Dragoslav Herceg Đorđe Herceg

NUMERIČKA MATEMATIKA

SYMBOL

Novi Sad, 2009.

Autori: Dr Dragoslav Herceg,
redovni profesor Prirodno-matematičkog fakulteta u Novom Sadu
Dr Đorđe Herceg,
vanredni profesor Prirodno-matematičkog fakulteta u Novom Sadu

Recenzenti: Dr Nataša Krejić, redovni profesor
Prirodno-matematičkog fakulteta u Novom Sadu
Dr Zorana Lužanin, redovni profesor
Prirodno-matematičkog fakulteta u Novom Sadu

Izdavač: "Symbol", Novi Sad

Za izdavača: Veselin Stefanović

Štampa: "Symbol", Novi Sad

Štampano u 150 primeraka

Predgovor

Ovaj udžbenik numeričke matematike rađen je prema programu predmeta Numerička analiza za studente matematike i informatike Prirodno-matematičkog fakulteta. Namenjen je prvenstveno studentima, ali može korisno poslužiti i učenicima srednjih škola, u kojima se uči numerička matematika.

Za razumevanje sadržine ovog udžbenika dovoljno je poznavanje gradiva iz analize, algebre i linearne algebre, sa kojima su učenici upoznati u okviru ovih predmeta.

Dodatak B je posvećen mašinskim brojevima i odgovarajućoj aritmetici. Uglavnom je informativnog karaktera i trebalo bi da posluži upoznavanju čitalaca sa ovom oblašću.

Većina teorema data je bez dokaza. Dokazi koji su dati uz druge teoreme mogu se izostaviti pri prvom čitanju. Oni služe za upoznavanje sa idejama koje su dovele do određenih numeričkih postupaka.

Tekst sadrži veći broj rešenih primera, slika, tabela i zadataka koji ilustruju izloženo gradivo i olakšavaju njegovo usvajanje.

Na korisnim sugestijama, primedbama i pažljivom čitanju teksta naše knjige zahvaljujemo se recenzentima dr Nataši Krejić i dr Zorani Lužanin, redovnim profesorima Prirodno-matematičkog fakulteta u Novom Sadu.

Novi Sad, februar 2009.

Autori

Sadržaj

Glava 1

Uvod

1.1 Numerička matematika

Matematika se može posmatrati kao disciplina koja na temelju datog sistema aksioma (za svaku oblast drugog) sa datom logikom podiže deduktivnu građevinu. U klasičnoj matematici taj se postupak sprovodi bez povezivanja sa praksom. Klasična matematika bavi se samo pitanjem koliko daleko mogu da nose njene aksiome. U primenjenoj matematici posmatraju se matematički modeli problema iz prakse i razvijaju se postupci za njihovo rešavanje. Znači, klasična i primenjena matematika razlikuju se samo po povodu za rešavanje određenih problema.

Numerička matematika bavi se rešavanjem i postupcima (metodama) za rešavanje numeričkih problema. Pri tome se matematički problem smatra numeričkim ako se određivanje njegovog rešenja sastoji iz obrade brojčanih podataka.

Termini "numerička matematika" i "numerička analiza" javljaju se uglavnom kao sinonimi. Razvoj numeričke matematike doveo je do njene podele na više oblasti. Dve osnovne oblasti su numerički postupci analize i numerički postupci algebre.

Za mnoge matematičke probleme može se dokazati da imaju rešenje, čak da je ono jedinstveno, ali se ne mogu navesti postupci za određivanje tog rešenja. Sa gledišta klasične matematike dokaz postojanja rešenja često se smatra konačnim rezultatom. Međutim, u primenjenoj matematici ili tehničkim disciplinama to nije dovoljno. Traži se i rešenje. Rešenje može biti broj, na primer kod izračunavanja vrednosti funkcije, ili $n-$torka, kao kod rešavanja sistema linearnih jednačina sa n nepoznatih, ili funkcija, na primer kod određivanja položaja klatna posle određenog vremenskog perioda.

Ako se rešava sistem n linearnih jednačina sa n nepoznatih, koji ima jedinstveno rešenje, može se koristiti postupak determinanata, Kramerov postupak, za izračunavanje rešenja tog sistema. Ali već za $n = 3$ to je dosta komplikovano. Potrebno je, naime, 48 množenja i 3 deljenja, kada se vrednosti determinanata izračunavaju prema definiciji. Ako se operacije množenja i deljenja posmatraju kao operacije približno istog vremenskog trajanja i označe sa "množenje", onda je u opštem slučaju za rešavanje sistema n linearnih jednačina sa n nepoznatih postupkom determinanata broj "množenja"

$$(n^2 - 1)n! + n.$$

Ako se operacije sabiranja i oduzimanja posmatraju kao operacije približno istog vremenskog trajanja i označe sa "sabiranje", onda je u opštem slučaju za rešavanje sistema n linearnih jednačina sa n nepoznatih postupkom determinanata broj "sabiranja"

$$(n + 1)(n! - 1).$$

Rešavajući isti sistem linearnih jednačina Gausovim postupkom eliminacije broj operacija "množenja" je

$$\frac{n(n^2 + 3n - 1)}{3},$$

a broj "sabiranja" je

$$\frac{n(2n^2+3n-5)}{6}.$$

Za $n=3$ broj "množenja" kod Gausovog postupka eliminacije je 17. Već za $n=3$ upoređivanje broja operacija "množenja" opravdava numerički Gausov postupak rešavanja sistema linearnih jednačina. Ovo se još bolje vidi u slučaju $n=4$ gde je broj "množenja" kod postupka determinanata 364, a kod Gausovog postupka svega 36. Drugim rečima, iako se postupkom determinanata može rešiti sistem linearnih jednačina, postoji potreba za drugim postupcima koji će brže rešiti zadati problem. Ovaj primer i mnogi drugi, od kojih ćemo neke kasnije navesti, ukazuju na potrebu razvijanja numeričkih postupaka za rešavanje numeričkih problema.

1.2 Zadaci i postupci numeričke matematike

Osnovni zadatak numeričke matematike je dobijanje numeričkog rešenja određenog problema. Postupak za njegovo dobijanje naziva se numerički postupak ili numerička metoda.

Da bi se jedan problem mogao numerički rešiti mora biti korektan. To znači, za zadate brojeve i uslove, koji se zajedno nazivaju polazni podaci, problem treba da ima jedinstveno rešenje. Pored toga rešenje treba da bude stabilno, tj. da male promene polaznih podataka imaju za posledicu malu promenu rešenja.

Primer 1.1. *Rešenje sistema linearnih jednačina*

$$\begin{aligned} 3x_1 + x_2 &= -3 \\ 0.9998x_1 + \frac{1}{3}x_2 &= -0.9994 \end{aligned}$$

je $x_1 = -3$, $x_2 = 6$. *Ako u drugoj jednačini navedenog sistema umesto* $\frac{1}{3}$ *napišemo* 0.3333, *dobićemo sistem*

$$\begin{aligned} 3x_1 + x_2 &= -3 \\ 0.9998x_1 + 0.3333x_2 &= -0.9994 \end{aligned}$$

čije je rešenje $x_1 = -5$, $x_2 = 12$. *Promena u rešenju posmatranog problema može se smatrati velikom u odnosu na promenu polaznog podatka, pa ovaj problem nije stabilan.*

Očigledno, pojam stabilnosti je relativan. Zavisi od toga kako se mere promene polaznih podataka i rešenja.

Mnogi postupci numeričke matematike sastoje se iz višestrukog ponavljanja istih operacija, čiji se rezultati korak po korak približavaju traženom rešenju. Takvi postupci se nazivaju iterativni postupci. Iz mnogih razloga, koji se u svakom posebnom slučaju navode, dobija se samo numeričko rešenje koje odstupa od tačnog rešenja. Zbog toga je cilj numeričkog postupka dobijanje približnog rešenja koje će što manje odstupati od tačnog rešenja.

Tok numeričkog postupka opisuje se pomoću algoritma.

Definicija 1.1. ***Algoritam.*** *Algoritam je konačan skup tačno opisanih uputstava koja, uz korišćenje polaznih podataka, treba sprovoditi utvrđenim redom jedno za drugim da bi se dobilo rešenje nekog problema.*

U opštem slučaju za rešavanje postavljenog problema postoje različiti postupci. Da bi se izabrao najpogodniji, potrebno je formirati kriterijum za izbor. Takođe je važno poznavati uslove pod kojima se jedan numerički postupak može primeniti na rešavanje datog problema. Na primer, treba znati

uslove koje moraju da zadovoljavaju konstante, parametri ili promenljive da bi postupak konvergirao, odnosno, da bi dao približno rešenje traženog problema sa zadatom tačnošću.

Potrebno je, zatim, oceniti grešku nastalu aproksimacijom tačnog rešenja polaznog problema rešenjem problema koji polazni problem zamenjuje u efektivnom računanju. Brzina konvergencije postupka je takođe veoma interesantna. Često se daje prednost komplikovanijem ali bržem postupku, u odnosu na jednostavan i spor. Ne sme se, u opštem slučaju, smatrati da će brzina računara nadoknaditi sporost postupka.

Pri izboru numeričkog postupka zahteva se određena ekono–mičnost tog postupka u odnosu na matematički postupak i u odnosu na primenjeno pomoćno sredstvo u računanju. Kao mera ekonomično–sti u odnosu na matematički postupak može poslužiti broj operacija (sabiranje, množenje,...) ili, kada je postupak komplikovaniji, brzina njegove konvergencije. Jasno je da je bolji onaj postupak koji ima manji broj operacija ili koji brže konvergira.

Možemo reći da su prvi koraci u matematici bili numerički vezani za prebrojavanja i merenja. Mnogi slavni matematičari bavili su se numeričkim rešavanjem određenih problema, o čemu svedoče nazivi pojedinih postupaka. Tako imamo postupke Euklida, Ojlera, Njutna, Gausa,....

Potreba za numeričkim rešavanjem zadataka bila je prisutna uvek, naročito kod inženjera i naučnika iz oblasti prirodnih i tehničkih nauka. Međutim, tek poslednjih pedesetak godina numerička matematika počinje da se brže razvija i dobija poseban značaj. Zasluga za to dobrim delom pripada razvoju računara - kompjutera. Od početka razvoja numeričke matematike iz vremena Ojlera, oko 1750. godine, do 1940. godine produktivnost računskih sredstava jedva je postala deset puta veća. Od pojave prvog računara sa programskim upravljanjem (1941. godine) produktivnost računskih sredstava veoma je brzo rasla. Sada je nekoliko milijardi puta veća nego pre pojave računara.

Računari mnogo pomažu za brzo i tačno računanje, ali je potreba za novim numeričkim postupcima itekako prisutna. Velike mogućnosti primene matematike uslovile su matematizaciju mnogih oblasti u hemiji, ekonomiji, biologiji, medicini, geologiji, geografiji, psihologiji itd.

U nekim granama nauke, kao što što su fizika ili mehanika, dati su mnogi matematički modeli za opisivanje prirodnih pojava i njihovo proučavanje radi objašnjavanja starih i otkrivanja novih efekata. Kao primer uspešnog proučavanja matematičkog modela i rezultata dobijenih posmatranjem može da posluži Leverijeovo otkriće, do tada nepoznate planete, Neptuna.

Na osnovu navedenog može se zaključiti da je prisutna potreba za razradom postojećih postupaka i njihovim prilagođavanjem radu na računaru kao i potreba za razvojem i formiranjem novih postupaka. Sve češće se postavljaju zadaci određivanja postupaka važnih za veoma uske klase problema.

Široka i uspešna primena u gotovo svim tehničkim i društvenim naukama doprinosi današnjem ugledu numeričke matematike.

1.3 Zadaci

1.1. *Izračunaj za $n = 5, 10, 100$ potreban broj "množenja" i "sabiranja" za rešavanje sistema n linearnih jednačina sa n nepoznatih Kramerovim postupkom determinanata.*

1.2. *Izračunaj za $n = 5, 10, 100$ potreban broj "množenja" i "sabiranja" za rešavanje sistema n linearnih jednačina sa n nepoznatih Gausovim postupkom eliminacije.*

1.3. *Reši sisteme*

$$\begin{aligned} 3x_1 + x_2 &= -3 \\ 0.9998x_1 + \frac{1}{3}x_2 &= -0.9994 \end{aligned}$$

$$\begin{aligned} 3x_1 + x_2 &= -3 \\ 0.9998x_1 + 0.33x_2 &= -0.9994 \end{aligned}$$

$$\begin{aligned} 3x_1 + x_2 &= -3 \\ 0.9998x_1 + 0.333x_2 &= -0.9994 \end{aligned}$$

$$\begin{aligned} 3x_1 + x_2 &= -3 \\ 0.9998x_1 + 0.333333x_2 &= -0.9994 \end{aligned}$$

Kramerovim postupkom i uporedi rešenja.

1.4. *Poznato je da važi*

$$1 - \cos x = 2\sin^2\left(\frac{x}{2}\right).$$

Koristeći funkciju $N[f]$ *izračunaj u Mathematica-i*

$$1 - \cos 1^\circ \quad i \quad 2\sin^2\left(\frac{1^\circ}{2}\right)$$

sa 20 decimalnih mesta i uporedi rezultate.

1.5. *Rešenja kvadratne jednačine* $ax^2 + bx + c = 0$ *su*

$$x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}, \qquad x_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}.$$

Ova rešenja pod pretpostavkom da je $-b + \sqrt{b^2 - 4ac} \neq 0$ *i* $-b - \sqrt{b^2 - 4ac} \neq 0$*, možemo zapisati i na sledeći način*

$$x_3 = \frac{2c}{-b + \sqrt{b^2 - 4ac}}, \qquad x_4 = \frac{2c}{-b - \sqrt{b^2 - 4ac}}.$$

Ako je $b = 1234567,\ a = 0.001,\ c = 0.002$ *izračunaj* $x_1,\ x_2,\ x_3$ *i* x_4 *i uporedi dobijene vrednosti. Ponovi prethodno računanje sa* $b = -1234567,\ a = 0.001,\ c = 0.002$.

1.6. *Površina trougla čije su stranice* a, b *i* c *je*

$$P_1 = \sqrt{s(s-a)(s-b)(s-c)},$$

gde je

$$s = \frac{a+b+c}{2}.$$

Površinu trougla je takođe

$$P_2 = \frac{\sqrt{(a+(b+c))(c-(a-b))(c+(a-b))(a+(b-c))}}{4}.$$

Ako je $a = 9.0$, $b = c = 4.5000001$, *izračunaj u Mathematica-i* $N[P_1, 20]$, $N[P_2, 20]$ *i* $N[P_1 - P_2, 20]$. *Uporedi dobijene rezultate.*

Glava 2

Približni brojevi i greške funkcije

2.1 Približni brojevi

2.1.1 Pojam približnog broja i izvori grešaka

Pri rešavanju numeričkih problema radimo sa različitim brojevima, koji mogu biti tačni ili približni. Tačni brojevi daju pravu vrednost veličine, na primer, kvadrat ima 4 stranice i 4 ugla, knjiga ima 159 stranica itd. Sa druge strane, brojevi koji se dobijaju merenjem su približni. Na primer, dužina automobila je 4.16 m, masa praznog automobila je 960 kg itd. Neki brojevi mogu biti i tačni i približni. Na primer, broj 3 je tačan kao broj uglova trougla, ali je približan za broj π.

Rešenje numeričkog problema je u opštem slučaju približna vrednost x^* tražene veličine x. Šta više, sve aritmetičke operacije sa realnim brojevima se izvode tako što se koriste približni brojevi (ako nisu u pitanju realni brojevi koji se mogu predstaviti tačno sa konačno mnogo cifara, na primer 2.0). Pri tome se koriste **zaokruživanje** i **odsecanje**. Zaokruživanje na k cifara u brojnom sistemu sa osnovom 10 se vrši tako što se $k-$ta cifra poveća za 1 ukoliko je $(k+1)-$va cifra veća od 5, a ukoliko je $(k+1)-$va cifra manja od 5 onda se $k-$ta cifra ne menja. U graničnom slučaju, kada je $(k+1)-$va cifra jednaka 5, $k-$ta cifra povećava se za 1 ukoliko je neparna, a ako je parna, onda se ne menja. Odsecanjem na k cifara se sve cifre desno od $k-$te jednostavno zanemare. Na opisani način se svaki realan broj može predstaviti kao približan broj sa k cifara. Zaokruživanje i odsecanje se analogno definišu i u drugim brojnim sistemima. Za podatke koji su dobijeni merenjem je jasno da je njihova tačnost ograničena preciznošću mernog instrumenta, te oni takođe predstavljaju približne brojeve.

Pod približnim brojem x^* podrazumeva se broj koji se neznatno razlikuje od tačnog broja x i zamenjuje ga u računanjima. Kako nije potpuno jasno šta se podrazumeva pod "neznatno se razlikuje", to je potrebno dati još neke kvalitativne odnose koji će doprineti boljem upoznavanju pojma približnog broja. O tome će biti više reči kasnije.

Rešavanje velikog broja praktičnih problema odvija se u sledeće dve etape:

- formiranje matematičkog opisa posmatranog problema,
- rešavanje dobijenog matematičkog problema.

U prvoj etapi srećemo se sa dva osnovna izvora grešaka. Prvi je nemogućnost da se realni procesi opišu matematički tačno, bez "idealizacije". Naime, pri opisu pojave uprošćava se situacija tako što se manje važni podaci zanemaruju ili se posmatraju specijalni uslovi. Na primer, da bi se dobila jednačina koja opisuje kretanje klatna, najpre se umesto realnog klatna posmatra matematičko, a zatim se kao jednostavnija situacija posmatra ona u kojoj koeficijent trenja sredine u kojoj se nalazi klatno linearno zavisi od brzine klatna (što inače nije slučaj). Slične situacije nastupaju i kada se u opisu fizičkih, hemijskih i drugih procesa zanemaruje otpor vazduha, gasovi posmatraju kao idealni i slično.

Drugi izvor grešaka je netačnost početnih podataka koji se najčešće dobijaju iz eksperimenata. Greške matematičkog modela i greške početnih podataka nazivaju se greške polaznih informacija. Po svom karakteru ove greške su neotklonjive i njih je moguće smanjiti samo boljim merenjem, ali se, u opštem slučaju, nikako ne mogu izbeći.

Dobijanje tačnog rešenja matematičkog problema često nije moguće, čak ni kada se dokaže da postoji i da je jedinstveno. Zbog toga se formiraju numerički modeli ili numerički algoritmi, čija rešenja su aproksimacije rešenja matematičkih problema. Pri ovoj zameni tačnog rešenja matematičkog problema numeričkim rešenjem javlja se greška koja se naziva greška postupka ili greška aproksimacije.

U procesu izračunavanja rešenja matematičkog problema, tj. pri realizaciji numeričkog algoritma, bez obzira da li se radi sa računarima ili bez njih, javljaju se računske greške. Ove greške nastaju zbog

- rada sa konačnim decimalnim brojevima, tj. zaokružuju se brojevi sa kojima se radi,
- grešaka koje se javljaju pri izvođenju aritmetičkih operacija sa približnim brojevima,
- grešaka sa kojima se računaju vrednosti funkcija,
- grešaka koje nastaju izvođenjem pseudoaritmetičkih operacija, koje zamenjuju aritmetičke operacije kada se radi sa računarima.

Da bi se za postavljen problem dobilo upotrebljivo rešenje, potrebno je sve navedene greške poznavati bar toliko da znamo njihov red veličine.

Na osnovu rečenog, ukupna greška, tj. razlika između rešenja $\tilde{x}$ realnog problema i računarskog rešenja x^*, može se prikazati kao

$$\tilde{x} - x^* = \tilde{x} - x + x - x_n + x_n - x^*,$$

gde je x rešenje matematičkog modela, a x_n je tačno rešenje numeričkog problema.

Na osnovu relacije trougla, sledi

$$|\tilde{x} - x^*| \leq |\tilde{x} - x| + |x - x_n| + |x_n - x^*|\,.$$

Neotklonjiva greška $\tilde{x} - x$ nastaje pri formiranju matematičkog modela za realni problem i nju u daljem radu nećemo posmatrati.

Greška postupka $x - x_n$ i njena ocena su predmet proučavanja numeričke matematike (numeričke analize) ili, preciznije, analize numeričkih postupaka. U sledećim poglavljima pri opisu postupaka i algoritama posebno ćemo se baviti njihovim greškama.

Računska greška $x_n - x^*$ zavisi kako od primene računara tako i od računarskog algoritma. Naime, za jedan matematički algoritam moguće je formirati više računarskih algoritama, od kojih neki mogu biti dobro uslovljeni i davati dobra računarska rešenja, a drugi mogu biti loše uslovljeni i davati loša računarska rešenja. Pored toga, računarski algoritmi mogu se razlikovati i po ekonomičnosti (vremenu izvršenja, memorijskim zahtevima itd.).

Pri rešavanju konkretnih problema obično se kao cilj postavlja dobijanje rezultata sa određenom tačnošću, odnosno sa greškom manjom od unapred datog broja. Zbog toga je pri radu sa približnim brojevima potrebno:

- dati matematičke karakteristike tačnosti približnih brojeva,
- oceniti tačnost rezultata, kada je poznata tačnost početnih podataka,
- odrediti tačnost početnih podataka da bi se obezbedila željena tačnost rezultata,
- uskladiti tačnost početnih podataka da bi se izbegao nepotreban rad oko određivanja ili izračunavanja jednih podataka, ako su drugi podaci dati suviše grubo,
- održavati tačnost u procesu računanja, stalno kontrolisati međurezultate da bi se krajnji rezultat dobio sa željenom tačnošću i, u isto vreme, uprošćavati računanje.

2.1.2 Apsolutna i relativna greška

Numeričko rešavanje nekog problema u opštem slučaju daje samo jednu približnu vrednost x^* umesto tačnog broja x. Uzroke te pojave upoznali smo u prethodnom paragrafu. Određivanje odstupanja približne vrednosti x^* od tražene vrednosti x jedan je od važnih zadataka numeričke matematike. Za opis ovog odstupanja uvedimo nekoliko veličina koje ga karakterišu.

Definicija 2.1. ***Apsolutna greška.*** *Neka je x^* približna vrednost broja x. Razlika*

$$\Delta(x^*) = x - x^*$$

je ***greška*** *približnog broja x^*, a njena apsolutna vrednost*

$$|\Delta(x^*)| = |x - x^*|$$

je ***apsolutna greška*** *približnog broja x^*.*

U većini slučajeva broj x je nepoznat, pa se ne mogu odrediti greška i apsolutna greška približnog broja x^*. Zbog toga se određuje što je moguće manji broj Δ_{x^*}, za koji važi

$$|\Delta(x^*)| = |x - x^*| \leq \Delta_{x^*}.$$

Definicija 2.2. ***Granica apsolutne greške.*** *Neka je $|\Delta(x^*)|$ apsolutna greška približnog broja x^*. Svaki broj Δ_{x^*} koji zadovoljava nejednakost*

$$|\Delta(x^*)| \leq \Delta_{x^*}$$

naziva se ***granica apsolutne greške*** *približnog broja x^*.*

Primer 2.1. *Neka je $x = \sqrt{2}$, $x^* = 1.414$. Tada je*

$$|\Delta(x^*)| = \left|\sqrt{2} - 1.414\right| = 0.0002135\ldots.$$

Za granicu apsolutne greške može se uzeti

$$\Delta_{x^*} = 0.00022.$$

Ako je poznata granica apsolutne greške Δ_{x^*} približnog broja x^*, onda je $|x - x^*| \leq \Delta_{x^*}$, odnosno

$$x^* - \Delta_{x^*} \leq x \leq x^* + \Delta_{x^*}.$$

Da je x^* približna vrednost broja x sa granicom apsolutne greške Δ_{x^*}, često se zapisuje na sledeći način

$$x = x^* \pm \Delta_{x^*}.$$

Primer 2.2. *Zapisi $x = 1.127 \pm 0.005$ i $x = 1.127 \pm 5 \cdot 10^{-3}$ označavaju da je*

$$1.122 = 1.127 - 0.005 \leq x \leq 1.127 + 0.005 = 1.132.$$

Primer 2.3. *Neka je $x^* = 1.414 \approx \sqrt{2}$. Tada je granica apsolutne greške*

$$\Delta_{x^*} = 0.00022.$$

Sledeća tri zapisa su ekvivalentna

$$x^* = 1.414, \quad |\Delta(x^*)| \leq 0.22 \cdot 10^{-3},$$

$$x = 1.414 \pm 0.22 \cdot 10^{-3},$$

$$1.41378 \leq x \leq 1.41422.$$

Ukoliko su podaci sa kojima se radi različitog reda veličine, apsolutna greška nije pogodna za upoređivanje njihove tačnosti, pa se zato definiše relativna greška.

Pri nekom merenju određene veličine x najčešće se dobija samo približna vrednost x^*. Za upoređivanje različitih merenja apsolutna greška nije pogodna, što vidimo iz sledećeg primera. Ako su izmerene dve dužine $x = 15\ cm$ i $y = 15\ m$ sa istom apsolutnom greškom $|\Delta(x^*)| = |\Delta(y^*)| = 5\ mm$, onda je kvalitet drugog merenja očigledno bolji. Uvođenjem relativne greške omogućeno je upoređivanje različitih merenja, a tačnijim merenjem smatra se ono čija je relativna greška manja.

Definicija 2.3. *Relativna greška. Neka je $|\Delta(x^*)|$ apsolutna greška približnog broja $x^* \neq 0$. Količnik*

$$\delta(x^*) = \frac{|\Delta(x^*)|}{|x^*|} = \frac{|x - x^*|}{|x^*|}$$

*je **relativna greška** približnog broja x^*.*

Kao i apsolutna greška, tako i relativna greška ne može da se odredi tačno. Zbog toga uvodimo pojam granice relativne greške.

Definicija 2.4. ***Granica relativne greške.*** *Neka je $\delta(x^*)$ relativna greška približnog broja x^*. Svaki broj δ_{x^*} za koji važi*

$$\delta(x^*) \leq \delta_{x^*}$$

*naziva se **granica relativne greške** približnog broja x^*.*

Poznavajući granicu apsolutne greške Δ_{x^*} približnog broja $x^* \neq 0$, može se odrediti granica relativne greške istog broja. Naime, kako je

$$\delta(x^*) = \frac{|\Delta(x^*)|}{|x^*|} = \frac{|x - x^*|}{|x^*|} \leq \frac{\Delta_{x^*}}{|x^*|},$$

za granicu relativne greške može se uzeti

$$\delta_{x^*} = \frac{\Delta_{x^*}}{|x^*|}.$$

Č esto se granica relativne greške množi sa 100 i daje u procentima (%). Tako nastala veličina naziva se **procentualna greška** približnog broja x^*

$$\delta_{x^*} = 100\delta_{x^*}\%.$$

Ako je x^* približna vrednost broja x sa granicom relativne greške δ_{x^*}, pisaćemo

$$x = x^*(1 \pm \delta_{x^*}).$$

Primer 2.4. *Za broj $x = \pi$ je $x^* = 3.14$ približna vrednost. Neka su $\Delta_{x^*} = 0.002$ i $\Delta_{x^*} = 0.0016$ dve granice apsolutne greške. Tada važi*

$$3.138 = 3.14 - 0.002 \leq \pi \leq 3.14 + 0.002 = 3.142$$

i

$$3.1384 = 3.14 - 0.0016 \leq \pi \leq 3.14 + 0.0016 = 3.1416.$$

Imamo sledeće dve granice relativne greške

$$\delta_{x^*} = \frac{0.002}{3.14} = 0.000636943\ldots$$

$$\delta_{x^*} = \frac{0.0016}{3.14} = 0.000509554\ldots.$$

Kako treba da važi $\delta(x^*) \leq \delta_{x^*}$, uzimamo da je u prvom slučaju $\delta_{x^*} = 0.00064$, a u drugom $\delta_{x^*} = 0.00051$. Za procentualne greške dobijamo $\delta_{x^*} = 0.064\%$, odnosno $\delta_{x^*} = 0.051\%$.

Primer 2.5. *Zapisi*

$$x = 1.127(1 \pm 0.005) = 1.127(1 \pm 5 \cdot 10^{-3})$$

i

$$x = 1.127(1 \pm 0.5\%)$$

označavaju da je

$$1.121365 = 1.127(1 - 0.005) \leq x \leq 1.127(1 + 0.005) = 1.132635.$$

Često se umesto o određivanju granice apsolutnih i relativnih grešaka približnih brojeva govori o tačnosti, odnosno greškama merenja ili računanja. U tom smislu izraze "sa tačnošću do 0.01" ili "sa greškom do 0.1%" smatraćemo za sinonime sa izrazima "sa granicom apsolutne greške ne većom od 0.01", odnosno "sa granicom relativne greške ne većom od 0.001".

2.1.3 Značajne i sigurne cifre

Kod računara broj se predstavlja, slično decimalnom predstavljanju, nizom diskretnih fizičkih veličina. Svakoj cifri odgovara jedna vrednost fizičke veličine. Kako decimalni sistem ima deset različitih cifara, to je veoma lako predstaviti sve cifre pomoću različitih vrednosti fizičke veličine. Pri tome nema problema sa tačnošću predstavljanja jedne cifre, jer vrednost fizičke veličine iz jednog intervala može odgovarati jednoj cifri. Na primer, ako cifri 8 odgovara napon od 8 V, onda se za tu cifru može tolerisati i napon iz intervala $[7.6\ V, 8.4\ V]$. Na taj način tačnost digitalnih računara nije ograničena fizičkom tačnošću merenja.

Iz tehničkih razloga računari ne koriste nama dobro poznat decimalni (dekadni) sistem predstavljanja brojeva.

Neka je dat prirodan broj $b \geq 2$, koji nazivamo baza. Pomoću cifara koje su elementi skupa $\{0, 1, \ldots, b-1\}$ svaki realan broj x se može predstaviti u obliku

$$x = \pm \sum_{i=-\infty}^{r} a_i b^i = \pm \left(a_r b^r + a_{r-1} b^{r-1} + \cdots + a_0 b^0 + a_{-1} b^{-1} + \cdots \right)$$

gde je $a_i \in \{0, 1, \ldots, b-1\}, i = r, r-1, \ldots$. Ovaj zapis nazivamo $b-$arni zapis broja x. Za $b = 2$ imamo binarni zapis, sa $b = 8$ oktalni, sa $b = 16$ heksadecimalni i sa $b = 10$ decimalni zapis.

Poznato je da svaki realan broj ima bar jedan $b-$arni zapis. Prema tome sledi da pojedini realni brojevi mogu imati i više od jednog $b-$arnog zapisa. Tako, u decimalnom zapisu, broj 5 možemo zapisati i kao $4.9999\ldots$ sa beskonačno mnogo cifara 9. Da bi se ovakve situacije izbegle, u daljem radu posmatraćemo samo zapise sa konačno mnogo cifara. To je potpuno prirodno, jer smo pri radu sa računarima ili bez računara, prinuđeni da koristimo samo brojeve sa konačno mnogo cifara.

U daljem radu koristimo samo decimalni zapis brojeva. Zapis realnog broja x može se prikazati kao

$$x = \pm a_r a_{r-1} \ldots a_0 . a_{-1} a_{-2} \ldots,$$

gde je tačka iza a_0 decimalna tačka. Cifre a_i nazivaju se još i mesta. Ako su sve cifre a_{-i}, $i = 1, 2, \ldots$ jednake nuli, broj x je ceo broj. Ako je za neko $i \geq 1$ cifra a_{-i} različita od nule, broj je **razlomljen**. Mesta desno od decimalne tačke nazivaju se **decimalna mesta** ili **decimale** .

Ako počev od nekog $m \geq 1$ važi $a_{-i} = 0$, $i > m$, decimalni razlomak je **konačan** . U suprotnom slučaju je **beskonačan.**

Definicija 2.5. ***Važeće cifre.*** *Sve cifre decimalnog zapisa počev od prve cifre sleva različite od nule nazivaju se važeće (značajne) cifre.*

Nule na kraju broja uvek su značajne cifre (u protivnom ih ne pišemo). Brojevi 0.01235, 0.0123500, 12345700 imaju 4, 6 i 8 značajnih cifara respektivno.

Pri zapisu celih brojeva može doći do nesporazuma. Tako, ako broju $x = 300000$ tri poslednje nule nisu značajne cifre, dati broj treba zapisati u obliku $x = 300 \cdot 10^3$ ili $x = 3.00 \cdot 10^5$ ili $x = 0.300 \cdot 10^6$. Na taj način su sve cifre broja x, koje su napisane, i značajne cifre.

U raznim računanjima pojavljuju se beskonačni decimalni razlomci, koje je potrebno zameniti konačnim, bez obzira da li se računanje izvodi sa ili bez pomoći računara. Znači, tačan broj x zamenjuje se nekim približnim brojem x^*. Da bi se informacija o apsolutnoj greški te aproksimacije dala istovremeno sa približnim brojem, uvodi se pojam sigurne cifre.

Definicija 2.6. ***Sigurne cifre.*** *Neka je*

$$x^* = \pm a_r a_{r-1} \dots a_0 . a_{-1} a_{-2} \dots a_{-k}$$

približna vrednost broja x. Kažemo da je cifra a_{r-m+1} ($m-$ta cifra sleva) broja x^* sigurna ako za granicu apsolutne greške važi

$$\Delta_{x^*} \leq \omega 10^{r-m+1}$$

za neko $\omega \in [0.5, 1]$.

Za cifre približnog broja koje nisu sigurne kažemo da su nesigurne. Očigledno, ako je neka cifra sigurna, sigurne su i sve cifre ispred (levo od) nje. Ako je $\omega = 0.5$, govorimo o sigurnim ciframa u **užem smislu**, a ako je $\omega = 1$, govorimo o sigurnim ciframa u **širem smislu**.

U daljem radu, ako se drugačije ne naglasi, smatraćemo da je poslednja važeća cifra približnog broja sigurna u užem smislu. To znači, da sa približnim brojem x^* dobijamo i granicu apsolutne greške Δ_{x^*}.

Primer 2.6. *Neka broj $x^* = 34.082928$ ima pet sigurnih cifara. Tada je*

$$\Delta_{x^*} = 0.5 \cdot 10^{1-5+1} = 0.0005$$

granica apsolutne greške.

Primer 2.7. *Neka je $x^* = 3.14$ približna vrednost broja $\pi = 3.14159265 \dots$. Kako je i*

$$|\pi - x^*| < 0.0016 \leq 0.5 \cdot 10^{-2},$$

sve tri važeće cifre broja x^* su sigurne. Ako bismo broj x^* zapisali kao 3.140 (a odlučili smo da svaki približan broj pišemo sa svim sigurnim ciframa), onda bi trebalo da važi

$$\Delta_{x^*} \leq 0.5 \cdot 10^{0-4+1} = 0.0005$$

što nije tačno. Opet su samo 3 važeće cifre i sigurne cifre.

Pri sastavljanju tablica sa vrednostima funkcija (logaritamske i trigonometrijske tablice) obično se uzima $\omega = 0.5$, a pri davanju rezultata merenja uzima se najčešće $\omega = 1$.

Ako približan broj x^* broja x ima m sigurnih cifara, to ne znači da je tih m cifara jednako odgovarajućim ciframa broja x. Na primer, neka je $x^* = 99.886$ i $x = 100$. Očigledno brojevi x^* i x nemaju jednakih cifara, ali kako je $\Delta_{x^*} = 0.014$ broj x^* ima tri sigurne cifre u užem smislu.

Pri praktičnom računanju sa približnim brojevima kada se koriste osnovne aritmetičke operacije, broj sigurnih cifara rezultata određuje se prema sledećim pravilima.

- Pri sabiranju i oduzimanju približnih brojeva rezultat zadržava najviše onoliko sigurnih cifara koliko ih ima približni broj sa najmanjim brojem sigurnih cifara.
- Kod množenja i deljenja približnih brojeva rezultat ima najviše onoliko sigurnih cifara koliko i broj sa najmanjim brojem sigurnih cifara (najčešće jednu ili dve sigurne cifre manje).

- Pri stepenovanju rezultat ima najviše onoliko sigurnih cifara koliko i broj koji se stepenuje.
- Pri korenovanju rezultat ima najviše onoliko sigurnih cifara koliko i potkorena veličina.

Ako se približni brojevi koji učestvuju u računanju mogu uzimati sa proizvoljnom tačnošću, da bi se dobio rezultat sa k sigurnih cifara, treba te brojeve birati tako da prema prethodnim pravilima obezbeđuju $k+1$ sigurnu cifru u rezultatu. U slučaju da neki brojevi imaju više važećih cifara od drugih, potrebno ih je zaokružiti pre računanja. Broj važećih cifara na koji se zaokružuju ti brojevi treba da je za jedan veći od broja važećih cifara broja sa najmanje važećih cifara.

2.2 Greške funkcije

Neka je potrebno izračunati vrednost funkcije $f(x)$ za realno x pomoću računara. Kao konačan rezultat dobićemo vrednost $\tilde{f}(\tilde{x})$ gde je $\tilde{x}$ mašinski broj na koji se redukcijom preslikao broj x, a $\tilde{f}$ je računarska aproksimacija funkcije f. Greška koja nas interesuje je $f(x) - \tilde{f}(\tilde{x})$. Na osnovu relacije

$$f(x) - \tilde{f}(\tilde{x}) = f(x) - f(\tilde{x}) + f(\tilde{x}) - \tilde{f}(\tilde{x})$$

dobijamo ocenu

$$\left| f(x) - \tilde{f}(\tilde{x}) \right| \leq |f(x) - f(\tilde{x})| + \left| f(\tilde{x}) - \tilde{f}(\tilde{x}) \right|.$$

Prvi sabirak na desnoj strani poslednje relacije je greška nastala zamenom argumenta x njegovom približnom vrednošću $\tilde{x}$. Drugi sabirak je greška nastala aproksimacijom funkcije f funkcijom $\tilde{f}$. Kako se u svakom računaru aproksimacije elementarnih funkcija, kao što su, $\sqrt{x}$, $\sin x$, $\cos x$, itd. dobijaju uz računar, to se obično uz računar dobija i uputstvo o dozvoljenoj oblasti argumenta tih funkcija i o tačnosti aproksimacije tih funkcija za tu oblast. Analiza greške aproksimacije složenijih funkcija može biti veoma komplikovana i time se nećemo više baviti. U svakom slučaju, greške elementarnih funkcija ograničene su sa $k \cdot eps$, gde je eps tačnost s kojom se izračunavaju aritmetičke operacije.

Računari daju veoma dobre aproksimacije elementarnih funkcija, pa je mnogo važnije znati oceniti greške nastale zamenom argumenata. Na primer, ako je $g(x) = \sqrt{e^x + 1}$, a $e^x + 1$ se izračunava pomoću računara kao y, onda približna vrednost y^* za $g(x)$ i tačnost te vrednosti zavisi od tačnosti vrednosti y. Sličan problem javlja se kada se računa vrednost funkcije više argumenta koji se zamenjuje približnim vrednostima. Na primer, zapremina valjka je $r^2\pi H$, gde je r poluprečnik, H visina valjka, a π je pri konkretnom računanju takođe potrebno zameniti približnim brojem, u zavisnosti od preciznosti zadavanja poluprečnika i visine. Ovde greške izvođenja operacije množenja ne mora ni postojati, ali se ipak javlja greška rezultata uzrokovana greškom argumenta.

U daljem radu posmatraćemo grešku rezultata neke funkcije pod pretpostavkom da se operacije koje se pojavljuju izvode egzaktno, bez novonastalih grešaka zbog prelaska na pseudooperacije.

Neka je tražena veličina

$$y = f(x_1, x_2, \ldots, x_n)$$

funkcija promenljivih $x_1, x_2, \ldots, x_n$ i neka je za svaku promenljivu x_i, $i = 1, 2, \ldots, n$, poznat interval G_i kojem pripada. Pretpostavimo da je potrebno izračunati približnu vrednost y^* za y i oceniti njenu grešku. Ovakav zadatak naziva se **direktan problem greške**.

Određivanje granica apsolutnih grešaka sa kojima je potrebno uzeti približne vrednosti promenljivih da bi se postigla zadata tačnost vrednosti funkcije naziva se **obrnut problem**.

2.2.1 Direktan problem

Granice apsolutne i relativne greške približne vrednosti y^* za y definišemo na isti način kao i granice greške približnog broja x^*.

Definicija 2.7. ***Granica apsolutne greške.*** *Neka je y^* približna vrednost za $y = f(x_1, x_2, \ldots, x_n)$. Svaki broj $A(y^*)$ za koji važi*

$$|y - y^*| \leq A(y^*), \qquad za\ x_i \in G_i,\ i = 0, 1, \ldots, n,$$

nazivamo granica apsolutne greške približne vrednosti y^* funkcije $y = f(x_1, x_2, \ldots, x_n)$.

Definicija 2.8. ***Granica relativne greške.*** *Neka je $y^* \neq 0$ približna vrednost funkcije $y = y(x_1, x_2, \ldots, x_n)$. Svaki broj $R(y^*)$ za koji važi*

$$\frac{|y - y^*|}{|y^*|} \leq R(y^*), \qquad za\ x_i \in G_i,\ i = 0, 1, \ldots, n,$$

nazivamo granica relativne greška približne vrednosti y^* funkcije y.

Očigledno, ako je poznato $A(y^*)$ i $y^* \neq 0$ možemo uzeti

$$R(y^*) = \frac{A(y^*)}{|y^*|},$$

što ćemo činiti u daljem radu. Ukoliko je $A^0(y^*)$ aproksimacija za $A(y^*)$, onda se količnik

$$R^0(y^*) = \frac{A^0(y^*)}{|y^*|}$$

uzima kao aproksimacija za $R(y^*)$.

Neka su poznate približne vrednosti x_i^* argumenata x_i sa granicama apsolutnih greška $\Delta_{x_i^*}$. Za približnu vrednost y^* uzimamo $y^* = f(x_1^*, x_2^*, \ldots, x_n^*)$, a intervali G_i se definišu na sledeći način

$$G_i = \left\{x : |x - x_i^*| \leq \Delta_{x_i^*}\right\}, \qquad i = 1, 2, \ldots, n.$$

U daljem radu smatraćemo da su y^* i G_i određeni upravo tako.

U sledećem primeru pokazaćemo kako se može oceniti greška približne vrednosti funkcije **metodom granica.**

Primer 2.8. *Potrebno je izračunati približnu vrednost y^* funkcije*

$$y = f(x_1, x_2, x_3) = \frac{x_1 + x_2^2}{x_3},$$

granicu apsolutne greške $A(y^*)$ i granicu relativne greške $R(y^*)$, pri čemu je

$$x_1^* = 3.25, \quad \Delta_{x_1^*} = 0.03, \quad x_2^* = 1.34, \quad \Delta_{x_2^*} = 0.008, \quad x_3^* = 2.11, \quad \Delta_{x_3^*} = 0.02.$$

Intervali G_i su određeni sa

$$G_1 = [3.22, 3.28], \quad G_2 = [1.332, 1.438], \quad G_3 = [2.09, 2.13].$$

Funkcija y je u posmatranim intervalima rastuća po x_1 i x_2 a opadajuća po x_3. Za $x_1 \in G_1$, $x_2 \in G_2$ i $x_3 \in G_3$ važi

$$y_d = y(3.22, 1.332, 2.13) \leq y = y(x_1, x_2, x_3) \leq y(3.28, 1.348, 2.09) = y_g.$$

Neposrednim računanjem dobijamo

$$y_d = \frac{3.22 + 1.332^2}{2.13} > 2.3447, \quad y_g = \frac{3.28 + 1.348^2}{2.09} < 2.4390, \quad y^* = \frac{3.25 + 1.34^2}{2.11} \approx 2.3913.$$

Znači,

$$2.3447 < y < 2.4390,$$

odnosno

$$|y - y^*| \leq \max\{y_g - y^*, y^* - y_d\} < \max\{0.0466, 0.0478\} = 0.0478.$$

Za granicu apsolutne greške uzimamo $A(y^*) = 0.0478$, *a za granicu relativne greške* $R(y^*) = 0.02$, *jer je*

$$\frac{A(y^*)}{|y^*|} < \frac{0.0478}{2.3912} < 0.02.$$

Da bi odredili broj koji je veći od i broj koji je manji od y, koristili smo činjenicu da funkcija y raste po x_1 i x_2 a opada po x_3. Takođe smo sve međurezultate korigovali tako da budu obezbeđene tražene nejednakosti i da $y_g - y_d$ bude što manje.

Princip ocene apsolutne greške funkcije izložen u prethodnom primeru može se koristiti u relativno malom broju slučajeva. Takođe, granica apsolutne greške dobijena na taj način obično je gruba. Bolju granicu apsolutne greške možemo dobiti primenjujući Lagranžovu teoremu.

Funkcija jedne promenljive

Teorema 2.1. ***Lagranžova teorema.*** *Ako je funkcija f definisana i neprekidna na intervalu $[a, b]$ i ima izvod u svakoj tački intervala $[a, b]$, onda postoji tačka $\theta \in (a, b)$ takva da je*

$$f(a) - f(b) = f'(\theta)(b - a).$$

Tačka θ je nepoznata, zna se samo da postoji i da pripada intervalu (a, b). Kao neposredna posledica Lagranžove teoreme dobija se

$$|f(a) - f(b)| = |f'(\theta)\|b - a| \leq B|b - a|,$$

gde je konstanta B određena tako da važi

$$B \geq \max\{|f'(x)| \ : x \in [a, b]\}.$$

Kao posledicu Lagranžove teoreme možemo dobiti granicu apsolutne greške približne vrednosti $y^* = f(x^*)$ funkcije $y = f(x)$. Neka je data približna vrednost x^* za x, njena granica apsolutne greške Δ_{x^*} i neka je G interval definisan sa

$$G = \{x : x^* - \Delta_{x^*} \leq x \leq x^* + \Delta_{x^*}\}.$$

Ako je funkcija f definisana i neprekidna na intervalu G i ima izvod u svakoj njegovoj tački, onda za svako $x \in G$ postoji neko $\tau \in G$ takvo da važi

$$|y - y^*| = |f(x) - f(x^*)| = |f'(\tau)|\,|x - x^*| \leq B|x - x^*|,$$

gde je B određeno tako da važi

$$B \geq \max\{|f'(x)| \ : x \in G\}.$$

Za svako $x \in G$ važi $|x - x^*| \leq \Delta_{x^*}$ i

$$|f(x) - f(x^*)| \leq B\Delta_{x^*}.$$

Broj B posmatramo kao granicu apsolutne greške $A(y^*)$ približne vrednosti y^* funkcije f. Kada je određivanje konstante B komplikovano, umesto nje uzima se konstanta

$$b = |f'(x^*)| \approx B,$$

a kao aproksimacija granice apsolutne greške koristi se

$$A^0(y^*) = b\Delta_{x^*}.$$

Važno je napomenuti da nije uvek $|f(x) - f(x^*)| \leq A^0(y^*)$ za svako $x \in G$, što pokazuje i sledeći primer.

Primer 2.9. *Neka je* $y = x^{10}$, $x^* = 1$ *i* $\Delta_{x^*} = 0.1$. *Tada je* $y^* = 1$ *i* $G = [0.9, 1.1]$. *Direktnim računanjem dobijamo*

$$|y - y^*| = \left|x^{10} - 1\right| \leq 1.1^{10} - 1 < 1.59375.$$

Dalje je

$$B = \max\{|y'(x)| : x \in G\} = 10 \cdot 1.1^9 < 23.58,$$
$$b = |y'(x^*)| = 10,$$
$$A(y^*) = 23.56 \cdot 0.1 = 2.358,$$

i

$$A^0(y^*) = 10 \cdot 0.1 = 1.$$

Vidimo da je broj $A(y^*)$ *granica apsolutne greške približne vrednosti posmatrane funkcije u datom intervalu, a broj* $A^0(y^*)$ *nije.*

Broj $A^0(y^*)$ nije uvek granica apsolutne greške, već je samo linearna ocena granice apsolutne greške. Iako može biti $A^0(y^*) < |y - y^*|$ ipak se linearna ocena često koristi kao aproksimacija granice apsolutne greške, naročito u slučajevima kada je granica apsolutne greške promenljive mnogo manja od 1.

Primer 2.10. *Neka je* $y = \ln x$, $G = \{x : x^* - \Delta_{x^*} \leq x \leq x^* + \Delta_{x^*}\}$ *i* $x^* - \Delta_{x^*} > 0$. *Tada je* $b = \frac{1}{x^*}$ *i*

$$A^0(y^*) = b\Delta_{x^*} = \frac{\Delta_{x^*}}{x^*} = \delta_{x^*}.$$

Na osnovu linearne ocene granice apsolutne greške dobija se sledeća aproksimacija granice relativne greške

$$R^0(y^*) = \frac{A^0(y^*)}{|y^*|} = \frac{\Delta_{x^*}}{|y^*|x^*} = \frac{\delta_{x^*}}{|y^*|}.$$

Funkcija više promenljivih

Neka su poznate približne vrednosti x_i^* argumenata x_i i njihove granice apsolutnih greška $\Delta_{x_i^*}$. Uzimamo da je $y = f(x_1, x_2, \ldots, x_n)$, $y^* = f(x_1^*, x_2^*, \ldots, x_n^*)$ i da su intervali G_i definisani sa

$$G_i = \left\{x : |x_i - x_i^*| \leq \Delta_{x_i^*}\right\}, \qquad i = 1, 2, \ldots, n.$$

Ocenu za

$$|y - y^*| = |f(x_1, x_2, \ldots, x_n) - f(x_1^*, x_2^*, \ldots, x_n^*)|$$

možemo dobiti na osnovu Lagranžove teoreme za funkcije više promenljivih.

Prvi izvod funkcije $y = f(x_1, x_2, \ldots, x_n)$ po promenljivoj x_j, dok ostale promenljive posmatramo kao konstante, označavamo sa

$$\frac{\partial f}{\partial x_j}(x_1, x_2, \ldots, x_n).$$

Teorema 2.2. ***Lagranžova teorema za funkciju više promenljivih.*** *Neka funkcija* $f(x_1, x_2, \ldots, x_n)$ *ima neprekidne parcijalne izvode* $\frac{\partial f}{\partial x_j}(x_1, x_2, \ldots, x_n)$, $j = 1, 2, \ldots n$, *za* $x_j \in G_j$, $j = 1, 2, \ldots, n$. *Tada postoje* $\tau_j \in G_j$, $j = 1, 2, \ldots, n$, *takvi da je*

$$f(x_1, x_2, \ldots, x_n) - f(x_1^*, x_2^*, \ldots, x_n^*) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\tau_1, \tau_2, \ldots, \tau_n)\left(x_j - x_j^*\right).$$

Kako su $\tau_j \in G_j$, $j = 1, 2, \ldots, n$, nepoznati brojevi, granica apsolutne greške funkcije određujemo pomoću relacije

$$\left| \sum_{j=1}^{n} \frac{\partial f}{\partial x_j} (\tau_1, \tau_2, \ldots, \tau_n) \left(x_j - x_j^* \right) \right| \leq \sum_{j=1}^{n} \left| \frac{\partial f}{\partial x_j} (\tau_1, \tau_2, \ldots, \tau_n) \right| \Delta_{x_j^*}$$

i

$$B_j = \max_{x \in G} \left| \frac{\partial f}{\partial x_j} (x_1, x_2, \ldots, x_n) \right|, \quad j = 1, 2, \ldots, n.$$

Naime, važi

$$|f(x_1, x_2, \ldots, x_n) - f(x_1^*, x_2^*, \ldots, x_n^*)| \leq \sum_{j=1}^{n} B_j \Delta_{x_j^*}.$$

Na osnovu ovoga za granicu apsolutne greške funkcije može se uzeti

$$A(y^*) = \sum_{j=1}^{n} B_j \Delta_{x_j^*}.$$

2.2.2 Linearne ocene granica apsolutne i relativne greške

Kako određivanje brojeva B_i može biti dosta komplikovano, u praksi se često koristi aproksimacija

$$B_i \approx b_i = \left| \frac{\partial y}{\partial x_i} (x_1^*, x_2^*, \ldots, x_n^*) \right|, \quad i = 1, 2, \ldots, n.$$

Za apsolutnu grešku funkcije tada važi

$$|y(x_1, x_2, \ldots, x_n) - y(x_1^*, x_2^*, \ldots, x_n^*)| \lesssim \sum_{i=1}^{n} b_i \Delta_{x_i^*} = A^0(y^*),$$

pri čemu znak „$\lesssim$" treba čitati kao "manje ili približno jednako". Veličina $A^0(y^*)$ se naziva **linearna ocena granice** apsolutne greške i može biti nesigurna, ali je mnogo jednostavnija za primenu. Po istom principu se definiše i linearna ocena za relativnu grešku funkcije, ako je $y^* \neq 0$,

$$R^0(y^*) = \frac{A^0(y^*)}{|y^*|}.$$

Primer 2.11. *Potrebno je izračunati približnu vrednost y^* funkcije*

$$y = f(x_1, x_2, x_3) = \frac{x_1 + x_2^2}{x_3},$$

granicu apsolutne greške $A(y^*)$ i granicu relativne greške $R(y^*)$, ako je

$$x_1^* = 3.25, \quad \Delta_{x_1^*} = 0.03, \quad x_2^* = 1.34, \quad \Delta_{x_2^*} = 0.008, \quad x_3^* = 2.11, \quad \Delta_{x_3^*} = 0.02.$$

Intervali G_i su određena sa

$$G_1 = [3.22, 3.28], \quad G_2 = [1.332, 1.348], \quad G_3 = [2.09, 2.13].$$

Funkcija y ima sledeće parcijalne izvode

$$\frac{\partial f}{\partial x_1} = \frac{1}{x_3}, \quad \frac{\partial f}{\partial x_2} = \frac{2x_2}{x_3}, \quad \frac{\partial f}{\partial x_3} = -\frac{x_1 + x_2^2}{x_3^2}.$$

Na osnovu ovih izvoda dobijamo

$$\begin{aligned}
B_1 &= \max\left\{\left|\frac{1}{x_3}\right| : 2.09 \le x_3 \le 2.13\right\} = \frac{1}{2.09} < 0.48,\\
B_2 &= \max\left\{\left|\frac{2x_2}{x_3}\right| : 2.09 \le x_3 \le 2.13,\ 1.332 \le x_2 \le 1.348\right\} = \frac{2\cdot 1.348}{2.09} < 1.29,\\
B_3 &= \max\left\{\left|-\frac{x_1+x_2^2}{x_3^2}\right| : 2.09 \le x_3 \le 2.13,\ 1.332 \le x_2 \le 1.348,\ 3.22 \le x_1 \le 3.28\right\}\\
&= \frac{3.28+1.348^2}{2.09^2} < 1.17.
\end{aligned}$$

Dalje je

$$\begin{aligned}
A(y^*) &= B_1\Delta_{x_1^*} + B_2\Delta_{x_2^*} + B_3\Delta_{x_3^*}\\
&= 0.48\cdot 0.03 + 1.29\cdot 0.008 + 1.17\cdot 0.02 < 0.04813
\end{aligned}$$

$$R(y^*) = \frac{A(y^*)}{|y^*|} = \frac{0.04813}{2.39128} < 0.021.$$

Linearne ocene granica apsolutne i relativne greške su

$$A^0(y^*) = b_1\Delta_{x_1^*} + b_2\Delta_{x_2^*} + b_3\Delta_{x_3^*} < 0.04707$$

$$R^0(y^*) = \frac{A^0(y^*)}{|y^*|} < \frac{0.04707}{2.39128} < 0.0197,$$

jer je

$$\begin{aligned}
b_1 &= \left|\frac{1}{x_3^*}\right| = \frac{1}{2.11} < 0.474,\\
b_2 &= \left|\frac{2x_2^*}{x_3^*}\right| = \frac{2\cdot 1.34}{2.11} < 1.271,\\
b_3 &= \left|-\frac{x_1^*+x_3^{*2}}{x_3^{*2}}\right| = \frac{3.25+1.34^2}{2.11^2} < 1.134.
\end{aligned}$$

Za jednostavnije funkcije granice apsolutnih i relativnih grešaka mogu se dobiti na osnovu prethodnih razmatranja.

Teorema 2.3. *Neka je* $a_k \in \{-1, 1\}, k = 1, 2, \ldots, n$ *i*

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n.$$

Tada je

$$|y - y^*| \le \Delta_{x_1^*} + \Delta_{x_2^*} + \cdots + \Delta_{x_n^*}.$$

Dokaz. Parcijalni izvod unkcije y po x_k je a_k, pa imamo za $k = 1, 2, \ldots, n$

$$\frac{\partial y}{\partial x_k}(x_1, x_2, \ldots, x_n) = a_k,\ B_k = b_k = \left|\frac{\partial y}{\partial x_k}(x_1^*, x_2^*, \ldots, x_n^*)\right| = |a_k| = 1.$$

Na osnovu ovoga tvrđenje teoreme sledi neposredno. ■

Ponekad se dobija veoma velika granica relativne greške, iako su granice relativnih grešaka promenljivih male. U tom slučaju kažemo da je došlo do gubitka tačnosti.

Primer 2.12. *Neka je* $y = x_1 - x_2$, $x_1^* = 13.456 \pm 0.0005$, $x_2^* = 13.451 \pm 0.0005$. *Tada je prema prethodnoj teoremi*

$$A(y^*) = 0.0005 + 0.0005 = 0.001.$$

Granice relativnih grešaka promenljivih su

$$\delta_{x_1^*} = 0.00004 > \frac{0.0005}{13.456}, \qquad \delta_{x_2^*} = 0.00004 > \frac{0.0005}{13.451},$$

a granica relativne greške funkcije je

$$R(y^*) = \frac{A(y^*)}{|x_1^* - x_2^*|} = \frac{0.001}{0.005} = 0.2.$$

Vidimo da je greška približne vrednosti funkcije 5000 *puta veća od granica relativnih grešaka promenljivih.*

Navedeni primer ukazuje na potrebu izbegavanja oduzimanja dvaju bliskih brojeva.

U sledećim teoremama granice relativnih grešaka uzimamo kao količnik granice apsolutne greške i apsolutne vrednosti približnog broja (različitog od nule).

Teorema 2.4. *Neka je* $x_k > 0,\ k = 1, 2, \ldots, n$ *i*

$$y = x_1 + x_2 + \cdots + x_n.$$

Tada je

$$\min\left\{\delta_{x_1^*}, \delta_{x_2^*}, \ldots, \delta_{x_n^*}\right\} \le R(y^*) \le \max\left\{\delta_{x_1^*}, \delta_{x_2^*}, \ldots, \delta_{x_n^*}\right\}.$$

Dokaz. Kako je

$$\delta_{x_k^*} = \frac{\Delta_{x_k^*}}{x_k^*}, \qquad \Delta_{x_k^*} = x_k^* \delta_{x_k^*}$$

i

$$|y - y^*| \le A(y^*) = \Delta_{x_1^*} + \Delta_{x_2^*} + \cdots + \Delta_{x_n^*} = \sum_{k=1}^{n} x_k^* \delta_{x_k^*},$$

dobijamo

$$\min\left\{\delta_{x_1^*}, \delta_{x_2^*}, \ldots, \delta_{x_n^*}\right\} \sum_{k=1}^{n} x_k^* \le A(y^*) = \sum_{k=1}^{n} x_k^* \delta_{x_k^*} \le \max\left\{\delta_{x_1^*}, \delta_{x_2^*}, \ldots, \delta_{x_n^*}\right\} \sum_{k=1}^{n} x_k^*.$$

Iz ove relacije tvrđenje teoreme sledi jednostavno, imajući u vidu da je $y^* = x_1^* + x_2^* + \cdots + x_n^*$,

$$\min\left\{\delta_{x_1^*}, \delta_{x_2^*}, \ldots, \delta_{x_n^*}\right\} \le R(y^*) = \frac{A(y^*)}{y^*} \le \max\left\{\delta_{x_1^*}, \delta_{x_2^*}, \ldots, \delta_{x_n^*}\right\}.$$

■

Koristeći linearnu ocenu granice apsolutne greške dobijaju se sledeći rezultati:

- Neka je $y = x_1 \cdot x_2 \cdot \cdots \cdot x_n$, $x_k^* \neq 0, k = 1, 2, \ldots, n$. Tada je

$$A^0(y^*) = |y^*| \sum_{k=1}^{n} \delta_{x_k^*}, \qquad R^0(y^*) = \sum_{k=1}^{n} \delta_{x_k^*}.$$

- Za $y = \dfrac{x_1}{x_2}$, $x_1^* x_2^* \neq 0$, dobijamo

$$A^0(y^*) = \left|\frac{x_1^*}{x_2^*}\right| \left(\delta_{x_1^*} + \delta_{x_2^*}\right), \qquad R^0(y^*) = \left(\delta_{x_1^*} + \delta_{x_2^*}\right).$$

- Za $y = x^m$, $x^* \neq 0$ m je prirodan broj dobijamo

$$A^0(y^*) = m|x^*|^m \delta_{x^*}, \qquad R^0(y^*) = m\delta_{x^*}.$$

- Za $y = \sqrt[m]{x}$, $x^* > 0$, m je prirodan broj, dobijamo

$$A^0(y^*) = \frac{1}{m}\sqrt[m]{x^*}\delta_{x^*}, \qquad R^0(y^*) = \frac{1}{m}\delta_{x^*}.$$

Prethodne relacije daju linearne ocene granice apsolutne greške i granice relativne greške za proizvod približnih brojeva, količnik dva približna broja, stepen i koren približnog broja.

Primer 2.13. *Neka približni brojevi $x_1^* = 3.72$ i $x_2^* = 4.37$ imaju sve cifre sigurne u užem smislu. Izračunajmo*

a) $y^* = x_1^* x_2^*$,

b) $y^* = \frac{x_1^*}{x_2^*}$,

c) $y^* = (x_2^*)^4$,

d) $y^* = \sqrt[5]{x_1^*}$ *i odgovarajuće granice apsolutnih i relativnih grešaka. Kako su sve cifre približnih brojeva sigurne, dobijamo*

$$\begin{aligned} \Delta_{x_1^*} &= \Delta_{x_2^*} \leq \frac{1}{2}10^{0-3+1} = 0.005, \\ \delta_{x_1^*} &= 0.00135 > \frac{0.005}{3.72}, \\ \delta_{x_2^*} &= 0.00115 > \frac{0.005}{4.37}. \end{aligned}$$

Na osnovu ranije izvedenih relacija dobijamo

a) $y^* = 16.2564$, $A^0(y^*) = y^*(0.00135 + 0.00115) = 0.40641$, $R(y^*) = 0.00250$.

b) $y^* = 0.851259$, $A^0(y^*) = 0.00211814$, $R(y^*) = 0.00250$.

c) $y^* = 364.692$, $A^0(y^*) = 4 \cdot 0.00115 \cdot y^* = 1.67758$, $R(y^*) = 0.0046$.

d) $y^* = 1.30049$, $A^0(y^*) = \frac{1}{5}0.00135 \cdot y^* = 0.000351134$, $R(y^*) = 0.00027$.

2.2.3 Obrnut problem

Do sada je na osnovu granica apsolutnih grešaka promenljivih određivana granica (ili njena aproksimacija) apsolutne greške funkcije. Za primenjenu matematiku takođe je važan problem određivanja granica grešaka argumenata funkcije tako da granica apsolutne greške funkcije ne premaši zadatu veličinu ili da joj je približno jednaka.

Taj zadatak nije jedinstveno rešiv, jer zadatu granicu apsolutne greške funkcije možemo dobiti sa različitim granicama apsolutnih grešaka njenih argumenata. Na primer, ako je

$$y = x_1 + x_2, \qquad y = x_1^* + x_2^*$$

onda je

$$|y - y^*| \leq \Delta_{x_1^*} + \Delta_{x_2^*}.$$

Očigledno, uslov

$$|y - y^*| \leq \varepsilon,$$

za neko pozitivno ε, ispunjavaju sve nenegativne vrednosti $\Delta_{x_1^*}$ i $\Delta_{x_2^*}$ za koje važi $\Delta_{x_1^*} + \Delta_{x_2^*} \leq \varepsilon$. Tako je za $\varepsilon = 0.05$ prethodni uslov ispunjen i za $\Delta_{x_1^*} = 0.02$ i $\Delta_{x_2^*} = 0.03$, kao i za $\Delta_{x_1^*} = \Delta_{x_2^*} = 0.025$ i za bezbroj drugih izbora za $\Delta_{x_1^*}$ i $\Delta_{x_2^*}$.

Problem postaje jednostavniji ako se uvedu još neke pretpostavke i ograničenja. Tako ćemo, kao prvo, pretpostaviti da je potrebno odrediti granice apsolutnih grešaka argumenata funkcije tako da linearna ocena granice apsolutne greške bude manja od unapred datog pozitivnog broja ε. Dakle, ako je $y^* = f(x_1^*, x_2^*, \ldots, x_n^*)$ približna vrednost funkcije $y = f(x_1, x_2, \ldots, x_n)$ i ako je $A^0(y^*)$ linearna ocena granice apsolutne greške približne vrednosti y^*, onda je potrebno odrediti granice apsolutnih grešaka približnih vrednosti $\Delta_{x_1^*}$, $\Delta_{x_2^*}, \ldots, \Delta_{x_n^*}$. tako da važi

$$A^0(y^*) \leq \varepsilon.$$

Videli smo da ni ovako postavljen problem nije jednoznačno rešiv, pa u daljem radu uvodimo nove pretpostavke, da bi ovaj problem bio jednoznačno rešiv. U zavisnosti od novih pretpostavki razlikuju se sledeća tri slučaja.

Princip jednakih doprinosa

Pretpostavlja se da svaki od sabiraka $b_1\Delta_{x_1^*}$, $b_2\Delta_{x_2^*}, \ldots, b_n\Delta_{x_n^*}$, jednako utiče na $A^0(y^*)$, tj. da važi

$$b_1\Delta_{x_1^*} = b_2\Delta_{x_2^*} = \cdots = b_n\Delta_{x_n^*}.$$

U ovom slučaju važi

$$A^0(y^*) = b_1\Delta_{x_1^*} + b_2\Delta_{x_2^*} + \cdots + b_n\Delta_{x_n^*} = nb_k\Delta_{x_k^*}$$

za svako $k = 1, 2, \ldots, n$. Očigledno, ako se granice apsolutnih grešaka odrede tako da važi

$$\Delta_{x_k^*} \leq \frac{\varepsilon}{nb_k}, \quad k = 1, 2, \ldots, n,$$

dobićemo da je

$$A^0(y^*) \leq \varepsilon.$$

Primer 2.14. *Poluprečnik veće osnove zarubljene kupe je $R \approx 2$ m, poluprečnik manje osnove je $r \approx 1.5$ m, a visina je $H \approx 3.5$ m. Odredićemo sa kakvim granicama apsolutnih grešaka je potrebno odrediti poluprečnike osnova kupe i njenu visinu, da bi se zapremina zarubljene kupe izračunala sa granicom apsolutne greške $A(y^*) \approx 0.1$ m^3. Zapremina zarubljene kupe je funkcija četiri argumenta $x_1 = R$, $x_2 = r$, $x_3 = H$, i $x_4 = \pi$:*

$$V = \frac{\pi}{3}H\left(R^2 + Rr + r^2\right).$$

Sada je

$$x_1^* = R^* = 2 \text{ m}, \quad x_2^* = r^* = 1.5 \text{ m}, \quad x_3^* = H^* = 3.5 \text{ m}, \quad x_4^* = \pi^* = 3.14,$$

$$b_1 = \frac{1}{3}H^*\left(r^* + 2R^*\right)\pi^* < 20.15, \qquad b_2 = \frac{1}{3}H^*\left(2r^* + R^*\right)\pi^* < 18.32,$$

$$b_3 = \frac{1}{3}\left(r^{*2} + R^*r^* + R^{*2}\right)\pi^* < 9.69, \quad b_4 = \frac{1}{3}H^*\left(r^{*2} + R^*r^* + R^{*2}\right) < 10.80.$$

Na osnovu principa jednakih doprinosa dobijamo

$$\Delta_{x_1^*} \leq \frac{0.1}{4b_1} < 0.001241, \quad \Delta_{x_2^*} \leq \frac{0.1}{4b_2} < 0.001365,$$

$$\Delta_{x_3^*} \leq \frac{0.1}{4b_3} < 0.002580, \quad \Delta_{x_4^*} \leq \frac{0.1}{4b_4} < 0.002315.$$

Kao što se vidi poluprečnike osnova zarubljene kupe i njenu visinu treba odredit veoma precizno, sa greškama za koje važi

$$|\Delta(R^*)| < 1.241\ mm, \quad |\Delta(r^*)| < 1.365\ mm, \quad |\Delta(H^*)| < 2.580\ mm,$$

Približnu vrednost π^* *za* π *treba uzeti tako da je* $|\Delta(\pi^*)| < 0.002315$. *Kako je* $0 < \pi - 3.14 < 0.0016$, *očigledno možemo uzeti* $\pi^* = 3.14$. *Približna vrednost zarubljene kupe sa datim vrednostima je* $V^* = 33.8858\ m^3$. *Ako navedene veličine izmerimo sa zahtevanom preciznošću, a za približnu vrednost broja* π *uzmemo* 3.14, *dobićemo da približna vrednost zapremine zarubljene kupe odstupa od tačne vrednosti zapremine približno za* $0.1\ m^3$, *odnosno za* 100 *litara.*

U prethodnom primeru broj π smo posmatrali kao promenljivu. To je opravdano, jer u računanju uvek koristimo neku njegovu aproksimaciju koja ima konačno mnogo decimala. Isto tako, i sve druge brojeve, na primer, $\sqrt{2}$, e, $\sqrt{3}$ koje u računanjima zamenjujemo konačnim decimalnim razlomcima, treba posmatrati kao promenljive.

Princip jednakih granica apsolutnih grešaka

Pretpostavljamo da su granice apsolutnih grešaka $\Delta_{x_1^*}, \Delta_{x_2^*}, \ldots, \Delta_{x_n^*}$ jednake, tj. da važi

$$\Delta_{x_1^*} = \Delta_{x_2^*} = \cdots = \Delta_{x_n^*}.$$

U ovom slučaju važi

$$A^0(y^*) = \Delta_{x_k^*}(b_1 + b_2 + \cdots + b_n)$$

za svako $k = 1, 2, \ldots, n$. Očigledno, ako se granice apsolutnih grešaka odrede tako da važi

$$\Delta_{x_k^*} \leq \frac{\varepsilon}{b_1 + b_2 + \cdots + b_n}, \quad k = 1, 2, \ldots, n,$$

dobićemo da je

$$A^0(y^*) \leq \varepsilon.$$

Primer 2.15. *Rešavajući problem iz prethodnog primera, po principu jednakih granica apsolutnih grešaka dobijamo*

$$|\Delta(R^*)| = |\Delta(r^*)| = |\Delta(H^*)| \leq \frac{0.1}{58.96}\ m < 0.001697\ m = 1.697\ mm.$$

Približnu vrednost π^* *za* π *treba uzeti tako da je* $|\Delta(\pi^*)| < 0.001697$. *Kako je* $0 < \pi - 3.14 < 0.0016$, *možemo uzeti* $\pi^* = 3.14$.

Princip jednakih granica relativnih grešaka

Pretpostavljamo da su granice relativnih grešaka $\delta_{x_1^*}, \delta_{x_2^*}, \ldots, \delta_{x_n^*}$ svih argumenata jednake, tj. da važi

$$\delta_{x_1^*} = \delta_{x_2^*} = \cdots = \delta_{x_n^*},$$

da je $x_1^* \neq 0$, $x_2^* \neq 0$, ..., $x_n^* \neq 0$ i da su granice relativnih grešaka određene sa

$$\delta_{x_k^*} = \frac{\Delta_{x_k^*}}{|x_k^*|}, \ k = 1, 2, \ldots, n.$$

U ovom slučaju važi

$$A^0(y^*) = \delta_{x_k^*}(b_1|x_1^*| + b_2|x_2^*| + \cdots + b_n|x_n^*|)$$

za svako $k = 1, 2, \ldots, n$. Očigledno, ako se granice apsolutnih grešaka odrede tako da važi

$$\Delta_{x_k^*} \leq \frac{\varepsilon\,|x_k^*|}{(b_1\,|x_1^*| + b_2|x_2^*| + \cdots + b_n|x_n^*|)},\ k = 1, 2, \ldots, n,$$

dobićemo da je

$$A^0\,(y^*) \leq \varepsilon.$$

Primer 2.16. *Rešavajući problem iz prethodna dva primera, po principu jednakih granica relativnih grešaka dobijamo*

$$|\Delta\,(R^*)| = |\Delta\,(r^*)| = |\Delta\,(H^*)| \leq \frac{0.1}{135.61}\ m < 0.0007375\ m = 0.7375\ mm.$$

Približnu vrednost π^* za π treba uzeti tako da je $|\Delta(\pi^*)| < 0.0007375$. Kako je $0 < \pi - 3.142 < 0.00041$, u ovom slučaju treba uzeti $\pi^* = 3.142$.

2.3 Zadaci

2.1. *Odredi procentualnu grešku približnog broja $x^* = 27.322$, ako je $\Delta_{x^*} = 0.00053.51\%$.*

2.2. *Nađi granicu apsolutne greške približnog broja $a = 8.321$, ako je $\Delta_a = 0.01\%$.*

2.3. *Zaokruži brojeve $a = 3.141593$, $b = 2.754649$, $c = 0.144851$ na 4, 3, 2 i 1 značajnu cifru.*

2.4. *Zaokruži brojeve $x = 1.1426$ i $y = 0.1245$ na tri značajne cifre i odredi procentualnu grešku dobijenih približnih brojeva x^* i y^*.*

2.5. *Odredi broj sigurnih cifara u užem i širem smislu sledećih približnih brojeva $a = 45.385 \pm 0.034$, $b = 1.2785 \pm 0.0006$, $c = 193.3 \pm 0.1$.*

2.6. *Neka su prvih m cifara približnog broja*

$$x^* = \pm\left(a_n 10^n + a_{n-1}10^{n-1} + \cdots + a_0 10^0 + a_{-1}10^{-1} + \cdots\right), \quad a_n \neq 0$$

sigurne cifre za dato $\omega \in [0.5, 1]$. Dokaži da je

$$\delta\,(x^*) \leq \frac{\omega}{a_n} 10^{1-m}.$$

2.7. *Neka je x^* približna vrednost realnog broja*

$$x = \pm x_0 \cdot 10^n, \quad 0.1 \leq x_0 < 1, \quad n \in \mathbb{Z}.$$

Ako je m najveći prirodan broj za koji važi

$$|x - x^*| \leq 0.5 \cdot 10^{n-m}.$$

onda su prvih m važećih cifara broja x^* sigurne cifre. Dokaži.

2.8. *Sa koliko sigurnih cifara u užem smislu treba uzeti približnu vrednost za $\sqrt{27}$ da bi procentualna greška bila manja od 0.1%?*

2.9. *Koristeći tablice trigonometrijskih funkcija sa 5 sigurnih cifara u užem smislu izračunaj približnu vrednost za $1 - cos1^\circ$ i oceni grešku.*

2.10. *Logaritamske tablice sa pet decimala sadrže dekadne logaritme brojeva sa tačnošću od $5 \cdot 10^{-6}$. Kolika je procentualna greška približnog broja x^* (uzetog iz tablice) za $\log x$ ako $x \in [300, 400]$?*

2.11. *Odredi broj sigurnih cifara u užem i širem smislu sledećih brojeva sa datim granicama njihovih apsolutnih grešaka:* $x^* = 0.3941$, $\Delta_{x^*} = 0.25 \cdot 10^{-2}$, $y^* = 38.2345$, $\Delta_{y^*} = 0.32 \cdot 10^{-2}$, $z^* = 0.00381$, $\Delta_{z^*} = 0.1 \cdot 10^{-4}$, $w^* = -0.2113$, $\Delta_{w^*} = 0.5 \cdot 10^{-2}$.

2.12. *Odredi broj sigurnih cifara u užem i širem smislu sledećih brojeva sa datim granicama njihovih relativnih grešaka:* $x^* = 9.8542$, $\delta_{x^*} = 0.1\%$, $y^* = 1.3452$, $\delta_{y^*} = 0.1 \cdot 10^{-2}$, $z^* = 0.3421$, $\delta_{z^*} = 0.2 \cdot 10^{-4}$, $w^* = 18.3742$, $\delta_{w^*} = 1\%$.

2.13. *Odredi šta je tačnije:* $\frac{6}{25} \approx \frac{1}{4}$ *ili* $\frac{1}{3} \approx 0.333$; $\frac{1}{9} \approx 0.1$ *ili* $\frac{1}{3} \approx 0.3$; $\pi \approx \frac{22}{7}$ *ili* $\pi \approx 3.14$; $\sqrt{10} \approx 3.1623$ *ili* $\frac{6}{7} \approx 0.86$.

2.14. *Izračunaj vrednosti sledećih izraza i oceniti tačnost rezultata, pri čemu svi brojevi sadrže samo sigurne cifre u užem smislu:*

$$a) \quad y = \frac{1.03 \cdot 43.7}{34.7 \cdot 40.2}; \qquad c) \quad y = \frac{37.421 - 8.31}{65.3 \cdot 101};$$

$$b) \quad y = \frac{47.2 \cdot 35.18}{13.7 + 14.2}; \qquad d) \quad y = \frac{34.2 + 18.3}{15.4 - 18.9}.$$

2.15. *Neka je* $y = \dfrac{\ln x_1 - x_2^2}{\sin x_3 + x_4}$. *Naći granicu apsolutne greške* $A(y^*)$ *i granicu relativne greške* $R(y^*)$, *ako je*

$$\begin{aligned} x_1^* &= 12.7, \quad \Delta_{x_1^*} = 0.03, \quad x_2^* = 1.43, \quad \Delta_{x_2^*} = 0.005, \\ x_3^* &= 0.72, \quad \Delta_{x_3^*} = 0.03, \quad x_4^* = 1.9, \quad \Delta_{x_4^*} = 0.07. \end{aligned}$$

2.16. *Izračunaj približnu vrednost za* $\ln(10.3 + \sqrt{4.4})$ *ako su sve cifre približnih brojeva* $x = 10.3$ *i* $y = 4.4$ *i sigurne u užem smislu.*

2.17. *Izračunaj približne vrednosti sledećih funkcija za datu vrednost promenljive* x *i odredi procentualnu grešku:*

$a)$ $y = x^3 \sin x$ *za* $x = \sqrt{2}$ *uzimajući* $\sqrt{2} \approx 1.414$

$b)$ $y = x \sin x$ *za* $x = \pi$ *uzimajući* $\pi \approx 3.142$

$c)$ $y = e^x \cos x$ *za* $x = \sqrt{3}$ *uzimajući* $\sqrt{3} \approx 1.732$

2.18. *Izračunaj* $\frac{1}{(2+\sqrt{3})^4}$ *koristeći se aproksimacijom za* $\sqrt{3}$. *Uporedi procentualne greške direktnog računanja i ekvivalentnog izraza* $97 - 56\sqrt{3}$.

2.19. *Izračunaj približnu vrednost sledećih funkcija za datu vrednost promenljivih. Odredi granice apsolutnih i relativnih grešaka ako su svi brojevi dati sa sigurnim ciframa u užem smislu:*

$a)$ $y = \ln\left(x_1 + x_2^2\right)$ $\quad x_1 = 0.97 \quad x_2 = 1.152$

$b)$ $y = \dfrac{(x_1 + \cos x_2)}{x_3}$ $\quad x_1 = 8.37 \quad x_2 = -2.11 \quad x_3 = 0.57$

$c)$ $y = x_1x_2 + x_3x_4 + x_2x_3$ $\quad x_1 = 1.05 \quad x_2 = 2.03 \quad x_3 = -0.7 \quad x_4 = -0.73$

$d)$ $y = 6x_1^2(\ln x_2 - \cos x_3)$ $\quad x_1 = 2.73 \quad x_2 = 4.73 \quad x_3 = 0.71$

2.20. *Odredi približnu vrednost funkcije i proceniti apsolutnu i relativnu grešku ako su sve cifre zadatih brojeva sigurne u užem smislu:*

a) $y = \ln\left(x_1 + x_2^2\right)$ $x_1 = 0.97$ $x_2 = 1.132$

b) $y = \dfrac{x_1 + x_2^2}{x_3}$ $x_1 = 3.28$ $x_2 = 0.932$ $x_3 = 1.132$

c) $y = x_1x_2 + x_1x_3 + x_2x_3$ $x_1 = 2.104$ $x_2 = 1.935$ $x_3 = 0.845$

2.21. *Sa koliko sigurnih cifara treba uzeti rezultate sledećih operacija:*

a) $x = \dfrac{1}{3}$; *d)* $x = \ln 13.7$; *g)* $x = e^{2.34}$;

b) $x = \sqrt{29}$; *e)* $x = 0.34^5$; *h)* $x = \operatorname{sh} 3.14$;

c) $x = \sqrt[3]{349}$; *f)* $x = \sin 1.3$;

tako da granica relativne greške rezultata ne bude veća od 0.1%*?*

2.22. *Sa koliko sigurnih cifara u užem smislu treba uzeti približnu vrednost za* x *da bi apsolutna greška približne vrednosti sledećih funkcija bila manja od* 10^{-6}*?*

a) $y = x^3 \sin x$ $x = \sqrt{2}$

b) $y = x \ln x$ $x = \pi$

c) $y = e^x \cos x$ $x = \sqrt{3}$

2.23. *Kod merenja za određivanje dužine budućeg mosta, na jednoj obali je izmerena baza, duž* AB*, dužine* $x = 200\ m \pm 0.01\ m$*. Izmereni su i uglovi,* α *i* β*, između baze i pravaca iz njenih krajeva kroz tačku* C *sa druge strane reke. Oni iznose* $\alpha = 80^\circ \pm 3'$ *i* $\beta = 60^\circ \pm 3'$*. S kolikom se tačnošću, na osnovu ovih podataka, može izračunati dužina visine trougla* ABC *iz temena* C *na bazu* AB*?*

2.24. *Neka brojevi* $x^* = 1.3134$ *i* $y^* = 0.3761$ *imaju sve cifre sigurne u užem smislu. Izračunaj sledeće proizvode tako da rezultati imaju tri sigurne cifre u užem smislu:*

a) $x\pi$ *b)* ye *c)* πe

2.25. *Reši obrnut problem za funkciju* $y = \ln x_1 + e^{x_2 + \sqrt{x_3}}$ *da bi se za* $x_1 \approx 1.93$, $x_2 \approx 0.341$, $x_3 \approx 12.506$ *približna vrednost funkcije dobila sa četiri sigurne cifre u užem smislu.*

2.26. *Odredi dozvoljene apsolutne greške argumenata koje omogućavaju da se vrednosti datih funkcija izračunavaju sa četiri sigurne cifre u užem smislu:*

a) $y = \ln\left(x_1 + x_2^2\right)$ $x_1 = 0.9731$ $x_2 = 1.13214$

b) $y = \dfrac{x_1 + x_2^2}{x_3}$ $x_1 = 3.2835$ $x_2 = 0.93221$ $x_3 = 1.13214$

c) $y = x_1x_2 + x_3x_4 + x_2x_3$ $x_1 = 2.10415$ $x_2 = 1.93521$ $x_3 = 0.84542$

Glava 3

Interpolacija

3.1 Egzistencija i greška interpolacionog polinoma

Neka je $\Phi(x; a_0, a_1, \ldots, a_n)$, funkcija koja zavisi od $n+1$ parametra $a_0, a_1, \ldots, a_n$. Problem **interpolacije** za funkciju Φ sastoji se u određivanju parametara a_i tako da za zadate tačke $(x_i, y_i),\ i = 0, 1, \ldots, n$, važi

$$\Phi(x_i; a_0, a_1, \ldots, a_n) = y_i, \quad i = 0, 1, \ldots, n.$$

Funkcija Φ se naziva **interpolaciona funkcija**, a ako je $y_i = f(x_i)$ kažemo da je Φ interpolaciona funkcija funkcije f, tj. Φ **interpolira** f.

Ako je Φ linearna funkcija u odnosu na parametre a_i

$$\Phi(x; a_0, a_1, \ldots, a_n) = a_0\Phi_0(x) + a_1\Phi_1(x) + \cdots + a_n\Phi_n(x),$$

govorimo o **linearnoj interpolaciji**. Funkcije $\Phi_i(x)$ su neke pogodno izabrane funkcije, odnosno pripadaju nekom **dopustivom skupu funkcija**. Za

$$\Phi_i(x) = x^i, \quad i = 0, 1, \ldots, n,$$

dobijamo **polinomnu interpolaciju**

$$\Phi(x; a_0, a_1, \ldots, a_n) = a_0 + a_1x + \cdots + a_nx^n,$$

a funkcija Φ se u ovom slučaju naziva interpolacioni polinom. U daljem radu govorićemo samo o polinomnoj interpolaciji.

Interpolacija se sastoji u izračunavanju približne vrednosti funkcije f za x različito od datih vrednosti $x_0, x_1, \ldots, x_n$, bez izračunavanja vrednosti $f(x)$. Ovo je potrebno naročito u slučajevima kada je funkcija f složena i komplikovana ili je čak nepoznata, a njene vrednosti u posmatranim tačkama dobijene su nekim postupkom, na primer merenjem. Da bi se odredila približna vrednost za $f(x)$, traži se neka druga funkcija koja je jednostavna za izračunavanje. Mi smo se opredelili da to bude polinom.

Brojeve $x_0, x_1, \ldots, x_n$ nazivamo **čvorovima interpolacije,** parove $(x_i, y_i),\ i = 0, 1, \ldots, n$, **čvornim tačkama**. Polinom

$$p_n(x) = a_0 + a_1x + \cdots + a_nx^n$$

sa osobinom

$$p_n(x_i) = y_i,\ i = 0, 1, \ldots, n,$$

nazivamo **interpolacioni polinom** za čvorne tačke $(x_i, y_i),\ i = 0, 1, \ldots, n$**.**

Ako je data funkcija f onda interpolacioni polinom za čvorne tačke $(x_i, f(x_i))$, $i = 0, 1, \ldots, n$, nazivamo **interpolacioni polinom funkcije** f i označavamo ga sa $P_n(f, x)$. Očigledno, za interpolacioni polinom funkcije f važi

$$P_n(f, x_i) = f(x_i), \ i = 0, 1, \ldots, n.$$

Za interpolaciju osnovna su sledeća tri pitanja:

- Da li postoji interpolacioni polinom za zadate čvorne tačke ?
- Ako postoji interpolacioni polinom, kako ga odrediti na osnovu datih podataka?
- Kako oceniti $|f(x) - P_n(f, x)|$, za $x \notin \{x_0, x_1, \ldots, x_n\}$?

Odgovor na treće pitanje možemo dati samo kada je funkcija f poznata i kada u posmatranom intervalu ima neprekidne sve izvode od prvog do $(n+1)-$og. U slučaju da imamo samo vrednosti funkcije f u čvorovima interpolacije može se naći samo približna ocena te veličine, ali time se nećemo baviti.

3.1.1 Egzistencija interpolacionog polinoma

U sledećoj teoremi dajemo uslove koji garantuju egzistenciju i jedinstvenost interpolacionog polinoma.

Teorema 3.1. ***Egzistencija interpolacionog polinoma.*** *Ako su x_0, x_1, ..., x_n međusobno različiti brojevi, onda za proizvoljne brojeve y_0, y_1, ..., y_n postoji jedan i samo jedan interpolacioni polinom za čvorne tačke (x_i, y_i), $i = 0, 1, \ldots, n$.*

Dokaz. Dokazaćemo da postoji jedan i samo jedan polinom $p_n(x) = a_0 + a_1 x + \cdots + a_n x^n$ za koji važi $p_n(x_i) = y_i$, $i = 0, 1, \ldots, n$, tj. dokazaćemo da se koeficijenti a_0, $a_1, \ldots,$ a_n polinoma p_n mogu odrediti na jedinstven način iz sistema

$$a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_n x_i^n = y_i, \quad i = 0, 1, \ldots, n.$$

Ovaj sistem od $n+1$ linearne jednačine sa $n+1$ nepoznatom a_i može se zapisati i u obliku

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^n \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix}.$$

Determinanta matrice ovog sistema, označimo je sa $V(x_0, x_1, \ldots, x_n)$ je Vandermondova determinanta

$$V(x_0, x_1, \ldots, x_n) = \prod_{i>j} (x_i - x_j) = \prod_{j=0}^{n-1} \left(\prod_{i=j+1}^{n} (x_i - x_j) \right).$$

Očigledno je

$$\begin{aligned} V(x_0, x_1, \ldots, x_n) \ = \ & (x_1 - x_0)(x_2 - x_0)(x_3 - x_0) \cdots (x_n - x_0) \cdot \\ & (x_2 - x_1)(x_3 - x_1) \cdots (x_n - x_1) \cdot \\ & (x_3 - x_2)(x_4 - x_2) \cdots (x_n - x_2) \cdot \\ & \vdots \\ & (x_{n-1} - x_{n-2})(x_n - x_{n-2}) \cdot \\ & (x_n - x_{n-1}) \end{aligned}$$

i $V(x_0, x_1, \ldots, x_n)$ je različito od nule, jer su po pretpostavci vrednosti x_i međusobno različite. To znači da ovaj sistem ima jedinstveno rešenje, odnosno da postoji jedinstven polinom p_n sa osobinom $p_n(x_i) = y_i, \;\; i = 0, 1, \ldots, n.$ ■

U daljem radu kada se govori o čvorovima interpolacije podrazumeva se da su oni međusobno različiti.

3.1.2 Greška interpolacije i njena ocena

Neka su za datu funkciju f i međusobno različitie brojeve x_0, x_1, ..., x_n poznate vrednosti $y_i = f(x_i)$, $i = 0, 1, \ldots, n$. Za interpolacioni polinom funkcije f važi

$$P_n(f, x_i) = f(x_i), \quad i = 0, 1, \ldots, n.$$

Za x različito od čvorova interpolacije grešku interpolacionog polinoma (grešku interpolacije) definišemo kao razliku funkcije i njenog interpolacionog polinoma

$$R_n(f, x) = f(x) - P_n(f, x).$$

Očigledno je greška interpolacionog polinoma u čvorovima interpolacije jednaka nuli. Pod pretpostavkom da funkcija f u nekom intervalu, kojem pripadaju svi čvorovi interpolacije i tačka x, ima prvi, drugi, ..., $n+1-$vi izvod i da su ti izvodi neprekidni (kraće rečeno, funkcija je $n+1$ put neprekidno diferencijabilna) može se dobiti izraz za grešku interpolacije. Dokaz odgovarajuće teoreme zasniva se na Rolovoj teoremi.

Teorema 3.2. ***Rolova teorema.*** *Neka je funkcija f definisana i neprekidna na intervalu $[a, b]$, neka ima prvi izvod u svakoj tački intervala (a, b) i neka je $f(a) = f(b)$. Tada postoji tačka $c \in (a, b)$ takva da je $f'(c) = 0$.*

Teorema 3.3. ***Greška interpolacije.*** *Neka međusobno različiti interpolacioni čvorovi $x_0, x_1, \ldots, x_n$ pripadaju intervalu $[a, b]$ i neka je funkcija f u tom intervalu $n + 1$ put neprekidno diferencijabilna. Tada za svako $x \in [a, b]$ postoji*

$$\alpha \in (\min\{x, x_0, x_1, \ldots, x_n\}, \max\{x, x_0, x_1, \ldots, x_n\}),$$

takvo da je

$$R_n(f, x) = \frac{f^{(n+1)}(\alpha)}{(n+1)!} \prod_{j=0}^{n} (x - x_j).$$

Dokaz. Očigledno je tvrđenje teoreme tačno ako je x jednako jednom od čvorova interpolacije. Za α se u ovom slučaju može uzeti bilo koji broj iz posmatranog intervala $[a, b]$. Posmatrajmo slučaj kada je x različito od svih čvorova interpolacije. Dokazaćemo tvrđenje teoreme samo za $n = 1$. Za $n > 1$ dokaz analogan. Posmatrajmo pomoćnu funkciju

$$\varphi(s) = f(s) - p_1(s) - \frac{(s - x_0)(s - x_1)}{(x - x_0)(x - x_1)} (f(x) - p_1(x)),$$

gde je p_1 interpolacioni polinom prvog stepena određen tačkama $(x_0, f(x_0))$ i $(x_1, f(x_1))$. Pošto je funkcija f dvaput neprekidno diferencijabilna u intervalu $[a, b]$, u istom intervalu je dvaput neprekidno diferencijabila i funkcija φ. Funkcija φ ima u posmatranom intervalu tri međusobno različite nule. To su čvorovi interpolacije x_0 i x_1 i x. Prema Rolovoj teoremi između svake dve nule funkcije φ nalazi se bar po jedna nula njenog prvog izvoda. Znači, ako pretpostavimo da je $x_0 < x < x_1$ (u ostalim slučajevima dokaz je potpuno isti), postoje dve nule α_0 i α_1 funkcije φ takve da je $x_0 < \alpha_0 < x < \alpha_1 < x_1$. Posmatrajmo sada funkciju φ', koja ima dve međusobno različite nule α_0 i α_1. Primenjujući

Rolovu teoremu dobijamo da φ'' ima bar jednu nulu $\alpha \in (\alpha_0, \alpha_1)$. Sada dobijamo (izvod tražimo po promenljivoj s)

$$0 = \varphi''(\alpha) = f''(\alpha) - \frac{2}{(x - x_0)(x - x_1)} \left(f(x) - p_1(x)\right),$$

jer je $p_1''(s) = 0$. Iz poslednje relacije dobijamo

$$R_1(f, x) = f(x) - p_1(x) = \frac{f''(\alpha)}{2}(x - x_0)(x - x_1).$$

■

Važno je naglasiti da α iz prethodne teoreme zavisi od funkcije f, čvorova interpolacije $x_0, x_1, \ldots, x_n$ i od x.

Kao posledicu prethodne teoreme imamo ocenu greške interpolacije. Ako pretpostavimo da funkcija f ima neprekidan $k-$ti izvod u intervalu $[a, b]$, onda postoji konstanta M_k takva da važi

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

Teorema 3.4. *Neka su ispunjene pretpostavke prethodne teoreme. Tada je*

$$|R_n(f, x)| \leq \frac{M_{n+1}}{(n+1)!} \left|\prod_{j=0}^{n} (x - x_j)\right|.$$

Primer 3.1. *Posmatraćemo aproksimaciju funkcije e^x na intervalu $[0, 0.5]$ interpolacionim polinomom prvog stepena ako su čvorovi interpolacije $x_0 = 0$ i $x_1 = 0.5$. Polinom prvog stepena određen tačkama $(0, 1)$, $(0.5, \sqrt{e})$ je jednačina prave koja sadrži te tačke, tj.*

$$p_1(x) = 1 + \frac{\sqrt{e} - 1}{0.5} x = 1 + 2x\left(\sqrt{e} - 1\right).$$

Za funkciju e^x je $M_2 = \sqrt{e} < 1.65$. Funkcija $(x - x_0)(x - x_1)$ je na intervalu $[x_0, x_1] = [0, 0.5]$ nepozitivna i njen lokalni minimum je u tački $\frac{x_0 + x_1}{2}$. Zbog toga je

$$|(x - x_0)(x - x_1)| \leq \left|\left(\frac{x_0 + x_1}{2} - x_0\right)\left(\frac{x_0 + x_1}{2} - x_1\right)\right| = \frac{(x_1 - x_0)^2}{4}, \text{ za } x \in [x_0, x_1],$$

odnosno

$$|(x - x_0)(x - x_1)| \leq \frac{1}{16}$$

jer je $(x_1 - x_0)^2 = \frac{1}{4}$. Na osnovu toga dobijamo za $x \in [0, 0.5]$

$$|e^x - p_1(x)| = |R_1(e^x, x)| \leq \frac{\sqrt{e}}{32} < 0.052.$$

3.2 Oblici interpolacionog polinoma

3.2.1 Lagranžov oblik

Interpolacioni polinom je za date čvorne tačke jedinstven, ali se može zapisati u više oblika. Jedan od njih je **Lagranžov oblik**

$$L_n(x) = \sum_{i=0}^{n} y_i \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j},$$

koji se može zapisati i na sledeći način

$$\begin{aligned} L_n(x) &= y_0\frac{(x-x_1)(x-x_2)\cdots(x-x_n)}{(x_0-x_1)(x_0-x_2)\cdots(x_0-x_n)} + y_1\frac{(x-x_0)(x-x_2)(x-x_3)\cdots(x-x_n)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)\cdots(x_1-x_n)} + \cdots \\ &+ y_k\frac{(x-x_0)(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)} + \cdots \\ &+ y_n\frac{(x-x_0)(x-x_1)\cdots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\cdots(x_n-x_{n-1})} \end{aligned}$$

Lako se vidi da je $L_n(x)$ polinom stepena ne većeg od n i da važi

$$\begin{aligned} L_n(x_0) &= y_0, \\ L_n(x_1) &= y_1, \\ &\vdots \\ L_n(x_k) &= y_k, \\ &\vdots \\ L_n(x_n) &= y_n. \end{aligned}$$

Primer 3.2. *Lagranžov polinom drugog stepena za čvorne tačke* $(0,1)$, $(0.5,0.8)$, $(1,0.5)$ *je*

$$L_2(x) = 1\cdot\frac{(x-0.5)(x-1)}{(0-0.5)(0-1)} + 0.8\cdot\frac{(x-0)(x-1)}{(0.5-0)(0.5-1)} + 0.5\cdot\frac{(x-0)(x-0.5)}{(1-0)(1-0.5)},$$

odnosno

$$p_2(x) = -0.2x^2 - 0.3x + 1.$$

Ako se čvornim tačkama iz prethodnog primera $(0,1)$, $(0.5,0.8)$, $(1,0.5)$ dodamo još jednu tačku i odredimo interpolacioni polinom za ove četiri čvorne tačke, dobićemo interpolacioni polinom trećeg stepena. Ceo postupak njegovog izračunavanja se mora ponoviti. Tako dopuna čvornom tačkom $(1.5,2.0)$ daje

$$p_3(x) = \frac{4}{3}(1-x)\left(x-\frac{3}{2}\right)\left(x-\frac{1}{2}\right) - 2\left(x-\frac{3}{2}\right)x\left(x-\frac{1}{2}\right) + \frac{8}{3}(x-1)x\left(x-\frac{1}{2}\right) - \frac{16}{5}\left(\frac{3}{2}-x\right)(x-1),$$

odnosno

$$p_3(x) = \frac{38x^3}{15} - 4x^2 + \frac{29x}{30} + 1.$$

Da bi se takav postupak izbegao, u slučajevima kada želimo da pređemo sa polinoma stepena n na polinom stepena $n+1$, koristi se predstavljanje interpolacionog polinoma pomoću podeljenih razlika. Kako je interpolacioni polinom jedinstven, govorimo samo o drugom zapisu istog polinoma.

Primer 3.3. *Polinom drugog stepena koji interpolira funkciju* $f(x) = \sin x$ *u tačkama* $x = 0.4$, $x = 0.7$ *i* $x = 1$ *glasi*

$$L_2(x) = \sin(0.4)\frac{(x-0.7)(x-1)}{(-0.3)\cdot(-0.6)} + \sin(0.7)\frac{(x-0.4)(x-1)}{0.3\cdot(-0.3)} + \sin(1.0)\frac{(x-0.4)(x-0.7)}{0.6\cdot 0.3}.$$

Na slici 1 *prikazani su grafik funkcije* $\sin x$ *i interpolacionog polinoma* $L_2(x)$. *Pojednostavljeni oblik interpolacionog polinoma je*

$$L_2(x) = -0.31970026x^2 + 1.20100144x - 0.03983019.$$

Za $x = 0.5$ *je*

$$L_2(0.5) = 0.48074546, \qquad f(0.5) = 0.47942553\ldots,$$

i

$$|L_2(0.5) - f(0.5)| < 1.32 \cdot 10^{-3}.$$

Kako je

$$f'''(x) = -\sin x, \qquad |f'''(x)| \leq 1, \quad x \in [0.4, 1],$$

i

$$|(x - 0.4)(x - 0.7)(x - 1)| < 0.0104, \quad x \in [0.4, 1],$$

sa $M_3 = 1$, *na osnovu teoreme o grešci interpolacije dobijamo da za svako* $x \in [0.4, 1]$ *važi*

$$|R_2(\sin x, x)| \leq \frac{M_3}{3!} \max_{x \in [0.4,1]} |(x - 0.4)(x - 0.7)(x - 1)| \leq \frac{1}{6} \cdot 0.0104 < 1.74 \cdot 10^{-3}.$$

3.2.2 Podeljene razlike. Njutnov oblik

U praksi je često potrebno odrediti nekoliko interpolacionih polinoma različitog stepena i zatim koristiti najbolji s obzirom na postavljene zahteve. Ukoliko se koristi Lagranžov oblik interpolacionog polinoma, svaki polinom se računa posebno, što zahteva veliki broj računanja kod polinoma većeg stepena. Ovaj problem se može izbeći korišćenjem rekurzivne veze između interpolacionih polinoma u **Njutnovom obliku**. Za zapis interpolacionih polinoma u ovom obliku koriste se **podeljene razlike**.

Definicija 3.1. *Neka su date čvorne tačke* (x_i, y_i), $i = 0, 1, \ldots, n$, *pri čemu su* $x_0, x_1, \ldots, x_n$ *međusobno različiti brojevi. Podeljene razlike k-tog reda označavaju se sa*

$$y[x_i, x_{i+1}, \ldots, x_{i+k}]$$

i određuju na sledeći način:
za $k = 0$

$$y[x_i] = y_i, \qquad i = 0, 1, \ldots, n,$$

a za $k = 1, 2, \ldots, n$ *je*

$$y[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{y[x_{i+1}, \ldots, x_{i+k}] - y[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}, \qquad i = 0, 1, \ldots n - k.$$

Podeljene razlike se ponekad pišu u vidu tabele:

i	x_i	$y[x_i]$	$y[x_i, x_{i+1}]$	$y[x_i, x_{i+1}, x_{i+2}]$	$y[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$	$y[x_i, x_{i+1}, x_{i+2}, x_{i+3}, x_{i+4}]$
0	x_0	y_0	$y[x_0, x_1]$	$y[x_0, x_1, x_2]$	$y[x_0, x_1, x_2, x_3]$	$y[x_0, x_1, x_2, x_3, x_4]$
1	x_1	y_1	$y[x_1, x_2]$	$y[x_1, x_2, x_3]$	$y[x_1, x_2, x_3, x_4]$	
2	x_2	y_2	$y[x_2, x_3]$	$y[x_2, x_3, x_4]$		
3	x_3	y_3	$y[x_3, x_4]$			
4	x_4	y_4				

Ako se podeljene razlike računaju za čvorne tačke (x_i, y_i), $i = 0, 1, \ldots, n$, pri čemu su $x_0, x_1, \ldots, x_n$ međusobno različiti brojevi, a $y_i = f(x_i)$ govorimo o podeljenim razlikama funkcije f. U tom slučaju pišemo $f[x_0, x_1, \ldots, x_k]$ umesto $y[x_0, x_1, \ldots, x_k]$.

Primer 3.4. *Izračunaćemo podeljene razlike za sledeće podatke*

k	0	1	2	3	4
x_k	0	2	3	5	8
y_k	1	4	2	0	3

$$y[x_0, x_1] = \frac{y_1 - y_0}{x_1 - x_0} = \frac{4 - 1}{2 - 0} = \frac{3}{2}, \qquad y[x_1, x_2] = \frac{y_2 - y_1}{x_2 - x_1} = \frac{2 - 4}{3 - 2} = -2,$$

$$y\left[x_2,x_3\right]=\frac{y_3-y_2}{x_3-x_2}=\frac{0-2}{5-3}=-1,\qquad y\left[x_3,x_4\right]=\frac{y_4-y_3}{x_4-x_3}=\frac{3-0}{8-5}=1.$$

$$y\left[x_0,x_1,x_2\right]=\frac{y\left[x_1,x_2\right]-y\left[x_0,x_1\right]}{x_2-x_0}=\frac{-2-\frac{3}{2}}{3-0}=-\frac{7}{6},$$

$$y\left[x_1,x_2,x_3\right]=\frac{y\left[x_2,x_3\right]-y\left[x_1,x_2\right]}{x_3-x_1}=\frac{-1+2}{5-2}=\frac{1}{3},$$

$$y\left[x_2,x_3,x_4\right]=\frac{y\left[x_3,x_4\right]-y\left[x_2,x_3\right]}{x_4-x_2}=\frac{1+1}{8-3}=\frac{2}{5},$$

$$y\left[x_0,x_1,x_2,x_3\right]=\left(y\left[x_1,x_2,x_3\right]-y\left[x_0,x_1,x_2\right]\right)/\left(x_3-x_0\right)=\frac{\frac{1}{3}+\frac{7}{6}}{5-0}=\frac{3}{10},$$

$$y\left[x_1,x_2,x_3,x_4\right]=\left(y\left[x_2,x_3,x_4\right]-y\left[x_1,x_2,x_3\right]\right)/\left(x_4-x_1\right)=\frac{\frac{2}{5}-\frac{1}{3}}{8-2}=\frac{1}{90},$$

$$y\left[x_0,x_1,x_2,x_3,x_4\right]=\left(y\left[x_1,x_2,x_3,x_4\right]-y\left[x_0,x_1,x_2,x_3\right]\right)/\left(x_4-x_0\right)=\frac{\frac{1}{90}-\frac{3}{10}}{8-0}=-\frac{13}{360}.$$

i	x_i	$y\left[x_i\right]$	$y\left[x_i,x_{i+1}\right]$	$y\left[x_i,x_{i+1},x_{i+2}\right]$	$y\left[x_i,x_{i+1},x_{i+2},x_{i+3}\right]$	$y\left[x_i,x_{i+1},x_{i+2},x_{i+3},x_{i+4}\right]$
0	0	1	3/2	−7/6	3/10	−13/360
1	2	4	−2	1/3	1/90	
2	3	2	−1	2/5		
3	5	0	1			
4	8	3				

Teorema 3.5. ***Podeljene razlike.*** *Za podeljene razlike važi*

$$y[x_0,x_1,\ldots,x_k]=\sum_{i=0}^{k}y_i\prod_{\substack{j=0\\j\neq i}}^{k}\frac{1}{x_i-x_j},\quad k=1,2,\ldots,n.$$

Dokaz. Dokaz se izvodi indukcijom po k. Za $k=1$ treba da važi

$$y[x_0,x_1]=y_0\frac{1}{x_0-x_1}+y_1\frac{1}{x_1-x_0}=\frac{y_1-y_0}{x_1-x_0},$$

što je vrednost koja se dobija iz definicije podeljenih razlika. Iz pretpostavke da je tvrđenje tačno za podeljene razlike reda $k-1$ pokazuje se da važi i za k. Na osnovu induktivne pretpostavke, dobija se

za $r = y[x_1, x_2, \ldots, x_k] - y[x_0, x_1, \ldots, x_{k-1}]$

$$\begin{aligned}
r &= \sum_{i=1}^{k} y_i \prod_{\substack{j=1 \\ j\neq i}}^{k} \frac{1}{x_i - x_j} - \sum_{i=0}^{k-1} y_i \prod_{\substack{j=0 \\ j\neq i}}^{k-1} \frac{1}{x_i - x_j} \\
&= \sum_{i=1}^{k-1} y_i \prod_{\substack{j=1 \\ j\neq i}}^{k} \frac{1}{x_i - x_j} + y_k \prod_{j=1}^{k-1} \frac{1}{x_k - x_j} - y_0 \prod_{j=1}^{k-1} \frac{1}{x_0 - x_j} - \sum_{i=1}^{k-1} y_i \prod_{\substack{j=0 \\ j\neq i}}^{k-1} \frac{1}{x_i - x_j} \\
&= \sum_{i=1}^{k-1} y_i \left(\prod_{\substack{j=1 \\ j\neq i}}^{k} \frac{1}{x_i - x_j} - \prod_{\substack{j=0 \\ j\neq i}}^{k-1} \frac{1}{x_i - x_j} \right) + y_k \prod_{j=1}^{k-1} \frac{1}{x_k - x_j} - y_0 \prod_{j=1}^{k-1} \frac{1}{x_0 - x_j} \\
&= \sum_{i=1}^{k-1} y_i \frac{x_k - x_0}{(x_i - x_0)(x_i - x_k)} \prod_{\substack{j=1 \\ j\neq i}}^{k-1} \frac{1}{x_i - x_j} + y_k \prod_{j=1}^{k-1} \frac{1}{x_k - x_j} - y_0 \prod_{j=1}^{k-1} \frac{1}{x_0 - x_j} \\
&= (x_k - x_0) \left(\sum_{i=1}^{k-1} y_i \prod_{\substack{j=0 \\ j\neq i}}^{k} \frac{1}{x_i - x_j} + y_k \prod_{j=0}^{k-1} \frac{1}{x_k - x_j} + y_0 \prod_{j=1}^{k} \frac{1}{x_0 - x_j} \right) \\
&= (x_k - x_0) \sum_{i=0}^{k} y_i \prod_{\substack{j=0 \\ j\neq i}}^{k} \frac{1}{x_i - x_j}.
\end{aligned}$$

Sada je

$$\begin{aligned}
\frac{r}{x_k - x_0} &= \sum_{i=0}^{k} y_i \prod_{\substack{j=0 \\ j\neq i}}^{k} \frac{1}{x_i - x_j} \\
&= \frac{y[x_1, x_2, \ldots, x_k] - y[x_0, x_1, \ldots, x_{k-1}]}{x_k - x_0} = y[x_0, x_1, \ldots, x_k],
\end{aligned}$$

pa tvrđenje sledi direktno. ■

Posledica 3.6. *Podeljena razlika $y[x_0, x_1, \ldots, x_k]$ je simetrična funkcija argumenata x_i, tj. ako je $x_{i_0}, x_{i_1}, \ldots, x_{i_k}$ proizvoljna permutacija brojeva $x_0, x_1, \ldots, x_k$, onda je*

$$y[x_0, x_1, \ldots, x_k] = y[x_{i_0}, x_{i_1}, \ldots, x_{i_k}].$$

Teorema 3.7. *Interpolacioni polinom određen čvornim tačkama (x_i, y_i), $i = 0, 1, \ldots, n$, sa međusobno različitim $x_0, x_1, \ldots, x_n$, može se zapisati u obliku*

$$N_n(x) = \sum_{i=0}^{n} y[x_0, x_1, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j).$$

Dokaz. Dovoljno je dokazati da je $L_n(x) = N_n(x)$, što se može postići indukcijom po n. Za $n = 1$ ova jednakost se lako proverava,

$$L_1(x) = y_0 \frac{x - x_1}{x_0 - x_1} + y_1 \frac{x - x_0}{x_1 - x_0} = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) = N_1(x).$$

Po induktivnoj pretpostavci je $L_{n-1}(x) = N_{n-1}(x)$ i kako je

$$N_n(x) - N_{n-1}(x) = y[x_0, x_1, \ldots, x_n] \prod_{j=0}^{n-1} (x - x_j),$$

ostaje da se dokaže da važi

$$L_n(x) - L_{n-1}(x) = y[x_0, x_1, \ldots, x_n] \prod_{j=0}^{n-1}(x - x_j).$$

Razlika $L_n(x) - L_{n-1}(x)$ je polinom stepena ne većeg od n, a njegove nule su x_i, $i = 0, 1, \ldots, n-1$, pa važi

$$L_n(x) - L_{n-1}(x) = A \prod_{j=0}^{n-1}(x - x_j).$$

Ako se dokaže da je $A = y[x_0, x_1, \ldots, x_n]$, sledi tvrđenje. Sa $x = x_n$ iz poslednje relacije se dobija

$$\begin{aligned} A &= \prod_{j=0}^{n-1} \frac{1}{x_n - x_j} \left(y_n - \sum_{i=0}^{n-1} y_i \prod_{\substack{j=0 \\ j\neq i}}^{n-1} \frac{x_n - x_j}{x_i - x_j} \right) = y_n \prod_{j=0}^{n-1} \frac{1}{x_n - x_j} - \sum_{i=0}^{n-1} y_i \frac{1}{x_n - x_i} \prod_{\substack{j=0 \\ j\neq i}}^{n-1} \frac{1}{x_i - x_j} \\ &= \sum_{i=0}^{n} y_i \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{1}{x_i - x_j} = y[x_0, x_1, \ldots, x_n]. \end{aligned}$$

■

Polinom $N_n(x)$ je **Njutnov interpolacioni polinom**, a predstavlja samo jedan oblik interpolacionog polinoma.

Primer 3.5. *U prethodnom primeru izračunali smo podeljene razlike za čvorne tačke* $(0,1)$, $(2,4)$, $(3,2)$, $(5,0)$, $(8,3)$. *Njutnov oblik interpolacionog polinoma za prve četiri čvorne tačke je*

$$p_3(x) = 1 + \frac{3}{2}(x-0) - \frac{7}{6}(x-0)(x-2) + \frac{3}{10}(x-0)(x-2)(x-3),$$

odnosno

$$p_3(x) = 1 + \frac{169}{30}x - \frac{8}{3}x^2 + \frac{3}{10}x^3.$$

Interpolacioni polinom za sve date čvorne tačke, tj. p_4, *sada se lako dobija*

$$p_4(x) = p_3(x) - \frac{13}{360}(x-0)(x-2)(x-3)(x-5),$$

odnosno

$$p_4(x) = 1 + \frac{403}{60}x - \frac{1363}{360}x^2 + \frac{119}{180}x^3 - \frac{13}{360}x^4.$$

Greška interpolacije $f(x) - N_n(x)$ može se iskazati i preko podeljenih razlika.

Teorema 3.8. *Neka je funkcija* f *definisana u intervalu* $[a,b]$. *Ako je* $N_n(x)$ *interpolacioni polinom funkcije* f *određen međusobno različitim čvorovima* $x_0, \ldots, x_n \in [a,b]$, *onda za svako* $x \in [a,b] \setminus \{x_0, x_1, \ldots, x_n\}$ *važi*

$$f(x) - N_n(x) = f[x_0, x_1, \ldots, x_n, x]\omega_n(x).$$

Dokaz. Vrednosti x, x_i su međusobno različite. Neka je $N_{n+1}(s)$ interpolacioni polinom funkcije f određen čvorovima $x_0, x_1, \ldots, x_n, x_{n+1}$, pri čemu je $x_{n+1} = x$. Tada važi

$$f(x) = N_{n+1}(x), \qquad f(x_i) = N_{n+1}(x_i), \quad i = 0, 1, \ldots, n.$$

Kako je

$$N_{n+1}(s) = N_n(s) + f[x_0, x_1, \ldots, x_n, x] \prod_{j=0}^{n} (s - x_j),$$

dobija se

$$f(x) = N_{n+1}(x) = N_n(x) + f[x_0, x_1, \ldots, x_n, x] \prod_{j=0}^{n} (x - x_j),$$

odnosno

$$f(x) - N_n(x) = f[x_0, x_1, \ldots, x_n, x]\omega_n(x).$$

■

Ovaj oblik greške ne zahteva poznavanje izvoda funkcije f, ali nije ništa praktičniji od ranije određenog oblika greške, jer je vrednost $f(x)$ nepoznata (inače bi se greška mogla direktno izračunati).

Sada se lako može dokazati sledeća teorema.

Teorema 3.9. *Neka je $f \in C^k[a, b]$ i neka $f^{(k+1)}(x)$ postoji u svakoj tački intervala (a, b). Tada za međusobno različite $x, x_0, x_1, \ldots, x_k \in [a, b]$ u intervalu*

$$(\min\{x, x_0, x_1, \ldots, x_n\}, \max\{x, x_0, x_1, \ldots, x_n\})$$

postoji $\xi = \xi(x)$ takvo da važi

$$f[x_0, x_1, \ldots, x_k, x] = \frac{f^{(k+1)}(\xi)}{(k+1)!}.$$

Dokaz. Na osnovu prethodne teoreme važi

$$f(x) - N_k(x) = f[x_0, x_1, \ldots, x_k, x]\omega_k(x),$$

gde je $N_k(s)$ interpolacioni polinom funkcije f određen međusobno različitim čvorovima $x_0, x_1, \ldots, x_k$. Na osnovu teoreme o grešci interpolacije dobija se da za neko $\xi = \xi(x)$ iz intervala

$$(\min\{x, x_0, x_1, \ldots, x_n\}, \max\{x, x_0, x_1, \ldots, x_n\})$$

važi

$$f(x) - N_k(x) = \frac{f^{(k+1)}(\xi)}{(k+1)!}\omega_k(x).$$

Upoređivanjem ova dva izraza za grešku, direktno se dobija tvrđenje. ■

Rezultat prethodne teoreme može se iskoristiti za dobijanje aproksimacije izraza

$$\frac{f^{(n+1)}(\xi)}{(n+1)!},$$

koji se javlja u izrazu za grešku interpolacije. Naime, ako su vrednosti $x_0, \ldots, x_{n+1}$ međusobno različite i ako su poznate vrednosti funkcije f u ovim tačkama, može se uzeti

$$\frac{f^{(n+1)}(\xi)}{(n+1)!} \approx f[x_0, x_1, \ldots, x_n, x_{n+1}].$$

3.2.3 Konačne razlike. Njutn-Gregorijevi oblici

U slučaju kada su čvorovi interpolacije ekvidistantni, tj. kada je

$$x_i = x_0 + ih, \quad i = 0, 1, \ldots, n,$$

pri čemu je h pozitivan broja a x_0 dato, interpolacioni polinom za čvorne tačke (x_i, y_i), $i = 0, 1, \ldots, n$, može se zapisati i u jednostavnijem obliku. Pri tome se umesto podeljenih razlika koriste opadajuće i rastuće konačne razlike.

Definicija 3.2. ***Opadajuće razlike.*** *Opadajuće razlike k-tog reda za skup* $\{y_0, y_1, \ldots, y_n\}$ *označavaju se sa* $\Delta^k y_i$ *i određuju na sledeći način:*
za $k = 0$

$$\Delta^0 y_i = y_i, \quad i = 0, 1, \ldots, n,$$

a za $k = 1, 2, \ldots, n$

$$\Delta^k y_i = \Delta^{k-1} y_{i+1} - \Delta^{k-1} y_i, \quad i = 0, 1, \ldots, n-k.$$

Definicija 3.3. ***Rastuće razlike.*** *Rastuće razlike k-tog reda za skup* $\{y_0, y_1, \ldots, y_n\}$ *označavaju se sa* $\nabla^k y_i$ *i određuju na sledeći način:*
za $k = 0$

$$\nabla^0 y_i = y_i, \quad i = 0, 1, \ldots, n,$$

a za $k = 1, 2, \ldots, n$

$$\nabla^k y_i = \nabla^{k-1} y_i - \nabla^{k-1} y_{i-1}, \quad i = k, k+1, \ldots, n.$$

Opadajuće i rastuće razlike zajednički se nazivaju **konačne razlike**. Konačne razlike prvog reda se obeležavaju sa Δy_i i ∇y_i.Iz definicije je očigledno da se pri definisanju konačnih razlika koriste samo vrednosti y_i. Slično podeljenim razlikama i konačne razlike pišemo u vidu tabele.

i	x_i	$\Delta^0 y_i \ (\nabla^0 y_i)$	$\Delta^1 y_i \ (\nabla^1 y_i)$	$\Delta^2 y_i \ (\nabla^2 y_i)$	$\Delta^3 y_i \ (\nabla^3 y_i)$	$\Delta^4 y_i \ (\nabla^4 y_i)$
0	x_0	y_0	$\Delta y_0 \ (\nabla y_1)$	$\Delta^2 y_0 \ (\nabla^2 y_2)$	$\Delta^3 y_0 \ (\nabla^3 y_3)$	$\Delta^4 y_0 \ (\nabla^4 y_4)$
1	x_1	y_1	$\Delta y_1 \ (\nabla y_2)$	$\Delta^2 y_1 \ (\nabla^2 y_3)$	$\Delta^3 y_1 \ (\nabla^3 y_4)$	
2	x_2	y_2	$\Delta y_2 \ (\nabla y_3)$	$\Delta^2 y_2 \ (\nabla^2 y_4)$		
3	x_3	y_3	$\Delta y_3 \ (\nabla y_4)$			
4	x_4	y_4				

U istoj tabeli prikazali smo i opadajuće i rastuće (u zagradama) konačne razlike. To su, dakle, isti brojevi, ali se označavaju na različite načine. Već iz date tabele konačnih razlika vidimo da važi sledeća teorema, koja se lako dokazuje matematičkom indukcijom.

Teorema 3.10. *Neka su dati brojevi* y_i, $i = 0, 1, , n$. *Za* $i = 0, 1, \ldots, n-k$ *važi*

$$\Delta^k y_i = \nabla^k y_{i+k}.$$

Dokaz. Dokaz se izvodi indukcijom po k. Očigledno je

$$\Delta^0 y_i = \nabla^0 y_i = y_i,$$

a iz

$$\Delta^{k-1} y_i = \nabla^{k-1} y_{i+k-1}$$

sledi

$$\Delta^k y_i = \Delta^{k-1} y_{i+1} - \Delta^{k-1} y_i = \nabla^{k-1} y_{i+1+k-1} - \nabla^{k-1} y_{i+k-1} = \nabla^{k-1} y_{i+k} - \nabla^{k-1} y_{i+k-1} = \nabla^k y_{i+k}.$$

■

Primer 3.6. *Izračunaćemo podeljene razlike za sledeće podatke*

k	0	1	2	3	4
y_k	1	4	12	10	32

i	y_i	Δy_i (∇y_i)	$\Delta^2 y_i$ $(\nabla^2 y_i)$	$\Delta^3 y_i$ $(\nabla^3 y_i)$	$\Delta^4 y_i$ $(\nabla^4 y_i)$
0	1	3	5	-15	49
1	4	8	-10	34	
2	12	-2	24		
3	10	22			
4	32				

Ako se konačne razlike računaju za brojeve $y_i = f(x_i)$, $i = 0, 1, , \ldots, n$, govorimo o konačnim razlikama funkcije f.

Konačne razlike mogu se koristiti za prikazivanje interpolacionog polinoma u jednostavnijem obliku, ako su čvorovi interpolacije ekvidistantni. Pri tome se polazi od Njutnovog oblika interpolacionog polinoma i veze koja postoji između konačnih i podeljenih razlika, date u sledećoj teoremi.

Teorema 3.11. ***Veza podeljenih i konačnih razlika.*** *Ako je $x_i = x_0 + ih$, $i = 0, 1, \ldots, n$, za dato x_0 i neko pozitivno h, onda za čvorne tačke (x_i, y_i), $i = 0, 1, \ldots, n$, za $k = 0, 1, \ldots, n$ i svako $i = 0, 1, \ldots, n-k$ važi*

$$\Delta^k y_i = \nabla^k y_{i+k} = k! h^k y[x_i, x_{i+1}, \ldots, x_{i+k}].$$

Dokaz. U prethodnoj teoremi je dokazano da važi $\Delta^k y_i = \nabla^k y_{i+k}$. Ostaje da se dokaže samo da važi $\Delta^k y_i = k! h^k y[x_i, x_{i+1}, \ldots, x_{i+k}]$. Dokaz se izvodi indukcijom po k. Očigledno je

$$\Delta^0 y_i = y_i = y[x_i].$$

Iz

$$\Delta^{k-1} y_i = (k-1)! h^{k-1} y[x_i, x_{i+1}, \ldots, x_{i+k-1}]$$

sledi

$$\begin{aligned} \Delta^k y_i &= \Delta^{k-1} y_{i+1} - \Delta^{k-1} y_i = (k-1)! h^{k-1} \left(y[x_{i+1}, \ldots, x_{i+k}] - y[x_i, \ldots, x_{i+k-1}] \right) \frac{x_{i+k} - x_i}{x_{i+k} - x_i} \\ &= (k-1)! h^{k-1} (x_{i+k} - x_i) y[x_i, \ldots, x_{i+k}] = k! h^k y[x_i, \ldots, x_{i+k}], \end{aligned}$$

jer je

$$x_{i+k} - x_i = kh.$$

■

Posmatraćemo sada specijalan slučaj koji se dobija za $i = 0$. Naime, tada na osnovu prethodne teoreme za svako $k = 0, 1, \ldots, n$ važi

$$\Delta^k y_0 = k! h^k y[x_0, x_1, \ldots, x_k].$$

Zamenujući u Njutnovom obliku interpolacionog polinoma

$$\begin{aligned} p_n(x) &= y_0 + y[x_0, x_1](x - x_0) + y[x_0, x_1, x_2](x - x_0)(x - x_1) + \cdots \\ &\quad + y[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1) \cdots (x - x_{n-1}) \end{aligned}$$

$y[x_0, x_1, \ldots, x_k]$ sa $\dfrac{\Delta^k y_0}{k! h^k}$ dobijamo

$$p_n(x) = y_0 + \frac{\Delta y_0}{1! h}(x - x_0) + \frac{\Delta^2 y_0}{2! h^2}(x - x_0)(x - x_1) + \cdots + \frac{\Delta^n y_0}{n! h^n}(x - x_0)(x - x_1) \cdots (x - x_{n-1}).$$

Ovaj oblik interpolacionog polinoma nazivamo **Njutn-Gregorijev oblik** sa opadajućim razlikama i možemo ga zapisati i na sledeći način

$$p_n(x) = \sum_{k=0}^{n} \frac{\Delta^k y_0}{k!h^k} \prod_{j=0}^{k-1}(x - x_j).$$

Imajući u vidu da su interpolacioni čvorovi ekvidistantni i da važi

$$x - x_i = x - x_0 - ih = h\left(\frac{x - x_0}{h} - i\right),\ i = 0, 1, \ldots, n,$$

smenom

$$s = \frac{x - x_0}{h},$$

odakle sledi $x - x_i = h(s - i)$, dobijamo

$$p_n(x) = y_0 + \frac{\Delta y_0}{1!}s + \frac{\Delta^2 y_0}{2!}s(s-1) + \cdots + \frac{\Delta^n y_0}{n!}s(s-1)\cdots(s-(n-1)),$$

što možemo zapisati i na sledeći način

$$p_n(x) = \sum_{k=0}^{n} \frac{\Delta^k y_0}{k!} \prod_{j=0}^{k-1}(s - j).$$

Interpolacionoi polinom za čvorne tačke $(x_i, y_i)\,,\ i = 0, 1, \ldots, n$, je jedinstven, ali se može zapisati na različite načine. Posle smene

$$X_i = x_{n-i}, \qquad Y_i = y_{n-i}, \qquad i = 0, 1, \ldots, n,$$

formiramo interpolacioni polinom $p_n\,(x)$ sa podeljenim razlikama za čvorne tačke (X_i, Y_i), $i = 0, 1, \ldots, n$,

$$p_n(x) = \sum_{k=0}^{n} Y[X_0, X_1, \ldots, X_k] \prod_{j=0}^{k-1}(x - X_j).$$

Imajući u vidu smene, dobijamo

$$p_n(x) = \sum_{k=0}^{n} y[x_n, x_{n-1}, \ldots, x_{n-k}] \prod_{j=0}^{k-1}(x - x_{n-j}).$$

Sada se iskoristi osobina simetričnosti podeljenih razlika i dobija se

$$p_n(x) = \sum_{k=0}^{n} y[x_{n-k}, x_{n-k+1}, \ldots, x_n] \prod_{j=0}^{k-1}(x - x_{n-j}).$$

Na osnovu specijalanog slučaja prethodne teoreme, koji se dobija za $i = n - k$, imamo za svako $k = 0, 1, \ldots, n$

$$\nabla^k y_n = k!h^k y\left[x_{n-k}, x_{n-k+1}, \ldots, x_n\right].$$

Znači,

$$p_n(x) = \sum_{k=0}^{n} \frac{\nabla^k y_n}{k!h^k} \prod_{j=0}^{k-1}(x - x_{n-j}),$$

odnosno

$$p_n(x) = y_n + \frac{\nabla y_n}{1!h}(x - x_n) + \frac{\nabla^2 y_n}{2!h^2}(x - x_n)(x - x_{n-1}) + \cdots + \frac{\nabla^n y_n}{n!h^n}(x - x_n)(x - x_{n-1})\cdots(x - x_1).$$

Imajući u vidu da su interpolacioni čvorovi ekvidistantni i da važi

$$x - x_i = x - x_0 - ih = x - x_0 - nh + (n-i)h = h\left(\frac{x - x_n}{h} + (n-i)\right),\ i = 0, 1, \ldots, n,$$

smenom

$$s = \frac{x - x_n}{h},$$

odakle sledi $x - x_i = h(s + (n-i))$, dobijamo

$$p_n(x) = y_n + \frac{\nabla y_n}{1!}s + \frac{\nabla^2 y_n}{2!}s(s+1) + \cdots + \frac{\nabla^n y_n}{n!}s(s+1)\cdots(s+(n-1)).$$

Ovaj oblik interpolacionog polinoma nazivamo **Njutn-Gregorijev oblik** sa rastućim razlikama i možemo ga zapisati i na sledeći način

$$p_n(x) = \sum_{k=0}^{n} \frac{\nabla^k y_n}{k!} \prod_{j=0}^{k-1}(j - s).$$

Primer 3.7. *Za sledeće podatke naći ćemo Njutn-Gregorijeve oblike interpolacionih polinoma*

k	0	1	2	3	4
x_k	10	12	14	16	18
y_k	1	4	12	10	32

Očigledno je $x_k = x_0 + kh$ sa $x_0 = 10$ i $h = 2$. U prethodnom primeru izračunali smo za date vrednosti y_i, $i = 0, 1, \ldots, n$, konačne razlike. Koristeći te rezultate dobijamo Njutn-Gregorijev oblik interpolacionog polinoma sa opadajućim konačnim razlikama

$$p_4(x) = 1 + \frac{3}{1!2}(x-10) + \frac{5}{2!4}(x-10)(x-12) - \frac{15}{3!8}(x-10)(x-12)(x-14) + \frac{49}{4!16}(x-10)(x-12)(x-14)(x-16),$$

odnosno

$$p_4(x) = 1 + \frac{3}{2}(x-10) + \frac{5}{8}(x-10)(x-12) - \frac{5}{16}(x-10)(x-12)(x-14) + \frac{49}{384}(x-10)(x-12)(x-14)(x-16).$$

Njutn-Gregorijev oblik interpolacionog polinoma sa rastućim konačnim razlikama je

$$p_4(x) = 32 + \frac{22}{1!2}(x-18) + \frac{24}{2!4}(x-18)(x-16) + \frac{34}{3!8}(x-18)(x-16)(x-14) + \frac{49}{4!16}(x-18)(x-16)(x-14)(x-12),$$

odnosno

$$p_4(x) = 32 + 11(x-18) + 3(x-18)(x-16) + \frac{17}{24}(x-18)(x-16)(x-14) + \frac{49}{384}(x-18)(x-16)(x-14)(x-12).$$

Pri izračunavanju vrednosti interpolacionog polinoma u tački x, koja je različita od svih čvorova interpolacije, ne traži se prvo interpolacioni polinom a zatim njegova vrednost u toj tački. Brži način je direktno računanje vrednosti interpolacionog polinoma u toj tački. Za to su posebno pogodni Njutn-Gregorijevi oblici interpolacionog polinoma.

Neka je za podatke iz prethodnog primera potrebno izračunati približnu vrednost za $x = 11.8$. Tada možemo uzeti

$$s = \frac{x - x_0}{h} = \frac{11.8 - 10}{2} = 0.9,$$

a zatim pomoću Njutn-Gregorijevog oblika sa opadajućim konačnim razlikama, dobijamo

$$p_4(11.8) = 1 + 3 \cdot 0.9 + \frac{5}{2} \cdot 0.9(-0.1) - \frac{15}{6} \cdot 0.9(-0.1)(-1.1) + \frac{49}{24} \cdot 0.9(-0.1)(-1.1)(-2.1),$$

odnosno $p_4(11.8) = 2.80304$.

Ako je x na početku ili na kraju tabele, onda se za izračunavanje približne vrednosti funkcije (zadate tabelom) u tački x često koristi samo nekoliko od raspoloživih tačaka sa početka, odnosno sa kraja tabele. Na taj način se sa interpolacionim polinomom reda manjeg od n izračunava vrednost za dato x. Ako je x na početku tabele, koristi se Njutn-Gregorijev oblik interpolacionog polinoma sa opadajućim konačnim razlikama, a ako je x na kraju tabele, koristi se Njutn-Gregorijev oblik interpolacionog polinoma sa rastućim konačnim razlikama.

Primer 3.8. *Približnu vrednost funkcije zadate tabelom*

k	0	1	2	3	4
x_k	10	12	14	16	18
y_k	1	4	12	10	32

u tački $x = 11.8$ izračunaćemo pomoću Njutn-Gregorijevog oblika interpolacionog polinoma sa opadajućim konačnim razlikama. Koristićemo interpolacioni polinom drugog stepena, određen sa prve tri tačke sa početka tabele. Iz tabele konačnih razlika

i	y_i	Δy_i (∇y_i)	$\Delta^2 y_i$ $(\nabla^2 y_i)$	$\Delta^3 y_i$ $(\nabla^3 y_i)$	$\Delta^4 y_i$ $(\nabla^4 y_i)$
0	1	3	5	-15	49
1	4	8	-10	34	
2	12	-2	24		
3	10	22			
4	32				

dobijamo

$$p_2(x) = 1 + \frac{3}{2}(x - 10) + \frac{5}{8}(x - 10)(x - 12),$$

i

$$p_2(11.8) = 1 + \frac{3}{2}1.8 + \frac{5}{8}1.8(-0.2) = 3.475.$$

Približnu vrednost funkcije zadate tabelom u tački $x = 17.3$ izračunaćemo pomoću Njutn-Gregorijevog oblika interpolacionog polinoma sa rastućim konačnim razlikama. Koristićemo interpolacioni polinom drugog reda, određen sa poslednje tri tačke tabele. Koristeći se prethodnom tabelom konačnih razlika dobijamo

$$p_2(x) = 32 + \frac{22}{1! \cdot 2}(x - 18) + \frac{24}{2! \cdot 4}(x - 18)(x - 16)$$

i

$$p_2(17.3) = 32 + 11(-0.7) + 3(-0.7)\ 1.3 = 21.57.$$

3.3 Linearna i kvadratna interpolacija

3.3.1 Linearna interpolacija

Neka su za međusobno različite vrednosti $x_0, \ldots, x_n$ poznate vrednosti funkcije f, tj. neka su poznate vrednosti

$$y_i = f(x_i), \quad i = 0, 1, \ldots, n.$$

Za izračunavanje približne vrednosti funkcije f u tački α, ako se α nalazi između dve date vrednosti x_k i x_{k+1}, najčešće se koristi linearna interpolacija. To znači da se za čvorne tačke

$$(x_k, f(x_k)), \quad (x_{k+1}, f(x_{k+1}))$$

određuje interpolacioni polinom prvog stepena i da se vrednost tog polinoma u tački α uzima kao približna vrednost za $f(\alpha)$.

Interpolacioni polinom određen čvornim tačkama $(x_k, f(x_k))$ i $(x_{k+1}, f(x_{k+1}))$ može se prikazati u Njutnovom obliku

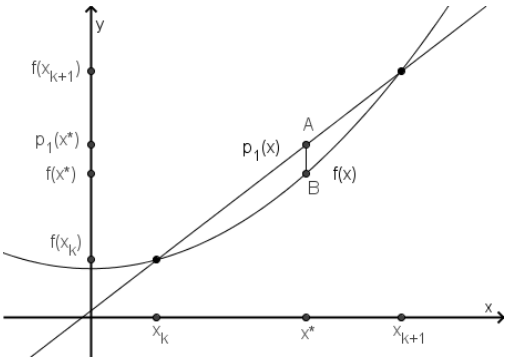
$$p_1(x) = f(x_k + f[x_k, x_{k+1}](x - x_k) = f(x_k) + \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k}(x - x_k).$$

Uzima se

$$f(\alpha) \approx p_1(\alpha) = f(x_k) + \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k}(\alpha - x_k).$$

Geometrijski posmatrano, p_1 je prava koja sadrži tačke $(x_k, f(x_k))$ i $(x_{k+1}, f(x_{k+1}))$. Na slici 1 dužina duži AB odgovara razlici

$$p_1(x^*) - f(x^*).$$



Slika 1.

U logaritamskim tablicama dato je $log7300 \approx 3.86332$ i $log7305 \approx 3.86362$. Približnu vrednost za $log7300.94$, koja se ne nalazi u tablicama, izračunaćemo pomoću polinoma prvog stepena

$$p_1(x) = 3.86332 + \frac{3.86362 - 3.86332}{7305 - 7300}(x - 7300) = 3.86332 + \frac{0.00030}{5}(x - 7300).$$

Dobijamo

$$p_1(7300.94) = 3.86332 + \frac{0.00030}{5} \cdot 0.94 = 3.86338.$$

Tačna vrednost je $log7300.94 \approx 3.86337877937134702\ldots$.

3.3.2 Kvadratna interpolacija

Neka su date tri tačke $(a, f(a))$, $(b, f(b))$ i $(c, f(c))$. Pretpostavimo da je $a \neq b \neq c \neq a$, tj. da su a, b i c međusobno različiti brojevi. Tada možemo odrediti interpolacioni polinom drugog stepena za funkciju f. Taj polinom u Lagranžovom obliku je

$$p_2(x) = \frac{(x-b)(x-c)}{(a-b)(a-c)}f(a) + \frac{(x-a)(x-c)}{(b-a)(b-c)}f(b) + \frac{(x-a)(x-b)}{(c-a)(c-b)}f(c),$$

a u Njutnov obliku

$$p_2(x) = f(a) + \left(\frac{f(a)}{a-b} + \frac{f(b)}{b-a}\right)(x-a) + \left(\frac{f(a)}{(a-b)(a-c)} + \frac{f(b)}{(b-a)(b-c)} + \frac{f(c)}{(c-a)(c-b)}\right)(x-a)(x-b).$$

Ako pretpostavimo da je

$$c = \frac{a+b}{2}$$

dobićemo

$$p_2(x) = \frac{1}{(a-b)^2}\left(f(a)(b-x)(a+b-2x) + 4f\left(\frac{a+b}{2}\right)(a-x)(x-b) + f(b)(a-x)(a+b-2x)\right).$$

Pod pretpostavkom da je funkcija f tri puta neprekidno diferencijabilna, greška kvadratne interpolacije je

$$R_2(f,x) = \frac{f^{(3)}(\alpha)}{6}(x-a)(x-b)(x-c)$$

Ako je $c = \frac{a+b}{2}$ greška $R_2(f,x)$ se može oceniti na intervalu $[a,b]$ na sledeći način. Grafik funkcije

$$g(x) = (x-a)(x-b)(x-c) = (x-a)(x-b)\left(x - \frac{a+b}{2}\right)$$

na intervalu $[a,b]$ prikazan je na slici 2. Ova funkcija ima maksimum

$$\frac{(b-a)^3}{12\sqrt{3}}$$

za

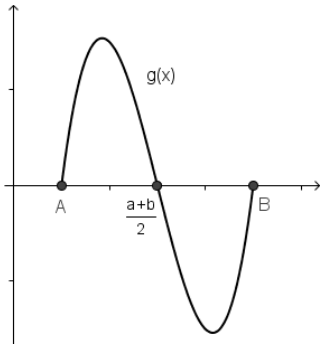
$$x_1 = \frac{1}{6}\left(3a + 3b - \sqrt{3}(b-a)\right)$$

i minimum

$$-\frac{(b-a)^3}{12\sqrt{3}}$$

za

$$x_2 = \frac{1}{6}\left(3a + 3b + \sqrt{3}(b-a)\right).$$



Slika 2.

U oba slučaja je

$$|R_2(f,x)| \leq \frac{M_3}{72\sqrt{3}}(b-a)^3,$$

gde je konstanta M_3 određena tako da važi

$$\left|f^{(3)}(x)\right| \leq M_3 \quad \text{za} \quad x \in [a,b].$$

3.4 Zadaci

3.1. *Za sledeću tabelu odrediti interpolacioni polinom u Lagranžovom obliku.*

i	0	1	2	3
x_i	1	2	3	5
y_i	1	5	14	81

3.2. *Funkcija* $y = f(x)$ *je data tabelom*

x	7	8	9	10
y	3	1	1	9

.

Odrediti $f(9.5)$, *koristeći Lagranžov interpolacioni polinom.*

3.3. *Sa kojom tačnošću se može izračunati* $\sqrt{117}$ *pomoću Lagranžovog interpolacionog polinoma za funkciju* $y = \sqrt{x}$, *ako su čvorovi interpolacije* $x_0 = 100$, $x_1 = 121$ *i* $x_2 = 144$?

3.4. *Za funkciju* $f(x) = x \ln x + xe^x$ *data je sledeća tabela*

i	0	1	2	3
x_i	0.2	0.9	1.7	2.3
$f(x_i)$	−0.0776	2.1188	10.2078	24.8563

.

Odredi odgovarajući Lagranžov interpolacioni polinom $L_3(x)$ *i oceni grešku* $|f(1.9) - L_3(1.9)|$.

3.5. *Nađi interpolacioni polinom za funkciju* $f(x) = e^{-x}$, *ako su čvorovi interpolacije* $x_0 = 1$, $x_2 = 2$ *i* $x_3 = 3$. *Oceni grešku za* $x = 1.5$.

3.6. *Nađi interpolacione polinome za funkciju*

$$f(x) = \ln x - \frac{x-1}{x}$$

sa sledećim čvorovima interpolacije
a) $2, 4, 6, 10$;
b) $4, 8, 10$;
c) $2, 4, 8$.
U svakom slučaju izračunati približnu vrednost za $\ln 5.25$ *i oceniti grešku.*

3.7. *Odredi polinom trećeg stepena koji interpolira polinom*

$$p(x) = x^4 - 4x^3 + 4x^2 + 1$$

u čvorovima $-1, 0, 1, 2$. *Nacrtaj grafike oba polinoma u intervalu* $[-2, 3]$ *i diskutuj aproksimaciju polinoma* $p(x)$ *interpolacionim polinomom.*

3.8. *Dokaži da važi*

$$y_{k+i} = \sum_{j=0}^{k} \binom{k}{j} \Delta^j y_i.$$

3.9. *Neka je* $y_i \in \mathbb{R}$, $i = 0, 1, \ldots, n$. *Dokaži da za* $m = 0, 1, \ldots, n$ *važi*

a)

$$\Delta^m y_k = \sum_{j=0}^{m} (-1)^{m-j} \binom{m}{j} y_{k+j}, \quad k = 0, 1, \ldots, n-m.$$

b)

$$\nabla^m y_k = \sum_{j=0}^{m} (-1)^{j} \binom{m}{j} y_{k-j}, \quad k = m, m+1, \ldots, n.$$

3.10. *Izračunaj vrednosti y_k koje nedostaju u tabeli*

k	y_k	Δy_k
0	0	1
1	·	2
2	·	4
3	·	7
4	·	11
5	·	16
6	·	

3.11. *Izračunaj vrednosti y_k i Δy_k koje nedostaju u tabeli*

k	y_k	Δy_k	$\Delta^2 y_k$
0	·	·	1
1	·	·	4
2	·	5	13
3	6	·	18
4	·	·	24
5	·	·	
6	·		

.

3.12. *Znajući vrednosti funkcije sin x u tačkama* 0, $(\pi/6)$, $(\pi/4)$, $(\pi/3)$, *i* $\pi/2$ *nađi približnu vrednost za sin* $(\pi/12)$ *pomoću interpolacionog polinoma četvrtog stepena.*

Glava 4

Numeričko rešavanje jednačina

U ovom delu se posmatraćemo neke postupke za numeričko rešavanje jednačine $f(x) = 0$, gde je f realna funkcija realne promenljive. Pri tome vaki broj ξ za koji važi $f(\xi) = 0$ nazivamo **rešenjem** ili **korenom** jednačine $f(x) = 0$. Koren jednačine $f(x) = 0$ naziva se još i **nula** funkcije f. Rešenje jednačina tražićemo samo u skupu realnih brojeva ili nekom njegovom podskupu.

Ako je funkcija f polinom stepena n

$$f(x) = p_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

čiji su koeficijenti a_n, $a_1, \ldots, a_n$ realni brojevi i $a_n \neq 0$, jednačina $f(x) = 0$ je **algebarska. Stepen polinoma je i stepen algebarske jednačine.**

Svaka jednačina koja nije algebarska naziva se transendentna jednačina. Takve su na primer jednačine

$$e^x + 3x - 2 = 0, \qquad x + \cos x = 0, \qquad x^2 + 2 + \log x = 0.$$

Da bi se moglo pristupiti određivanju korena jednačine $f(x) = 0$, potrebno je utvrditi da li oni postoje. Ako je f polinom, a skup u kojem se traže njegove nule skup kompleksnih brojeva, problem egzistencije nula je rešen. Naime, prema osnovnoj teoremi algebre svaka algebarska jednačina $n-$tog stepena ima u skupu kompleksnih brojeva tačno n rešenja. Ova rešenja nisu obavezno međusobno različita. U daljem radu, ako se posebno ne naglasi, govorićemo samo o nulama polinoma koje su realni brojevi. Kada se traže realna rešenja algebarske jednačine $n-$tog stepena, onda za broj m tih rešenja važi $0 \leq m \leq n$.

Algebarska jednačina prvog stepena

$$a_1 x + a_0 = 0, \ a_1 \neq 0,$$

kao što je poznato, ima rešenje

$$\xi = -\frac{a_0}{a_1}.$$

Za algebarsku jednačinu drugog stepena

$$a_2 x^2 + a_1 x + a_0 = 0, \ a_2 \neq 0,$$

možemo navesti formule za određivanje rešenja

$$\xi_1 = \frac{-a_1 - \sqrt{a_1^2 - 4a_0 a_2}}{2a_2}, \qquad \xi_2 = \frac{-a_1 + \sqrt{a_1^2 - 4a_0 a_2}}{2a_2}.$$

Još za $n = 3$ i $n = 4$ postoje formule pomoću kojih se mogu izračunati rešenja algebarske jednačine trećeg i četvrtog stepena, ali su suviše komplikovane. Zbog toga se, sem u specijalnim slučajevima,

algebarske jednačine trećeg i četvrog stepena rešavaju numerički. Algebarske jednačine reda $n \geq 5$ takođe se rešavaju numerički, osim u posebnim slučajevima.

Za proizvoljnu transcendentnu jednačinu egzistencija rešenja se ne može utvrditi. I kad rešenje postoji, često ga nije moguće dati u zatvorenom obliku, pomoću neke formule. To je još jedan razlog više za razvijanje numeričkih postupaka za rešavanje jednačina.

Osnovna ideja kod većine numeričkih postupaka se sastoji u određivanju **početne aproksimacije** x_0 i **iterativnog pravila** po kom se, polazeći od x_0, generiše **niz aproksimacija** $x_0, x_1, \ldots$ (iterativni niz) koji **konvergira** ka rešenju ξ posmatrane jednačine. Kada iterativni niz konvergira, kažemo da iterativni postupak koji proizvodi iterativni niz konvergira.

Za određivanje početne aproksimacije x_0 potrebno je **lokalizovati** rešenja jednačine, odnosno odrediti skup koji sadrži jedno ili više rešenja posmatrane jednačine. Iterativna pravila se mogu definisati na različite načine. Važan kriterijum za izbor iterativnog pravila je brzina konvergencije koja se meri **redom konvergencije**. U praksi se izračunava samo konačan broj aproksimacija, pa je potrebno oceniti i **grešku** aproksimacije. Pored toga, bitne karakteristike nekog iterativnog pravila su njegova složenost i osetljivost na greške zaokruživanja do kojih dolazi usled rada sa mašinskim brojevima.

Da bi se numeričkim postupkom dobila približna rešenja jednačine, potrebno je

- odrediti dovoljno male intervale koji sadrže rešenja posmatrane jednačine,
- izračunati rešenja sa zadatom tačnošću.

U daljem radu sa $[a, b]$ ćemo označavati zatvoren interval u skupu realnih brojeva, a sa $C[a, b]$ skup neprekidnih funkcija na intervalu $[a, b]$. Skup neprekidnih funkcija čiji su svi izvodi od prvog do $k-$tog zaključno neprekidni na intervalu $[a, b]$ označavaćemo sa $C^k[a, b]$.

4.1 Lokalizacija rešenja

4.1.1 Grafička lokalizacija rešenja

Za određivanje oblasti u kojoj se nalazi samo jedan koren jednačine $f(x) = 0$ ponekad su pogodni grafički postupci. Naročito se koriste sledeći.

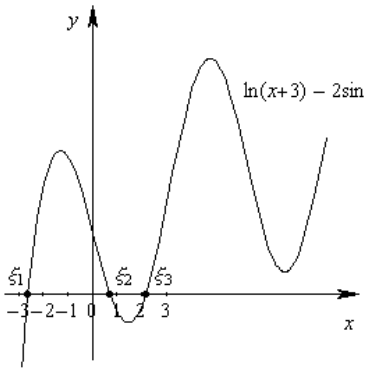
- Nacrta se grafik funkcije f. Apscise tačaka preseka grafika sa $x-$osom su približne vrednosti korena jednačine $f(x) = 0$. Ukoliko je crtež preciznije urađen, utoliko se sa sve većom tačnošću dobijaju približna rešenja posmatrane jednačine.
- Umesto jednačine $f(x) = 0$ posmatrati se njoj ekvivalentna jednačina $g(x) = h(x)$ za pogodno izabrane funkcije f i g . Apscise presečnih tačaka grafika funkcije f i h su približna rešenja jednačine $f(x) = 0$.

Definicija 4.1. *Za dve jednačine kažemo da su ekvivalentne na intervalu $[a, b]$ ako su rešenja koja pripadaju intervalu $[a, b]$ jedne jednačine rešenja druge jednačine i obrnuto.*

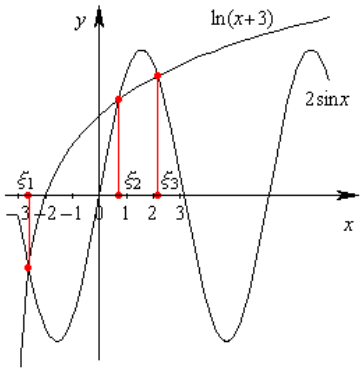
Primer 4.1. *Grafički lokalizovati rešenja jednačine*

$$\ln(x + 3) - 2 \sin x = 0.$$

Grafik funkcije $f(x) = \ln(x+3) - 2\sin x$ *je prikazan na slici* 3.



Slika 3.



Slika 4.

Jednačina $f(x) = 0$ *se može napisati u ekvivalentnom obliku* $\ln(x+3) = 2\sin x$, *a funkcije* $h(x) = \ln(x+3)$ *i* $g(x) = 2\sin x$ *su jednostavnije za crtanje. Njihovi grafici su prikazani na slici* 4. *Posmatrana jednačina ima tri rešenja koje se nalaze u intervalima:* $(-3,-2)$, $(0,1)$ *i* $(2,3)$.

Za određivanje intervala u kojima se nalaze rešenja jednačine, pored grafičkih postupaka koriste se i analitički. Tako možemo koristiti poznatu osobinu neprekidnih funkcija

Teorema 4.1. *O* ***suprotnim znacima.*** *Neka je* $f(x)$ *neprekidna funkcija na intervalu* $[a,b]$. *Ako* $f(a)$ *i* $f(b)$ *imaju suprotne znake, tj. ako je* $f(a)f(b) < 0$, *onda postoji bar jedno rešenje jednačine* $f(x) = 0$ *u intervalu* (a,b).

Koristeći navedenu teoremu i druge osobine funkcije f u mnogim slučajevima lako se mogu prciznije lokalizovati koreni jednačine $f(x) = 0$.

Primer 4.2. *Funkcija*

$$f(x) = \ln(x+3) - 2\sin x$$

je definisana i neprekidna za $x > -3$. *Kako je*

$$f(-2.7) < 0 \quad i \quad f(-2) > 0$$

funkcija f *ima bar jednu nulu u intervalu* $(-2.7,-2)$. *Očigledno u datom intervalu funkcija* $2\sin x$ *monotono opada, a funkcija* $\ln(x+1)$ *za* $x > -3$ *monotono raste, što znači da u intervalu* $(-2.7,-2)$ *postoji samo jedan koren jednačine* $f(x) = 0$.

Teorema o suprotnim znacima daje dovoljan uslov za egzistenciju rešenja jednačine na posmatranom intervalu. Taj uslov nije i potreban, o čemu treba voditi računa posebno kod **višestrukih** rešenja jednačine.

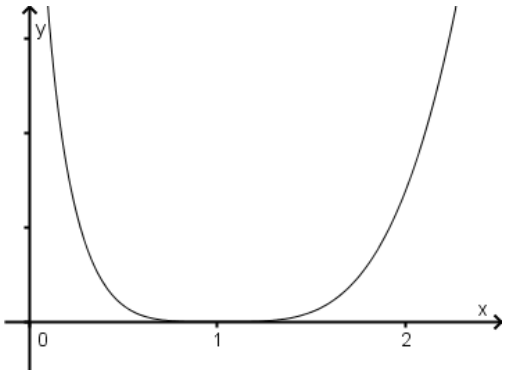
Definicija 4.2. *Broj* ξ *je rešenje višestrukosti* k *jednačine* $f(x) = 0$ *ako je*

$$f(x) = (x-\xi)^k g(x), \qquad g(\xi) \neq 0,$$

pri čemu je funkcija $g(x)$ *ograničena u tački* ξ.

Ukoliko su izvodi funkcije f, $f', f'', \ldots, f^{(k-1)}$, defisani u tački ξ, onda je $f'(\xi) = f''(\xi) = \cdots = f^{(k-1)}(\xi) = 0$. Ako je $k = 1$, broj ξ se naziva **prosto** rešenje, a ako je $k > 1$ naziva se višestruko rešenje. Tabeliranjem funkcije nije uvek moguće lokalizovati rešenje, kao što pokazuje sledeći primer.

Primer 4.3. *Neka je data jednačina* $(x-1)^3 \ln x = 0$. *Ova jednačina ima rešenje* $\xi = 1$ *višestrukosti* 4. *Na osnovu teoreme o suprotnim znacima u okolini tačke* $\xi = 1$ *ne može se zaključiti da rešenje postoji. Grafička lokalizacija rešenja ove jednačine međutim daje dobar rezultat, kao što je prikazano na slici* 5.



Slika 5.

4.1.2 Lokalizacija nula polinoma

Jednačina $f(x) = 0$, gde je

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

a koeficijenti a_n, $a_1, \ldots,$ a_n realni brojevi i $a_n \neq 0$, očigledno, može da se zameni ekvivalentnom jednačinom $p_n(x) = 0$, gde je polinom p_n normalizovan

$$p_n(x) = x^n + b_{n-1} x^{n-1} + \cdots + b_1 x + b_0,$$

sa koeficijentima

$$b_k = \frac{a_k}{a_n}, \qquad k = 0, 1, \ldots n-1.$$

Takođe, pretpostavićemo da je $b_0 \neq 0$. Ovaj uslov ne umanjuje opštost razmatranja, jer u slučaju $b_0 = 0$ jedno rešenje posmatrane jednačine je $\xi = 0$, pa se problem lokalizacije nula polinoma p_n svodi na lokalizaciju nula polinoma p_{n-1}, koji je određen tako da važi $p_n(x) = x p_{n-1}(x)$:

$$p_{n-1}(x) = x^{n-1} + b_{n-1} x^{n-2} + \cdots + b_2 x + b_1.$$

Ako je $b_1 \neq 0$ imamo polazni problem, ali stepena $n-1$. U slučaju kada je $b_1 = 0$, postupajući kao i ranije, dobijamo da je $p_{n-1}(x) = x p_{n-2}(x)$, gde je

$$p_{n-2}(x) = x^{n-2} + b_{n-1} x^{n-3} + \cdots + b_3 x + b_2.$$

Sada posmatramo lokalizaciju nula polinoma p_{n-2}. Ako je $b_2 \neq 0$ imamo polazni problem, ali stepena $n-2$.

Nastavljajući tako ili ćemo dobiti da su svi koeficijenti b_0, $b_1, \ldots, b_{n-1}$ jednaki nuli, pa je $\xi = 0$ jedina nula polaznog polinoma, ili će neko b_k biti različito od nule, pa polazni polinom ima i nule različite od 0. Ove nule su tada nule polinoma

$$p_{n-k}(x) = x^{n-k} + b_{n-1} x^{n-k-1} + \cdots + b_{k+1} x + b_k.$$

U daljem radu posmatraćemo lokalizaciju nula polinoma p_n uz uslov $b_0 \neq 0$. Na osnovu fundamentalne teoreme algebre, svaki polinom stepena n ima n nula u skupu kompleksnih brojeva, pri čemu se višestruke nule broje onoliko puta kolika im je višestrukost.

Teorema 4.2. ***Konjugovane nule.*** *Ako je kompleksni broj $\xi = \alpha + i\beta$ nula polinoma višestrukosti k, onda je i broj $\overline{\xi} = \alpha - i\beta$ nula istog polinoma višestrukosti k.*

Posledica 4.3. ***Realne nule.*** *Polinom neparnog stepena ima bar jednu realnu nulu.*

U sledećoj teoremi dajemo krug koji sadrži sve nule polinoma. Centar kruga je u koordinatnom početku, a poluprečnik se određuje pomoću koeficijenata polinoma.

Teorema 4.4. ***Krug sa nulama polinoma.*** *Neka je ξ ma koja nula polinoma*

$$p_n(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0 = 0, \qquad b_0 \neq 0.$$

Tada je

$$|\xi| < 1 + A,$$

gde je

$$A = \max\left\{|b_i| : i = 0, 1, \ldots n-1\right\}.$$

Dokaz. Pretpostavićemo, suprotno tvrđenju teoreme, da za neko rešenje ξ važi $|\xi| \geq 1 + A$. Kako je $b_0 \neq 0$, sledi da je $A \neq 0$, odnosno $|\xi| \geq 1$. Tada je

$$0 = |p_n(\xi)| = \left|\xi^n + b_{n-1}\xi^{n-1} + b_{n-2}\xi^{n-2} + \cdots + b_1\xi + b_0\right| \geq |\xi|^n - |b_{n-1}\xi^{n-1} + b_{n-2}\xi^{n-2} + \cdots + b_1\xi + b_0|.$$

Kako je

$$\left|b_{n-1}\xi^{n-1} + b_{n-2}\xi^{n-2} + \cdots + b_1\xi + b_0\right| \leq A\left(|\xi|^{n-1} + |\xi|^{n-2} + \cdots + |\xi| + 1\right),$$

to je

$$0 = |p_n(\xi)| \ \geq |\xi|^n - A\frac{|\xi|^n - 1}{|\xi| - 1} > |\xi|^n - A\frac{|\xi|^n}{|\xi| - 1} = \frac{|\xi|^n\left(|\xi| - 1 - A\right)}{|\xi| - 1} \geq 0.$$

Iz poslednje relacije sledi da je $|p_n(\xi)| > 0$, što je suprotno sa pretpostavkom da je ξ rešenje posmatrane jednačine. Tako je dokazano da za svako rešenje važi $|\xi| < 1 + A$. ■

Neka je

$$q_n(y) = 1 + b_{n-1}y + \cdots + b_1 y^{n-1} + b_0 y^n.$$

Nule ovog polinoma su recipročne vrednosti nula polinoma $p_n(x)$. Za proizvoljnu nulu ξ polinoma $p_n(x)$ broj $\mu = 1/\xi$ je nula polinoma $q_n(y)$. Jednačina $q_n(y) = 0$ se može zapisati u normalizovanom obliku

$$y^n + c_{n-1}y^{n-1} + \cdots + c_1 y + c_0 = 0,$$

gde je

$$c_0 = \frac{1}{b_0}, \qquad c_{n-i} = \frac{b_i}{b_0}, \quad i = 1, 2, \ldots, n-1.$$

Na osnovu prethodne teoreme sledi da za svaku nulu μ polinoma q_n važi

$$|\mu| < 1 + \max\left\{|c_i| : i = 0, 1, \ldots n-1\right\} = 1 + \frac{d}{|a_0|},$$

gde je

$$d = \max\left\{1, \max\left\{|b_i| : i = 1, 2, \ldots n-1\right\}\right\}.$$

Na osnovu toga je

$$|\xi| = \frac{1}{|\mu|} > \frac{|b_0|}{|b_0| + d} = 1 - a.$$

Ako uzmemo da je

$$a = \frac{d}{|a_0| + d}$$

dobijamo

$$1 - a < |\xi| .$$

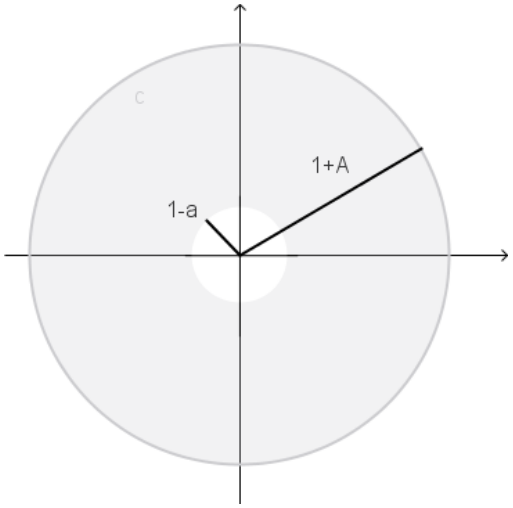
Imajući u vidu ovaj rezultat i prethodnu teoremu, možemo zaključiti da za ma koju nulu ξ polinoma p_n važi

$$1 - a < |\xi| < 1 + A.$$

Primer 4.4. *Polinom* $p_4(x) = x^4 - x^3 - x^2 - x + 0.8$ *ima u skupu kompleksnih brojeva četiri nule:* $x_1 = -0.621761 - 0.733811i, \quad x_2 = -0.621761 + 0.733811i, \quad x_3 = 0.494431, \quad x_4 = 1.74909$. *Rešenja jednačine* $p_4(x) = 0$ *na osnovu prethodne teoreme pripadaju prstenu koji određuju poluprečnici*

$$1 + A = 2 \quad i \quad 1 - a = \frac{4}{9}$$

jre je $A = 1$ *i* $a = \frac{5}{9}$. *Realna rešenja pripadaju skupu* $(-2, -\frac{4}{9}) \cup (\frac{4}{9}, 5)$. *Na slici 6 prikazane su nule polinoma* p_4 *i prsten kojem pripadaju.*



Slika 6.

Primer 4.5. *Polinom* $p_4(x) = 8x^4 - 8x^2 - 32x + 1$ *ima u skupu kompleksnih brojeva četiri nule. Da bismo mogli primeniti prethodnu teoremu posmatrajmo polinom*

$$q_4(x) = \frac{1}{8} p_4(x) = x - x^2 - 4x + \frac{1}{8}.$$

Polinomi p_4 *i* q_4 *imaju iste korene. Neka je*

$$A = \max\left\{\frac{1}{8}, 4, 1\right\} = 4 \quad i \quad d = \max\{1, 4, 1\} = 4.$$

Tada, prema prethodnoj teoremi, moduli svih korena polinoma q_4 *pripadaju intervalu*

$$\left(1 - \frac{4}{\frac{1}{8} + 4}, 1 + 4\right) = \left(\frac{1}{33}, 5\right).$$

Na osnovu toga znamo da realni koreni polinoma p_4, *ako ih ima, pripadaju skupu*

$$(-5, -\frac{1}{33}) \cup (\frac{1}{33}, 5).$$

Realni koreni polinoma p_4 *su* $0.0310\ldots$, *i* $1.788\ldots$, *a* $\frac{1}{33} \approx 0.030303$.

Pomoću sledeće teoreme mogu se utvrditi intervali koji sadrže sve realne nule polinoma, što je pokazano u sledećem primeru. Dokaz ove teoreme je sličan dokazu prethodne, te ga izostavljamo.

Teorema 4.5. *Neka je b maksimum apsolutnih vrednosti negativnih koeficijenata polinoma*

$$p_n(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0 = 0, \qquad b_0 \neq 0,$$

a b_{n-m} neka je prvi negativan koeficijent u nizu

$$b_{n-1},\ b_{n-2}, \ldots, b_1,\ b_0.$$

Tada je svaka pozitivna nula polinoma p_n manja od

$$1 + \sqrt[m]{b}.$$

Ako polinom p_n nema negativnih koeficijenata, nema ni pozitivnih nula.

Primer 4.6. *Koristeći prethodnu teoremu odredićemo intervale kojima pripadaju pozitivne nule polinoma*

$$p_4(x) = x^4 + 2x^3 - 6x^2 - 2x + 1.$$

Sa oznakama iz prethodne teoreme imamo u ovom slučaju $b = 6$ i $m = 2$, te za svaki pozitivan koren x datog polinoma važi $x < 1 + \sqrt{6} < 3.45$.
Recipročne vrednosti nula polinoma p_4 (0 nije nula polinoma ovog polinoma) dobićemo iz jednačine

$$x^4 p_4\left(\frac{1}{x}\right) = 0,$$

tj. one su nule polinoma

$$q_4(x) = 1 + 2x - 6x^2 - 2x^3 + x^4.$$

Za svaku pozitivnu nulu z polinoma q_4 važi $z < 1 + 6 = 7$, jer je u ovom slučaju $b = 6$ i $m = 1$. Neka je z najveća pozitivna nula polinoma q_4. Tada je $1/z$ najmanja pozitivna nula polinoma p_4 i za nju važi $z > 1/7 > 0.142$. Na taj način dobijamo da pozitivne nule polinoma p_4 pripadaju intervalu $(1/7, 1 + \sqrt{6})$. Umesto navedenog, zamenjujući granice intervala približnim decimalnim razlomcima, možemo uzeti nešto širi interval $(0.14, 3.45)$ kao interval koji sadrži sve pozitivne nule polinoma p_4

Negativne nule polinoma p_n su pozitivna rešenja jednačine $p_n(-x) = 0$. Znači, negativne nule polinoma $p_4(x) = x^4 + 2x^3 - 6x^2 - 2x + 1$ su su pozitivne nule polinoma $q_4(x) = p_4(-x) = x^4 - 2x^3 - 6x^2 + 2x + 1$. Koristeći ponovo prethodnu teoremu dobijamo da pozitivni koreni ove jednačine pripadaju intervalu

$$\left(\frac{1}{1+\sqrt{6}}, 1 + 6\right).$$

Odatle sledi da negativne nule polinoma p_4 pripadaju intervalu

$$\left(-7, -\frac{1}{1+\sqrt{6}}\right),$$

ili nešto širem intervalu $(-7, -0.289)$. Radi upoređenja, navedimo približne vrednosti nula polinoma p_4: $-3.52015\ldots,\ -0.557537\ldots,\ 0.284079\ldots,\ 1.7936\ldots$.

4.2 Pojam približnog rešenja. Lagranžova ocena greške

Rešenja neke jednačine nije uvek moguće izračunati tačno. U takvim slučajevima tražimo približno rešenje jednačine kao broj koji se ne razlikuje mnogo od tačnog rešenja. Da bismo to postigli moramo, pored postupka kojim određujemo približno rešenje, znati da ocenimo grešku s kojom smo izračunali približno rešenje. Na primer, ako znamo da jednačina $\ln(x-1)-\sin x=0$ ima u intervalu $(2,3)$ rešenje α, možemo bilo koji broj x_0 iz tog intervala uzeti kao približno rešenje. Za greške $|\alpha-x_0|$ možemo reći samo da je manja od dužine intervala u kojem se nalaze i α i x_0. U konkretnom slučaju ta dužina je 1. Svakako, znajući još neke osobine funkcije čije nule tražimo, možemo dobiti i bolju ocenu greške $|\alpha-x_0|$. Jednu od mogućnosti da se oceni apsolutna vrednost razlike tačnog rešenja i približnog rešenja (bez obzira kako je približno rešenje određeno) pruža Lagranžova teorema uz dodatni uslov.

Teorema 4.6. ***Lagranžova ocena greške.*** *Neka je funkcija f diferencijabilna na intervalu $[a,b]$ i neka je za pozitivnu konstantu m*

$$|f'(x)|\geq m, \qquad za\ x\in[a,b].$$

Ako je $\alpha\in[a,b]$ nula funkcije f, onda za svako $x\in[a,b]$ važi

$$|x-\alpha|\leq\frac{|f(x)|}{m}.$$

Dokaz. Prema Lagranžovoj teoremi postoji $\tau\in(\min\{x,\alpha\},\max\{x,\alpha\})$ takvo da važi

$$f(x)-f(\alpha)=f'(\tau)(x-\alpha).$$

Kako je $f(\alpha)=0$, dobijamo

$$|f(x)|=|f'(\tau)|\,|x-\alpha|\geq m|x-\alpha|,$$

a odatle tvrđenje teoreme, jer je m pozitivan broj. ■

Primer 4.7. *Već smo rekli da jednačina $\ln(x-1)-\sin x=0$ ima u intervalu $(2,3)$ rešenje α. Neka je x_0 neki broj iz tog intervala. Kako je $f'(x)=\frac{1}{x-1}-\cos x$ prvi izvod funkcije $f(x)=\ln(x-1)-\sin x$ i*

$$\left|\frac{1}{x-1}-\cos x\right|>1.4 \qquad za\ \ x\in[2,3],$$

sledi prema prethodnoj teoremi da je

$$|\alpha-x_0|\leq\frac{|f(x_0)|}{1.4}.$$

Za $x_0=2.5$ dobijamo

$$|\alpha-2.5|\leq\frac{|-0.193007...|}{1.4}<0.138.$$

Za $x_0=2.6$ dobijamo

$$|\alpha-2.6|\leq\frac{|-0.0454977...|}{1.4}<0.0325.$$

Tačno rešenje posmatrane jednačine je $2.630664\ldots$.

Ako približno rešenje jednačine tražimo nekim određenim postupkom, možemo dobiti i bolje ocene apsolutne vrednosti razlike tačnog i približnog rešenja. Jedan od načina je da sužavamo interval u kojem se nalazi tačno rešenje.

4.3 Postupak polovljenja

Ako je f neprekidna funkcija na intervalu $[a,b]$ i ako važi $f(a)f(b)<0$, tada f ima bar jednu nulu u (a,b). Ovo poznato tvrđenje leži u osnovi postupka polovljenja. Naime, polazeći od intervala $[a_0,b_0]=[a,b]$, formira se niz novih intervala $[a_k,b_k]$ za koje važi

$$[a_k,b_k]\subset[a_{k-1},b_{k-1}] \quad \text{i} \quad f(a_k)f(b_k)<0.$$

Tada bar jedna nula funkcije f pripada svim intervalima.

Pretpostavimo da je funkcija f neprekidna na intervalu $[a,b]$ i da je $f(a)f(b)<0$. Radi jednostavnosti uzmimo da je $f(a)<0$ i $f(b)>0$, što ne umanjuje opštost razmatranja. Naime, ako bi bilo $f(a)>0$, a $f(b)<0$, umesto jednačine $f(x)=0$, posmatrali bi jednačinu $-f(x)=0$ i funkcija $F(x)=-f(x)$ bila bi neprekidna na istom intervalu $[a,b]$ i za nju bi važilo $F(a)<0$ i $F(b)<0$.

Neka je

$$a_0=a, \qquad b_0=b.$$

Posmatrajmo tačku

$$c_0=\frac{a_0+b_0}{2}$$

koja polovi interval $[a_0,b_0]$. Ako je $f(c_0)=0$ računanje prekidamo jer smo odredili jedno rešenje jednačine $f(x)=0$. Ako je $f(c_0)\neq 0$ određujemo novi interval $[a_1,b_1]$ na sledeći način:

$$\begin{aligned} a_1 &= c_0,\ b_1=b_0 \quad \text{ako je } f(c_0)<0,\\ a_1 &= a_0,\ b_1=c_0 \quad \text{ako je } f(c_0)>0. \end{aligned}$$

Sada imamo $f(a_1)<0$, $f(b_1)>0$, što znači da je jedna nula funkcije f u intervalu $[a_1,b_1]$ i da važi

$$a_0\le a_1<b_1\le b_0.$$

Posmatrajmo tačku

$$c_1=\frac{a_1+b_1}{2}$$

koja polovi interval $[a_1,b_1]$. Ako je $f(c_1)=0$ računanje prekidamo jer smo odredili jedno rešenje jednačine $f(x)=0$.Ako je $f(c_1)\neq 0$ određujemo novi interval $[a_2,b_2]$ na sledeći način:

$$\begin{aligned} a_2 &= c_1, \quad b_2=b_1 \quad \text{ako je } f(c_1)<0,\\ a_2 &= a_1, \quad b_2=c_1 \quad \text{ako je } f(c_1)>0. \end{aligned}$$

Sada imamo $f(a_2)<0$, $f(b_2)>0$, što znači da je jedna nula funkcije f u intervalu $[a_2,b_2]$ i da važi

$$a_0\le a_1\le a_2<b_2\le b_1\le b_0.$$

Produžavajući na isti način ili ćemo dobiti da je za neko k

$$c_k=\frac{a_k+b_k}{2}$$

rešenje jednačine $f(x)=0$, ili da za sve $k=0,1,\ldots$ važi $f(c_k)\neq 0$. Ako je $f(c_k)=0$ postupak se prekida. U drugom slučaju, produžavajući postupak dobijamo dva beskonačna niza brojeva $a_0,\ a_1,\ldots,$ i $b_0,\ b_1,\ldots$ za koje važi

$$a_0\le a_1\le\cdots\le a_k<b_k\le\cdots\le b_1\le b_0,$$
$$f(a_k)<0,\ f(b_k)>0,$$

i

$$b_k-a_k=\frac{1}{2}(b_{k-1}-a_{k-1}),$$

za svako $k = 1, 2, \ldots$. Nizovi $a_0,\ a_1, \ldots,$ i $b_0,\ b_1, \ldots$ su beskonačni, monotoni i ograničeni, te postoje njihove granične vrednosti. Neka je

$$\alpha = \lim_{k\to\infty} a_k, \qquad \beta = \lim_{k\to\infty} b_k.$$

Zbog

$$b_k - a_k = \frac{1}{2}\left(b_{k-1} - a_{k-1}\right) = \frac{1}{4}\left(b_{k-2} - a_{k-2}\right) = \cdots = \frac{1}{2^k}(b-a),$$

dobijamo

$$\lim_{k\to\infty}\left(b_k - a_k\right) = \lim_{k\to\infty}\frac{1}{2^k}(b-a) = 0,$$

odnosno

$$\lim_{k\to\infty}\left(b_k - a_k\right) = \lim_{k\to\infty} b_k - \lim_{k\to\infty} a_k = \beta - \alpha = 0.$$

Znači, granične vrednosti posmatranih nizova su jednake, tj. $\alpha = \beta$.

Zbog neprekidnosti funkcije f sledi

$$\lim_{k\to\infty} f\left(a_k\right) = f\left(\lim_{k\to\infty} a_k\right) = f(\alpha),$$

a zbog $f\left(a_k\right) < 0,\ k = 0, 1, \ldots$ sledi $f(\alpha) \le 0$.

Analogno zaključujemo da važi

$$\lim_{k\to\infty} f\left(b_k\right) = f\left(\lim_{k\to\infty} a_k\right) = f(\beta),$$

a zbog $f\left(b_k\right) > 0,\ k = 0, 1, \ldots$ sledi $f(\beta) \ge 0$. Očigledno iz $\alpha = \beta$, $f(\alpha) \le 0$ i $f(\beta) \ge 0$ sledi $f\left(\alpha\right) = 0$.

Za niz $c_0,\ c_1, \ldots$ važi

$$a_k < c_k < b_k, \qquad k = 0, 1, \ldots$$

i

$$\alpha = \lim_{k\ \to\ \infty} a_k \le \lim_{k\to\infty} c_k \le \lim_{k\ \to\ \infty} b_k = \beta,$$

što znači da je granična vrednost niza $c_0,\ c_1, \ldots$ takođe α, odnosno nula funkcije f.

Bilo koji član niza $c_0,\ c_1, \ldots$ možemo uzeti kao aproksimaciju rešenja α jednačine $f\left(x\right) = 0$. Prema dokazanom je

$$\alpha \in \left(a_k, b_k\right), \qquad c_k = \frac{a_k + b_k}{2}, \qquad k = 0, 1, \ldots.$$

Odatle sledi

$$|\alpha - c_k| < \frac{b_k - a_k}{2} = \frac{1}{2^{k+1}}(b-a), \qquad k = 0, 1, \ldots.$$

Teorema 4.7. ***Postupak polovljenja.*** *Neka je funkcija f neprekidna na intevalu $[a, b]$, $f(a)f(b) < 0$ i neka je niz $c_0,\ c_1, \ldots$ dobijen postupkom polovljenja. Tada je ili za neko k broj c_k nula funkcije f ili je granična vrednost niza $c_0,\ c_1, \ldots$ nula funkcije f. U oba slučaja za neku nulu α funkcije f važi*

$$|\alpha - x_k| < \frac{1}{2^{k+1}}\left|b - a\right|, \quad k = 0, 1, \ldots.$$

Uslovi za konvergenciju postupka polovljenja su lako proverljivi, ali je postupak linearno konvergentan. Pored toga, ako na posmatranom intervalu jednačina ima više rešenja, postupkom polovljenja dobija se samo jedno od njih.

Primer 4.8. *Jednačina* $x^4 - 5x^3 - 12x^2 + 76x - 79 = 0$ *ima u intervalu* $(1.7, 1.8)$ *rešenje. U sledećoj tabeli prikazane su vrednosti za* a_k, c_k, b_k, $k = 1, 2, \ldots, 12$, *dobijene postupkom polovljenja.*

k	a_k	c_k	b_k	$f(c_k)$
1	1.7000000000	1.7500000000	1.8000000000	$-1.7 \cdot 10^{-1}$
2	1.7500000000	1.7750000000	1.8000000000	$5.7 \cdot 10^{-2}$
3	1.7500000000	1.7625000000	1.7750000000	$-5.2 \cdot 10^{-2}$
4	1.7625000000	1.7687500000	1.7750000000	$3.2 \cdot 10^{-3}$
5	1.7625000000	1.7656250000	1.7687500000	$-2.4 \cdot 10^{-2}$
6	1.7656250000	1.7671875000	1.7687500000	$-1.1 \cdot 10^{-2}$
7	1.7671875000	1.7679687500	1.7687500000	$-3.7 \cdot 10^{-3}$
8	1.7679687500	1.7683593750	1.7687500000	$-2.4 \cdot 10^{-4}$
9	1.7683593750	1.7685546880	1.7687500000	$1.5 \cdot 10^{-3}$
10	1.7683593750	1.7684570310	1.7685546880	$6.2 \cdot 10^{-4}$
11	1.7683593750	1.7684082030	1.7684570310	$1.9 \cdot 10^{-4}$
12	1.7683593750	1.7683837890	1.7684082030	$-2.7 \cdot 10^{-5}$

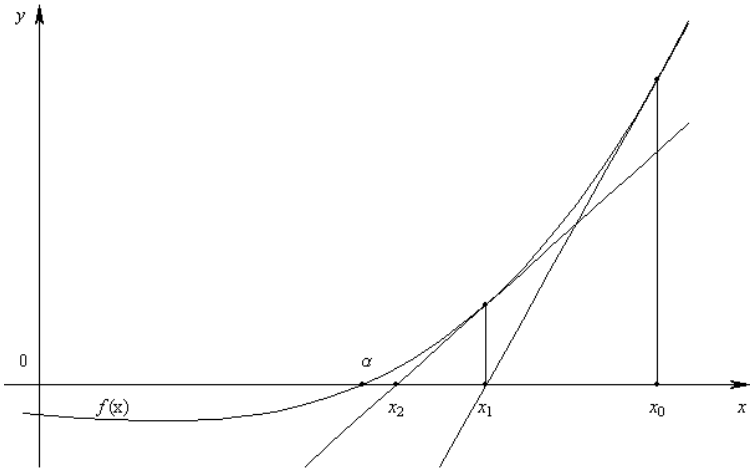
4.4 Njutnov postupak

4.4.1 Definisanje postupka

Jedan od najpopularnijih postupaka za rešavanje jednačina oblika $f(x) = 0$ je **Njutnov** (ili **Njutn-Rafsonov**) postupak. U osnovi Njutnovog postupka je aproksimacija funkcije f, čije se nule traže, linearnom funkcijom. Ova aproksimacija se posmatra u blizini nule koja se traži. Kao aproksimaciona funkcija se koristi linearna funkcija čiji grafik je tangenta funkcije f u izabranoj tački $(x_0, f(x_0))$. Presek x_1 tangente sa x osom se posmatra kao približna vrednost nule funkcije f, a zatim se u tački $(x_1, f(x_1))$ postavlja nova tangenta itd. Na primeru funkcije

$$f(x) = x^3 - 2x - 5$$

prikazaćemo Njutnov postupak, a grafički prikaz dat je na slici 7.



Slika 7.

Vidimo da je jedini realan koren date jednačine na intervalu $(2, 4)$. Posmatrajmo tangentu zadate funkcije koja sadrži tačku $(x_0, f(x_0))$, tj. $(4, f(4))$. Jednačina te tangente je.

$$y(x) = f(x_0) + f'(x_0)(x - x_0).$$

Kao aproksimaciju rešenja α jednačine $f(x) = 0$ uzimamo rešenje x_1 jednačine $y(x) = 0$. To rešenje se izračunava iz

$$f(x_0) + f'(x_0)(x - x_0) = 0$$

i glasi

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)},$$

pod pretpostavkom da je $f'(x_0) \neq 0$. Za izabrano $x_0 = 4$ na ovaj način dobijamo $x_1 = 2.8913\ldots$. Kako je sada

$$f(x_1) \neq 0 \quad \text{i} \quad f'(x_1) \neq 0,$$

opisani postupak možemo ponoviti sa tangentom na funkciju f u tački $(x_1, f(x_1))$. Tako dobijamo

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 2.31122\ldots.$$

Očigledno, ovaj postupak se može ponavljati sve dok je $f'(x_k) \neq 0$. Polazeći od $k-$te aproksimacije x_k rešenja posmatrane jednačine, sledeće aproksimacije se dobijaju po Njutnovom iterativnom pravilu

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \ldots.$$

Primer 4.9. *Za jednačinu* $x^3 - 2x - 5 = 0$, *je*

$$f(x) = x^3 - 2x - 5, \qquad f'(x) = 3x^2 - 2,$$

a Njutnov postupak glasi

$$x_k = x_k - \frac{x_k^3 - 2x_k - 5}{3x_k^2 - 2}, \quad k = 0, 1, \ldots .$$

Sa $x_0 = 4$ *posle šestog iterativnog koraka dobija se sledeća tabela*

k	0	1	2	3	4	5	6
x_k	4.000000	2.891304	2.311223	2.117035	2.094831	2.094552	2.094551
$f(x_k)$	$5.1 \cdot 10^1$	$1.3 \cdot 10^1$	$2.7 \cdot 10^0$	$2.5 \cdot 10^{-1}$	$3.1 \cdot 10^{-3}$	$4.9 \cdot 10^{-7}$	$1.4 \cdot 10^{-14}$

Radi upoređenja navodimo približnu vrednost rešenja posmatrane jednačine sa 52 decimalna mesta:

$$\alpha_0 = 2.094\ 551\ 481\ 542\ 326\ 591\ 482\ 386\ 540\ 579\ 302\ 963\ 857\ 306\ 105\ 628\ 239.$$

Ocenjujući $\delta = |\alpha_0 - x_k|$ *direktno i računajući Lagranžovu ocenu greške, sa* $m = 10$, *(u intervalu* $[2, 4]$ *prvi izvod* $f'(x) = 3x^2 - 2$ *je pozitivan i važi* $|f'(x)| \geq f'(2) = 10$*) dobijamo sledeću tabelu*

k	0	1	2	3	4	5	6
x_k	4.000000	2.891304	2.311223	2.117035	2.094831	2.094552	2.094551
$\frac{\lvert f(x_k)\rvert}{m}$	$5.1 \cdot 10^0$	$1.3 \cdot 10^0$	$2.7 \cdot 10^{-1}$	$2.5 \cdot 10^{-2}$	$3.1 \cdot 10^{-4}$	$4.9 \cdot 10^{-8}$	$1.4 \cdot 10^{-15}$
δ	$1.91 \cdot 10^0$	$7.97 \cdot 10^{-1}$	$2.17 \cdot 10^{-1}$	$2.25 \cdot 10^{-2}$	$2.80 \cdot 10^{-4}$	$4.40 \cdot 10^{-8}$	$4.45 \cdot 10^{-16}$

Pod određenim pretpostavkama niz $x_0,\ x_1,\ x_2, \ldots$ dobijen Njutnovim iterativnim postupkom imaće graničnu vrednost α, koja je rešenje jednačine $f(x) = 0$.

Da bi se Njutnov postupak mogao sprovesti, potrebno je da prvi izvod f' postoji u svim tačkama posmatranog intervala i da je $f'(x) \neq 0$ za svako x iz tog intervala.

Ako je prvi izvod f' neprekidna funkcija i ako je $f'(x) \neq 0$ na posmatranom intervalu, onda zaključujemo na osnovu teoreme o suprotnim znacima da je f' ili pozitivna ili negativna funkcija na tom intervalu. Znači, važi $f'(x) > 0$ ili $f'(x) < 0$ za x iz posmatranog intervala. Kao posledicu ovoga dobijamo da funkcija f može imati najviše jednu nulu u posmatranom intervalu.

Kad kažemo da neka funkcija na određenom intervalu ne menja znak, podrazumevaćemo da je ona ili samo pozitivna ili samo negativna na tom intervalu.

4.4.2 Konvergencija

U daljem radu će se razmotriti neki od uslova za konvergenciju Njutnovog iterativnog postupka. Pri tome se koristi i sledeća lema.

Lema 4.8. *Ako postoji konstanta γ takva da je za svako $x, y \in [a, b]$*

$$|f'(y) - f'(x)| \leq \gamma |y - x|,$$

onda važi

$$|f(y) - f(x) - f'(x)(y - x)| \leq \frac{\gamma}{2}(y - x)^2.$$

Dokaz. Na osnovu Njutn-Lajbnicove formule je

$$f(y) - f(x) = \int_x^y f'(t)dt,$$

i

$$f(y) - f(x) - f'(x)(y - x) = \int_x^y (f'(t) - f'(x))\, dt.$$

Kako je $|f'(t) - f'(x)| \leq \gamma |t - x|$, dobijamo

$$|f(y) - f(x) - f'(x)(y - x)| \leq \left| \int_x^y (f'(t) - f'(x))\, dt \right| \leq \gamma \int_x^y |t - x|\, dz = \frac{\gamma (y - x)^2}{2}.$$

■

Ako je $f \in C^2[a, b]$ i ako se γ odredi tako da je za svako $x \in [a, b]$ $\gamma \geq |f''(x)|$, tvrđenje leme sledi direktno iz Tejlorovog razvoja funkcije $f(x)$:

$$f(y) - f(x) - f'(x)(y - x) = \frac{1}{2} f''(\tau)(y - x)^2$$

za neko $\tau \in [a, b]$ i

$$|f(y) - f(x) - f'(x)(y - x)| \leq \frac{\gamma}{2}(y - x)^2.$$

Teorema 4.9. ***Lokalna konvergencija Njtnovog postupka.*** *Neka postoje konstante γ i $m > 0$ takve da svako $x, y \in (a, b)$ važi*

$$|f'(y) - f'(x)| \leq \gamma |y - x| \quad i \quad |f'(x)| \geq m.$$

Ako jednačina $f(x) = 0$ ima rešenje $\xi \in (a, b)$, onda postoji $\rho > 0$ takvo da za x_0 sa osobinom $|x_0 - \xi| \leq \rho$, niz $x_0, x_1, \ldots$ definisan Njutnovim iterativnim postupkom

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \ldots$$

postoji i konvergira ka ξ. Pri tome važi

$$|x_{k+1} - \xi| \leq \frac{\gamma}{2m} |x_k - \xi|^2, \quad k = 0, 1, \ldots.$$

Dokaz. Na osnovu pretpostavki teoreme zaključujemo da je prvi izvod f' neprekidna funkcija na intervalu $[a,b]$ i da funkcija f može imati najviše jednu nulu u tom intervalu. Neka je q, proizvoljan pozitivan broj manji od 1, a r broj za koji važi $[\xi - r, \xi + r] \subset (a,b)$. Neka je ρ određeno sa

$$\rho = \min\left\{r, q\frac{2m}{\gamma}\right\}.$$

Dokazaćemo indukcijom da je za svako $x_0 \in (\xi-\rho, \xi+\rho)$ postoji Njtnov iterativni niz $x_0,\ x_1, \ldots$ čiji svi članovi pripadaju intervalu $(\xi-\rho, \xi+\rho)$ i koji konvergira tačnom rešenju ξ jednačine $f(x) = 0$. Za $k = 0$ je

$$x_1 - \xi = x_0 - \xi - \frac{f(x_0) - f(\xi)}{f'(x_0)} = \frac{f(\xi) - f(x_0) - f'(x_0)(\xi - x_0)}{f'(x_0)}.$$

Na osnovu prethodne leme sledi

$$|f(\xi) - f(x_0) - f'(x_0)(\xi - x_0)| \leq \frac{\gamma}{2}(x_0 - \xi)^2,$$

a kako je $|f'(x_0)| \geq m$ i

$$|x_0 - \xi| < \rho \leq q\frac{2m}{\gamma},$$

to je

$$|x_1 - \xi| \leq \frac{\gamma}{2m}|x_0 - \xi|^2 \leq \frac{\gamma\rho}{2m}|x_0 - \xi| \leq q|x_0 - \xi| < \rho.$$

Znači, x_1 pripada intervalu $(\xi-\rho, \xi+\rho)$. Na isti način se dokazuje da iz $x_k \in (\xi-\rho, \xi+\rho)$ sledi $x_{k+1} \in (\xi-\rho, \xi+\rho)$ i da važi

$$|x_{k+1} - \xi| \leq \frac{\gamma}{2m}|x_0 - \xi|^2 \leq q|x_k - \xi|.$$

Na osnovu toga sledi da je niz $x_0,\ x_1, \ldots$ dobro definisan, da za svako $k = 0, 1, \ldots$ važi

$$|x_{k+1} - \xi| \leq \frac{\gamma}{2m}|x_k - \xi|^2,$$

i

$$|x_{k+1} - \xi| \leq q|x_k - \xi| \leq \cdots \leq q^{k+1}|x_0 - \xi|.$$

Kako je $0 < q < 1$ sledi

$$\lim_{k\to\infty} q^{k+1} = 0$$

i

$$\lim_{k\to\infty} |x_{k+1} - \xi| \leq \lim_{k\to\infty} q^{k+1}|x_0 - \xi|) = |x_0 - \xi| \lim_{k\to\infty} q^{k+1} = 0.$$

To znači da je ξ granična vrednost Njutnovog iterativnog niza i da važi data ocena greške. ■

Prethodna teorema daje uslove za konvergenciju Njutnovog postupka. Ovaj postupak je kvadratno konvergentan. Umesto uslova $|f'(y) - f'(x)| \leq \gamma|y - x|$ može se uzeti i jači uslov $|f''(x)| \leq \gamma$, za svako $x \in [a,b]$.

Pomoću

$$|x_{k+1} - \xi| \leq \frac{\gamma}{2m}|x_k - \xi|^2$$

može se nešto reći i o sigurnim ciframa aproksimacije x_{k+1}. Naime, ako je $\gamma \leq 2m$, a za $k-$tu aproksimaciju važi

$$|x_k - \xi| < 10^{-s},$$

onda je

$$|x_{k+1} - \xi| < 10^{-2s},$$

odnosno, u slučaju konvergentnog Njutnovog niza broj sigurnih cifara aproksimacije se udvostručuje počevši od nekog x_k. Greška se može oceniti na sledeći način.

Teorema 4.10. *Neka je funkcija f dvaput neprekidno diferencijabilna na intervalu $[a,b]$ i neka postoji konstanta $m>0$ takva da svako $x\in[a,b]$ važi*

$$|f'(x)|\geq m.$$

Ako jednačina $f(x)=0$ ima rešenje $\xi\in(a,b)$, onda postoji $\rho>0$ takvo da za x_0 sa osobinom $|x_0-\xi|\leq\rho$, niz $x_0, x_1,\ldots$ definisan Njutnovim iterativnim pravilom

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)},\quad k=0,1,\ldots$$

postoji i konvergira ka ξ. Pri tome važi

$$|x_{k+1}-\xi|\leq\frac{\gamma}{2m}|x_k-x_{k-1}|^2,\quad k=0,1,\ldots,$$

gde je γ određeno tako da za svako $x\in[a,b]$ važi

$$|f''(x)|\leq\gamma.$$

Dokaz. Na osnovu prethodne teoreme sledi da iterativni niz postoji i da konvergira rešenju posmatrane jednačine. Na osnovu Tejlorove teoreme dobija se

$$f(x_k)=f\left(x_{k-1}+(x_k-x_{k-1})\right)=f(x_{k-1})+f'(x_{k-1})(x_k-x_{k-1})+\frac{1}{2}f''(\tau_{k-1})(x_k-x_{k-1})^2,$$

gde je $\tau_{k-1}\in\left(\min\{x_{k-1},x_k\},\max\{x_{k-1},x_k\}\right)$. Po definiciji postupka je

$$f(x_{k-1})+f'(x_{k-1})(x_k-x_{k-1})=0.$$

Zbog $|f''(x)|\leq\gamma$, za svako $x\in[a,b]$, sledi

$$|f(x_k)|\leq\frac{1}{2}|f''(\tau_{k-1})|\,|x_k-x_{k-1}|^2\leq\frac{\gamma}{2}|x_k-x_{k-1}|^2.$$

Na osnovu toga i Lagranžove ocene greške dobijamo

$$|x_k-\xi|\leq\frac{|f(x_k)|}{m}\leq\frac{\gamma}{2m}|x_k-x_{k-1}|^2.$$

■

Vidimo iz teoreme da je za konvergenciju Njutnovog postupka potrebno odrediti x_0 dovoljno dobro, odnosno startna vrednost mora biti dovoljno blizu tačne vrednosti nule koji se traži. To nije uvek moguće lako postići.U sledećoj teoremi je pokazano da uz dodatnu pretpostavku o funkciji f možemo dati interval iz kojeg treba birati x_0.

Teorema 4.11. ***Globalna konvergencija.*** *Neka je funkcija f dvaput neprekidno diferencijabilan na intervalu $[a,b]$. Ako je*

$$f(a)f(b)<0,$$

a f' i f'' ne menjaju znak na intervalu $[a,b]$, onda za svako $x_0\in[a,b]$ za koje važi

$$f(x_0)f''(x_0)>0$$

Njutnov iterativni postupak konvergira jedinstvenom rešenju $\xi\in(a,b)$ jednačine $f(x)=0$ i važi

$x\in[a,b]$	$f'(x)>0$	$f'(x)<0$
$f''(x)>0$	$x_k>x_{k+1}$	$x_k<x_{k+1}$
$f''(x)<0$	$x_k<x_{k+1}$	$x_k>x_{k+1}$

Dokaz. Zbog $f(a)f(b) < 0$ i neprekidnosti funkcije f sledi da ta funkcija ima bar jednu nulu u (a, b). To rešenje je jedinstveno jer je prvi izvod konstantnog znaka na $[a, b]$, pa je funkcija f monotona. Posmatrajmo sada slučaj $f'(x) > 0$ i $f''(x) > 0$. U preostala tri slučaja dokaz se izvodi analogno.
U posmatranom slučaju mora biti $f(a) < 0$, a x_0 se bira iz intervala $(\xi, b]$, .gde je ξ jedina nula funkcije f u intervalu (a, b). Za svako $x_0 \in (\xi, b)$ je $f(x_0) > 0$ i $f'(x_0) > 0$ pa je

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} < x_0.$$

Prema Lagranžovoj teoremi je

$$0 = f(\xi) = f(x_0) + f'(\tau)(\xi - x_0),$$

za neko $\tau \in (\xi, x_0)$. Iz ove relacije dobijamo

$$\xi = x_0 - \frac{f(x_0)}{f'(\tau)}.$$

Kako je $f''(x) > 0$ funkcija f' je rastuća, te iz $x_0 > \tau$ sledi $f'(\tau) < f'(x_0)$. Kako je i $f'(x_0) > 0$ imamo

$$\xi = x_0 - \frac{f(x_0)}{f'(\tau)} < x_0 - \frac{f(x_0)}{f'(x_0)} = x_1.$$

Znači, dobili smo $\xi < x_1 < x_0$. Kako je $f(x_1) > 0$, postupajući kao u prethodnom slučaju dobijamo $\xi < x_2 < x_1 < x_0$, itd. Niz $x_0, x_1, \ldots$ je monotono opadajući i ograničen, pa ima graničnu vrednost. Zbog neprekidnosti funkcija f i f' sledi da je ta granična vrednost nula funkcije f, tj. jednaka je ξ. ■

4.4.3 Izračunavanje kvadratnog korena

Neka je dat pozitivan broj c. Kvadratni koren ovog broja možemo izračunati kao nulu finkcije $f(x) = x^2 - c$. Aproksimaciju te nule možemo izračunati Njutnovim iterativnim postupkom. U ovom slučaju Njutnov iterativni postupak glasi

$$x_{k+1} = x_k - \frac{x_k^2 - c}{2x_k} = \frac{1}{2}\left(x_k + \frac{c}{x_k}\right).$$

Očigledno za pozitivno x važi

$$f'(x) = 2x > 0 \quad \text{i} \quad f''(x) = 2 > 0.$$

Birajući interval $[a, b]$ tako da važi

$$0 < a < \sqrt{c} < b,$$

uslovi prethodne teoreme će biti ispunjeni za svako $x_0 \in [\sqrt{c}, b]$.

Primer 4.10. *Izračunaćemo približnu vrednost kvadratnog korena broja $c = 12$ Njutnovim postupkom. Neka je $x_0 = 4$. Dobijamo tabelu*

k	0	1	2	3	4
x_k	4.000000000	3.500000000	3.464285714	3.464101620	3.464101615
$f(x_k)$	$4.0 \cdot 10^{0}$	$2.5 \cdot 10^{-1}$	$1.3 \cdot 10^{-3}$	$3.4 \cdot 10^{-8}$	$-1.8 \cdot 10^{-15}$

Kako je $x_4 = x_5$ sa navedenim brojem cifara, možemo uzeti kao konačnu približnu vrednost za $\sqrt{12}$ broj $x_4 = 3.464101615$.

4.5 Postupak sečice

Njutnov postupak zahteva izračunavanje vrednosti prvog izvoda u svakoj iteraciji. Ako želimo to da izbegnemo, možemo koristiti **postupak sečice.** U osnovi ovog postupka je aproksimacija funkcije f, čije se nule traže, linearnom funkcijom. Ova aproksimacija se posmatra u blizini nule koja se traži. Kao aproksimaciona funkcija se koristi linearna funkcija čiji grafik prava koja sadrži tačke $(x_0, f(x_0))$ i $(x_1, f(x_1))$. Presek x_2 ove prave sa x osom se posmatra kao približna vrednost nule funkcije f. Zatim se određuje nova prava koja sdrži tačke $(x_1, f(x_1))$ i $(x_2, f(x_2))$ i određuje njen presek x_3 sa $x-$osom, itd.

Jednačina prave koja sadrži tačke $(x_0, f(x_0))$ i $(x_1, f(x_1))$ je

$$y(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0).$$

Kao aproksimaciju rešenja α jednačine $f(x) = 0$ uzimamo rešenje x_2 jednačine $y(x) = 0$. To rešenje se izračunava iz

$$f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) = 0$$

i glasi

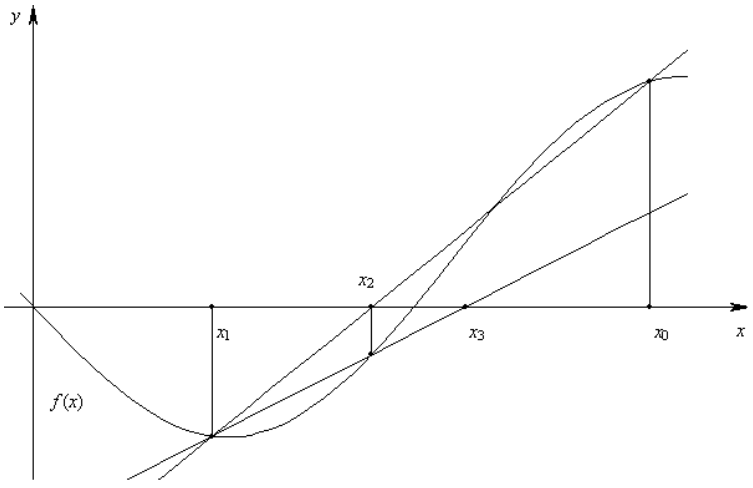
$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1),$$

pod pretpostavkom da je $x_1 \neq x_0$ i $f(x_1) \neq f(x_0)$. Na isti način, polazeći od prave koja sadrži tačke $(x_1, f(x_1))$ i $(x_2, f(x_2))$ odredićemo tačku x_3 kao presek te prave i $x-$ose itd. Znači, za svako $k = 1, 2, \ldots$ odredićemo presek x_{k+1} prave koja sadrži tačke $(x_k, f(x_k))$ i $(x_{k-1}, f(x_{k-1}))$ i $x-$ose prema

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k),$$

pod pretpostavkom da je $x_k \neq x_{k-1}$ i $f(x_k) \neq f(x_{k-1})$.

Geometrijsko tumačenje postupka sečice je sledeće. Linearni interpolacioni polinom sa čvorovima x_0 i x_1 je sečica krive $f(x)$. U preseku te prave i $x-$ose dobija se aproksimacija x_2, kao što je prikazano na slici 8.



Slika 8.

Uslovi za konvergenciju postupka sečice dati su u sledećoj teoremi.

Teorema 4.12. *Neka je funkcija f dvaput neprekidno diferencijabilan na intervalu $[a, b]$ i neka postoji konstanta $m > 0$ takva da svako $x \in [a, b]$ važi*

$$|f'(x)| \geq m.$$

Ako jednačina $f(x) = 0$ *ima rešenje* $\xi \in (a, b)$, *onda postoji pozitivno* $\rho < \frac{2m}{\gamma}$ *takvo da za svako* x_0 *i* x_1 *za koje važi* $x_0 \neq x_1$, $|x_0 - \xi| \leq \rho$, *i* $|x_0 - \xi| \leq \rho$ *niz* $x_0, x_1, \ldots$ *definisan postupkom sečice*

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k), \quad k = 1, 2, \ldots$$

postoji i konvergira ka ξ. *Pri tome važi*

$$|x_k - \xi| \leq \frac{2m}{\gamma} q^{F_k}, \quad k = 0, 1, \ldots,$$

i

$$|x_k - \xi| \leq \frac{\gamma}{2m} |x_{k-1} - x_k| \, |x_{k-2} - x_k|, \quad k = 2, 3, \ldots,$$

gde je

$$q = \frac{\gamma}{2m} \rho,$$

konstanta γ *je određena tako da za svako* $x \in [a, b]$ *važi*

$$|f''(x)| \leq \gamma,$$

a F_k *su Fibonačijevi brojevi,*

$$F_0 = F_1 = 1, \quad F_k = F_{k-1} + F_{k-2}, \quad k = 2, 3, \ldots.$$

Kako za Fibonačijeve brojeve važi

$$F_k = \frac{1}{\sqrt{5}} \left(\lambda_1^{k+1} - \lambda_2^{k+1} \right), \qquad \lambda_1 = \frac{1}{2} \left(1 + \sqrt{5} \right), \qquad \lambda_2 = \frac{1}{2} \left(1 - \sqrt{5} \right),$$

broj λ_1 postaje dominantan za veliko k, pa se može pokazati da je red konvergencije postupka sečice $\lambda_1 \approx 1.618\ldots$.

Ukoliko se za funkciju f pretpostavi samo neprekidnost i egzistencija bar jednog rešenja na intervalu $[a, b]$, onda se rešenje može odrediti i **primitivnim postupkom sečice,** definisanim pomoću iterativnog pravila

$$x_{k+1} = x_k - \frac{x_k - x_s}{f(x_k) - f(x_s)} f(x_k), \quad k = 1, 2, \ldots, \tag{4.1}$$

gde je

$$s = s(k) = \max \{ i \in \{0, 1, \ldots, k-1\} : f(x_k) f(x_i) < 0 \}.$$

Teorema 4.13. *Neka je* $f \in C(D)$ *i* $f(a)f(b) < 0$. *Tada primitivni postupak sečice dat iterativnim pravilom, sa* $x_0 = a$, $x_1 = b$, *konvergira ka rešenju* $\xi \in D$ *jednačine* $f(x) = 0$.

Sledeća teorema o konvergenciji postupka sečice je analogna odgovarajućoj teoremi za Njutnov postupak, pa je dokaz izostavljen.

Teorema 4.14. ***Globalna konvergencija.*** *Neka je funkcija* f *dvaput neprekidno diferencijabilna na intervalu* $[a, b]$. *Ako je*

$$f(a)f(b) < 0,$$

a f' *i* f'' *ne menjaju znak na intervalu* $[a, b]$, *onda za svako* $x_0 \in [a, b]$ *za koje važi*

$$f(x_0)f''(x_0) > 0 \quad i \quad f(x_1)f''(x_1) < 0$$

primitivni postupak sečice

$$x_{k+1} = x_k - \frac{x_k - x_0}{f(x_k) - f(x_0)} f(x_k), \quad k = 1, 2, \ldots,$$

konvergira jedinstvenom rešenju $\xi \in (a, b)$ *jednačine* $f(x) = 0$ *i važi*

$x \in [a, b]$	$f'(x) > 0$	$f'(x) < 0$
$f''(x) > 0$	$x_k < x_{k+1}$	$x_k > x_{k+1}$
$f''(x) < 0$	$x_k > x_{k+1}$	$x_k < x_{k+1}$

4.6 Opšti iterativni postupak

4.6.1 Teorema o nepokretnoj tački

U ovom delu posmatraćemo jednačinu oblika $x = \varphi(x)$, pri čemu je φ neprekidna realna funkcija na zatvorenom intervalu $[a, b]$. Broj $\alpha \in [a, b]$ za koji važi $\alpha = \varphi(\alpha)$ je rešenje jednačine $x = \varphi(x)$ i naziva se **nepokretna tačka** funkcije φ. Kako smo do sada posmatrali jednačinu oblika $f(x) = 0$, to ćemo u sledećoj teoremi dati uslove za ekvivalentnost jednačina $f(x) = 0$ i $x = \varphi(x)$.

Teorema 4.15. ***Ekvivalencija jednačina.*** *Neka je $g(x) \neq 0$ za $x \in [a, b]$. Tada su jednačine $f(x) = 0$ i $x = \varphi(x)$ sa $\varphi(x) = x - g(x)f(x)$ ekvivalentne na intervalu [a,b].*

Dokaz. Ako je $\alpha \in [a, b]$ rešenje jednačine $f(x) = 0$, tada sledi $\varphi(\alpha) = \alpha$. Obrnuto, ako je $\alpha \in [a, b]$ rešenje jednačine $x = \varphi(x)$, sa $\varphi(x) = x - g(x)f(x)$ sledi $g(\alpha)f(\alpha) = 0$. Odatle, zbog $g(\alpha) \neq 0$, imamo $f(\alpha) = 0$. ■

Primer 4.11. *Za jednačinu*

$$f(x) = 5x^3 - 20x + 3 = 0$$

navešćemo nekoliko ekvivalentnih jednačina oblika $x = \varphi(x)$:

- $x = x + 5x^3 - 20x + 3$,
- $x = \dfrac{5x^3 + 3}{20}$,
- $x = \sqrt[3]{\dfrac{20x - 3}{5}}$,
- $x = \dfrac{-3}{5x^2 - 20}, \quad x \notin \{-2, 2\}$,
- $x = x - a\left(5x^3 - 20x + 3\right), \quad a \neq 0.$

Neka je x_0 proizvoljan broj iz intervala $[a, b]$. Obrazujmo niz $x_0, x_1, \ldots$ prema

$$x_{k+1} = \varphi(x_k).$$

Kako je φ definisano na intervalu $[a, b]$, ovaj niz je moguće formirati ako je $\varphi(x_k) \in [a, b]$ za svako $k = 0, 1, \ldots$.Ovaj uslov je ispunjen ako funkcija φ preslikava interval $[a, b]$ u samog sebe. Ako je uz to funkcija φ neprekidna, a niz $x_0, x_1, x_2, \ldots$ ima graničnu vrednost, tj. za neko α važi

$$\lim_{k\to\infty} x_k = \alpha,$$

onda je α rešenje jednačine $x = \varphi(x)$. Zaista, iz $\varphi(x) \in [a, b]$ za svako $x \in [a, b]$ sledi da α pripada intervalu $[a, b]$, a zbog neprekidnosti funkcije φ sledi

$$\alpha = \lim_{k\to\infty} x_k = \lim_{k\to\infty} \varphi(x_{k-1}) = \varphi(\lim_{k\to\infty} x_{k-1}) = \varphi(\alpha).$$

Znači, ako niz $x_0, x_1, x_2, \ldots$ konvergira, njegova granična vrednost α je rešenje jednačine $x = \varphi(x)$, a članovi tog niza aproksimiraju to rešenje.

Teorema 4.16. ***Egzistencija nepokretne tačke.*** *Neka je φ neprekidna funkcija na intervalu $[a, b]$ i $\varphi(a), \varphi(b) \in [a, b]$. Tada funkcija φ ima nepokretnu tačku $\alpha \in [a, b]$.*

Dokaz. Kako je $\varphi(a) \in [a,b]$ i $\varphi(b) \in [a,b]$, sledi da je $\varphi(a) \geq a$ i $\varphi(b) \leq b$. Ako je $\varphi(a) = a$ ili $\varphi(b) = b$, tvrđenje je dokazano. Neka je $\varphi(a) = A > a$ i $\varphi(b) = B < b$. Funkcija $f(x) = x - \varphi(x)$ je neprekidna na intervalu $[a,b]$ i

$$f(a) = a - A < 0, \quad f(b) = b - B > 0,$$

pa na osnovu teoreme o suprotnim znacima sledi da jednačina $f(x) = 0$ ima bar jedno rešenje u $[a,b]$. Neka je to rešenje α. Tada je $0 = f(\alpha) = \alpha - \varphi(\alpha)$, odnosno $\alpha = \varphi(\alpha)$. ■

Posledica 4.17. *Neka funkcija φ preslikava interval $[a,b]$ u samog sebe, tj. za svako $x \in [a,b]$ je $\varphi(x) \in [a,b]$. Tada postoji nepokretna tačka $\alpha \in [a,b]$ funkcije φ.*

Veliki značaj u teoriji nepokretne tačke imaju kontraktivna preslikavanja.

Definicija 4.3. ***Lipšicov uslov.*** *Funkcija φ je zadovoljava Lipšicov uslov na intervalu $[a,b]$ ako postoji konstanta γ takva da za svako $x, y \in [a,b]$ važi*

$$|\varphi(x) - \varphi(y)| \leq \gamma |x - y|.$$

Konstanta γ iz prethodne definicije naziva se **Lipšicova konstanta.** Ako je $\gamma <$ onda se ova konstanta naziva **konstanta kontrakcije**, a funkcija φ se naziva **kontrakcija** ili **kontraktivno preslikavanje.** Ako funkcija φ ima u intervalu $[a,b]$ neprekidan prvi izvod, tj. ako je za svako $x \in [a,b]$

$$|\varphi'(x)| \leq \gamma,$$

onda je na osnovu Lagranžove teoreme sledi

$$|\varphi(x) - \varphi(y)| \leq \gamma |x - y|,$$

za svako $x, y \in [a,b]$. Drugim rečima, ako je funkcija φ neprekidno diferencijabilna na intervalu $[a,b]$, onda ona na tom intervalu zadovoljava Lipšicov uslov.

Teorema 4.18. ***Jedinstvenost neprekidne tačke.*** *Neka je φ kontrakcija na $[a,b]$. Jednačina $\varphi(x) = x$ ima najviše jedno rešenje u $[a,b]$.*

Dokaz. Ako tvrđenje teoreme nije tačno, onda postoje bar dva različita rešenja $\alpha, \beta \in [a,b]$ jednačine $x = \varphi(x)$. Tada je

$$|\alpha - \beta| = |\varphi(\alpha) - \varphi(\beta)| \leq \gamma |\alpha - \beta| < |\alpha - \beta|,$$

što je kontradikcija. Znači, jednačina $x = \varphi(x)$ ima najviše jedno rešenje. ■

Teorema 4.19. *Neka funkcija φ preslikava interval $[a,b]$ u samog sebe i neka je kontrakcija sa konstantom γ. Tada iterativni niz određen iterativnim pravilom*

$$x_{k+1} = \varphi(x_k), \quad k = 0, 1, \ldots$$

sa proizvoljnim $x_0 \in [a,b]$ konvergira ka jedinstvenom rešenju u intervalu $[a,b]$ jednačine $x = \varphi(x)$.

Dokaz. Na osnovu posledice teoreme *o* egzistenciji nepokretne tačke postoji jedinstveno $\alpha \in [a,b]$ takvo da je $\alpha = \varphi(\alpha)$. Neka je $x_0 \in [a,b]$ proizvoljno i

$$x_{k+1} = \varphi(x_k), \quad k = 0, 1, \ldots .$$

Tada je

$$|x_{k+1} - \alpha| = |\varphi(x_k) - \varphi(\alpha)| \leq \gamma |x_k - \alpha| \leq \gamma^2 |x_{k-1} - \alpha| \leq \cdots \leq \gamma^{k+1} |x_0 - \alpha|.$$

Kako je $\gamma < 1$, sledi

$$\lim_{k\to\infty} \gamma^{k+1} = 0,$$

odnosno

$$\lim_{k\to\infty} |x_{k+1} - \alpha| \leq |x_0 - \alpha| \lim_{k\to\infty} \gamma^{k+1} = 0$$

i

$$\lim_{k\to\infty} x_k = \alpha.$$

Kako je φ neprekidna funkcija sledi $\alpha = \varphi(\alpha)$, a time i tvrđenje teoreme. ■

Prethodna teorema je poznata pod imenom Banahov princip kontrakcije.

Primer 4.12. *Neka je data jednačina* $f(x) = 0$ *iz prethodnog primera*

$$4 - x + \sqrt[3]{\frac{x-1}{x+1}} = 0.$$

Ova jednačina ima rešenje u intervalu $[4, 5]$. *Neka je*

$$\varphi(x) = 4 + \sqrt[3]{\frac{x-1}{x+1}}.$$

Jednačine $f(x) = 0$ *i* $x = \varphi(x)$ *su ekvivalentne na intervalu* $[4, 5]$. *Za ovako definisano* $\varphi(x)$ *ispunjen je i uslov* $\varphi([4, 5]) \subset [4, 5]$. *Kako je*

$$\varphi'(x) = \frac{2}{3\sqrt[3]{(x-1)^2(x+1)^4}},$$

za $x \in [4, 5]$ *je*

$$|\varphi'(x)| \leq \frac{2}{15\sqrt[3]{45}} < 1,$$

pa su ispunjeni svi uslovi prethodne teoreme. Postupkom sukcesivnih aproksimacija dobijaju se približne vrednosti rešenja posmatrane jednačine date u tabeli.

k	0	1	2	3	4
x_k	4	4.84343	4.86966	4.87034	4.87035
$f(x_k)$	$-8.4\ 10^{-1}$	$-2.6\ 10^{-2}$	$-6.7\ 10^{-4}$	$-1.7\,10^{-5}$	$-4.39\ 10^{-7}$

Za primenu prethodne teoreme potrebno je proveriti da li funkcija preslikava interval koji sadrži nepokretnu tačku u taj isti interval i da li je kontrakcija. U praksi se drugi uslov svodi na određivanje konstante $\gamma < 1$ za koju je $|\varphi'(x)| \leq \gamma$ za sve vrednosti x na posmatranom intervalu. Nijedan od ovih uslova nije lako proverljiv ako je jednačina složena, pa je u sledećoj teoremi pokazano kako se oni mogu oslabiti. Pri tome će se fiksirati početna aproksimacija x_0, što ne otežava sprovođenje postupka.

Teorema 4.20. *Neka je funkcija* φ *kontrakcija sa konstantom* γ, *i neka* $\varphi(a), \varphi(b) \in [a, b]$. *Tada iterativni niz određen iterativnim pravilom*

$$x_{k+1} = \varphi(x_k), \quad k = 0, 1, \ldots$$

sa

$$x_0 = \begin{cases} a, & \text{ako je } \varphi(a) + \varphi(b) < a + b, \\ b, & \text{ako je } \varphi(a) + \varphi(b) \geq a + b. \end{cases}$$

konvergira ka jedinstvenom rešenju u intervalu $[a, b]$ *jednačine* $x = \varphi(x)$.

4.6.2 Ocena greške

Nepokretna tačka funkcije $\varphi(x)$, odnosno rešenje ekvivalentne jednačine $f(x) = 0$, može se dobiti kao granična vrednost konvergentnog niza sukcesivnih aproksimacija. Kako se uvek izračunava samo konačan broj članova iterativnog niza potrebno je oceniti grešku aproksimacije x_k. Sledeća teorema daje ocenu greške nezavisno od iterativne funkcije φ.

Teorema 4.21. ***Lagranžova ocena greške.*** *Neka je $f \in C^1[a,b]$ i $|f'(x)| \geq m > 0$ za $x \in [a,b]$. Ako je $\xi \in [a,b]$ rešenje jednačine $f(x) = 0$ i $x_k \in [a,b]$, onda je*

$$|x_k - \xi| \leq \frac{|f(x_k)|}{m}.$$

Greška postupka sukcesivnih aproksimacija zavisi od funkcije φ na sledeći način.

Teorema 4.22. *Neka je $\varphi \in Lip_\gamma[a,b]$, $\gamma < 1$ i*

$$x_0 \in [a,b], \qquad x_{k+1} = \varphi(x_k) \in [a,b], \quad k = 0, 1, \ldots .$$

Ako je $\alpha \in [a,b]$ rešenje jednačine $x = \varphi(x)$ onda je

$$|x_k - \alpha| \leq \frac{\gamma}{1-\gamma} |x_{k-1} - x_k| \leq \frac{\gamma^k}{1-\gamma} |x_1 - x_0|, \quad k = 1, 2, \ldots .$$

Dokaz. Pošto je φ kontrakcija sa konstatom γ i $\alpha = \varphi(\alpha)$, važi

$$|x_k - \alpha| \leq |\varphi(x_{k-1}) - \varphi(x_k)| + |\varphi(x_k) - \varphi(\alpha)| \leq \gamma |x_{k-1} - x_k| + \gamma |x_k - \alpha| .$$

Odatle, zbog $1 - \gamma > 0$, sledi

$$|x_k - \alpha| \leq \frac{\gamma}{1-\gamma} |x_{k-1} - x_k| .$$

Takođe je

$$|x_k - x_{k-1}| = |\varphi(x_{k-1}) - \varphi(x_{k-2})| \leq \gamma |x_{k-1} - x_{k-2}| \leq \cdots \leq \gamma^{k-1} |x_1 - x_0| ,$$

odnosno

$$|x_k - \alpha| \leq \frac{\gamma^k}{1-\gamma} |x_1 - x_0| .$$

■

Vrednost

$$A_k = \frac{\gamma}{1-\gamma} |x_k - x_{k-1}| \tag{4.2}$$

naziva se **aposteriorna** ocena greške, a

$$B_k = \frac{\gamma^k}{1-\gamma} |x_0 - x_1| \tag{4.3}$$

apriorna ocena greške. Očigledno je $A_k \leq B_k$, odnosno aposteriorna ocena greške je bolja od apriorne, ali se apriorna greška može izračunati na početku iterativnog postupka, na osnovu prve dve aproksimacije.

Primer 4.13. *Neka je data jednačina $f(x) = 0$, gde je*

$$f(x) = x^3 + 4x^2 + 4x - 1.$$

Ova jednačina ima rešenje na intervalu $[0, 0.4]$. *Rešenje se određuje postupkom sukcesivnih aproksimacija* $x_{k+1} = \varphi(x_k)$, *gde je*

$$\varphi(x) = \frac{1}{(x+2)^2}.$$

Kako je

$$|\varphi'(x)| = \left|\frac{-2}{(x+2)^3}\right| < \frac{1}{4} = 0.25, \quad x \in [0, 0.4],$$

može se uzeti $\gamma = 0.25$. *Kako je*

$$f'(x) = 3x^2 + 8x + 4,$$

na posmatranom intervalu je $|f'(x)| \geq 4$, *pa su uslovi za primenu Lagranžove teoreme o oceni greške zadovoljeni za* $m = 4$. *Za*

$$C_k = \frac{|f(x_k)|}{m},$$

aproksimacije i ocene grešaka aproksimacija su date u sledećoj tabeli.

k	x_k	A_k	B_k	C_k
0	0.4	–	–	0.326000
1	0.173611	0.075463	0.075463	0.044940
2	0.211659	0.012683	0.018866	0.001630
3	0.204439	0.002407	0.004716	0.000304
4	0.205780	0.000447	0.001179	0.000057
5	0.205530	0.000083	0.000295	0.000011
6	0.205577	0.000016	0.000074	0.000002

4.7 Red konvergencije

Brzina konvergencije iterativnog postupka meri se **redom konvergencije.**

Definicija 4.4. *Neka je* $\lim_{k\to\infty} x_k = \alpha$. *Ako postoji konstanta* $C \in [0, 1)$ *i ceo broj* $K \geq 0$ *takav da za* $k \geq K$ *važi*

$$|x_{k+1} - \alpha| \leq C\,|x_k - \alpha|$$

kaže se da je niz $x_0, x_1, \ldots$ *linearno konvergentan.*
Ako postoji niz c_k *takav da je*

$$\lim_{k\to\infty} c_k = 0$$

i važi

$$|x_{k+1} - \alpha| \leq c_k\,|x_k - \alpha|\,,$$

onda je niz $x_0, x_1, \ldots$ *superlinearno konvergentan.*
Ako postoje konstante $p > 1$, $C \geq 0$ *i ceo broj* $K \geq 0$ *takav da za* $k \geq K$ *važi*

$$|x_{k+1} - \alpha| \leq C\,|x_k - \alpha|^p$$

kaže se da niz $x_0, x_1, \ldots$ *konvergira ka* α *sa redom bar* p. *Za* $p = 2$ *konvergencija je kvadratna, a za* $p = 3$ *kubna.*

Red konvergencije opšteg iterativnog postupka može se odrediti na osnovu sledeće teoreme.

Teorema 4.23. *Neka je* $\varphi \in C^p(D)$. *Ako je*

$$\begin{aligned} &x_0 \in [a, b]\,, \quad x_{k+1} = \varphi(x_k) \in [a, b]\,, \quad k = 0, 1, \ldots, \quad \lim_{k\to\infty} x_k = \alpha, \\ &\varphi(\alpha) = \alpha, \quad \varphi'(\alpha) = \varphi''(\alpha) = \cdots = \varphi^{(p-1)}(\alpha) = 0, \quad \varphi^{(p)}(\alpha) \neq 0, \end{aligned}$$

onda je niz $x_0, x_1, \ldots$ *konvergentan sa redom* p.

Dokaz. Na osnovu Tejlorovog razvoja funkcije φ je

$$x_{k+1} = \varphi(x_k) = \varphi(\alpha) + \varphi'(\alpha)(x_k - \alpha) + \cdots + \frac{\varphi^{(p)}(\tau)}{p!}(x_k - \alpha)^p,$$

gde je $\tau \in (\min\{x_k, \alpha\}, \max\{x_k, \alpha\})$. Kako je

$$\varphi(\alpha) = \alpha, \qquad \varphi^{(i)}(\alpha) = 0, \quad i = 1, \ldots, p-1,$$

sledi

$$x_{k+1} = \alpha + \frac{\varphi^{(p)}(\tau)}{p!}(x_k - \alpha)^p.$$

Po uslovu teoreme je $\varphi^{(p)}(\alpha) \neq 0$ i $\lim\limits_{k\to\infty} x_k = \alpha$, te je

$$\lim_{k\to\infty} \varphi^{(p)}(\tau) = \varphi^{(p)}(\alpha) \qquad \text{i} \qquad \lim_{k\to\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^p} = \frac{\left|\varphi^{(p)}(\alpha)\right|}{p!},$$

odnosno postoji konstanta $\eta \geq 0$ i ceo broj $K \geq 0$ takav da je

$$|x_{k+1} - \alpha| \leq \eta |x_k - \alpha|^p, \quad k \geq K.$$

■

Da bi se odredio red konvergencije često se posmatra konstanta

$$\eta = \lim_{k\to\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^p}.$$

Ukoliko ovakva konstanta postoji postupak je reda bar p. Ako je $\eta \neq 0$, postupak je reda p i η se naziva **asimptotska konstanta postupka**. Ako je $\eta = 0$, posmatra se granična vrednost sa $p+1$ u eksponentu imenioca. Ako je opet $\eta = 0$, uzima se $p+2$ itd, dok se ne dobije konstanta različita od nule.

Red konvergencije Njutnovog postupka je 2. Postupak sečice ima red konvergencije $\frac{1+\sqrt{5}}{2} \approx 1.618$, a primitivni postupak sečice i postupak polovljenja imaju red konvergencije 1.

4.8 Zadaci

4.1. *Lokalizuj rešenja jednačine* $x + \sqrt{x} - 1 - x^2 = 0$.

4.2. *Lokalizuj korene jednačine*

$$\ln(x+3) - \sin x = 0.$$

4.3. *Odredi intervale koji sadrže realne nule polinoma*

$$p(x) = x^4 - 2x^3 - 7x^2 + 8x + 12.$$

4.4. *Odredi intervale koji sadrže realne nule polinoma*

$$p(x) = x^5 - 4x^3 + 6x - 9.$$

4.5. *Na osnovu teoreme o krugu sa nulama polinoma oceni module nula polinoma*

$$p(x) = x^5 - 4x^3 + 6x - 9.$$

4.6. *Ispitaj da li jednačina* $x = \varphi(x)$ *ima rešenje u intervalu* $[0, 1]$ *ako je:*

a) $\varphi(x) = x^3 + 0.5$;

b) $\varphi(x) = \begin{cases} 1 - x^2, & x \in [0, 0.5], \\ (x-1)^2, & x \in [0.5, 1]. \end{cases}$

4.7. *Neka je* $\varphi(x) \in [a,b]$ *za* $x \in [a,b]$. *Dokaži da je iterativni niz* $x_0, x_1, \ldots,$ *definisan sa*

$$x_{k+1} = \varphi(x_k), \quad k = 0, 1, 2, \ldots$$

i proizvoljnim $x_0 \in [a,b]$ *monoton, ako važi* $0 < \varphi'(x) < 1$, *za svako* $x \in [a,b]$, *a oscilujući ako važi* $-1 < \varphi'(x) < 0$, $x \in [a,b]$.

4.8. *Izračunaj*

$$\sqrt{2 + \sqrt{2 + \sqrt{2 + \cdots}}}.$$

4.9. *Data je jednačina*

$$x + \ln x = 0$$

i sledeći iterativni postupci

$$x_{k+1} = -\ln x_k, \quad k = 0, 1, \ldots,$$
$$x_{k+1} = e^{-x_k}, \quad k = 0, 1, \ldots,$$
$$x_{k+1} = \frac{x_k + e^{-x_k}}{2}, \quad k = 0, 1, \ldots.$$

a) *Lokalizuj nulu* ξ *date jednačine.*

b) *Koji od navedenih iterativnih postupka i za koje* x_0 *konvergira?*

c) *Koji je postupak od navedenih najbolji?*

d) *Jednim od ovih postupaka izračunaj* ξ *sa tri sigurne cifre.*

4.10. *Jednačina* $x^3 - 2x + 2 = 0$ *ima jedan realan koren. Grafički pokaži da Njutnov postupak sa* $x_0 = 0$ *ne konvergira.*

4.11. *Odredi granice intervala* $[a,b]$ *sa tri sigurne cifre tako da za svako* $x_0 \in [a,b]$ *Njutnov postupak konvergira ka rešenju* $\alpha = 0$ *jednačine* $\sin x = 0$.

4.12. *Neka za funkciju* $f \in C^2[a,b]$ *važi*

i) $f(a) < 0, \quad f(b) > 0,$

ii) $f'(x) > 0, \quad f''(x) > 0$, $x \in [a,b]$,

iii) $f'(b) \leq 2f'(a)$.

Dokaži da za svako $x_0 \in [a,b]$, *takvo da je* $f(x_0) > 0$, *Njutnov iterativni postupak konvergira ka rešenju* $\xi \in [a,b]$ *jednačine* $f(x) = 0$ *i da važi*

$$|x_k - \xi| < |x_k - x_{k-1}|, \quad k = 0, 1, \ldots.$$

4.13. *Rešenje jednačine* $xc = 1$, $c > 0$ *je recipročna vrednost broja c. Iterativni niz*

$$x_{k+1} = x_k(2 - cx_k), \quad k = 0, 1, \ldots$$

konvergira ka $1/c$ *za svako* $x_0 \in \left(0, \frac{2}{c}\right)$. *Dokaži.*

4.14. *Reši jednačinu*

$$2x\cos x - \sin x = 0$$

Njutnovim postupkom sa greškom manjom od 10^{-9}.

4.15. *Jednačina*

$$2e^{-x} = \frac{1}{x+2} + \frac{1}{x+1}$$

ima dva rešenja veća od -1. *Izračunaj ta rešenja primitivnim postupkom sečice sa greškom manjom od* 10^{-6}.

4.16. *Odredi rešenje jednačine*

$$x^4 - 5x^3 - 12x^2 + 76x - 79 = 0$$

postupkom polovljenja sa greškom $0.2 \cdot 10^{-3}$.

Glava 5

Numerička integracija

U ovom delu ćemo posmatrati izračunavanje približne vrednosti određenog integrala

$$I(f;a,b)=\int_a^b f(x)dx.$$

Pri tome pretpostavljamo da su granice integracije a i b realni brojevi i da je podintegralna funkcija f realna funkcija jedne realne promenljive.

Na osnovu **Njutn-Lajbnicove** teoreme, ako postoji primitivna funkcija F za funkciju f (odnosno, funkcija F koja je diferencijabilna na intervalu (a,b) i za koju je $F'(x)=f(x)$, $x\in(a,b)$), onda je

$$\int_a^b f(x)dx=F(b)-F(a).$$

Određivanje primitivne funkcije za zadatu podintegralnu funkciju f nije uvek jednostavno. Postoje brojne tablice integrala gde su za određene klase podintegralnih funkcije date primitivne funkcije ili postupci za njihovo izračunavanje. Međutim, za veliki broj podintegralnih funkcija postupak nalaženja primitivnih funkcija je veoma komplikovan ili se primitivna funkcija čak i ne može odrediti u zatvorenoj formi (kao kombinacija konačno mnogo jednostavnijih funkcija). U tom slučaju, kao i u slučaju kada je podintegralna funkcija zadata skupom tačaka$(x_i,f(x_i))$, $i=0,1,\ldots,n$, a njen analitički izraz je nepoznat, traži se približna vrednost za $I(f;a,b)$.

Približnu vrednost posmatranog određenog integrala tražićemo kao linearnu kombinaciju vrednosti podintegralne funkcije f

$$Q_n(f;a,b)=\sum_{i=0}^{n}A_if(x_i),$$

koju nazivamo **kvadraturna formula.** Znači, imamo

$$\int_a^b f(x)dx\approx\sum_{i=0}^{n}A_if(x_i).$$

Za izračunavanje vrednosti $Q_n(f;a,b)$ potrebno je poznavati tačke x_i, koje nazivamo **čvorovi integracije**, i brojeve A_i, koje nazivamo **koeficijenti kvadraturne formule.**

Greška aproksimacije integrala kvadraturnom formulom naziva se **greška kvadraturne formule** i označava se sa $E_n\left(f;a,b\right)$. Definisana je sa

$$E_n\left(f;a,b\right)=I(f;a,b)-Q_n(f;a,b)=\int_a^b f(x)dx-\sum_{i=0}^{n}A_if(x_i).$$

Motivaciju za izračunavanje približne vrednosti integrala $I(f;a,b)$ pomoću $Q_n(f;a,b))$ nalazimo u samoj definiciji određenog integrala, koju ćemo ovde ukratko ponoviti.

Pretpostavimo da je funkcija f ograničena. Podelimo interval $[a,b]$ na n podintervala tačkama.

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b.$$

Označimo tu podelu sa P. Kažemo da je f integrabilna funkcija sa integralom $I(f;a,b)$ ako i samo ako za svako $\varepsilon > 0$ postoji $\delta(\varepsilon) > 0$ takvo da za bilo koju podelu P intervala sa osobinom

$$\max_{1\le i\le n} \{x_i - x_{i-1} : i = 1,2,\ldots,n\} < \delta$$

i svako $c_i \in [x_{i-1}, x_i]$, $i = 1,2,\ldots,n$, važi

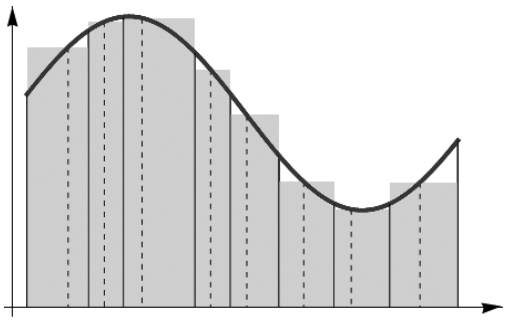
$$\left| \sum_{i=1}^{n} f(c_i)(x_i - x_{i-1}) - \int_a^b f(x)dx \right| < \varepsilon.$$

Broj

$$R(f;a,b) = \sum_{i=1}^{n} f(c_i)(x_i - x_{i-1})$$

nazivamo **Rimanova suma**. Očigledno je da se za svaki prirodan broj n i svaku podelu P intervala $[a,b]$ odgovarajuća Rimanova suma može uzeti kao aproksimacija integrala $I(f;a,b)$. Rimanova suma zavisi i od izbora tačaka c_i, što znači da se za fiksno n i fiksnu podelu P mogu formirati različite aproksimacije posmatranog integrala.

Ako je $f(x) \ge 0$ za $x \in [a,b]$, onda se Rimanova suma geometrijski tumači kao ukupna površina pravougaonika prikazanih na slici 9



Slika 9.

. Zbog ovog tumačenja određenog integrala, formule za izračunavanje njegovih aproksimacija nazivaju se **kvadraturne formule**.

Za uprošćavanje pojedinih izraza koji se dobijaju pri analizi greške integracije koristimo sledeću teoremu.

Teorema 5.1. *Neka je $f \in C[a,b]$ i neka su a_i, $i = 1,2,\ldots,n$, realni brojevi istog znaka. Ako $x_i \in [a,b]$, $i = 1,2,\ldots,n$, onda postoji $x \in [a,b]$ takvo da je*

$$\sum_{i=1}^{n} a_i f(x_i) = f(x) \sum_{i=1}^{n} a_i.$$

Dokaz. Bez umanjenja opštosti možemo pretpostaviti da su svi a_i pozitivni brojevi. Funkcija $f(x)$ je neprekidna u intervalu $[a,b]$, te dostiže svoj maksimum i minimum

$$m = \min\{f(x) \mid x \in [a,b]\}, \qquad M = \max\{f(x) \mid x \in [a,b]\}.$$

Kako je

$$m\,a_i \le a_i f(x_i) \le M a_i, \quad i = 1, 2, \ldots, n,$$

to je

$$m \sum_{i=1}^{n} a_i \le \sum_{i=1}^{n} a_i f(x_i) \le M \sum_{i=1}^{n} a_i,$$

odnosno

$$m \le \frac{\sum_{i=1}^{n} a_i f(x_i)}{\sum_{i=1}^{n} a_i} \le M.$$

Pošto svaka neprekidna funkcija u zatvorenom intervalu dostiže svaku vrednost izmedju svog minimuma i maksimuma (teorema o medjuvrednosti), to postoji tačka $x \in [a, b]$ takva da važi

$$f(x) = \frac{\sum_{i=1}^{n} a_i f(x_i)}{\sum_{i=1}^{n} a_i},$$

a odatle sledi tvrđenje teoreme. ■

5.1 Primitivne kvadraturne formule

5.1.1 Formule levih i desnih pravougaonika

Pod primitivnim kvadraturnim formulama podrazumevamo Rimanove sume dobijene za specijalne izbore tačaka c_i. Za izbor $c_i = x_{i-1}$ dobija se **formula levih pravougaonika**,

$$L_n(f; a, b) = \sum_{i=1}^{n} f(x_{i-1})(x_i - x_{i-1}),$$

za $c_i = x_i$ **formula desnih pravougaonika**,

$$D_n(f; a, b) = \sum_{i=1}^{n} f(x_i)(x_i - x_{i-1}),$$

a za $c_i = \frac{1}{2}(x_{i-1} + x_i)$ dobija se **formula srednjih pravougaonika**.

$$M_n(f; a, b) = \sum_{i=1}^{n} f\left(\frac{x_{i-1} + x_i}{2}\right)(x_i - x_{i-1}).$$

Ako tačke x_i dele interval $[a, b]$ na n podintervala jednake dužine $h = \frac{b-a}{n}$, tj. ako je

$$x_i = a + ih, \quad i = 0, 1, \ldots, n,$$

primitivne kvadraturne formule možemo zapisati u jednostavnijem obliku

$$L_h(f, a, b) = h \sum_{i=0}^{n-1} f(a + ih), \qquad D_h(f, a, b) = h \sum_{i=1}^{n} f(a + ih),$$

$$M_h(f,a,b) = h\sum_{i=1}^{n} f\left(a + \left(i - \frac{1}{2}\right)h\right).$$

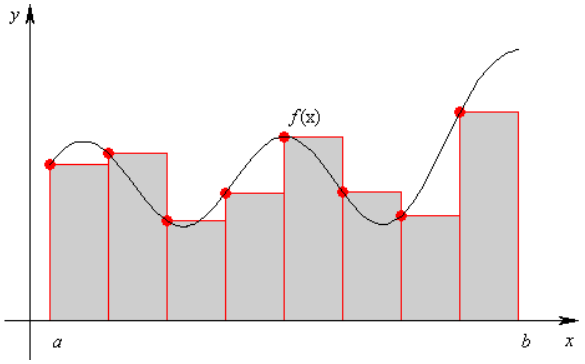
Primitivne kvadraturne formule sa neekvidistantnom raspodelom čvorova integracije obeležavamo indeksom n, a iste formule sa ekvidistantnom raspodelom čvorova integracije, indeksom h.

Nazivi primitivnih kvadraturnih formula potiču od njihove geometrijske interpretacije. Na sledećim slikama posmatramo grafik funkcije

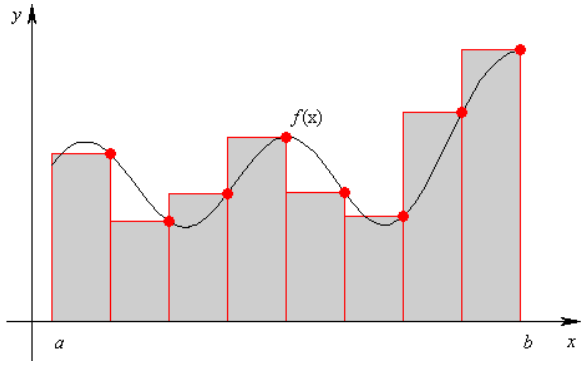
$$f(x) = 2 + \frac{1}{3}(x - 1.5)2 + sin(3x) + sin(x)$$

za $x \in [0.2, 5]$ i podelu tog intervala na $n = 8$ jednakih podintervala ($h = 0.6$). Tačke $(x_i, f(x_i))$ obeležene su posebno u svakom od posmatrana tri slučaja.

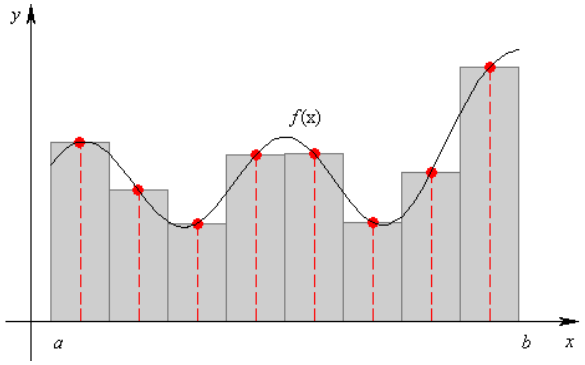
Formula levih pravougaonika kao približnu vrednost određenog integrala daje sumu površina pravougaonika sa slike 10. Formula desnih pravougaonika daje kao približnu vrednost integrala sumu površina pravougaonika sa slike 11, a formula srednjih pravougaonika sumu površina pravougaonika sa slike 12.



Slika 10.



Slika 11.



Slika 0.12

Ocene grešaka primitivnih kvadratnih formula date su u sledećim teoremama.

Teorema 5.2. *Neka su funkcije f i f' neprekidne na intervalu $[a,b]$ Tada postoji $\alpha \in (a,b)$ takvo da je*

$$I(f;a,b) - L_n(f;a,b) = \frac{1}{2} f'(\alpha) \sum_{i=1}^{n} (x_i - x_{i-1})^2 .$$

Dokaz. Prema Tejlorovoj teoremi za $x \in [x_{i-1}, x_i]$ postoji $\sigma_i \in (x_{i-1}, x)$ takvo da važi

$$f(x) = f(x_{i-1}) + f'(\sigma_i)(x - x_{i-1}) .$$

Odavde se dobija

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = f(x_{i-1})(x_i - x_{i-1}) + \int_{x_{i-1}}^{x_i} f'(\sigma_i)(x - x_{i-1})\,dx.$$

Kako je f' neprekidna funkcija i $x - x_{i-1} \geq 0$ za $x \in [x_{i-1}, x_i]$, prema teoremi o srednjoj vrednosti integrala postoji $\alpha_i \in (x_{i-1}, x_i)$ takvo da je

$$\int_{x_{i-1}}^{x_i} f'(\sigma_i)(x - x_{i-1})\,dx = f'(\alpha_i) \int_{x_{i-1}}^{x_i} (x - x_{i-1})\,dx = f'(\alpha_i) \left. \frac{(x - x_{i-1})^2}{2} \right|_{x_{i-1}}^{x_i} = \frac{1}{2} f'(\alpha_i)(x_i - x_{i-1})^2 .$$

Znači,

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = f(x_{i-1})(x_i - x_{i-1}) + \frac{1}{2} f'(\alpha_i)(x_i - x_{i-1})^2$$

i

$$\int_a^b f(x)\,dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,dx = \sum_{i=1}^{n} f(x_{i-1})(x_i - x_{i-1}) + \frac{1}{2} \sum_{i=1}^{n} f'(\alpha_i)(x_i - x_{i-1})^2 .$$

Zbog neprekidnosti funkcije f', na osnovu teoreme 5.1. sledi da postoji $\alpha \in (a,b)$ takvo da je

$$\sum_{i=1}^{n} f'(\alpha_i)(x_i - x_{i-1})^2 = f'(\alpha) \sum_{i=1}^{n} (x_i - x_{i-1})^2 .$$

■

Posledica 5.3. *Neka su funkcije f i f' neprekidne na intervalu $[a,b]$ Tada postoji $\alpha \in (a,b)$ takvo da je*

$$I(f;a,b) - L_h(f;a,b) = \frac{b-a}{2} f'(\alpha)\, h.$$

Dokaz. U ovom slučaju je

$$\sum_{i=1}^{n} (x_i - x_{i-1})^2 = nh^2 = n\frac{b-a}{n} h = (b-a)\, h,$$

jer je $h = \frac{b-a}{n}$ i tvrđenje sledi direktno na osnovu prethodne teoreme. ■

Za ocenu greške aproksimacije određenog integrala formulom levih pravougaonika koristimo tvrđenja iz sledeće teoreme, koja se lako dokazuje na osnovu rezultata prethodne dve teoreme.

Teorema 5.4. *Neka su funkcije f i f' neprekidne na intervalu $[a,b]$ i neka je M_1 konstanta određena tako da važi $M_1 \geq \max\{|f'(x)| : x \in [a,b]\}$. Tada važi*

$$|I(f;a,b) - L_n(f;a,b)| \leq \frac{M_1}{2}\sum_{i=1}^{n}(x_i - x_{i-1})^2$$

i

$$|I(f;a,b) - L_h(f;a,b)| \leq \frac{M_1}{2}(b-a)\,h.$$

Potpuno analogno, polazeći od Tejlorovog razvoja

$$f(x) = f(x_i) + f'(\sigma_i)(x - x_i), \qquad \sigma_i \in (x, x_i),$$

dokazuje se sledeća teorema i njene posledice.

Teorema 5.5. *Neka su funkcije f i f' neprekidne na intervalu $[a,b]$ Tada postoji $\beta \in (a,b)$ takvo da je*

$$I(f;a,b) - D_n(f;a,b) = -\frac{1}{2}f'(\beta)\sum_{i=1}^{n}(x_i - x_{i-1})^2.$$

Posledica 5.6. *Neka su funkcije f i f' neprekidne na intervalu $[a,b]$ Tada postoji $\beta \in (a,b)$ takvo da je*

$$I(f;a,b) - D_h(f;a,b) = -\frac{b-a}{2}f'(\beta)\,h.$$

Posledica 5.7. *Neka su funkcije f i f' neprekidne na intervalu $[a,b]$ i neka je M_1 konstanta određena tako da važi $M_1 \geq \max\{|f'(x)| : x \in [a,b]\}$. Tada važi*

$$|I(f;a,b) - D_n(f;a,b)| \leq \frac{M_1}{2}\sum_{i=1}^{n}(x_i - x_{i-1})^2$$

i

$$|I(f;a,b) - D_h(f;a,b)| \leq \frac{M_1}{2}(b-a)\,h.$$

5.1.2 Formula srednjih pravougaonika

Teorema 5.8. *Neka su funkcije f, f' i f'' neprekidne na intervalu $[a,b]$ Tada postoji $\gamma \in (a,b)$ takvo da je*

$$I(f;a,b) - M_n(f;a,b) = \frac{1}{24}f''(\gamma)\sum_{i=1}^{n}(x_i - x_{i-1})^3.$$

Dokaz. Prema Tejlorovoj teoremi za $x \in [x_{i-1}, x_i]$ postoji $\sigma_i \in (x_{i-1}, x)$ takvo da važi

$$f(x) = f\left(\frac{x_i + x_{i-1}}{2}\right) + f'\left(\frac{x_i + x_{i-1}}{2}\right)\left(x - \frac{x_i + x_{i-1}}{2}\right) + \frac{1}{2}f''(\sigma_i)\left(x - \frac{x_i + x_{i-1}}{2}\right)^2.$$

Na osnovu toga je

$$\begin{aligned}\int_{x_{i-1}}^{x_i} f(x)\,dx &= f\left(\frac{x_i + x_{i-1}}{2}\right)(x_i - x_{i-1}) + f'\left(\frac{x_i + x_{i-1}}{2}\right)\int_{x_{i-1}}^{x_i}\left(x - \frac{x_i + x_{i-1}}{2}\right)dx \\ &\quad + \frac{1}{2}\int_{x_{i-1}}^{x_i} f''(\sigma_i)\left(x - \frac{x_i + x_{i-1}}{2}\right)^2 dx.\end{aligned}$$

Kako je

$$\int_{x_{i-1}}^{x_i}\left(x-\frac{x_i+x_{i-1}}{2}\right)dx=\frac{1}{2}\left(x-\frac{x_i+x_{i-1}}{2}\right)^2\Bigg|_{x_{i-1}}^{x_i}=\frac{1}{2}\left(\left(\frac{x_i-x_{i-1}}{2}\right)^2-\left(\frac{x_{i-1}-x_i}{2}\right)^2\right)=0$$

i prema teoremi o srednjoj vrednosti integrala za neko $\gamma_i\in(x_{i-1},x_i)$

$$\begin{aligned}\int_{x_{i-1}}^{x_i}f''(\sigma_i)\left(x-\frac{x_i+x_{i-1}}{2}\right)^2dx &= f''(\gamma_i)\int_{x_{i-1}}^{x_i}\left(x-\frac{x_i+x_{i-1}}{2}\right)^2dx\\ &= \frac{1}{3}f''(\gamma_i)\left(x-\frac{x_i+x_{i-1}}{2}\right)^3\Bigg|_{x_{i-1}}^{x_i}\\ &= \frac{1}{3}f''(\gamma_i)\left(\left(\frac{x_i-x_{i-1}}{2}\right)^3-\left(\frac{x_{i-1}-x_i}{2}\right)^3\right)\\ &= \frac{1}{12}f''(\gamma_i)(x_i-x_{i-1})^3,\end{aligned}$$

dobijamo

$$\int_{x_{i-1}}^{x_i}f(x)\,dx=f\left(\frac{x_i+x_{i-1}}{2}\right)(x_i-x_{i-1})+\frac{1}{24}f''(\gamma_i)(x_i-x_{i-1})^3.$$

Na osnovu ovog rezultata dobijamo

$$\int_a^b f(x)\,dx=\sum_{i=1}^n\int_{x_{i-1}}^{x_i}f(x)\,dx=\sum_{i=1}^n f\left(\frac{x_i+x_{i-1}}{2}\right)(x_i-x_{i-1})+\frac{1}{24}\sum_{i=1}^n f''(\gamma_i)(x_i-x_{i-1})^3.$$

Zbog neprekidnosti funkcije f', na osnovu teoreme 5.1. sledi da postoji $\gamma\in(a,b)$ takvo da je

$$\int_a^b f(x)\,dx=\sum_{i=1}^n f\left(\frac{x_i+x_{i-1}}{2}\right)(x_i-x_{i-1})+\frac{f''(\gamma)}{24}\sum_{i=1}^n(x_i-x_{i-1})^3.$$

■

Posledica 5.9. *Neka su funkcije f, f' i f'' neprekidne na intervalu $[a,b]$ Tada postoji $\gamma\in(a,b)$ takvo da je*

$$I(f;a,b)-M_h(f;a,b)=\frac{b-a}{24}f''(\gamma)\,h^2.$$

Dokaz. Dovoljno je primetiti da je u ekvidistantnom sliučaju

$$x_i=a+ih,\quad i=0,1,\ldots,n,\qquad h=\frac{b-a}{n},$$

i

$$\sum_{i=1}^n(x_i-x_{i-1})^3=nh^3=n\frac{b-a}{n}h^2=(b-a)\,h^2.$$

■

Posledica 5.10. *Neka su funkcije f, f' i f'' neprekidne na intervalu $[a,b]$ i neka je M_2 konstanta određena tako da važi $M_2 \geq \max\{|f''(x)| : x \in [a,b]\}$. Tada važi*

$$|I(f;a,b) - M_n(f;a,b)| \leq \frac{M_2}{24}\sum_{i=1}^{n}(x_i - x_{i-1})^3$$

i

$$|I(f;a,b) - M_h(f;a,b)| \leq \frac{b-a}{24}M_2h^2.$$

Primer 5.1. *Približne vrednosti integrala $I(f;0,1)$, gde je*

$$f(x) = \sin^3(16x) + 16x + 1,$$

izračunate pomoću $L_h(f;0,1)$, $D_h(f;0,1)$ i $M_h(f;0,1)$ sa $h = \frac{1}{8}$ su

$$L_h(f;0,1) = 8.24019, \quad D_h(f;0,1) = 10.2372, \quad M_h(f;0,1) = 8.92745.$$

Tačna vrednost posmatranog integrala je $9.08322\ldots$.

5.2 Interpolacione kvadraturne formule

5.2.1 Njutn-Kotesove formule

U ovom paragrafu se posmatraju kvadraturne formule oblika sa pretpostavkom da za čvorove integracije x_i važi

$$a \leq x_0 < x_1 < \cdots < x_n \leq b.$$

Pri izboru čvorova integracije x_i i koeficijenata A_i kvadraturne formule $Q_n(f;a,b)$ potrebno je postići da $Q_n(f;a,b)$ bude dobra aproksimacija za $I(f;a,b)$ za što širu klasu podintegralnih funkcija f. Kao i kod aproksimacije funkcija i za određivanje greške numeričke integracije

$$E_n(f;a,b) = I(f;a,b) - Q_n(f;a,b)$$

postoje razni postupci.

Definicija 5.1. *Kvadraturna formula $Q_n(f;a,b)$ je reda tačnosti k ako je*

$$E_n(x^i;a,b) = 0, \quad i = 0,1,\ldots,k, \qquad E_n(x^{k+1};a,b) \neq 0.$$

Prema ovoj definiciji formule levih i desnih pravougaonika su reda tačnosti 0, a formula srednjih pravougaonika je reda tačnosti 1.

Definicija 5.2. *Kvadraturna formula $Q_n(f;a,b)$ je interpolaciona kvadraturna formula $n-$tog reda ako za sve $f \in C[a,b]$ važi*

$$\sum_{i=0}^{n} A_i f(x_i) = \int_a^b P_n(f,x)dx,$$

gde je $P_n(f,x)$ interpolacioni polinom $n-$tog stepena funkcije f određen čvorovima $x_0, x_1, \ldots, x_n$.

Za interpolacione kvadraturne formule $n-$tog reda važi $E_n(q;a,b) = 0$ za svaki polinom q stepena manjeg ili jednakog sa n, jer je $P_n(q,x) = q(x)$.

Teorema 5.11. *Koeficijenti interpolacione kvadraturne formule $n-$tog reda dati su sa*

$$A_i = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{1}{x_i - x_j} \int_a^b \prod_{\substack{j=0 \\ j\neq i}}^{n} (x - x_j)\, dx, \qquad i = 0, 1, \ldots, n.$$

Dokaz. Neka je $P_n(f, x)$ interpolacioni polinom u Lagranžovom obliku određen interpolacionim čvorovima $x_0, x_1, \ldots, x_n$. Integracijom tog polinoma dobija se

$$\int_a^b P_n(f,x)dx = \int_a^b \left(\sum_{i=0}^{n} f(x_i) \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j} \right) dx = \sum_{i=0}^{n} \left(\prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{1}{x_i - x_j} \int_a^b \prod_{\substack{j=0 \\ j\neq i}}^{n} (x - x_j) \right) f(x_i)dx.$$

Iz prethodne jednakosti sledi tvrđenje teoreme. ■

Vezu između formula $n-$tog reda tačnosti i interpolacionih formula $n-$tog reda daje sledeća teorema.

Teorema 5.12. *Interpolaciona kvadraturna formula reda n ima red tačnosti veći ili jednak sa n.*

Dokaz. Neka je $P_n(x^k, x)$ interpolacioni polinom funkcije x^k određen čvorovima $x_0, x_1, \ldots, x_n$. Tada je

$$x^k = P_n(x^k, x),$$

i

$$E_n(x^k; a, b) = E_n(P_n(x^k, x); a, b) = 0, \quad k = 0, 1, \ldots, n,$$

što znači da je kvadraturna formula reda tačnosti bar n. Može se desiti da je i

$$E_n(x^k; a, b) = 0 \qquad \text{za} \qquad k = n + 1, n + 2, ...$$

i tada interpolaciona kvadraturna formula $n-$tog reda može biti reda tačnosti većeg od n. ■

Definicija 5.3. *Interpolacione kvadraturne formule $n-$tog reda sa ekvidistantnim čvorovima*

$$x_i = a + ih, \quad i = 0, 1, \ldots, n, \qquad h = \frac{b-a}{n}, \qquad n \in \mathbb{N},$$

*nazivaju se **Njutn-Kotesove** kvadraturne formule zatvorenog tipa.*

Kod Njutn-Kotesovih kvadraturnih formula zatvorenog tipa svi čvorovi integracije pripadaju intervalu $[a, b]$, a $x_0 = a$ i $x_n = b$ su prvi i poslednji čvor integracije, što opravdava naziv "zatvorenog tipa".

Teorema 5.13. *Njutn-Kotesova kvadraturna formula reda n je*

$$N_n(f; a, b) = h \sum_{i=0}^{n} A_{n,i} f(x_i),$$

sa koeficijentima

$$A_{n,i} = \frac{(-1)^{n-i}}{i!(n-i)!} \int_0^n \prod_{\substack{j=0 \\ j\neq i}}^{n} (t - j)dt, \quad i = 0, 1, \ldots, n.$$

Dokaz. Pošto su čvorovi integracije ekvidistantni, smenom $x = x_0 + th$ dobija se interpolacioni polinom za funkciju f u obliku

$$L_n(x_0 + th) = \sum_{i=0}^{n} f(x_i) \frac{(-1)^{n-i}}{i!\,(n-i)!} \prod_{\substack{j=0 \\ j\neq i}}^{n} (t-j).$$

Sada je

$$\int_a^b L_n(x)\,dx = h\int_0^n L_n(x_0+th)dt = h\sum_{i=0}^{n} f(x_i)\frac{(-1)^{n-i}}{i!\,(n-i)!}\int_0^n \prod_{\substack{j=0\\ j\neq i}}^{n}(t-j)\,dt,$$

odakle neposredno sledi tvrđenje teoreme. ■

Posledica 5.14. *Za koeficijente iz prethodne teoreme važi za svako* $i = 0, 1, \ldots, n$

$$A_{n,i} = A_{n,n-i}.$$

5.2.2 Greška Njutn-Kotesovih formula

Za interpolacione kvadraturne formule $Q_n(f; a, b)$ greška integracije je

$$E_n(f; a, b) = \int_a^b f(x)dx - Q_n(f; a, b) = \int_a^b R_n(f, x)dx,$$

gde je $R_n(f, x)$ greška interpolacije

$$R_n(f, x) = f(x) - P_n(f, x) = \frac{f^{(n+1)}(\tau)}{(n+1)!}\omega_n(x), \qquad \omega_n(x) = \prod_{j=0}^{n}(x - x_j).$$

U opštem slučaju, direktna integracija greške $R_n(f, x)$ nije moguća, jer τ zavisi od x na nepoznat način. Takođe, u opštem slučaju, teorema o srednjoj vrednosti integrala može se primeniti tek kada se ispita da li funkcija $\omega_n(x)$ menja znak u posmatranom intervalu.

Teorema 5.15. *Postoji* $\tau \in [a, b]$ *takvo da za grešku Njutn-Kotesove kvadraturne formule* $n-$*tog reda i zatvorenog tipa važi*

$$E_n(f; a, b) = \begin{cases} h^{n+3} f^{(n+2)}(\tau) \displaystyle\int_0^n \binom{s}{n+2} ds, \text{ za } f \in C^{n+2}[a, b] \text{ i parno } n, \\ \\ h^{n+2} f^{(n+1)}(\tau) \displaystyle\int_0^n \binom{s}{n+1} ds, \text{ za } f \in C^{n+1}[a, b] \text{ i neparno } n. \end{cases}$$

Ocena greške $E_n(f; a, b)$, na osnovu teoreme, je

$$|E_n(f; a, b)| \leq \begin{cases} h^{n+3} M_{n+2} \displaystyle\int_0^n \left|\binom{s}{n+2}\right| ds, \text{ za } f \in C^{n+2}[a, b] \text{ i parno } n, \\ \\ h^{n+2} M_{n+1} \displaystyle\int_0^n \left|\binom{s}{n+1}\right| ds, \text{ za } f \in C^{n+1}[a, b] \text{ i neparno } n. \end{cases}$$

5.3 Trapezna kvadraturna formula

5.3.1 Prosta trapezna formula

Posmatra se Njutn-Kotesova kvadraturna formula reda 1. Tada je $x_0 = a,\ x_1 = b,\ h = b - a$. Odgovarajuća Njutn-Kotesova formula

$$N_1(f; a, b) = A_{1,0} f(a) + A_{1,1} f(b)$$

dobija se integracijom interpolacionog polinoma prvog reda određenog tačkama $(a, f(a))$ i $(b, f(b))$. Znači,

$$A_{1,0} f(a) + A_{1,1} f(b) = \int P_1(f, x)\, dx,$$

gde je

$$P_1(f, x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a).$$

Lako se izračunava

$$\int_a^b P_1(f,x)dx = \int_a^b f(a)dx + \int_a^b \frac{f(b)-f(a)}{b-a}(x-a)dx = f(a)\int_a^b dx + \frac{f(b)-f(a)}{b-a}\int_a^b (x-a)dx$$

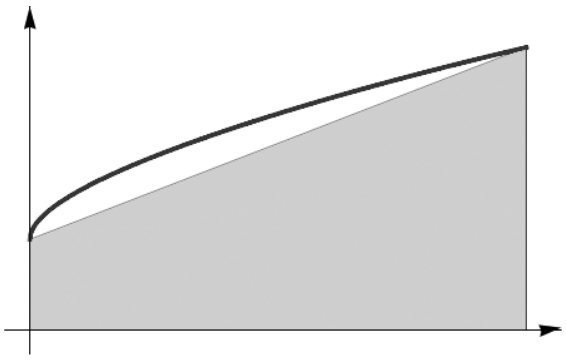
odnosno

$$\int_a^b P_1(f,x)dx = f(a)(b-a) + \frac{f(b)-f(a)}{b-a}\frac{(b-a)^2}{2} = \frac{b-a}{2}(f(a)+f(b)).$$

Na ovaj način smo dobili

$$N_1(f; a, b) = \frac{b-a}{2}(f(a) + f(b)).$$

Ova formula se naziva **trapezna formula**, što je motivisano geometrijskom interpretacijom vrednosti $N_1(f; a, b)$. Neka je $f(x) \geq 0$, $x \in [a, b]$. Tada je $N_1(f; a, b)$ površina trapeza na slici 13.



Slika 13.

Odgovarajuća greška je

$$E_1(f; a, b) = \int_a^b R_1(f, x)\, dx = -\frac{h^3}{12} f''(\tau), \quad \tau \in (a, b).$$

gde je

$$R_1(f, x) = f(x) - p_1(x) = \frac{f''(\alpha)}{2}(x - a)(x - b).$$

Na intervalu $[a,b]$ je $(x-a)(x-b) \leq 0$. Ako je funkcija f'' neprekidna na intervalu $[a,b]$ onda, prema teoremi o srednjoj vrednosti integrala važi

$$\begin{aligned}\int_a^b R_1(f,x)\,dx &= \int_a^b \frac{f''(\alpha)}{2}(x-a)(x-b)\,dx = \frac{1}{2}f''(\tau)\int_a^b (x-a)(x-b)\,dx \\ &= \frac{1}{2}f''(\tau)\left(\frac{x^3}{3} - \frac{a+b}{2}x^2 + abx\right)\Big|_a^b = \frac{1}{2}f''(\tau)\left(\frac{ab^2}{2} - \frac{b^3}{6} - \left(-\frac{a^3}{6} + \frac{a^2b}{2}\right)\right) \\ &= -\frac{1}{12}f''(\tau)\left(b^3 + 3a^2b - 3ab^2 - a^3\right) = -\frac{1}{12}f''(\tau)(b-a)^3.\end{aligned}$$

Na ovaj način smo dokazali sledeću teoremu.

Teorema 5.16. *Ako je funkcija f dva puta neprekidno diferencijabilna na intervalu $[a,b]$, onda postoji $\tau \in (a,b)$ takvo da je*

$$\int_a^b f(x)\,dx - N_1(f;a,b) = -\frac{1}{12}f''(\tau)(b-a)^3.$$

Posledica 5.17. *Neka je funkcija f dva puta neprekidno diferencijabilna na intervalu $[a,b]$ i neka je M_2 konstanta određena tako da važi $M_2 \geq \max\{|f''(x)| : x \in [a,b]\}$. Tada je*

$$\left|\int_a^b f(x)\,dx - N_1(f;a,b)\right| \leq \frac{M_2}{12}(b-a)^3.$$

5.3.2 Složena trapezna formula

Interpolacione kvadraturne formule za aproksimaciju integrala $I(f;a,b)$ dobijaju se integracijom interpolacionog polinoma za funkciju f sa granicama integracije a i b. Ako je interval $[a,b]$ velik ne može se u opštem slučaju dobiti dobra tačnost interpolacione kvadraturne formule nastale integracijom interpolacionog polinoma nižeg stepena. S druge strane, korišćenje interpolacionih polinoma višeg stepena ne mora dovesti do interpolacionih formula zadovoljavajuće tačnosti. Takođe, formiranje formula velikog reda tačnosti dovodi do komplikovanog izračunavanja koeficijenata integracije. Oba ova problema, integracija interpolacionih polinoma višeg stepena i izračunavanje koeficijenata formula visokog reda tačnosti, mogu se izbeći formiranjem složenih kvadraturnih formula. Ove formule se dobijaju podelom intervala integracije na više podintervala i primenom neke od kvadraturnih formula nižeg reda na svakom od tih podintervala.

Posmatra se integral $I(f;a,b)$ i podela intervala $[a,b]$ na m podintervala tačkama x_i za koje važi

$$a = x_0 < x_1 < \cdots < x_m = b.$$

Trapezna formula primenjena na integrale sa granicama x_{i-1} i x_i daje

$$\begin{aligned}\int_a^b f(x)dx &= \sum_{i=1}^m \int_{x_{i-1}}^{x_i} f(x)dx = \sum_{i=1}^m \left(N_1(f;x_{i-1},x_i) + E_1(f;x_{i-1},x_i)\right) \\ &= \sum_{i=1}^m \frac{x_i - x_{i-1}}{2}\left(f(x_{i-1}) + f(x_i)\right) - \frac{1}{12}\sum_{i=1}^m f''(\tau_i)(x_i - x_{i-1})^3,\end{aligned}$$

sa $\tau_i \in (x_{i-1}, x_i)$, $i = 1, 2, \ldots, n$. Na osnovu teoreme 5.1. dobijamo da postoji $\theta \in (a, b)$ takvo da je

$$\int_a^b f(x)dx = T_m(f; a, b) - \frac{f''(\tau)}{12} \sum_{i=1}^{m} (x_i - x_{i-1})^3,$$

gde je

$$T_m(f; a, b) = \frac{x_1 - x_0}{2} f(x_0) + \sum_{i=1}^{m-1} \frac{x_{i+1} - x_{i-1}}{2} f(x_i) + \frac{x_m - x_{m-1}}{2} f(x_m).$$

Formula $T_m(f; a, b)$ je neekvidistantna složena trapezna formula i pogodna je za integraciju funkcija zadatih tabelarno. U slučaju kada su čvorovi ekvidistantni, tj kada je $h = \dfrac{b-a}{m}$ i

$$x_i = a + ih, \quad i = 0, 1, \ldots, m,$$

dobija se **složena trapezna formula**

$$T_m(f; a, b) = \frac{h}{2} \left(f(a) + 2 \sum_{i=1}^{m-1} f(a + ih) + f(b) \right)$$

i za neko $\tau \in (a, b)$

$$I(f; a, b) = T_m(f; a, b) - \frac{b-a}{12} h^2 f''(\tau),$$

jer je

$$\sum_{i=1}^{m} (x_i - x_{i-1})^3 = mh^3 = m \frac{b-a}{m} h^2 = (b-a)\, h^2.$$

5.4 Simpsonova kvadraturna formula

5.4.1 Prosta Simsonova formula

Posmatra se Njutn-Kotesova kvadraturna formula reda 2. Tada je

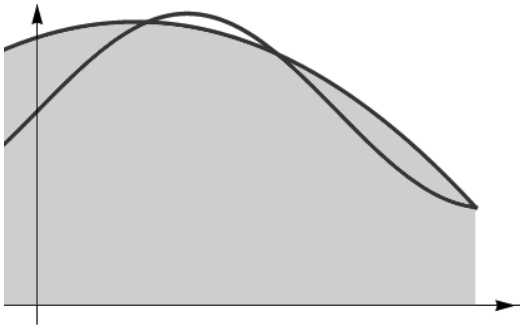
$$h = \frac{b-a}{2}, \qquad x_0 = a, \qquad x_1 = \frac{a+b}{2}, \qquad x_2 = b.$$

Odgovarajuća Njutn-Kotesova formula

$$N_2(f; a, b) = A_{2,0} f(a) + A_{2,1} f(\left(\frac{a+b}{2} \right) + A_{2,2} f\,(b)$$

dobija se integracijom interpolacionog polinoma drugog stepena određenog tačkama

$$(a, f\,(a))\,, \qquad \left(\frac{a+b}{2}, f\left(\frac{a+b}{2} \right) \right), \qquad (b, f\,(b))\,.$$



Slika 14.

Znači,

$$A_{2,0}f(a) + A_{2,1}f(\left(\frac{a+b}{2}\right) + A_{2,2}f\left(b\right) = \int P_2\left(f,x\right)dx,$$

gde je

$$P_2\left(f,x\right) = \frac{1}{(a-b)^2}\left(f(a)(b-x)(a+b-2x) + 4f\left(\frac{a+b}{2}\right)(a-x)(x-b) + f(b)(a-x)(a+b-2x)\right).$$

Koeficijente $A_{2,0}, A_{2,1}$ i $A_{2,2}$ ćemo lako odrediti ako izračunamo sledeće integrale

$$\int_a^b (b-x)(a+b-2x)dx,$$

$$\int_a^b (a-x)(x-b)dx$$

i

$$\int_a^b (a-x)(a+b-2x)dx.$$

Primenom Njutn-Lajbnicove formule dobijamo

$$\int_a^b (b-x)(a+b-2x)dx = \frac{2x^3}{3} - \frac{a+3b}{2}x^2 + b\left(a+b\right)x\Big|_a^b = \frac{ab^2}{2} + \frac{b^3}{6} - \left(\frac{a^3}{6} - \frac{a^2b}{2} + ab^2\right) = \frac{1}{6}\left(b-a\right)^3$$

$$\int_a^b (a-x)(b-x)dx == \left(-\frac{x^3}{3} + \frac{a+b}{2}x^2 - abx\right)\Big|_a^b = -\frac{ab^2}{2} + \frac{b^3}{6} - \left(\frac{a^3}{6} - \frac{a^2b}{2}\right) = \frac{1}{6}\left(b-a\right)^3$$

$$\int_a^b (a-x)(a+b-2x)dx = \frac{2x^3}{3} - \frac{b+3a}{2}x^2 + a\left(a+b\right)x\Big|_a^b = a^2b - \frac{ab^2}{2} + \frac{b^3}{6} - \left(\frac{a^3}{6} + \frac{a^2b}{2}\right) = \frac{1}{6}\left(b-a\right)^3.$$

Sada imamo

$$A_{2,0} = \frac{1}{(a-b)^2}\int_a^b (b-x)(a+b-2x)dx = \frac{b-a}{6},$$

$$A_{2,1} = \frac{4}{(a-b)^2}\int_a^b (a-x)(b-x)dx = \frac{4}{6}\left(b-a\right),$$

$$A_{2,2} = \frac{1}{(a-b)^2}\int_a^b (b-x)(a+b-2x)dx = \frac{b-a}{6}$$

i

$$\int_a^b P_2(f,x)dx = \frac{b-a}{6}\left(f\left(a\right) + 4f\left(\frac{a+b}{2}\right) + f\left(b\right)\right).$$

Znači, Njutn-Kotesova kvadraturna formula reda 2, koju nazivamo **Simpsonova formula,** je

$$N_2(f;a,b) = \frac{b-a}{6}\left(f\left(a\right) + 4f\left(\frac{a+b}{2}\right) + f\left(b\right)\right).$$

Odgovarajuća greška je

$$E_2(f;a,b)=\int_a^b f(x)\,dx-\frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right).$$

Teorema 5.18. *Neka je funkcija f četiri puta neprekidno diferencijabilna. Tada postoji $\sigma\in(a,b)$ takvo da je*

$$E_2(f;a,b)=-\frac{1}{90}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\sigma).$$

Dokaz. Neka je $s=\frac{a+b}{2}$ i $h=\frac{b-a}{2}$. Tada je $a=s-h$ i $b=s+h$, a greška $E_2(f;a,b)$ se može posmatrati kao funkcija po h (radi jednostavnosti označimo je sa φ). Imamo

$$\varphi(h)=\int_{s-h}^{s+h} f(x)\,dx-\frac{h}{3}\left(f(s-h)+4f(s)+f(s+h)\right).$$

Uzastopnim diferenciranjem po h dobijamo

$$\begin{aligned}\varphi'(h) &= f(s+h)-f(s-h)-\frac{1}{3}\left(f(s-h)+4f(s)+f(s+h)\right)-\frac{h}{3}\left(-f'(s-h)+f'(s+h)\right)\\ &= \frac{2}{3}\left(f(s+h)-f(s-h)\right)-\frac{4}{3}f(s)-\frac{h}{3}\left(f'(s+h)-f'(s-h)\right),\end{aligned}$$

$$\varphi''(h)=\frac{1}{3}\left(f(s+h)-f(s-h)\right)-\frac{h}{3}\left(f''(s+h)+f''(s-h)\right),$$

$$\varphi'''(h)=-\frac{h}{3}\left(f'''(s+h)-f'''(s-h)\right).$$

Očigledno na osnovu Lagranžove teoreme sledi

$$\varphi'''(h)=-\frac{h}{3}\left(f'''(s+h)-f'''(s-h)\right)=-\frac{2}{3}h^2f^{(4)}(\alpha),$$

gde α zavisi od h i pripada intervalu $(s-h,s+h)$. Kako je $\varphi''(0)=0$, sledi

$$\varphi''(h)=\int_0^h\varphi'''(t)\,dt=-\frac{2}{3}\int_0^h t^2f^{(4)}(\sigma)\,dt=-\frac{2}{9}h^3f^{(4)}(\beta)$$

gde β zavisi od h i intervalu $(s-h,s+h)$. Na sličan način zbog $\varphi'(0)=0$ i $\varphi(0)=0$ imamo

$$\varphi'(h)=\int_0^h\varphi''(t)\,dt=-\frac{2}{9}\int_0^h t^3f^{(4)}(\beta)\,dt=-\frac{1}{18}h^4f^{(4)}(\gamma),$$

gde γ zavisi od h i intervalu $(s-h,s+h)$ i

$$\varphi(h)=\int_0^h\varphi'(t)\,dt=-\frac{1}{18}\int_0^h t^4f^{(4)}(\gamma)\,dt=-\frac{1}{90}h^5f^{(4)}(\sigma),$$

što je i trebalo dokazati. ■

5.4.2 Složena Simpsonova formula

Posmatrajmo integral $I(f;a,b)$ i podelimo interval $[a,b]$ na $2m$ podintervala tačkama x_i za koje važi

$$a = x_0 < x_1 < \cdots < x_m = b.$$

Složena Simpsonova kvadraturna formula se dobija na sledeći način. Interval $[a,b]$ se deli na $2m$ podintervala tačkama x_i za koje važi

$$a = x_0 < x_1 < \cdots < x_{2m} = b \quad i \quad x_{2i-1} = \frac{x_{2i-2} + x_{2i}}{2}, \quad i = 1, 2, \ldots, m.$$

Sada je

$$\int_a^b f(x)dx = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + \cdots + \int_{x_{2m-4}}^{x_{2m-2}} f(x)dx + \int_{x_{2m-2}}^{x_{2m}} f(x)dx.$$

Primenjujući Simpsonovu formulu na svaki od integrala sa desne strane, dobijamo

$$\int_{x_{2i-2}}^{x_{2i}} f(x)dx \approx \frac{x_{2i} - x_{2i-2}}{6} \left(f\left(x_{2i-2}\right) + 4f\left(x_{2i+1}\right) + f\left(x_{2i}\right)\right)$$

i

$$\int_a^b f(x)dx \approx \frac{1}{6} \sum_{i=1}^{m} (x_{2i} - x_{2i-2}) \left(f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})\right),$$

odnosno

$$I(f;a,b) \approx S_m(f;a,b)$$

gde je

$$S_m(f;a,b) = \frac{1}{6} \sum_{i=1}^{m} (x_{2i} - x_{2i-2}) \left(f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})\right).$$

Formula $S_m(f;a,b)$ se naziva **složena neekvidistantna Simpsonova formula**. U ekvidistantnom slučaju, tj. kada je

$$x_i = a + iH, \quad i = 0, 1, \ldots, 2m, \qquad H = \frac{b-a}{2m},$$

dobija se

$$S_m(f;a,b) = \frac{H}{3} \left(f(a) + 2 \sum_{i=1}^{m-1} f(a + i\frac{b-a}{m}) + 4 \sum_{i=1}^{m} f(a + (2i-1)\frac{b-a}{2m}) + f(b) \right).$$

Formula $S_m(f;a,b)$ je **ekvidistantna složena Simpsonova formula**, ili kraće **složena Simpsonova formula**.

Teorema 5.19. *Neka je funkcija f četiri puta neprekidno diferencijabilna. Tada postoji $\theta \in (a,b)$ takvo da je*

$$I(f;a,b) - S_m(f;a,b) = -\frac{f^{(4)}(\theta)}{2880} \sum_{i=1}^{m} \left(x_{2i} - x_{2i-2}\right)^5,$$

a u ekvidistantnom slučaju

$$I(f;a,b) = S_m(f;a,b) - \frac{b-a}{180} H f^{(4)}(\theta).$$

Dokaz. Grešku složene Simpsonove formule određujemo kao zbir grešaka

$$E_2(f; x_{2i-1}, x_{2i}) = -\frac{1}{90}\left(\frac{x_{2i-1} - x_{2i}}{2}\right)^5 f^{(4)}(\sigma_i),$$

tj.

$$I(f; a, b) - S_m(f; a, b) = -\frac{1}{90}\sum_{i=1}^{m}\left(\frac{x_{2i} - x_{2i-2}}{2}\right)^5 f^{(4)}(\sigma_i),$$

odnosno

$$I(f; a, b) - S_m(f; a, b) = -\frac{f^{(4)}(\theta)}{2880}\sum_{i=1}^{m}(x_{2i} - x_{2i-2})^5.$$

U ekvidistantnom slučaju je

$$I(f; a, b) - S_m(f; a, b) = -\frac{f^{(4)}(\theta)}{2880} m (2H)^5 = -\frac{b-a}{180} f^{(4)}(\theta) H^4,$$

jer je $mH = \dfrac{b-a}{2}$.

$$I(f; a, b) = S_m(f; a, b) - \frac{b-a}{180} H^4 f^{IV}(\theta), \quad \theta \in (a, b).$$

■

Primer 5.2. *Složenom trapeznom formulom sa* $m = 8$ *izračunava se približna vrednost integrala* $I(f; 0, 1)$, *gde je*

$$f(x) = \frac{1}{1 + x^2}.$$

Kako je

$$f''(x) = \frac{6x^2 - 2}{(1 + x^2)^3}$$

i $|f''(x)| \leq M_2 = 4$, $x \in [0, 1]$, *to je*

$$|I(f; 0, 1) - T_8(f; 0, 1)| \leq \frac{1}{12} \cdot \frac{1}{64} \cdot 4 < 0.00521.$$

Prema tome, mogu se koristiti približne vrednosti za $f(x_k)$ *izračunate sa* 3 *sigurne cifara, a da računarske greške ne budu veće od greške ostatka kvadraturne formula. Tako se dobija*

$$T(f; 0, 1) \approx T_8^* = \frac{1}{16} \cdot 12.556 = 0.78475.$$

Približne vrednosti za $f(x_k)$ *su date sa tri sigurne cifre u užem smislu, pa je*

$$|T_8(f : 0, 1) - T_8^*| \leq h \cdot 8 \cdot 0.5 \cdot 10^{-3} = 0.5 \cdot 10^{-3}.$$

Sada je

$$|I - T_8^*| < 0.00521 + 0.5 \cdot 10^{-3} = 0.00571.$$

Znajući da je $I(f; 0, 1) = \arctan 1$, *lako se dobija ocena*

$$|\arctan 1 - T_8^*| \leq 0.000650.$$

5.5 Zadaci

5.1. *Dužina d elipse*

$$\left\{(x,y) : \frac{x^2}{a^2} + y^2 = 1\right\}, \quad a > 0,$$

izračunava se prema

$$d = 4\int_0^{\pi/2} \sqrt{1 - (1-a^2)\sin^2 t}\, dt.$$

Ovaj integral se za $a \neq 1$ *ne može izračunati preko Njutn-Lajbnicove formule. Izračunaj njegovu približnu vrednost za* $a = 2$ *pomoću Simpsonove kvadraturne formule sa* $n = 16$.

5.2. *Iz*

$$\int_0^1 \frac{4}{1+x^2} dx = \pi$$

pomoću formula za numeričku integraciju izračunaj broj π *sa tri sigurne cifre.*

5.3. *Izračunaj približnu vrednost integrala*

$$\int_{\frac{\pi}{6}}^{\frac{\pi}{2}} (1 + \ln(\sin x))\, dx$$

formulom levih pravougaonika sa $n = 8$ *i proceniti grešku.*

5.4. *Izračunaj približnu vrednost integrala*

$$\int_0^1 \frac{1}{x^2+1} dx$$

a) pomoću formule desnih pravougaonika za $n = 10$;

b) pomoću formule srednjih pravougaonika za $n = 10$;

c) pomoću Simpsonove formule za $n = 10$. U sva tri slučaja proceni grešku.

5.5. *Koristeći približne vrednosti integrala*

$$I = \int_{0.5}^{0.5} \frac{1}{1-x^2} dx$$

M_8 *i* T_8 *dobijene pomoću formule srednjih pravougaonika i trapezne formule, izračunaj približnu vrednosti* $S = 0.5\,(M_8 + T_8)$ *i oceniti*

$$R = |I - S|\,.$$

5.6. *Koristeći Simpsonovu formulu sa* $n = 4$ *izračunaj približnu vrednost integrala*

$$\int_0^{0.8} f(x)\, dx,$$

pri čemu se vrednosti funkcije izračunavaju iz jednačine

$$x = f(x)\, e^{f(x)}.$$

5.7. *Izračunaj približne vrednosti sledećih integrala sa tačnošću* ε.

a) $\int_0^2 \frac{\sin x}{x} dx, \quad \varepsilon = 10^{-6};$

b) $\int_0^2 \frac{\ln(1+\sqrt{x})}{\sqrt[3]{x}} dx, \quad \varepsilon = 10^{-5}.$

5.8. *Izračunaj približne vrednosti sledećih integrala sa tačnošću* 10^{-4}.

a) $\int_0^1 \frac{dx}{(1+x)\sqrt{x}};$

b) $\int_0^1 \frac{xdx}{\sqrt{1-x^2}};$

c) $\int_1^2 \frac{dx}{x\sqrt{x^2-1}};$

d) $\int_0^1 \frac{dx}{\sqrt{x}\sqrt[4]{(1-x)^3}}.$

5.9. *Izračunaj približnu vrednost integrala*

a) $\int_2^\infty \frac{dx}{1+x^3};$

b) $\int_1^\infty \frac{xe^{-x^2}}{2+\sin x} dx.$

Glava 6

Sistemi linearnih jednačina

6.1 Uvod

U ovoj glavi ćemo se baviti numeričkim rešavanjem **sistema linearnih jednačina**. Posmatraćemo Gausov postupak eliminacije, koji spada u grupu direktnih postupaka za rešavanje sistema algebarskih jednačina, i iterativno rešavanje sistema linearnih jednačina. Pre toga ćemo navesti osnovne pojmove iz ove oblasti i posmatrati vektorske i matrične norme.

Skup realnih matrica formata $m \times n$ označavamo sa $\mathbb{R}^{m,n}$, a skup realnih $n-$dimenzionalnih vektora sa $\mathbb{R}^n$.

Definicija 6.1. ***Linearna jednačina.*** *Linearna jednačina sa nepoznatim $x_1, x_2, \ldots, x_n$ je jednačina*

$$x_1 a_1 + x_2 a_2 + \cdots + x_n a_n = b,$$

gde su $a_1, a_2, \ldots, a_n$ realni brojevi koje nazivamo koeficijenti jednačine, a b je realan broj koji nazivamo slobodni član jednačine.

Primer 6.1. *Jednačine*

$$2x_1 + x_2 - 4x_3 = 6 \qquad i \qquad -x_1 + x_2 - x_3 = 0$$

su linearne.

Ako je slobodni član b u linearnoj jednačini jednak nuli onda je ta jednačina **homogena**. Druga jednačina u prethodnom primeru je homogena.

Definicija 6.2. ***Sistem linearnih jednačina.*** *Sistem od m linearnih jednačina sa n nepoznatih $x_1, x_2, \ldots, x_n$ je skup linearnih jednačina oblika*

$$\begin{array}{ccccccccc}
a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\
a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\
\vdots & & \vdots & & & & \vdots & & \vdots \\
a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m
\end{array}$$

gde su a_{ij}, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$ i b_i, $i = 1, 2, \ldots, m$ realni brojevi. Konstante a_{ij} su ***koeficijenti sistema****, a vektor $b = [b_1 \;\; b_2 \;\; \cdots \;\; b_m]^\top$ je* ***slobodni vektor****.*

Ako je svaka komponenta $b_1, b_2, \ldots, b_m$ slobodnog vektora jednaka nuli, sistem je **homogen**.
Prethodni sistem linearnih jednačina može se zapisati i u **matričnom obliku**

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

ili kraće

$$Ax = b,$$

gde je

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}_{m\times n}, \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{n\times 1} \quad \text{i} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}_{m\times 1}.$$

Matrica A se naziva **matrica koeficijenata sistema**, ili kraće, **matrica sistema**.

Rešenje sistema linearnih jednačina je vektor $y = [y_1 \;\; y_2 \;\; \cdots \;\; y_n]^\top$ takav da je svaka jednačina sistema $Ax = b$ zadovoljena za vrednosti

$$x_i = y_i, \qquad i = 1, 2, \ldots, n.$$

Svaki sistem linearnih jednačina se može posmatrati u matričnom obliku $Ax = b$, ali i kao linearna kombinacija vektora

$$x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

ili kompaktnije,

$$x_1 a_1 + x_2 a_2 + \cdots + x_n a_n = b,$$

gde su a_i, $i = 1, 2, \ldots, n$, vektori kolone matrice koeficijenata A. Odatle, vidimo da se rešenje $x = [x_1 \;\; x_2 \;\; \cdots \;\; x_n]^\top$ može posmatrati kao vektor skalara takvih da je linearna kombinacija vektora kolona matrice A jednaka konstantnom vektoru b.

Primer 6.2. *Posmatrajmo sledeći sistem jednačina:*

$$\begin{array}{rcrcl} x_1 & + & x_2 & = & 3 \\ 3x_1 & + & x_2 & = & 5 \end{array}.$$

Ovaj sistem se može zapisati i u obliku

$$x_1 \begin{bmatrix} 1 \\ 3 \end{bmatrix} + x_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$$

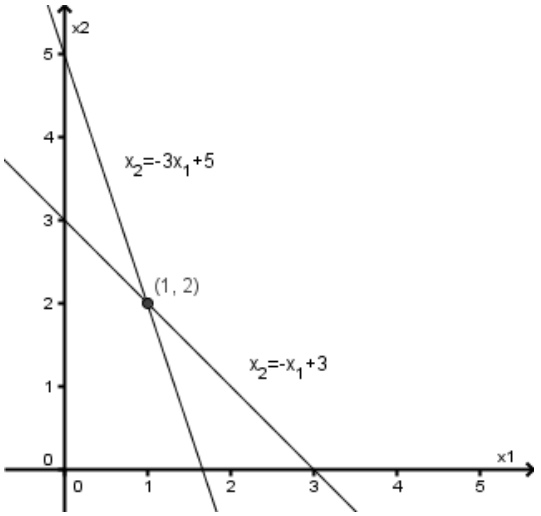
ili

$$x_1 a_1 + x_2 a_2 = b.$$

Posmatrani sistem se može rešiti i grafički. U tu svrhu posmatraju se prave

$$x_2 = -x_1 + 3 \qquad i \qquad x_2 = -3x_1 + 5.$$

Ove prave seku se u tački $(1, 2)$, *slika 15,*



Slika 15.

pa je traženo rešenje $x_1 = 1, x_2 = 2$. *Istovremeno smo dobili i prikaz vektora* $b = [3\ \ 5]^\top$ *kao linearne kombinacije vektora* $a_1 = [1\ \ 3]^\top$ *i* $a_2 = [1\ \ 1]^\top$:

$$1\begin{bmatrix} 1 \\ 3 \end{bmatrix} + 2\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}.$$

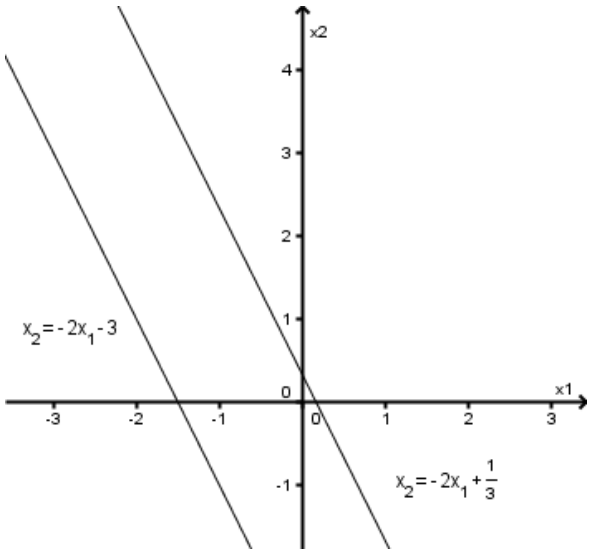
Primer 6.3. *Posmatrajmo sistem jednačina*

$$\begin{array}{rcrcr} 2x_1 & + & x_2 & = & -3 \\ 6x_1 & + & 3x_2 & = & 1 \end{array}.$$

Prave

$$x_2 = -2x_1 - 3 \qquad i \qquad x_2 = -2x_1 + \frac{1}{3}$$

su paralelne, slika 16,



Slika 16.

Znači, sistem nema rešenje, jer nema presečne tačke koja bi zadovoljila obe jednačine. Znači, ne postoje brojevi x_1 i x_2 takvi da se vektor $b = [3 \quad 5]^\top$ može prikazati kao linearna kombinacija

$$x_1 \begin{bmatrix} 2 \\ 6 \end{bmatrix} + x_2 \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

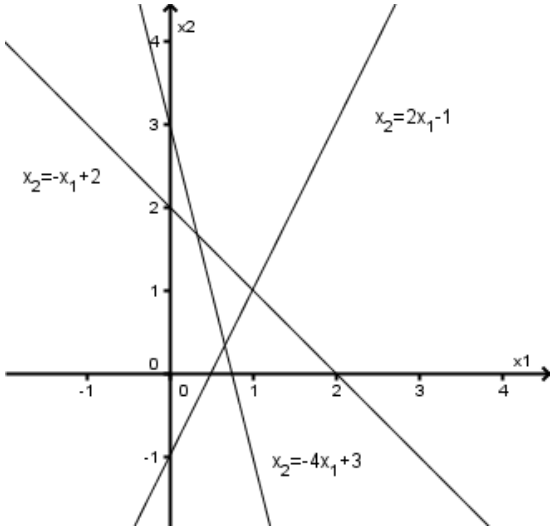
Primer 6.4. *Neka je dat sistem od tri jednačine sa dve nepoznate*

$$\begin{array}{rcrcl} x_1 & + & x_2 & = & 2 \\ 2x_1 & - & x_2 & = & 1 \\ 8x_1 & + & 2x_2 & = & 6 \end{array}.$$

Posmatrajmo prave

$$x_2 = -x_1 + 2, \qquad x_2 = 2x_1 - 1 \qquad i \qquad x_2 = -4x_1 + 3.$$

Vidimo da se prave ne seku u istoj tački, slika 17.



Slika 17.

To znači da posmatrani sistem nema rešenje.

Primer 6.5. *Posmatrajmo sistem*

$$\begin{array}{rcrcrcl} x_1 & + & x_2 & - & x_3 & = & 1 \\ 2x_1 & + & 3x_2 & + & x_3 & = & 3 \end{array}.$$

Ako prvu jednačinu pomnožimo sa -2 i dodamo drugoj dobićemo novu jednačinu

$$x_2 + 3x_3 = 1.$$

Iz ove jednačine sledi

$$x_2 = -3x_3 + 1.$$

Zamenjujući u prvoj jednačini x_2 sa $-3x_3 + 1$ dobijamo

$$x_1 + (-3x_3 + 1) - x_3 = 1,$$

tj.

$$x_1 = 4x_3.$$

Vidimo da za svaki proizvoljan broj x_3 dobijamo vrednosti za x_1 i x_2, takve da je vektor $[x_1 \;\; x_2 \;\; x_3]^\top$ rešenje posmatranog sistema. Na primer, za $x_3 = 0$ dobijamo $x_1 = 0$ i $x_2 = 1$. Za $x_3 = 1$ dobijamo $x_1 = 4$ i $x_2 = -2$.

Iz navedenih primera se vidi da sistemi linearnih jednačina mogu imati rešenja, ali i ne moraju imati rešenje.

- Sistem linearnih jednačina je **saglasan** (moguć, konzistentan) ako ima bar jedno rešenje. Ukoliko sistem nema ni jedno rešenje on je **nesaglasan** (protivrečan, nemoguć, kontradiktoran).
- Saglasan sistem je **određen** ako ima jedno rešenje, a **neodređen** ako ima više od jednog rešenja.
- Dva sistema linearnih jednačina su **ekvivalentna** ako je svako rešenje prvog sistema rešenje drugog i obrnuto, svako rešenje drugog sistema je rešenje prvog. Drugim rečima, sistemi koji imaju rešenja su ekvivalentni ako su im skupovi rešenja isti.
- Svaka dva nesaglasna sistema smatramo ekvivalentnim, tj. sistemi koji nemaju rešenja su ekvivalentni.

Primer 6.6. *Sistemi*

$$\begin{array}{rcrcr} x_1 & + & 2x_2 & = & 3 \\ x_1 & - & 2x_2 & = & -1 \end{array} \quad i \quad \begin{array}{rcrcr} 2x_1 & + & x_2 & = & 3 \\ 2x_1 & - & x_2 & = & 1 \end{array}$$

su određeni, tj. imaju po jedno rešenje. Rešenje prvog sistema $x_1 = 1, x_2 = 1$ je istovremeno i rešenje drugog sistema, pa su ovi sistemi ekvivalentni.

Primer 6.7. *Sistem*

$$\begin{array}{rcrcr} x_1 & + & 3x_2 & = & 5 \\ x_1 & - & x_2 & = & 1 \end{array}$$

ima rešenje $x_1 = 2, x_2 = 1$ koje je rešenje i sistema

$$\begin{array}{rcrcr} 2x_1 & + & 6x_2 & = & 10 \\ 3x_1 & + & 9x_2 & = & 15 \end{array}.$$

Drugi sistem pored rešenja $x_1 = 2, x_2 = 1$ ima još beskonačno mnogo rešenja oblika $x_1 = 5-3s, x_2 = s$, gde je s proizvoljan realan broj, koja nisu rešenja prvog sistema. Znači, posmatrani sistemi nisu ekvivalentni.

Sa teorijskog stanovišta sve je poznato o linearnim sistemima. Navešćemo Kroneker-Kapelijevu teoremu na osnovu koje možemo utvrditi da li je neki sistem linearnih jednačina saglasan, određen ili neodređen. U tu svrhu posmatramo **proširenu matricu** sistema linearnih jednačina i rang matrice sistema i rang proširene matrice.

Definicija 6.3. ***Rang matrice.*** *Maksimalan broj linearno nezavisnih vektora vrsta matrice naziva se rang matrice A i označava $r(A)$.*

Definicija 6.4. ***Saglasan linearni sistem.*** *Sistem linearnih jednačina je saglasan ako ima bar jedno rešenje. Ukoliko sistem nema ni jedno rešenje on je protivrečan. Saglasan sistem je određen ako ima jedno i samo jedno rešenje, a neodređen ako ima više rešenja.*

Definicija 6.5. ***Proširena matrica sistema.*** *Ako posmatramo linearni sistem $Ax = b$, gde je*

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad i \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix},$$

onda matricu

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right]$$

formata $m \times (n+1)$ *nazivamo* ***proširena matrica*** *sistema* $Ax = b$ *i obeležavamo je sa* $[A|b]$. *Slobodni članovi su odvojeni vertikalnom crtom, da bi se lakše razlikovali od koeficijenata sistema.*

Teorema 6.1. ***Teorema Kroneker-Kapelija.*** *Sistem linearnih jednačina* $Ax = b$ *je saglasan ako i samo ako je rang matrice* A *jednak rangu proširene matrice* $[A|b]$. *Ako je dati sistem saglasan, a zajednički rang matrice sistema* A *i proširene matrice jednak* r, *onda je sistem određen ako i samo ako je* $r = n$, *sistem je neodređen ako i samo ako je* $r < n$ *i tada se umesto nekih* $n - r$ *promenljivih mogu uzeti proizvoljne vrednosti, a preostale nepoznate su jednoznačno određene.*

6.2 Vektorske i matrične norme

Poznavanje vektorskih i matričnih normi nam je potrebno da bismo analizirali grešku Gausovog postupka eliminacije i da bismo se kasnije bavili iterativnim postupcima.

Definicija 6.6. ***Vektorska norma.*** *Vektorska norma je preslikavanje* $\|\cdot\|$ *skupa* $n-$*dimenzionalnih vektora u skup realnih brojeva sa osobinama:*

1) $\|x\| \geq 0, \qquad \|x\| = 0$ *ako i samo ako je* $x = 0$;

2) $\|\alpha x\| = |\alpha| \|x\|, \quad \alpha$ *je realan broj*

3) $\|x + y\| \leq \|x\| + \|y\|$.

Vektorska norma predstavlja uopštenje apsolutne vrednosti u skupu realnih brojeva na višedimenzionalne prostore.

Teorema 6.2. ***Vektorske*** $p-$***norme.****Preslikavanja*

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \quad 1 \leq p < \infty,$$

su vektorske norme.

Najčešće se koriste $p-$norme za $p = 1, 2$ i u graničnom slučaju $p = \infty$,

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \qquad \|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}, \qquad \|x\|_\infty = \max\{|x_i| : 1 \leq i \leq n\}.$$

Teorema 6.3. ***Ekvivalencija normi.*** *Neka su* $\|\cdot\|$ *i* $\|\cdot\|'$ *dve vektorske norme na* $\mathbb{R}^n$. *Tada postoje konstante* $c_2 \geq c_1 > 0$ *takve da je za svako* $x \in \mathbb{R}^n$

$$c_1 \|x\| \leq \|x\|' \leq c_2 \|x\|.$$

Definicija 6.7. ***Matrična norma.*** *Ako su $\|\cdot\|$ i $\|\cdot\|'$ vektorske norme definisane na $\mathbb{R}^n$ i $\mathbb{R}^m$ respektivno, onda preslikavanje $\|\cdot\|$ skupa matrica formata $m \times n$ na skup realnih brojeva definisana sa*

$$\|A\| = \sup_{\|x\|=1} \|Ax\|'.$$

nazivamo matrična norma u $\mathbb{R}^{m,n}$.

Teorema 6.4. ***Osobine matrične norme.*** *Matrična norma na $\mathbb{R}^{m,n}$ ima sledeće osobine:*

1) $\|A\| \geq 0, \qquad \|A\| = 0$ *ako i samo ako je* $A = 0$,

2) $\|\alpha A\| = |\alpha| \, \|A\|, \quad \alpha$ *je realan broj,*

3) $\|A + B\| \leq \|A\| + \|B\|,$

Teorema 6.5. *Za proizvoljne matrice $A, B \in \mathbb{R}^{n,n}$ je*

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

Vidimo da u odnosu na vektorsku normu ovde zahtevamo jednu osobinu više. Primetimo da za vektorsku i matričnu normu koristimo istu oznaku, što ne dovodi do zabune. Preslikavanje koje se definiše analogno vektorskoj normi $\|\cdot\|_\infty$

$$\|A\|'_\infty = \max\{|a_{ij}| : 1 \leq i \leq n, 1 \leq j \leq n\}$$

nije matrična norma. Mi ćemo ovde koristiti **prirodne matrične norme** koje su **indukovane** nekom vektorskom normom:

$$\|A\| = \sup_{\|x\|=1} \|Ax\|.$$

Ako je $\|\cdot\|$ $p-$norma na $\mathbb{R}^n$, onda su indukovane $p-$matrične norme

$$\|A\|_p = \sup_{\|x\|_p=1} \|Ax\|_p.$$

Ovi izrazi nisu pogodni za određivanje tih matričnih normi, ali postoje mnogo praktičniji oblici koje ćemo navesti u sledećim teoremama.

Teorema 6.6. ***Norma 1 i beskonačno.*** *Za matrične norme $\|A\|_1$ i $\|A\|_\infty$ važi*

$$\|A\|_1 = \max\left\{\sum_{i=1}^{n} |a_{ij}| : 1 \leq j \leq n\right\}, \qquad \|A\|_\infty = \max\left\{\sum_{j=1}^{n} |a_{ij}| : 1 \leq i \leq n\right\}.$$

Da bismo odredili matričnu normu $\|A\|_2$ potrebno je da poznajemo spektralni radijus matrice, tj. najveći moduo karakterističnih korena matrice $A^\top A$.

Ako je A matrica formata $n \times n$, tada se realan ili kompleksan broj λ i vektor $x \neq 0$ nazivaju **karakteristični koren** i **karakteristični vektor** matrice A, ako važi $Ax = \lambda x$. Otuda sledi da je $(A - \lambda E)\, x = 0$, gde je E jedinična matrica

$$E = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n}$$

pa kako je $x \neq 0$, sledi da je λ karakteristični koren matrice A ako i samo ako je $\det(A - \lambda E) = 0$. Ova jednačina se naziva **karakteristična jednačina** matrice A, a polinom $\det(A - \lambda E)$ je polinom $n-$tog stepena po λ i naziva se **karakteristični polinom.** On ima n korena $\lambda_1, \lambda_2, \ldots, \lambda_n$, ne obavezno različitih. Skup ovih korena se naziva **spektar** matrice A, a

$$\rho(A) = \max\{|\lambda_i| : 1 \leq i \leq n\}$$

naziva se **spektralni radijus** matrice A.

Teorema 6.7. ***Norma 2.*** *Za matričnu nromu* $\|A\|_2$ *važi*

$$\|A\|_2 = \sqrt{\rho(A^\top A)},$$

gde je $\rho\left(A^\top A\right)$ *spektralni radijus matrice* $A^\top A$.

Matrična norma je neprekidna funkcija i takođe važi teorema o ekvivalenciji normi.

Definicija 6.8. ***Saglasnost normi.*** *Ako za neku matričnu i neku vektorsku normu važi*

$$\|Ax\| \leq \|A\| \, \|x\|$$

za svaku matricu formata $n \times n$ *i svaki* $n-$*dimenzionalni vektor, kažemo da su vektorska i matrična norma* ***saglasne.***

Prirodna matrična norma je uvek saglasna sa vektorskom normom koja ju je indukovala. Pomenimo još jednu osobinu prirodnih normi. Za svaku prirodnu normu važi $\|E\| = 1$, gde je E jedinična matrica.

Teorema 6.8. ***Spektralni radijus i norma.*** *Za svaku prirodnu matričnu normu* $\|\cdot\|$ *i kvadratnu matricu* A *važi*

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

6.3 Gausov postupak eliminacije

Posmatrajmo sistem

$$\begin{array}{rcrcrcrcr} 6x_1 & + & x_2 & + & 4x_3 & - & x_4 & = & 9 \\ & & 7x_2 & - & 3x_3 & + & 2x_4 & = & -6 \\ & & & & -2x_3 & + & x_4 & = & 2 \\ & & & & & & 5x_4 & = & -20 \end{array}. \tag{6.1}$$

Iz poslednje jednačine

$$5x_4 = -20$$

lako se određuje x_4, tj.

$$x_4 = \frac{-20}{5} = -4.$$

Znajući x_4 iz pretposlednje jednačine

$$-2x_3 \quad + \quad x_4 \quad = \quad 2$$

određuje se x_3 :

$$x_3 = \frac{2 - x_4}{-2} = -\frac{2+4}{2} = -3.$$

Iz druge jednačine

$$7x_2 \quad - \quad 3x_3 \quad + \quad 2x_4 \quad = \quad -6$$

dobija se

$$x_2 = \frac{-6 + 3x_3 - 2x_4}{7} = \frac{-6 - 9 + 8}{7} = -1.$$

Konačno, iz prve jednačine

$$6x_1 \quad + \quad x_2 \quad + \quad 4x_3 \quad - \quad x_4 \quad = \quad 9$$

dobija se

$$x_1 = \frac{9 - x_2 - 4x_3 + x_4}{6} = \frac{9 + 1 + 12 - 4}{6} = 3.$$

Gausov postupak eliminacije sastoji se u određivanju sistema linearnih jednačina oblika (6.1) koji je ekvivalentan sistemu m jednačina sa n nepoznatih

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array} . \tag{6.2}$$

Jedan od načina da se za sistem (6.2) dobije ekvivalentan sistem oblika (6.1), ukoliko je to moguće, je primena **elementarnih transformacija** na sistem (6.2).

Elementarne transformacije sistema linearnih jednačina su sledeće transformacije tog sistema:

- Međusobna zamena bilo koje dve jednačine.
- Množenje bilo koje jednačine brojem različitim od nule.
- Dodavanje jedne jednačine pomnožene bilo kojim brojem nekoj drugoj jednačini.

Primenom konačnog broja elementarnih transformacija na sistem (6.2) dobija se njemu ekvivalentan sistem.

Može se desiti da se posle vršenja određenog broja elementarnih transformacija dođe do sistema u kome su u jednoj jednačini svi koeficijenti jednaki nuli. Ako je i slobodan član te jednačine jednak nuli, onda je svaka $n-$torka brojeva rešenje te jednačine, pa se izostavljajući tu jednačinu, dobija sistem ekvivalentan sa polaznim. Ako je slobodan član te jednačine različit od nule, onda ona uopšte nema rešenje, pa je dobijeni sistem protivrečan, a takođe i polazni sistem, pa se dalji postupak obustavlja.

Primer 6.8. *Neka je dat sledeći sistem linearnih jednačina*

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & - & 2x_3 & = & 5 \\ 2x_1 & + & 4x_2 & + & 2x_3 & = & 7 \\ -x_1 & - & 5x_2 & + & 8x_3 & = & 20 \end{array} .$$

Ako pomnožimo prvu jednačinu sa -2 *i dodamo je drugoj dobićemo sistem*

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & - & 2x_3 & = & 5 \\ & - & 2x_2 & + & 6x_3 & = & -3 \\ -x_1 & - & 5x_2 & + & 8x_3 & = & 20 \end{array} .$$

Ako sada prvu jednačinu dodamo trećoj dobićemo sistem

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & - & 2x_3 & = & 5 \\ & - & 2x_2 & + & 6x_3 & = & -3 \\ & - & 2x_2 & + & 6x_3 & = & 25 \end{array} .$$

Posle dodavanja druge jednačine prethodno pomnožene sa -1 trećoj, iz poslednjeg sistema dobijamo

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & - & 2x_3 & = & 5 \\ & - & 2x_2 & + & 6x_3 & = & -3 \\ & & 0x_2 & + & 0x_3 & = & 22 \end{array} .$$

Poslednji sistem ima koeficijente u trećoj jednačini jednake nuli, a slobodan član je različit od nule. To znači da je taj sistem, pa prema tome i polazni koji je sa njim ekvivalentan, protivrečan.

Neka je dat sistem od m linearnih jednačina (6.2) sa n nepoznatih. Pretpostavimo da je koeficijent $a_{11} \neq 0$. Ukoliko bi bilo $a_{11} = 0$ onda se vršeći elementarne transformacije može prva jednačina zameniti jednačinom u kojoj je koeficijent uz x_1 različit od nule, a onda menjajući numeraciju koeficijenata dobijamo sistem, ekvivalentan sa polaznim, u kome je $a_{11} \neq 0$.

Pomnožimo prvu jednačinu najpre sa $-\dfrac{a_{21}}{a_{11}}$ i dodajmo je drugoj, zatim je pomnožimo sa $-\dfrac{a_{31}}{a_{11}}$ i dodajmo trećoj itd., na kraju prvu jednačinu množimo sa $-\dfrac{a_{m1}}{a_{11}}$ i dodajmo poslednjoj jednačini. Tako dobijamo sistem

$$\begin{array}{rcrcccrcr} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ & & c_{22}x_2 & + & \cdots & + & c_{2n}x_n & = & d_2 \\ & & \vdots & & & & \vdots & & \vdots \\ & & c_{m2}x_2 & + & \cdots & + & c_{mn}x_n & = & d_m \end{array} \tag{6.3}$$

koji je ekvivalentan sa (6.2). Može se desiti da su u nekoj jednačini sistema (6.3) svi koeficijenti jednaki nuli. Ako je slobodan član u toj jednačini različit od nule, onda je sistem (6.3) protivrečan pa je njemu ekvivalentan sistem (6.2) takođe protivrečan. Ako je slobodan član jednak nuli, onda s obzirom na ono što je ranije rečeno, ta se jednačina može odbaciti. To znači da sitem (6.3) ima m ili manje od m jednačina. Ako sistem (6.3) ima više od dve jednačine nastavićemo njegovo transformisanje.

Ukoliko se desi da su svi koeficijenti $c_{i2} = 0$, $i = 2, 3, \ldots, m$, tj. da nepoznate x_2 nema ni u jednoj jednačini sistema (6.3) sem prve, onda ćemo promeniti numeraciju nepoznatih tako da se nova nepoznata x_2 pojavi još u nekoj jednačini sem prve.

U daljem postupku nećemo dirati prvu jednačinu, a sa ostalim ćemo postupati kao što smo na početku postupali sa sistemom (6.2) (množimo drugu jednačinu sa $-\dfrac{c_{32}}{c_{22}}$ i dodajemo je trećoj itd.). Tako dobijamo sistem

$$\begin{array}{rcrcrcccrcr} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ & & c_{22}x_2 & + & c_{23}x_3 & + & \cdots & + & c_{2n}x_n & = & d_2 \\ & & & & e_{33}x_3 & + & \cdots & + & e_{3n}x_n & = & f_3 \\ & & & & \vdots & & & & \vdots & & \vdots \\ & & & & e_{m3}x_3 & + & \cdots & + & e_{mn}x_n & = & f_m \end{array} \tag{6.4}$$

koji je ekvivalentan sa (6.3), a to znači i sa (6.2). Na isti način kao i u prethodnom koraku zaključujemo da ako u sistemu (6.4) neka od jednačina ima sve koeficijente jednake nuli a slobodan član te jednačine je različit od nule, onda je sistem (6.4), a to znači i polazni sistem (6.2), protivrečan. Ako je slobodan član te jednačine jednak nuli, onda tu jednačinu odbacujemo.

Produžavajući tako dalje ili ćemo dokazati da je sistem (6.2) protivrečan, ili ćemo doći do jednog od sledeća dva sistema

$$\begin{array}{rcrcccrcr} g_{11}x_1 & + & g_{12}x_2 & + & \cdots & + & g_{1n}x_n & = & h_1 \\ & & g_{22}x_2 & + & \cdots & + & g_{2n}x_n & = & h_2 \\ & & & & \ddots & & \vdots & & \vdots \\ & & & & & & g_{nn}x_n & = & h_n \end{array} \tag{6.5}$$

gde je $g_{ii} \neq 0$, $i = 1, 2, \ldots, n$, (ovakav sistem se naziva **trougaoni**), ili

$$\begin{array}{ccccccccccccc} p_{11}x_1 & + & p_{12}x_2 & + & \cdots & + & p_{1k}x_k & + & \cdots & + & p_{1n}x_n & = & q_1 \\ & & p_{22}x_2 & + & \cdots & + & p_{2k}x_k & + & \cdots & + & p_{2n}x_n & = & q_2 \\ & & & & \ddots & & \vdots & & & & \vdots & & \vdots \\ & & & & & & p_{kk}x_k & + & \cdots & + & p_{kn}x_n & = & q_k \end{array} \tag{6.6}$$

gde je $p_{ii} \neq 0$, $i = 1, 2, \ldots, k$, $k < n$.

Uvodni primer se odnosio na sistem oblika (6.5), a sledeći primer ilustruje slučaj (6.6).

Primer 6.9. *Neka je u trećoj jednačini sistema iz prethodnog primera slobodan član* 20 *zamenjen sa* -8. *Tada imamo novi sistem*

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & - & 2x_3 & = & 5 \\ 2x_1 & + & 4x_2 & + & 2x_3 & = & 7 \\ -x_1 & - & 5x_2 & + & 8x_3 & = & -8 \end{array} .$$

Ako pomnožimo prvu jednačinu sa -2 *i dodamo je drugoj dobićemo sistem*

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & - & 2x_3 & = & 5 \\ & - & 2x_2 & + & 6x_3 & = & -3 \\ -x_1 & - & 5x_2 & + & 8x_3 & = & 20 \end{array} .$$

Ako sada prvu jednačinu dodamo trećoj dobićemo sistem

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & - & 2x_3 & = & 5 \\ & - & 2x_2 & + & 6x_3 & = & -3 \\ & - & 2x_2 & + & 6x_3 & = & -3 \end{array} .$$

Posle dodavanja druge jednačine prethodno pomnožene sa -1 *trećoj, iz poslednjeg sistema dobijamo*

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & - & 2x_3 & = & 5 \\ & - & 2x_2 & + & 6x_3 & = & -3 \\ & & 0x_2 & + & 0x_3 & = & 0 \end{array} .$$

Poslednji sistem ima koeficijente u trećoj jednačini jednake nuli, a i slobodan član je jednak nuli. To znači da se treća jednačina može odbaciti, tako da se dobija sistem

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & - & 2x_3 & = & 5 \\ & - & 2x_2 & + & 6x_3 & = & -3 \end{array} .$$

U slučaju kada se dobije sistem (6.5) imamo jedinstveno rešenje. Zaista, iz poslednje jednačine se odmah dobija jedinstveno određeno x_n. Zamenjujući tu dobijenu vrednost za x_n u pretposlednjoj jednačini dobijamo jedinstveno određeno x_{n-1}. Produžavajući tako, dobijamo da sistem (6.5), a to znači i njemu ekvivalentan polazni sistem (6.2), ima jedinstveno rešenje.

U slučaju da smo dobili sistem (6.6), možemo ga napisati u obliku

$$\begin{array}{ccccccccccccccc} p_{11}x_1 & + & p_{12}x_2 & + & \cdots & + & p_{1k}x_k & = & q_1 & - & p_{1,k+1}x_{k+1} & - & \cdots & - & p_{1n}x_n \\ & & p_{22}x_2 & + & \cdots & + & p_{2k}x_k & = & q_2 & - & p_{2,k+1}x_{k+1} & - & \cdots & - & p_{2n}x_n \\ & & & & \ddots & & \vdots & & \vdots & & \vdots & & & & \vdots \\ & & & & & & p_{kk}x_k & = & q_k & - & p_{k,k+1}x_{k+1} & - & \cdots & - & p_{kn}x_n \end{array} \tag{6.7}$$

Ako umesto nepoznatih $x_{k+1}, \ldots, x_n$ u sistemu (6.7) stavimo proizvoljne brojeve, dobićemo sistem oblika (6.5) za koji smo videli da ima jedinstveno rešenje po nepoznatim $x_1, x_2, \ldots, x_k$. S obzirom da vrednosti za nepoznate $x_{k+1}, \ldots, x_n$ možemo birati na beskonačno mnogo načina, sledi da sistem (6.7), odnosno njemu ekvivalentan sistem (6.2), ima beskonačno mnogo rešenja. Na taj način se i dobijaju sva rešenja sistema (6.7), jer su za date vrednosti nepoznatih $x_{k+1}, x_{k+2}, \ldots, x_n$ vrednosti ostalih nepoznatih jednoznačno određene.

Primer 6.10. *Sistem iz prethodnog primera*

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & - & 2x_3 & = & 5 \\ & - & 2x_2 & + & 6x_3 & = & -3 \end{array}$$

možemo zapisati u obliku

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & = & 5 & + & 2x_3 \\ & - & 2x_2 & = & -3 & - & 6x_3 \end{array}.$$

Izaberimo $x_3 = a$. *Tada iz druge jednačine* $-2x_2 + 6a = -3$ *dobijamo*

$$x_2 = \frac{-3 - 6a}{-2} = \frac{3}{2}(1 + 2a).$$

Stavimo li ovu vrednost u prvu jednačinu dobijamo

$$x_1 + 3 \cdot \frac{3}{2}(1 + 2a) - 2a = 5,$$

a odatle

$$x_1 = \frac{1}{2} - 7a.$$

Prema tome, dati sistem je neodređen, ima beskonačno mnogo rešenja

$$x = \begin{bmatrix} 0.5 - 7a \\ 1.5 + 3a \\ a \end{bmatrix},$$

gde je a proizvoljan realan broj.

Dakle, zaključili smo sledeće:

- Gausov metod eliminacije se može primeniti na svaki sistem linearnih jednačina.
- Ako opisanim postupkom eliminacije dođemo do jednačine u kojoj su svi koeficijenti jednaki nuli, a slobodan član nije, dati sistem je protivrečan. Ukoliko se do takve jednačine ne dođe, sistem je saglasan.
- Saglasan sistem je određen ako se na kraju eliminacije dođe do sistema oblika (6.5) i iz njega se dobija rešenje datog sistema. Saglasan sistem je neodređen ako se dođe do sistema oblika (6.6) i iz tog sistema se dobijaju sva rešenja datog sistema.

Prilikom praktičnog rešavanja sistema linearnih jednačina Gausovom eliminacijom, radi bržeg i jednostavnijeg pisanja, zgodno je pisati samo koeficijente i slobodne članove sistema bez nepoznatih i oznaka za operacije. Tako, umesto sistema (6.2) od m jednačina sa n nepoznatih, možemo pisati samo njegove koeficijente i slobodne članove u obliku matrice tipa $m \times (n + 1)$, koja se naziva **proširena matrica** sistema $Ax = b$ i obeležava sa $[A|b]$,

$$[A|b] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right]. \tag{6.8}$$

Da bi se lakše razlikovali od koeficijenata sistema, slobodni članovi su odvojeni vertikalnom crtom. Svi elementi transformacije koje bismo vršili na datom sistemu vršimo na matrici (6.8). Tako, na primer, ako prvu jednačinu sistema treba pomnožiti sa $-\frac{a_{21}}{a_{11}}$ i dodati drugoj, onda ćemo elemente prve vrste matrice (6.8) pomnožiti sa $-\frac{a_{21}}{a_{11}}$ i dodati ih odgovarajućim elementima druge vrste.

Primer 6.11. *Neka je dat sistem*

$$\begin{array}{rcrcrcrcr} x_1 & + & 2x_2 & - & 12x_3 & + & 8x_4 & = & -10 \\ 5x_1 & + & 4x_2 & - & 7x_3 & - & 2x_4 & = & 8 \\ -3x_1 & + & 6x_2 & + & 9x_3 & + & 5x_4 & = & 17 \\ 6x_1 & - & 12x_2 & - & 18x_3 & + & 3x_4 & = & -21 \end{array}.$$

Transformisaćemo proširenu matricu $[A|b]$

$$[A|b] = \left[\begin{array}{rrrr|r} 1 & 2 & -12 & 8 & -10 \\ 5 & 4 & -7 & -2 & 8 \\ -3 & 6 & 9 & 5 & 17 \\ 6 & -12 & -18 & 3 & -21 \end{array}\right] \sim \left[\begin{array}{rrrr|r} 1 & 2 & -12 & 8 & -10 \\ 0 & -6 & 53 & -42 & 58 \\ 0 & 12 & -27 & 29 & -13 \\ 0 & -24 & 54 & -45 & 39 \end{array}\right]$$

$$\sim \left[\begin{array}{rrrr|r} 1 & 2 & -12 & 8 & -10 \\ 0 & -6 & 53 & -42 & 58 \\ 0 & 0 & 79 & -55 & 103 \\ 0 & 0 & -158 & 123 & -193 \end{array}\right] \sim \left[\begin{array}{rrrr|r} 1 & 2 & -12 & 8 & -10 \\ 0 & -6 & 53 & -42 & 58 \\ 0 & 0 & 79 & -55 & 103 \\ 0 & 0 & 0 & 13 & 13 \end{array}\right].$$

Sada dobijamo

$$x_4 = \frac{13}{13} = 1, \qquad x_3 = \frac{103 + 55x_4}{79} = \frac{158}{79} = 2,$$

$$x_2 = \frac{58 + 42x_4 - 53x_3}{-6} = \frac{-6}{-6} = 1 \qquad i \qquad x_1 = -10 - 8x_4 + 12x_3 - 2x_2 = 4.$$

6.4 Analiza greške postupka eliminacije

U ovom odeljku ćemo posmatrati rešavanje kvadratnog sistema linearnih jednačina $Ax = b$ sa regularnom matricom sistema A. Kao što smo već rekli kod postupka eliminacije javlljaju se greške zaokrugljivanja, dakle ovde nemamo greške postupka već isključivo greške izračunavanja. Neka je x^* približno rešenje sistema $Ax = b$ i neka je

$$r = b - Ax^*.$$

Teorema 6.9. ***Ocena greške.*** *Za saglasne norme važi*

$$\|x - x^*\| \leq \left\|A^{-1}\right\| \|r\|.$$

Dokaz. Očigledno je

$$x - x^* = A^{-1}b - A^{-1}Ax^* = A^{-1}\left(b - Ax^*\right) = A^{-1}r$$

i

$$\|x - x^*\| = \left\|A^{-1}r\right\| \leq \left\|A^{-1}\right\| \|r\|.$$

■

Ova teorema pokazuje da $\|r\|$ ne mora biti dobar pokazatelj odstupanja približnog rešenja x^* od tačnog rešenja x. Na grešku ima uticaja i A^{-1}. Ako je $\left\|A^{-1}\right\|$ veliko, onda i greška može biti velika čak i kad je $\|r\|$ malo. Sistemi kod kojih je to slučaj su loše uslovljeni. Merilo dobre ili loše uslovljenosti može biti upravo $\left\|A^{-1}\right\|$, ali je bolje koristiti **uslovni broj matrice**

$$k\left(A\right) = \|A\| \left\|A^{-1}\right\|,$$

koji se definiše za neku prirodnu matričnu normu. Očigledno važi

$$k\left(A\right) \geq \left\|AA^{-1}\right\| = \|E\| = 1. \|x - x^*\| \leq \left\|A^{-1}\right\| \|r\|$$

Uslovni broj se javlja u sledećoj teoremi o relativnoj greški približnog rešenja x^*.

Teorema 6.10. ***Ocena relativne greške.*** *Za $x \neq 0$ važi*

$$\frac{\|x - x^*\|}{\|x\|} \leq k(A) \frac{\|r\|}{\|b\|},$$

gde je matrična norma indukovana vektorskom normom $\|\cdot\|$.

Dokaz. Na osnovu prethodne teoreme sledi

$$\frac{\|x - x^*\|}{\|x\|} \leq \frac{\left\|A^{-1}\right\|}{\|x\|} \|r\|,$$

pa ostaje da dokažemo da je

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|},$$

što dobijamo iz

$$\|b\| = \|Ax\| \leq \|A\| \|x\|.$$

■

Primetimo da važi

$$\frac{\|r\|}{\|b\|} = \frac{\|b - b^*\|}{\|b\|}$$

za $b^* = Ax^*$, tako da je količnik $\frac{\|r\|}{\|b\|}$ relativna greška vektora b^*.

Da bismo ocenili grešku $\|x - x^*\|$ trebalo bi da znamo $\left\|A^{-1}\right\|$. Međutim, najčešće nemamo tačnu inverznu matricu. Neka je C približna inverzna matrica za matricu A, tj. neka je $C \approx A^{-1}$ rešenje jednačine

$$AX = E.$$

Teorema 6.11. ***Norma inverzne matrice.*** *Neka je $R = E - AC$ ili $R = E - CA$. Ako je*

$$\|R\| < 1$$

onda za prirodnu matričnu normu važi

$$\left\|A^{-1}\right\| \leq \frac{\|C\|}{1 - \|R\|}.$$

Dokaz. Neka je na primer $R = E - AC$, neka je λ karakteristični koren matrice R. Tada je $1 - \lambda$ karakteristični koren matrice $E - R$. Kako je $\rho(R) \leq \|R\|$, sledi da je $|\lambda| < 1$, tj. $1 - |\lambda| > 0$, pa je matrica $E - R$ regularna. Zbog toga imamo

$$A^{-1} = C(E - R)^{-1}$$

i

$$\left\|A^{-1}\right\| \leq \|C\| \left\|(E - R)^{-1}\right\|.$$

Dalje iz

$$(E - R)(E - R)^{-1} = E$$

nalazimo

$$(E - R)^{-1} = E + R(E - R)^{-1},$$

$$\left\|(E - R)^{-1}\right\| \leq 1 + \left\|R(E - R)^{-1}\right\| \leq 1 + \|R\| \left\|(E - R)^{-1}\right\|$$

$$\left\|(E - R)^{-1}\right\| (1 - \|R\|) \leq 1.$$

Odavde i iz ocene za $\left\|A^{-1}\right\|$ sledi tvrđenje. ■

Primetimo da u prethodnoj teoremi nije trebalo pretpostavljati da je A regularna matrica. To sledi iz $\|R\| < 1$, jer je tada $E - R = AC$ (ili $E - R = CA$) regularna matrica, pa i A i C moraju biti regularne matrice.

Na osnovu prethodne dve teoreme dobijamo

$$\|x - x^*\| = \left\|A^{-1}r\right\| \leq \left\|A^{-1}\right\| \|r\| \leq \frac{\|C\| \, \|r\|}{1 - \|R\|},$$

gde je matrična norma indukovana datom vektorskom normom i $\|R\| < 1$.

Na osnovu prethodne teoreme možemo oceniti grešku približne inverzne matrice C. Zbog

$$C - A^{-1} = A^{-1}(E - R) - A^{-1} = -A^{-1}R$$

imamo

$$\left\|C - A^{-1}\right\| = \left\|A^{-1}\right\| \|R\| \leq \frac{\|C\| \, \|R\|}{1 - \|R\|}.$$

6.5 Iterativni postupci

6.5.1 Opšti iterativni postupak

U ovom delu će se razmatrati iterativni postupci za rešavanje sistema linearnih jednačina $Ax = b$, gde je $A \in \mathbb{R}^{n,n}$ regularna matrica i $b \in \mathbb{R}^n$. Jedinstveno rešenje $x = A^{-1}b$ posmatranog sistema određuje se kao rešenje ekvivalentnog sistema

$$x = Gx + d. \tag{6.9}$$

Dva sistema linearnih jednačina su ekvivalentna ako je svako rešenje jednog sistema rešenje i drugog sistema i obrnuto, što u ovom slučaju znači da matrica $E - G$ treba da bude regularna i da važi

$$(E - G)^{-1} d = A^{-1}b, \qquad \text{odnosno} \qquad d = (E - G) A^{-1}b.$$

Za regularnu matricu A postoji proizvoljno mnogo matrica G takvih da su sistemi $Ax = b$ i $x = Gx + d$ ekvivalentni, ali su samo neke matrice pogodne za iterativnu matricu. U opštem slučaju se ne može reći kako birati matricu G tako da se dobije konvergentan iterativni postupak, već su potrebne neke dodatne pretpostavke o matrici A. Najčešće se posmatra **razlaganje** matrice A,

$$A = M - N, \tag{6.10}$$

pri čemu je M regularna matrica, pa se zatim sistem $Ax = b$ transformiše u sistem

$$x = M^{-1}Nx + M^{-1}b,$$

odnosno na oblik $x = Gx + d$, gde je

$$G = M^{-1}N, \quad d = M^{-1}b.$$

Za neke specijalne klase matrica je poznato razlaganje matrice A na matrice M i N koje daje konvergentan postupak, kao što će se pokazati u narednim paragrafima.

Polazeći od sistema $Ax = b$ i proizvoljnog **startnog** vektora $x^0 \in \mathbb{R}^n$ izračunava se **iterativni niz** $x^0, x^1, \ldots$ po **iterativnom pravilu**

$$x^{k+1} = Gx^k + d, \quad k = 0, 1, \ldots. \tag{6.11}$$

Pri tome je za svako $k = 0, 1, \ldots,$

$$x^k = \begin{bmatrix} x_1^k \\ x_2^k \\ \vdots \\ x_n^k \end{bmatrix}.$$

Interesuje nas šta se dešava kada $k \to \infty$, tj. kakav se niz vektora dobija.

Definicija 6.9. ***Granična vrednost niza vektora.*** *Neka je dat niz vektora* $x^k \in \mathbb{R}^n$, $k = 0, 1, \ldots$. *Ako za svako* $i = 1, 2, \ldots, n$ *postoji*

$$\lim_{k\to\infty} x_i^k = x_i,$$

onda se vektor

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

naziva granična vrednost niza $x^0, x^1, \ldots$, *i za taj niz se kaže da konvergira ka* x. *To zapisujemo sa*

$$\lim_{k\to\infty} x^k = x \quad \textit{ili} \quad x^k \to x.$$

Lako je proveriti da za konvergenciju niza vektora važi sledeća teorema.

Teorema 6.12. *Ako je* $x^0, x^1, \ldots$ *niz vektora, onda je* $\lim_{k\to\infty} x^k = x$ *ako i samo ako je za neku vektorsku normu*

$$\lim_{k\to\infty} \left\|x^k - x\right\| = 0.$$

Imajući u vidu definiciju granične vrednosti niza vektora, lako se dokazuje sledeća teorema.

Teorema 6.13. *Ako je niz vektora* $x^0, x^1, \ldots$ *konvergentan i* $\lim_{k\to\infty} x^k = x$, *onda za svaku matricu* G *i svaki vektor* d *važi*

$$\lim_{k\to\infty} \left(Gx^k + d\right) = Gx + d.$$

Nas će naravno interesovati kada niz $x^0, x^1, \ldots$ konvergira ka rešenju sistema $x = Gx + d$. Kada je **niz** $x^0, x^1, \ldots$ **konvergentan,** kaže se da **iterativni postupak konvergira**. Matrica G se naziva **iterativna matrica.** Pokazuje se da je pod navedenim pretpostavkama o matrici G i vektoru d granična vrednost niza $x^0, x^1, \ldots$, ako postoji, rešenje sistema $x = Gx + d$, odnosno $Ax = b$.

Potreban i dovoljan uslov za konvergenciju iterativnog postupka dat je u sledećoj teoremi.

Teorema 6.14. ***Potreban i dovoljan uslov za konvergenciju.*** *Niz* $x^0, x^1, \ldots$ *dat iterativnim pravilom konvergira ka jedinstvenom rešenju sistema* $x = Gx + d$ *za proizvoljno* $x^0 \in \mathbb{R}^n$ *ako i samo ako je* $\rho(G) < 1$.

Potreban i dovoljan uslov za konvergenciju iterativnog postupka dat u prethodnoj teoremi nije lako proverljiv, jer je izračunavanje karakterističnih korena matrica većih dimenzija složen numerički problem. Zbog toga se često koristi sledeća teorema koja daje dovoljan uslov za konvergenciju iterativnog postupka.

Teorema 6.15. ***Dovoljan uslov za konvergenciju.*** *Neka je za neku matričnu normu* $\|G\| = \gamma < 1$, *gde je* $\gamma \in [0, 1)$. *Tada sistem* $x = Gx + d$ *ima jedinstveno rešenje* x *koje je granična vrednost niza* $x^0, x^1, \ldots$ *datog iterativnim pravilom*

$$x^{k+1} = Gx^k + d, \quad k = 0, 1, \ldots.$$

sa proizvoljnim $x^0 \in \mathbb{R}^n$ *i važi*

$$\left\|x^k - x\right\| \le \frac{\gamma}{1-\gamma}\left\|x^k - x^{k-1}\right\| \le \frac{\gamma^k}{1-\gamma}\left\|x^1 - x^0\right\|, \quad k = 1, 2, \ldots.$$

Dokaz. Ako je je λ karakteristični koren matrice G, onda je $1-\lambda$ karakteristični koren matrice $E-G$. Kako je $\rho(G) \leq \|G\| = \gamma < 1$, sledi da je $|\lambda| < 1$, tj. $1-|\lambda| > 0$, pa je matrica $E-G$ regularna, što znači da sistem $x = Gx + d$ ima jedinstveno rešenje x. Pošto radimo sa saglasnim matričnim i vektorskim normama, za proizvoljne vektore u i v važi

$$\|Gu - Gv\| = \|G(u-v)\| \leq \|G\| \, \|u-v\| \leq \gamma \, \|u-v\| .$$

Neka je $x^0 \in \mathbb{R}^n$ proizvoljan vektor i

$$x^{k+1} = Gx^k + d, \quad k = 0, 1, \ldots .$$

Ako je $x^1 = x^0$, onda je taj vektor rešenje sistema $x = Gx + d$ i niz x^0, $x^1, \ldots$ je stacionarni niz sa granicom $x = x^1$. U tom slučaju tvrđenje teoreme je dokazano. Neka je $x^1 \neq x^0$. Tada je za $k = 0, 1, \ldots$

$$\left\|x^{k+1} - x\right\| = \left\|Gx^k + d - Gx - d\right\| \leq \gamma \left\|x^k - x\right\| \leq \gamma^2 \left\|x^{k-1} - x\right\| \leq \cdots \leq \gamma^k \left\|x^1 - x\right\| .$$

Kako je $\gamma < 1$, sledi $\lim\limits_{k\to\infty} \gamma^{k+1} = 0$, odnosno

$$\lim_{k\to\infty} \left\|x^{k+1} - x\right\| = 0.$$

To znači da je

$$\lim_{k\to\infty} x^k = x.$$

Sada iz

$$x = \lim_{k\to\infty} x^{k+1} = \lim_{k\to\infty} \left(Gx^k + d\right) = Gx + d$$

vidimo da je iterativni niz konvergentan i da je njegova granična vrednost x rešenje polaznog sistema linearnih jednačina.
Pošto je $\|G\| = \gamma < 1$ i $x = Gx + d$, važi

$$\left\|x^k - x\right\| = \left\|Gx^{k-1} - Gx^k\right\| + \left\|Gx^k - Gx\right\| \leq \gamma \left\|x^{k-1} - x^k\right\| + \gamma \left\|x^{k-1} - x\right\| .$$

Odatle, zbog $1-\gamma > 0$, sledi

$$\left\|x^k - x\right\| \leq \frac{\gamma}{1-\gamma} \left\|x^{k-1} - x^k\right\| .$$

Takođe je

$$\left\|x^k - x^{k-1}\right\| = \left\|Gx^{k-1} - Gx^{k-2}\right\| \leq \gamma \left\|x^{k-1} - x^{k-2}\right\| \leq \cdots \leq \gamma^{k-1} \left\|x^1 - x^0\right\| ,$$

odnosno

$$\left\|x^k - x\right\| \leq \frac{\gamma^k}{1-\gamma} \left\|x^1 - x^0\right\| .$$

■

Vrednost

$$A_k = \frac{\gamma}{1-\gamma} \left\|x^k - x^{k-1}\right\|$$

naziva se **aposteriorna** ocena greške, a

$$B_k = \frac{\gamma^k}{1-\gamma} \left\|x^1 - x^0\right\|$$

apriorna ocena greške. Očigledno je $A_k \leq B_k$, odnosno aposteriorna ocena greške je bolja od apriorne, ali se apriorna greška može izračunati na početku iterativnog postupka, na osnovu prve dve aproksimacije.

Smisao iterativnih postupaka je u tome što direktni postupci mogu biti veoma osetljivi na greške zaokrugljivanja, dok se kod iterativnih postupaka svaka novodobijena iteracija, iako je pri njenom izračunavanju zaokrugljivanje prisutno, može smatrati za novu početnu iteraciju. Prednost iterativnih postupaka se ogleda i u tome što oni često zahtevaju manje memorijskog prostora od direktnih postupaka. Ako je matrica G retka (veliki broj njenih elemenata je jednak nuli) dovoljno je u memoriji računara čuvati samo elemente različite od nule. Pri tome se matrica G ne menja u toku iterativnog postupka. S druge strane, Gausov postupak u opštem slučaju ne očuvava osobinu matrice da je retka. U svakom koraku iterativnog postupka potrebno je u opštem slučaju izvršiti oko n^2 množenja i sabiranja. To znači da ako je broj iteracija manji od $n/3$, iterativni postupak ima manje operacija od Gausovog. Osnovni nedostatak iterativnih postupaka je u tome što im je primena uglavnom ograničena na pojedine klase matrica, za koje se unapred može garantovati konvergencija.

6.5.2 Jakobijev i Gaus-Zajdelov postupak

Ovde ćemo govoriti o dva iterativna postupka kod kojih se matrica iterativnog koraka G određuje na poseban način. To su **Jakobijev** i **Gaus-Zajdelov postupak.**

Neka je A regularna matrica sa nenula dijagonalnim elementima i

$$A = D - T - S$$

njeno **standardno razlaganje** na dijagonalnu matricu

$$D = \operatorname{diag}\left(a_{11}, a_{22}, \ldots, a_{nn}\right),$$

strogo donju trougaonu matricu T i strogo gornju trougaonu matricu S, tj.

$$T = -\begin{bmatrix} 0 & & & & & & \\ a_{21} & 0 & & & & & \\ a_{31} & a_{32} & 0 & & & & \\ \vdots & \vdots & & \ddots & & & \\ a_{n-1,1} & a_{n-1,2} & a_{n-1,3} & \cdots & 0 & & \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{n,n-1} & 0 \end{bmatrix},$$

$$S = -\begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ & 0 & a_{23} & \cdots & a_{2n} \\ & & \ddots & & \vdots \\ & & & 0 & a_{n-1,n} \\ & & & & 0 \end{bmatrix}.$$

Matrica D je tada regularna matrica i njena inverzna matrica je

$$D^{-1} = \operatorname{diag}(\frac{1}{a_{11}}, \frac{1}{a_{22}}, \ldots, \frac{1}{a_{nn}}).$$

Sistem $Ax = b$ može se zapisati u ekvivalentnom obliku

$$Dx = (T + S)x + b,$$

odnosno

$$x = D^{-1}\left(T + S\right)x + D^{-1}b.$$

Sada se može definisati iterativni postupak

$$x^{k+1} = B_J x^k + d, \quad k = 0, 1, \ldots$$

gde je

$$B_J = D^{-1}(T+S), \qquad d = D^{-1}b.$$

Ovaj postupak se naziva **Jakobijev iterativni postupak**, a matrica B_J je **Jakobijeva iterativna matrica**. Komponente vektora x^{k+1} se izračunavaju po formuli

$$x_i^{k+1} = \frac{-1}{a_{ii}} \left(\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^k - b_i \right), \quad i = 1, 2, \ldots, n.$$

Vidi se da se za izračunavanje x^{k+1} koriste komponente vektora x^k, pa se Jakobijev postupak naziva i **postupak zajedničkog koraka**.

Prirodno se nameće ideja da se pri izračunavanju $i-$te komponente vektora x^{k+1} koriste već izračunate komponente tog vektora x_j^{k+1}, $j = 1, 2, \ldots, i-1$, što daje iterativni postupak

$$x_i^{k+1} = \frac{-1}{a_{ii}} \left(\sum_{j=1}^{i-1} a_{ij} x_j^{k+1} + \sum_{j=i+1}^{n} a_{ij} x_j^k - b_i \right), \quad i = 1, 2, \ldots, n.$$

Ovako definisan postupak ima prednost u odnosu na Jakobijev postupak jer se ne moraju memorisati obe aproksimacije, x^k i x^{k+1}. Matrični zapis iterativnog pravila dobija se iz

$$(D-T)\,x = Sx + b.$$

Kako su dijagonalni elementi matrice A različiti od nule, matrica $D-T$ je regularna, pa se iz prethodne jednačine dobija

$$x = (D-T)^{-1} Sx + (D-T)^{-1} b.$$

Na osnovu ovoga imamo iterativni postupak

$$x^{k+1} = B_G x^k + f, \quad k = 0, 1, \ldots,$$

gde je

$$B_G = (D-T)^{-1} S, \qquad f = (D-T)^{-1} b.$$

Postupak se naziva **Gaus-Zajdelov iterativni postupak** ili **postupak pojedinačnog koraka**.

Jakobijev i Gaus-Zajdelov postupak su najstariji iterativni postupci za rešavanje sistema linearnih jednačina. Njihovim uopštavanjem dobijaju se relaksacioni postupci, kod kojih se pomoću relaksacionog parametra utiče na brzinu konvergencije. Poznato je da su oba postupka konvergentna za neke klase matrica koje se često javljaju u rešavanju matematičkih modela, a u nekim od tih slučajeva se zna i odnos njihove brzine konvergencije, no generalno se ne može reći da je jedan postupak brži od drugog. Ovde je dat dokaz teorema o konvergenciji oba postupka za **strogo dijagonalno dominantne** matrice.

Definicija 6.10. ***Strogo dijagonalno dominantna matrica.*** *Matrica $A = [a_{ij}] \in \mathbb{R}^{n,n}$ je strogo dijagonalno dominantna po vrstama ako je*

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, 2, \ldots, n,$$

a strogo dijagonalno dominantna po kolonama ako je

$$|a_{jj}| > \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}|, \quad j = 1, 2, \ldots, n.$$

Najčešće se pod strogo dijagonalno dominantnim (SDD) matricama podrazumevaju SDD matrice po vrstama, pa se posebno naglašava ako se radi o SDD matricama po kolonama. Očigledno, matrice koje su strogo dijagonalne ili po vrstama ili po kolonoma imaju na glavnoj dijagonali elemente različite od nule.

Teorema 6.16. ***Konvergencija Jakobijevog postupka.*** *Neka je A strogo dijagonalno dominantna matrica po vrstama ili kolonama. Tada Jakobijev postupak konvergira ka jedinstvenom rešenju sistema $Ax = b$ za proizvoljno $x^0 \in \mathbb{R}^n$.*

Dokaz. Ako je A SDD po vrstama, onda je

$$\|B_J\|_\infty = \left\|D^{-1}(T+S)\right\|_\infty = \max_{1\le i\le n} \sum_{\substack{j=1\\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1,$$

pa je Jakobijev postupak konvergentan za svako $x^0 \in \mathbb{R}^n$. U slučaju da je A SDD po kolonama važi

$$\left\|(T+S)\,D^{-1}\right\|_1 = \max_{1\le j\le n} \sum_{\substack{i=1\\ i\neq j}}^{n} \left|\frac{a_{ij}}{a_{jj}}\right| < 1.$$

Neka je $x^k = D^{-1}y^k$. Tada je Jakobijev postupak dat sa

$$D^{-1}y^k = B_J D^{-1}y^{k-1} + D^{-1}b, \quad k = 0, 1, \ldots$$

odnosno

$$y^k = (T+S)D^{-1}y^{k-1} + b, \quad k = 0, 1, \ldots.$$

Kako je

$$\left\|(T+S)\,D^{-1}\right\|_1 < 1,$$

niz $y^0, y^1, \ldots$ konvergira ka rešenju y sistema

$$y = (T+S)\,D^{-1}y + b,$$

pa i niz $x^0, x^1, \ldots$ konvergira ka vektoru $x = D^{-1}y$, koji je rešenje sistema

$$x = D^{-1}\,(T+S)\,x + D^{-1}b.$$

■

U dokazu konvergencije Gaus-Zajdelovog postupka koristi se sledeća lema.

Lema 6.17. *Neka je A strogo dijagonalno dominantna matrica po vrstama ili kolonama. Tada je A regularna matrica.*

Dokaz. Neka je $A = D - T - S$ standardno razlaganja matrice A. Kako je, u slučaju stroge dijagonalne dominacije po vrstama,

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \geq 0, \quad i = 1, 2, \ldots, n,$$

tj. svi elementi a_{ii} su različiti od nule, matrica D je regularna i važi

$$\left\|D^{-1}\,(T+S)\right\|_\infty < 1.$$

Tada je i matrica $E - D^{-1}\,(T+S)$ regularna, a kako je

$$E - D^{-1}\,(T+S) = D^{-1}\,(D - T - S) = D^{-1}A,$$

sledi da je A regularna matrica. ■

Ako je A SDD po kolonama, onda je

$$\left\|(T+S)\,D^{-1}\right\|_1 < 1,$$

a odatle sledi da su matrice D i

$$E-(T+S)\,D^{-1} = (D-T-S)\,D^{-1} = AD^{-1}$$

regularne, pa je A regularna matrica.

Teorema 6.18. ***Konvergencija Gaus-Zajdelovog postupka.*** *Neka je A strogo dijagonalno dominantna matrica po vrstama. Tada Gaus-Zajdelov postupak konvergira ka rešenju sistema $Ax = b$ za svako $x^0 \in \mathbb{R}^n$.*

Dokaz. Po definiciji prirodnih matričnih normi postoji vektor $u \in \mathbb{R}^n$ sa osobinama

$$\left\|(D-T)^{-1}\,S\right\|_\infty = \left\|(D-T)^{-1}\,Su\right\|_\infty \qquad \text{i} \qquad \|u\|_\infty = 1.$$

Neka je

$$z = (D-T)^{-1}\,Su.$$

Tada je $\|z\|_\infty = \left\|(D-T)^{-1}\,S\right\|_\infty$ i

$$(D-T)\,z = Su,$$

odnosno

$$z_i = \frac{1}{a_{ii}}\left(-a_{i1}z_1 - a_{i2}z_2 - \cdots - a_{i,i-1}z_{i-1} - a_{i,i+1}u_{i+1} - a_{i,i+2}u_{i+2} - \cdots - a_{in}u_n\right).$$

Dokazaćemo indukcijom da je za svako $i = 1, 2, \ldots, n$

$$|z_i| \le q,$$

gde je

$$q = \max_{1\le i\le n}\sum_{\substack{j=1\\ j\ne i}}^{n}\left|\frac{a_{ij}}{a_{ii}}\right| < 1.$$

Za $i = 1$ imamo

$$|z_1| = \left|\frac{1}{a_{ii}}\left(a_{i2}u_2 + \cdots + a_{in}u_n\right)\right| \le \left|\frac{1}{|a_{ii}|}\left(|a_{i2}| + |a_{i3}| + \cdots + |a_{in}|\right)\right|,$$

jer je $|u_i| \le 1$ ya svako $i = 1, 2, \ldots, n$. Zbog toga što je matrica strogo dijagonalno dominantna po vrstama sledi $|z_1| \le q < 1$. Neka važi

$$|z_j| \le q < 1 \qquad \text{za} \qquad j = 1, 2, \ldots, i-1.$$

Tada sledi

$$|z_i| \le \frac{1}{|a_{ii}|}\left(|a_{i1}|\,|z_1| + |a_{i2}|\,|z_2| + \cdots + |a_{i,i-1}|\,|z_{i-1}| + |a_{i,i+1}|\,|u_{i+1}| + |a_{i,i+2}|\,|u_{i+2}| + \cdots + |a_{in}|\,|u_n|\right)$$

i

$$|z_i| \le \frac{1}{|a_{ii}|}\left(|a_{i1}| + |a_{i2}| + \cdots + |a_{in}|\right) \le q < 1.$$

Dakle dokazali smo da je $\|z\|_\infty \le q < 1$, tj. $\left\|(D-T)^{-1}\,S\right\|_\infty \le q < 1$. ■

Primer 6.12. *Neka je dat sistem linearnih jednačina $Ax = b$ sa matricom*

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

Treba ispitati konvergenciju Jakobijevog i Gaus-Zajdelovog postupka. Matrica Jakobijevog postupka je

$$B_J = D^{-1}(T+S) = T+S = \begin{bmatrix} 0 & 2 & -2 \\ 1 & 0 & 1 \\ 2 & 2 & 0 \end{bmatrix}.$$

Njen karakteristični polinom matrice B_J je

$$\det(\lambda E - B_J) = \lambda^3,$$

pa je $\rho(B_J) = 0$ i Jakobijev postupak je konvergentan. Matrica Gaus-Zajdelovog postupka je

$$B_G = (D-T)^{-1} S = \begin{bmatrix} 0 & 2 & -2 \\ 0 & 2 & -1 \\ 0 & 8 & -6 \end{bmatrix},$$

pa je njen karakteristični polinom

$$\det(\lambda E - B_G) = \lambda\left(\lambda^2 + 4\lambda - 4\right)$$

sa korenima

$$\lambda_1 = 0, \qquad \lambda_{2,3} = -2 \pm 2\sqrt{2}$$

i spektralnim radijusom $\rho(B_G) = |\lambda_2| > 1$, tj. Gaus-Zajdelov postupak je divergentan.

Primer 6.13. *Neka je*

$$A = \begin{bmatrix} 1 & 0.5 & -0.5 \\ -1 & 1 & -1 \\ -0.5 & 0.5 & 1 \end{bmatrix}.$$

Kako je

$$B_J = T + S = \begin{bmatrix} 0 & -0.5 & 0.5 \\ 1 & 0 & 1 \\ 0.5 & 0.5 & 0 \end{bmatrix}$$

i

$$\det(\lambda E - B_J) = (\lambda + 0.5)\left(\lambda^2 - 0.5\lambda + 1\right),$$

to su karakteristični koreni matrice B_J

$$\lambda_1 = 0.5, \qquad i \qquad \lambda_{2,3} = 0.25\left(1 \pm \sqrt{-15}\right),$$

pa je $\rho(B_J) = 1$, odakle sledi da je Jakobijev postupak za rešavanje sistema $Ax = b$ divergentan. U ovom slučaju je

$$B_G = \begin{bmatrix} 0 & -0.5 & 0.5 \\ 0 & -0.5 & 1.5 \\ 0 & 0 & -1.5 \end{bmatrix} \qquad i \qquad \det(\lambda E - B_G) = \lambda(\lambda + 0.5)^2,$$

pa je $\rho(B_G) = 0.5$, odnosno Gaus-Zajdelov postupak je konvergentan.

Iz prethodna dva primera vidi se da se ne može generalno govoriti o tome koji je postupak brži. Takva tvrđenja se mogu dati samo za specijalne klase matrica, na primer za tridijagonalne matrice. Naime u ovom slučaju ili oba postupka konvergiraju ili divergiraju, a u slučaju konvergencije Gaus-Zajdelov postupak je brži. Brzina konvergencije nekog iterativnog postupka za rešavanje sistema linearnih jednačina meri se spektralnim radijusom matrice koraka. Ako su matrice koraka dva iterativna postupka M_1 i M_2 a za njihove spektralne radijuse važi $\rho(M_1) < \rho(M_2)$, onda kažemo da je prvi iterativni postupak (sa matricom koraka M_1) brži od drugog iterativnog postupka.

6.6 Zadaci

6.1. *Neka je dat sistem* $Ax = b$, *gde je*

$$A = \begin{bmatrix} 5 & -1 & -1 & -0.25 \\ -1 & 5 & -0.25 & -1 \\ -1 & -0.25 & 5 & -1 \\ -0.25 & -1 & -1 & 5 \end{bmatrix}, \quad b = \begin{bmatrix} 0.25 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

a) *Pokazati da za dati sistem konvergira i Jakobijev i Gaus–Zajdelov iterativni postupak.*

b) *Sa dve iteracije Jokobijevim, odnosno Gaus–Zajdelovim postupkom odrediti približna rešenja datog sistema, uzimajući* $x^0 = [1, 1, 1, 1]^\top$.

c) *Oceniti greške približnih rešenja dobijenih pod b).*

d) *Koliko je potrebno iteracija Jakobijevim a koliko Gaus–Zajdelovim postupkom da bi se dobilo približno rešenje datog sistema sa greškom manjom od* 10^{-6}, *ako nema grešaka zaokrugljivanja?*

6.2. *Dat je sistem linearnih jednačina*

$$\begin{aligned} 11x_1 &+ 2x_2 &+ x_3 &= 15, \\ x_1 &+ 10x_2 &+ 2x_3 &= 16, \\ 2x_1 &+ 3x_2 &- 8x_3 &= 1. \end{aligned}$$

a) Dokaži da za dati sistem konvergira i Jakobijev i Gaus–Zajdelov iterativni postupak.

b) Izračunaj sa četiri iteracije približno rešenje datog sistema Gaus–Zajdelovim postupkom uzimajući $x^0 = [0, 0, 0]^\top$.

c) Oceni grešku približnog rešenja dobijenog pod b).

Sva računanja izvodi sa 3 cifre iza decimalne tačke.

6.3. *Neka je dat sistem* $Ax = b$, *gde je*

$$A = \begin{bmatrix} 5 & -1 & -1 & -0.25 \\ -1 & 5 & -0.25 & -1 \\ -1 & -0.25 & 5 & -1 \\ -0.25 & -1 & -1 & 5 \end{bmatrix}, \quad b = \begin{bmatrix} 0.25 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

a) Pokaži da za dati sistem konvergira i Jakobijev i Gaus–Zajdelov iterativni postupak.

b) Sa dve iteracije Jokobijevim, odnosno Gaus–Zajdelovim postupkom odredi približna rešenja datog sistema uzimajući $x^0 = [1, 1, 1, 1]^\top$.

c) Oceni greške približnih rešenja dobijenih pod b).

d) Koliko je potrebno iteracija Jakobijevim, a koliko Gaus–Zajdelovim postupkom da bi se dobilo približno rešenje datog sistema sa greškom manjom od 10^{-6}, ako nema grešaka zaokrugljivanja?

6.4. *Reši sistem*

$$\begin{aligned} x_1 &+ 2x_2 &- 12x_3 &+ 8x_4 &= 27, \\ 5x_1 &+ 4x_2 &+ 7x_3 &- 2x_4 &= 4, \\ -3x_1 &+ 7x_2 &+ 9x_3 &+ 5x_4 &= 11, \\ 6x_1 &- 12x_2 &- 8x_3 &+ 3x_4 &= 49. \end{aligned}$$

6.5. *Reši sistem*

$$\begin{aligned} x_1 &+ 3x_2 - 2x_3 = 5,\\ 2x_1 &+ 4x_2 + 2x_3 = 7,\\ -x_1 &- 5x_2 + 8x_3 = 20. \end{aligned}$$

6.6. *Reši sistem*

$$\begin{aligned} x_1 &+ 3x_2 - 2x_3 = 5,\\ 2x_1 &+ 4x_2 + 2x_3 = 7,\\ -x_1 &- 5x_2 + 8x_3 = -8. \end{aligned}$$

6.7. *Gausovim postupkom eliminacije reši sisteme $Ax = b$ i $Ay = c$, gde je*

$$A = \begin{bmatrix} 4 & -9 & 2 \\ 2 & -4 & 6 \\ 1 & -1 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} 5 \\ 3 \\ 4 \end{bmatrix}, \quad c = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}.$$

6.8. *Izračunaj $\|x\|_p$, $p = 1, 2, \infty$ za $x = [-5, 3, 2, 0, 1]^T$.*

6.9. *Gausovim postupkom eliminacije reši sistem*

$$\begin{aligned} 3x_1 &+ x_2 = -3,\\ 0.9998x_1 &+ \tfrac{1}{3}x_2 = -0.9994, \end{aligned}$$

vršeći sva računanja tačno. Zatim umesto $\frac{1}{3}$ uzmi približnu vrednost 0.3333 i reši tako izmenjeni sistem. Da li je sistem dobro uslovljen? Odredi uslovni broj matrice sistema.

6.10. *Neka je*

$$C = \begin{bmatrix} 5 & -1 & 0.1 \\ -6.5 & 3.5 & 1.5 \\ 3 & -2 & -1 \end{bmatrix}$$

približna inverzna matrica za matricu

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 10 & 15 \\ -5 & -14 & -22 \end{bmatrix}$$

Oceni u normi $\|\cdot\|_\infty$ odstupanje C od A^{-1}.

6.11. *Sistem $x = Mx + d$ sa*

$$M = \begin{bmatrix} 0.6 & -0.2 & -0.1 \\ 0.1 & 0.2 & 0.3 \\ -0.1 & -0.2 & -0.2 \end{bmatrix}, \quad d = \begin{bmatrix} 0 \\ 17 \\ 0.3 \end{bmatrix}$$

reši iterativnim postupkom za $x^0 = [0, 0, 0]^T$.

6.12. *Nađi iteraciju x^3 Jakobijevog postupka sa $x^0 = [0, 0, 0]^T$ za sistem*

$$\begin{aligned} 10x_1 &+ x_2 + x_3 = 1,\\ x_1 &+ 10x_2 + x_3 = 2,\\ x_1 &+ x_2 - 8x_3 = 3. \end{aligned}$$

i oceni grešku.

6.13. *Primer iz prethodnog zadatka reši Gaus-Zajdelovim postupkom.*

6.14. *Gausovim postupkom eliminacije reši sistem*

$$\begin{aligned} 3x_1 &+ 2x_2 + x_3 = 5,\\ x_1 &+ x_2 - x_3 = 0,\\ 4x_1 &- x_2 + 5x_3 = 3. \end{aligned}$$

6.15. *Gausovim postupkom eliminacije reši sistem*

$$\begin{aligned} 36.47x_1 &+ 5.28x_2 + 6.34x_3 = 12.26,\\ 7.33x_1 &+ 28.74x_2 - 5.86x_3 = 15.15,\\ 4.63x_1 &+ 6.31x_2 + 26.17x_3 = 25.22, \end{aligned}$$

vršeći sva računanja sa 4 *cifre iza decimalne tačke.*

6.16. *Ispitaj konvergenciju iterativnog postupka* $x^{k+1} = Mx^k + d$, $k = 0, 1, \ldots$, $x^0 = d$ *sa*

$$M = \begin{bmatrix} 0.12 & 0.14 & -0.25 \\ -0.04 & -0.45 & -0.32 \\ 0.08 & -0.04 & 0.11 \end{bmatrix}, \quad d = \begin{bmatrix} 0.63 \\ -0.28 \\ -0.97 \end{bmatrix}.$$

Ako je x *rešenje sistema* $x = Mx + d$ *oceni* k *za koje važi* $\left\|x - x^k\right\| \leq \varepsilon$, *.gde je* $\varepsilon > 0$ *unapred zadato.*

6.17. *Izračunaj iteraciju* x^2 *iz prethodnog zadatka i oceni njenu grešku.*

6.18. *Dat je sistem linearnih jednačina* $Ax = b$, *gde je*

$$A = \begin{bmatrix} 2 & 0.3 & 0.5 \\ 0.1 & 3 & 0.4 \\ 0.1 & 0.1 & 4.8 \end{bmatrix}, \quad b = \begin{bmatrix} 4.1 \\ 7.3 \\ 14.7 \end{bmatrix}.$$

Dokaži da iterativni postupak

$$x^{k+1} = (E - \omega A)\, x^k + \omega b, \qquad k = 0, 1, \ldots,$$

konvergira za $\omega \in (0, 0.4)$ *i proizvoljno* x^0.

6.19. *Ispitaj konvergenciju Jakobijevog postupka za sistem linearnih jednačina sa matricom*

$$A = \begin{bmatrix} 1 & 2 & 0.1 \\ 2 & 5 & 0 \\ 0.1 & 0 & 1 \end{bmatrix}.$$

6.20. *Odredi iteraciju* x^3 *Gaus-Zajdelovog postupka sa* $x^0 = 0$ *za sistem*

$$\begin{aligned} 1.2x_1 &- 1.5x_2 + 7.2x_3 = 16.80,\\ 2.2x_1 &+ 5.5x_2 - 1.5x_3 = 10.55, \ .\\ 6.1x_1 &+ 2.2x_2 + 1.2x_3 = 16.55, \end{aligned}$$

i oceni grešku.

6.21. *Formiraj konvergentan iterativni postupak za sistem*

$$\begin{aligned} 1.02x_1 &- 0.05x_2 - 0.10x_3 = 0.795, \\ -0.11x_1 &+ 1.03x_2 - 0.05x_3 = 0.849, \\ -0.11x_1 &- 0.12x_2 + 1.04x_3 = 1.389, \end{aligned} \quad .$$

nađi x^3 *za* $x^0 = [0.80, 0.85, 1.40]^T$. *Oceni grešku.*

Glava 7

Sistemi nelinearnih jednačina

7.1 Uvod

Problem rešavanja sistema nelinearnih jednačina koji će se razmatrati u ovom poglavlju može se formulisati na sledeći način. Neka je $D \subseteq \mathbb{R}^n$ zatvoren skup, $D_0 \subset \mathbb{R}^n$ otvoren skup, a $F : D \to \mathbb{R}^n$ dato nelinearno preslikavanje. Potrebno je odrediti $x \in D$ tako da je $F(x) = 0$. Ako su komponente preslikavanja F označene sa $f_1, f_2, \ldots, f_n$, gde

$$f_i : D \subset \mathbb{R}^n \to \mathbb{R}, \quad i = 1, 2, \ldots, n,$$

a $x = [x_1, x_2, \ldots, x_n]^\top$, sistem $F(x) = 0$ se može zapisati u obliku

$$f_i(x_1, x_2, \ldots, x_n) = 0, \quad i = 1, 2, \ldots, n.$$

Bitno je primetiti da sistem nelinearnih jednačina ne mora imati rešenje ili može imati proizvoljno mnogo rešenja. Za razliku od sistema linearnih jednačina, gde se na osnovu vrednosti determinante matrice sistema i ranga proširene matrice može utvrditi broj rešenja sistema, kod sistema nelinearnih jednačina ne može se generalno formulisati tvrđenje o broju rešenja koje bi bilo primenljivo u praktičnom rešavanju. Ako preslikavanje F ima neke specijalne osobine, na primer ako je F kontraktivno preslikavanje, onda se može utvrditi egzistencija jedinstvene nepokretne tačke preslikavanja F. Takođe, ako je determinanta Jakobijana različita od nule na nekom skupu, uz još neke dodatne pretpostavke, sistem $F(x) = 0$ ima jedinstveno rešenje.

Složenost problema utvrđivanja broja rešenja sistema nelinearnih jednačina pokazuje sledeći primer preslikavanja $F : \mathbb{R}^2 \to \mathbb{R}^2$.

Primer 7.1. *Za različite vrednosti realnog parametra α sistem*

$$f_1(x_1, x_2) = x_1^2 + (x_2 - \alpha)^2 - 4, \qquad f_2(x_1, x_2) = \cos(2x_1) - x_2 - \alpha,$$

ima jedno ili više rešenja ili ih nema.

Utvrđivanje egzistencije rešenja sistema nelinearnih jednačina znatno prevazilazi okvire ove knjige, pa će jedna od standardnih pretpostavki u daljem radu biti postojanje rešenja posmatranog sistema jednačina.

Za razliku od sistema linearnih jednačina, direktni postupci za rešavanje sistema nelinearnih jednačina se mogu primeniti samo na veoma usku klasu sistema. Zbog toga će se ovde razmatrati **iterativni postupci** za rešavanje sistema nelinearnih jednačina. Iterativnim postupkom se, polazeći od nekog $x^0 \in D$, formira **iterativni niz** vektora x^0, $x^1, \ldots$, pri čemu je cilj da ovaj niz bude **konvergentan** i da je njegova **granica** rešenje posmatranog sistema nelinearnih jednačina.

Pri konstrukciji iterativnog pravila susreće se nekoliko problema. Prvo, iterativni niz mora biti dobro definisan. Na primer, ako se x^k izračunava pomoću $F\left(x^{k-1}\right)$, kao što je slučaj kod svih postupaka koji će ovde biti analizirani, potrebno je obezbediti da članovi iterativnog niza ostaju u domenu funkcije F. Ovo se obično postiže dobrim izborom početne aproksimacije x^0.

Drugi problem je konvergencija niza $x^0, x^1, \ldots$, a zatim se, u slučaju konvergentnog niza, postavlja pitanje da li je njegova granica rešenje posmatranog sistema. Razlikuju se **lokalna i globalna** tvrđenja o konvergenciji. Teoreme o lokalnoj konvergenciji sadrže pretpostavku da rešenje x^* postoji i definišu okolinu tačke x^* iz koje treba birati početnu aproksimaciju x^0. Uslovi pod kojima se rešenje može dobiti kao granica niza $x^0, x^1, \ldots$ sa proizvoljnim x^0 dati su u teoremama o globalnoj konvergenciji.

Treći problem je **brzina konvergencije** iterativnog postupka, koja se, kao i kod jedne jednačine i sistema linearnih jednačina, meri **redom konvergencije**. Pored toga **efikasnost** svakog iterativnog postupka zavisi i od njegove **računske složenosti**, odnosno broja aritmetičkih operacija potrebnih za izračunavanje jedne iteracije.

Ovde će se razmatrati nekoliko najpoznatijih iterativnih postupaka za rešavanje sistema nelinearnih jednačina. Prvu grupu čine postupci zasnovani na principu kontrakcije i teoremama o nepokretnoj tački. Njutnov postupak i njegove modifikacije čine drugu grupu postupaka koji će se ovde proučavati.

Brzina konvergencije iterativnog postupka meri se redom konvergencije. Red konvergencije niza se definiše na sledeći način.

Definicija 7.1. *Neka je* $x^0, x^1, \ldots, x^* \in \mathbb{R}^n$, $\lim_{k\to\infty} x^k = x^*$. *Kaže se da niz* $x^0, x^1, \ldots$ *konvergira ka* x^*

i) kvadratno ako za dovoljno veliko k *postoji* $K > 0$ *takvo da je*

$$\left\|x^{k+1} - x^*\right\| \leq K \left\|x^k - x^*\right\|^2,$$

ii) superlinearno sa redom $\alpha > 1$ *ako za dovoljno veliko* k *postoji* $K > 0$ *takvo da je*

$$\left\|x^{k+1} - x^*\right\| \leq K \left\|x^k - x^*\right\|^\alpha,$$

iii) superlinearno ako je

$$\lim_{k\to\infty} \frac{\left\|x^{k+1} - x^*\right\|}{\left\|x^k - x^*\right\|} = 0,$$

iv) linearno sa faktorom $\sigma \in (0,1)$ *ako je za dovoljno veliko* k

$$\left\|x^{k+1} - x^*\right\| \leq \sigma \left\|x^k - x^*\right\|.$$

Na osnovu prethodne definicije može se govoriti o redu lokalne konvergencije postupka.

Definicija 7.2. *Iterativni postupak za izračunavanje* x^* *je lokalno kvadratno (superlinearno, linearno) konvergentan ako iterativni niz, za početnu aproksimaciju* x^0 *dovoljno blizu* x^*, *konvergira kvadratno (superlinearno, linearno) ka* x^*.

Očigledno je superlinearno konvergentan niz ili iterativni postupak istovremeno i linearno konvergentan sa faktorom σ, za svako $\sigma > 0$. Takođe je kvadratno konvergentan niz ili postupak istovremeno i superlinearno konvergentan sa redom 2.

Kako efikasnost iterativnog postupka zavisi od njegove brzine i računske složenosti, može se zaključiti da superlinearno konvergentni postupci imaju prednost u odnosu na linearno konvergentne postupke ukoliko je složenost izračunavanja jedne iteracije u oba postupka približno jednaka. No, postoje primeri linearno konvergentnih postupaka kod kojih je izračunavanje jedne iteracije dovoljno jednostavno, pa su oni u nekim slučajevima pogodniji za primenu od bržih, ali računski složenijih superlinearno konvergentnih postupaka.

7.2 Opšti iterativni postupak

Mnogi nelinearni problemi se prirodno formulišu u obliku nepokretne tačke

$$x = T(x),$$

gde je $T : D \subset \mathbb{R}^n \to \mathbb{R}^n$ nelinearno preslikavanje. Ako je pak dat sistem nelinearnih jednačina $F(x) = 0$, onda je za svaku regularnu matricu $C \in \mathbb{R}^{n,n}$ rešenje tog sistema nepokretna tačka preslikavanja

$$T(x) = x - CF(x),$$

i obrnuto, tj. sistemi $F(x) = 0$ i $x = T(x)$ su ekvivalentni na skupu D. Kao i kod jedne jednačine i sistema linearnih jednačina, iterativni postupak za rešavanje sistema $x = T(x)$ se može definisati na sledeći način

$$x^0 \in D, \qquad x^{k+1} = T\left(x^k\right), \quad k = 0, 1, \ldots .$$

Ovaj postupak se naziva **postupak sukcesivnih aproksimacija** ili postupak **nepokretne tačke**, a po analogiji sa iterativnim postupcima za rešavanje sistema linearnih jednačina, koristi se i naziv postupak zajedničkog koraka ili **Jakobijev postupak**. Ukoliko su $t_1, t_2, \ldots, t_n$, gde je $t_i : D \subseteq \mathbb{R}^n \to \mathbb{R}$, komponente preslikavanja T, iterativno pravilo može se zapisati kao

$$\left[x_1^0, x_2^0, \ldots, x_n^0\right]^\top \in D, \quad x_i^{k+1} = t_i\left(x_1^k, x_2^k, \ldots, x_n^k\right), \quad i = 1, 2, \ldots, n,\ k = 0, 1, \ldots .$$

Ovde će se prvo razmotriti neki od uslova koji garantuju egzistenciju rešenja sistema $x = T(x)$, a zatim uslovi za konvergenciju Jakobijevog postupka.

Definicija 7.3. ***Kontraktivno preslikavanje.*** *Preslikavanje $T : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ je kontrakcija sa konstantom kontrakcije γ na skupu D ako postoji $\gamma \in [0, 1)$ takvo da je*

$$\|T(x) - T(y)\| \leq \gamma \|x - y\|, \quad x, y \in D.$$

Ako je sa $J(x)$ označen Jakobijan kontraktivnog preslikavanja T

$$J(x) = \begin{bmatrix} \frac{\partial t_1}{\partial x_1}(x) & \frac{\partial t_1}{\partial x_2}(x) & \ldots & \frac{\partial t_1}{\partial x_n}(x) \\ \frac{\partial t_2}{\partial x_1}(x) & \frac{\partial t_2}{\partial x_2}(x) & \ldots & \frac{\partial t_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial t_n}{\partial x_1}(x) & \frac{\partial t_n}{\partial x_2}(x) & \ldots & \frac{\partial t_n}{\partial x_n}(x) \end{bmatrix}$$

konstanta kontrakcije se može odrediti kao

$$\gamma = \max_{x \in D} \|J(x)\|.$$

Sledeća teorema daje uslove za egzistenciju jedinstvenog rešenja sistema $x = T(x)$ i konvergenciju postupka sukcesivnih aproksimacija.

Teorema 7.1. ***Teorema o nepokretnoj tački.*** *Neka je $T : D \to D$ kontraktivno preslikavanje sa konstantom kontrakcije γ. Tada postoji jedinstvena nepokretna tačka $x^* \in D$ preslikavanja T i niz sukcesivnih aproksimacija $x^0, x^1, \ldots$ konvergira linearno ka x^* sa faktorom γ za proizvoljno $x^0 \in D$.*

Dokaz. Neka je $x^0 \in D$ proizvoljna početna aproksimacija. Kako je $T(D) \subseteq D$ niz $x^0, x^1, \ldots$ definisan pravilom pripada skupu D. Za svako $i = 1, 2, \ldots$ je

$$\left\|x^{i+1} - x^i\right\| = \left\|T\left(x^i\right) - T\left(x^{i-1}\right)\right\| \leq \gamma \left\|x^i - x^{i-1}\right\| \leq \gamma^2 \left\|x^{i-1} - x^{i-2}\right\| \leq \cdots \leq \gamma^i \left\|x^1 - x^0\right\|,$$

a na osnovu toga sledi

$$\left\|x^k - x^0\right\| = \left\|\sum_{i=0}^{k-1}\left(x^{i+1} - x^i\right)\right\| \leq \sum_{i=0}^{k-1}\left\|x^{i+1} - x^i\right\| \leq \left\|x^1 - x^0\right\| \sum_{i=0}^{k-1}\gamma^i \leq \frac{\left\|x^1 - x^0\right\|}{1-\gamma}.$$

Za svako $m, k > 0$ važi

$$\begin{aligned}\left\|x^{m+k} - x^m\right\| &= \left\|T\left(x^{m+k-1}\right) - T\left(x^{m-1}\right)\right\| \leq \gamma \left\|x^{m+k-1} - x^{m-1}\right\| \\ &\leq \gamma^2 \left\|x^{m+k-2} - x^{m-2}\right\| \leq \cdots \leq \gamma^m \left\|x^k - x^0\right\| \leq \frac{\gamma^m}{1-\gamma}\left\|x^1 - x^0\right\|.\end{aligned}$$

Kako je $\gamma < 1$, to za svako $\varepsilon > 0$ i svako k, a za dovoljno veliko m, važi

$$\left\|x^{k+m} - x^m\right\| < \varepsilon$$

što znači da je niz $x^0, x^1, \ldots$ Košijev i prema tome konvergentan. Neka je $\lim\limits_{k\to\infty} x^k = x^* \in D$. Kako je T neprekidno preslikavanje, to je

$$x^* = \lim_{k\to\infty} x^{k+1} = \lim_{k\to\infty} T\left(x^k\right) = T\left(\lim_{k\to\infty} x^k\right) = T\left(x^*\right).$$

Neka su $x^* \neq y^*$ dve nepokretne tačke preslikavanja T u skupu D. Tada je

$$\left\|x^* - y^*\right\| = \left\|T\left(x^*\right) - T(y^*)\right\| \leq \gamma \left\|x^* - y^*\right\| < \left\|x^* - y^*\right\|,$$

što je kontradikcija, pa je x^* jedinstvena nepokretna tačka preslikavanja T. Očigledno je

$$\left\|x^* - x^{k+1}\right\| = \left\|T\left(x^*\right) - T\left(x^k\right)\right\| \leq \gamma \left\|x^* - x^k\right\|,$$

odnosno, niz $x^0, x^1, \ldots$ je linearno konvergentan sa faktorom γ. ■

Neposredna posledica prethodne teoreme su apriorna i aposteriorna ocena greške.

Posledica 7.2. *Neka su ispunjene pretpostavke prethodne teoreme. Tada je*

$$\left\|x^* - x^k\right\| \leq \frac{\gamma}{1-\gamma}\left\|x^k - x^{k-1}\right\| \leq \frac{\gamma^k}{1-\gamma}\left\|x^1 - x^0\right\|, \quad k = 1, 2, \ldots.$$

Dokaz. Za svako $m, k > 0$ je

$$\begin{aligned}\left\|x^{m+k} - x^k\right\| &\leq \sum_{i=0}^{k-1}\left\|x^{m+i+1} - x^{m+i}\right\| \leq \left\|x^{k+1} - x^k\right\| \sum_{i=0}^{k-1}\gamma^i \\ &\leq \frac{\gamma}{1-\gamma}\left\|x^k - x^{k-1}\right\| \leq \frac{\gamma^k}{1-\gamma}\left\|x^1 - x^0\right\|.\end{aligned}$$

Kako je $\lim\limits_{m\to\infty} x^{m+k} = x^*$ a norma je neprekidno preslikavanje, slede tražene ocene. ■

Kao i kod sistema linearnih jednačina, može se i za sistem nelinearnih jednačina definisati **Gaus-Zajdelov postupak** ili **postupak pojedinačnog koraka**. Za sistem $x = T(x)$ sa $\left[x_1^0, x_2^0, \ldots, x_n^0\right]^\top \in D$ ovaj postupak ima oblik

$$\begin{aligned}x_1^{k+1} &= t_1\left(x_1^k, x_2^k, \ldots, x_n^k\right), \\ x_i^{k+1} &= t_i\left(x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, x_{i+1}^k, \ldots, x_n^k\right), \quad i = 2, 3, \ldots, n,\end{aligned}$$

pri čemu je $x^0 \in D$ i $k = 0, 1, \ldots$. Ako se na skupu D definiše preslikavanje

$$H = (h_1, h_2, \ldots, h_n)$$

sa

$$h_j(x_1, x_2, \ldots, x_n) = z_j, \quad j = 1, 2, \ldots, n,$$

gde je

$$\begin{aligned} z_1 &= t_1(x_1, x_2, \ldots, x_n), \\ z_j &= t_j(z_1, z_2, \ldots, z_{j-1}, x_j, x_{j+1}, \ldots, x_n), \quad j = 2, 3, \ldots, n, \end{aligned}$$

i pretpostavi da za $x \in D$ važi

$$[z_1, z_2, \ldots, z_{j-1}, x_j, x_{j+1}, \ldots, x_n]^\top \in D, \quad j = 2, 3, \ldots, n,$$

onda se Gaus-Zajdelov postupak može zapisati u obliku

$$x^{k+1} = H\left(x^k\right), \quad k = 0, 1, \ldots.$$

Sistemi $x = T(x)$ i $x = H(x)$ su očigledno ekvivalentni na D.

Teorema 7.3. *Neka je*

$$c = [c_1, c_2, \ldots, c_n]^\top, \qquad \rho = [\rho_1, \rho_2, \ldots, \rho_n]^\top \in \mathbb{R}^n$$

i

$$D = \left\{x \in \mathbb{R}^n : |x_j - c_j| \leq \rho_j, \quad j = 1, 2, \ldots, n\right\}.$$

Ako je $T(D) \subseteq D$ i za neko $\gamma \in [0, 1)$ važi

$$\|T(x) - T(y)\|_\infty \leq \gamma \|x - y\|_\infty, \quad x, y \in D,$$

onda Gaus-Zajdelov postupak konvergira ka rešenju $x^* \in D$ sistema $x = T(x)$ za proizvoljno $x^0 \in D$ i važe ocene

$$\left\|x^* - x^k\right\|_\infty \leq \frac{\gamma}{1-\gamma} \left\|x^k - x^{k-1}\right\|_\infty \leq \frac{\gamma^k}{1-\gamma} \left\|x^1 - x^0\right\|_\infty, \quad k = 1, 2, \ldots.$$

Primena Jakobijevog i Gaus-Zajdelovog postupka pokazana je na sledećem primeru.

Primer 7.2. *Neka je preslikavanje $F : \mathbb{R}^2 \to \mathbb{R}^2$ dato sa*

$$\begin{aligned} f_1(x_1, x_2) &= -x_1 + x_2^2, \\ f_2(x_1, x_2) &= -x_2 + x_1^2. \end{aligned}$$

Sistem $F(x) = 0$ je ekvivalentan sa sistemom $T(x) = x$ gde je

$$\begin{aligned} t_1(x_1, x_2) &= x_2^2, \\ t_2(x_1, x_2) &= x_1^2. \end{aligned}$$

Neka je

$$D = \left\{ x \in \mathbb{R}^2 : |x_i| \leq 0.3, \quad i = 1, 2 \right\}.$$

Jakobijan preslikavanja T je matrica

$$J(x) = \begin{bmatrix} 0 & 2x_2 \\ 2x_1 & 0 \end{bmatrix},$$

pa je

$$\gamma = \max_{x \in D} \|J(x)\|_\infty = \max \left\{ \max_{x \in D} 2|x_2|, \max_{x \in D} 2|x_1| \right\} \leq 0.6.$$

Lako se vidi da za $x \in D$ važi

$$t_1(x) \in [0, 0.09], \qquad t_2(x) \in [0, 0.09],$$

pa je zadovoljen i uslov $T(D) \subseteq D$. Pošto su ispunjene sve pretpostavke teorema i sistem $T(x) = x$, odnosno $F(x) = 0$, ima jedinstveno rešenje u D. Polazeći od

$$x^0 = [0.3, 0.3]^\top$$

Jakobijevim i Gaus Zajdelovim postupkom dobijaju se nizovi dati u sledećoj tabeli.

	Jakobijev postupak		*Gaus-Zajdelov postupak*	
i	x_1^i	x_2^i	x_1^i	x_2^i
0	0.3	0.3	0.3	0.3
1	0.09	0.09	0.09	0.0081
2	0.0081	0.0081	$0.6561 \cdot 10^{-4}$	$0.430467 \cdot 10^{-8}$

7.3 Njutnov postupak

Osnovna ideja Njutnovog postupka je, kao i pri rešavanju jedne jednačine, linearizacija odnosno zamena originalnog preslikavanja linearnom aproksimacijom. Sukcesivnim rešavanjem linearnih aproksimacija generiše se iterativni niz. Neka je x^k aproksimacija rešenja x^* sistema $F(x) = 0$. Ako preslikavanje F ima prvi izvod

$$F'(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \dots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(x) & \frac{\partial f_n}{\partial x_2}(x) & \dots & \frac{\partial f_n}{\partial x_n}(x) \end{bmatrix}$$

u tački x^k, onda je na osnovu Tejlorove formule

$$0 = F(x^*) = F\left(x^k\right) + J\left(x^k\right)\left(x^* - x^k\right) + R\left(x^* - x^k\right),$$

pri čemu je

$$J\left(x^k\right) = F'\left(x^k\right)$$

i

$$\lim_{h \to 0} \frac{R(h)}{\|h\|} = 0.$$

Ako je x^k blizu rešenja x^*, onda se ostatak $R\left(x^* - x^k\right)$ u Tejlorovoj formuli može zanemariti i razlika $x^* - x^k$ se aproksimira rešenjem s^k sistema linearnih jednačina

$$J\left(x^k\right) s^k = -F\left(x^k\right). \tag{7.1}$$

Nova aproksimacija x^{k+1} se dobija kao

$$x^{k+1} = x^k + s^k, \tag{7.2}$$

odnosno

$$x^{k+1} = x^k - J\left(x^k\right)^{-1} F\left(x^k\right). \tag{7.3}$$

Iterativno pravilo, definiše Njutnov postupak za rešavanje sistema jednačina $F(x) = 0$.

Njutnov postupak je teorijski veoma atraktivan jer je kvadratno konvrgentan za dovoljno dobru početnu aproksimaciju x^0. Pri praktičnoj realizaciji Njutnovog postupka javlja se nekoliko problema. Svaki korak Njutnovog postupka zahteva rešavanje jednog sistema linearnih jednačina (inverzna matrica Jakobijana se ne izračunava eksplicitno), a to može predstavljati problem, pogotovo kod sistema velikih dimenzija kakvi su najčešće sistemi nastali diskretizacijom parcijalnih diferencijalnih jednačina. Šta više, u svakom koraku je potrebno izračunati n komponenti vektora $F\left(x^k\right)$ i n^2 komponenti matrice $J\left(x^k\right)$, što značajno povećava računsku složenost postupka. Kako je u većini slučajeva Njutnov postupak lokalno konvergentan, potrebno je odrediti i dovoljno dobru početnu aproksimaciju x^0.

Primer 7.3. *Sistem jednačina*

$$f_1(x_1, x_2) = 2(x_1 + x_2)^2 + (x_1 - x_2)^2 - 8, \qquad f_2(x_1, x_2) = 5x_1^2 + (x_2 - 3)^2 - 9,$$

rešava se Njutnovim postupkom. Neka je $x^0 = [2, 0]^\top$. *Matrica Jakobijana za ovaj sistem je*

$$J(x) = \begin{bmatrix} 6x_1 + 2x_2 & 2x_1 + 6x_2 \\ 10x_1 & 2(x_2 - 3) \end{bmatrix}.$$

Članovi iterativnog niza su dati u sledećoj tabeli. Tačno rešenje sistema je vektor $[1, 1]^\top$.

k	0	1	2	3	4
x_1^k	2.	1.31579	1.0371	1.00061	1.000000
x_2^k	0.	1.05263	0.996613	0.999862	1.000000

Uslovi pod kojima Njutnov postupak konvergira ako je početna aproksimacija x^0 u dovoljno blizu rešenja x^* sistema $F(x) = 0$ dati su u sledećoj teoremi.

Teorema 7.4. *Neka su zadovoljene standardne pretpostavke.*
1) *Sistem* $F(x) = 0$ *ima rešenje* $x^* \in D$;
2) *Preslikavanje* $J : D \to \mathbb{R}^{n,n}$ *je Lipšic neprekidno na* D *sa konstantom* γ, *tj. važi*

$$\|J(x) - J(y)\| \le \gamma \|x - y\|, \quad x, y \in D,$$

3) *Matrica* $J(x^*)$ *je regularna. Tada postoji* $\varepsilon > 0$ *takvo da za svako* $x^0 \in \mathcal{B}(x^*, \varepsilon)$, *gde je* $\mathcal{B}(x^*, \delta)$ *lopta sa centrom u* x^* *i poluprečnikom* δ

$$\mathcal{B}(x^*, \delta) = \{x \in \mathbb{R}^n : \|x - x^*\| \le \delta\},$$

niz $x^0, x^1, \ldots$ *definisan iterativnim pravilom*

$$x^{k+1} = x^k - J\left(x^k\right)^{-1} F\left(x^k\right), \quad k = 0, 1, \ldots$$

konvergira kvadratno ka rešenju x^* *sistema* $F(x) = 0$.

U prethodnoj teoremi se početna aproksimacija x^0 bira tako da važi $x^0 \in \mathcal{B}(x^*, \varepsilon)$ za neko $\varepsilon > 0$. Ovaj zahtev je na prvi pogled dosta restriktivan. Međutim, u mnogim slučajevima primene Njutnovog postupka početna aproksimacija se može odrediti tako da je vrlo blizu rešenja. Na primer, kod sistema nastalih diskretizacijom konturnih problema početna aproksimacija se može dobiti interpolacijom onog rešenja za isti problem koje je dobijeno sa manje finom mrežom diskretizacije, tj. sa mrežom koja ima manje čvorova. Ukoliko je pak dovoljno dobra početna aproksimacija nepoznata koriste se globalno konvergentni postupci u početnoj fazi rešavanja, a teorema opisuje ponašanje iterativnog niza u blizini rešenja.

7.4 Zadaci

7.1. *Grafički lokalizuj rešenja sistema*

$$\begin{aligned} x_1^2 + x_2^2 &= 1, \\ x_1^3 - x_2 &= 0. \end{aligned}$$

7.2. *Grafički lokalizuj rešenja sistema*

$$\begin{aligned} x_1^3 + x_2^3 - 6x_1 + 3 &= 0, \\ x_1^3 - x_2^3 - 6x_2 + 2 &= 0. \end{aligned}$$

7.3. *Grafički lokalizuj rešenja sistema*

$$\begin{aligned} x_1^2 - x_2^2 &= 1, \\ (x_1 - 1)\, x_2 &= 1. \end{aligned}$$

7.4. *Grafički lokalizuj rešenja sistema*

$$\begin{aligned} 2x_1^2 + x_2^2 &= 1, \\ x_1^3 + 6x_1^2 x_2 &= 1. \end{aligned}$$

7.5. *Grafički lokalizuj rešenja sistema*

$$\begin{aligned} 1.5x_1^2 - 2.5x_2^2 &= 1, \\ \cos\left(0.4x_2 + x_1^2\right) + x_1^2 &= 1.6 - x_2^2. \end{aligned}$$

7.6. *Jakobijevim postupkom izračunaj sa sedam sigurnih cifara približno rešenje sistema*

$$\begin{aligned} x_1^3 + x_2^3 - 6x_1 + 3 &= 0, \\ x_1^3 - x_2^3 - 6x_2 + 2 &= 0, \end{aligned}$$

koje pripada skupu $D = \{(x_1, x_2) : 0.5 \le x_1 \le 0.8,\ 0 \le x_2 \le 0.5\}$.

7.7. *Naći sistem oblika $x = G(x)$ ekvivalentan sistemu*

$$\begin{aligned} x_1^2 + x_2^2 &= 1, \\ x_1^3 - x_2 &= 0, \end{aligned}$$

za $x \in D = \{(x_1, x_2) : 0.6 \leq x_1 \leq 0.9,\ 0.45 \leq x_2 \leq 0.75\}$, tako da važi $\|G'(x)\|_\infty \leq L < 1$, $x \in D$. Zatim, sa $x^0 = [0.75, 0.6]^\top$ izračunaj aproksimaciju rešenja posmatranog sistema sa osam sigurnih cifara.

7.8. *Gaus-Zajdelovim postupkom izračunaj sa sedam sigurnih cifara približno rešenje sistema*

$$\begin{aligned} x_1^3 + x_2^3 - 6x_1 + 3 &= 0, \\ x_1^3 - x_2^3 - 6x_2 + 2 &= 0, \end{aligned}$$

koje pripada skupu $D = \{(x_1, x_2) : 0.5 \leq x_1 \leq 0.8,\ 0 \leq x_2 \leq 0.5\}$.

7.9. *Pojednostavljenim Njutnovim postupkom reši sledeći sistem*

7.10. *Jednačina $z^3 - 1 = 0$ ima 3 kompleksna rešenja $z_1 = (1, 0)$, $z_2 = \left(-1/2, -\sqrt{3}/2\right)$, $z_3 = \left(-1/2, \sqrt{3}/2\right)$. Ova rešenja se mogu dobiti primenjujući Njutnov postupak ili pojednostavljeni Njutnov postupak za rešavanje sistema $F(x) = 0$, gde je*

$$\begin{aligned} f_1(x) &= x_1^3 - 3x_1x_2^2 - 1, \\ f_2(x) &= 3x_1^2x_2 - x_2^3. \end{aligned}$$

Izračunaj prvih 6 iteracija Njutnovog postupka i pojednostavljenog Njutnovog postupka uzimajući redom

a) $x^0 = (1, 1)$
b) $x^0 = (-0.5, -0.5)$
c) $x^0 = (-1, 0.5)$.

7.11. *Njutnovim postupkom reši sistem $F(u) = 0$, gde je $u = [x, y, z]^\top$ i*

$$\begin{aligned} f_1(u) &= x^2 + y^2 + z^2 - 1, \\ f_2(u) &= 2x^2 + y^2 - 4z, \\ f_3(u) &= 3x^2 - 4y + z^2. \end{aligned}$$

7.12. *Njutnovim postupkom reši sistem $F(x) = 0$, ako je*

$$\begin{aligned} x_0 &= 0, \quad x_n = 1, \\ f_i(x) &= x_i - \frac{1}{2}(x_{i-1} + x_{i+1}) + \frac{h}{4}\sqrt{4h^2 + (x_{i+1} - x_{i-1})^2}, \quad i = 1, 2, \ldots n-1, \end{aligned}$$

za $n = 5$ i $h = \dfrac{1}{n}$.

Glava 8

Numeričko diferenciranje

U ovoj glavi se posmatra izračunavanje numeričkih aproksimacija izvoda realne funkcije realne promenljive u zadatoj tački. Veoma je mnogo problema koji se svode na numeričko diferenciranje. Numeričke aproksimacije izvoda koriste se i pri numeričkom rešavanju početnih i konturnih problema običnih i parcijalnih diferencijalnih jednačina. Numeričko diferenciranje se koristi i kada je analitički izraz funkcije poznat, ali je veoma komplikovan, pa nije jednostavno odrediti odgovarajući izvod i kada je funkcija zadata pomoću tabele, odnosno njen analitički izraz je nepoznat, a tabela je formirana na osnovu nekih eksperimenata.

Formule za numeričko diferenciranje se izvode pod pretpostavkom da su poznate vrednosti funkcije u zadatim tačkama. Posmatra se aproksimacija $m-$tog izvoda funkcije f u tački x pomoću $n-$tačkastog **diferencnog količnika**

$$d_{m,n}(x) = h^{-m} \sum_{i=1}^{n} \alpha_i f(x + \beta_i h).$$

Sa zadatim prirodnim brojevima m i n i realnim brojevima x, h i β_i određuju se **koeficijenti** α_i diferencnog količnika $d_{m,n}(x)$. Ovi koeficijenti se uvek traže u skupu realnih brojeva.

U daljem radu se pretpostavlja da je $m = 1$ ili $m = 2$, da je $n = 2$ ili $n = 3$ da su realni brojevi $\beta_1, \beta_2, \ldots, \beta_n$ međusobno različiti i da je $h > 0$.

Ako je funkcija f poznata određuje se i greška numeričkog diferenciranja

$$R_{m,n}(x) = f^{(m)}(x) - d_{m,n}(x)$$

i njena ocena. Pri tome sa M_k označavamo konstantu za koju važi

$$M_k \geq \left| f^{(k)}(x) \right|, \qquad x \in [a, b].$$

8.1 Diferencni količnici za prvi izvod

Koristeći se Tejlorovim razvojem funkcije f u pogodno izabranim tačkama mogu se dobiti aproksimacije za $f'(x)$ i $f''(x)$ i odgovarajuće greške sa ocenama.

Pretpostavimo da je $f \in C^2[a, b]$. Tada za $x, x - h \in [a, b]$, $h > 0$, postoji $\tau \in (x - h, x)$ takvo da je

$$f(x - h) = f(x) - f'(x)\, h + \frac{f''(\tau)}{2} h^2.$$

Odatle je

$$f'(x) = \frac{f(x) - f(x - h)}{h} + \frac{f''(\tau)}{2} h.$$

Analogno, za neko $\theta \in (x, x+h)$ je

$$f(x+h) = f(x) + f'(x)\,h + \frac{f''(\theta)}{2}h^2,$$

a odatle se dobija

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{f''(\theta)}{2}h.$$

Za aproksimaciju $f'(x)$ može se uzeti **diferencni količnik unazad**

$$D_- f(x) = \frac{f(x) - f(x-h)}{h},$$

ili **diferencni količnik unapred**

$$D_+ f(x) = \frac{f(x+h) - f(x)}{h}.$$

Pri tome važi sledeća teorema.

Teorema 8.1. *Neka je* $f \in C^2[a,b]$. *Tada je za* $x, x \pm h \in [a,b]$, $h > 0$,

$$|f'(x) - D_\pm f(x)| \leq \frac{M_2}{2}h.$$

Iz Tejlorovih razvoja za $f(x+h)$ i $f(x-h)$ dobija se **centralni diferencni količnik za prvi izvod**

$$D_0 f(x) = \frac{f(x+h) - f(x-h)}{2h}$$

koji takođe aproksimira $f'(x)$.

Teorema 8.2. *Neka je* $f \in C^2[a,b]$. *Tada je za* $x, x \pm h \in [a,b]$, $h > 0$,

$$|f'(x) - D_0 f(x)| \leq \frac{M_2}{2}h.$$

Dokaz. Iz Tejlorovih razvoja za $f(x+h)$ i $f(x-h)$ sledi direktno

$$f(x+h) - f(x-h) = 2f'(x)\,h + \frac{h^2}{2}\left(f''(\theta) - f''(\tau)\right),$$

odnosno,

$$|f'(x) - D_0 f(x)| = \frac{h}{4}\left|f''(\theta) - f''(\tau)\right| \leq \frac{M_2}{2}h.$$

■

Ako se pretpostavi da je $f \in C^3[a,b]$ može se dobiti bolja ocena aproksimacije prvog izvoda centralnim diferencnim količnikom. Pri tome se koristi rezultat sledeće teoreme, koja je dokazana u glavi Numerička integracija.

Teorema 8.3. *Neka je* $f \in C[a,b]$ *i neka su* a_i, $i = 1, 2, \ldots, n$, *realni brojevi istog znaka. Ako* $x_i \in [a,b]$, $i = 1, 2, \ldots, n$, *onda postoji* $x \in [a,b]$ *takvo da je*

$$\sum_{i=1}^{n} a_i f(x_i) = f(x) \sum_{i=1}^{n} a_i.$$

Teorema 8.4. *Neka je* $f \in C^3[a,b]$. *Tada je za* $x, x \pm h \in [a,b]$, $h > 0$,

$$|f'(x) - D_0 f(x)| \leq \frac{M_3}{6}h^2.$$

Dokaz. Kako je za neko $\tau \in (x-h, x)$

$$f(x-h) = f(x) - f'(x)\,h + \frac{f''(x)}{2}h^2 - \frac{f'''(\tau)}{6}h^3,$$

i za neko $\theta \in (x, x+h)$

$$f(x+h) = f(x) + f'(x)\,h + \frac{f''(x)}{2}h^2 + \frac{f'''(\theta)}{6}h^3,$$

sledi

$$f(x+h) - f(x-h) = 2f'(x)\,h + \frac{h^3}{6}\left(f'''(\theta) + f'''(\tau)\right) = 2f'(x)\,h + \frac{h^3}{3}f'''(\gamma),$$

za neko $\gamma \in (x-h, x+h)$. Sada je

$$|f'(x) - D_0 f(x)| = \frac{h^2}{6}\,|f'''(\varepsilon)| \leq \frac{M_3}{6}h^2.$$

■

8.2 Diferencni količnici za drugi izvod

Na osnovu

$$f(x-h) = f(x) - f'(x)\,h + \frac{f''(x)}{2}h^2 - \frac{f'''(\tau)}{6}h^3,$$

za neko $\tau \in (x-h, x)$ i

$$f(x+h) = f(x) + f'(x)\,h + \frac{f''(x)}{2}h^2 + \frac{f'''(\theta)}{6}h^3,$$

za neko $\theta \in (x, x+h)$ sledi

$$f(x-h) - 2f(x) + f(x+h) = h^2 f''(x) + \frac{f'''(\theta)}{6}h^3 - \frac{f'''(\tau)}{6}h^3,$$

a odatle

$$f''(x) = D''f(x) - \frac{f'''(\theta)}{6}h + \frac{f'''(\tau)}{6}h,$$

gde je

$$D''f(x) = \frac{f(x-h) - 2f(x) + f(x+h)}{h^2}$$

centralni diferencni količnik za drugi izvod. Imajući to u vidu, lako se dokazuje sledeća teorema.

Teorema 8.5. *Neka je $f \in C^3[a,b]$. Tada je za svako $x, x \pm h \in [a,b]$, $h > 0$,*

$$|f''(x) - D''f(x)| \leq \frac{M_3}{3}h.$$

Sa $f \in C^4[a,b]$ može se dokazati i bolja ocena za aproksimaciju $f''(x)$ centralnim diferencnim količnikom za drugi izvod $D''f(x)$.

Teorema 8.6. *Ako je $f \in C^4[a,b]$, onda je za svako $x, x \pm h \in [a,b]$, $h > 0$,*

$$|f''(x) - D''f(x)| \leq \frac{M_4}{12}h^2.$$

Dokaz. Postoje $\tau \in (x-h, x)$ i $\theta \in (x, x+h)$ za koje važi

$$f(x-h) = f(x) - f'(x)\,h + \frac{f''(x)}{2}h^2 - \frac{f'''(x)}{6}h^3 + \frac{f^{IV}(\tau)}{24}h^4,$$

i

$$f(x+h) = f(x) + f'(x)\,h + \frac{f''(x)}{2}h^2 + \frac{f'''(x)}{6}h^3 + \frac{f^{IV}(\theta)}{24}h^4.$$

Sada je

$$f(x-h) - 2f(x) + f(x+h) = h^2 f''(x) + \frac{f^{IV}(\tau)}{24}h^4 + \frac{f^{IV}(\theta)}{24}h^4,$$

i za neko $\gamma \in (x-h, x+h)$

$$f''(x) - D''f(x) = -\frac{h^2}{24}\left(f^{IV}(\tau) + f^{IV}(\theta)\right) = -\frac{h^2}{12}f^{IV}(\gamma),$$

a odatle tvrđenje sledi direktno. ■

8.3 Zadaci

8.1. *Neka je $f \in C^2[a,b]$. Dokazati da za svako $x, x \pm h \in [a,b]$, $h > 0$, postoje*

$$\alpha, \beta \in (x, x+h) \quad i \quad \gamma, \delta \in (x-h, x)$$

takvi da važi:

a) $f'(x) - \dfrac{f(x+h) - f(x)}{h} = -\dfrac{f''(\alpha)}{2}h;$

b) $f'(x) - \dfrac{f(x) - f(x-h)}{h} = \dfrac{f''(\gamma)}{2}h;$

c) $f'(x) - \dfrac{f(x+h) - f(x-h)}{2h} = \dfrac{f''(\delta) - f''(\beta)}{2}h.$

8.2. *Neka je $f \in C^3[a,b]$. Dokazati da za svako $x, x \pm h \in [a,b]$, $h > 0$, postoje $\alpha \in (x-h, x+h)$, $\beta \in (x-h, x)$ i $\gamma \in (x, x+h)$ takvi da važi*

$$f'(x) - \frac{f(x+h) - f(x-h)}{2h} = -\frac{f''(\alpha)}{6}h^2,$$

$$f''(x) - \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} = \frac{f'''(\gamma) - f'''(\beta)}{6}h.$$

8.3. *Za funkciju f, čije su približne vrednosti f^* date u tabeli,*

x	0.8	0.9	1.0	1.1	1.2	1.3	1.4
$f^*(x)$	0.6967	0.6216	0.5403	0.4536	0.3624	0.2675	0.1700

,

važi $\left|f^{IV}(x)\right| \le 1$.

a) *Na osnovu date tabele izračunati tri približne vrednosti za $f''(1.1)$ koristeći se formulom*

$$f''(x) \approx \frac{f^*(x-h) - 2f^*(x) + f^*(x+h)}{h^2},$$

za $h = 0.1,\ 0.2,\ 0.3$. Sve cifre u tabeli su sigurne.

b) *Naći $h_{\min}$ za koje se dobija najbolja aproksimacija.*

Glava 9

Početni problemi

U ovoj glavi posmatramo numeričko rešavanje početnog problema

$$y' = f(x, y), \qquad y(a) = \alpha.$$

Na isti način se mogu rešavati i početni problemi oblika

$$\begin{aligned}
y_1'(x) &= f_1\left(x, y_1(x), y_2(x), \ldots, y_n(x)\right), && y_1(a) = \alpha_1,\\
y_2'(x) &= f_2\left(x, y_1(x), y_2(x), \ldots, y_n(x)\right), && y_2(a) = \alpha_2,\\
&\vdots\\
y_n'(x) &= f_n\left(x, y_1(x), y_2(x), \ldots, y_n(x)\right), && y_n(a) = \alpha_n,
\end{aligned}$$

pri čemu se traži n realnih funkcija y_i jedne realne promenljive koje zadovoljavaju dati sistem i početne uslove. Ovaj problem se može zapisati u obliku

$$Y' = F(x, Y), \qquad Y(a) = \alpha,$$

gde je

$$\begin{aligned}
Y(x) &= \begin{bmatrix} y_1(x)\\ y_2(x)\\ \vdots\\ y_n(x) \end{bmatrix}, \qquad Y'(x) = \begin{bmatrix} y_1'(x)\\ y_2'(x)\\ \vdots\\ y_n'(x) \end{bmatrix},\\
F\left(x, Y\right) &= \begin{bmatrix} f_1(x, y_1, y_2, \ldots, y_n)\\ f_2(x, y_1, y_2, \ldots, y_n)\\ \vdots\\ f_n(x, y_1, y_2, \ldots, y_n) \end{bmatrix}, \qquad \alpha = \begin{bmatrix} \alpha_1\\ \alpha_2\\ \vdots\\ \alpha_n \end{bmatrix}.
\end{aligned}$$

Na sličan način početni problem

$$y^{(m)} = f\left(x, y(x), y'(x), y''(x), \ldots, y^{(m-1)}(x)\right), \qquad y^{(k)}(a) = \alpha_k, \quad k = 0, 1, \ldots, m-1,$$

uvođenjem pomoćnih funkcija

$$\begin{aligned} z_1(x) &= y(x), \\ z_2(x) &= y'(x), \\ &\vdots \\ z_m(x) &= y^{(m-1)}(x), \end{aligned}$$

može se zapisati kao

$$Z' = F(x, Z), \qquad Z(a) = \alpha,$$

gde je

$$Z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_{m-1} \\ z_m \end{bmatrix}, \qquad Z' = \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_{m-1}' \\ z_m' \end{bmatrix},$$

$$F(x, Z) = \begin{bmatrix} z_2 \\ z_3 \\ \vdots \\ z_m \\ f(x, z_1, z_2, \ldots, z_m) \end{bmatrix}, \qquad \alpha = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{m-1} \end{bmatrix}.$$

Od brojnih postupaka koji daju približno rešenje ovde se posmatraju samo postupci zasnovani na **diskretizaciji.** Kod ovih postupaka se dobijaju približne vrednosti rešenja na skupu $\{x_0, x_1, \ldots, x_n\}$. Najčešće, ali ne i uvek, vrednosti x_i su ekvidistantne, tj.

$$x_i = x_0 + ih, \quad i = 0, 1, \ldots, n.$$

U tom slučaju veličina h se naziva **dužina koraka** ili jednostavno **korak diskretizacije.** Uopšteno govoreći, numerički postupak se sastoji u tome da se svakoj tački x_i dodeli vrednost y_i koja se zatim posmatra kao aproksimacija vrednosti $y(x_i)$ tačnog rešenja u tački x_i. Postupci za numeričko rešavanje početnih problema dele se na **jednokoračne** i **višekoračne.** Kod jednokoračnih postupaka za izračunavanje vrednosti y_{i+1} dovoljno je poznavanje vrednosti y_i. Višekoračni postupci za izračunavanje vrednosti y_{i+1} zahtevaju poznavanje određenog broja prethodno izračunatih vrednosti $y_i, y_{i-1}, \ldots$. Postupak je $k-$koračan ako zahteva k prethodnih vrednosti.

Kod jednokoračnih postupaka početna tačka nema specijalnu ulogu. Tačnije, svaka tačka skupa $\{x_0, x_1, \ldots, x_{n-1}\}$ može se posmatrati kao početna. To omogućava da se korak diskretizacije može menjati od tačke do tačke. Postupci koji su $k-$koračni postupci zahtevaju specijalno računanje u prvih k tačaka $x_0, x_1, \ldots, x_{k-1}$. Isti takav način računanja je potreban i pri promeni dužine koraka diskretizacije. Računanje sa višekoračnim postupcima je nešto složenije, ali zato i tačnije u odnosu na jednokoračne.

U daljem radu se pretpostavlja da postoji rešenje y početnog problema

$$y' = f(x, y), \qquad y(a) = \alpha,$$

da je

$$f \in C^k\left([a, b] \times \mathbb{R}\right),$$

dakle $y \in C^{k+1}\left[a, b\right]$.

9.1 Jednokoračni postupci

U ovom delu se pretpostavlja da posmatrani početni problem

$$y' = f(x, y), \qquad y(a) = \alpha,$$

ima jednoznačno određeno rešenje y iz klase $C^1(D)$, tj. da postoji rešenje y diferencijalne jednačine $y' = f(x, y)$ koje ima neprekidan prvi izvod i koje zadovoljava početni uslov $y(a) = \alpha$.

Sledi nekoliko osnovnih pojmova koji se koriste pri definisanju jednokoračnih postupaka.

Za izabrani korak diskretizacije

$$h \in (0, b - a],$$

odredi se prirodan broj n iz uslova

$$\frac{b-a}{h} - 1 < n \leq \frac{b-a}{h},$$

i definiše **mreža diskretizacije**

$$I_h = \{a + ih : i = 0, 1, \ldots, n\}.$$

Tačke $x \in I_h$ nazivaju se **čvorovi** mreže diskretizacije ili čvorovi diskretizacije i označavaju se i sa x_i. Naime, kada se kaže da je x čvor podrazumeva se da je $x = a + ih$ za neko $i \in \{0, 1, \ldots, n\}$. Očigledno, za čvorove diskretizacije važi

$$a = x_0 < x_1 < \cdots < x_n \leq b < x_n + h.$$

Neka je

$$I'_h = I_h \backslash \{x_n\} = \{a + ih : i = 0, 1, \ldots, n - 1\}.$$

Integracijom jednačine $y'(x) = f(x, y(x))$ od x do $x + h$, za $x \in I'_h$, i deljenjem dobijenog rezultata sa h dobija se

$$\frac{1}{h}(y(x + h) - y(x)) = \frac{1}{h}\int_x^{x+h} f(s, y(s))\, ds, \quad x \in I'_h.$$

Postupci za numeričko rešavanje početnog problema $y' = f(x, y)$, $y(a) = \alpha$, dobijaju se aproksimacijom određenog integrala u prethodnoj relaciji. Jednokoračni postupci se dobijaju aproksimacijom navedenog integrala nekom funkcijom $f_h(x, z)$ dve promenljive $(x, z) \in D \times \mathbb{R}$ tako da vrednosti rešenja $y(x)$ posmatranog početnog problema zadovoljavaju jednačine

$$y(a) = \alpha, \qquad \frac{1}{h}(y(x + h) - y(x)) = f_h(x, y(x)) + T_h(x), \quad x \in I'_h.$$

Sa

$$T_h(x) = \frac{1}{h}\int_x^{x+h} f(s, y(s))\, ds - f_h(x, y(x)), \quad x \in I'_h,$$

označavamo **grešku odsecanja,** odnosno **lokalnu greška odsecanja,** jer se menja u zavisnosti od tačke $x \in I'_h$. Očigledno, lokalna greška odsecanja se može zapisati i na sledeći način

$$T_h(x) = \frac{1}{h}\int_x^{x+h} y'(s)ds - f_h(x, y(x)) = \frac{1}{h}(y(x + h) - y(x)) - f_h(x, y(x)) \quad x \in I'_h.$$

Pri formiranju jednokoračnog postupka uvek se traži da lokalna greška odsecanja za svako $x \in I_h'$ teži nuli kada $h \to 0$. Imajući to u vidu mogu se kao aproksimacije za $y(x)$ uzeti vrednosti $u(x)$ definisane kao rešenja sistema jednačina

$$u(a) = \alpha_h, \qquad \frac{1}{h}\left(u(x+h) - u(x)\right) = f_h\left(x, u(x)\right), \quad x \in I_h',$$

gde je α_h aproksimacija za α za koju važi

$$\lim_{h\to 0} |\alpha - \alpha_h| = 0.$$

Prethodni sistem označavamo kao **diskretni početni problem** i očigledno ima jedinstveno rešenje

$$u(a) = \alpha_h, \qquad u(x+h) = u(x) + h f_h\left(x, u(x)\right), \quad x \in I_h'.$$

Za definisanje jednokoračnih postupaka i kasnije, za ispitivanje njihove konzistencije, potrebno je da rešenje y početnog problema $y' = f(x, y)$ ima neprekidne izvode višeg reda na intervalu $[a, b]$. Za to je dovoljno da je funkcija f neprekidna i da ima neprekidne parcijalne izvode po obe promenljive x, y u nekoj pogodno izabranoj okolini $G \subset [a, b] \times \mathbb{R}$ rešenja y početnog problema.

9.1.1 Ojler–Košijev postupak

Pretpostavimo da je $f \in C^1(G)$, tj. da je $y \in C^2[a, b]$. Ako za aproksimaciju integrala

$$\frac{1}{h}\int_x^{x+h} f\left(s, y(s)\right) ds = \frac{1}{h}\int_x^{x+h} y'(s) ds$$

koristimo formulu levih pravougaonika, dobijamo

$$\frac{1}{h}\int_x^{x+h} f\left(s, y(s)\right) ds = y'(x) + \frac{h}{2} y''(\tau), \quad \tau \in (x, x+h).$$

Kako je $y'(x) = f\left(x, y(x)\right)$, sa

$$f_h(x, z) = f(x, z)$$

dobijamo jednokoračni postupak

$$u(a) = \alpha_h, \qquad u(x+h) = u(x) + h f\left(x, u(x)\right), \quad x \in I_h',$$

koji se naziva **Ojler–Košijev.** Lokalna greška odsecanja ovog postupka je

$$T_h(x) = \frac{h}{2} y''(\tau), \quad \tau \in (x, x+h).$$

9.1.2 Poboljšani Ojlerov postupak

Pretpostavimo da je $f \in C^2(G)$, tj. $y \in C^3[a, b]$. Na osnovu formule srednjih pravougaonika dobijamo za neko $\tau \in (x, x+h)$

$$\frac{1}{h}\int_x^{x+h} f\left(s, y(s)\right) ds = y'\left(x + \frac{h}{2}\right) + \frac{1}{24} h^2 y'''(\tau).$$

U izrazu

$$y'\left(x + \frac{h}{2}\right) = f\left(x + \frac{h}{2}, y(x + \frac{h}{2})\right)$$

koristi se aproksimacija

$$y(x+\frac{h}{2}) \approx y(x)+\frac{h}{2}y'(x),$$

za koju važi

$$y(x+\frac{h}{2})-y(x)-\frac{h}{2}y'(x)=\frac{h^2}{8}y''(\theta)$$

za neko, $\theta \in \left(x, x+\frac{h}{2}\right)$. Na osnovu Tejlorove teoreme za neko $\sigma \in (\max\{s, s+\varepsilon\}, \min\{s, s+\varepsilon\})$ važi

$$f(t, s+\varepsilon) = f(t,s)+f_y(t,\sigma)\,\varepsilon.$$

Sada je

$$f\left(x+\frac{h}{2}, y(x+\frac{h}{2})\right) = f\left(x+\frac{h}{2}, y(x)+\frac{h}{2}f(x,y(x))\right)+\frac{h^2}{8}f_y(t,\sigma)\,y''(\theta).$$

Ako se definiše

$$f_h(x,z) = f\left(x+\frac{h}{2}, z+\frac{h}{2}f(x,z)\right)$$

dobija se **poboljšani Ojlerov postupak**

$$u(a)=\alpha_h, \qquad u(x+h)=u(x)+hf\left(x+\frac{h}{2}, u(x)+\frac{h}{2}f(x,u(x))\right), \quad x\in I'_h.$$

Lokalna greška odsecanja ovog postupka je

$$T_h(x)=\frac{1}{24}h^2y'''(\tau)+\frac{h^2}{8}f_y\left(x+\frac{h}{2},\sigma\right)y''(\theta).$$

9.1.3 Poboljšani Ojler–Košijev postupak

Pretpostavimo da je $f\in C^2(G)$, tj. $y\in C^3[a,b]$. Slično kao kod poboljšanog Ojlerovog postupka pomoću trapezne formule i Tejlorovih razvoja dobija se **poboljšani Ojler–Košijev postupak.** Prema trapeznoj formuli je za neko $\tau\in(x,x+h)$

$$\frac{1}{h}\int_x^{x+h} f(s,y(s))\,ds = \frac{1}{2}\left(y'(x)+y'(x+h)\right)-\frac{1}{12}h^2y'''(\tau).$$

Kako je za neko $\theta\in(x,x+h)$

$$y(x+h)=y(x)+y'(x)h+\frac{h^2}{2}y''(\theta)$$

postoji tačka $(x+h,\sigma)\in G$ takva da je

$$f(x+h,y(x+h)) = f(x+h, y(x)+hf(x,y(x)))+\frac{h^2}{2}f_y(x+h,\sigma)\,y''(\theta).$$

Zbog

$$y'(x)=f(x,y(x)) \qquad \text{i} \qquad y'(x+h)=f(x+h,y(x+h)),$$

sa

$$f_h(x,z)=\frac{1}{2}\left(f(x,z)+f(x+h,z+hf(x,z))\right),$$

dobija se poboljšani Ojler–Košijev postupak

$$u(a)=\alpha_h, \qquad u(x+h)=u(x)+\frac{h}{2}\left(f(x,u(x))+f(x+h,u(x)+hf(x,u(x)))\right), \quad x\in I'_h,$$

i odgovarajuća lokalna greška odsecanja

$$T_h(x)=-\frac{1}{12}h^2y'''(\tau)+\frac{h^2}{4}f_y(x+h,\sigma)\,y''(\theta).$$

9.1.4 Postupak Runge–Kuta

Postupak Runge–Kuta, koji se ovde posmatra, pripada jednoj široj klasi jednokoračnih postupaka visokog reda tačnosti. Često se cela ova klasa naziva imenom postupci Runge–Kuta. Postupci te klase razlikuju se uglavnom po redu tačnosti, odnosno redu konzistencije.

Pretpostavimo da je $f \in C^4(G)$, što ima za posledicu da je $y \in C^5[a,b]$. Prema Simpsonovoj kvadraturnoj formuli je za neko $\tau \in (x, x+h)$

$$\frac{1}{h}\int_x^{x+h} y'(s)ds = \frac{1}{6}\left(y'(x) + 4y'(x+\frac{h}{2}) + y'(x+h)\right) - \frac{h^4}{2880}y^{(5)}(\tau).$$

Za

$$f_h(x,s) = \frac{1}{6}\left(k_0 + 2k_1 + 2k_2 + k_3\right),$$

gde je

$$\begin{aligned} k_0 &= f(x,s), \qquad k_1 = f\left(x+\frac{h}{2}, s+\frac{h}{2}k_0\right), \\ k_2 &= f\left(x+\frac{h}{2}, s+\frac{h}{2}k_1\right), \qquad k_3 = f\left(x+h, s+hk_2\right), \end{aligned}$$

važi

$$f_h(x,y(x)) = \frac{1}{6}\left(y'(x) + 4y'(x+\frac{h}{2}) + y'(x+h)\right) + \mathcal{O}(h^4).$$

Na taj način, sa ovako definisanom funkcijom $f_h(x,s)$, dobija se postupak

$$u(a) = \alpha_h, \qquad u(x+h) = u(x) + hf_h(x,u(x)), \quad x \in I_h',$$

za čiju lokalnu grešku odsecanja važi $|T_h(x)| \leq Kh^4$, gde je K konstanta nezavisna od h.

9.2 Zadaci

9.1. *Napisati diferencijalnu jednačinu*

$$4y''' - 5xy'' + \cos y = f(x)$$

kao sistem diferencijalnih jednačina prvog reda.

9.2. *Neka su dati početni problemi*

$$\begin{aligned} y' &= y, \quad y(0) = 1, \\ y' &= -y, \quad y(0) = 1. \end{aligned}$$

Odrediti za oba početna problema tačna rešenja.

9.3. *Ojler-Košijevim postupkom izračunati približnu vrednost rešenja početnog problema*

$$y' = xy, \quad y(0) = 1,$$

na intervalu $[0,1]$ *sa* $h = 0.1$. *Uporediti dobijene vrednosti sa odgovarajućim tačnim vrednostima, znajući da je tačno rešenje posmatranog početnog problema* $y(x) = e^{x^2/2}$.

9.4. *Poboljšanim Ojlerovim, poboljšanim Ojler-Košijevim i postupkom Runge-Kuta izračunati približnu vrednost rešenja početnog problema iz prethodnog zadatka na intervalu* $[0,1]$ *sa* $h = 0.1$ *i uporediti dobijene vrednosti sa odgovarajućim tačnim vrednostima.*

9.5. *Dat je početni problem*

$$y' = y, \quad y(0) = 1.$$

Ojler-Košijevim postupkom naći aproksimaciju rešenja $y(x)$ *tog problema koristeći korak* $h = x$, *a zatim uraditi isto i postupkom Runge–Kuta. Uporediti dobijene rezultate sa Maklorenovim razvojem tačnog rešenja* $y(x)$.

9.6. *Poznato je tačno rešenje sistema*

$$y' = y - z + 1, \qquad y_1(0) = 0,$$

$$z' = y + 3z + e^{-x}, \quad y_2(0) = 1,$$

$$\begin{aligned} y(x) &= -\frac{3}{4} - \frac{1}{9}e^{-x} + \frac{31}{36}e^{2x} - \frac{11}{6}xe^{2x}, \\ z(x) &= \frac{1}{4} - \frac{2}{9}e^{-x} + \frac{35}{36}e^{2x} + \frac{11}{6}xe^{2x}. \end{aligned}$$

Ojler-Košijevim postupkom izračunaj približne vrednosti rešenja datog sistema na intervalu $[0, 1]$ *sa* $h = 0.1$. *Uporedi dobijene vrednosti sa odgovarajućim tačnim vrednostima.*

9.7. *Dat je početni problem*

$$y'' = 4y\left(2 + y^2\right), \quad y(0) = 2, \quad y'(0) = 1.$$

Napiši datu jednačinu kao sistem diferencijalnih jednačina prvog reda i rešavajući odgovarajući početni problem izračunaj aproksimaciju za $y(0.5)$ *pomoću Ojlerovog postupka sa* $h = 0.1$.

9.8. *Poboljšanim Ojler-Košijevim postupkom sa* $h = 0.25$ *izračunaj aproksimaciju za* $y(2)$, *gde je* $y(x)$ *rešenje početnog problema*

$$y' = 4x - 2y, \quad y(0) = 1.$$

Glava 10

Konturni problemi

Konturni problem običnih diferencijalnih jednačina se dobija kada se od rešenja diferencijalne jednačine zahteva da zadovoljava dodatne uslove u dve ili više tačaka. Za početne probleme je poznato da se zadavanjem uslova u jednoj tački mogu odrediti dovoljni uslovi za egzistenciju rešenja za veoma široku klasu diferencijalnih jednačina. Za konturne probleme je karakteristično da se u zavisnosti samo od konturnih uslova može desiti da jedna jednačina ima jedinstveno rešenje, ima više rešenja ili uopšte nema rešenja. U slučajevima kada se konturni uslovi zadaju samo u dve tačke, što je najčešće slučaj u primeni, može se razviti teorija za mnoge specijalne klase jednačina. Teorija koja daje egzistenciju i jedinstvenost ima izuzetnu ulogu i u analizi numeričkih postupaka. U ovom delu posmatraju se samo konturni problemi **prvog reda**.

U intervalu $[a, b]$ traži se rešenje y diferencijalne jednačine drugog reda

$$y''(x) = f(x, y, y')$$

koje zadovoljava konturne uslove

$$y(a) = \alpha, \qquad y(b) = \beta,$$

gde su α i β date konstante.

Posmatrani konturni problem je u opštem slučaju **nelinearan.** Ako je $\alpha = \beta = 0$, konturni uslovi su **homogeni.** Diferencijalna jednačina sa konturnim uslovima daje konturni problem **prvog reda**. Konturni uslovi se nazivaju **Dirihleovi uslovi.**

Početni problem se obično rešava tako (kada je to moguće) što se odredi opšte rešenje u kojem se javljaju slobodne integracione konstante koje se određuju iz početnih uslova. Na prvi pogled čini se da se takav postupak može ponoviti i kod konturnih problema. Međutim, to nije tako, jer konturni problem nije uvek rešiv. Nastaju mnoge teškoće koje su karakteristične samo za konturne probleme. Postupci za rešavanje konturnih problema razlikuju se znatno od postupaka za rešavanje početnih problema. To se jednako odnosi i na numeričke postupke. Karakteristične teškoće u rešavanju konturnih problema vide se i u sledećim primerima.

Primer 10.1. *Neka je dat konturni problem*

$$y'' + y = 0, \qquad y(0) = 1, \qquad y(2) = 0.$$

Opšte rešenje posmatrane diferencijalne jednačine je

$$y(x) = A \sin x + B \cos x.$$

Iz konturnih uslova dobijaju se jednačine za određivanje konstanti A i B :

$$1 = B, \qquad 0 = A \sin 2 + B \cos 2.$$

Iz ovih jednačina se dobija

$$A = -\operatorname{ctg} 2, \qquad B = 1,$$

tako da je rešenje konturnog problema

$$y(x) = -\operatorname{ctg} 2 \sin x + \cos x.$$

Ako se konturni uslovi promene u

$$y(0) = 1, \qquad y(\pi) = 0,$$

dobija se sledeći sistem za određivanje konstanti A i B :

$$1 = A \cdot 0 + B, \qquad 0 = A \cdot 0 - B.$$

Kako ovaj sistem nema rešenje, to ni odgovarajući konturni problem nema rešenja. Nova promena konturnih uslova

$$y(0) = 0, \qquad y(\pi) = 0,$$

daje jednačine

$$0 = A \cdot 0 + B, \qquad 0 = A \cdot 0 - B,$$

iz kojih se dobija $B = 0$, a A ostaje neodređeno. To znači da posmatrani konturni problem ima beskonačno mnogo rešenja oblika

$$y(x) = A \sin x.$$

Posebno su važni i interesantni **linearni konturni problemi**, tj. takvi problemi kod kojih je difrencijalna jednačina linearna. Linearna diferencijalna jednačina drugog reda se može zapisati u obliku

$$Ly = r(x), \quad x \in D,$$

gde je

$$Ly = -y'' + p(x)\, y' + q(x)\, y,$$

a funkcije $p(x)$, $q(x)$ i $r(x)$ su definisane na intervalu $[a, b]$. Diferencijalna jednačina je **homogena** ako je $r(x) = 0$, za $x \in [a, b]$, a **konturni problem je homogen** ako su i jednačina i konturni uslovi homogeni.

U sledećim paragrafima posmatra se numeričko rešavanje konturnih problema

$$y'' = f(x, y, y'), \quad x \in [a, b], \qquad y(a) = \alpha, \qquad y(b) = \beta,$$

pod pretpostavkama koje garantuju egzistenciju i jedinstvenost rešenja y posmatranog konturnog problema. Na **mreži diskretizacije**

$$I_h = \{x_i = a + ih : i = 0, 1, \ldots, n\}$$

koja je određena **korakom** diskretizacije

$$h = \frac{b-a}{n}, \quad n \in \mathbb{N},$$

formira se **diskretni analogon** za posmatrani konturni problem. U zavisnosti od toga da li se posmatra linearni ili nelinearni konturni problem, diskretni analogon je sistem linearnih, odnosno nelinearnih jednačina. U opštem slučaju diskretni analogon može se zapisati kao sistem

$$F_i(w) = 0, \quad i = 0, 1, \ldots, n,$$

pri čemu je $w = [w_0, w_1, \ldots, w_n]^\top \in \mathbb{R}^{n+1}$. Ovaj sistem se može zapisati kraće kao $F(w) = 0$, gde je preslikavanje F je definisano preko svojih komponenti $F_0, F_1, \ldots, F_n$. Rešenje diskretnog analogona označava se sa

$$w^* = [w_0^*, w_1^*, \ldots, w_n^*]^\top \in \mathbb{R}^{n+1},$$

a restrikcija tačnog rešenja y posmatranog konturnog problema na mreži diskretizacije I_h sa y_h. U daljem radu koriste se oznake

$$y_i = y(x_i), \quad i = 0, 1, \ldots, n,$$

tako da je

$$y_h = [y_0, y_1, \ldots, y_n]^\top = [\alpha, y_1, y_2, \ldots, y_{n-1}, \beta]^\top.$$

Rešenje w^* diskretnog analogona je numeričko rešenje posmatranog konturnog problema, odnosno w^* je aproksimacija za y_h, $w^* \approx y_h$, tj.

$$w_i^* \approx y(x_i), \quad i = 0, 1, \ldots, n.$$

Posebno je značajno utvrditi kada postoji jedinstveno rešenje w^* posmatranog sistema i šta se može reći za grešku $\|w^* - y_h\|$.

10.1 Linearni konturni problemi

Neka je dat linearni konturni problem drugog reda

$$-y'' + p(x)y' + q(x)y = r(x), \quad x \in [a, b], \qquad y(a) = \alpha, \qquad y(b) = \beta.$$

Njegov diskretni analogon se formira tako što se u svakoj unutrašnjoj tački mreže I_h izvodi funkcije y aproksimiraju diferencnim količnicima. Tako se dobija **diskretni analogon** problema posmatranog konturnog problema

$$w_0 = \alpha, \quad -D''w_i + p(x_i)D'w_i + q(x_i)w_i = r(x_i), \quad i = 1, 2, \ldots, n-1, \quad w_n = \beta,$$

gde je D'' diferencna šema za aproksimaciju drugog izvoda, a D' jedna od šema za aproksimaciju prvog izvoda (D_-, D_+ ili D_0).

Umesto $D''y(x_i)$ piše se kraće $D''y_i$ i analogno za diferencne šeme za aproksimaciju prvog izvoda. Kako je pretpostavljeno da postoji rešenje y konturnog problema i da je $y_i = y(x_i)$, dobija se

$$-D''y_i + p(x_i)D'y_i + q(x_i)y_i \approx r(x_i), \quad i = 1, 2, \ldots, n-1,$$

gde znak $\approx$ stoji zbog toga što su izvodi rešenja y u tačkama x_i, zamenjeni odgovarajućim aproksimacijama. Greška koja se pri tome javlja je

$$\tau_i[y] = -D''y_i + p(x_i)D'y_i + q(x_i)y_i - r(x_i), \quad i = 1, 2, \ldots, n-1,$$

i naziva se **lokalna greška konzistencije.** Vektor

$$\tau[y] = [0, \tau_1[y], \tau_2[y], \ldots, \tau_{n-1}[y], 0]^\top \in \mathbb{R}^{n+1}$$

se naziva **vektor greške konzistencije**.

U slučaju $p \equiv 0$ diskretni analogon može se zapisati u obliku

$$w_0 = \alpha, \quad -\frac{1}{h^2}w_{i-1} + \left(\frac{2}{h^2} + q(x_i)\right)w_i - \frac{1}{h^2}w_{i+1} = r(x_i), \quad i = 1, 2, \ldots, n-1, \quad w_n = \beta.$$

Ovo je tridijagonalni sistem linearnih jednačina koji se može zapisati u obliku $Aw = r$, gde je

$$A = \frac{1}{h^2}\begin{bmatrix} h^2 & 0 & & & & & & \\ -1 & 2+h^2q_1 & -1 & & & & & \\ & -1 & 2+h^2q_2 & -1 & & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & & -1 & 2+h^2q_{n-1} & -1 \\ & & & & & & 0 & h^2 \end{bmatrix}, \quad r = \begin{bmatrix} \alpha \\ r_1 \\ r_2 \\ \vdots \\ r_{n-1} \\ \beta \end{bmatrix}$$

i

$$r_i = r(x_i), \qquad q_i = q(x_i), \qquad i = 0, 1, \ldots, n.$$

Važi sledeća teorema.

Teorema 10.1. *Neka je $q, r \in C^k[a,b]$ i $q(x) \geq 0,\ x \in [a,b]$. Tada diskretni analogon konturnog problema*

$$-y'' + q(x)y = r(x), \quad x \in [a,b], \qquad y(a) = \alpha, \qquad y(b) = \beta.$$

ima jedinstveno rešenje w^* i važi

$$\|w^* - y_h\|_\infty = \begin{cases} \mathcal{O}(h), & k = 1, \\ \mathcal{O}(h^2), & k = 2. \end{cases}$$

Definicija 10.1. *Diskretni analogon $Aw = r$ je stabilan ako je A regularna matrica i ako važi*

$$\left\|A^{-1}\right\|_\infty \leq C_1,$$

gde je C_1 konstanta nezavisna od n.

Definicija 10.2. *Diskretni analogon $Aw = r$ je konzistentan sa konturnim problemom ako postoji pozitivan broj m takav da važi*

$$\|\tau[y]\|_\infty \leq C_2 h^m,$$

gde je C_2 konstanta nezavisna od n. Broj m je red konzistencije.

Definicija 10.3. *Neka postoji rešenje w^* diskretnog analogona $Aw = r$. Ako važi*

$$\|y_h - w^*\|_\infty \leq C_3 h^s,$$

gde je konstanta C_3 nezavisna od n, kaže se da w^* konvergira ka y_h. Broj s je red konvergencije.

Teorema 10.2. *Neka je diskretni analogon $Aw = r$ stabilan i konzistentan sa redom konzistencije m. Tada je m i red konvergencije i $C_3 = C_1C_2$.*

Dokaz. Matrica A je regularna, postoji jedinstveno rešenje w^* diskretnog analogona i važi

$$\tau[y] = Ay_h - r = Ay_h - Aw^* = A(y_h - w^*).$$

Odatle sledi

$$\|y_h - w^*\| = \left\|A^{-1}\tau[y]\right\| \leq \left\|A^{-1}\right\| \|\tau[y]\| \leq C_1C_2h^m.$$

■

Ova teorema daje princip dokazivanja konvergencije koji se može izraziti kratko na sledeći način

$$\textit{stabilnost} \quad + \quad \textit{konzistencija} \quad \Rightarrow \quad \textit{konvergencija}.$$

Neka je sada u

$$w_0 = \alpha, \quad -D''w_i + p(x_i)D'w_i + q(x_i)w_i = r(x_i), \quad i = 1, 2, \ldots, n-1, \quad w_n = \beta,$$

D' jednaka D_0. Tada dobijamo diskretni analogon

$$w_0 = \alpha,$$

$$a_i w_{i-1} + b_i w_i + c_i w_{i+1} = r(x_i), \qquad i = 1, 2, \ldots n-1,$$

$$w_n = \beta,$$

gde je

$$a_i = -\frac{1}{h^2} - \frac{p(x_i)}{2h}, \qquad b_i = \frac{2}{h^2} + q(x_i), \qquad c_i = -\frac{1}{h^2} + \frac{p(x_i)}{2h}.$$

Ovo je **diskretni analogon sa centralnom šemom** (za prvi izvod) i možemo ga zapisati kao sistem $Aw = r$ gde je

$$A = \begin{bmatrix} 1 & 0 & & & & & & \\ a_1 & b_1 & c_1 & & & & & \\ & a_2 & b_2 & c_2 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & & & 0 & 1 \end{bmatrix}.$$

Teorema 10.3. *Neka je dat konturni problem*

$$-y'' + p(x)y' + q(x)y = r(x), \quad x \in [a,b], \qquad y(a) = \alpha, \qquad y(b) = \beta,$$

sa $p, q, r \in C^2[a,b]$ *i* $q(x) \geq Q_* > 0$, $x \in [a,b]$. *Ako je*

$$|p(x)| \leq P^*, \quad x \in [a,b], \qquad i \qquad n \geq \frac{P^*}{2}(b-a),$$

tada diskretni analogon

$$w_0 = \alpha,$$

$$a_i w_{i-1} + b_i w_i + c_i w_{i+1} = r(x_i), \qquad i = 1, 2, \ldots n-1,$$

$$w_n = \beta,$$

ima jedinstveno rešenje w^* *i važi*

$$\|w^* - y_h\|_\infty = Ch^2,$$

gde je konstanta C *nezavisna od* n.

Teorema 10.4. *Neka je dat konturni problem*

$$-y'' + p(x)y' + q(x)y = r(x), \quad x \in [a,b], \qquad y(a) = \alpha, \qquad y(b) = \beta.$$

sa $p, q, r \in C^1[a,b]$ *i* $q(x) \geq Q_* > 0$, $x \in [a,b]$. *Tada diskretni analogon sa* D' *određenim tako da važi*

$$D'w_i = \begin{cases} D_- w_i, & p(x_i) > 0, \\ D_+ w_i, & p(x_i) < 0 \end{cases}$$

ima jedinstveno rešenje w^* *i važi*

$$\|w^* - y_h\|_\infty = \mathcal{C}h,$$

gde je konstanta C *nezavisna od* n.

Diskretni analogon iz prethodne teoreme naziva se **diskretni analogon sa jednostranim šemama** (za prvi izvod).

Iz prethodnih teorema se vidi da je diskretni analogon sa centralnom šemom bolji jer daje grešku proporcionalnu sa h^2, dok diskretni analogon sa jednostranim šemama ima grešku proporcionalnu sa h. Međutim, pretpostavke u teoremi koja daje bolju grešku su mnogo jače. Posebno pretpostavka

$$n \geq \frac{P^*}{2}(b-a)$$

može da bude nezgodna jer za veliko P^* ili veliko $b-a$ mora se koristiti i veliko n, a to znači da je odgovarajući sistem linearnih jednačina glomazan.

10.2 Nelinearni konturni problemi

Diferencne aproksimacije za izvode mogu se koristiti i za formiranje diskretnih analogona **nelinearnih konturnih problema** oblika

$$\mathcal{L}y(x) = -y'' + f(x, y, y') = 0, \quad x \in [a, b], \qquad y(a) = \alpha, \qquad y(b) = \beta.$$

Pretpostavlja se da posmatrani konturni problem ima rešenje i da funkcija $f(x, y, z)$ na skupu

$$S = \{(x, y, z) : x \in [a, b], \quad \text{i} \quad y, z \in \mathbb{R}\}$$

ima neprekidne parcijalne izvode koji zadovoljavaju uslove

$$\left|\frac{\partial f(x, y, z)}{\partial z}\right| \leq P^*, \qquad 0 < Q_* \leq \left|\frac{\partial f(x, y, z)}{\partial y}\right| \leq Q^*, \qquad (x, y, z) \in S,$$

za neke pozitivne konstante P^*, Q_* i Q^*.

Ako se u unutrašnjim tačkama x_i ekvidistantne mreže, vrednost $\mathcal{L}y(x_i)$ aproksimira pomoću

$$\mathcal{L}_h y(x_i) = -D''y_i + f(x_i, y_i, D_0 y_i),$$

diskretni analogon za

$$-y'' + f(x, y, y') = 0, \qquad y(a) = \alpha, \qquad y(b) = \beta,$$

se tada može zapisati kao sistem

$$F(w) = 0.$$

Komponente F_i preslikavanja F definisane su sa

$$F_0(w) = w_0 - \alpha, \qquad F_n(w) = w_n - \beta,$$

a za $i = 1, 2, \ldots, n-1$ sa

$$F_i(w) = \frac{-w_{i-1} + 2w_i - w_{i+1}}{h^2} + f\left(x_i, w_i, \frac{w_{i+1} - w_{i-1}}{2h}\right).$$

Potpuno analogno linearnom slučaju definiše se vektor greške konzistencije $\tau[y]$

$$\tau_i[y] = \begin{cases} 0, & i = 0, \\ \mathcal{L}_h y(x_i) - \mathcal{L}y(x_i), & i = 1, 2, \ldots, n-1, \\ 0, & i = n. \end{cases}$$

za koju važi

$$\tau_i[y] = -\frac{h^2}{12}\left[y^{IV}(\theta_i) - 2\frac{\partial f(x_i, y(x_i), y''(\eta_i))}{\partial z} y'''(\mu_i)\right],$$

za neke $\theta_i, \eta_i, \mu_i \in (x_{i-1}, x_{i+1}), \quad i = 1, 2, \ldots, n-1.$

Analogno linearnom slučaju definiše se i stabilnost nelinearnog diskretnog analogona.

Definicija 10.4. *Diskretni analogon $F(w) = 0$ je stabilan ako važi*

$$\|u - v\| \leq C_1 \|F(u) - F(v)\|, \quad u, v \in \mathbb{R}^{n+1},$$

gde je C_1 konstanta nezavisna od n.

Konzistencija i konvergencija se u nelinearnom slučaju definišu potpuno isto kao i u linearnom slučaju.

Teorema 10.5. *Neka je diskretni analogon $F(w) = 0$ stabilan, konzistentan sa redom konzistencije m i neka ima rešenje w^*. Tada je m i red konvergencije.*

Teorema 10.6. *Neka funkcija $f(x, y, z)$ ima neprekidne parcijalne izvode koji zadovoljavaju*

$$\left|\frac{\partial f(x,y,z)}{\partial z}\right| \leq P^*, \qquad 0 < Q_* \leq \left|\frac{\partial f(x,y,z)}{\partial y}\right| \leq Q^*, \qquad (x,y,z) \in S,$$

i neka važi

$$hP^* \leq 2.$$

Tada je diskretni analogon $F(w) = 0$ stabilan.

Kao posledicu prethodnih teorema imamo sledeću teoremu.

Teorema 10.7. *Neka su ispunjene pretpostavke prethodne teoreme. Tada za rešenje w^* sistema $F(w) = 0$ važi*

$$\|y_h - w\|_\infty \leq C_1 \|\tau[y]\|_\infty, \qquad sa \qquad C_1 = \max\left\{1, \frac{1}{Q^*}\right\}.$$

Ako je još $y \in C^4([a,b])$, onda važi

$$\|y_h - w\|_\infty \leq C_1 \frac{h^2}{12}(M_4 + 2P^* M_3),$$

pri čemu su konstante M_3 i M_4 određene tako da važi

$$\max\left\{\left|y^{(k)}(x)\right| \mid x \in [a,b]\right\} \leq M_k, \quad k = 3, 4.$$

Prethodna teorema daje ocenu greške numeričkog i tačnog rešenja konturnog problema pod pretpostavkom da rešenje w nelinearnog sistema $F(w) = 0$ postoji.

10.3 Zadaci

10.1. *Numerički rešiti konturni problem*

$$-y'' + (2 + x)y = e^x(x + 1), \quad y(0) = 1, \quad y(1) = e,$$

sa korakom $h = 0.1$.

10.2. *Numerički reši konturni problem*

$$-y'' + (2+x)\,y = e^x\,(x+1)\,, \quad y\,(0) = 1, \quad y\,(1) = e,$$

sa korakom $h = 0.1$ *i uporedi dobijene rezultate sa tačnim.*

10.3. *Numerički reši konturni problem*

$$-y'' + y = 0, \quad y\,(0) = 0, \quad y\,(1) = 1,$$

sa korakom $h = 1/16$ *i uporedi dobijene rezultate sa tačnim.*

10.4. *Numerički reši konturni problem*

$$-y'' + y = -x, \quad y\,(0) = 1, \quad y\,(1) = 2,$$

sa korakom $h = 0.1$ *i uporedi dobijene rezultate sa tačnim.*

10.5. *Numerički reši konturni problem*

$$-y'' + y' = 2e^x \sin x, \quad y\,(0) = 3 - e^{-\pi}, \quad y\,(\pi) = 0,$$

sa korakom $h = 1/16$ *i uporedi dobijene rezultate sa tačnim.*

10.6. *Numerički reši konturni problem*

$$\begin{aligned} -y'' + y' \sin x + y \cos x &= -4e^{2x} + \left(e^{2x} + x\right) \cos x + \sin x + 2e^{2x} \sin x, \\ y\,(0) &= 1, \quad y\,(1) = 1 + e^2 \end{aligned}$$

sa korakom $h = 1/16$ *i uporedi dobijene rezultate sa tačnim.*

10.7. *Sa korakom* $h = 0.1$ *numerički reši konturni problem*

$$-y'' - \frac{y'}{x} + \frac{y}{x^2} = -8x, \quad y\,(1) = 0, \quad y\,(2) = 0.$$

10.8. *Numerički reši konturni problem*

$$-y'' - \frac{y'}{x} = -1, \quad y\,(1) = 0, \quad y\,(2) = 0,$$

sa korakom $h = 1/16$ *i uporedi dobijene rezultate sa tačnim.*

10.9. *Numerički reši konturni problem*

$$-y'' + \frac{1}{x^2} - y^2 - \frac{y'}{x} + \frac{\ln^2 x}{4} = 0,$$

$$y\,(1) = 0, \quad y\,(1.4) = \frac{\ln^2 (1.4)}{2}$$

sa korakom $h = 0.05$ *i uporedi dobijene rezultate sa tačnim.*

10.10. *Numerički reši konturni problem*

$$-y'' + 2 + y^2 = 0, \quad y\,(0) = y\,(1) = 0,$$

sa korakom $h = 0.1$.

10.11. *Numerički reši konturni problem*

$$-y'' + 2\,(2+9x)\,y = 2\,(2+9x)\,e^x, \quad y\,(0) = 0, \quad y\,(1) = 1.$$

Glava 11

Dodatak A

11.1 Definicije, teoreme i oznake

Realna matrica $A \in \mathbb{R}^{m,n}$ definiše linearno preslikavanje iz $\mathbb{R}^m$ u $\mathbb{R}^n$, te će se ubuduće takvo preslikavanje i matrica označavati na isti način.

Za svako $A \in \mathbb{R}^{n,n}$ kompleksni broj λ se naziva **karakteristični koren** matrice A ako jednačina

$$Ax = \lambda x$$

ima nenula rešenje x koje se naziva **karakteristični vektor** matrice A. Karakteristični koreni su rešenja **karakteristične jednačine**

$$\det(\lambda E - A) = 0.$$

Skup svih karakterističnih korena matrice A se naziva **spektar matrice** i obeležava $\sigma(A)$. Karakteristični koren matrice maksimalnog modula naziva se **spektralni radijus** i obeležava $\rho(A)$,

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

Teorema 11.8. *Neka je $A \in \mathbb{R}^{n,n}$ L - matrica i neka postoji vektor $v \in \mathbb{R}^n$, takav da važi*

$$v_i > 0, \quad (Av)_i \geq c = const. > 0, \quad i = 1, 2, \ldots, n.$$

Tada je A M -matrica i važi

$$\left\|A^{-1}\right\|_\infty \leq \frac{1}{c} \|v\|_\infty.$$

U specijalnom slučaju, ako je $v_i = 1$, $i = 1, 2, \ldots, n$, dobija se

$$\left\|A^{-1}\right\|_\infty \leq \frac{1}{c}.$$

Teorema 11.9. *Ako za matricu $A = [a_{ij}] \in \mathbb{R}^{n,n}$ važi*

$$|a_{ii}| - \sum_{j=1}^{n} |a_{ij}| \geq \gamma > 0, \quad i = 1, 2, \ldots, n,$$

onda je matrica A regularna i

$$\left\|A^{-1}\right\|_\infty \leq \frac{1}{\gamma}.$$

Definicija 11.1. *Vektori $x, y \in \mathbb{R}^n$ su ortogonalni ako je $(x, y) = 0$.*

Definicija 11.2. *Maksimalan broj linearno nezavisnih vektora vrsta matrice naziva se rang matrice A i označava $r(A)$.*

Definicija 11.3. *Sistem linearnih jednačina je saglasan ako ima bar jedno rešenje. Ukoliko sistem nema ni jedno rešenje on je protivrečan. Saglasan sistem je određen ako ima jedno i samo jedno rešenje, a neodređen ako ima više rešenja.*

Teorema 11.10. ***Teorema Kroneker-Kapelija.*** *Sistem linearnih jednačina $Ax = b$, gde je $A \in \mathbb{R}^{n,m}$, $b \in \mathbb{R}^n$, $x \in \mathbb{R}^m$ je saglasan ako i samo ako je rang matrice A jednak rangu proširene matrice $[A|b]$ koja se dobija dodavanjem vektora b kao $(m+1)$-ve kolone matrici A. Ako je dati sistem saglasan, a zajednički rang matrice sistema A i proširene matrice jednak r, onda je sistem određen ako i samo ako je $r = m$, sistem je neodređen ako i samo ako je $r < m$ i tada se umesto nekih $m - r$ promenljivih mogu uzeti proizvoljne vrednosti, a preostale nepoznate su jednoznačno određene.*

Definicija 11.4. *Preslikavanje $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ koje zadovoljava:*

1) $\|x\| \geq 0, \quad x \in \mathbb{R}^n; \quad \|x\| = 0$ *samo ako je* $x = 0$;

2) $\|\alpha x\| = |\alpha| \|x\|, \quad x \in \mathbb{R}^n, \alpha \in \mathbb{R}$;

3) $\|x + y\| \leq \|x\| + \|y\|, \quad x, y \in \mathbb{R}^n$

naziva se vektorska norma.

Najčešće se koriste p-norme,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad 1 \leq p < \infty,$$

za $p = 1, 2$ i u graničnom slučaju $p = \infty$,

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \qquad \|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}, \qquad \|x\|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

Teorema 11.11. *Vektorska norma je neprekidna funkcija.*

Definicija 11.5. *Skalarni proizvod u $\mathbb{R}^n$ je preslikavanje $(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ koje zadovoljava:*

1) $(x, x) \geq 0, \quad x \in \mathbb{R}^n; \quad (x, x) = 0$ *samo ako je* $x = 0$,

2) $(x, y) = (y, x), \quad x, y \in \mathbb{R}^n$,

3) $(x + y, z) = (x, z) + (y, z), \quad x, y, z \in \mathbb{R}^n$,

4) $(\alpha x, y) = \alpha (x, y), \quad x, y \in \mathbb{R}^n, \alpha \in \mathbb{R}$.

Skalarni proizvod u $\mathbb{R}^n$ definiše normu $\|x\|^2 = (x, x)$. Vektorska norma $\|x\|_2$, koja se naziva Euklidska norma, definisana je skalarnim proizvodom $(x, y) = x^\top y$. Za skalarni proizvod važi Koši-Švarcova nejednakost

$$|(x, y)| \leq \|x\| \|y\|, \quad x, y \in \mathbb{R}^\kappa.$$

Teorema 11.12. ***Ekvivalencija normi.*** *Neka su $\|\cdot\|$ i $\|\cdot\|'$ dve vektorske norme na $\mathbb{R}^n$. Tada postoje konstante $c_2 \geq c_1 > 0$ takve da je*

$$c_1 \|x\| \leq \|x\|' \leq c_2 \|x\|, \quad x \in \mathbb{R}^n.$$

Definicija 11.6. *Neka je $A \in \mathbb{R}^{n,m}$, a $\|\cdot\|$ i $\|\cdot\|'$ norme definisane na $\mathbb{R}^n$ i $\mathbb{R}^m$ respektivno. Tada je matrična norma u $\mathbb{R}^{n,m}$ definisana sa*

$$\|A\| = \sup_{\|x\|=1} \|Ax\|'.$$

Teorema 11.13. *Matrična norma na $\mathbb{R}^{n,m}$ ima sledeće osobine:*

1) $\|A\| \geq 0, \quad A \in \mathbb{R}^{n,m}, \qquad \|A\| = 0$ *samo ako je* $A = 0$,

2) $\|\alpha A\| = |\alpha| \|A\|, \quad \alpha \in \mathbb{R}, A \in \mathbb{R}^{n,m}$,

3) $\|A + B\| \leq \|A\| + \|B\|$.

Teorema 11.14. *Za proizvoljne matrice $A, B \in \mathbb{R}^{n,n}$ je*

$$\|AB\| \leq \|A\| \|B\|.$$

Ako su $\|\cdot\|$ i $\|\cdot\|'$ iz definicije matrične norme ista p-norma na $\mathbb{R}^n$, onda se kaže da je matrična norma za $A \in \mathbb{R}^{n,n}$ **indukovana** vektorskom normom. Ovakve norme nazivaju se i **prirodne** matrične norme. Indukovane p-matrične norme su

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|, \qquad \|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|, \qquad \|A\|_2 = \sqrt{\rho(A^\top A)}.$$

Matrična norma je neprekidna funkcija i takođe važi teorema o ekvivalenciji normi.

Ako je $\|\cdot\|$ proizvoljna norma na $\mathbb{R}^n$, a $P \in \mathbb{R}^{n,n}$ regularna matrica, onda su preslikavanja

$$\|x\|' = \|Px\| \qquad \text{i} \qquad \|A\|' = \left\|PAP^{-1}\right\|$$

takođe norma na $\mathbb{R}^n$ i $\mathbb{R}^{n,n}$ respektivno.

Teorema 11.15. *Za svaku matricu $A \in \mathbb{R}^{n,n}$ i za svako $\varepsilon > 0$ postoji prirodna matrična norma $\|\cdot\|$ takva da je*

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

Teorema 11.16. *Neka je $A_1, A_2, \ldots$ konvergentan niz matrica iz $\mathbb{R}^{n,m}$ i*

$$\lim_{k\to\infty} A_k = A$$

Ako su $B \in \mathbb{R}^{m,k}$ i $C \in \mathbb{R}^{p,n}$ proizvoljne matrice onda je

$$\lim_{k\to\infty} A_k B = AB$$

i

$$\lim_{k\to\infty} CA_k = CA.$$

Teorema 11.17. *Neka je $A_1, A_2, \ldots$ niz matrica iz $\mathbb{R}^{n,m}$. Tada je*

$$\lim_{k\to\infty} A_k = A$$

ako i samo ako je za neku matričnu normu

$$\lim_{k\to\infty} \|A_k - A\| = 0.$$

Teorema 11.18. *Neka je $A \in \mathbb{R}^{n,n}$. Tada je $\rho(A) < 1$ ako i samo ako je*

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

Teorema 11.19. *Neka je $A \in \mathbb{R}^{n,n}$. Ako je za neku prirodnu matričnu normu $\|A\| < 1$ onda su matrice $(E - A)$ i $(E + A)$ regularne i*

$$\frac{1}{1+\|A\|} \leq \left\|(E \pm A)^{-1}\right\| \leq \frac{1}{1-\|A\|}.$$

Definicija 11.7. ***Vandermondova matrica.*** *Matrica*

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}$$

je Vandermondova matrica. Njena determinanta

$$V(x_0, x_1, \ldots, x_n) = \begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{vmatrix}$$

je Vandermondova determinanta.

Teorema 11.20. ***Vandermondova determinanta.*** *Vandermondova determinanta je*

$$\begin{aligned} V(x_0, x_1, \ldots, x_n) &= \prod_{i>j}(x_i - x_j) = \prod_{j=0}^{n-1}\left(\prod_{i=j+1}^{n}(x_i - x_j)\right) \\ &= (x_1 - x_0)(x_2 - x_0)(x_3 - x_0)\cdots(x_n - x_0)\cdot \\ &\quad (x_2 - x_1)(x_3 - x_1)\cdots(x_n - x_1)\cdot \\ &\quad (x_3 - x_2)\cdots(x_n - x_2)\cdot \\ &\quad \vdots \\ &\quad \cdot(x_n - x_{n-1}). \end{aligned}$$

Dokaz. Posmatrajmo funkciju

$$V(x) = V(x_0, x_1, \ldots, x_{n-1}, x) = \begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^n \\ 1 & x & x^2 & \cdots & x^n \end{vmatrix}. \tag{11.1}$$

V(x) je očigledno polinom reda n. Njegove nule su $x_0, x_1, \ldots, x_{n-1}$. Otuda je

$$V(x_0, x_1, \ldots, x_{n-1}, x) = A(x - x_0)(x - x_1)\cdots(x - x_{n-1}),$$

gde A zavisi samo od $x_0, x_1, \ldots, x_{n-1}$. Da bi odredili A, razvijmo determinantu iz 11.1 po minorima njene poslednje kolone. Tada vidimo da je koeficijent uz x^n detrminanta $V(x_0, x_1, \ldots, x_{n-1})$. Otuda sledi

$$V(x_0, x_1, \ldots, x_{n-1}, x) = V(x_0, x_1, \ldots, x_{n-1})(x - x_0)(x - x_1) \cdots (x - x_{n-1})$$

i rekurentna formula

$$V(x_0, x_1, \ldots, x_{n-1}, x_n) = V(x_0, x_1, \ldots, x_{n-1}) \cdot (x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1}). \tag{11.2}$$

Kako je $V(x_0, x_1) = x_1 - x_0$, dobijamo iz 11.2

$$V(x_0, x_1, x_2) = (x_1 - x_0)(x_2 - x_0)(x_2 - x_1)$$

a posle višestruke primene formule 11.2 konačno

$$V(x_0, x_1, \ldots, x_n) = \prod_{i>j} (x_i - x_j).$$

■

Definicija 11.8. ***Elementarni simetrični polinomi.*** *Neka su dati brojevi $x_i, i = 1, 2, \ldots, n$. Zbir svih proizvoda formiranih od $x_1, x_2, \ldots, x_n$ kao faktora tako da svaki proizvod ima tačno k međusobno različitih faktora uzetih među $x_1, x_2, \ldots, x_n$ naziva se elementarni simetrični polinom reda k, i označava se sa σ_k. Pored toga, definiše se $\sigma_0 = 1$. Znači ,*

$$\begin{aligned}
\sigma_0 &= 1, \\
\sigma_1 &= x_1 + x_2 + \cdots + x_n, \\
\sigma_2 &= x_1x_2 + x_1x_3 + \cdots + x_{n-1}x_n, \\
\sigma_3 &= x_1x_2x_3 + x_1x_2x_4 + \cdots + x_{n-2}x_{n-1}x_n, \\
&\vdots \\
\sigma_k &= x_1x_2 \cdots x_k + \cdots + x_{n-k+1}x_{n-k+2} \cdots x_n, \\
&\vdots \\
\sigma_n &= x_1x_2 \cdots x_n.
\end{aligned}$$

Teorema 11.21. ***Vijetove formule.*** *Ako su $x_1, x_2, \ldots, x_n$ nule polinoma*

$$p_n(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n, \quad a_0 \neq 0,$$

onda važi

$$a_k = a_0 (-1)^k \sigma_k, \quad k = 0, 1, 2, \ldots, n,$$

gde je σ_k elementarni simetrični polinom reda k formiran od $x_1, x_2, \ldots, x_n$.

Teorema 11.22. *Ako su $x_1, x_2, \ldots, x_n$ nule polinoma*

$$p_n(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n, \quad a_0 \neq 0,$$

onda važi

$$p_n(x) = a_0 \left(x^n - \sigma_1 x^{n-1} + \sigma_2 x^{n-2} + \cdots + (-1)^k \sigma_k x^{n-k} + \cdots + (-1)^{n-1} \sigma_{n-1} x + (-1)^n \sigma_n \right).$$

Definicija 11.9. *Ako je funkcija $f(x)$ n puta neprekidno diferencijabilna u intervalu $[a, b]$ zapisuje se $f \in C^n[a, b]$.*

Definicija 11.10. *Za funkciju $f : [a, b] \to \mathbb{R}$ primitivna funkcija je neprekidna funkcija $F : [a, b] \to \mathbb{R}$ koja je diferencijabilna u (a, b) i za koju važi $F'(x) = f(x)$ za $x \in (a, b)$.*

Teorema 11.23. ***Teorema o međuvrednosti.*** *Neka je $f \in C[a,b]$ i neka je $A = f(a)$, $B = f(b)$. Tada za svako $C \in \mathrm{In}(A,B)$ postoji $\tau \in (a,b)$ sa osobinom $C = f(\tau)$.*

Teorema 11.24. ***Darbuova teorema.*** *Neka je $f \in C[a,b]$. Pretpostavimo da $f'(x)$ postoji za svako $x \in [a,b]$ i da je $A = f'(a)$, $B = f'(b)$. Tada za svako $C \in \mathrm{In}(A,B)$ postoji $\tau \in (a,b)$ takvo da je $C = f'(\tau)$.*

Dokaz. Bez gubitka opštosti možemo prepostaviti da je $A < B$, jer je u suprotnom slučaju dovoljno $f(x)$ zameniti sa $-f(x)$. Neka je C proizvoljan broj za koji važi $A < C < B$. Posmatrajmo funkciju $F(x) = f(x) - Cx$. Tada je

$$F'(x) = f'(x) - C, \qquad F'(a) = A - C < 0, \qquad F'(b) = B - C > 0.$$

Otuda vidimo da neprekidna funkcija $F(x)$ ima negativan izvod u a i pozitivan u b. Zbog toga $F(x)$ uzima u (a,b) vrednosti koje su manje od $F(a)$ i $F(b)$, te ima minimum u nekoj tački $\tau \in (a,b)$ u kojoj je njen prvi izvod jednak nuli, tj. $F'(\tau) = f'(\tau) - C = 0$, odnosno $f'(\tau) = C$. ■

Teorema 11.25. ***Teorema o srednjoj vrednosti.*** *Ako je funkcija $f : [a,b] \subset \mathbb{R} \to \mathbb{R}$ neprekidna na $[a,b]$ i diferencijabilna na (a,b), onda postoji $t \in (a,b)$ takvo da je*

$$f(a) - f(b) = f'(t)(b-a).$$

Definicija 11.11. ***Lipšicov uslov.*** *Ako zafunkciju $f : [a,b] \subset \mathbb{R} \to \mathbb{R}$ važi*

$$|f(x) - f(y)| \le \gamma |x - y|, \quad x, y \in [a,b],$$

kaže se da f zadovoljava Lipšicov uslov i piše se $f \in Lip_\gamma [a,b]$.

Teorema 11.26. ***Rolova teorema.*** *Neka je $f \in C[a,b]$ diferencijabilna funkcija u intervalu (a,b). Ako je $f(a) = f(b)$ onda postoji tačka $\tau \in (a,b)$ za koju je $f'(\tau) = 0$.*

Teorema 11.27. ***Uopštena Rolova teorema.*** *Neka je $n \ge 2$. Pretpostavimo da je $f \in C[a,b]$ i da $f^{(n-1)}(x)$ postoji za svako $x \in (a,b)$. Ako je $f(x_1) = f(x_2) = \cdots = f(x_n) = 0$ za međusobno različite vrednosti $x_i \in [a,b], i = 1,2,\ldots,n$, onda postoji tačka $\tau \in \mathrm{In}(x_1, x_2, \ldots, x_n)$ takva da je $f^{(n-1)}(\tau) = 0$.*

Dokaz. Bez umanjenja opštosti možemo da pretpostavimo da važi $a \le x_1 < x_2 < \cdots < x_n \le b$. Neka je $x_i^{(0)} = x_i, i = 1,2,\ldots,n$. Tada važi

$$.x_1^{(0)} < x_2^{(0)} < \cdots < x_n^{(0)}.$$

Pošto je $f(x)$ difrencijabilno, prema Rolovoj teoremi $f'(x)$ ima $n-1$ međusobno različitu nulu, označimo ih sa $x_i^{(1)}, i = 1,2,\ldots,n-1$, za koje važi

$$x_i^{(0)} < x_i^{(1)} < x_{i+1}^{(0)}, \quad i = 1,2,\ldots,n-1.$$

Kako $f''(x)$ takođe postoji, primena Rolove teoreme na $f'(x)$ daje egzistenciju tačaka $x_i^{(2)}, i = 1,2,\ldots,n-2$, za koje važi

$$x_i^{(1)} < x_i^{(2)} < x_{i+1}^{(1)}, \quad i = 1,2,\ldots,n-2,$$

$$f''(x_i^{(2)}) = 0, \quad i = 1,2,\ldots,n-2.$$

Primenjujući Rolovu teoremu redom na $f''(x), f'''(x), \ldots, f^{(n-1)}(x)$ dobijamo da $(n-1)$-vi izvod funkcije $f(x)$ ima jednu nulu τ sa osobinom

$$x_1^{(n-2)} < \tau < x_2^{(n-2)},$$

tj. $\tau \in \mathrm{In}(x_1, x_2, \ldots, x_n)$ i $f^{(n-1)}(\tau) = 0$. ■

Teorema 11.28. ***Tejlorova teorema.*** *Neka je $f \in C^{n+1}[a,b]$ i neka je $x_0 \in [a,b]$. Tada je za svako $x \in [a,b]$*

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \frac{1}{n!}\int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n dt.$$

Teorema 11.29. ***Tejlorova teorema.*** *Neka je $f \in C^n[a,b]$ i neka $f^{(n+1)}(x)$ postoji za $x \in (a,b)$. Tada postoji $\tau \in (a,b)$ takvo da je*

$$f(b) = f(a) + f'(a)(b-a) + \frac{f''(a)}{2!}(b-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(b-a)^n + \frac{f^{(n+1)}(\tau)}{(n+1)!}(b-a)^{n+1}.$$

Teorema 11.30. *Neka je dato preslikavanje $f : A \subset \mathbb{R}^n \to \mathbb{R}^m$. Sledeća tvrđenja su ekvivalentna.*

(i) f je neprekidno na A.

(ii) Za svaki konvergentan niz $x_k \to x_0$ u A važi $f(x_k) \to f(x_0)$.

Definicija 11.12. *Preslikavanje $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ je diferencijabilno (Freše-diferencijabilno) u tački $x \in D$ ako postoji linearno preslikavanje $A : \mathbb{R}^n \to \mathbb{R}^m$ takvo da je*

$$\lim_{h \to 0} \frac{\|F(x+h) - F(x)\|}{\|h\|} = 0.$$

Linearni operator A se označava sa $F'(x)$ i naziva $F-$izvod preslikavanja F u tački x, a njegova matrična reprezentacija je data matricom jakobijana

$$F'(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}.$$

Teorema 11.31. ***Teorema o srednjoj vrednosti.*** *Neka je funkcija $f : D \subset \mathbb{R}^n \to \mathbb{R}$ diferencijabilna u svakoj tački konveksnog skupa $D_0 \in D$. Tada za svake dve tačke $x, y \in D_0$ postoji $t \in (0,1)$ takvo da je*

$$f(x) - f(y) = f'(x + t(y-x))(x-y).$$

Neka je funkcija $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ diferencijabilna u svakoj tački konveksnog skupa $D_0 \in D$. Tada za svake dve tačke $x, y \in D_0$ postoje $t_1, t_2, \ldots, t_m \in (0,1)$ takvi da je

$$Fx - Fy = B(x,y)(x-y),$$

gde je matrica $B(x,y) \in \mathbb{R}^{n,m}$ dobijena pomoću komponenti $f_1, f_2, \ldots, f_m$ preslikavanja F

$$B(x,y) = \begin{bmatrix} f_1'(x + t_1(y-x))(x-y) \\ f_2'(x + t_2(y-x))(x-y) \\ \vdots \\ f_m'(x + t_m(y-x))(x-y) \end{bmatrix}.$$

Teorema 11.32. *(Tejlorova teorema za dimenziju dva). Neka je funkcija $f\,(x)$ definisana i $n+1$ put diferencijabilana u okolini $S(x_0,\rho)$, tačke $x_0=[x_1^0,x_2^0]^\top$ tada za svako $x=[x_1,x_2]^\top\in S(x_0,\rho)$ važi*

$$f(x)=f(x_0)+\sum_{k=1}^{n}\frac{1}{k!}\sum_{j=0}^{k}\binom{k}{j}\frac{\partial^k f\,(x_0)}{\partial x_1^{k-j}\partial x_2^j}(x_1-x_1^0)^{k-j}(x_2-x_2^0)^j+R_n\,(x)\,,$$

gde je za neko $\theta\in(0,1)$ $R_n\,(x)$ dato sa

$$\frac{1}{(n+1)!}\sum_{j=0}^{n+1}\binom{n+1}{j}\frac{\partial^{n+1}f\,(x_0+\theta(x-x_0))}{\partial x_1^{n+1-j}\partial x_2^j}(x_1-x_1^0)^{n+1-j}(x_2-x_2^0)^j.$$

Teorema 11.33. ***Tejlorova teorema za više dimenzija****. Neka je funkcija $f\,(x)$ definisana i $n+1$ put diferencijabilana u okolini $S(x_0,\rho)$, tačke $x_0=[x_1^0,x_2^0,\ldots,x_m^0]^\top$ tada za svako $x=[x_1,x_2,\ldots,x_m]^\top\in S(x_0,\rho)$ važi*

$$f(x)=f(x_0)+\sum_{k=1}^{n}\frac{1}{k!}\left((x_1-x_1^0)\frac{\partial}{\partial x_1}+\cdots+(x_m-x_m^0)\frac{\partial}{\partial x_m}\right)^k f\,(x_0)+R_n\,(x)\,,$$

gde je za neko $\theta\in(0,1)$ $R_n\,(x)$ dato sa

$$\frac{1}{(n+1)!}\left((x_1-x_1^0)\frac{\partial}{\partial x_1}+\cdots+(x_m-x_m^0)\frac{\partial}{\partial x_m}\right)^{n+1} f\,(x_0+\theta(x-x_0))\,.$$

Teorema 11.34. ***Osnovna teorema integralnog računa****. Neka je funkcija $f:[a,b]\to\mathbb{R}$ neprekidna . Tada postoji primitivna funkcija F i važi*

$$\int_a^b f(x)dx=F(b)-F(a).$$

Teorema 11.35. *Neka je funkcija $f:[a,b]\to\mathbb{R}$ integrabilna funkcija i neka x_1, x_2 i $x_3\in[a,b]$. Tada važi*

$$\int_{x_1}^{x3} f\,(x)\,dx=\int_{x_1}^{x_2} f\,(x)\,dx+\int_{x_2}^{x_3} f\,(x)\,dx.$$

Teorema 11.36. ***Teorema o srednjoj vrednosti integrala.*** *Neka je funkcija $f:[a,b]\to\mathbb{R}$ integrabilna i neka je*

$$m=\inf\{f\,(x):x\in[a,b]\}\,,\quad M=\inf\{f\,(x):x\in[a,b]\}\,.$$

Ako je funkcija $g:[a,b]\to\mathbb{R}$ integrabilna i ako je $g\,(x)\geq 0$ ili $g\,(x)\leq 0$ za $x\in[a,b]$ onda postoji $\mu\in[m,M]$ takvo da je

$$\int_a^b f\,(x)\,g\,(x)\,dx=\mu\int_a^b g\,(x)\,dx.$$

Ako je još $f\in C\,[a,b]$ onda postoji $\xi\in[a,b]$ takvo da je $\mu=f\,(\xi)$.

11.1.1 Oznake

- $\mathbb{N}$ – skup prirodnih brojeva
- $\mathbb{Z}$– skup celih brojeva
- $\mathbb{R}$– skup realnih brojeva
- $\mathbb{C}$– skup komleksnih brojeva
- $\mathcal{M}$– skup mašinskih brojeva
- $\mathcal{M}(\beta, t, e_{\min}, e_{\max})$ – skup mašinskih brojeva određenih osnovom β, preciznošću t, minimalnim eksponentom $e_{\min}$ i maksimalnim eksponentom $e_{\max}$
- $\mathbb{R}^n$– vektorski prostor uređenih n–torki nad poljem realnih brojeva
- $\mathbb{C}^n$– vektorski prostor uređenih n–torki nad nad poljem kompleksnih brojeva
- $x = [x_1, x_2, \ldots, x_n]^\top$– vektor kolona sa komponentama x_i
- $x^\top$– transponovani vektor vektopra x
- $e_1, e_2, \ldots, e_n$– vektori standardne baze vektorskog prostora
- $\left\{x^1, x^2, \ldots, x^m\right\}$ – skup m vektora
- $\|x\|$ – proizvoljna vektorska norma
- $\|x\|_p$ – vektorska p–norma , $1 \leq p \leq \infty$
- (x, y) – skalarni proizvod vektora x i y
- $\mathcal{S}(x, \rho)$ – otvorena lopta $\{y \in \mathbb{R}^n : \|y - x\| < \rho\}$
- $\mathcal{B}(x, \rho)$ – zatvorena lopta $\{y \in \mathbb{R}^n : \|y - x\| \leq \rho\}$
- $x \leq (<)\ y$– parcijalno uredjenje $x_i \leq (<)\ y_i$
- $|x|$ – vektor sa komponentama $|x_i|$
- $C[a, b], C(D)$ – skup svih na $[a, b]$, odnosno D neprekidnih funkcija
- $C^n[a, b], C^n(D)$ – skup svih na $[a, b]$, odnosno D – neprekidno diferencijabilnih funkcija
- $A = [a_{ij}]$ – matrica formata $m \times n$ sa elementima a_{ij}
- A^{-1}– inverzna matrica matrice A
- $A^\top$– transponovana matrica matrice A
- $\bar{A}$– konjugovana matrica matrice A
- A^H– konjugovano-transponovana matrica matrice A
- $\Delta(x)$ – greška broja (vektora) x
- Δ_x– granica apsolutne greške broja x
- $\sigma(A)$ – spektar matrice A
- $\rho(A)$ – spektralni radijus matrice A

- $\|A\|$ – proizvoljna matrična norma
- $\|A\|_p$ – matrična p–norma, $1 \leq p \leq \infty$
- $A \leq (<)\ B$– parcijalno uređenje $a_{ij} \leq (<)\, b_{ij}$
- $\operatorname{diag}(a_1, a_2, \ldots, a_n)$ – dijagonalna matrica sa elementima a_i
- $f_x(x,y) = \dfrac{\partial f}{\partial x}(x,y)$
- $f_y(x,y) = \dfrac{\partial f}{\partial y}(x,y)$
- $f_{xx}(x,y) = \dfrac{\partial^2 f}{\partial x^2}(x,y)$
- $f_{yy}(x,y) = \dfrac{\partial^2 f}{\partial y^2}(x,y)$
- $\Delta(x)$ – greška broja (vektora) x
- Δ_x– granica apsolutne greške broja x
- $\delta(x)$ – relativna greška broja x
- δ_x– granica relativne grteške broja x
- $k(A)$ – uslovni broj matrice A
- $f^{(k)}$ – k-ti izvod funkcije f. $f^{(0)} = f$.
- $\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$ – Kronekerov simbol
- Za $n \in \mathbb{Z}$, $s \in \mathbb{C}$ je

$$\binom{s}{n} = \begin{cases} \dfrac{s(s-1)\cdots(s-n+1)}{n!}, & n \in \mathbb{N}, \\ 1, & n = 0, \\ 0, & n < 0. \end{cases}$$

- $\operatorname{sgn} x = \begin{cases} 1, & x \geq 0, \\ -1, & x < 0. \end{cases}$
- M_k–konstanta za koju važi $\max\left\{\left|f^{(k)}(x)\right| \mid x \in D\right\} \leq M_k$.

Додатак Б. МАШИНСКИ БРОЈЕВИ

1. Увод

Познато нам је да се код рачунара сваки симбол кодира помоћу низа битова. Ако је дужина тог низа L, онда се може кодирати највише 2^L различитих симбола, што је коначан број, без обзира колико је велико L.

За представљање бројева у рачунару, тј. за њихово кодирање, користи се низ битова фиксиране дужине, формата N. Значи, у рачунару се може представити највише 2^N различитих реалних бројева, тј. сваки рачунар може тачно да представи само један подскуп скупа реалних бројева. Због тога се и појављује проблем како представити остале реалне бројеве, како **симулирати** аритметику и како контролисати грешке настале због тога.

Скуп машинских бројева, као и сваки његов подскуп, је дискретан скуп. Због тога, када кажемо да одређени подскуп машинских бројева покрива неки интервал реалних бројева, подразумевамо да за сваки број из интервала постоји одговарајући машински број. О пресликавању скупа реалних бројева у скуп машинских бројева говорићемо касније.

Пример 1.

Ако је формат $N = 32$ ради се о бројевима за чије се представљање користи 32 бита. Таквих бројева може бити највише

$$2^{32} = 4 \times 294 \times 967 \times 296.$$

2. Представљање целих бројева

Увод

Најприроднији начин да се низ битова $d_{N-1}\, d_{N-2} \cdots d_2\, d_1\, d_0$ интерпретира као реалан број је да се тај низ посматра као број у бинарном бројном систему, тј.

$$d_{N-1}\, d_{N-2} \cdots d_2\, d_1\, d_0 \doteq \sum_{j=0}^{N-1} d_j\, 2^j, \quad d_j \in \{0, 1\}.$$

Знак $\doteq$ се користи у смислу "одговара". На овај начин могу се представити сви цели бројеви од 0 до $2^N - 1$:

$$\begin{aligned} 0000 \;\cdots\; 0000 &\doteq 0 \\ 0000 \;\cdots\; 0001 &\doteq 1 \\ &\vdots \\ 1111 \;\cdots\; 1111 &\doteq 2^N - 1. \end{aligned}$$

За представљање негативних целих бројева користе се углавном следећа три начина: посебно кодирање знака, комплемент до један и комплемент до два.

Посебно кодирање знака

Код приказа целог броја знак се може кодирати посебно. Ако се бит $s \in \{0, 1\}$ користи за приказ знака, низ битова дужине N интерпретира се на следећи начин:

$$s\, d_{N-2}\, d_{N-3} \cdots d_2\, d_1\, d_0 \doteq (-1)^s \sum_{j=0}^{N-2} d_j\, 2^j, \quad d_j \in \{0, 1\}.$$

На овај начин могу се представити сви цели бројеви од $-\left(2^{N-1} - 1\right)$ до $2^{N-1} - 1$:

$$\begin{array}{lllll} 0000 & \cdots & 0000 & \doteq & 0 \\ 0000 & \cdots & 0001 & \doteq & 1 \\ & & \vdots & & \\ 0111 & \cdots & 1111 & \doteq & 2^{N-1} - 1 \end{array} \qquad \begin{array}{lllll} 1000 & \cdots & 0000 & \doteq & -0 \\ 1000 & \cdots & 0001 & \doteq & -1 \\ & & \vdots & & \\ 1111 & \cdots & 1111 & \doteq & -\left(2^{N-1} - 1\right). \end{array}$$

У овом случају број нула није кодиран једнозначно, већ као $+0$ и као -0.

Комплемент до један

Комплемент до један користи

$$\overline{x} = \sum_{j=0}^{N-1} \left(1 - d_j\right) 2^j, \quad d_j \in \{0, 1\}$$

за приказ негативног броја $-x$:

$$\begin{array}{lllll} 0000 & \cdots & 0000 & \doteq & 0 \\ 0000 & \cdots & 0001 & \doteq & 1 \\ & & \vdots & & \\ 0111 & \cdots & 1111 & \doteq & 2^{N-1} - 1 \end{array} \qquad \begin{array}{lllll} 1111 & \cdots & 1111 & \doteq & -0 \\ 1111 & \cdots & 1110 & \doteq & -1 \\ & & \vdots & & \\ 1000 & \cdots & 0000 & \doteq & -\left(2^{N-1} - 1\right). \end{array}$$

Ни у овом случају нула није кодирана једнозначно.

Комплемент до два

Комплемент до два користи приказ

$$\overline{\overline{x}} = 1 + \sum_{j=0}^{N-1} \left(1 - d_j\right) 2^j, \quad d_j \in \{0, 1\}$$

за приказ негативног броја $-x$:

$$\begin{array}{lllll} 0000 & \cdots & 0000 & \doteq & 0 \\ 0000 & \cdots & 0001 & \doteq & 1 \\ & & \vdots & & \\ 0111 & \cdots & 1111 & \doteq & 2^{N-1} - 1 \end{array} \qquad \begin{array}{lllll} 1111 & \cdots & 1111 & \doteq & -1 \\ & & \vdots & & \\ 1000 & \cdots & 0001 & \doteq & -\left(2^{N-1} - 1\right) \\ 1000 & \cdots & 0000 & \doteq & -2^{N-1}. \end{array}$$

У овом случају нула је кодирана једнозначно, али интервал $\left[-2^{N-1}, 2^{N-1} - 1\right]$ није симетричан у односу на нулу.

Пример 2. (Intel)

Intel микропроцесори користе комплемент до два за кодирање целих бројева. При томе користе формате $N = 16$, $N = 32$ и $N = 64$, које називају short integer, word integer и long integer респективно. Интервал којем припадају word integer бројеви је

$$\left[-2^{31},\, 2^{31}-1\right] = [-2\,147 \times 483 \times 648,\ 2 \times 147 \times 483 \times 647].$$

3. Бројеви са непокретном тачком

Интервал реалних бројева покривен машинским бројевима може се повећати или смањити множењем сваког броја са 2^k где је k неки цео број. На овај начин повећавамо или смањујемо густину бројева које приказујемо. Множењу са 2^k, где је k негативан број одговара низ битова $s\, d_{N-2} \cdots d_2\, d_1\, d_0$ као број у бинарном бројном систему са знаком, код кога је $-k > 0$ цифара иза децималне тачке:

$$s\, d_{N-2}\, d_{N-3} \cdots d_2\, d_1\, d_0 \ \doteq\ (-1)^s\, 2^k \sum_{j=0}^{N-2} d_j\, 2^j, \qquad d_j \in \{0, 1\},$$

односно

$$s\, d_{N-2}\, d_{N-3} \cdots d_2\, d_1\, d_0 \ \doteq\ (-1)^s\, d_{N-2} \cdots d_{-k}.\, d_{-k-1}\, d_{-k-2} \cdots d_0.$$

Скуп бројева добијем множењем са 2^k при чему је k фиксно, назива се систем бројева са непокретном (фиксном) тачком.

Пример 3.

Систем бројева са непокретном тачком формата $N = 16$ битова добијен множењем са 2^k, $k = -4$, одређује скуп бројева облика

$$s\, d_{14}\, d_{13} \cdots d_1\, d_0 \ \doteq\ (-1)^s\, 2^{-4} \sum_{j=0}^{14} d_j\, 2^j.$$

На пример,

$$1000 \times 0010\ 0101 \times 0101 \doteq -100101.0101_2 = -37.3125.$$

За $k = -(N-1)$ децимална тачка је испред водеће цифре d_{N-2}. Као и код целих бројева један бит $s \in \{0, 1\}$ служи за кодирање знака броја $(-1)^s$.

Основни недостатак система бројева са непокретном тачком је тај што покрива само бројеве одређене величине, а у току рачунања сложенијих израза тешко је предвидети величине свих међурезултата и тако омогућити њихово ваљано представљање у изабраном систему бројева.

4. Бројеви са покретном тачком

Увод

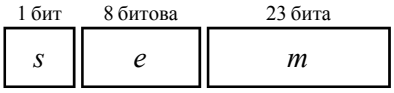
Много значајнији, посебно за стручна и научна рачунања, су рачунари код којих се бројеви представљају помоћу **покретне тачке**. Овде није фиксан положај децималне тачке и због тога се мора пазити после које цифре она долази. Положај те тачке задаје се помоћу **експонента**.

Низ битова којим се представља број у систему са покретном тачком састоји се из три дела: **знак** s, **експонент** e и **мантиса** m (значајне цифре). Формат таквог низа битова, који даје дужину сваког од три наведена дела, фиксан је за сваки систем бројева са покретном тачком.

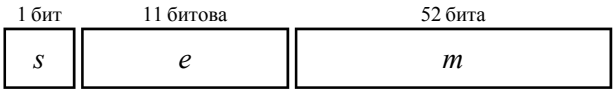
Пример 4. (IEC/IEEE систем)

Стандард IEC 559 из 1989. године за бинарну аритметику са покретном тачком предвиђа два основна формата:

Обична прецизност: формат $N = 32$ бита



Двострука прецизност: формат $N = 64$ бита



Мантиса m *је* ненегативан реалан број у систему са непокретном тачком (без знака)

$$m = d_1\, b^{-1} + d_2\, b^{-2} + \cdots + d_p\, b^{-p}$$

са фиксираном базом b бројног система и децималном тачком испред водеће цифре d_1. Цифра d_1 је **најзначајнија** (водећа) цифра, а d_p је **најмање значајна** цифра.

Природан број p дефинише прецизност представљања бројева у систему са покретном тачком.

Експонент e је цео број. Реалан број x приказује се као

$$x = (-1)^s\, b^e\, m,$$

при чему $s \in \{0, 1\}$ одређује знак броја x.

Скуп реалних бројева који се могу представити у систему са покретном тачком означавамо са $\mathcal{M}$. Овај скуп зависи од избора **базе** b бројног система са непокретном тачком у којем се кодира **мантиса** m, **прецизности** p и од дозвољених вредности за **експонент** e.

За нумеричку обраду података ирелевантно је како је имплементиран систем са покретном тачком у конкретном рачунару. Важан је само распоред бројева у систему са покретном тачком на реалној бројевној оси и аритметика у скупу $\mathcal{M}$. Због тога се дискусија о рачунарској аритметици заснива на систему са покретном тачком, а не на одговарајућем кодирању.

Систем са покретном тачком који има базу b, прецизност p и чији експонент припада скупу $[e_{\min}, e_{\max}] \cap \mathcal{Z}$ садржи следеће реалне бројеве

$$x = (-1)^s b^e \sum_{j=1}^{p} d_j b^{-j},$$

где је

$$s \in \{0, 1\},$$
$$e \in \{e_{\min}, e_{\min+1}, \ldots, e_{\max}\},$$
$$d_j \in \{0, 1, \ldots \mathrm{b} - 1\}, \qquad j = 1, 2, \ldots, p.$$

Цифре d_j су цифре мантисе.

Пример 5.

У децималном систему са покретном тачком са шест децималних места ($b = 10$ и $p = 6$) број 0.1 представља се као $0.100000 \cdot 10^0$. У бинарном систему са покретном тачком ($b = 2$) број 0.1 не може се представити тачно. Уместо тога користи се његово представљање помоћу 24 бинарне цифре

$$0.1100 \times 1100 \times 1100 \times 1100 \times 1100 \times 1101 \cdot 2^{-3}.$$

Нормализовани и денормализовани бројеви

Очигледно, представљање бројева са покретном тачком није једнозначно. Тако у децималном систему са покретном тачком са прецизношћу $p = 6$ имамо следећа представљања броја 0.123:

$$0.123000\,10^0, \quad 0.012300 \cdot 10^1, \quad 0.001230 \cdot 10^2, \quad 0.000123 \cdot 10^3.$$

Да би се ово избегло уводи се **нормализовано** представљање бројева са покретном тачком, увођењем додатног услова $d_1 \neq 0$. Дакле, број је представљен нормализовано ако је прва цифра мантисе различита од нуле. Одатле следи да за мантису важи $b^{-1} \leq m < 1$ за све нормализоване бројеве. На тај начин се сви бројеви различити од нуле могу представити у нормализованом облику на јединствен начин. Број нула представља се посебно, да би се избегло њено вишезначно представљање ($0 = 0 \cdot b^2 = 0 \cdot b^{-10} = 0 \cdot b^{25}$, итд.). Нула има мантису $m = 0$. Скуп чији су елементи нормализовани бројеви и нула означавамо са $\mathcal{M}_N$.

Бројеви који се не могу записати као нормализовани бројеви (бројеви чија је апсолутна вредност веома мала) могу се придружити скупу нормализованих бројева тако да јединствено представљање бројева буде сачувано. За ове бројеве узима се да је $d_1 = 0$ само за најмањи експонент $e = e_{\min}$. Сви бројеви са мантисом $m \in \left[b^{-p}, b^{-1} - b^{-p}\right]$ припадају интервалу $\left(-b^{e_{\min}-1}, b^{e_{\min}-1}\right)$ и називају се **денормализовани бројеви**. Скуп денормализованих бројева означавамо са $\mathcal{M}_D$.

Да би се обезбедило јединствено представљање нуле узима се да је бит који одређује знак нуле $s = 0$. Број нула (који има мантису $m = 0$) припада скупу $\mathcal{M}_N$.

У следећој табели приказана је обострано једнозначна коресподенција између ненегативних бројева из скупа $\mathcal{M}$ и парова (e,m). За ове бројеве је $s = 0$. Аналогна релација важи и за негативне бројеве из скупа $\mathcal{M}$, али, у овом случају је $s = 1$.

бројеви	експонент	мантиса
нормализовани	$e \in [e_{\min}, e_{\max}]$	$m \in \left[b^{-1}, 1 - b^{-p}\right]$
денормализовани	$e = e_{\min}$	$m \in \left[b^{-p}, b^{-1} - b^{-p}\right]$
нула	$e = e_{\min}$	$m = 0$

Систем бројева са покретном тачком окарактерисан је са четири целобројна параметра и једним логичким:

основа бројног система	$b \geq 2$
прецизност	$p \geq 2$
најмањи експонент	$e_{\min} < 0$
највећи експонент	$e_{\max} > 0$
индикатор нормализације	*denorm.*

Ако је *denorm* = *true* (тачно), онда су у скупу машинских бројева и нормализовани и денормализовани бројеви. Ако је *denorm* = *false* (нетачно), онда денормализовани бројеви не припадају скупу машинских бројева.

Ако са $\mathcal{M}(b, p, e_{\min}, e_{\max}, denorm)$ означимо скуп машинских бројева са покретном тачком, онда важи

$$\mathcal{M}(b, p, e_{\min}, e_{\max}, true) = \mathcal{M}_N(b, p, e_{\min}, e_{\max}) \bigcup \mathcal{M}_D(b, p, e_{\min}, e_{\max})$$
$$\mathcal{M}(b, p, e_{\min}, e_{\max}, false) = \mathcal{M}_N(b, p, e_{\min}, e_{\max}).$$

Данашњи рачунари користе за базу углавном 2 или 10, а ређе 16.

Пример 6. (Intel)

У Intel-овим микропроцесорима углавном се користе системи

$$\mathcal{M}(2, 24, -125, 128, \mathit{true}) \quad \text{и} \quad \mathcal{M}(2, 53, -1021, 1024, \mathit{true})$$

за обичну и двоструку тачност респективно. Intel користи и повећану тачност са

$$\mathcal{M}(2, 64 - 16\,381, 16\,384, \mathit{true}).$$

Пример 7. (IBM System/390)

IBM System/390 користи хексадецимални систем: кратка прецизност $\mathcal{M}(16, 6, -64, 63, \mathit{false})$, дугачка прецизност $\mathcal{M}(16, 14 - 64, 63, \mathit{false})$ и проширена прецизност $\mathcal{M}(16, 28, -64, 63, \mathit{false})$.

Пример 8. (Cray)

Cray рачунари користе два бројна система: $\mathcal{M}(2, 48, -16\,384, 8191, \mathit{false})$ и $\mathcal{M}(2, 96, -16\,384, 8191, \mathit{false})$.

Пример 9. (Џепни рачунари)

Џепни рачунари обично користе бројни систем $\mathcal{M}(10, 10, -98, 100, \mathit{false})$. Неки рачунари користе интерно и проширену тачност, али резултате приказују само са десет децималних места.

Број бројева у систему са покретном тачком

Укупан број машинских бројева у скупу $\mathcal{M}_N(b, p, e_{\min}, e_{\max})$ је

$$2(b-1)\,b^{p-1} \cdot (e_{\max} - e_{\min} + 1) + 1.$$

Укупан број могућих мантиса је

$$2(b-1)\,b^{p-1}$$

јер d_1 може бити један од бројева 1, 2, …, $b-1$. На преосталих $p-1$ места мантисе може се појавити ма која од b цифара, а све се множи са 2 због два могућа предзнака. Укупан број могућих експонената је $(e_{\max} - e_{\min} + 1)$. На крају додата јединица представља 0 која је такође машински број.

Пример 10. (IEC/IEEE систем)

IEC/IEEE систем има два система са следећим бројем бројева:

$\mathcal{M}_N(2, 24, -125, 128)$: $2^{24}\cdot 254+1 \approx 4.26141 \cdot 10^9$,

$\mathcal{M}_N(2, 53, -1021, 1024)$: $2^{53} \cdot 2046 + 1 \approx 1.84287 \cdot 10^{19}$.

Скуп нормализованих машинских бројева, за разлику од скупа реалних бројева, има најмањи и највећи позитиван машински број. Позитиван нормализован машински број може се записати у облику

$$b^e \left(d_1\, b^{-1} + d_2\, b^{-2} + \cdots + d_p\, b^{-p}\right).$$

Најмањи позитиван нормализован машински број добија се када је прва цифра $d_1 = 1$, а остале цифре су нуле и експонент је најмањи, тј. $e = e_{\min}$. Дакле,

$$x_{\min} = b^{e_{\min}-1}.$$

Највећи позитиван нормализован машински број добија се када су све цифре $b - 1$, а експонент је највећи, тј. $e = e_{\max}$. Дакле,

$$x_{\max} = b^{e_{\max}} (b-1)\left(b^{-1} + b^{-2} + \cdots + b^{-p}\right) = b^{e_{\max}} (1 - b^{-p}).$$

Пример 11. (IEC/IEEE систем)

IEC/IEEE систем са покретном тачком има два система са следећим највећим и најмањим позитивним нормализованим бројевима:

$\mathcal{M}_N(2, 24, -125, 128)$:

$x_{\min} = 2^{-126} \approx 1.17549 \cdot 10^{-38}$, $x_{\max} = 2^{128}\left(1 - 2^{-24}\right) \approx 3.40282 \cdot 10^{38}$,

$\mathcal{M}_N(2, 53, -1021, 1024)$:

$x_{\min} = 2^{-1022} \approx 2.22507 \cdot 10^{-308}$, $x_{\max} = 2^{1024}\left(1 - 2^{-53}\right) \approx 1.79769 \cdot 10^{308}$.

Знак броја $(-1)^s$ представља се посебно, па је због тога представљање машинских бројева симетрично у односу на нулу, тј.

$$x \in \mathcal{M} \iff -x \in \mathcal{M}.$$

На основу тога закључујемо да су $-x_{\max}$ и $-x_{\min}$ најмањи и највећи негативан нормализован број.

Ако се посматрају и денормализовани бројеви, онда су $\overline{x}_{\min}$ и $-\overline{x}_{\min}$ највећи и најмањи позитиван денормализован број, где је

$$\overline{x}_{\min} = b^{e_{\min}-p}.$$

Размак између бројева у систему са покретном тачком

За сваки експонент $e \in (e_{\min}, e_{\max})$ најмања $m_{\min}$ и највећа $m_{\max}$ мантиса нормализованих машинских бројева имају следеће цифре

$$d_1 = 1,\ d_2 = d_3 = \cdots =\ \ d_p = 0, \qquad d_1 =\ d_2 = d_3 = \cdots =\ \ d_p = b - 1$$

респективно. Нека је $\delta = b - 1$. Тада је

$$m_{\min} = .100\ldots00_b = b^{-1},$$

$$m_{\max} =. \delta\delta\delta\ldots\delta\delta_b = \sum_{j=1}^{p} (b-1)\, b^{-j} = 1 - b^{-p}.$$

У овом случају мантиса се мења од $m_{\min}$ до $m_{\max}$ са кораком b^{-p}. Овај основни корак, који одговара вредности последњег цифарског места, назива се *ulp* (unit of last position). У даљем раду *ulp* означавамо са *u:*

$$u = b^{-p}.$$

Апсолутно растојање између суседних нормализованих машинских бројева у интервалу $\left[b^e, b^{e+1}\right]$ је константно

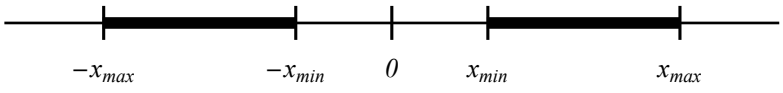
$$\Delta x = b^{e-p} = u\, b^e.$$

На овај начин сваки интервал облика $\left[b^e, b^{e+1}\right]$ окарактерисан је константном густином нормализованих бројева. Ако се експонент e смањи за један, растојање суседних бројева ће се смањити за b пута, што значи да се густина нормализованих бројева повећава за исти фактор. Аналогно томе, ако се експонент e повећа за један, растојање суседних нормализованих бројева се повећава b пута, а густина се смањује за исти фактор. Овде се понавља низ од $(b-1)\, b^{p-1}$ еквидистантних бројева, а сваки нови низ добијен је од претходног множењем са b.

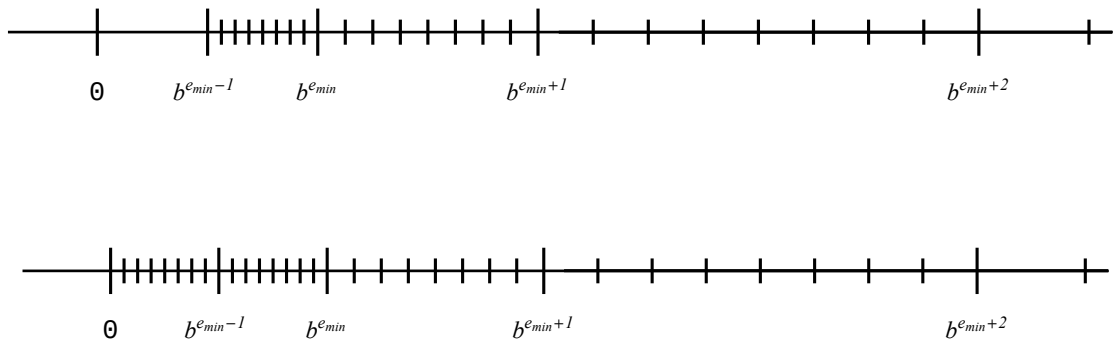
Скуп $\mathcal{M}_N$ има празан простор око нуле, слика 1. Наиме, у случају да је *denorm* = *false* постоје само два нормализована машинска броја у скупу $[0, x_{\min}]$ и то су баш 0 и $x_{\min}$. Већ следећи интервал, $[x_{\min}, b\, x_{\min}]$ (који је исте дужине ако је $b = 2$) садржи $1 + b^{p-1}$ нормализованих бројева, слика 2. Тако IEC/IEEE систем са покретном тачком има у скупу $\mathcal{M}_N(2, 24, -125, 128)$ $8 \times 388 \times 609$ нормализованих бројева који припадају интервалу $[x_{\min}, b\, x_{\min}]$.

Ако посматрамо денормализоване бројеве у интервалу $(0, x_{\min})$, видећемо да их има $b^{p-1} - 1$ и да су еквидистантно распоређени са размаком $u\, b^{e_{\min}}$ између суседних бројева, слика 3. Такође видимо да је најмањи позитиван денормализован број $\overline{x}_{\min} = b^{e_{\min}-p}$ много ближи нули него најмањи позитиван нормализован број $x_{\min} = b^{e_{\min}-1}$.

Негативни денормализовани бројеви су у односу на нулу симетрично пресликани позитивни денормализовани бројеви.



Слика 1.



Слика 3.

Апсолутна разлика Δx броја $x \in \mathcal{M}_N$ и првог броја већег од x из истог скупа расте са порастом експонента e: $\Delta x = u\, b^e$. Релативно растојање $\dfrac{\Delta x}{x}$ остаје скоро непромењено, не зависи од експонента e већ само од мантисе $m(x)$:

$$\frac{\Delta x}{x} = \frac{(-1)^s\, u\, b^e}{(-1)^s\, m(x)\, b^e} = \frac{u}{m(x)} = \frac{b^{-p}}{m(x)}.$$

Ako je $x \in \left[b^e, b^{e+1}\right]$, онда релативно растојање опада од $b\,u$ до u када x расте, јер је $b^{-1} \le m(x) < 1$. За $x = b^{e+1}$, када је $m = b^{-1}$, релативно растојање постаје $b\,u$ и опет почиње да опада. Овакво понашање понавља се у сваком од интервала $\left[b^e, b^{e+1}\right]$. У овом смислу релативно растојање је константно у $\mathcal{M}_N$. Очигледно да релативно растојање зависи од b, и што је b веће веће је и релативно растојање. Да би се релативно растојање смањило, најбоље је узети $b = 2$.

Код денормализованих бројева $\mathcal{M}_D$ релативна униформност се губи. Пошто $m(x) \to 0$, релативно растојање расте брзо када $x \to 0$. Апсолутно растојање $\Delta x = b^{e_{\min}-1}$ и у овом случају остаје константно.

Систем бројева са покретном тачком $\mathcal{M}(b, p, e_{\min}, e_{\max}, true)$ има у сваком од следећих скупова различит распоред бројева:

$$\begin{aligned}
\mathcal{R}_N &= [-x_{\max}, -x_{\min}] \cup [x_{\min}, x_{\max}] \\
\mathcal{R}_D &= (-x_{\min}, x_{\min}) \\
\mathcal{R}_\infty &= (-\infty, -x_{\max}) \cup (x_{\max}, \infty).
\end{aligned}$$

Ово нам сугерише и поделу скупа реалних бројева у три дисјунктна подскупа:

- $\mathcal{R}_N$ је подскуп који покривају машински бројеви који имају скоро униформну релативну густину,
- $\mathcal{R}_D$ је подскуп који покривају нула и денормализовани машински бројеви који имају униформну апсолутну густину,
- $\mathcal{R}_\infty$ је подскуп који не покрива ниједан машински број.

5. Стандардизација бројева са покретном тачком

Увод

Као и у многим другим областима живота и у програмирању и раду са рачунарима појављују се стандарди који имају за циљ да се олакша рад и споразумевање међу различитим субјектима. Међународна организација за стандард (International Standardization Organization - ISO) обухвата око сто националних организација за стандарде. Стандарди за програмске језике, такође, спадају у ISO стандарде.

За електро инжењерство и електронику стандарде је развила међународна електротехничка комисија (International Electrotechnical Commission - IEC).

Скуп машинских бројева је коначан и већ једноставни примери показују да резултат аритметичких операција са машинским бројевима не мора бити машински број. Да би се обезбедила затвореност скупа машинских бројева у односу на аритметичке операције, дефинишу се псеудоаритметичке операције. Проблеми који се при томе јављају решавани су неким стандардима и пре него што је уведен стандард IEEE 754. Ови стандарди су се односили на:

- формат и кодирање основног и проширеног бројног система (обична и двострука тачност),
- елементарне операције и правила заокруживања,
- конверзију између различитих представљања бројева, као и између децималног бројног система и бинарног бројног система,
- обраду изузетака, као што су underflow, overflow, дељење нулом и слично.

Као резултат дужих преговарања, Америчко рачунарско удружење IEEE (скраћеница од Institute of Electronical and Electronics Engineers) усвојило је IEEE Standard 754-1985. IEC је одлучила 1989. године да овај стандард постане међународни: IEC 559:1989 Binary Floating-Point Arithmetic for Microprocessor Systems.

IEEE формати

Основни формат за обичну тачност је $\mathcal{M}(2, 24, -125, 128, \mathit{true})$ а за двоструку тачност $\mathcal{M}(2, 53, -1021, 1024, \mathit{true})$.

За проширене формате дефинишу се само границе за параметре:

проширен IEEE стандард

	обична тачност	двострука тачност
дужина формата	$N \geq 43$	$N \geq 79$
тачност	$p \geq 32$	$p \geq 64$
најмањи експонент	$e_{\min} \leq -1021$	$e_{\min} \leq -16\,381$
највећи експонент	$e_{\max} \geq 1024$	$e_{\max} \geq 16\,384$

Многе имплементације користе само један проширен формат, који се односи на бројни систем $\mathcal{M}(2, 64, -16\,381, 16\,384, \mathit{true})$. На пример, HP Workstations користе $\mathcal{M}(2, 113, -16\,381, 16\,384, \mathit{true})$.

При имплементацији бројева са покретном тачком у IEC/IEEE стандарду користи се **скривени бит**, тј. водећа цифра $d_1 = 1$ код нормализованих бројева и водећа цифра $d_1 = 0$ код денормализованих бројева се не кодирају. На тај начин, за кодирање мантисе довољно је 23 односно 52 бита. За кодирање знака броја користи се још један бит.

Да би се нормализовани бројеви разликовали од денормализованих, као експонент денормализованих бројева користи се вредност $e_{\min} - 1$.

IEC/IEEE стандард користи системе са форматом $N = 32$, $N = 64$ или $N = 80$. Не користе се сви низови битова само за бројеве. Неки се користе за ± 0 и $\pm \infty$, као и за кодирање симбола који означавају NaN (Not a Number) бројеве.

вредност која се кодира	експонент	мантиса
$(-1)^s \cdot 0 = \pm 0$	$e_{\min} - 1$	нула
$(-1)^s \cdot \infty = \pm \infty$	$e_{\min} + 1$	нула
NaN	$e_{\min} + 1$	$\neq$ нуле

NaN је специјалан симбол који се јавља само код недозвољених операција. Овај симбол није упоредив, тј. ако се јавља као резултат две различите операције не представља исту величину. Његова мантиса је било који број различит од нуле. Симбол $\pm\infty$ има један бит за знак и исти експонент као NaN. Нула има такође један бит за знак, али је операција поређења тако дефинисана да важи $+0 = -0$.

Следећа табела садржи још нека решења наведених стандарда:

изузетак	пример	резултат
недозвољена операција	$\frac{0}{0},\ 0 \cdot \infty,\ \sqrt{-1}$	NaN
overflow		$\pm \infty$
дељење нулом	$\frac{c}{0},\ c \neq 0$	$\pm \infty$
underflow		денормализован број
нетачност	$fl\,(x \text{ op } y) \neq x \text{ op } y$	заокружен број

Са **op** означена је једна од операција $+,\ -,\ \cdot,\ /$, а ***fl*** је ознака за заокруживање.

Специјалисти за програмске језике и компајлере нису били укључени у рад на стандардима. Због тога програмски језици још увек не користе све погодности које пружају ови стандарди. У том погледу највише је учињено са програмским језиком FORTRAN 90. Сада се IEC/IEEE стандард користи код побољшања многих програмских језика: Ada, BASIC, Common Lisp, FORTRAN, Modula-2, Pascal и PL/I.

6. Аритметика система са покретном тачком

Увод

При практичном раду можемо користити само оне реалне бројеве који су на располагању у рачунару. То значи да и операције са овим бројевима морају бити тако дефинисане да се као резултат добијају бројеви који се могу представити у рачунару. Посматраћемо нумеричке операције у скупу $\mathcal{M}(b, p, e_{\min}, e_{\max}, denorm)$, ради једноставности. Овај скуп је подскуп скупа реалних бројева. Сем у ретким случајевима, резултат примене неке операције на елементе скупа $\mathcal{M}$ не припада том скупу.

Резултат аритметичких операција у скупу $\mathcal{M}(b, p, e_{\min}, e_{\max}, denorm)$ обично захтева више од p цифара мантисе и експонент који не припада интервалу $[e_{\min}, e_{\max}]$ да би могао бити представљен у посматраном скупу. У општем случају, резултати дељења и многих стандардних операција не могу се представити у рачунару, поготово не са коначно много цифара мантисе.

Проблем који се појављује код дефинисања операција у скупу $\mathcal{M}$ сличан је проблему рада са децималним разломцима, тачан резултат потребно је заокружити тако да буде број у скупу $\mathcal{M}$. Термин "тачан резултат" означава резултат у скупу реалних бројева $\mathcal{R}$ који је добијен неком операцијом са реалним бројевима. Како је $\mathcal{M} \subset \mathcal{R}$, тачан резултат је увек добро дефинисан за све машинске бројеве.

Пример 12.

Посматрамо скуп

$\mathcal{M}(10, 6, -9, 9, true)$

и оператор $\square : \mathcal{R} \to \mathcal{M}$, који сваком реалном броју додељује најближи машински број из датог скупа. Нека је

$$x = .135797 \cdot 10^6, \quad y = .246864 \cdot 10^0.$$

тачан резултат			заокружен резултат		
$x + y$	$=$	$.135797 \times 24\,686 \cdot 10^6$	$x + y$	$=$	$.135797 \cdot 10^6$
$x - y$	$=$	$.135796 \times 75\,314 \cdot 10^6$	$x - y$	$=$	$.135797 \cdot 10^6$
$x \cdot y$	$=$	$.335233 \times 90\,608 \cdot 10^5$	$x \cdot y$	$=$	$.335234 \cdot 10^5$
x / y	$=$	$.550088 \times 30\,773 \cdot 10^6$	x / y	$=$	$.550088 \cdot 10^6$
$\sqrt{x}$	$=$	$.368506 \times 44\,499 \cdot 10^3$	$\sqrt{x}$	$=$	$.368506 \cdot 10^6$

Функција заокруживања

Заокруживање или редукционо пресликавање је пресликавање

$$\square : \mathcal{R} \to \mathcal{M},$$

које сваком реалном броју x додељује машински број по одређеним правилима.

Ако је функција заокруживања позната, онда се аритметичка операција може дефинисати на следећи начин.

За аритметичку операцију у скупу реалних бројева

$$\circ : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$$

одговарајућа операција у скупу машинских бројева је $\boxed{\circ}$

$$\boxed{\circ} : \mathcal{M} \times \mathcal{M} \to \mathcal{M}$$

дефинисана са

$$x \boxed{\circ} y = \square(x \circ y).$$

Овако дефинисана функција прво израчунава тачан резултат $x \circ y$, а затим тај резултат заокруживањем пресликава у скуп $\mathcal{M}$. Значи, $x \boxed{\circ} y \in \mathcal{M}$.

Слично се дефинишу и функције једне промењиве, какве су скоро све стандардне функције у рачунарима. За функцију $f : \mathcal{R} \to \mathcal{R}$ одговарајућа функција је

$$\boxed{f} : \mathcal{M} \to \mathcal{M},$$

дефинисана са

$$\boxed{f}(x) = \square f(x).$$

Функција заокруживања мора задовољавати још и следеће особине:
пројективност

$$\square x = x \quad \text{за } x \in \mathcal{M}$$

и монотоност

$$x \le y \iff \square x \le \square y \qquad \text{за } x, y \in \mathcal{M}.$$

Оптимално заокруживање

Нека су $x_1,\ x_2 \in \mathcal{M}$ два суседна машинска броја и нека је

$$\hat{x} = \frac{x_1 + x_2}{2}.$$

Ако је $x \in [x_1, x_2]$, онда се заокруживање на најближи машински број, оптимално заокруживање, дефинише на следећи начин:

$$\square x = \begin{cases} x_1 & \text{ако је } x_1 \le x < \hat{x} \\ x_2 & \text{ако је } \hat{x} < x \le x_2. \end{cases}$$

Број $x \in \mathcal{R}_N \cup \mathcal{R}_D$ заокружује се на њему најближи број из скупа $\mathcal{M}$. У случају када је $x = \hat{x}$ постоје две стратегије.

У првој се за $\square x$ узима број који је даљи од нуле. Ова стратегија се користи код већине електронских рачунара.

У другој варијанти као резултат заокруживања узима се онај машински број који се завршава цифром која је, када се посматра као број, паран број. У овом случају база b мора бити паран број.

Пример 13.

У скупу $\mathcal{M}(10, 6, -9, 9, true)$ број 0.1000005 се при заокруживању даље од нуле, заокружује на $.100001 \cdot 10^0$. Код заокруживања на паран број, 0.1000005 се заокружује на $.100000 \cdot 10^0$ (а не на $.100001 \cdot 10^0$), а број 0.1000015 се заокружује на $.100002 \cdot 10^0$.

Код одсецања, тј. код заокруживања према нули, тачка $\hat{x}$ зависи од знака броја x ($\text{sign}(x) = -1$, ако је $x < 0$ и $\text{sign}(x) = 1$ ако је $x \geq 0$) и одређује се према

$$\hat{x} = \text{sign}\,(x) \max\{|x_1|, |x_2|\}.$$

У овом случају $\hat{x}$ је, дакле, онај од бројева x_1 и x_2 који има већу апсолутну вредност, а број $\square x$ је онај од бројева x_1 и x_2 који има мању апсолутну вредност. Увек важи

$$x \in \mathcal{R} \setminus \mathcal{M} \;\Rightarrow |\square x| < |x|.$$

Директно заокруживање

Поред оптималног заокруживања дефинише се и директно заокруживање, или заокруживање према плус бесконачно и заокруживање према минус бесконачно. У овим случајевима је

$$\hat{x} = \min\{x_1, x_2\} \quad \text{и} \quad \hat{x} = \max\{x_1, x_2\}.$$

Значи, код директног заокруживања резултат $\square x$ је увек следећи мањи или следећи већи број, без обзира какав је знак броја $x \notin \mathcal{M}$. За број $x \notin \mathcal{M}$ увек је $\square x > x$ код заокруживања према плус бесконачно, односно $\square x < x$ код заокруживања ка минус бесконачно.

Оптимално заокруживање и одсецање су симетричне функције, тј. за њих важи

$$\square(-x) = -(\square x).$$

Директно заокруживање нема ову особину.

Пример 14.

У скупу $\mathcal{M}(10, 6, -9, 9, true)$ број 0.112256788 заокружује се на следеће начине:

$$\square x = \begin{cases} .112257, & \text{за оптимално заокруживање и заокруживање према } +\infty \\ .112256, & \text{за одсецање и заокруживање према } -\infty. \end{cases}$$

До сада су разматране функције заокруживања под претпоставком да је $x \in \mathcal{R} \setminus \mathcal{M}$ и да постоје $x_1,\ x_2 \in \mathcal{M}$ такви да је $x_1 < x < x_2$. То је случај када је $x \in \mathcal{R}_N \bigcup \mathcal{R}_D \setminus \mathcal{M}$, али не и када је $x \in \mathcal{M}_\infty$.

Overflow

Ако је $x \in \mathcal{R}_\infty = (-\infty, -x_{\max}) \bigcup (x_{\max}, +\infty)$ онда је вредност $\square x$ недефинисана. Ако је x тачан резултат неке операције са аргументима из скупа машинских бројева $\mathcal{M}$, онда се код већине рачунара појављује порука да је наступио overflow или exponent overflow и програм се прекида. Понекад понашање рачунара у таквим случајевима може дефинисати сам корисник.

Underflow

Ако је $x \in \mathcal{R}_D = (-x_{\min}, x_{\min})$ вредност $\square x$ је увек добро дефинисана, без обзира да ли су денормализовани бројеви укључени у разматрање или нису. У овом случају појављује се порука да је наступио underflow или exponent underflow. Обично се као међурезултат узима број нула и користи се за наставак рачунања.

Грешка заокруживања

Разлика

$$\varepsilon(x) = \square x - x$$

заокружене вредности $\square x \in \mathcal{M}$ и тачног броја $x \in \mathcal{R}$ назива се апсолутна грешка заокруживања. Количник

$$\rho(x) = \frac{\square x - x}{x} = \frac{\varepsilon(x)}{x}$$

је релативна грешка заокруживања броја x.

За број $x \in \mathcal{R}_N$ вредност $\square x$ је добро дефинисана и има јединствено представљање у облику

$$\square x = (-1)^s \, m(x) \, b^{e(x)},$$

где је $m(x)$ мантиса а $e(x)$ експонент броја $\square x$. Дужина најмањег интервала $[x_1, x_2]$, $x_1, x_2 \in \mathcal{R}_N$, који садржи $x \in \mathcal{R}_N \setminus \mathcal{M}_N$ је

$$\Delta x = u \, b^{e(x)}$$

тако да је

$$|e(x)| \begin{cases} < u \, b^{e(x)}, & \text{код директног заокруживања и одсецања} \\ \leq \dfrac{u}{2} \, b^{e(x)}, & \text{код оптималног заокруживања.} \end{cases}$$

Ако је $x \in \mathcal{R}_D$ и заокруживањем се добија денормализован број, узима се да је $e(x) = e_{\min}$.

Граница релативне грешке заокруживања за број $x \in \mathcal{R}_N$ добија се на основу претходног

$$|\rho(x)| \begin{cases} < \dfrac{u}{|m(x)|} \le b\,u, & \text{код директног заокруживања и одсецања} \\ \le \dfrac{u}{2\,|m(x)|} \le \dfrac{b}{2}\,u, & \text{код оптималног заокруживања.} \end{cases}$$

Граница релативне грешке често се назива и машинско епсилон, у ознаци *eps*

$$eps = \begin{cases} b\,u = b^{1-p}, & \text{код директног заокруживања и одсецања} \\ \dfrac{b}{2}\,u = \dfrac{1}{2}\,b^{1-p}, & \text{код оптималног заокруживања.} \end{cases}$$

Пример 15. (IEC/IEEE аритметика)

За обичну прецизност машинско епсилон дато је са

$$eps = \begin{cases} 2^{-23} \approx 1.19 \cdot 10^{-7}, & \text{код директног заокруживања и одсецања} \\ 2^{-24} \approx 5.96 \cdot 10^{-8}, & \text{код оптималног заокруживања.} \end{cases}$$

За двоструку тачност је

$$eps = \begin{cases} 2^{-52} \approx 2.2 \cdot 10^{-16}, & \text{код директног заокруживања и одсецања} \\ 2^{-53} \approx 1.1 \cdot 10^{-16}, & \text{код оптималног заокруживања.} \end{cases}$$

Пример 16. (IBM System/360 аритметика)

За обичну прецизност је машински епсилон дато са

$$eps = \begin{cases} 16^{-5} \approx 9.54 \cdot 10^{-7}, & \text{код директног заокруживања и одсецања} \\ \frac{1}{2}\,16^{-5} \approx 4.77 \cdot 10^{-7}, & \text{код оптималног заокруживања.} \end{cases}$$

Теорема 1. Грешка заокруживања

За свако $x \in \mathcal{R}_N$ постоји неко $\rho \in \mathcal{R}$ такво да важи

$$\square x = x(1+\rho), \quad |\rho| \le eps.$$

На основу претходне теореме следи да за границу релативне грешке заокруживања $\rho(x)$ важи $|\rho(x)| \le eps$, ако је $x \in \mathcal{R}_N$.

Заокруживање и аритметичке операције

При практичном рачунању, сем у ретким случајевима, сви бесконачни децимални разломци морају бити замењени коначним децималним разломцима. Поред тога, при коришћењу рачунара јављају се проблеми и са записом података који се учитавају, међурезултата и коначних резултата. Посматраћемо сада, као општију, ситуацију која настаје при раду с рачунаром.

Скуп машинских бројева $\mathcal{M}$ је коначан. Тиме се као први проблем појављује апроксимација броја x, који није машински број, неким машинским бројем. Овај проблем се не јавља само код уношења података већ и у току рачунања. Као што једноставни примери показују, већ резултати једноставних аритметичких операција не морају припадати скупу машинских бројева $\mathcal{M}$, иако сами бројеви x и y припадају том скупу.

Као директну последицу претходне теореме имамо следеће две теореме.

Теорема 2. Грешка псеудоаритметичких операција

Предпоставимо да тачан резултат операције $\circ$ припада скупу $\mathcal{R}_N$. Тада за свако $x, y \in \mathcal{R}_N$ постоји неко $\rho \in \mathcal{R}$ такво да важи

$$x \circ y = (x \circ y)(1 + \rho), \quad |\rho| \leq eps.$$

Док је ток аритметичких псеудооперација у скупу машинских бројева данас готово у свим рачунарима хардверски уређен, извођење нерационалних операција (нпр. $\sqrt{\ }$, sin, cos, exp, итд.) скоро увек је решено софтверски. Извођење ових операција своди се обично на извођење коначног низа рационалних операција.

У многим разматрањима рационалних и нерационалних операција у једном нумеричком алгоритму, често се не прави разлика. Наиме, при имплементацији и једне и друга прелазе у псеудооперације које у општем случају дају рачунску грешку. Наравно, важно је и код нерационалних операција познавати границу рачунских грешака. Идеално би било да за имплментацију сваке нерационалне операције важи следећа теорема.

Теорема 3. Грешка функције

Претпоставимо да тачан резултат функције $f : D \to \mathcal{R}$ аргумента $x \in D \cap \mathcal{M}_N$ припада скупу $\mathcal{R}_N$. Тада за свако x постоји неко $\rho \in \mathcal{R}$ такво да важи

$$\boxed{f}\left(x\right) = f(x)(1 + \rho), \quad |\rho| \leq eps.$$

Код данашњих добрих рачунара израчунавање вредности одређених функција тако је имплементирано да тврђење претходне теореме важи за веће подручје скупа машинских бројева. За екстремне вредности аргумента могу наступити и веће релативне грешке ρ.

Пример 17.

Посматраћемо збир три машинска броја $x, y, z \in \mathcal{M}_N$, под претпоставком да су тачни резултати $x+y$, $y+z$, $x+y+z$ у скупу $\mathcal{R}_N$. Збир $x+y+z$ може се израчунати на следећа два начина:

$$\begin{aligned}\left(x \boxed{+} y\right) \boxed{+} z &= (x+y)(1+\rho_1) \boxed{+} z \\ &= ((x+y)(1+\rho_1)+z)(1+\rho_2) \\ &= x+y+z+(x+y)(\rho_1+\rho_2+\rho_1\rho_2+z\rho_2),\end{aligned}$$

$$\begin{aligned}x \boxed{+} \left(y \boxed{+} z\right) &= x \boxed{+} (y+z)(1+\rho_3) \\ &= (x+(y+z)(1+\rho_3))(1+\rho_4) \\ &= x+y+z+x\rho_4+(y+z)(\rho_3+\rho_4+\rho_3\rho_4).\end{aligned}$$

При томе важи $|\rho_k| \le eps$, за $k = 1, 2, 3, 4$. Ако занемаримо eps^2, тј. ако посматрамо само линеарну оцену границе апсолутне грешке, добијамо

$$\left|\left(x \boxed{+} y\right) \boxed{+} z - (x+y+z)\right| \le C_1 \quad \text{где је} \quad C_1 \approx (2|x+y|+|z|)\,eps$$

$$\left|x \boxed{+} \left(y \boxed{+} z\right) - (x+y+z)\right| \le C_2 \quad \text{где је} \quad C_2 \approx (|x|+2|y+z|)\,eps.$$

Ако је $|x| \gg |y|$ и $|x| \gg |z|$ приближна вредност границе апсолутне грешке мања је у другом случају. Одатле проистиче и правило да се више бројева сабира тако што се прво сортирају у неопадајући низ, а затим се сабирају редом први и други број, њихов збир и трећи број итд.

Посматрајући претходни пример, можемо закључити да у скупу машинских бројева $\mathcal{M}$ не важе сви закони аритметике. Наиме, очигледно је

$$\left(x \boxed{+} y\right) \boxed{+} z \neq x \boxed{+} \left(y \boxed{+} z\right), \qquad \left(x \boxed{\cdot} y\right) \boxed{\cdot} z \neq x \boxed{\cdot} \left(y \boxed{\cdot} z\right)$$

и

$$x \boxed{\cdot} \left(y \boxed{+} z\right) \neq \left(x \boxed{\cdot} y\right) \boxed{+} \left(x \boxed{\cdot} z\right).$$

Комутативни закон за сабирање и множење важи, јер је

$$\left(x \boxed{+} y\right) = \square(x+y) = \square(y+x) = y \boxed{+} x$$

и

$$\left(x \boxed{\cdot} y\right) = \square(x \cdot y) = \square(y \cdot x) = y \boxed{\cdot} x.$$

Видимо да математички еквивалентни изрази не морају имати исту вредност ако се рачуна у систему бројева са покретном тачком. Због тога имплементацију аритметичких операција називамо **псеудоаритметичке операције**.

Свака аритметичка операција, свака функција и сваки аргумент, односно сваки избор аргумената, има своје ρ.

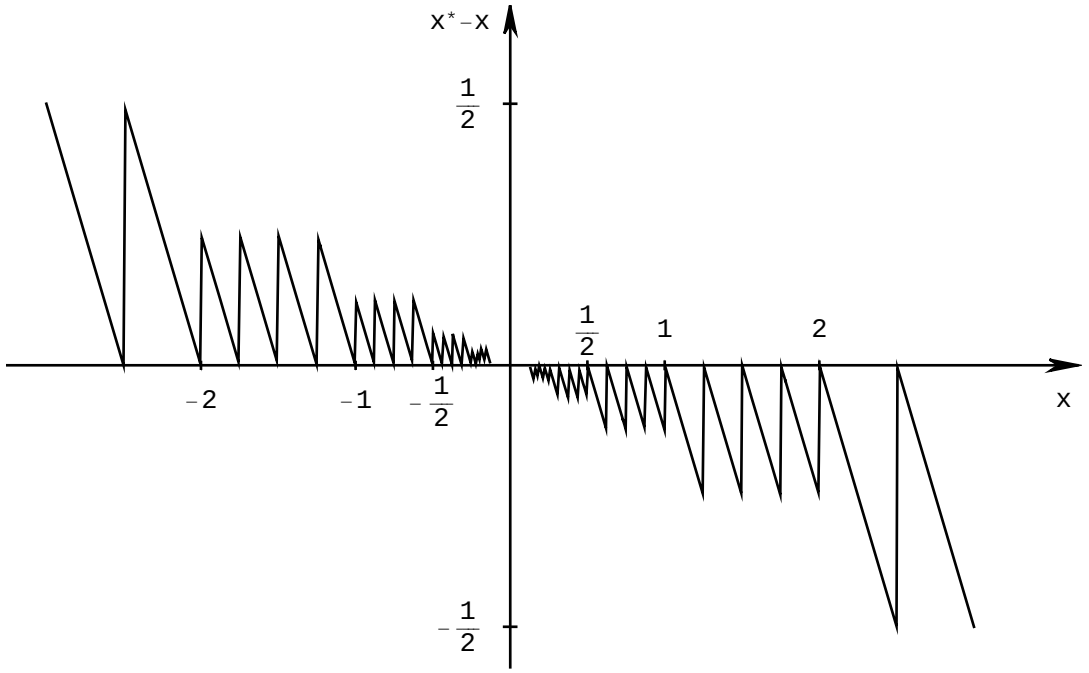
7. Вишеструка тачност

Скоро сви рачунари нуде поред "нормалног" кодирања у систему са покретном тачком и форме кодирања са већом мантисом (већом дужином мантисе) и са већим подручјем за експоненте. Иако дужине мантиса нису увек вишеструке дужине основне мантисе, ипак се говори о "двоструком", "троструком" итд. тачности.

Са становишта нумеричке математике могућност преласка на један шири скуп $\mathcal{M}'$ са мањим *eps* има значајне предности. Основно је да и у $\mathcal{M}'$ важе иста **квалитативна** ограничења и одступања од реалне математике као и у $\mathcal{M}$. Значајно побољшање би наступило када би се дужина k мантисе могла **динамички** мењати на основу критеријума који би се испробавали за време рачунања. Међутим, ово је тешко реализовати, а рачунари са таквим могућностима срећу се само у истраживачким центрима.

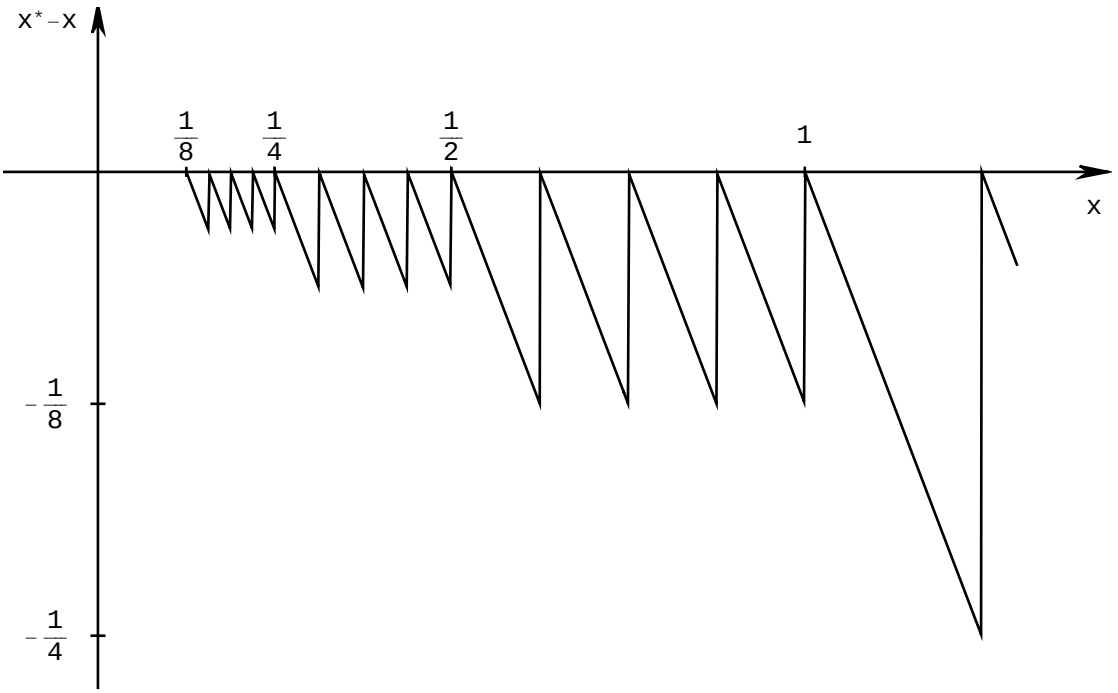
Вишеструка тачност доста се одомаћила због једноставне примене и користи се увек када је потребно постићи одређену тачност без већег анализирања самог нумеричког алгоритма. Ипак, губитак тачности наступа увек само на неколико критичних места у алгоритму, која није тешко открити. Брижљивије анализирање места, величина и операција које доприносе губитку тачности често доводи до исте коначне тачности као и рад са двоструком тачношћу, али са много мањим утрошком времена и мањим меморијским захтевима. На ово треба посебно пазити у случајевима када се праве програми.

За реалан број $x \in [-2.5, 2.5]$ и њему одговарајући машински број x^* из скупа $\mathcal{M}(3, 2, -3, 3)$, са одсецањем, посматра се разлика $x^* - x$.



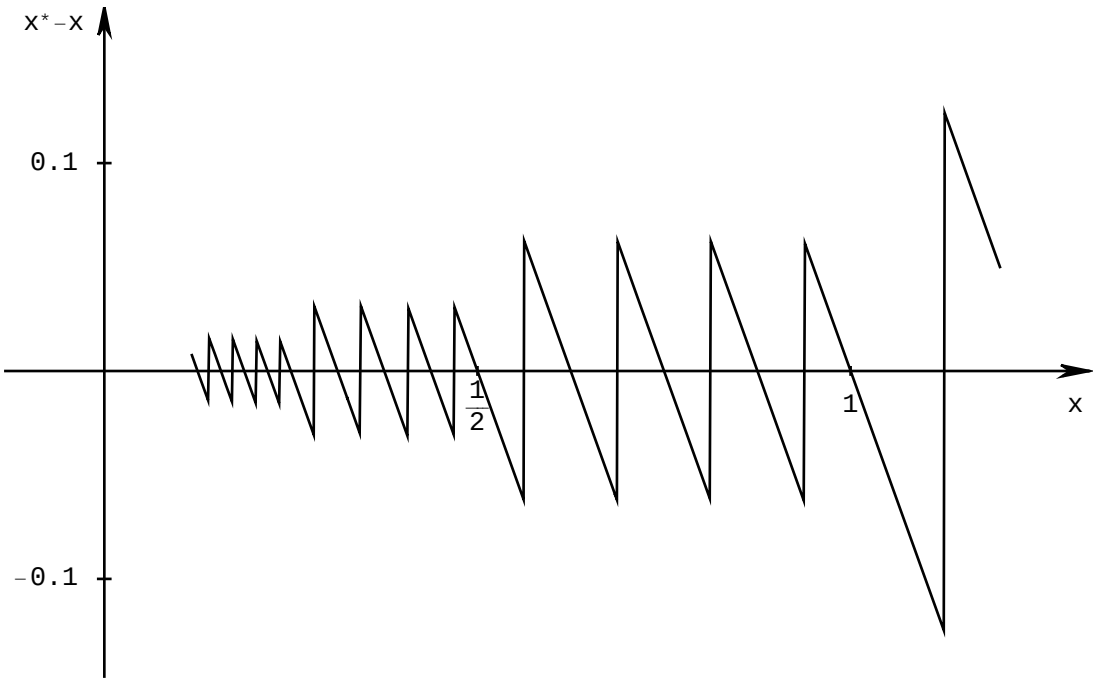
Слика 4.

За реалан број $x \in [0, 1.5]$ и њему одговарајући машински број x^* из скупа $\mathcal{M}(3, 2, -3, 3)$, са одсецањем, посматра се разлика $x^* - x$.



Слика 5.

За реалан број $x \in [0, 1.5]$ и њему одговарајући машински број x^* из скупа $\mathcal{M}(3, 2, -3, 3)$, са заокругљивањем, посматра се разлика $x^* - x$.



Слика 6.

Додатак Ц. АПРОКСИМАЦИЈЕ ФУНКЦИЈА ПОМОЋУ ДИФЕРЕНЦИЈАЛА

1. Увод

У претходном поглављу при разматрању грешака наглашено је да се у рачунару "уграђене" стандардне функције принципијелно не разликују од рационалних операција. Наиме, и функције и рационалне операције се замењују псеудооперацијама при имплементацији алгоритама у којима се појављују. При томе је област на којој су ове псеудооперације дефинисане само скуп машинских бројева. За функције које се не могу једноставно свести на рационалне операције или стандардне функције потребно је дати апроксимације помоћу којих ће се у оквиру машинске аритметике израчунавати приближне вредности тих функција. Значи, за функцију f одређује се функција F која ће је замењивати у рачунањима. Затим се за функцију F формира алгоритам за израчунавање вредности за аргументе из скупа машинских бројева и тај се алгоритам имплементира.

У овом поглављу нас првенствено интересује одређивање функције F у једној тачки или интервалу. Основни захтеви које ова функција треба да испуни су следећи:

- Вредности функције F у одређеном интервалу треба да се незнатно разликују од одговарајућих вредности функције f.

- Структура функције F требало би да буде подеснија за рад у односу на полазну функцију f. Ту се пре свега мисли на то да израчунавање вредности функције F треба да буде лакше од израчунавања вредности функције f.

Без сумње други захтев најбоље испуњавају полиноми, те постоји више начина за апроксимају функције f полиномима. Један од њих је интерполација, о чему ће бити више речи у следећем поглављу. Код апроксимације полиномима у принципу се посматра апроксимација функције у једном интервалу. Уз додатне захтеве, као што су захтев за равномерном грешком на целом интервалу и слично, могу се формирати специјални полиноми.

Сасвим друга врста апроксимације је локална апроксимација. Ту се захтева израчунавање приближне вредности функције у околини фиксне вредности аргумента $x = a$ и то тако да је грешка апроксимације у околини ове променљиве мала и да се веома мало мења када је x у тој околини. Због тога се као прво захтева да важи

$$f(a) = F(a).$$

Будући да је промена вредности функције одређена променом извода, може се очекивати да ће се разлика функција f и F лагано мењати при удаљавању од тачке $x = a$ ако се одговарајуће вредности извода поклапају, тј. ако је

$$f'(a) = F'(a).$$

Релативно једноставан начин за апроксимације вредности функције у тачки добија се коришћењем **диференцијала функције.**

2. Диференцијал функције

Уско везан са појмом извода функције је и појам диференцијала функције.

Дефиниција 1. Диференцијал функције

Функција $y = f(x)$ је диференцијабилна у тачки x_0 ако се њен прираштај у тој тачки

$$\Delta y = f(x_0 + \Delta x) - f(x_0)$$

може записати у облику

$$\Delta y = A\,\Delta x + \alpha(\Delta x)\,\Delta x,$$

где је A независно од Δx (тј. A је константа за дато x_0), а $\alpha(\Delta x) \to 0$ када $x \to x_0$.

Величина $A\,\Delta x$ представља "главни" део прираштаја функције и назива се **диференцијал функције**.

Диференцијал функције f у тачки x_0 је линеарна функција променљиве Δx, и означава се са dy или $df(x_0)$.

Пример 1.

Посматрајмо функцију $y = x^3$. Прираштај ове функције у тачки x_0 је

$$\Delta y = (x_0 + \Delta x)^3 - x_0^3 = 3x_0^2\,\Delta x + \left(3x_0\,\Delta x + (\Delta x)^2\right)\Delta x.$$

Коефицијент $A = 3x_0^2$ у прираштају функције не зависи од Δx, а други члан прираштаја

$$\alpha(\Delta x) = \left(3x_0\,\Delta x + (\Delta x)^2\right)\Delta x$$

очигледно тежи нули када $\Delta x \to 0$. Значи, $3x_0^2\,\Delta x$ је диференцијал функције $y = x^3$ у тачки x_0:

$$dy = 3x_0^2\,\Delta x \qquad \text{или} \qquad d\left(x_0^3\right) = 3x_0^2\,\Delta x.$$

Теорема 1. Диференцијабилност функције

Потребан и довољан услов да функција $f(x)$ буде диференцијабилна у тачки x_0 је да има први извод у тој тачки.

Дакле, ако је функција f диференцијабилна у тачки x_0, тј. ако постоји $f'(x_0)$, важи

$$f'(x_0) = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} = \lim_{\Delta x \to 0} \frac{A\,\Delta x + \alpha(\Delta x)\,\Delta x}{\Delta x} = A + \lim_{\Delta x \to 0} \alpha(\Delta x) = A.$$

Сада диференцијал можемо изразити као

$$dy = f'(x_0)\,\Delta x \qquad \text{или} \qquad df(x_0) = f'(x_0)\,\Delta x.$$

Извод функције $y = x$ једнак је јединици, па за ову функцију важи

$$dy = dx = \Delta x\,.$$

Сада можемо закључити да је **диференцијал независне променљиве** једнак његовом прираштају. Зато можемо писати

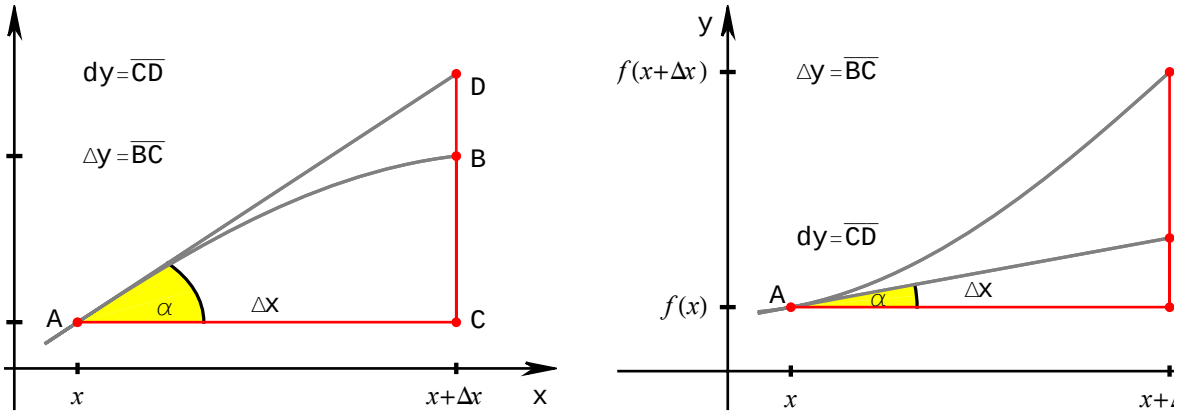
$$df(x) = f'(x)\,dx, \qquad \text{односно} \qquad \frac{df(x)}{dx} = f'(x)$$

или само

$$\frac{dy}{dx} = f'(x),$$

што значи да је извод функције једнак количнику диференцијала функције и диференцијала аргумента.

Погледајмо геометријску интерпретацију диференцијала функције (слика 1).



Слика 1.

Прираштај функције $y = f(x)$ је

$$\overline{\mathrm{BC}} = f(x + \Delta x) - f(x).$$

Из правоуглог троугла $\triangle$ACD је

$$\operatorname{tg}\alpha = f'(x) = \frac{\overline{\mathrm{CD}}}{\overline{\mathrm{AC}}} = \frac{\overline{\mathrm{CD}}}{\Delta x},$$

односно

$$\overline{\mathrm{CD}} = f'(x)\,\Delta x = \quad dy.$$

Дакле, диференцијал функције једнак је прираштају ординате тангенте функције у тачки x.

Ако променљива x представља време, а $y = f(x)$ је ордината тачке на правој у моменту x, тада је диференцијал једнак оној промени ординате, која би се добила у времену Δx ако би брзина тачке у временском интервалу $[x, x + \Delta x]$ била константна и једнака $f'(x)$. Када брзина тачке није константна, промена координате у времену Δx једнака је Δy. Дакле, у општем случају је $dy \neq \Delta y$.

Међутим, за довољно мало Δx прираштај функције и диференцијал функције су блиске величине. То се види из једнакости

$$\Delta y = f'(x)\,\Delta x + \alpha \Delta x .$$

Наиме, величине α и Δx теже нули кад $\Delta x \to 0$ па њихов производ брже тежи нули него Δx (када $\Delta x \to 0$). Уколико се $\alpha \Delta x$, као бесконачно мала величина, занемари, добија се апроксимација

$$\Delta y \approx f'(x)\,\Delta x ,$$

односно,

$$f(x + \Delta x) - f(x) \approx f'(x)\,\Delta x,$$

или

$$f(x + \Delta x) \approx f(x) + f'(x)\,\Delta x.$$

Последња релација даје могућност налажења приближне вредности функције, када је прираштај аргумента довољно мали.

Пример 2.

Замењујући прираштај функције њеним диференцијалом нађимо приближну вредност за $\sin 31^\circ$.

Посматрајмо функцију $y = \sin x$. Kako je

$$y(30^\circ) = \frac{1}{2}, \quad y'(30^\circ) = \cos 30^\circ = \frac{\sqrt{3}}{2}, \qquad 1^\circ = \frac{2\pi}{360} \approx 0.0175$$

na основу релације

$$f(x + \Delta x) \approx f(x) + f'(x)\,\Delta x$$

добијамо

$$\sin 31^\circ \approx \frac{1}{2} + \frac{\sqrt{3}}{2}\,\frac{2\pi}{360} \approx 0.5151.$$

Приближна вредност за $\sin 31^\circ$ добијена помоћу *Mathematica*-е је 0.515038.

Пример 3.

Одредићемо приближну вредност за $\sqrt[3]{1.1}$.

Нека је

$$y(x) = \sqrt[3]{x}, \qquad x_0 = 1, \qquad \Delta x = 0.1.$$

Сада је

$$y'(x) = \frac{1}{3\sqrt[3]{x^2}}, \qquad y'(1) = \frac{1}{3}$$

и

$$y(1 + 0.1) \approx y(1) + y'(1)\,0.1 = 1 + \frac{1}{3}\,0.1 = \frac{31}{30} \approx 1.033.$$

Приближна вредност за $\sqrt[3]{1.1}$ добијена помоћу *Mathematica*-е је 1.03228.

Ако се функција рачуна по тачној формули али је аргумент задат само приближно, тада, због $\Delta y \approx f'(x)\,\Delta x$ важи

$$|\Delta y| = |f(x + \Delta x) - f(x)| \approx |f'(x)| \cdot |\Delta x|.$$

Сада, стављајући најпре да је $x = x^*$, а затим да је $\Delta x = x - x^*$, добијамо

$$|f(x) - f(x^*)| \approx |f'(x^*)| \cdot |\Delta x|.$$

Ако је x^* приближна вредност за x, а $y^* = f(x^*)$ приближна вредност за $y = f(x)$, онда за приближну вредност границе апсолутне грешке можемо узети

$$A(y^*) = |f'(x^*)| \cdot |\Delta x|.$$

На основу тога, ако је $f(x^*) \neq 0$, добијамо приближну вредност границе релативне грешке

$$R(y^*) = \frac{A(y^*)}{|f(x^*)|} = \frac{|f'(x^*)| \cdot |\Delta x|}{|f(x^*)|}.$$

Ако је Δ_{x^*} граница апсолутне грешке приближне вредности x^*, онда се на основу приближних вредности граница апсолутне и релативне грешке могу добити линеарне оцене границе апсолутне грешке

$$A^0(y^*) = |f'(x^*)|\,\Delta_{x^*}$$

и линеарне оцене границе релативне грешке

$$R^0(y^*) = \frac{A^0(y^*)}{|f(x^*)|} = \frac{|f'(x^*)|\,\Delta_{x^*}}{|f(x^*)|}.$$

Очигледно је

$$A(y^*) \le A^0(y^*), \qquad R(y^*) \le R^0(y^*).$$

Пример 4.

Мерењем је добијена приближна вредност дужине страница квадрата $x^* = 46\,m$ са границом апсолутне грешке $\Delta_{x^*} = 0.1\,m$. Наћи ћемо линеарну оцену границе апсолутне грешке приближне вредности површине квадрата. Приближна вредност површине квадрата је $y^* = (x^*)^2$. Одатле је

$$A^0(y^*) = 2\,x^*\,\Delta_{x^*} = 9.2\,m^2.$$

За линеарну оцену границе релативне грешке добијамо

$$R^0(y^*) = \frac{A^*(y^*)}{y^*} = \frac{9.2}{46^2} \approx 0.0043,$$

односно 0.43%.

3. Неке особине диференцијала

Следеће особине диференцијала лако се доказује:

- Диференцијал константе једнак је нули:

$$y = c \quad \Rightarrow \quad dy = 0.$$

- Диференцијал линеарне функције једнак је њеном прираштају:

$$y = a\,x + b \quad \Rightarrow \quad dy = a\Delta x.$$

- Диференцијал степене функције:

$$y = x^n \quad \Rightarrow \quad dy = n\,x^{n-1}\,\Delta x.$$

- Диференцијал функције $c\,f(x)$, где је c константа:

$$y = c\,f(x) \quad \Rightarrow \quad dy = c\,df(x).$$

- Диференцијал збира неколико функција једнак је збиру њихових диференцијала:

$$y = f_1(x) + f_2(x) - f_3(x) \quad \Rightarrow \quad dy = df_1(x) + df_2(x) - df_3(x).$$

Израз $f'(x)\,\Delta x$ је диференцијал $df(x)$ функције f када се x посматра као променљива. Ако се и x посматра као функција неке друге променљиве t, израз $f'(x)\,\Delta x$, у општем случају, не мора бити диференцијал (што се може видети у следећем примеру). Изузетак од овога је случај када постоји линеарна зависност $x = a\,t + b$.

Пример 5.

Израз $2\,x\,\Delta x$ је диференцијал функције $y = x^2$ када је x променљива. Нека је сада $x = t^2$ и нека је t променљива. Тада је

$$y = x^2 = t^4$$

и

$$dy = 4\,t^3\,\Delta t.$$

Из $x = t^2$ следи

$$\Delta x = (t+\Delta t)^2 - t^2 = 2\,t\,\Delta t + (\Delta t)^2.$$

Значи,

$$2\,x\,\Delta x = 2\,t^2\left(2\,t\,\Delta t + (\Delta t)^2\right) = 4\,t^3\,\Delta t + 2\,t^2(\Delta t)^2.$$

.

Упоређујући $2\,x\,\Delta x$ са dy видимо да $2\,x\,\Delta x$ није диференцијал. Наиме, ове две величине разликују се за $2\,t^2(\Delta t)^2$.

Другачија је ситуација са $df(x) = f'(x)\,dx$. То је израз за диференцијал и када је x променљива и када је функција неке друге променљиве t, као што се види у следећем примеру. Ова особина диференцијала назива се **инваријантност** (неизмењеност).

Пример 6.

Израз $2\,x\,dx$ је диференцијал функције $y = x^2$ за сваку променљиву t када је x функција променљиве t. Посматраћемо случај $x = t^2$. Тада је

$$dx = 2\,t\Delta t$$

и

$$2\,x\,dx = 2\,t^2(2\,t\,\Delta t) = 4\,t^3\,\Delta t.$$

.

Упоређујући $2\,x\,dx$ са $dy = 4\,t^3\,\Delta t$ видимо да је $dy = 2\,x\,dx$.

С обзиром на инваријантност, диференцијали сложених функција могу се једноставно одређивати.

Пример 7.

Нађимо диференцијал функције $y = \left(1 + x^2\right)^4$. Посматраћемо ову функцију као сложену функцију: $y = u^4,\ u = 1 + x^2$. Имамо

$$dy = 4\,u^3\,du, \quad du = 2\,x\,dx.$$

Одатле је

$$dy = 4\left(1 + x^2\right)^3 2\,x\ dx = 8\,x\left(1 + x^2\right)^3 dx.$$

4. Апроксимационе формуле

Полазећи од релације

$$f(x+\Delta x) \approx f(x) + f'(x)\,\Delta x,$$

претпостављајући да је Δx довољно мало, могу се добити корисне формуле за релативно брзо рачунање приближних вредности функција.

Нека је $y = x^n$ и $x = 1$. Тада за $\Delta x = \alpha$ и $\Delta x = -\alpha$ добијамо респективно

$$(1+\alpha)^n \approx 1 + n\,\alpha \quad \text{и} \quad (1-\alpha)^n \approx 1 - n\,\alpha.$$

На основу тога, за различите вредности експонента n, добијају се нове апроксимације:

$$\frac{1}{1+\alpha} \approx 1-\alpha \qquad \frac{1}{1-\alpha} \approx 1+\alpha$$

$$\frac{1}{(1+\alpha)^2} \approx 1-2\,\alpha \qquad \frac{1}{(1-\alpha)^2} \approx 1+2\,\alpha$$

$$\sqrt{1+\alpha} \approx 1+\frac{\alpha}{2} \qquad \sqrt{1-\alpha} \approx 1-\frac{\alpha}{2}$$

$$\frac{1}{\sqrt{1+\alpha}} \approx 1-\frac{\alpha}{2} \qquad \frac{1}{\sqrt{1-\alpha}} \approx 1+\frac{\alpha}{2}$$

$$\sqrt[3]{1+\alpha} \approx 1+\frac{\alpha}{3} \qquad \sqrt[3]{1-\alpha} \approx 1-\frac{\alpha}{3}.$$

На сличан начин, полазећи од функција

$$\ln(1+x), \quad \ln(1-x), \quad e^x, \quad e^{-x}, \quad \sin x, \quad \cos x, \quad \operatorname{tg} x \quad \text{и} \quad b^x,\ b>0,$$

са $x = 0$ и $\Delta x = \alpha$ добијају се следеће апроксимације:

$$\ln(1+\alpha) \approx \alpha \qquad \ln(1-\alpha) \approx -\alpha$$

$$e^{\alpha} \approx 1+\alpha \qquad e^{-\alpha} \approx 1-\alpha$$

$$\sin\alpha \approx \alpha \qquad \cos\alpha \approx 1$$

$$\operatorname{tg}\alpha \approx \alpha \qquad b^{\alpha} \approx 1+\alpha \ln b.$$

Нека је

$$f(x) = \sqrt[n]{a^n + x}\,.$$

Тада је

$$f(0) = a$$

и

$$f'(0) = \frac{1}{n\,a^{n-1}}.$$

На основу овога, поступајући као и раније, добијамо

$$\sqrt[n]{a^n + x} \approx a + \frac{x}{n\,a^{n-1}}$$

за мале вредности x, односно када је $|x|$ довољно мало у односу на a.

Наведене формуле и њима сличне, које се могу извести на исти начин, могу се употребити за приближно рачунање вредности функција за које су оне прављене. У следећем примеру то је показано за последњу формулу.

Пример 8.

Израчунаћемо приближну вредност за $\sqrt[7]{130}$. Користећи претходну релацију са $n = 7$, $a = 2$ и $x = 2$ добијамо

$$\sqrt[7]{130} \approx 2 + \frac{2}{7 \cdot 64} \approx 2.0045.$$

Приближна вредност за $\sqrt[7]{130}$ добијена помоћу *Mathematica*-е је 2.00446.

Задаци

1. Нађи диференцијал функције $y = f(x) = x^2 - x + 3$ у тачки $x_0 = 2$:
   а) издвајањем линеарног дела у односу на Δx у прираштају функције Δy,
   б) према $dy = f'(x)\,dx$.

> **Решење:** а) $dy = 3\,\Delta x$, б) $dy = 3\,dx = 3\,\Delta x$.

2. Нађи диференцијал функције y у тачки x ако је:

а) $y = \dfrac{1}{x}$, б) $y = x\,e^{2x}$, ц) $y = x\sin x + \cos x$.

3. Нађи диференцијал функције $y = \sin x^2$:

а) у тачки $x = x_0$, б) у тачки $x = \sqrt{\pi}$, ц) у тачки $x = \sqrt{\pi}$, за $dx = -2$.

> **Решење:** а) $dy = 2\,x_0 \cos x_0^2\,dx$, б) $dy = -2\sqrt{\pi}\,dx$, ц) $dy = 4\sqrt{\pi}$.

4. Помоћу диференцијала израчунај приближне вредности за

а) $\cos 151^\circ$, б) $\operatorname{arctg} 1.1$, ц) $e^{0.2}$.

> **Решење:** а) $-\dfrac{\sqrt{3}}{2} - \dfrac{\pi}{360} \approx -0.874752$,
>
> б) $\dfrac{1}{20} + \dfrac{\pi}{4} \approx 0.835398$,
>
> ц) $\dfrac{6}{5} = 1.2$.
>
> Одговарајуће вредности добијене помоћу *Mathematica*-е су:
>
> а) -0.87462, б) 0.835398, ц) 1.2214.

5. Нађи приближне вредности за:

а) $\sqrt[3]{9}$, б) $\sqrt{255}$.

> **Решење:** а) $\dfrac{25}{12} \approx 2.08333, \left(y = \sqrt[3]{x},\ x_0 = 2,\ \Delta x = 1\right)$,
>
> б) $\dfrac{511}{32} \approx 15.9688, \left(y = \sqrt{x},\ x_0 = 256,\ \Delta x = -1\right)$.
>
> Одговарајуће вредности добијене помоћу *Mathematica*-е су:

а) 2.08008 б) 15.9687.

6. Нека је $s = 2t^2 + t + 1$.Нађи прираштај и диференцијал за s у тачки $t = 1$ и упореди их за:

а) $\Delta t =$ 0.1, б) $\Delta t =$ 0.2, ц) $\Delta t =$ 1.

Решење: $\Delta s = 5\,\Delta t + 2\,\Delta t^2$, $ds = 5\,\Delta t$,

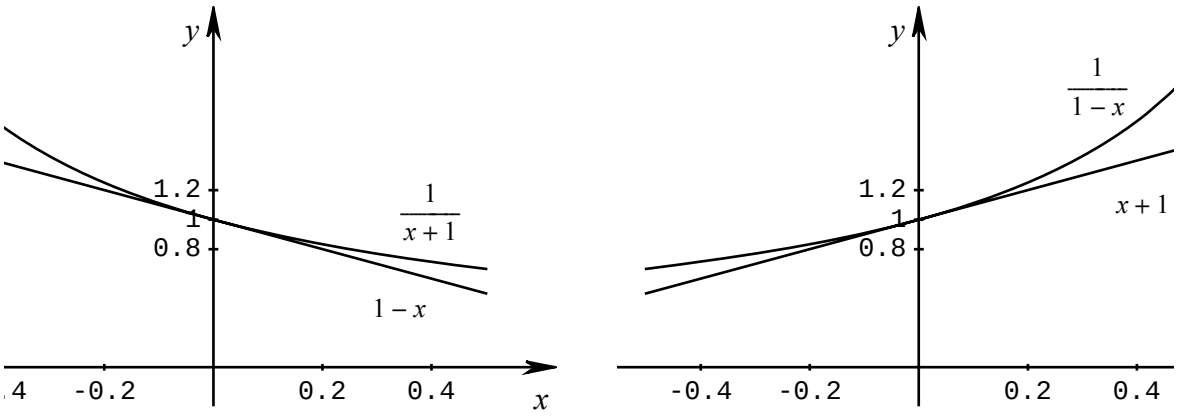
а) $\Delta s = 0.52$, $ds = 0.5$, б) $\Delta s = 1.08$, $ds = 1$, ц) $\Delta s = 7$, $ds = 5$.

7. Нека је $y = \sin x$ и $x = \cos t$. Да ли је тачно:

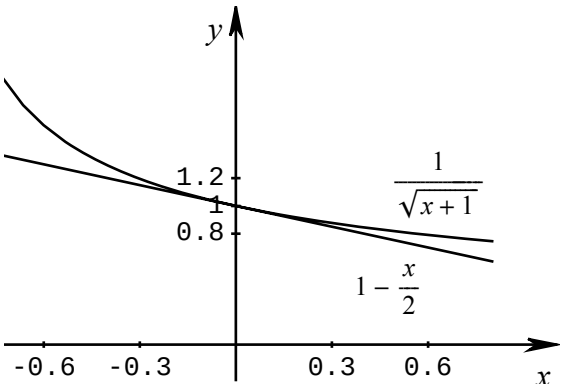
а) $dy =$ 0, за $t = \frac{\pi}{2}$, б) $dy = dx$, за $t = \frac{\pi}{2}$, ц) $dy = -dt$, за $t = \frac{\pi}{2}$?

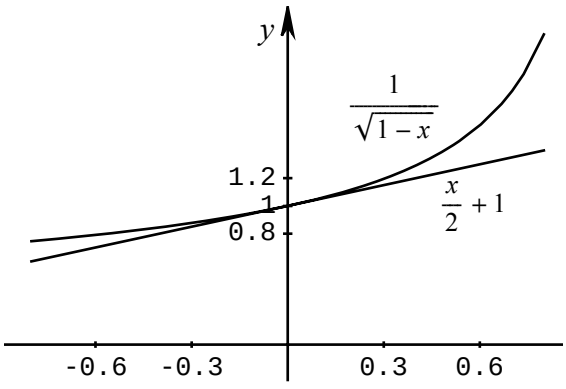
Решење: а) није б) јесте, ц) јесте.

Апроксимације помоћу диференцијала.



Слика 2.





x	$\sqrt{x+1}$	$\frac{x}{2}+1$	$\left\lvert-\frac{x}{2}+\sqrt{x+1}-1\right\rvert$
–0.5000	0.707107	0.750000	4.29×10^{-2}
–0.4375	0.750000	0.781250	3.13×10^{-2}
–0.3750	0.790569	0.812500	2.19×10^{-2}
–0.3125	0.829156	0.843750	1.46×10^{-2}
–0.2500	0.866025	0.875000	8.97×10^{-3}
–0.1875	0.901388	0.906250	4.86×10^{-3}
–0.1250	0.935414	0.937500	2.09×10^{-3}
–0.0625	0.968246	0.968750	5.04×10^{-4}
0.0000	1.000000	1.000000	0.00×10^{0}
0.0625	1.030776	1.031250	4.74×10^{-4}
0.1250	1.060660	1.062500	1.84×10^{-3}
0.1875	1.089725	1.093750	4.03×10^{-3}
0.2500	1.118034	1.125000	6.97×10^{-3}
0.3125	1.145644	1.156250	1.06×10^{-2}
0.3750	1.172604	1.187500	1.49×10^{-2}
0.4375	1.198958	1.218750	1.98×10^{-2}
0.5000	1.224745	1.250000	2.53×10^{-2}
0.5625	1.250000	1.281250	3.13×10^{-2}
0.6250	1.274755	1.312500	3.77×10^{-2}
0.6875	1.299038	1.343750	4.47×10^{-2}
0.7500	1.322876	1.375000	5.21×10^{-2}
0.8125	1.346291	1.406250	6.00×10^{-2}
0.8750	1.369306	1.437500	6.82×10^{-2}
0.9375	1.391941	1.468750	7.68×10^{-2}
1.0000	1.414214	1.500000	8.58×10^{-2}

Табела 1.

Literatura

[1] Ascher, U.M., Matheij, R.M.M., Russell, R.D., Numerical solution of boundary value problems for ordinary differential equations, SIAM, Philadelphia, 1995.

[2] Björck, A., Dahlquist, G., Numerische Methoden, Oldenbourg Verlag, München, Wien, 1979.

[3] Bohl, E., Finite Modelle gewohnlicher Randwertaufgaben, Teubner, Stuttgart, 1981.

[4] Brosowski, B.,Kreß, R., Einführung in die Numerische Mathematik I, II, Bibliographisches Institut, Mannheim, Wien, Zürich, 1975.

[5] Collatz, L., The Numerical Treatment of Differential Equations, Springer Verlag, Berlin, Heidelberg, New York, 1966

[6] Davis, P.J., Interpolation and approximation. Blaisdell, New York, 1963.

[7] Demidovich, B.P., Maron I.A.S., Computational Mathematics, Mir, Moskva, 1973.

[8] Hagender, N., Sundblad, Y., Aufgabensammlung Numerische Methoden, 2. Aufgaben, Oldenbourg Verlag, München Wien, 1972.

[9] Henrici, P., Discrete Variable Methods in Ordinary Differential Equantions, New York, London, 1962.

[10] Henrici, P., Elements of numerical analysis, John Wiley and Sons, New York, 1964.

[11] Herceg, D., Numeričke i statističke metode za obradu eksperimentalnih podataka, Institut za matematiku, Novi Sad, 1992.

[12] Herceg, D., Numerička analiza za IV razred srednjeg obrazovanja i vaspitanja srednjeg stupnja prirodno-matematičke struke, Zavod za izdavanje udžbenika, Novi Sad, Naučna knjiga, Beograd, 1990.

[13] Herceg, D., Stojaković, Z., Numeričke metode linearne algebre - zbirka zadataka, Građevinska knjiga, Beograd, 1988.

[14] Herceg, D., Herceg, Đ., Elementi numeričke analize i Mathematica, Symbol, Novi Sad, 2008.

[15] Herceg, D., Herceg, Đ., Numerička matematika, Stylos, Novi Sad, 2003.

[16] Herceg, D., Vulanović, R., Numerička matematika, teorija, problemi i programiranje. VŠOR, Novi Sad, 1994.

[17] Herceg, D., Krejić, N., Numerička analiza. Univerzitet u Novom Sadu, Stylos, Novi Sad, 1997.

[18] Herceg, D., Krejić, N., Numerička analiza. Zbirka zadataka I. Univerzitet u Novom Sadu, Institut za matematiku, Novi Sad, 1998.

[19] Herceg, D., Krejić, N., Numerička analiza. Zbirka zadataka II. Univerzitet u Novom Sadu, Institut za matematiku, Novi Sad, 1998.

[20] Isaacon, E., Keller, H.B., Analysis of numerical methods, John Wiley & sons, Inc., New York, London, Sydney, 1966.

[21] Jordan-Engeln, G., Reutter, F., Numerische Mathematik für Ingenieure, Bibliograph. Inst. Mannheim, 1973.

[22] Mathews, J.H., Numerical Methods for Computer Science, Engineering and Mathematics, Prentice-Hall, inc. Englewood Cliffs, New Jersey, 1987.

[23] Ortega, J.M., Rheinboldt, W.C., Iterative Solution of Nonlinear Equations of Several Variables, Academic Press, New York, 1970.

[24] Philips, G.M., Taylor, P.J., Theory and applications of numerical analisys, Academic Press, London, 1996.

[25] Stetter, H.J., Numerik für Informatiker, R. Oldenbourg Verlag, München, Wien, 1976.

[26] Stoer, J., Einführung in die Numerische Mathematik I, Springer Verlag, Berlin; Heidelberg, New York, 1979.

[27] Stoer, J., Bulirsch, R., Einführung in die Numerische Matematik II, Springer Verlag, Berlin; Heidelberg, New York, 1979.

[28] Stojaković, Z., Herceg, D., Numeričke metode linearne algebre, Građevinska knjiga, Beograd, 1988.

[29] Stummel, F., Hainer, K., Praktische Mathematik, Teubner, Stuttgart, 1971.

[30] Ueberhuber, C.W., Numerical Computation I, II, Springer, Berlin, 1997.

[31] Vilenkin, N.J., Methoda sukcesivnih aproksimacija, Školska knjiga, Zagreb, 1980.