešćem slučaju sve težine su jednake jedinici, tj. W je jedinična matrica. Tada je

$$\overrightarrow{a} = \left(X^T X\right)^{-1} X^T \overrightarrow{f}.$$

U slučaju kada se (bazisne) funkcije φ_i izaberu sa $\varphi_i(x) = x^i$ ($i = 0, 1, \ldots, n$), imamo

$$X = \begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_m & x_m^2 & \ldots & x_m^n \end{bmatrix}.$$

Posebno je interesantan slučaj poslednje metode najmanjih kvadrata kada je $n = 1$, tj. kada je aproksimaciona funkcija oblika $\varphi(x) = a_0 + a_1 x$. Tada sistem jednačina (5.5) postaje

$$\begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix},$$

gde su

$$s_{11} = \sum_{j=0}^{m} w_j, \quad s_{12} = s_{21} = \sum_{j=0}^{m} w_j x_j, \quad s_{22} = \sum_{j=0}^{m} w_j x_j^2,$$

$$b_0 = \sum_{j=0}^{m} w_j f_j, \quad b_1 = \sum_{j=0}^{m} w_j x_j f_j.$$

Traženi aproksimacioni parametri su tada

$$a_0 = \frac{1}{D}(s_{22} b_0 - s_{12} b_1), \quad a_1 = \frac{1}{D}(s_{11} b_1 - s_{21} b_0),$$

gde je $D = s_{11}s_{22} - s_{12}^2$.

PRIMER 5.1. Metodom najmanjih kvadrata odredićemo pravu $y = a_0 + a_1x$ na osnovu eksperimentalnih podataka prikazanih tabelom

x_j	1	2	3	4
y_j	3.5	3.0	2.8	2.3

Odredimo vrednosti parametara a_0 i a_1 za koje funkcija

$$\begin{aligned} F(a_0, a_1) &= \sum_{i=1}^{4}[y_i - (a_0 + a_1x)]^2 = (3.5 - a_0 - a_1)^2 \\ &+ (3 - a_0 - 2a_1)^2 + (2.8 - a_0 - 3a_1)^2 + (2.3 - a_0 - 4a_1)^2 \end{aligned}$$

postiže minimum. Rešavanjem sistema jednačina

$$\frac{\partial F}{\partial a_0} = 0, \quad \frac{\partial F}{\partial a_1} = 0,$$

odnosno

$$\begin{cases} 27.1 - 10a_0 - 30a_1 = 0, \\ 11.6 - 4a_0 - 10a_1 = 0, \end{cases}$$

dobijamo

$$a_0 = 3.85, \quad a_1 = -0.38.$$

Konačno je

$$y = \varphi(x) = 3.85 - 0.38x.$$

PRIMER 5.2. Metodom najmanjih kvadrata aproksimiraćemo sledeći skup podataka

x_j	-2	-1	0	1	2
f_j	-0.1	0.1	0.4	0.9	1.6

pomoću aproksimacione funkcije $\varphi(x) = a_0 + a_1x + a_2x^2$.

Ovde imamo

$$X = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}, \qquad \overrightarrow{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}, \quad \overrightarrow{f} = \begin{bmatrix} -0.1 \\ 0.1 \\ 0.4 \\ 0.9 \\ 1.6 \end{bmatrix}.$$

Kako je

$$X^T X = \begin{bmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{bmatrix} \quad \text{i} \quad X^T \overrightarrow{f} = \begin{bmatrix} 2.9 \\ 4.2 \\ 7.0 \end{bmatrix},$$

iz sistema jednačina $(X^T X)\overrightarrow{a} = X^T \overrightarrow{f}$ nalazimo

$$a_0 = 0.4086\,, \qquad a_1 = 0.42\,, \qquad a_2 = 0.0857\,,$$

odnosno

$$y = \varphi(x) = 0.4086 + 0.42x + 0.0857x^2.$$

Međutim, ako je φ nelinearna aproksimaciona funkcija, tada je odgovarajući normalni sistem jednačina

$$\frac{\partial F}{\partial a_i} = 0 \quad (i = 0, 1, \ldots, n)$$

nelinearan. Za njegovo se rešavanje tada zahteva primena neke numeričke metode, recimo metode Newton-Kantoroviča, čime se postupak određivanja aproksimacionih parametara dosta komplikuje. U cilju lakšeg i bržeg određivanja parametara postoje neki uprošćeni metodi transformacije ovakvih problema na linearne aproksimacione probleme. Osnovna ideja je u uvođenju izvesnih smena

$$X = g(x), \quad Y = h(y),$$

pomoću kojih se nelinearni problem svodi na linearni.

Na primer, neka je $y = \varphi(x; a_0, a_1) = a_0 e^{a_1 x}$. Tada, logaritmovanjem i uvođenjem smena

$$X = x, \quad Y = \log y, \quad b_0 = \log a_0, \quad b_1 = a_1,$$

problem se svodi na linearan, jer je sada $Y = b_0 + b_1 X$. Dakle, minimizacijom veličine ($w_j = 1$)

$$G = G(b_0, b_1) = \sum_{j=0}^{m} (Y_j - b_0 - b_1 X_j)^2,$$

gde su $X_j = x_j$ i $\log Y_j = \log y_j$ $(j = 0, 1, \ldots, m)$, određujemo parametre b_0 i b_1, a zatim

$$a_0 = e^{b_0} \quad \text{i} \quad a_1 = b_1.$$

Ovakav postupak, međutim, ne dovodi do istih parametara koji se dobijaju minimizacijom funkcije

$$F = F(a_0, a_1) = \sum_{j=0}^{m} \left(y_j - a_0 e^{a_1 x_j}\right)^2 .$$

Dobijene vrednosti mogu znatno da odstupaju. Ova činjenica se javlja zato što se rešava problem različit od postavljenog, imajući u vidu nelinearnu transformaciju koju smo izvršili ($Y = \log y$). Ali, za mnoge potrebe ovako dobijeni parametri su zadovoljavajući.

Navedimo još neke tipične funkcionalne zavisnosti gde je moguća jednostavna transformacija promenljivih:

$1^0 \quad y = a_0 x^{a_1},\ X = \log x,\ Y = \log y,\ b_0 = \log a_0,\ b_1 = a_1;$

$2^0 \quad y = \dfrac{1}{a_0 + a_1 x},\ X = x,\ Y = \dfrac{1}{y},\ b_0 = a_0,\ b_1 = a_1;$

$3^0 \quad y = \dfrac{x}{a_0 + a_1 x},\ X = \dfrac{1}{x},\ Y = \dfrac{1}{y},\ b_0 = a_0,\ b_1 = a_1;$

$4^0 \quad y = \dfrac{1}{a_0 + a_1 e^{-x}},\ X = e^{-x},\ Y = \dfrac{1}{y},\ b_0 = a_0,\ b_1 = a_1.$

PRIMER 5.3. Metodom najmanjih kvadrata odredićemo aproksimacionu funkciju oblika

$$\varphi(x) = \log(a + e^{b+x})$$

za funkciju $f(x)$ koja je zadata skupom podataka

x	2.6	2.8	3.0	3.5
$f(x)$	$\log 2.22$	$\log 2.44$	$\log 2.67$	$\log 3.21$

.

Iz $\varphi(x) = \log\left(a + e^{b+x}\right)$ imamo da je $e^{\varphi(x)} = a + e^b \cdot e^x$, tj.

$$\phi(t) = A + Bt, \quad \text{gde su} \quad A = a,\ B = e^b,\ t = e^x.$$

Aproksimacioni uslov $\varphi(x_k) = f(x_k)$, tj. $e^{\varphi(x_k)} = e^{f(x_k)}$ daje

$$X = \begin{bmatrix} 1 & e^{2.6} \\ 1 & e^{2.8} \\ 1 & e^{3.0} \\ 1 & e^{3.5} \end{bmatrix}, \quad \overrightarrow{a} = \begin{bmatrix} A \\ B \end{bmatrix}, \quad \overrightarrow{f} = \begin{bmatrix} 2.22 \\ 2.44 \\ 2.67 \\ 3.21 \end{bmatrix}.$$

Sistem $X^T X \overrightarrow{a} = X^T \overrightarrow{f}$ tada postaje

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ e^{2.6} & e^{2.8} & e^{3.0} & e^{3.5} \end{bmatrix} \cdot \begin{bmatrix} 1 & e^{2.6} \\ 1 & e^{2.8} \\ 1 & e^{3.0} \\ 1 & e^{3.5} \end{bmatrix} \cdot \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ e^{2.6} & e^{2.8} & e^{3.0} & e^{3.5} \end{bmatrix} \times \begin{bmatrix} 2.22 \\ 2.44 \\ 2.67 \\ 3.21 \end{bmatrix}.$$

S obzirom na vrednosti

$$e^{2.6} \cong 13.464, \quad e^{2.8} \cong 16.445, \quad e^{3.0} \cong 20.086, \quad e^{3.5} \cong 33.115,$$

prethodni sistem se transformiše u sistem jednačina

$$\begin{bmatrix} 4 & 83.11 \\ 83.11 & 1951.768 \end{bmatrix} \cdot \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 10.54 \\ 229.945 \end{bmatrix},$$

odakle dobijamo

$$A = 1.596, \quad B = 0.05,$$

tj.

$$a = A = 1.596, \quad b = \log B = -2.996.$$

Aproksimaciona funkcija je

$$\varphi(x) = \log\left(1.596 + e^{-2.996+x}\right).$$

Glava 6

Numerička integracija

Ovo poglavlje posvećeno je konstrukciji metoda za približno izračunavanje Riemanovih integrala

$$I(f) = \int_a^b f(x)\,dx\,.$$

Ako nije drukčije naglašeno pretpostavljaćemo da je f neprekidna na $[a, b]$.

Potreba za numeričkom integracijom je veoma česta. Postoji mnogo veoma jednostavnih određenih integrala koje ne možemo izračunati analitički, jer ne možemo odrediti primitivnu funkciju integranda f. Primeri takvih integranada, koji se često javljaju u fizici i tehničkim nauka, su:

$$\frac{1}{\ln x}, \quad \frac{\sin x}{x}, \quad e^{-x^2}, \quad \frac{1}{(1-k^2\sin^2 x)^{1/2}}, \quad (1-k^2\sin^2 x)^{1/2} \quad \ldots$$

Čak i kad možemo analitički odrediti $I(f)$, često njegovu tačnu numeričku vrednost ne možemo predstaviti kao broj, tj. možemo je odrediti samo približno. Na primer,

$$\int_{-1}^{1} \frac{2x^2+2x+13}{(x-2)(x^2+1)^2}\,dx = \ln\frac{1}{3} - 8\arctan 1 - 2\,.$$

Pored toga, primitivne funkcije mogu biti jako komplikovane, a njihovo određivanje može biti veoma težak matematički zadatak. U praksi je čest slučaj da je f dobijena eksperimentalno, tj. sve što znamo o njoj su njene vrednosti na nekom konačnom skupu tačaka.

Definicija 6.0.1 *Kvadraturna formula*[1] *je svako numeričko pravilo za aproksimaciju određenog integrala, koje koristi samo podatke o integrandu na diskretnom skupu tačaka iz intervala integracije.*

Ove tačke su čvorovi kvadraturne formule. Za njih ćemo pretpostavljati

$$a \leq x_1 < x_2 < \ldots < x_n \leq b\,.$$

Podaci o integrandu su vrednosti funkcije f i njenih izvoda. S obzirom na linearnost određenog integrala, prirodno je zahtevati da i kvadraturna formula bude linearna u odnosu na f. Ako od podataka o integrandu koristimo samo vrednosti funkcije, onda kvadraturne formule imaju sledeći oblik

$$Q_n(f) = \sum_{i=1}^{n} A_i f(x_i)\,. \tag{6.1}$$

Brojevi A_i su (težinski) koeficijenti kvadraturne formule. Razlika

$$R_n(f) = I(f) - Q_n(f)$$

je ostatak (greška) kvadrature.

Postoji više različitih metoda za konstrukciju kvadraturnih formula, koje ponekad dovedu do iste kvadrature. Neke od njih su:

- **Geometrijska konstrukcija**. Cilj je površinu ispod krive aproksimirati zbirom površina jednostavnijih mnogouglova. Najčešće se to radi tako što se prvo izaberu čvorovi, a onda, nad intervalima između dva susedna čvora, konstruišu pravougaonici ili trapezi, tako da zbir njihovih površina bude što bliži vrednosti $I(f)$. Na primer, u prvom slučaju, ako stavimo $x_1 = a$, $x_n < b$, dobijamo formulu

$$Q_n(f) = (x_2 - x_1)f(x_1) + (x_3 - x_2)f(x_2) + \ldots + (x_n - x_{n-1})f(x_{n-1})\,, \tag{6.2}$$

[1] Često se koristi i termin mehanička kvadratura da se naglasi da aproksimacija nije dobijena analitički.

dok u drugom slučaju, za $x_1 = a$, $x_n = b$, dobijamo

$$\begin{aligned} Q_n(f) &= \frac{x_2 - x_1}{2}[f(x_1) + f(x_2)] + \frac{x_3 - x_2}{2}[f(x_2) + f(x_3)] + \dots \\ &\quad + \frac{x_n - x_{n-1}}{2}[f(x_{n-1}) + f(x_n)]\,. \end{aligned} \tag{6.3}$$

- **Metoda neodređenih koeficijenata**. Kvadraturnu formulu određujemo tako da ona bude tačna za svako f iz nekog linearno nezavisnog skupa funkcija $\{\varphi_0, \varphi_2, \dots, \varphi_m\}$. Najčešće se bira $\varphi_i(x) = x^i$, tj. kvadratura treba biti tačna za svako $f \in \mathcal{P}_m$.[2] Ako je još $R_n(x^{m+1}) \neq 0$, tada kažemo da kvadraturna formula ima algebarski stepen tačnosti m. Lako se može postići da je $m = n - 1$. Dovoljno je proizvoljno izabrati n različitih čvorova, a zatim rešiti regularni kvadratni sistem linearnih jednačina

$$x_1^i A_1 + x_2^i A_2 + \dots + x_n^i A_n = \int_a^b x^i\,dx\,, \quad i = 0, \dots, n-1\,. \tag{6.4}$$

  Za očekivati je da možemo postići i veći algebarski stepen tačnosti, ako dopustimo i proizvoljan izbor čvorova. Prethodni sistem tada više nije linearan, pa ne možemo ništa reći o njegovim rešenjima. Pokazaćemo kasnije da se može i na drugi način odgovoriti na ovo pitanje.

- **Interpolacione kvadrature**. Kvadraturnu formulu možemo dobiti analitičkom integracijom neke aproksimacije od f. Najčešći izbor aproksimacije je interpolacioni polinom.

- **Kompozitna pravila**. Često se javlja potreba da se u kvadraturi povećava broj čvorova kako bi se postigla dovoljna tačnost. To se najlakše može postići tako da podelimo interval integracije na nekoliko podintervala i primenimo kvadraturu na svakom od njih.

[2] $\mathcal{P}_m$ je skup svih algebarskih polinoma stepena $\leq m$.

PRIMER 6.1. Iako kvadraturna formula

$$Q_n(f) = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

ima samo dva čvora, njen algebarski stepen tačnosti je tri, na intervalu $[-1, 1]$.

PRIMER 6.2. Izračunaćemo

$$\int_{0.1}^{0.3} f(x)\,dx,$$

ako je funkcija f zadata vrednostima u sledećoj tablici.

x	0.20	0.27	0.30
$f(x)$	0.85148	0.81903	0.80749

Za približno izračunavanje datog integrala koristićemo kvadraturnu formulu koju konstruišemo metodom neodređenih koeficijenata. Odredićemo nepoznate koeficijente A_1, A_2, A_3 iz uslova da kvadraturna formula

$$\int_{0.1}^{0.3} f(x)\,dx \approx \sum_{k=1}^{3} A_k f(x_k) \tag{6.5}$$

bude tačna za algebarske polinome stepena ne većeg od dva, tj. za $f(x)$ uzimamo $1, x, x^2$.

Iz zahteva da kvadraturna formula (6.5) bude tačna za $f(x) = 1$ dobijamo uslov

$$A_1 + A_2 + A_3 = 0.1. \tag{6.6}$$

Iz zahteva da kvadraturna formula (6.5) bude tačna za $f(x) = x$ dobijamo uslov

$$0.2A_1 + 0.27A_2 + 0.3A_3 = 0.25. \tag{6.7}$$

Konačno, iz zahteva da kvadraturna formula (6.5) bude tačna za $f(x) = x^2$ dobijamo uslov

$$0.04A_1 + 0.0729A_2 + 0.09A_3 = 0.006333. \tag{6.8}$$

Sada, rešavanjem sistema jednačina formiranog pomoću (6.6), (6.7), (6.8), dobijamo

$$A_1 = 0.02622, \quad A_2 = 0.07929, \quad A_3 = -0.00551,$$

i

$$\int_{0.2}^{0.3} f(x)\,dx = A_1 f_1 + A_2 f_2 + A_3 f_3 = 0.08281.$$

Kvadrature oblika (6.1) možemo koristiti i za približno izračunavanje integrala

$$I_w(f) = \int_a^b f(x)\,w(x)\,dx\,, \tag{6.9}$$

gde je w težinska funkcija, za koju pretpostavljamo da je pozitivna[3] i integrabilna na (a, b). Ako je interval integracije beskonačan, uvodimo dodatnu pretpostavku da svi momenti težinske funkcije

$$\mu_k := \int_a^b x^k\,w(x)\,dx\,, \quad k \in \mathbb{N}_0\,,$$

postoje i da su konačni. Drugim rečima, posmatramo opšti problem jednodimenzionalne integracije zadane funkcije f, po zadanoj neprekidnoj meri $d\lambda$ generisanoj težinskom funkcijom w na zadanom domenu.

6.1 Newton-Cotesove kvadraturne formule

U ovoj sekciji razmatramo interpolacione kvadraturne formule zatvorenog tipa (prvi čvor je levi kraj intervala integracije a, a poslednji desni kraj tog

[3]U opštem slucaju težinska funkcija može imati i vrednost nula, ali samo na skupu mere nula.

intervala b) u kojima su interpolacioni čvorovi $x_k = x_0 + kh$ $(k = 0, 1, \ldots, n)$ uzeti ekvidistantno sa korakom $h = (b-a)/n$. To su formule oblika

$$\int_a^b w(x) f(x)\, dx = \sum_{k=0}^{n} A_k f(x_k) + R_{n+1}(f)\,. \tag{6.10}$$

Indeks $n+1$ u ostatku označava da se integral približno izračunava na osnovu vrednosti podintegralne funkcije u $n+1$ tačaka. Kako je kvadraturna formula (6.10) interpolacionog tipa, koeficijenti A_k $(k = 0, 1, \ldots, n)$ se određuju pomoću

$$A_k = \frac{1}{\omega'(x_k)} \int_a^b w(x)\, \frac{\omega(x)}{x - x_k} \quad (k = 0, 1, \ldots, n), \tag{6.11}$$

gde je $\omega(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$, a ostatak pomoću

$$R_{n+1}(f) = \int_a^b w(x) r_n(f; x)\, dx = \int_a^b w(x) \frac{f^{(n+1)}(\xi)}{(n+1)!}\, \omega(x)\, dx. \tag{6.12}$$

Do formula (6.11) i (6.12) se dolazi na osnovu formule za interpolaciju funkcije f algebarskim polinomom n-tog stepena π_n (recimo Lagrangeovim), tj.

$$f(x) = \pi_n(x) + r_n(f; x),$$

njenim množenjem sa težinskom funkcijom w a zatim integracijom po intervalu $[a, b]$.

Algebarski stepen tačnosti u ovom slučaju nije manji od n, jer je $R_{n+1}(f) = 0,\ \forall f \in \mathcal{P}_n$ (pošto je $r_n(f; x) \equiv 0$ na $[a, b]$, ako $f \in \mathcal{P}_n$).

Razmotrićemo najprostiji slučaj, kada je $w(x) \equiv 1$.

Ako uvedemo smenu $x - x_0 = ph$, imamo

$$\omega(x) = (x - x_0)(x - x_1) \cdots (x - x_n) = h^{n+1} p(p-1) \cdots (p-n) \tag{6.13}$$

i

$$\begin{aligned} \omega'(x_k) = \ & (x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n) \\ & = h^n (-1)^{n-k} k!(n-k)!\ . \end{aligned} \tag{6.14}$$

Uvođenjem oznake za uopšteni stepen $x^{(s)} = x(x-1)\cdots(x-s+1)$, na osnovu (6.13) i (6.14), i (6.11) (sa $w(x) \equiv 1$), dobijamo

$$A_k = \int_0^n \frac{(-1)^{n-k} p^{(n+1)} h}{(p-k)k!(n-k)!}\, dp \quad (k = 0, 1, \ldots, n),$$

tj.

$$A_k = (b-a)H_k \quad (k = 0, 1, \ldots, n),$$

gde smo stavili

$$H_k \equiv H_k(n) = \frac{(-1)^{n-k}}{n!\, n}\binom{n}{k}\int_0^n \frac{p^{(n+1)}}{p-k}\, dp \quad (k = 0, 1, \ldots, n). \tag{6.15}$$

Koeficijenti H_k su poznati kao Newton–Cotesovi koeficijenti, a odgovarajuće formule tipa (6.10),

$$\int_{x_0=a}^{x_n=b} f(x)\, dx \simeq (b-a)\sum_{k=0}^{n} H_k f\left(a + k\frac{b-a}{n}\right) \quad (n \in \mathbb{N}) \tag{6.16}$$

kao Newton–Cotesove kvadraturne formule.

Pošto je kvadraturna formula (6.16) tačna za $f(x) = 1$, to važi

$$\sum_{k=0}^{n} H_k = 1.$$

Za koeficijente H_k važe i jednakosti

$$H_k = H_{n-k} \quad \left(k = 0, 1, \ldots, \left[\frac{n}{2}\right]\right). \tag{6.17}$$

Dokažimo poslednje tvrđenje. Umesto k stavimo $n-k$ u (6.15). Tada dobijamo

$$H_{n-k} = \frac{(-1)^k}{n!\, n}\binom{n}{n-k}\int_0^n \frac{p^{(n+1)}}{p-n+k}\, dp\,. \tag{6.18}$$

Smenom $p := n - p$ ($\Rightarrow dp := -dp$) u integralu koji se pojavljuje na desnoj strani u (6.18) dobijamo

$$H_{n-k} = \frac{(-1)^k}{n!\, n}\binom{n}{n-k}\int_0^n \frac{(n-p)^{(n+1)}}{-p+k}\, dp\,.$$

Kako je

$$\begin{aligned}(n-p)^{(n+1)} &= (n-p)(n-p-1)\cdots(n-p-n)\\ &= (-p)(-p+1)\cdots(-p+n)\\ &= (-1)^{n+1}\,p^{(n+1)}\end{aligned}$$

i $(-1)^{n+k}=(-1)^{n-k}$ i $\binom{n}{n-k}=\binom{n}{k}$, na osnovu prethodnog zaključujemo da važi $H_k = H_{n-k}$ $\left(k=0,1,\ldots,\left[\frac{n}{2}\right]\right)$.

Razmotrimo sada dva najpoznatija podslučaja Newton-Cotesovih kvadraturnih formula.

a) Neka je $n=1$. Tada je, korišćenjem (6.15), $H_0 = H_1 = \dfrac{1}{2}$, pa je odgovarajuća Newton–Cotesova kvadraturna formula

$$\int_a^b f(x)\,dx = \frac{b-a}{2}(f(a)+f(b)) + R_2(f). \tag{6.19}$$

Ova formula je poznata kao trapezna formula ili trapezno pravilo.

Teorema 6.1.1 *Ako je $f \in C^2[a,b]$, za ostatak $R_2(f)$ u kvadraturnoj formuli* (6.19) *važi*

$$R_2(f) = -\frac{(b-a)^3}{12}f''(\xi_1) \quad (a<\xi_1<b). \tag{6.20}$$

Dokaz. Na osnovu (6.12), a za ovaj slučaj, imamo

$$R_2(f) = \frac{1}{2}\int_a^b (x-a)(x-b)f''(\xi_x)\,dx.$$

Pošto je funkcija f'' neprekidna na intervalu $[a,b]$ i $(x-a)(x-b)$ stalnog znaka na $[a,b]$, to korišćenjem poznate teoreme o srednjoj vrednosti određenog integrala, sledi da postoji $\xi_1 \in (a,b)$ tako da je

$$R_2(f) = \frac{f''(\xi_1)}{2}\int_a^b (x-a)(x-b)\,dx = -\frac{(b-a)^3}{12}f''(\xi_1),$$

što je i trebalo dokazati. □

Trapezna formula može imati zadovoljavajuću tačnost, ukoliko je h dovoljno malo. To možemo obezbediti podelom intervala integracije na niz podintervala. Ako interval $[a, b]$ podelimo na n podintervala oblika $[x_{k-1}, x_k]$, $k = 1, \ldots, n$, tako da je

$$a = x_1 < x_2 < \ldots < x_n = b\,,$$

i saberemo sve trapezne formule nad njima, dobijamo kompozitnu trapeznu formulu. Najjednostavnije je uzeti da su tačke x_k ekvidistantne. Tako dobijamo

$$\int_a^b f(x)\,dx = h\left(\frac{f(x_0)}{2} + f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + \frac{f(x_n)}{2}\right) + R_n^T(f)\,,$$

pri čemu grešku možemo zapisati kao zbir grešaka trapeznih formula na podintervalima

$$R_n^T(f) = \sum_{i=1}^{n} -f''(\xi_i)\,\frac{h^3}{12}\,.$$

Greška napisana na ovaj način nije korisna, pa ćemo je napisati drugačije

$$R_n^T(f) = -\frac{h^3\,n}{12}\left(\frac{1}{n}\sum_{i=1}^{n} f''(\xi_i)\right)\,.$$

Broj u zagradi je aritmetička sredina vrednosti drugog izvoda u tačkama ξ_i. Taj broj se sigurno nalazi između minimuma i maksimuma drugog izvoda na $[a, b]$. Budući da je f'' neprekidna funkcija na $[a, b]$, broj u zagradi je jednak $f''(\xi)$, za neko $\xi \in [a, b]$, pa važi

$$R_n^T(f) = -\frac{h^3\,n}{12} f''(\xi) = -\frac{(b-a)\,h^2}{12} f''(\xi)\,.$$

b) Neka je sada $n = 2$. Tada na osnovu (6.15) i (6.16) dobijamo kvadraturnu formulu

$$\int_a^b f(x)\,dx = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) + R_3(f). \qquad (6.21)$$

Ova formula je poznata kao Simpsonova formula ili pravilo. Primetimo da je ova formula, kao interpolaciona, korišćenjem (6.12), tačna za algebarske polinome stepena 2 ili manjeg. Štaviše, Simpsonova kvadraturna formula je tačna i za algebarski polinom $g(x) = \left(x - \frac{a+b}{2}\right)^3$, što se lako da zaključiti na osnovu osobine neparnosti funkcije g na intervalu $[a, b]$ u odnosu na sredinu intervala $(a+b)/2$, i na osnovu osobine Cotesovih koeficijenata za ovaj slučaj $H_0 = H_2$. Dakle, Simpsonova kvadraturna formula (6.21) tačna je bar za algebarske polinome trećeg stepena.

Teorema 6.1.2 *Ako je $f \in C^4[a, b]$, za ostatak $R_3(f)$ u kvadraturnoj formuli* (6.21) *važi*

$$R_3(f) = -\frac{(b-a)^5}{2880} f^{(4)}(\xi_2) \quad (a < \xi_2 < b). \tag{6.22}$$

Dokaz. Posmatrajmo Simpsonovu kvadraturnu formulu (6.21) i uvedimo oznake

$$I(f) = \int_a^b f(x)\,dx, \quad S(f) = \frac{x_2 - x_0}{6}\left(f(x_0) + 4f\left(\frac{x_0 + x_2}{2}\right) + f(x_2)\right),$$

gde je $x_0 = a,\ x_1 = \frac{a+b}{2},\ x_2 = b.$

Već smo konstatovali da je Simpsonova kvadraturna formula (6.21) tačna za algebarske polinome stepena ≤ 3.

Neka je P_3 polinom trećeg stepena za koji je

$$P_3(x_0) = f(x_0),\ P_3(x_1) = f(x_1),\ P_3(x_2) = f(x_2),\ P_3'(x_1) = f'(x_1).$$

Zaključujemo da je P_3 Hermiteov interpolacioni polinom za koji važi

$$I(P_3) = S(P_3).$$

Takođe, na osnovu formule za ostatak u Hermiteovoj interpolaciji, važi

$$f(x) - P_3(x) = \frac{f^{(4)}(\xi_x)}{4!}(x - x_0)(x - x_1)^2(x - x_2),$$

za neko $\xi_x \in (a,b)$. Sada je

$$\begin{aligned} I(f) - S(f) &= I(f) - I(P_3) = \int_a^b (f(x) - P_3(x))\,dx \\ &= \int_a^b \frac{f^{(4)}(\xi_x)}{4!}(x-x_0)(x-x_1)^2(x-x_2)\,dx \\ &= \frac{f^{(4)}(\xi_2)}{4!}\int_a^b (x-x_0)(x-x_1)^2(x-x_2)\,dx, \end{aligned}$$

za neko $\xi_2 \in (a,b)$, gde smo koristili teoremu o srednjoj vrednosti određenog integrala.

Ostaje nam da izračunamo integral

$$\int_a^b (x-x_0)(x-x_1)^2(x-x_2)\,dx.$$

Razmotrimo grešku Simpsonovog pravila (6.21) za funkciju

$$h(x) = (x-x_1)^4.$$

Imamo

$$\begin{aligned} I(h) &= \int_{x_0}^{x_2} (x-x_1)^4\,dx = \frac{2}{5}\left(\frac{b-a}{2}\right)^5, \\ S(h) &= \frac{2}{3}\left(\frac{b-a}{2}\right)^5. \end{aligned}$$

Kako je $h^{(4)}(x) = 4!$, imamo da je

$$\begin{aligned} \int_a^b (x-x_0)(x-x_1)^2(x-x_2)\,dx &= I(h) - S(h) \\ &= \frac{2}{5}\left(\frac{b-a}{2}\right)^5 - \frac{2}{3}\left(\frac{b-a}{2}\right)^5 \\ &= -\frac{4}{15}\left(\frac{b-a}{2}\right)^5. \end{aligned}$$

Zamenom poslednje jednakosti u izrazu za grešku $R_3(f) = I(f) - S(f)$, dobijamo

$$\begin{aligned} R_3(f) &= -\frac{f^{(4)}(\xi_2)}{4!}\cdot\frac{4}{15}\left(\frac{b-a}{2}\right)^5 = -\frac{f^{(4)}(\xi_2)(b-a)^5}{120\cdot 4!} \\ &= -\frac{f^{(4)}(\xi_2)}{2880}(b-a)^5, \end{aligned}$$

što je i trebalo dokazati. □

Slično kao kompozitnu trapeznu, izvodimo i kompozitnu Simpsonovu formulu, s malom razlikom što svaki od podintervala moramo prepoloviti još jednom tačkom da bismo nad njim mogli primeniti Simpsonovu formulu. Ako se ograničimo na ekvidistantni slučaj i zahtevamo podelu intervala na paran broj podintervala dobijamo

$$\begin{aligned}\int_a^b f(x)\,dx &= \frac{h}{3}\left[f(x_0)+4f(x_1)+2f(x_2)+4f(x_3)+2f(x_4)+\ldots\right.\\ &\quad \left.+4f(x_{n-1})+f(x_n)\right]+R_n^S(f)\,,\end{aligned}$$

gde je

$$R_n^S(f)=\sum_{i=1}^{n/2}-\frac{h^5}{90}\,f^{(4)}(\xi_i)=-\frac{h^5\,n}{2\cdot 90}\left(\frac{2}{n}\sum_{i=1}^{n/2}f^{(4)}(\xi_i)\right).$$

Izraz u zagradi možemo zameniti sa $f^{(4)}(\xi)$, $\xi\in[a,b]$, pa dobijamo

$$R_n^S(f)=-\frac{h^5 n}{180}\,f^{(4)}(\xi)=-\frac{(b-a)h^4}{180}f^{(4)}(\xi)\,.$$

Primer 6.3. Razmotrimo na koliko delova treba podeliti interval integracije, da bi se integral

$$\int_{0.7}^{1.3}\frac{dx}{\sqrt{2x^2+0.3}}$$

izračunao kompozitnom (uopštenom) trapeznom formulom sa tačnošću $\varepsilon = 0.5\cdot 10^{-3}$.

Za dostizanje tražene tačnosti dovoljno je odrediti $n\in\mathbb{N}$ tako da je apsolutna vrednost ostatka u složenoj trapeznoj formuli

$$\frac{(b-a)^3}{12n^2}\,|f''(\xi)|<0.5\cdot 10^{-3}.$$

Ovde je $a=0.7, b=1.3, f(x)=1/\sqrt{2x^2+0.3}$. Nalazimo

$$f'(x)=\frac{-2x}{\sqrt{(2x^2+0.3)^3}},\quad f''(x)=\frac{8x^2-0.6}{\sqrt{(2x^2+0.3)^5}},$$

$$\max_{0.7\le x\le 1.3} |f''(x)| < \frac{8\cdot 1.3^2 - 0.6}{\sqrt{(2\cdot 0.7^2+0.3)^5}} \approx 6.93 < 7.$$

Dakle, apsolutna vrednost ostatka u složenoj trapeznoj formuli je manja od ε, ako je

$$\frac{(1.3-0.7)^3}{12n^2}\cdot 7 < 0.0005,$$

što će biti ispunjeno za $n^2 \ge 252$, tj. za $n \ge 16$. Uzećemo da je $n = 20$. Približnu vrednost integrala ćemo izračunati po složenoj trapeznoj formuli

$$T_{20} = h\left(\frac{1}{2}f_0 + f_1 + \dots f_{19} + \frac{1}{2}f_{20}\right),$$

gde je $h = (b-a)/n = 0.6/20 = 0.003$, $y_i = f(x_i) = 1/\sqrt{2x_i^2+0.3}$, $x_i = 0.7 + ih$ $(i = 0, 1, \dots, 20)$.

Zamenom numeričkih vrednosti dobija se da je $T_n = 0.40418 \approx 0.404$.

Primer 6.4. Razmotrimo na koliko delova treba podeliti interval integracije da bi se integral

$$\int_0^1 \frac{\sin x}{x}\,dx,$$

izračunao složenom trapeznom formulom sa tačnošću $\varepsilon = 10^{-8}$.

Ovde je podintegralna funkcija

$$f(x) = \frac{\sin x}{x}.$$

Pošto je

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots\ ,$$

biće

$$f(x) = \frac{\sin x}{x} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \dots\ .$$

Diferenciranjem funkcije f dva puta dobija se

$$f''(x) = -\frac{1}{3} + \frac{1}{5}\frac{x^2}{2!} - \frac{1}{7}\frac{x^4}{4!} + \dots\ .$$

Pošto je red na desnoj strani znaka jednakosti Leibnitzov, važi

$$\forall x \in (0,1] : \ |f''(x)| \leq \frac{1}{3}.$$

Zbog toga je asolutna vrednost ostatka složene trapezne formule manja ili jednaka $1/(36n^2)$. Da bi se dostigla tražena tačnost dovoljno je podeliti interval integracije na n delova, pri čemu je

$$\frac{1}{36n^2} \leq 10^{-8}.$$

Iz poslednjeg uslova se dobija da je dovoljno uzeti $n \geq 1667$.

Primer 6.5. Izračunaćemo približnu vrednost integrala

$$I = \int_{1.2}^{1.6} \frac{\sin(2x - 2.1)}{x^2 + 1}\, dx,$$

deleći interval integracije na 8 jednakih delova i oceniti grešku. Koristićemo složenu Simpsonovu formulu.

Iz $2n = 8$ je $h = (b - a)/8 = (1.6 - 1.2)/8 = 0.05$. Formula za izračunavanje približne vrednosti integrala je

$$S_4 = \frac{h}{3}(y_0 + 4(y_1 + y_3 + y_5 + y_7) + 2(y_2 + y_4 + y_6) + y_8),$$

pri čemu je

$$y_i = f(x_i) = \frac{\sin(2x_i) - 2.1}{x_i^2 + 1}, \quad x_i = 1.2 + ih \quad (i = 0, 1, \ldots, 8).$$

Numerički rezultati su dati u sledećoj tabeli

i	x_i	y_0, y_8	y_1, y_3, y_5, y_7	y_2, y_4, y_6
0	1.20	0.1211		
1	1.25		0.1520	
2	1.30			0.1782
3	1.35		0.2000	
4	1.40			0.2176
5	1.45		0.2312	
6	1.50			0.2410
7	1.55		0.2473	
8	1.60	0.2503		
$\sum$		0.3714	0.8305	0.6368

Na osnovu tabele je

$$S_4 = \frac{0.05}{3}(0.3714 + 4 \cdot 0.835 + 2 \cdot 0.6368) = 0.088278.$$

Za ocenu greške R_{S_4} u odgovarajućoj složenoj Simpsonovoj formuli možemo iskoristiti konačne razlike četvrtog reda. Naime, pošto je

$$\frac{\Delta^4 y_i}{4!h^4} = [x_i, x_{i+1}, x_{i+2}, x_{i+3}, x_{i+4}; f] = \frac{f^{(4)}(\xi)}{4!},$$

za neko $\xi \in (1.2, 1.6)$, možemo smatrati da je

$$\max_{\xi \in (1.2,1.6)} \frac{|f^{(4)}(\xi)|}{4!} \approx \frac{\max|\Delta^4 y_i|}{4!h^4},$$

pa je

$$|R_{S_4}| \leq \frac{b-a}{180} \max|\Delta^4 y_i|.$$

i	x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\Delta^4 y_i$
0	1.20	0.1211				
			0.0309			
1	1.25	0.1520		−0.0047		
			0.0262		0.0003	
2	1.30	0.1782		−0.0014		−0.0001
			0.0218		0.0002	
3	1.35	0.2000		−0.0042		0.0000
			0.0176		0.0002	
4	1.40	0.2176		−0.0040		0.0000
			0.0136		0.0002	
5	1.45	0.2312		−0.0038		0.0001
			0.0098		0.0003	
6	1.50	0.2410		−0.0035		−0.0001
			0.0063		0.0002	
7	1.55	0.2473		−0.0033		
			0.0030			
8	1.60	0.2503				

Iz tabele se vidi da je

$$\max|\Delta^4 y_i| = 0.0001,$$

pa je

$$|R_{S_4}| \leq \frac{0.4 \cdot 0.0001}{180} \approx 0.0000003.$$

Izračunavanje je izvršeno sa četiri značajne cifre, te zbog toga greška metoda nema uticaja na konačan rezultat. Računsku grešku možemo oceniti na sledeći način

$$\Delta(\overline{I}) = (b-a)\Delta(\overline{y}) = 0.4 \cdot 0.0001 < 0.00005.$$

Dakle, $I \approx 0.0833$, i sve cifre su sigurne u užem smislu.

6.2 Rombergova integracija

Posmatrajmo kvadraturnu formulu sa ekvidistantnim čvorovima i ostatkom reda r, tj. $R(f) = Ch^r + O(h^{r+1})$, gde je h rastojanje između čvorova. Aproksimaciju integrala $I(f)$ dobijenu primenom ove formule, označima sa $I(h)$. Uzmimo za h dve različite vrednosti h_1 i h_2 ($h_2 < h_1$), i izračunajmo $I(h_1)$ i $I(h_2)$. Sledećim postupkom možemo dobiti novu aproksimaciju integrala, koja je tačnija od pomenute dve. Zanemarivanjem veličina $O(h_1^{r+1})$ i $O(h_2^{r+1})$ i eliminacijom faktora C iz jednakosti

$$I - I(h_1) = C\, h_1^r\,, \qquad I - I(h_2) = C\, h_2^r\,,$$

dobijamo formulu

$$I \approx \frac{\left(\dfrac{h_1}{h_2}\right)^r I(h_2) - I(h_1)}{\left(\dfrac{h_1}{h_2}\right)^r - 1} = I(h_2) + \frac{I(h_2) - I(h_1)}{\left(\dfrac{h_1}{h_2}\right)^r - 1}\,,$$

koja je poznata kao Richardsonova ekstrapolacija. Najčešće se uzima $h_2 = h_1/2$, i tada dobijamo

$$I \approx \frac{2^r I(h_2) - I(h_1)}{2^r - 1}\,. \tag{6.23}$$

Za posmatranu kvadraturu biramo kompozitnu trapeznu formulu[4] kako bismo iskoristili Euler-Maclaurinovu formulu koja daje asimptotski razvoj njene greške. Naime, ako je $f \in C^{2m+2}[a,b]$, za kompozitnu trapeznu formulu, važi

$$R(f) = c_1 h^2 + c_2 h^4 + c_3 h^6 + \ldots - (b-a)\frac{B_{2m+2} h^{2m+2}}{(2m+2)\,!} f^{(2m+2)}(\xi)$$

gde su B_{2k} Bernoullijevi brojevi, $\xi \in (a,b)$ i

$$c_k = -\frac{B_{2k}}{(2k)\,!}\left[f^{(2k-1)}(b) - f^{(2k-1)}(b)\right], \quad k = 1, \ldots, m\,.$$

Ako stavimo $h = h_k = (b-a)/2^k$ i odgovarajuću trapeznu aproksimaciju $T_n(f)$ $(n = 2^k)$ označimo sa $T_k^{(0)}$, tada, primenom Richardsonove ekstrapolacije na $T_k^{(0)}$ i $T_{k+1}^{(0)}$, dobijamo

$$I \approx \frac{4\,T_{k+1}^{(0)} - T_k^{(0)}}{3}\,.$$

Označimo novu aproksimaciju sa $T_k^{(1)}$ i odredimo njenu grešku. Kako je

$$I - T_k^{(1)} = \frac{4}{3}\left(I - T_{k+1}^{(0)}\right) - \frac{1}{3}\left(I - T_k^{(0)}\right),$$

primenom Euler-Maclaurinove formule, dobijamo

$$I - T_k^{(1)} = \frac{4}{3}\left[c_1\left(\frac{h_k}{2}\right)^2 + c_2\left(\frac{h_k}{2}\right)^4 + \ldots\right] - \frac{1}{3}\left[c_1 {h_k}^2 + c_2 {h_k}^2 + \ldots\right],$$

tj.

$$I - T_k^{(1)} = -\frac{1}{4}\,c_2\,h_k^4 + O(h_k^6)\,.$$

Vidimo da je r udvostručeno, kad $h_k \to 0$.

Ako je funkcija f dovoljan broj puta diferencijabilna, navedeni postupak sugeriše konstrukciju iterativnog procesa u obliku

$$T_k^{(m)} = \frac{4^m\,T_{k+1}^{(m-1)} - T_k^{(m-1)}}{4^m - 1}, \quad m = 1, 2, \ldots\,.$$

[4] U ovom slučaju je $r = 2$.

Ovaj postupak se zove Rombergova integracija. Možemo je predstaviti T-tabelom

$$\begin{array}{cccc} T_0^{(0)} & T_0^{(1)} & T_0^{(2)} & T_0^{(3)} \\ T_1^{(0)} & T_2^{(2)} & T_1^{(2)} & \vdots \\ T_2^{(0)} & T_3^{(1)} & \vdots & \\ T_3^{(0)} & \vdots & & \\ \vdots & & & \end{array}$$

U prvoj koloni nalaze se aproksimacije integrala dobijene uopštenom trapeznom formulom za $h_k = (b-a)/2^k$. Druga kolona dobija se na osnovu prve, treća na osnovu druge, itd.

Može se pokazati da nizovi $\{T_k^{(m)}\}_{k\in\mathbb{N}_0}$ (kolone u tabeli) i $\{T_k^{(m)}\}_{m\in\mathbb{N}_0}$ (vrste) konvergiraju ka $I(f)$. Kod praktične primene Rombergove integracije, iterativni proces se najčešće prekida kada je $|T_0^{(m)} - T_1^{(m-1)}| \leq \varepsilon$. Odnosi grešaka dva uzastopna broja u koloni se moraju ponašati kao

$$\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 4 & 16 & 64 & \vdots \\ 4 & 16 & \vdots & \\ 4 & \vdots & & \\ \vdots & & & \end{array}$$

Ova ponašanja brojeva u T-tabeli vrede samo ako je funkcija dovoljno glatka.

Primer 6.6 Funkcija e^x je ima beskonačno mnogo neprekidnih izvoda, pa je dobar primer za Rombergovu integraciju. Prikazaćemo odnose grešaka u kolonama prilikom računanja $\int_0^1 e^x dx$.

$$\begin{array}{ccccc} 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\ 3.9512 & 15.6517 & 62.4639 & 249.7197 & \\ 3.9875 & 15.9913 & 63.6087 & & \\ 3.9969 & 15.9777 & & & \\ 3.9992 & & & & \end{array}$$

Zamenimo prethodni integrand funkcijom $x^{3/2}$. Njen drugi izvod nije neprekidan u 0, pa bi već u drugoj koloni (funkcija je dovoljno glatka za ocenu greške u trapeznoj formuli) trebalo doći do neuobičajenog ponašanja.

1.0000	1.0000	1.0000	1.0000	1.0000
3.7346	5.4847	5.6484	5.6566	
3.8154	5.5912	5.6559		
3.8721	5.6331			
3.9112				

6.3 Gaussove kvadraturne formule

Videli smo da u kvadraturnoj formuli Q_n uvek možemo odrediti koeficijente A_i, tako da ona ima algebarski stepen tačnosti $n-1$. Primer 6.1. pokazuje da se pogodnim izvorom čvorova može postići algebarski stepen tačnosti od $2n-1$, pa se postavlja pitanje njegove maksimalne vrednosti, naravno, u zavisnosti od n. Odgovor na to pitanje dobićemo u ovoj sekciji.

Teorema 6.3.1 *Algebarski stepen tačnosti kvadraturne formule sa n čvorova ne može biti veći od* $2n-1$.

Dokaz. Dovoljno je naći jedan polinom stepena $2n$, takav da je $I_w(f) \neq Q_n(f)$. To važi za ω_n^2, gde je $\omega_n(x) = (x-x_1)(x-x_2)\dots(x-x_n)$. Naime, očigledno je $Q_n(\omega_n) = 0$, dok je, s obzirom na osobine težinske funkcije,

$$I_w(\omega_n) = \int_a^b \omega_n^2(x)\, w(x)\, dx > 0\,.$$

□

Kvadrature sa maksimalno mogućim algebarskim stepenom tačnosti zovemo Gaussovim kvadraturnim formulama. Dokazaćemo njihovu egzistenciju i pokazati način za njihovu konstrukciju.

Teorema 6.3.2 *Da bi kvadraturna formula sa n čvorova i minimalnim algebarskim stepenom tačnosti* $n-1$ *imala algebarski stepen tačnosti* $2n-1$,

potrebno je i dovoljno da su njeni čvorovi nule polinoma W_n, gde je $\{W_k\}_{k\in\mathbb{N}_0}$ niz ortogonalnih polinoma u odnosu na težinsku funkciju $w(x)$.

Dokaz. Neka su čvorovi x_k, $k = 1, \ldots, n$, nule polinoma W_n i neka je p proizvoljan polinom iz $\mathcal{P}_{2n-1}$. Tada se p može predstaviti u obliku

$$p(x) = u_{n-1}(x)\, W_n(x) + v_{n-1}(x)\,, \quad u_{n-1}, v_{n-1} \in \mathcal{P}_{n-1}\,,$$

pa iz $p(x_k) = v_{n-1}(x_k)$, $k = 1, \ldots, n$, sledi

$$\begin{aligned}\int_a^b p(x)\, w(x)\, dx &= \int_a^b W_n(x)\, u_{n-1}(x)\, dx + \int_a^b v_{n-1}(x)\, w(x)\, dx \\ &= \int_a^b v_{n-1}(x)\, w(x)\, dx = Q_n(v_{n-1}) = Q_n(p)\,.\end{aligned}$$

Dokažimo sada obrnuto tvrđenje. Neka Q_n ima algebarski stepen tačnosti $2n-1$. Za svaki od polinoma $x^m W_n$, $m = 0, 1, \ldots, n-1$, očigledno važi

$$Q_n(x^m W_n) = 0\,,$$

pa sledi

$$I_w(x^m W_n) = \int_a^b W_n(x)\, x^m\, w(x)\, dx = 0\,, \quad m = 0, 1, \ldots, n-1,$$

tj. W_n pripada nizu ortogonalnih polinoma u odnosu na težinsku funkciju $w(x)$. □

Prethodna teorema sugeriše metodu za konstrukciju Gaussovih kvadratura. Ukoliko prvo odredimo čvorove, koeficijente lako dobijamo iz sistema (6.4), i time obezbeđujemo neophodni stepen tačnosti $n-1$. Znatno teži deo problema je određivanje čvorova za koje smo videli da moraju biti nule polinoma W_n. Sad se nameće pitanje da li takav polinom uopšte postoji, i ako postoji, kakve su njegove nule. Za konstrukciju Gaussove kvadrature, neophodno je da one budu realne, proste i da pripadaju intervalu $[a, b]$.

S obzirom na pretpostavke o težinskoj funkciji, u prostor algebarskih polinoma nad $[a,b]$ možemo uvesti skalarni proizvod

$$\langle f,g\rangle = \int_a^b f(x)\,g(x)\,w(x)\,dx = 0\,,$$

pa egzistencija niza $\{W_k\}_{k\in\mathbb{N}_0}$ sledi iz Gram-Schmidtovog postupka ortogonalizacije. Naime, možemo uzeti da je

$$W_n = x^n - \sum_{i=0}^{n-1} \frac{\langle x^n, W_i\rangle}{\langle W_i, W_i\rangle}\, W_i\,.$$

Teorema 6.3.3 *Neka je $\{W_k\}_{k\in\mathbb{N}_0}$ niz ortogonalnih polinoma u odnosu na težinsku funkciju $w(x)$. Nule polinoma W_n su realne, proste i leže u (a,b).*

Dokaz. Neka su $t_1, t_2, \ldots, t_m$ $(m \le n)$ realne nule neparne višestrukosti polinoma $W_n(t)$ koje leže u intervalu (a,b). Polinom

$$Q_n(t) = \begin{cases} 1; & m = 0 \\ \prod\limits_{i=1}^{m}(t-t_i)\,; & m = 1,2,\ldots,n \end{cases}$$

na (a,b) menja znak u istim tačkama kao i polinom $\pi_n(t)$, odakle sledi da polinom $W_n(t)Q_m(t)$ ne menja znak na (a,b), pa važi

$$\int_a^b W_n(t)\,Q_m(t)\,w(t)\,dt \neq 0\,.$$

Ovo je moguće samo za $m = n$. □

Na osnovu algebarskog stepena tačnosti Gaussove kvadraturne formule G_n, zaključujemo da važi

$$G_n(f) = \int_a^b H_{2n-1}(x;f)\,dx\,, \tag{6.24}$$

gde je H_{2n-1} Hermiteov interpolacioni polinom funkcije f dobijen na osnovu vrednosti od f i f' u čvorovima kvadrature. Lako se proverava da važi

$$H_{2n-1}(x) = \sum_{i=1}^{n} \left(\mu_i(x)\,f(x_i) + \nu_i(x)\,f'(x_i)\right)\,,$$

gde je

$$\mu_i(x) = [1 - 2\ell_i'(x_i)\,(x - x_i)]\,\ell_i^2(x)\,, \quad \nu_i(x) = \ell_i^2(x)\,(x - x_i)\,.$$

Ako prethodno uvrstimo u (6.24) dobijamo

$$G_n(f) = \sum_{i=1}^{n} [A_i\, f(x_i) + B_i\, f'(x_i)]\,,$$

gde je

$$A_i = \int_a^b [1 - 2\ell_i'(x_i)\,(x - x_i)]\,\ell_i^2(x)\,w(x)\,dx\,, \quad B_i = \int_a^b \ell_i^2(x)\,(x - x_i)\,w(x)\,dx\,.$$

Kako se u kvadraturi G_n ne javljaju vrednosti od f', zaključujemo da je $B_i = 0$, $i = 1, \ldots, n$. Sada je

$$A_i = \int_a^b \ell_i^2(x)\,w(x)\,dx > 0\,.$$

Grešku $R_n(f)$ Gaussove kvadrature G_n dobijamo iz greške Hermiteovog interpolacionog polinoma H_{2n-1}. Ako $f \in C^{(2n)}[a, b]$, iz (3.26) sledi

$$R_n(f) = \frac{f^{(2n)}(\xi)}{(2n)\,!} \int_a^b W_n^2(x)\,w(x)\,dx\,.$$

Primer 6.7. Odredimo Gaussovu kvadraturnu formulu oblika

$$\int_{-1}^{1} f(x)\,(1 - x^2)^{3/2}\,dx = A_1\, f(x_1) + A_1\, f(x_1) + A_1\, f(x_1) + R_3(f)\,.$$

Polazeći od linearno nezavisnog niza $\{1, x, x^2, \ldots\}$, Gram-Schmidtovim postupkom ortogonalizacije dobijamo niz ortogonalnih polinoma u odnosu na težinsku funkciju $w(x) = (1 - x^2)^{3/2}$:

$$W_0(x) = 1\,, \quad W_1(x) = x\,, \quad W_2(x) = x^2 - \frac{1}{6}\,, \quad W_3(x) = x^3 - \frac{3}{8}x\,, \quad \ldots$$

Čvorovi x_1, x_2 i x_3 su nule polinoma W_3, dakle,

$$x_1 = -\sqrt{\frac{3}{8}}\,, \quad x_2 = 0\,, \quad x_3 = \sqrt{\frac{3}{8}}\,.$$

Rešavanjem linearnog sistema

$$\begin{array}{rcrcrcl} A_1 & + & A_2 & + & A_3 & = & I_w(1) \\ -\sqrt{\frac{3}{8}}A_1 & & & + & \sqrt{\frac{3}{8}}A_3 & = & I_w(x) \\ \frac{3}{8}A_1 & & & + & \frac{3}{8}A_3 & = & I_w(x^2)\,, \end{array}$$

dobijamo

$$A_1 = A_3 = \frac{\pi}{12}\,, \quad A_2 = \frac{5\,\pi}{24}\,.$$

Ako $f \in C^{(6)}[-1,1]$, imamo

$$R_3(f) = \frac{f^{(6)}(\xi)}{6\,!} \int_{-1}^{1} \left(x^3 - \frac{3}{8}x\right)^2 (1-x^2)^{3/2}\,dx = \frac{3\,\pi}{6!\,2^{10}}\, f^{(6)}(\xi)\,.$$

Glava 7

Numeričke metode za rešavanje diferencijalnih jednačina

Ova glava je posvećena uglavnom rešavanju Cauchyjevog problema kod običnih diferencijalnih jednačina, tj. problema sa početnim uslovima. Metode su razvrstane u dve opšte klase i to:

- Klasa linearnih višekoračnih metoda,
- Klasa metoda Runge–Kutta.

Ukazaćemo i na rešavanje konturnih problema kod običnih diferencijalnih jednačina.

7.1 Linearne višekoračne metode

7.1.1 Eulerova metoda

Eulerova metoda je najprostija numerička metoda za rešavanje Cauchyjevog problema

$$y' = f(x, y), \quad y(x_0) = y_0 \tag{7.1}$$

i bazira se na približnoj jednakosti

$$y(x) \approx y(x_0) + (x - x_0)y'(x_0),$$

tj.

$$y(x) \approx y(x_0) + (x - x_0)\, f(x_0, y_0), \tag{7.2}$$

s obzirom na (7.1). Ako sa y_1 označimo približnu vrednost za $y(x_1)$, na osnovu (7.2) imamo

$$y_1 = y_0 + (x_1 - x_0) f(x_0, y_0).$$

U opštem slučaju za proizvoljni skup tačaka

$$x_0 < x_1 < x_2 < \dots \ ,$$

približne vrednosti za $y(x_n)$, u oznaci y_n, možemo odrediti pomoću

$$y_{n+1} = y_n + (x_{n+1} - x_n) f(x_n, y_n) \quad (n = 0, 1, \dots). \tag{7.3}$$

Poslednja formula definiše Eulerovu metodu, čija je geometrijska interpretacija data na slici 7.1.

Poligonalna linija (x_0, y_0)—(x_1, y_1)—(x_2, y_2)— ... poznata je kao Eulerov poligon.

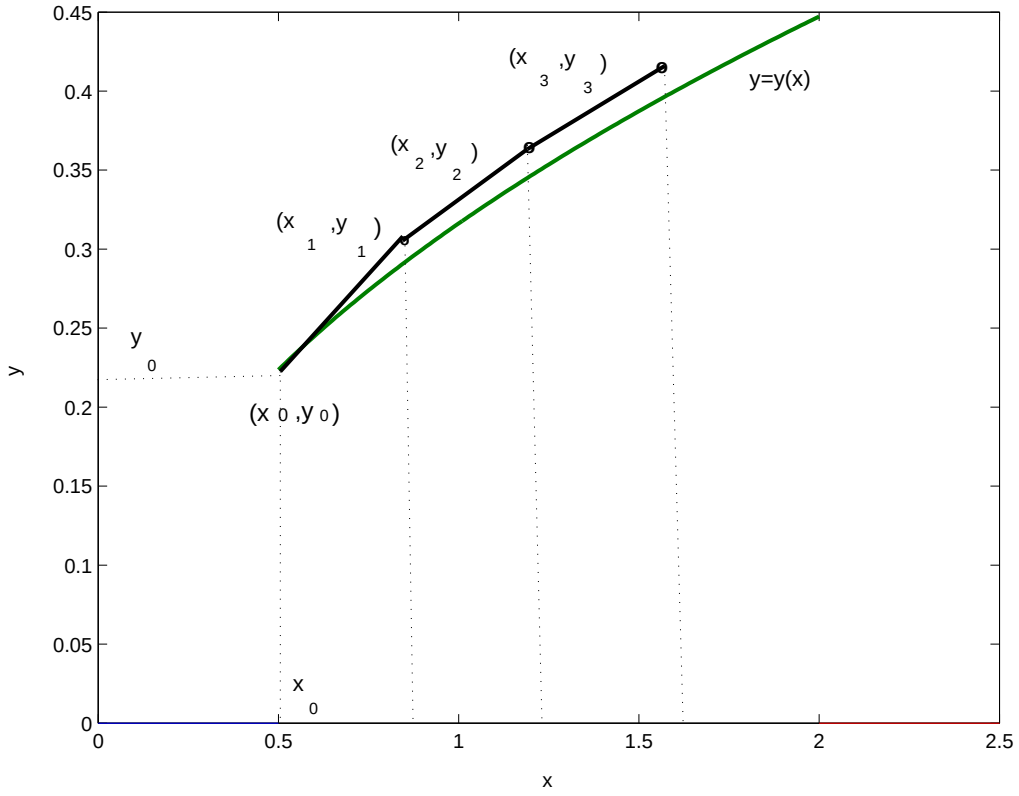
Najčešće se tačke x_n biraju ekvidistantno, tj $x_{n+1} - x_n = h = \text{const}$ (> 0) $(n = 0, 1, \dots)$ i u tom slučaju (7.3) se svodi na

$$y_{n+1} = y_n + h f(x_n, y_n) \quad (n = 0, 1, \dots).$$

7.1.2 Opšta linearna višekoračna metoda

U ovoj i narednim podsekcijama ove sekcije razmatramo jednu opštu klasu metoda za rešavanje Cauchyjevog problema

$$y' = f(x, y), \quad y(x_0) = y_0 \quad (x_0 \le x \le b). \tag{7.4}$$



Sl. 7.1.

Ako segment $[x_0, b]$ podelimo na N podsegmenata dužine $h = (b - x_0)/N$ dobijamo niz tačaka x_n određen sa

$$x_n = x_0 + nh \quad (n = 0, 1, \ldots, N).$$

Neka y_n označava niz približnih vrednosti rešenja problema (7.4) u tačkama x_n i neka je $f_n \equiv f(x_n, y_n)$. Pred nama se postavlja problem određivanja niza $\{y_n\}$. Za rešavanje ovog problema razvijen je veliki broj metoda. Jedna od njih je i Eulerova metoda razmatrana u prethodnoj podsekciji. Kod Eulerove metode niz $\{y_n\}$ se izračunava rekurzivno pomoću

$$y_{n+1} - y_n = hf_n \quad (n = 0, 1, \ldots), \tag{7.5}$$

pri čemu postoji linearna veza između y_n, y_{n+1} i f_n. U opštem slučaju za

izračunavanje niza y_n mogu se koristiti neke druge rekurentne relacije, koje su složenije nego što je (7.5). Među metodama, koje proizilaze iz pomenutih relacija, važnu ulogu igraju metode kod kojih postoji linearna veza između y_{n+i}, f_{n+i} $(i = 0, 1, \ldots, k)$, i one čine klasu takozvanih linearnih višekoračnih metoda ("multistep methods"na engleskom).

Opšta linearna višekoračna metoda može se predstaviti u obliku

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h \sum_{i=0}^{k} \beta_i f_{n+i} \quad (n = 0, 1, \ldots), \tag{7.6}$$

gde su α_i i β_i konstantni koeficijenti određeni sa tačnošću do na multiplikativnu konstantu. Da bismo obezbedili njihovu jednoznačnost uzimamo npr. $\alpha_k = 1$.

Ako je $\beta_k = 0$, kažemo da je metoda (7.6) otvorenog tipa ili da je eksplicitna, u protivnom metoda je zatvorenog tipa ili implicitna.

U opštem slučaju (7.6) predstavlja nelinearnu diferencnu jednačinu, s obzirom da je $f_{n+i} \equiv f(x_{n+i}, y_{n+i})$.

Za određivanje niza y_n, primenom metode (7.6) potrebno je poznavanje početnih vrednosti y_i $(i = 0, 1, \ldots, k-1)$. Kako nam je unapred poznata jedino vrednost y_0, poseban problem u primeni višekoračnih metoda (7.6) predstavlja određivanje ostalih početnih vrednosti. Ovom problemu biće posvećena posebna podsekcija.

Pod pretpostavkom da su poznate početne vrednosti y_i $(i = 0, 1, \ldots, k-1)$, kod eksplicitnih metoda direktno se izračunavaju $y_k, y_{k+1}, \ldots, y_N$ pomoću

$$y_{n+k} = h \sum_{i=0}^{k-1} \beta_i f_{n+i} - \sum_{i=0}^{k-1} \alpha_i y_{n+i} \quad (n = 0, 1, \ldots, N-k).$$

Međutim, kod implicitnih metoda za određivanje vrednosti y_{n+k} treba rešiti jednačinu

$$y_{n+k} = h\beta_k f(x_{n+k}, y_{n+k}) + \Phi, \tag{7.7}$$

gde je

$$\Phi = h\sum_{i=0}^{k-1}\beta_i f_{n+i} - \sum_{i=0}^{k-1}\alpha_i y_{n+i}.$$

Kada je $f(x,y)$ nelinearna funkcija koja zadovoljava Lipschitzov uslov po y sa konstantom L u nekoj oblasti D,

$$|f(x,y_1) - f(x,y_2)| \leq L|y_1 - y_2|, \quad (x,y_1),(x,y_2) \in D,$$

jednačina (7.7) se može rešiti iterativnim procesom

$$y_{n+k}^{[s+1]} = h\beta_k f\left(x_{n+k}, y_{n+k}^{[s]}\right) + \Phi \tag{7.8}$$

polazeći od proizvoljne vrednosti $y_{n+k}^{[0]}$ ako je

$$h|\beta_k|L < 1.$$

Uslov dat ovom nejednakošću obezbeđuje konvergenciju iterativnog procesa (7.8).

Za metodu (7.6) definiše se diferencni operator $L_h : C^1[x_0,b] \mapsto C[x_0,b]$ pomoću

$$L_h[y] = \sum_{i=0}^{k}[\alpha_i y(x+ih) - h\beta_i y'(x+ih)]. \tag{7.9}$$

Neka je funkcija $g \in C^\infty[x_0,b]$. Tada se $L_h[g]$ može predstaviti u obliku

$$L_h[g] = C_0 g(x) + C_1 h g'(x) + C_2 h^2 g''(x) + \dots\ , \tag{7.10}$$

gde su C_j $(j = 0,1,\dots)$ konstante, koje ne zavise od g i h.

Definicija 7.1.1 *Linearna višekoračna metoda* (7.6) *ima red* p *ako je u razvoju* (7.10)

$$C_0 = C_1 = \dots = C_p = 0,\ C_{p+1} \neq 0.$$

Neka je $y(x)$ tačno rešenje problema (7.4) i y_n niz približnih vrednosti ovog rešenja u tačkama $x_n = x_0 + nh$ $(n = 0, 1, \ldots, N)$ dobijen primenom metode (7.6), sa početnim vrednostima $y_i = s_i(h)$ $(i = 0, 1, \ldots, k-1)$.

Definicija 7.1.2 *Za linearnu višekoračnu metodu* (7.6) *se kaže da je konvergentna, ako je za svako* $x \in [x_0, b]$

$$\lim_{h\to 0, x-x_0=nh} y_n = y(x)$$

i ako za početne vrednosti važi

$$\lim_{h\to 0} s_i(h) = y_0 \ (i = 0, 1, \ldots, k-1).$$

Linearna višekoračna metoda (7.6) se može okarakterisati prvim i drugim karakterističnim polinomom koji su dati, tim redom, pomoću

$$\rho(\xi) = \sum_{i=0}^{k} \alpha_i \xi^i, \quad \sigma(\xi) = \sum_{i=0}^{k} \beta_i \xi^i.$$

Dve važne klase konvergentnih višekoračnih metoda koje se sreću u primenama su:

1^0 Metode kod kojih je $\rho(\xi) = \xi^k - \xi^{k-1}$,

2^0 Metode kod kojih je $\rho(\xi) = \xi^k - \xi^{k-2}$.

Ekplicitne metode prve klase nazivaju se Adams-Bashfortove, a implicitne Adams-Moultonove metode. Slično, eksplicitne metode druge klase su Nystromove, dok se odgovarajuće implicitne metode nazivaju generalisane Milne-Simpsonove metode.

7.1.3 Izbor početnih vrednosti

Kao što je ranije napomenuto, kod primene linearnih višekoračnih metoda na rešavanje problema (7.4) potrebno je poznavanje početnih vrednosti $y_i =$

$s_i(h)$, takvih da je $\lim_{h\to 0} s_i(h) = y_0$ $(i = 1, 2, \ldots, k-1)$. Naravno, ovaj problem se postavlja kada je $k > 1$.

Ako je metoda (7.6) reda p, tada početne vrednosti $s_i(h)$ treba birati tako da je

$$s_i(h) - y(x_i) = O\left(h^{p+1}\right) \quad (i = 1, 2, \ldots, k-1),$$

gde je $y(x)$ tačno rešenje problema (7.4).

U ovoj podsekciji navešćemo jednu klasu metoda za određivanje potrebnih početnih vrednosti.

Pretpostavimo da je funkcija f u diferencijalnoj jednačini (7.4) dovoljan broj puta diferencijabilna. Tada na osnovu Taylorovog metoda imamo

$$y(x_0 + h) = y(x_0) + hy'(x_0) + \frac{h^2}{2!}y''(x_0) + \ldots + \frac{h^p}{p!}y^{(p)}(x_0) + O\left(h^{p+1}\right).$$

Poslednja jednakost ukazuje na to da se može uzeti

$$s_1(h) = y(x_0) + hy'(x_0) + \frac{h^2}{2!}y''(x_0) + \ldots + \frac{h^p}{p!}y^{(p)}(x_0),$$

s obzirom da je tada $s_i(h) - y(x_1) = O\left(h^{p+1}\right)$ $(x_1 = x_0 + h)$. Isti postupak se može primeniti na određivanje ostalih početnih vrednosti. Naime, u opštem slučaju imamo

$$s_i(h) = y(x_{i-1}) + hy'(x_{i-1}) + \frac{h^2}{2!}y''(x_{i-1}) + \ldots + \frac{h^p}{p!}y^{(p)}(x_{i-1}) \quad (i = 1, \ldots, k-1),$$

pri čemu za $y(x_{i-1})$ uzimamo $s_{i-1}(h)$.

7.1.4 Prediktor-korektor metode

Kao što je navedeno primena implicitnih metoda je skopčana sa rešavanjem jednačine (7.7) na svakom koraku integracije, pri čemu se za ovo rešavanje koristi iterativni proces (7.8). Bez obzira na ovu teškoću implicitne metode se dosta koriste za rešavanje Cauchyjevog problema, s obzirom da

imaju niz prednosti nad eksplicitnim metodama (viši red, bolja numerička stabilnost). Početna vrednost $y_{n+k}^{[0]}$ u primenama se određuje korišćenjem neke eksplicitne metode, koju nazivamo prediktor. Implicitnu metodu (7.7) nazivamo korektor. Metodu dobijenu ovakvom kombinacijom nazivamo prediktor-korektor metodom.

Za određivanje y_{n+k}, iterativni proces (7.8) treba primenjivati sve dok ne bude ispunjen uslov

$$\left| y_{n+k}^{[s+1]} - y_{n+k}^{[s]} \right| < \varepsilon,$$

gde je ε dozvoljena greška, obično reda lokalne greške zaokruživanja. Tada se za y_{n+k} može uzeti $y_{n+k}^{[s+1]}$.

Međutim, ovakav način se najčešće ne primenjuje u praksi s obzirom da zahteva veliki broj izračunavanja vrednosti funkcije f u svakom koraku i uz to je ovaj broj promenljiv od koraka do koraka. Da bi se smanjio ovaj broj izračunavanja, broj iteracija u (7.8) se fiksira. Uzima se samo $s = 0, 1, \ldots, m-1$.

Primer 7.1. Posmatramo problem

$$y' = x^2 + y, \quad y(1) = 1 \quad (1 \leq x \leq 2), \tag{7.11}$$

čije je tačno rešenje $y(x) = 6e^{x-1} - x^2 - 2x - 2$.

Eulerova metoda

$$y_{n+1} - y_n = hf_n \quad (n = 0, 1, \ldots),$$

čiji je red $p = 1$, i Adams-Bashfortova metoda trećeg reda

$$y_{n+3} - y_{n+2} = \frac{h}{12}(23f_{n+2} - 16f_{n+1} + 5f_n) \quad (n = 0, 1, \ldots),$$

mogu se realizovati jednostavno korišćenjem nekog programskog jezika, recimo Fortran-a, ili u nekom softverskom paketu, recimo u Matlab-u. U formiranim programima testiranja su izvršena na primeru (7.11) sa $h = 0.1$.

Kod Eulerove metode $y(1)$ predstavlja datu početnu vrednost, a za Adamsovu metodu početne vrednosti se određuju primenom Taylorove metode za $p = 3$. Naime, kako je

$$y(1) = 1, \quad y'(1) = 2, \quad y''(1) = 4, \quad y'''(1) = 6, \quad h = 0.1,$$

dobijamo

$$y(1) = 1., \quad y(2) = 1.221, \quad y(3) = 1.48836.$$

Izračunate vrednosti funkcije y u tačkama x_i $(i = 0, 1, \ldots)$ dobijene Eulerovom, odnosno Adamsovom metodom, i označene sa $\bar{y}_i$, $\widetilde{y}_i$, respektivno, prikazane su u sledećoj tabeli.

i	x_i	$\bar{y}_i$	Greška (u %)	$\widetilde{y}_i$	Greška (u %)
0	1.0	1.00000	0.00	1.00000	0.00
1	1.1	1.20000	1.72	1.22100	0.00
2	1.2	1.44100	3.19	1.48836	0.00
3	1.3	1.72910	4.42	1.80883	0.02
4	1.4	2.07101	5.47	2.19028	0.03
5	1.5	2.47411	6.37	2.64126	0.04
6	1.6	2.94652	7.13	3.17116	0.05
7	1.7	3.49717	7.79	3.79040	0.06
8	1.8	4.13589	8.36	4.51045	0.06
9	1.9	4.87348	8.87	5.34403	0.07
10	2.0	5.72183	9.32	6.30518	0.07

Primer 7.2. Uzimajući Eulerovu metodu kao prediktor i trapezno pravilo $(p = 2)$

$$y_{n+1} - y_n = \frac{h}{2}(f_n + f_{n+1}) \quad (n = 0, 1, \ldots)$$

kao korektor (sa brojem iteracija $m = 2$), imamo metodu za rešavanje diferencijalne jednačine (7.11), čiji test rezultati su prikazani u sledećoj tabeli.

i	x_i	y_i	Greška (u %)
0	1.0	1.00000	0.00
1	1.1	1.22152	0.04
2	1.2	1.48952	0.07
3	1.3	1.81097	0.10
4	1.4	2.19363	0.12
5	1.5	2.64602	0.14
6	1.6	3.17760	0.15
7	1.7	3.79881	0.17
8	1.8	4.52118	0.18
9	1.9	5.35747	0.18
10	2.0	6.32177	0.19

7.2 Metode Runge-Kutta

Pre nego što pređemo na razmatranje navedenih metoda pomenućemo Taylorovu metodu, koji spada u grupu približnih metoda za rešavanje Cauchyjevog problema (7.4).

Ako je funkcija $f(x, y)$ analitička u tački (x_0, y_0), može se pokazati da Cauchyjev problem (7.4) tada ima jedinstveno rešenje $y = y(x)$ koje u okolini tačke x_0 ima izvode proizvoljnog reda, pa je

$$y(x) = y(x_0) + (x - x_0)y'(x_0) + \frac{1}{2!}(x - x_0)^2 y''(x_0) + \dots \ . \tag{7.12}$$

Na osnovu (7.4) možemo izračunati potrebne izvode $y^{(i)}(x_0)$ $(i = 1, 2, \dots)$. Naime, imamo redom

$$\begin{aligned} y(x_0) &= y_0, \\ y'(x_0) &= f(x_0, y_0), \\ y''(x_0) &= \left(\frac{\partial f}{\partial x} + y'\frac{\partial f}{\partial y}\right)_{x=x_0, y=y_0} \\ &= \frac{\partial f(x_0, y_0)}{\partial x} + f(x_0, y_0)\frac{\partial f(x_0, y_0)}{\partial y}, \quad \text{itd.} \end{aligned}$$

Tako, za konkretan problem

$$y' = x^2 + y^2, \quad y(0) = 1,$$

sukcesivnim diferenciranjem dobijamo

$$\begin{aligned}
&y' = x^2 + y^2, && y'_0 = x_0^2 + y_0^2 = 1,\\
&y'' = 2x + 2yy', && y''_0 = 2x_0 + 2y_0y'_0 = 2,\\
&y''' = 2 + 2yy'' + 2y'^2, && y'''_0 = 2 + 2y_0y''_0 + 2y'^2_0 = 8,\\
&y^{(4)} = 2yy''' + 6y'y'', && y^{(4)}_0 = 2y_0y'''_0 + 6y'_0y''_0 = 28,
\end{aligned}$$

gde smo stavili $y_0^{(i)} = y^{(i)}(x_0)$.

Zamenom dobijenih vrednosti u (7.12) dobijamo

$$y(x) = 1 + x + 2\frac{x^2}{2!} + 8\frac{x^3}{3!} + 28\frac{x^4}{4!} + \dots,$$

tj.

$$y(x) = 1 + x + x^2 + \frac{4}{3}x^3 + \frac{7}{6}x^4 + \dots\ .$$

U prethodnoj sekciji razmatrani su linearne višekoračne metode za rešavanje Cauchyjevog problema (7.4). Red ovih metoda se može povećati povećanjem broja koraka. Međutim, ukoliko se žrtvuje linearnost koju poseduju ove metode, moguće je konstruisati jednokoračne metode sa proizvoljnim redom.

Za rešavanje Cauchyjevog problema oblika (7.4) sa dovoljno puta diferencijabilnom funkcijom f, moguće je, takođe, konstruisati jednokoračne metode višeg reda (na primer, Taylorova metoda).

Posmatrajmo opštu eksplicitnu jednokoračnu metodu

$$y_{n+1} - y_n = h\Phi(x_n, y_n, h). \tag{7.13}$$

Definicija 7.2.1 *Metoda* (7.13) *je reda p ako je p najveći ceo broj za koji važi*

$$y(x+h) - y(x) - h\Phi(x, y(x), h) = O\left(h^{p+1}\right),$$

gde je $y(x)$ tačno rešenje problema (7.4).

Definicija 7.2.2 *Metoda* (7.13) *je konzistentna ako je* $\Phi(x, y, 0) \equiv f(x, y)$.

Primetimo da je Taylorova metoda specijalan slučaj metode (7.13). Naime, kod Taylorove metode reda p imamo

$$\Phi(x, y, h) = \Phi_T(x, y, h) = \sum_{i=0}^{p-1} \frac{h^i}{(i+1)!} \left(\frac{\partial}{\partial x} + f \frac{\partial}{\partial y} \right)^i f(x, y). \qquad (7.14)$$

U specijalnom slučaju, kod Eulerove metode je $\Phi(x, y, h) = f(x, y)$.

U ovoj sekciji razmatraćemo jednu specijalnu klasu metoda oblika (7.13), koja je nazvana metodama Runge-Kutta.

Kako ćemo kasnije videti sve ove metode sadrže slobodne parametre. S obzirom na vreme u kome su se pojavile ove metode, slobodni parametri su birani tako da se dobiju što jednostavnije formule za praktično računanje. Međutim, ovakve vrednosti parametara ne obezbeđuju optimalne karakteristike posmatranih metoda. U daljem tekstu ove metode zvaćemo klasičnim metodama Runge-Kutta.

Opšta eksplicitna metoda Runge-Kutta ima oblik

$$y_{n+1} - y_n = h\Phi(x_n, y_n, h), \qquad (7.15)$$

gde su

$$\Phi(x, y, h) = \sum_{i=1}^{m} c_i \, k_i,$$

$$\begin{aligned} k_1 &= f(x, y), \\ k_i &= f(x + a_i h, y + b_i h) \quad (i = 2, \ldots, m), \\ a_i &= \sum_{j=1}^{i-1} \alpha_{ij}, \quad b_i = \sum_{j=1}^{i-1} \alpha_{ij} k_j. \end{aligned} \qquad (7.16)$$

Primetimo da iz uslova konzistencije metode (7.15) sledi da je $\sum_{i=1}^{m} c_i = 1$.

Nepoznate koeficijente koji figurišu u ovoj metodi određujemo iz uslova da metoda ima maksimalni red. Pri tome koristimo sledeću činjenicu: Ako se $\Phi(x, y, h)$, razvijeno po stepenima od h, može predstaviti u obliku

$$\Phi(x, y, h) = \Phi_T(x, y, h) + O(h^p),$$

gde je Φ_T definisano pomoću (7.14), tada je metoda (7.15) reda p.

Najpre nađimo razvoj $\Phi_T(x, y, h)$ po stepenima od h. Korišćenjem Mongeovih oznaka za parcijalne izvode imamo

$$\left(\frac{\partial}{\partial x} + f\frac{\partial}{\partial y}\right) f = f_x + f f_y = F$$

i

$$\left(\frac{\partial}{\partial x} + f\frac{\partial}{\partial y}\right)^2 f = \left(\frac{\partial}{\partial x} + f\frac{\partial}{\partial y}\right) F = G + f_y F,$$

gde smo stavili $G = f_{xx} + 2f f_{xy} + f^2 f_{yy}$. Tada iz (7.14) sledi

$$\Phi_T(x, y, h) = f + \frac{1}{2}hF + \frac{1}{6}h^2(G + f_y F) + O(h^3). \tag{7.17}$$

Razmotrićemo sada samo metode Runge-Kutta čiji je red $p \leq 3$. Pokazuje se da je za dobijanje metoda trećeg reda dovoljno uzeti $m = 3$. U tom slučaju formule (7.15) se svode na

$$\begin{aligned} \Phi(x, y, h) &= c_1 k_1 + c_2 k_2 + c_3 k_3, \\ k_1 &= f(x, y), \\ k_2 &= f(x + a_2 h, y + b_2 h), \\ k_3 &= f(x + a_3 h, y + b_3 h) \end{aligned}$$

i

$$\begin{aligned} a_2 &= \alpha_{21}, \quad b_2 = \alpha_{21} k_1, \\ a_3 &= \alpha_{31} + \alpha_{32}, \quad b_2 = \alpha_{31} k_1 + \alpha_{32} k_2. \end{aligned}$$

Razvijanjem funkcije k_2 u Taylorov red, u okolini tačke (x, y), dobijamo

$$k_2 = f + a_2 F h + \frac{1}{2} a_2^2 G h^2 + O(h^3).$$

Kako je

$$b_3 = \alpha_{31} k_1 + \alpha_{32} k_2 = \alpha_{31} f + \alpha_{32} \left(f + a_2 F h + \frac{1}{2} G h^2 \right) + O(h^3),$$

imamo

$$b_3 = a_3 f + a_2 \alpha_{32} F h + O(h^2) \quad \text{i} \quad b_3^2 = a_3^2 f^2 + O(h).$$

Razvijanjem funkcije k_3 u okolini tačke (x, y) i korišćenjem poslednjih jednakosti imamo

$$k_3 = f + a_3 F h + \frac{1}{2}(2 a_3 \alpha_{32} F f_y + a_3^2 G) h^2 + O(h^3).$$

Konačno, zamenom dobijenih izraza za k_1, k_2, k_3 u izrazu za $\Phi(x, y, h)$ dobijamo

$$\begin{aligned} \Phi(x, y, h) &= (c_1 + c_2 + c_3) f + (c_2 a_2 + c_3 a_3) F h \\ &+ (c_2 a_2^2 G + 2 c_3 a_2 \alpha_{32} F f_y + c_3 a_3^2 G) \frac{h^2}{2} + O(h^3). \end{aligned}$$

Poslednja jednakost dozvoljava konstrukciju metode za $m = 1, 2, 3$.

Slučaj $m = 1$**.** Kako je $c_2 = c_3 = 0$ imamo

$$\Phi(x, y, h) = c_1 f + O(h^3).$$

Upoređivanjem sa (7.17) dobijamo

$$\Phi_T(x, y, h) - \Phi(x, y, h) = (1 - c_1) f + \frac{1}{2} h^2 (G + f_y F) + O(h^3),$$

odakle zaključujemo da se za $c_1 = 1$ dobija metoda

$$y_{n+1} - y_n = h f_n,$$

čiji je red $p = 1$. S obzirom da je ovo Eulerova metoda vidimo da ona pripada i klasi metoda Runge-Kutta.

Slučaj $m = 2$**.** Ovde je $c_3 = 0$ i

$$\Phi(x, y, h) = (c_1 + c_2)f + c_2 a_2 Fh + \frac{1}{2} c_2 a_2^2 Gh^2 + O(h^3).$$

Kako je

$$\begin{aligned} \Phi(x, y, h) - \Phi_T(x, y, h) &= (c_1 + c_2 - 1)f + \left(c_2 a_2 - \frac{1}{2}\right) Fh \\ &+ \frac{1}{6}[(3c_2 a_2^2 - 1)G - f_y F]h^2 + O(h^3), \end{aligned}$$

zaključujemo da se pod uslovima

$$c_1 + c_2 = 1 \quad \text{i} \quad c_2 a_2 = \frac{1}{2} \tag{7.18}$$

dobija metoda drugog reda sa jednim slobodnim parametrom. Naime, iz (7.18) sledi

$$c_2 = \frac{1}{2a_2} \quad \text{i} \quad c_1 = \frac{2a_2 - 1}{2a_2},$$

gde je a_2 ($\neq 0$) slobodan parametar. Dakle, za $m = 2$ imamo jednoparametarsku familiju metoda

$$\begin{aligned} y_{n+1} - y_n &= \frac{h}{2a_2}((2a_2 - 1)k_1 + k_2), \\ k_1 &= f(x_n, y_n), \\ k_2 &= f(x_n + a_2 h, y_n + a_2 k_1 h). \end{aligned}$$

U specijalnom slučaju, za $a_2 = \frac{1}{2}$, dobijamo Euler-Cauchyjevu metodu

$$y_{n+1} - y_n = hf\left(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hf(x_n, y_n)\right).$$

Slično, za $a_2 = 1$, dobijamo takozvanu poboljšanu Euler-Cauchyjevu metodu

$$y_{n+1} - y_n = \frac{1}{2}[f(x_n, y_n) + f(x_n + h, y_n + hf(x_n, y_n)].$$

Slučaj $m = 3$**.** Kako je

$$\begin{aligned}\Phi(x,y,h) &- \Phi_T(x,y,h) = (c_1+c_2+c_3-1)f + \left(c_2a_2+c_3a_3-\frac{1}{2}\right)Fh \\ &+ \left[\left(c_2a_2^2+c_3a_3^2-\frac{1}{3}\right)G + \left(2c_3a_2\alpha_{32}-\frac{1}{3}\right)Ff_y\right]\frac{h^2}{2}+O(h^3),\end{aligned}$$

zaključujemo da su za dobijanje metoda trećeg reda dovoljni uslovi

$$\begin{aligned} c_1+c_2+c_3 &= 1,\\ c_2a_2+c_3a_3 &= \frac{1}{2},\\ c_2a_2^2+c_3a_3^2 &= \frac{1}{3},\\ c_3a_2\alpha_{32} &= \frac{1}{6}\,. \end{aligned} \tag{7.19}$$

Kako imamo četiri jednačine sa šest nepoznatih to u slučaju $m = 3$ imamo dvoparametarsku familiju metoda Runge-Kutta. Može se pokazati da među metodama ove familije ne postoji ni jedna metoda čiji je red veći od tri.

U specijalnom slučaju kada je $a_2 = 1/3$ i $a_3 = 2/3$, iz (7.19) sledi da je $c_1 = 1/4$, $c_2 = 0$, $c_3 = 3/4$, $\alpha_{32} = 2/3$. Dakle, dobili smo metodu

$$\begin{aligned} y_{n+1} &- y_n = \frac{h}{4}(k_1+3k_3),\\ k_1 &= f(x_n,y_n),\\ k_2 &= f\left(x_n+\frac{h}{3},y_n+\frac{h}{3}\,k_1\right),\\ k_3 &= f\left(x_n+\frac{2h}{3},y_n+\frac{2h}{3}\,k_2\right), \end{aligned}$$

koja se u literaturi sreće kao Heunova metoda.

Za

$$a_2 = \frac{1}{2}, a_3 = 1 \ (\Longrightarrow c_1 = c_3 = 1/6,\ c_2 = 2/3,\ \alpha_{32} = 2)$$

dobijamo metodu

$$y_{n+1} - y_n = \frac{h}{6}(k_1+4k_2+k_3),$$

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ k_2 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}\, k_1\right), \\ k_3 &= f\left(x_n + h, y_n - h\, k_1 + 2h\, k_2\right), \end{aligned}$$

koja je najpopularnija među metodama trećeg reda sa stanovišta ručnog izračunavanja.

U slučaju kada je $m = 4$, dobijamo dvoparametarsku familiju metoda četvrtog reda. Ovde se analogno sistemu (7.19) javlja sistem od 11 jednačina sa 13 nepoznatih.

Navodimo, bez dokazivanja, metodu Runge-Kutta četvrtog reda koja se u primenama tradicionalno najviše koristi:

$$\begin{aligned} y_{n+1} - y_n &= \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4), \\ k_1 &= f(x_n, y_n), \\ k_2 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}\, k_1\right), \\ k_3 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}\, k_2\right), \\ k_4 &= f(x_n + h, y_n + hk_3). \end{aligned} \tag{7.20}$$

Za razliku od linearnih višekoračnih metoda, metode Runge-Kutta ne zahtevaju poznavanje početnih vrednosti (sem $y(x_0) = y_0$, koja inače definiše Cauchyjev problem), ali su za praktičnu primenu znatno komplikovaniji, s obzirom da zahtevaju m izračunavanja vrednosti funkcije f u svakom koraku.

Primer 7.3. Programska realizacija Euler-Cauchyjeve, poboljšane Euler-Cauchyjeve i metode četvrtog reda (7.20), respektivno, u Fortran-u ili Matlab-u je testirana na primeru (7.11).

Kao ulazne parametre za integraciju, kod obične Euler-Cauchyjeve metode, uzeli smo $h = 0.1, N = 10, m = 1$. Vrednosti rešenja datog Cauchyjevog problema (7.11), kao i odgovarajućih grešaka (izraženih u %) u odnosu na tačno rešenje, prikazane su u sledećoj tabeli.

i	x_i	y_i	Greška (u %)
0	1.0	1.000000	0.00000
1	1.1	1.220250	0.06349
2	1.2	1.486676	0.11696
3	1.3	1.806227	0.16179
4	1.4	2.186581	0.19934
5	1.5	2.636222	0.23109
6	1.6	3.164526	0.25814
7	1.7	3.781851	0.28130
8	1.8	4.499645	0.30134
9	1.9	5.330558	0.31909
10	2.0	6.288567	0.33481

Neka su sada ulazni parametri za integraciju, kod poboljšane Euler-Cauchyjeve metode, opet $h = 0.1, N = 10, m = 1$. Vrednosti rešenja datog Cauchyjevog problema (7.11), kao i odgovarajućih grešaka (izraženih u %) u odnosu na tačno rešenje, prikazane su u sledećoj tabeli.

i	x_i	y_i	Greška (u %)
0	1.0	1.000000	0.00000
1	1.1	1.220500	0.04302
2	1.2	1.487203	0.08161
3	1.3	1.807059	0.11583
4	1.4	2.187750	0.14598
5	1.5	2.637764	0.17274
6	1.6	3.166479	0.19656
7	1.7	3.784260	0.21778
8	1.8	4.502557	0.23682
9	1.9	5.334026	0.25424
10	2.0	6.292649	0.27012

Konačno, pod istim uslovima vrednosti rešenja datog Cauchyjevog problema (7.11), kao i odgovarajućih grešaka (izraženih u %) u odnosu na tačno rešenje, dobijene pomoću metode Runge-Kutta četvrtog reda (7.20), prikazane su u sledećoj tabeli.

i	x_i	y_i	Greška (u %)
0	1.0	1.000000	0.00000
1	1.1	1.221025	0.00000
2	1.2	1.488416	0.00008
3	1.3	1.809152	0.00014
4	1.4	2.190947	0.00009
5	1.5	2.642325	0.00011
6	1.6	3.172710	0.00019
7	1.7	3.792512	0.00018
8	1.8	4.513240	0.00012
9	1.9	5.347612	0.00019
10	2.0	6.309683	0.00015

7.3 Rešavanje sistema jednačina i jednačina višeg reda

Metode koje su razmatrane u prethodnim sekcijama mogu se uopštiti u tom smislu da budu primenljive za rešavanje Cauchyjevog problema za sistem od p jednačina prvog reda

$$y_i' = f(x; y_1, \ldots, y_p), \quad y_i(x_0) = y_{i0} \ (i = 1, 2, \ldots, p). \tag{7.21}$$

U ovom slučaju sistem jednačina (7.21) treba predstaviti u vektorskom obliku

$$\overrightarrow{y}' = \overrightarrow{f}(x, \overrightarrow{y}), \quad \overrightarrow{y}(x_0) = \overrightarrow{y}_0, \tag{7.22}$$

gde su

$$\overrightarrow{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}, \quad \overrightarrow{y}_0 = \begin{bmatrix} y_{10} \\ y_{20} \\ \vdots \\ y_{p0} \end{bmatrix}, \quad \overrightarrow{f}(x, \overrightarrow{y}) = \begin{bmatrix} f_1(x; y_1, \ldots, y_p) \\ f_2(x; y_1, \ldots, y_p) \\ \vdots \\ f_p(x; y_1, \ldots, y_p) \end{bmatrix}.$$

Od interesa je i rešavanje Cauchyjevog problema za diferencijalne jednačine višeg reda. Međutim ovaj se problem može svesti na prethodni. Naime, neka

je data diferencijalna jednačina reda p

$$y^{(p)} = f(x, y, y', \ldots, y^{(p-1)}) \tag{7.23}$$

sa početnim uslovima

$$y^{(i)}(x_0) = y_{i0} \quad (i = 0, 1, \ldots, p-1). \tag{7.24}$$

Tada se supstitucijama

$$z_1 = y,\ z_2 = y',\ \ldots\ , z_p = y^{(p-1)},$$

jednačina (7.23) sa uslovima (7.24) svodi na sistem

$$\begin{aligned} z_1' &= z_2, \\ z_2' &= z_3, \\ &\vdots \\ z_{p-1}' &= z_p, \\ z_p' &= f(x; z_1, z_2, \ldots, z_p), \end{aligned}$$

sa uslovima

$$z_i(x_0) = z_{i0} = y_{i0} \quad (i = 1, 2, \ldots, p).$$

Linearne višekoračne metode, koje smo do sada razmatrali, mogu se formalno generalisati na vektorski oblik

$$\sum_{i=0}^{k} \alpha_i \overrightarrow{y}_{n+i} = h \sum_{i=0}^{k} \beta_i \overrightarrow{f}_{n+i},$$

gde je $\overrightarrow{f}_{n+i} = \overrightarrow{f}(x_{n+i}, \overrightarrow{y}_{n+i})$, a zatim se kao takve mogu primeniti na rešavanje Cauchyjevog problema (7.22).

Takođe, metode Runge-Kutta za rešavanje Cauchyjevog problema (7.22) imaju oblik

$$\overrightarrow{y}_{n+1} - \overrightarrow{y}_n = h\overrightarrow{\Phi}(x_n, \overrightarrow{y}_n, h),$$

gde su

$$\overrightarrow{\Phi}(x, \overrightarrow{y}, h) = \sum_{i=1}^{m} c_i \overrightarrow{k}_i,$$
$$\overrightarrow{k}_1 = \overrightarrow{f}(x, \overrightarrow{y}),$$

i

$$\begin{aligned} \overrightarrow{k}_i &= \overrightarrow{f}(x + a_i h, \overrightarrow{y} + \overrightarrow{b}_i h), \\ a_i &= \sum_{j=1}^{i-1} \alpha_{ij}, \quad \overrightarrow{b}_i = \sum_{j=1}^{i-1} \alpha_{ij} \overrightarrow{k}_j \qquad (i = 2, 3, \ldots, m). \end{aligned}$$

Celokupna analiza, koja je data u prethodnim sekcijama ove glave, formalno se može preneti na ove vektorske metode.

Primer 7.4. Realizujmo standardnu metodu Runge-Kutta četvrtog reda (7.20) za rešavanje sistema od dve diferencijalne jednačine

$$y' = xyz, \quad z' = \frac{xy}{z},$$

pri uslovima

$$y(1) = 1/3, \ z(1) = 1,$$

na segmentu $[1, 2.5]$ uzimajući korak integracije $h = 0.01$. U sledećoj tabeli prikazaćemo test rezultate u vrednostima x sa korakom 0.1 i odgovarajuće vrednosti za y, y_T, z, z_T, gde su y_T i z_T tačna rešenja ovog zadatka i data su sa

$$y_T(x) = \frac{72}{(7 - x^2)^3} \quad \text{i} \quad z_T = \frac{6}{7 - x^2}\,.$$

x	y	y_T	z	z_T
1.00	0.3333333	0.3333333	1.0000000	1.0000000
1.10	0.3709342	0.3709341	1.0362694	1.0362694
1.20	0.4188979	0.4188979	1.0791367	1.0791367
1.30	0.4808936	0.4808936	1.1299436	1.1299435
1.40	0.5623944	0.5623943	1.1904763	1.1904762
1.50	0.6718182	0.6718181	1.2631581	1.2631578
1.60	0.8225904	0.8225905	1.3513514	1.3513515
1.70	1.0370675	1.0370678	1.4598541	1.4598541
1.80	1.3544686	1.3544689	1.5957446	1.5957447
1.90	1.8481333	1.8481344	1.7699113	1.7699116
2.00	2.6666656	2.6666667	1.9999998	2.0000000
2.10	4.1441259	4.1441321	2.3166018	2.3166029
2.20	7.1444836	7.1444917	2.7777767	2.7777779
2.30	14.3993673	14.3994160	3.5087693	3.5087738
2.40	37.7629280	37.7631035	4.8387012	4.8387108
2.50	170.6634674	170.6666718	7.9999280	8.0000000

7.4 Konturni problemi

U ovoj sekciji ukazaćemo na diferencnu metodu za rešavanje konturnog problema

$$y'' + p(x)y' + q(x) = f(x), \quad y(a) = A,\ y(b) = B, \tag{7.25}$$

gde su funkcije p, q, f neprekidne na $[a, b]$.

Segment $[a, b]$ podelimo na $N+1$ podsegmenata dužine $h = \dfrac{b-a}{N+1}$, tako da je $x_n = a + nh$ $(n = 0, 1, \ldots, N+1)$. U tačkama x_n $(n = 1, 2, \ldots, N)$ diferencijalnu jednačinu iz (7.25) aproksimirajmo sa (videti (3.34), (3.35))

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} + p_n \frac{y_{n+1} - y_{n-1}}{2h} + q_n y_n = f_n \quad (n = 1, 2, \ldots, N), \tag{7.26}$$

gde su $p_n = p(x_n)$, $q_n = q(x_n)$, $f_n = f(x_n)$.

Ako uvedemo smene

$$a_n = 1 - \frac{h}{2}\, p_n, \quad b_n = h^2 q_n - 2, \quad c_n = 1 + \frac{h}{2}\, p_n,$$

(7.26) se može predstaviti u obliku

$$a_n y_{n-1} + b_n y_n + c_n y_{n+1} = h^2 f_n \quad (n = 1, 2, \ldots, N). \tag{7.27}$$

S obzirom da su konturni uslovi $y_0 = A$ i $y_{N+1} = B$, pred nama se postavlja problem rešavanja sistema linearnih jednačina $T\overrightarrow{y} = \overrightarrow{d}$, gde su

$$\overrightarrow{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \overrightarrow{d} = \begin{bmatrix} h^2 f_1 - A a_1 \\ h^2 f_2 \\ \vdots \\ h^2 f_N - B c_N \end{bmatrix}, \quad T = \begin{bmatrix} b_1 & c_1 & 0 & \ldots & 0 \\ a_2 & b_2 & c_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & b_N \end{bmatrix}.$$

Matrica sistema je trodijagonalna. Za rešavanje ovog sistema pogodno je izvršiti dekompoziciju matrice T u obliku $T = LR$, čime se problem svodi na sukcesivno rešavanje dva trougaona sistema linearnih jednačina. Ovakav postupak za rešavanje konturnog problema (7.25) u literaturi je poznat kao matrična faktorizacija.

Primer 7.5. Izloženi postupak ima programsku realizaciju u, recimo, Fortran-u ili Matlab-u. U programu možemo predvideti i tabeliranje tačnog rešenja u tačkama u kojima određujemo rešenja pomoću ove metode. Međutim, jasno je da ovo poslednje ima smisla samo u školskim primerima, gde nam je rešenje poznato. Tako, ovde posmatramo konturni problem

$$y'' - 2xy' - 2y = -4x, \quad y(0) = 1,\ y(1) = 1 + e \cong 3.7182818,$$

čije je tačno rešenje $y = x + \exp(x^2)$. U slučaju $N = 4$ dobijeni su sledeći rezultati

i	0	1	2	3	4	5
x_i	0.00	0.20	0.40	0.60	0.80	1.00
y_i	1.000000	1.243670	1.577952	2.038018	2.699739	3.718282
y_i^T	1.000000	1.240811	1.573511	2.033330	2.696481	3.718282

Bibliografija

[1] I.S. Berezin, N.P. Žitkov, *Numerička analiza – numeričke metode*, Naučna knjiga, Beograd, 1963.

[2] W. Gautschi, *Numerical Analysis, An Introduction*, Birkhäuser, Boston·Basel·Berlin, 1997.

[3] W. Gautschi, *Orthogonal Polynomials, Computation and Approximation*, Oxford University Press, 2004.

[4] G.H. Golub, C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, 1983.

[5] B.S. Jovanović, *Numerička analiza*, PMF, Beograd, 1984.

[6] B.S. Jovanović, D. Radunović, *Numerička analiza*, Naučna knjiga, Beograd, 1993 (drugo izdanje: Matematički fakultet, Beograd, 2003.)

[7] G.V. Milovanović, *Numerička analiza, I deo*, Naučna knjiga, Beograd, 1985 (drugo izdanje 1988, treće izdanje 1991).

[8] G.V. Milovanović, *Numerička analiza, II deo*, Naučna knjiga, Beograd, 1985 (drugo izdanje 1988, treće izdanje 1991).

[9] G.V. Milovanović, *Numerička analiza, III deo*, Naučna knjiga, Beograd, 1988 (drugo izdanje 1991).

[10] G.V. Milovanović, M.A. Kovačević, M.M. Spalević, *Numerička matematika – zbirka rešenih problema,* Univerzitet u Nišu, Elektronski fakultet, 2003.