Gradimir V. Milovanović

NUMERIČKA ANALIZA I TEORIJA APROKSIMACIJA

UVOD U NUMERIČKE PROCESE I REŠAVANJE JEDNAČINA

Zavod za udžbenike

Recenzent
Prof. GIUSEPPE MASTROIANNI
University of Basilicata
Department of Mathematics and Computer Sciences
Potenza, Italy

Urednici
prof. dr LJUBOMIR PROTIĆ
dr MILOLJUB ALBIJANIĆ

Odgovorni urednik
dr UROŠ ŠUVAKOVIĆ

Za izdavača
DRAGOLJUB KOJČIĆ, v.d. direktora i glavnog urednika

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

519.61/.65(075.8)

МИЛОВАНОВИЋ, Градимир В., 1948-
Numerička analiza i teorija aproksimacija : uvod u numeričke procese i rešavanje jednačina / Gradimir V. Milovanović. - 1. izd. - Beograd : Zavod za udžbenike, 2014 (Beograd : Službeni glasnik). - XII, 456 str. : graf. prikazi ; 24 cm

Tiraž 700. - Bibliografija na kraju svakog poglavlja. - Registar.

ISBN 978-86-17-18875-5

a) Нумеричка анализа

COBISS.SR-ID 209318924

Mojoj ćerki IRENI

Predgovor

Prve knjige na srpskom jeziku iz oblasti numeričke matematike bili su prevodi, sa engleskog i ruskog jezika, poznatih klasičnih knjiga:

– E. WHITTAKER, G. ROBINSON, *Tečaj numeričke matematike*, Naučna knjiga, Beograd, 1955;

– I.S. BEREZIN, N.P. ŽITKOV, *Numerička analiza – numeričke metode*, Naučna knjiga, Beograd, 1963.

Petnaestak godina kasnije pojavljuju se i prve knjige domaćih autora. Do 1980. godine pubikovane su tri knjige:

– M. BERTOLINO, *Numerička analiza*, Naučna knjiga, Beograd, 1977;

– V. SIMONOVIĆ *Numeričke metode – skripta*, Mašinski fakultet, Beograd, 1979;

– G.V. MILOVANOVIĆ, *Numerička analiza, I deo*, Univerzitet u Nišu, Niš, 1979.

Drugi deo prethodne autorove knjige pojavljuje se dve godine kasnije pod naslovom:

– G.V. MILOVANOVIĆ, *Numerička analiza – Diferencijalne i integralne jednačine*, Institut za dokumentaciju zaštite na radu „Edvard Kardelj“ Centar za informativno izdavačku delatnost – Birotehnika, Niš, 1981.

Ove dve knjige su bile osnov za izradu mnogo kompletnijih knjiga – udžbenika, u tri toma, sa pratećom zbirkom problema, sve na preko 1100 stranica, po kojima su učile mnogobrojne generacije studenata širom bivše Jugoslavije:

– G.V. MILOVANOVIĆ, *Numerička analiza, I deo*, Naučna knjiga, Beograd, 1985 (drugo izdanje 1988; treće izdanje 1991);

– G.V. MILOVANOVIĆ, *Numerička analiza, II deo*, Naučna knjiga, Beograd, 1985 (drugo izdanje 1988; treće izdanje 1991);

– G.V. MILOVANOVIĆ, *Numerička analiza, III deo*, Naučna knjiga, Beograd, 1988 (drugo izdanje 1991);

– G.V. MILOVANOVIĆ, M.A. KOVAČEVIĆ, *Zbirka rešenih zadataka iz numeričke analize*, Naučna knjiga, Beograd, 1985 (drugo izdanje 1988; treće izdanje 1991).

Poslednja *Zbirka zadataka* je pre desetak godina ponovo objavljena, uz znatna proširenja,

– G.V. MILOVANOVIĆ, M.A. KOVAČEVIĆ, M.M. SPALEVIĆ, *Numerička matematika – Zbirka rešenih problema*, Elektronski fakultet u Nišu, Niš, 2003.

Numerička matematika je poslednjih dvadesetak godina doživela ekspanziju, zahvaljujući mnogobrojnim primenama u skoro svim oblastima nauke i inženjerstva, s jedne strane, i burnim razvojem računarskih arhitektura, s druge strane. Skorašnji progres u simboličkom izračunavanju i razvoj visokokvalitetnog softvera u mnogim domenima dodatno ubrzava pomenutu ekspanziju.

Numerička matematika se oslanja na više drugih matematičkih disciplina od kojih vidno mesto zauzimaju linearna algebra, funkcionalna analiza, teorija aproksimacija, optimizacija, teorija operatora, itd. Posebno, teorija aproksimacija je do te mere povezana i isprepletena sa numeričkom matematikom da ih nije moguće danas jasno razgraničiti.

Sve je to izazvalo i pojavu novih pravaca istraživanja koji su se postepeno utemeljili u posebne naučne discipline (npr. teorija ortogonalnosti, teorija splajnova i talasića, naučna izračunavanja, geometrijsko modeliranje, obrada signala, paralelna izračunavanja, itd.). U svetu se u poslednje vreme pojavio veliki broj knjiga i monografija, uglavnom na engleskom jeziku, koje tretiraju, svaka na svoj način, razne probleme iz pomenutih oblasti. Nedavno se pojavila i monografija iz oblasti teorije aproksimacija sa posebnim osvrtom na interpolacione procese i primene:

– G. MASTROIANNI, G.V. MILOVANOVIĆ, *Interpolation Processes – Basic Problems and Applications*, Springer Monographs in Mathematics, Springer Verlag, Berlin – Heidelberg – New York, 2008.

Nepostojanje odgovarajuće literature na srpskom jeziku, navela je autora da napiše seriju od nekoliko knjiga monografskog karaktera sa ovakvim sadržajem i pod opštim zajedničkim naslovom *Numerička analiza i teorija aproksimacija*, dok će se konkretan sadržaj knjige definisati podnaslovom.

Prva od tih knjiga je upravo ova sa podnaslovom *Uvod u numeričke procese i rešavanje jednačina*, čiji je zadatak da čitaoca uvede i upozna sa tzv. aritmetikom konačne dužine i osnovnim principima numeričkih procesa, uključujući rekurzivna izračunavanja i sumiranja, kao i sa rešavanjem nelinearnih jednačina i sistema (linearnih i nelinearnih) jednačina.

Knjiga je podeljena u pet glava. Sve glave su podeljene na poglavlja, a poglavlja na odeljke. Numeracija objekata (formula, teorema, definicija i sl.) u okviru jednog odeljka izvršena je pomoću tri broja od kojih prvi ukazuje na poglavlje, drugi na odeljak, a treći na redni broj tog objekta u odeljku. Na ovaj način uspostavljena je jednoznačna numeracija objekata u okviru jedne glave. Kraj dokaza pojedinih tvrđenja označen je standardnom oznakom $\square$, a kraj teksta u okviru primera je označen sa $\triangle$.

U posebnom odeljku svake glave dat je spisak citirane i korišćene literature.

Dodatni podaci o mnogim činjenicama, kao i korisne informacije o autorima mnogih metoda i pristupa u numeričkoj analizi i teoriji aproksimacija, a koji se spominju u tekstu, navedeni su u fusnotama.

U prvoj glavi ove knjige, pored analize uticaja aritmetike konačne dužine na numeričke procese i analize uslovljenosti i stabilnosti izračunavanja, dajemo bazična rekurzivna izračunavanja, uvod u sumacione procese, teoriju verižnih razlomaka i asimptotske razvoje, kao i metode za efikasno izračunavanje elementarnih funkcija i nekih specijalnih funkcija. Posebno je analiziran algoritam za efikasno izračunavanje gama funkcije za kompleksne vrednosti argumenta.

Druga glava je posvećena osnovnim elementima funkcionalne analize i linearne algebre, koji su neophodni za praćenje izlaganja u preostalim glavama ove knjige. Međutim, za uspešno praćenje izlaganja u preostalim knjigama iz ove serije, neohodan je dodatni matematički aparat i on će biti sastavni deo tih knjiga.

Opšta teorija iterativnih procesa se razmatra u trećoj glavi, uključujući opšte karakteristike iterativnih procesa, kao i metode za ubrzavanje procesa. Četvrta glava je posvećena numeričkim metodama u linearnoj algebri, gde se izučavaju kako direktni tako i iterativni metodi. Posebna pažnja je posvećena problemu sopstvenih vrednosti matrica. Najzad, u petoj glavi se najpre izučavaju metodi za nelinearne jednačine, a zatim i za nelinearne sisteme jednačina. Posebna pažnja je posvećena homotopskim metodama rešavanja sistema jednačina, koristeći princip neprekidnosti. Zbog svoje važnosti, algebarske jednačine se razmatraju u posebnom poglavlju. Kao komplementarna literatura za ovo poglavlje se preporučuje autorova knjiga:

– G.V. MILOVANOVIĆ, *Ektremalni problemi i nejednakosti za polinome*, Zavod za udžbenike, Beograd, 2012.

Na kraju knjige dat je indeks imena i pojmova.

Druga knjiga iz pomenute serije knjiga, sa podnaslovom *Ortogonalni sistemi*, biće posvećena modernoj teoriji ortogonalnosti, sa posebnim osvrtom na razne

koncepte ortogonalnosti, kao i na konstruktivnu teoriju ortogonalnih polinoma koja je razvijena poslednjih decenija prošlog veka.

U trećoj knjizi biće razmatrani razni aproksimacioni problemi, uključujući interpolacione procese. Kvadraturni procesi i primene u integralnim jednačinama biće predmet četvrte knjige, itd.

Ova knjiga, kao i ostale iz ove serije, su namenjene studentima prirodno-matematičkih i tehničkih fakulteta, mladim istraživačima, kao i već afirmisanim naučnicima u oblasti numeričke matematike, teorije aproksimacija, matematičkih optimizacija, i svima onima u oblasti prirodnih i tehničkih nauka u čijim istraživanjima se pojavljuju odgovarajući numerički i aproksimacioni problemi.

Tokom pripreme ove knjige delove rukopisa su čitali moji saradnici dr Marija P. Stanić, vanredni profesor Prirodno-matematičkog fakulteta u Kragujevcu, i dr Nenad P. Cakić, redovni profesor Elektrotehničkog fakulteta u Beogradu, i dali niz sugestija koje su uzete u obzir kod sastavljanja konačne verzije knjige. Koristim ovu priliku da im se najsrdačnije zahvalim na uloženom trudu.

Takođe, želim da se zahvalim Zavodu za udžbenike i uredniku matematičkih izdanja dr Miloljubu Albijaniću za stalno interesovanje i ohrabrenje da se istraje na izradi ove knjige.

Beograd, juni 2014. *Gradimir V. Milovanović*

Sadržaj

1. UVOD U NUMERIČKE PROCESE

1.1 UVODNE NAPOMENE

1.1.1 Zadatak numeričke matematike i teorije aproksimacija

Numerička analiza i *teorija aproksimacija* su matematičke oblasti koje se uzajamno prožimaju i koje su u manjoj ili većoj meri povezane sa skoro svim drugim matematičkim oblastima. Njihov glavni cilj je da obezbede rešavanje najsloženijih matematičkih problema kakve danas postavljaju savremena nauka i tehnika. Takvi problemi mogu biti formulisani na najrazličitije načine, na primer, u terminima algebarskih ili transcendentnih jednačina, običnih ili parcijalnih diferencijalnih jednačina, integralnih jednačina, itd. ili pak kao skup takvih jednačina. Čak i u najjednostavnijim slučajevima kada su analitička rešenja moguća, ona su vrlo često neupotrebljiva zbog svoje glomaznosti. U tim slučajevima je često prihvatljivije približno (aproksimativno) rešenje, čija se tačnost može dirigovati, tj. greška aproksimacije se može, zavisno od zahteva i potrebe, učiniti proizvoljno malom veličinom.

Poznato je, na primer, da se algebarske jednačine u opštem slučaju mogu rešiti pomoću radikala, samo kada su stepena ne višeg od četvrtog, pri čemu su već za jednačine trećeg stepena dobro poznate CARDANOve[1] formule prilično komplikovane. Međutim, u praksi se vrlo često sreće problem rešavanja numeričkih algebarskih jednačina[2] visokog stepena. Takođe, često se javlja i problem rešavanja sistema linearnih algebarskih jednačina sa stotinu, pa i više hiljada nepoznatih, gde klasični aparat linearne algebre postaje neupotrebljiv. Standardna teorija diferencijalnih jednačina omogućava rešavanje (nalaženje opštih rešenja) samo nekih uskih klasa diferencijalnih jednačina. Međutim, u praksi se javljaju

[1] GEROLAMO CARDANO (1501 – 1576) je bio čuveni lekar, astrolog, filozof i matematičar, koji je živeo u Milanu.

[2] Algebarska jednačina sa numeričkim koeficijentima.

CAUCHYevi[3] i konturni problemi sa diferencijalnim jednačinama koje ne pripadaju ovim klasama. Posebno važnu ulogu imaju problemi koji se svode na rešavanje parcijalnih diferencijalnih jednačina.

S druge strane, čak i u najelementarnijim problemima, rešenje koje je izraženo simbolički ne zadovoljava potrebe ako se traži numerički rezultat. Ilustrujmo ovo jednostavnim primerom.

Neka je potrebno odrediti pozitivan koren jednačine

$$x^2 - a = 0 \qquad (a > 0).$$

Traženo egzaktno rešenje je $x = \sqrt{a}$. Međutim, simbol $\sqrt{\ }$ ne rešava problem jer ne daje postupak za izračunavanje broja $\sqrt{a}$.

Formirajmo sada niz $\{x_n\}_{n\in\mathbb{N}}$ pomoću

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right), \qquad n = 0, 1, \ldots, \tag{1.1.1}$$

gde je $x_0 = a$. U standardnim udžbenicima matematičke analize dokazuje se konvergencija ovog niza ka vrednosti $\sqrt{a}$, kada $n \to +\infty$ (videti, na primer, [53, str. 47–49]).

Formula (1.1.1) predstavlja tzv. *iterativni proces* i specijalni je slučaj poznatog NEWTONovog[4] metoda. Za član niza x_n kažemo da je n-ta iteracija. Ako sa e_n označimo grešku u n-toj iteraciji, tj. $e_n = x_n - \sqrt{a}$, jednostavno se može dokazati da važi jednakost

$$e_{n+1} = \frac{1}{2x_n} e_n^2, \tag{1.1.2}$$

odakle zaključujemo da je greška u $(n+1)$-oj iteraciji srazmerna kvadratu greške u n-toj iteraciji. Prilikom određivanja kvadratnog korena iz broja a pomoću (1.1.1), iterativni proces se prekida kada je greška u n-toj iteraciji manja od unapred izabrane tačnosti ε i tada se uzima $\sqrt{a} \cong x_n$.

Primenimo sada iterativni proces (1.1.1) na određivanje vrednosti $\sqrt{2}$. Dakle, startujući sa $x_0 = a = 2$, dobijamo niz

[3] AUGUSTIN LUIS CAUCHY (1789 – 1857), veliki francuski matematičar, utemeljivač matematičke i kompleksne analize, kao i niza drugih oblasti.

[4] ISAAC NEWTON (1642 – 1727), veliki engleski fizičar, matematičar, astronom i filozof. Poznat je po nizu doprinosa u matematici, posebno kao jedan je od tvoraca infinitezimalnog računa, ali i po velikim doprinosima u mehanici, optici, itd.

$$\left\{\frac{3}{2}, \frac{17}{12}, \frac{577}{408}, \frac{665857}{470832}, \frac{886731088897}{627013566048}, \frac{1572584048032918633353217}{1111984844349868137938112}, \cdots\right\},$$

čiji članovi (iteracije), izračunati sa 50 decimala, imaju sledeće vrednosti

$$\begin{aligned}
x_0 &= 2,\\
x_1 &= \mathbf{1}.5,\\
x_2 &= \mathbf{1.41}66666666\ 6666666666\ 6666666666\ 6666666666\ 6666666666\ \ldots,\\
x_3 &= \mathbf{1.4142}56862\ 7450980392\ 1568627450\ 9803921568\ 6274509803\ \ldots,\\
x_4 &= \mathbf{1.4142135623\ 7}468991062\ 6295578890\ 1349101165\ 5962211574\ \ldots,\\
x_5 &= \mathbf{1.4142135623\ 7309504880\ 168}9623502\ 5302436149\ 8192577619\ \ldots,\\
x_6 &= \mathbf{1.4142135623\ 7309504880\ 1688724209\ 6980785696\ 7187537}723\ \ldots,
\end{aligned}$$

pri čemu se boldirane (zacrnjene) cifre poklapaju sa odgovarajućim ciframa u tačnoj vrednosti

$$\sqrt{2} = 1.4142135623\ 7309504880\ 1688724209\ 6980785696\ 7187537694\ \ldots\ .$$

Iteracije $x_1, \ldots, x_6$ imaju redom jednu, tri, šest, 12, 24 i 48 tačnih cifara, a odgovarajuće greške e_n su redom

$$8.58\times10^{-2},\ 2.45\times10^{-3},\ 2.12\times10^{-6},\ 1.60\times10^{-12},\ 8.99\times10^{-25},\ 2.86\times10^{-49},$$

što je u skladu sa formulom (1.1.2).

Navedimo sada još jedan primer. Neka je potrebno odrediti vrednost $\sin x$ za $x = 0.5$. Simbol sin, i u ovom slučaju, ne daje postupak za rešavanje problema. Da bismo izračunali traženu vrednost, funkciju $x \mapsto \sin x$ možemo da razvijemo, na primer, u TAYLORov[5] razvoj

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots,$$

odakle, uzimanjem dva, odnosno tri člana u razvoju, dobijamo

$$\sin 0.5 \cong 0.5 - \frac{0.5^3}{3!} = 0.479167$$

i

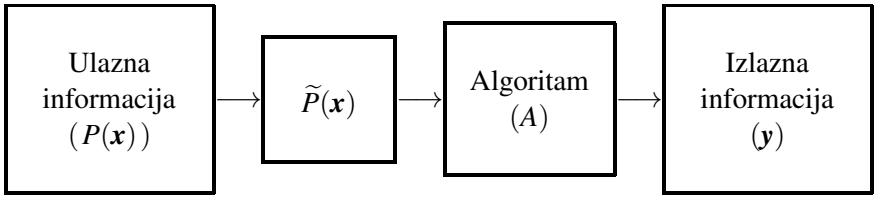
[5] BROOK TAYLOR (1685 – 1731), engleski matematičar.

$$\sin 0.5 \cong 0.5 - \frac{0.5^3}{3!} + \frac{0.5^5}{5!} = 0.479427\,.$$

Tačna vrednost za $\sin 0.5$ sa šest decimala je 0.479425. Greške koje ovde nastaju potiču usled uzimanja konačnog broja članova TAYLORovog reda i nazivamo ih *greškama odsecanja*. Drugim rečima, greška nastaje iz razloga što rešavamo problem različit od postavljenog, koji je znatno jednostavniji sa stanovišta računanja, a čije je rešenje u nekom smislu blisko rešenju postavljenog problema.

Neka je, u opštem slučaju, potrebno rešiti zadati problem sa datim ulaznim podacima, koje ćemo označiti sa $\boldsymbol{x}$, a sam problem sa $P(\boldsymbol{x})$. Na primer, ulazni podaci mogu biti definisani n-dimenzionalnim vektorom $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Problem $P(\boldsymbol{x})$ zvaćemo ulaznom informacijom, a odgovarajući rezultat $\boldsymbol{y}$, tj. rešenje problema $P(\boldsymbol{x})$, zvaćemo izlaznom informacijom. Na primer, izlazna informacija može biti m-dimenzionalni vektor $\boldsymbol{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m$. U postupku rešavanja, problem $P(\boldsymbol{x})$ se najčešće zamenjuje (aproksimira) nekim „bliskim“ problemom $\widetilde{P}(\boldsymbol{x})$.

Postupak transformacije modifikovane ulazne informacije $\widetilde{P}(\boldsymbol{x})$ u izlaznu informaciju $(\boldsymbol{y})$ zvaćemo *algoritmom* (A). Navedena transformacija može se predstaviti blok dijagramom



ili simbolički kao $P(\boldsymbol{x}) \cong \widetilde{P}(\boldsymbol{x}) \xrightarrow{A} \boldsymbol{y}$. Ponekad se pod algoritmom podrazumeva i prethodna aproksimacija problema $P(\boldsymbol{x})$ u modifikovani problem $\widetilde{P}(\boldsymbol{x})$, tako da u tom slučaju imamo transformaciju $P(\boldsymbol{x}) \xrightarrow{A} \boldsymbol{y}$.

Pri rešavanju nekog problema potrebno je izabrati pogodan algoritam koji najbrže dovodi do željenog rezultata. Izbor optimalnog algoritma u prethodnom smislu predstavlja vrlo složen problem, koji teorijski, u opštem slučaju, još uvek nije rešen.

Razradom i realizacijom algoritama i analizom greške u izlaznoj informaciji bavi se posebna oblast matematike, tzv. *numerička matematika*. Njen centralni deo čine numerički metodi, koji moraju biti takvi da su pogodni sa stanovišta primene digitalnih računara. Kako računari najčešće izvode samo četiri osnovne računske operacije (sabiranje, oduzimanje, množenje i deljenje), to numerički metodi moraju biti takvi da se svode na konačan niz takvih operacija.

Teorija aproksimacija se u osnovi bavi problemima zamene (aproksimacije) složenijih objekata (funkcija, funkcionela, operatora, ...) drugim objektima (obično jednostavnijim) uz određene uslove. Kako smo prethodno videli, aproksimacija funkcija može se, na primer, sprovoditi pomoću TAYLORovog razvoja, uzimanjem konačnog broja članova tako da se funkcija zamenjuje polinomom. Korišćenje polinomske aproksimacije ili tzv. racionalne aproksimacije[6] je pogodno za konstrukciju bibliotečkih funkcija u računarima, s obzirom na to da se za njihovo izračunavanje koriste samo prethodno pomenute četiri osnovne računske operacije. Teorija aproksimacija je usko povezana i dobrim delom se preklapa sa funkcionalnom analizom. Vidno mesto u teoriji aproksimacija danas zauzima *teorija ortogonalnosti*. U ruskoj literaturi teorija aprosimacija se često naziva *konstruktivna teorija funkcija*.

Na osnovu prethodnog, pogrešno bi bilo izvesti zaključak da su se numerička matematika i teorija aproksimacija počele razvijati sa pojavom računskih mašina. Poznato je, naime, da se u delima naučnika prošlih vekova nalazi niz približnih metoda koji danas čine osnovu numeričke matematike. Teorija aproksimacija ima korene u radovima ruskog matematičara ČEBIŠEVa[7] sredinom devetnaestog veka. Međutim, nagli razvoj računarske tehnike posle II svetskog rata uslovio je brzi i sistematski razvoj numeričke matematike i teorije aproksimacija. Rad koji su 1947. godine objavili VON NEUMANN[8] i GOLDSTINE[9] pod naslovom „*Numerical inverting of matrices of high order*“ u američkom časopisu *Bulletin of the American Mathematical Society*, **53** (1947), 1021–1099, često se uzima kao početak *moderne numeričke analize*[10]. U cilju obeležavanja 60 godina od ovog događaja, organizovan je simpozijum pod nazivom „*The birth of numerical analysis*“ na Katoličkom univerzitetu u Luvenu u Belgiji (K.U. Leuven, October 29–30, 2007),

[6] Aproksimacija pomoću racionalne funkcije.

[7] PAFNUTIĬ L'VOVIČ ČEBIŠEV (1821 – 1894), veliki ruski matematičar, poznat po značajnim doprinosima u verovatnoći, analizi, teoriji brojeva, mehanici, itd. U literaturi na engleskom jeziku njegovo ime se pojavljuje kao CHEBYSHEV, CHEBYCHEV ili CHEBYSHOV, na francuskom kao TCHEBYCHEFF, a u nemačkoj literaturi kao TSCHEBYSCHEFF.

[8] JOHN VON NEUMANN (1903 – 1957), mađarsko-američki matematičar, jedan od najvećih u dvadesetom veku, dao je niz značajnih doprinosa u mnogim oblastima, uključujući logiku i teoriju skupova, funkcionalnu analizu, numeričku analizu, kvantnu mehaniku, kompjuterske nauke, ekonomiju i teoriju igara, itd. Učestvovao je u izradi prve atomske bombe, radio na optimizaciji procesa snabdevanja savezničkih trupa u Evropi (1944), kao i na razvoju prvog elektronskog digitalnog računara ENIAC. Takođe, bio je glavni dizajner računara EDVAC (Electronic Discrete Variable Automatic Computer).

[9] HERMAN H. GOLDSTINE (1913 – 2004), američki matematičar i jedan od naučnika na razvoju prvog računara ENIAC (Electronic Numerator, Integrator, Analyzer, and Computer).

[10] Drugi deo ovog rada objavili su H.H. GOLDSTINE i J. VON NEUMANN u časopisu *Proc. Amer. Math. Soc.* **2** (1951), 188–202.

a 2010. godine se pojavio i odgovarajući zbornik sa tog skupa [11]. Poslednjih šezdesetak godina razrađen je niz metoda u teoriji jednačina svih vrsta i za njih su dobijeni novi fundamentalni rezultati, kako teorijski, tako i praktični. Među njima, posebno se ističu metodi za rešavanje konturnih (graničnih) problema u teoriji diferencijalnih jednačina (običnih i parcijalnih), metodi u teoriji integralnih i integro-diferencijalnih jednačina, kao i metodi za rešavanje apstraktnih operatorskih jednačina. Poslednjih godina dosta se radi na metodima za rešavanje graničnih problema u jednoj neiscrpnoj oblasti – teoriji nelinearnih parcijalnih jednačina. Takođe, vidno mesto zauzimaju razvoj interpolacionih i kvadraturnih procesa, kao i niz drugih aproksimacionih metoda (teorija splajnova, teorija malih talasa, itd.).

Među numeričkim metodama posebno su interesantni iterativni metodi, s obzirom na to da su veoma pogodni za primenu na računskim mašinama. Izučavanje numeričkih metoda uključuje analizu greške, stabilnost, konvergenciju (kod iterativnih metoda), kao i niz drugih svojstava. Zato se vrlo često ovaj deo numeričke matematike naziva numerička analiza.

U svom razvoju, numerička analiza se oslanja na više matematičkih disciplina od kojih u poslednje vreme vidno mesto zauzimaju linearna algebra, funkcionalna analiza, teorija operatora, itd. Posebno, teorija aproksimacija je do te mere povezana i isprepletena sa numeričkom matematikom da ih nije moguće jasno razgraničiti.

Problemima praktične realizacije algoritama bavi se oblast programiranja koja se u poslednje vreme uspešno razvija. Razvojem novih programskih jezika obezbeđuje se često jednostavan programski kôd. Danas su na raspolaganju mnogi programski paketi za različite namene i dosta od toga se može naći besplatno na *internetu*. Široko rasprostranjeni numerički softver[11] je *Netlib*, koji je dostupan preko URL adrese: `http://www.netlib.org`. Napomenimo još da se dosta numeričkog softvera može naći i u kolekciji TOMS (u okviru časopisa: *ACM Transactions on Mathematical Software*). Softver je najčešće realizovan na programskom jeziku FORTRAN, mada se može naći i na drugim programskim jezicima.

Najzad, pomenimo i komercijalne softverske pakete IMSL i NAG. Takođe, danas su veoma popularni tzv. interaktivni sistemi poput MAPLE, MACSYMA,

[11] Odomaćeno od engleske reči: *software*.

MATLAB[12] i MATHEMATICA[13]. Svi oni, pored numeričkih izračunavanja, nude i simbolička izračunavanja, kao i niz grafičkih mogućnosti.

Mada se radovi iz oblasti numeričke matematike i teorije aproksimacija objavljuju i u časopisima koji tretiraju opštu problematiku, ipak se u poslednjih pedesetak godina u svetu pojavio veliki broj specijalizovanih časopisa za numeričku matematiku i teoriju aproksimacija kao što su[14]:

- **ACM Trans. Math. Software** [Association for Computing Machinery. Transactions on Mathematical Software. ACM, New York. ISSN 0098–3500]
- **Adv. Comput. Math.** [Advances in Computational Mathematics. Springer, Dordrecht. ISSN 1019–7168]
- **Appl. Math. Comput.** [Applied Mathematics and Computation. Elsevier, Inc., New York. ISSN 0096–3003]
- **Appl. Numer. Math.** [Applied Numerical Mathematics. An IMACS Journal. Elsevier, Amsterdam. ISSN 0168–9274]
- **BIT** [BIT. Numerical Mathematics. Springer, Dordrecht. ISSN 0006–3835]
- **Calcolo** [Calcolo. A Quarterly on Numerical Analysis and Theory of Computation. Springer Italia, Milan. ISSN 0008–0624]
- **Comput. Math. Appl.** [Computers & Mathematics with Applications. An International Journal. Elsevier, Amsterdam. ISSN 0898–1221]
- **Comput. Math. Math. Phys.** [Computational Mathematics and Mathematical Physics. MAIK Nauka/Interperiod. Publ., Moscow. (Translation of Zh. Vychisl. Mat. Mat. Fiz.) ISSN 0965–5425]
- **Computing** [Computing. Archives for Scientific Computing. Springer, Vienna. ISSN 0010–485X]
- **Constr. Approx.** [Constructive Approximation. An International Journal for Approximations and Expansions. Springer, New York. ISSN 0176–427]
- **IMA J. Numer. Anal.** [IMA Journal of Numerical Analysis. Oxford Univ. Press, Oxford. ISSN 0272–4979]

[12] MATLAB je razvijen u kompaniji MathWorks i prilagođen za rad sa matricama kao osnovnim elementima u izračunavanjima. Naziv MATLAB dolazi od engleskih reči **mat**rix **lab**oratory. Prva verzija se pojavila 1984. godine, a sada je aktuelna verzija 8.3 (R2014a) za sve operativne sisteme.

[13] MATHEMATICA je razvijena u softverskoj kompaniji Wolfram Research. Prva verzija se pojavila 1988. godine, a 1989. verzija MATHEMATICA 2.0 za DOS operativni sistem. Verzija MATHEMATICA 2.2 vezana je za Windows 3.11. Aktuelna verzija 9.0.1, koja se pojavila početkom 2013. godine, razvijena je za sve operativne sisteme.

[14] Naslovi časopisa su dati po abecednom redu i to onako kako se citiraju u referentnom časopisu američkog matematičkog društva (AMS): *Mathematical Reviews*.

- **J. Approx. Theory** [Journal of Approximation Theory. Elsevier, Inc., San Diego, CA. ISSN 0021–9045]
- **J. Comput. Appl. Math.** [Journal of Computational and Applied Mathematics. Elsevier, Amsterdam. ISSN 0377–0427]
- **J. Comput. Math.** [Journal of Computational Mathematics. An International Journal on Numerical Methods, Analysis and Applications. Chinese Acad. Sci., Beijing. ISSN 0254–9409]
- **Math. Comp.** [Mathematics of Computation. Amer. Math. Soc., Providence, RI. ISSN 0025–571]
- **Math. Comput. Modelling** [Mathematical and Computer Modelling. Elsevier, Amsterdam. ISSN 0895–717]
- **Numer. Algorithms** [Numerical Algorithms. Springer, Dordrecht. ISSN 1017–1398]
- **Numer. Linear Algebra** [Appl. Numerical Linear Algebra with Applications. Wiley, Chichester. ISSN 1070–532]
- **Numer. Math.** [Numerische Mathematik. Springer, Heidelberg. ISSN 0029–599X]
- **Rev. Anal. Numér. Théor. Approx.** [Revue dAnalyse Numérique et de Théorie de l'Approximation. Ed. Acad. Române, Bucharest. ISSN 1222–902]
- **SIAM J. Matrix Anal. Appl.** [SIAM Journal on Matrix Analysis and Applications. SIAM, Philadelphia, PA. ISSN 0895–4798]
- **SIAM J. Numer. Anal.** [SIAM Journal on Numerical Analysis. SIAM, Philadelphia, PA. ISSN 0036–1429]
- **SIAM J. Sci. Comput.** [SIAM Journal on Scientific Computing. SIAM, Philadelphia, PA. ISSN 1064–8275]

Na kraju napomenimo da se u svetu svake godine organizuje veći broj simpozijuma posvećenih numeričkoj matematici i teoriji aproksimacija.

1.1.2 Razvoj novih naučnih disciplina i oblasti

Numerička matematika je poslednjih dvadesetak godina doživela ekspanziju, zahvaljujući mnogobrojnim primenama u skoro svim oblastima nauke i inženjerstva, s jedne strane, i burnim razvojem računarskih arhitektura, s druge strane. Sve je to izazvalo i pojavu novih pravaca istraživanja koji su se postepeno utemeljili u posebne naučne discipline. U ovom odeljku navešćemo samo nekoliko tipičnih primera.

Naučna izračunavanja[15] je posebna disciplina koja izučava primenu specijalnih numeričkih metoda u konkretnim problemima koji se pojavljuju u nauci i inženjerstvu. Na primer, tu spadaju numerička izračunavanja važnih konstanata, izračunavanje vrednosti elementarnih i specijalnih funkcija sa ekstremno visokom tačnošću, rešavanje jednačina, generisanje slučajnih brojeva, razvoj algoritama za teoriju brojeva, brze transformacije (diskretna FOURIERova[16], brza FOURIERova (FFT), diskretna kosinusna transformacija, itd.), fraktali, kao i konstrukcija niza algoritama za obradu signala. Ovo poslednje se, zajedno sa izučavanjem odgovarajućih tehničkih sredstava, može tretirati i kao posebna disciplina *obrada signala*.

Obrada signala[17] obuhvata analognu i digitalnu obradu svih vrsta signala koji se pojavljuju u realnom svetu, uključujući sintezu signala, detekciju, modeliranje, korelaciju i spektralnu analizu signala, konstrukciju i primenu odgovarajućih filtara, itd. Posebno važan deo je onaj koji se odnosi na kompresiju podataka bilo da se radi o zvučnim signalima ili slikama.

Teorija kompleksnosti,[18] a posebno *teorija kompleksnosti izračunavanja* obezbeđuje okvir za razumevanje cene rešavanja problema, merene zahtevima za resursima kakvi su vreme i memorijski prostor. Glavni pristup u teoriji kompleksnosti je razmatranje algoritama koji deluju na konačne nizove simbola iz jednog konačnog alfabeta (azbuke). Ti nizovi mogu predstavljati najrazličitije diskretne objekte poput celih brojeva ili algebarskih izraza, ali ne i realne (ili kompleksne) brojeve, osim ako oni nisu zaokrugljeni na približne vrednosti iz jednog diskretnog skupa. Glavni zadatak ove teorije je da se odredi broj (računskih) koraka neophodnih za rešavanje problema u funkciji dužine ulaznog niza. U vezi sa ovim, teorija kompleksnosti grupiše probleme u tzv. *klase kompleksnosti* i razmatra njihov odnos. Na primer, klasu P čine oni problemi koji se mogu rešiti u polinomijalnom vremenu, tj. broj koraka neophodnih za njihovo rešavanje je ograničen polinomijalnom funkcijom dužine ulaznog niza, dok je NP klasa problema čija se rešenja mogu proveriti (verifikovati) u polinomialnom vremenu. U novije vreme tretiraju se klase kompleksnosti i nad poljem realnih brojeva [17].

Geometrijsko modeliranje[19] je disciplina koja proučava metode konstruisanja geometrijskih i prirodnih formi sredstvima računarske grafike (videti [41]). U pozadini složenih grafičkih algoritama stoje sofisticirani numerički pristupi, neo-

[15] Na engleskom: *Scientific Computation*.
[16] JEAN-BAPTISTE JOSEPH FOURIER (1768 – 1830), poznati francuski fizičar i matematičar.
[17] Na engleskom: *Signal Processing*.
[18] Na engleskom: *Complexity Theory*.
[19] Na engleskom: *Computer Added Geometric Design (CAGD)*.

Slika 1.2.1. Primeri biološkog sveta

phodni u procesu optimizacije algoritamskih tokova i izbora najboljih modela. U svemu tome vodič i inspiracija je *priroda* i njene tvorevine od kojih neke interesantne primere možemo videti na slici 1.2.1, preuzete iz članka [73], a koji, bilo da su žive ili nežive strukture fasciniraju svojom racionalnom geometrijom kojoj u osnovi stoji hijerarhija samosličnosti i veoma složeni iterativni procesi.

Paralelna izračunavanja[20] predstavljaju disciplinu koja se bavi konstrukcijom takvih algoritama i procedura koji su mogu implementirati na višeprocesorskim sistemima, koji se, s druge strane, svakim danom sve više razvijaju i usavršavaju.

Simbolička izračunavanja,[21] se danas, takođe, mogu izdvojiti u posebnu disciplinu. Ona se često nazivaju i *kompjuterska algebra*, *algebarska izračunavanja*, *simbolička matematika*, itd. Za razliku od numeričkih izračunavanja koja se uglavnom realizuju u tzv. aritmetici konačne dužine (videti odeljak 1.2.3), simbo-

[20] Na engleskom: *Parallel Computation*.
[21] Na engleskom: *Simbolic Computation*.

lička izračunavanja se izvode sa brojevima, simbolima, izrazima i formulama na egzaktan način. Ona su nastala kao rezultat težnje da se sa numeričkih izračunavanja krene ka apstraktim izračunavanjima, što je omogućeno razvojem tzv. *veštačke inteligencije* i novih programskih jezika, poput jezika LISP i njegovih usavršenih naslednika. Danas su za simbolička izračunavanja najpopularniji i široko rasprostranjeni interaktivni paketi MAPLE, MACSYMA, MATLAB i MATHEMATICA, koji se, naravno, mogu koristiti i za numerička izračunavanja, kao i za grafičke prezentacije.

Ilustrovaćemo sada kako se u paketu MATHEMATICA mogu na jednostavan način, izdavanjem naredbe u komandnoj liniji, rešavati razni matematički problemi. Složeniji problemi zahtevaju pripremu odgovarajućih programa.

Na primer, da bismo dobili rešenja jednačine četvrtog stepena $x^4+2x+1=0$ u simboličkom obliku, pišemo jednostavno

```
In[1]:= Solve[x^4 + 2 x + 1 == 0, x]
```

$$\text{Out[1]}= \left\{\{x \to -1\},\ \left\{x \to \frac{1}{3}\left(1 - \frac{2}{\left(-17+3\sqrt{33}\right)^{1/3}} + \left(-17+3\sqrt{33}\right)^{1/3}\right)\right\},\right.$$

$$\left\{x \to \frac{1}{3} + \frac{1+i\sqrt{3}}{3\left(-17+3\sqrt{33}\right)^{1/3}} - \frac{1}{6}\left(1-i\sqrt{3}\right)\left(-17+3\sqrt{33}\right)^{1/3}\right\},$$

$$\left.\left\{x \to \frac{1}{3} + \frac{1-i\sqrt{3}}{3\left(-17+3\sqrt{33}\right)^{1/3}} - \frac{1}{6}\left(1+i\sqrt{3}\right)\left(-17+3\sqrt{33}\right)^{1/3}\right\}\right\}$$

Međutim, ako želimo samo drugo rešenje, izostavljajući pri tome vitičaste zagrade i simboliku $x \to$, dajemo naredbu

```
In[2]:= x /. Solve[x^4 + 2 x + 1 == 0, x] [[2]]
```

$$\text{Out[2]}= \frac{1}{3}\left(1 - \frac{2}{\left(-17+3\sqrt{33}\right)^{1/3}} + \left(-17+3\sqrt{33}\right)^{1/3}\right)$$

Numerička vrednost ovog rešenja, na primer sa 16 cifara, dobija se jednostavno pomoću:

```
In[3]:= N[%, 16]

Out[3]= -0.5436890126920764
```

Integracija funkcija, na primer, određivanje primitivne funkcije za $1/(1+x^3)$ sprovodi se takođe jednostavnom naredbom:

```
In[4]:= Integrate[1 / (1 + x^3), x]
```

$$\text{Out[4]}= \frac{\text{ArcTan}\left[\frac{-1+2x}{\sqrt{3}}\right]}{\sqrt{3}} + \frac{1}{3}\text{Log}[1+x] - \frac{1}{6}\text{Log}\left[1-x+x^2\right]$$

Štavise, MATHEMATICA može da rešava i neke složenije slučajeve, na primer, računanje nesvojstvenog integrala funkcije $1/(x^2+2ax+1)$ na $(0,+\infty)$ sa parametrom $a>0$,

```
In[5]:= Integrate[1 / (x^2 + 2 a x + 1)^2, {x, 0, Infinity},
          Assumptions → a > 0]
```

$$\text{Out[5]}= \frac{1}{4}\left(\frac{2a}{-1+a^2} + \left(\frac{1}{1-a^2}\right)^{3/2}\pi - \frac{2\,\text{ArcTan}\left[\frac{a}{\sqrt{1-a^2}}\right]}{\left(1-a^2\right)^{3/2}}\right)$$

Kombinovanje numeričkih i simboličkih izračunavanja može biti veoma korisno u mnogim primenama. Za numerička izračunavanja posebno su interesantne tzv. *aritmetike višestruke tačnosti.*[22] Navešćemo jedan jednostavan primer.

Primer 1.2.1. U časopisu *Quarterly Journal of Pure and Applied Mathematics* **45** (1913/14), str. 350, RAMANUJAN[23] je postavio hipotezu da je $e^{\pi\sqrt{163}}$ ceo broj, pri čemu je, radeći „ručno“, našao da je

$$e^{\pi\sqrt{163}} \cong 262\,537\,412\,640\,768\,743.999\,999\,999\,999.$$

Kako njegov metod nije omogućavao dobijanje sledeće decimale, on je pretpostavio da se cifra 9 stalno ponavlja, i da je onda $e^{\pi\sqrt{163}} = 262\,537\,412\,640\,768\,744$. Korišćenjem programskog paketa MATHEMATICA jednostavno dobijamo

[22] Na engleskom: *multi-precision arithmetics*.
[23] SRINIVASA RAMANUJAN (1887 – 1920), indijski matematičar.

$$e^{\pi\sqrt{163}} = 262537412640768743.99999999999925007259719818568887935\ldots,$$

što ukazuje da je hipoteza RAMANUJANa bila pogrešna. △

Najzad napomenimo da je *Obrada teksta*[24] danas oblast bez koje se ne može zamisliti bilo koja delatnost. Tokom sedamdesetih i osamdesetih godina prošloga veka intenzivno se počelo sa razvojem efikasnih algoritama i konstrukcijom tzv. *tekst procesora* za obradu svih vrsta tekstova, uključujući i matematički slog. Kada su u pitanju matematički tekstovi, glavni doprinos je učinio DONALD KNUTH[25] razvojem sistema TEX. Njegove pojednostavljene varijante, poput LATEX-a koji je, u stvari, makro paket i čije su komande definisane nizom TEX-komandi, danas je postao standard u matematičkoj komunikaciji.

1.2 ANALIZA GREŠAKA U NUMERIČKIM PROCESIMA

U ovom poglavlju se uvode osnovni pojmovi teorije grešaka i analizira efekat prostiranja grešaka kroz računski proces. Poseban tretman je dat mašinskim greškama zaokrugljivanja i efektu gubitaka značajnih cifara. Na kraju poglavlja su date izvesne napomene o statističkom pristupu u oceni grešaka.

1.2.1 Greške u numeričkim izračunavanjima

U skoro svim numeričkim problemima, izlazna informacija, ili kako se češće kaže numeričko rešenje, praćeno je greškama čiji izvori mogu biti različiti. Pre nego što pređemo na klasifikaciju grešaka uvešćemo neke osnovne pojmove i definicije.

Definicija 2.1.1. Svaka cifra broja, izuzimajući nule koje služe za fiksiranje decimalne tačke, naziva se *značajnom cifrom* tog broja.

[24] Na engleskom: *Text Processing*.

[25] DONALD ERVIN KNUTH (1938 –), profesor emeritus na Stanford univerzitetu (SAD), je danas najpoznatije ime u oblasti informatike i računarstva. Pored ogromnog doprinosa u matematičkom zasnivanju algoritama, zbog čega je poznat kao „otac algoritama“, KNUTH je razvio softverske sisteme poznate kao TEX i METAFONT, koji su promenili tehnologiju štampanja matematičkih i drugih publikacija, ali i način komunikacije među naučnicima (za detalje videti: `http://www-cs-faculty.stanford.edu/~knuth/`).

Primer 2.1.1. Svaki od navedenih brojeva

$$2.563, \quad 15.32, \quad 4.300, \quad 0.05050, \quad 0.2687, \quad 0.002649$$

ima četiri značajne cifre. $\triangle$

Približan broj $\overline{x}$ je broj koji zamenjuje tačan broj x u izračunavanjima i neznatno se razlikuje od njega. Odgovarajuća greška broja $\overline{x}$ je

$$e = \overline{x} - x,$$

ali se češće koristi tzv. *apsolutna greška*

$$|e| = |\overline{x} - x|.$$

U slučaju kada tačna vrednost x nije poznata, uvodi se pojam granice apsolutne greške približnog broja $\overline{x}$.

Definicija 2.1.2. Pod *granicom apsolutne greške* Δ_x približnog broja $\overline{x}$ podrazumeva se svaki broj ne manji od apsolutne greške tog broja.

Kako je $|e| = |\overline{x} - x| \leq \Delta_x$, imamo

$$\overline{x} - \Delta_x \leq x \leq \overline{x} + \Delta_x. \tag{2.1.1}$$

S obzirom na to da greška e nedovoljno karakteriše tačnost uvodi se i pojam relativne greške[26] kao

$$r = \frac{e}{x} = \frac{\overline{x} - x}{x} \quad (x \neq 0),$$

pri čemu se češće operiše sa njenom apsolutnom vrednošću $|r|$. Primetimo da je

$$\overline{x} = x(1 + r). \tag{2.1.2}$$

Slično se uvodi i pojam *granice relativne greške*

$$\varepsilon_x = \frac{\Delta_x}{|x|} \cong \frac{\Delta_x}{|\overline{x}|}.$$

[26] Relativna greška se često povezuje sa pojmom *korektne značajne cifre* u približnom broju $\overline{x}$. Broj korektnih značajnih cifara je jedna gruba mera tačnosti u odnosu na relativnu grešku, na primer, ako je relativna greška reda 10^{-3}, tada približni broj ima grubo rečeno tri korektne značajne cifre.

Nejednakosti (2.1.1) tada postaju

$$\overline{x}(1-\varepsilon_x \mathrm{sgn}\,(\overline{x})) \leq x \leq \overline{x}(1+\varepsilon_x \mathrm{sgn}\,(\overline{x}))\,.$$

Napomenimo da u mnogim izračunavanjima, posebno naučnim, veličine sa kojima radimo mogu dosta varirati po veličini, pa smo zbog toga prinuđeni da uvodimo tzv. normalizacione faktore, tj. da takve veličine *skaliramo*. Na primer, umesto sa x i $\overline{x}$ mi radimo sa αx i $\alpha\overline{x}$, respektivno, gde je α neki normalizacioni faktor. U tom slučaju, relativna greška se ne menja, tj. važi

$$r = \frac{\alpha\overline{x} - \alpha x}{\alpha x} = \frac{\overline{x} - x}{x}.$$

Po jednoj od mogućih klasifikacija, greške se mogu podeliti na:

1° neotklonjive greške;

2° greške metoda;

3° greške zaokrugljivanja.

Neotklonjive greške su one na koje ne možemo uticati u toku numeričkog procesa. To su uglavnom greške koje potiču od netačnosti ulaznih podataka, koji su dobijeni najčešće merenjima u nekom prethodnom postupku. Greške ulaznih podataka (ulazne informacije) mogu se u nekim slučajevima drastično manifestovati u izlaznoj informaciji i pri uslovu da sâm algoritam ne unosi grešku.

Primer 2.1.2. Sistem jednačina (ulazna informacija)

$$\begin{aligned} 2x + 6y &= 8, \\ 2x + 6.0001y &= 8.0001 \end{aligned}$$

ima rešenje (izlazna informacija) $x = 1,\ y = 1$. Ako se koeficijenti druge jednačine neznatno promene, tj. ako se uzme jednačina

$$2x + 5.99999y = 8.00002,$$

tada su su $x = 10,\ y = -2$. Dakle, ovaj sistem jednačina je jako osetljiv jer male promene (perturbacije) u ulaznim podacima izazivaju velike promene u rešenju. Za takve jako osetljive sisteme kažemo da su *slabo uslovljeni*[27]. $\triangle$

[27] Na engleskom: *ill-conditioned systems*.

Greške metoda se javljaju usled toga što se u primenama obično dati problem zamenjuje drugim problemom ($P(\boldsymbol{x}) \cong \widetilde{P}(\boldsymbol{x})$), koji je lakši za računanje, a čije se rešenje poklapa sa rešenjem postavljenog „teškog“ problema ili je u izvesnom smislu blisko njegovom rešenju. Dakle, jedan složeni problem $P(\boldsymbol{x})$ zamenjuje se jednostavnijim problemom $\widetilde{P}(\boldsymbol{x})$; na primer, jedan beskonačno dimenzionalni problem zamenjuje se konačno–dimenzionalnim, diferencijalni problem[28] algebarskim, nelinearan problem linearnim, itd. Tada, dobijeno rešenje problema $\widetilde{P}(\boldsymbol{x})$ predstavlja samo aproksimaciju rešenja originalnog problema $P(\boldsymbol{x})$.

Tokom procesa rešavanja problema $\widetilde{P}(\boldsymbol{x})$, usled zaokrugljivanja međurezultata u procesu računanja, nastaju *greške zaokrugljivanja*. U zavisnosti od tzv. *uslovljenosti* (*osetljivosti*) *problema*, eventualne greške ulaznih podataka, kao i greške koje se generišu tokom procesa računanja, mogu se značajno akumulirati i ozbiljno ugroziti tačnost krajnjeg rezultata (izlazne informacije).

Primer 2.1.3. Integral $\int_a^b f(x)\,\mathrm{d}x$ se može približno izračunati, na primer, zamenom date funkcije f nekim algebarskim polinomom P na segmentu $[a,b]$ koji je u izvesnom smislu „blizak“ datoj funkciji. Međutim, moguće je za približno izračunavanje integrala koristiti i konačnu sumu

$$\sum_{i=1}^{n} f(x_i)\Delta x_i. \quad \triangle$$

U oba slučaja iz prethodnog primera čine se greške metoda. Greške odsecanja su, takođe, greške metoda.

Zbir svih grešaka koje se pojavljuju u numeričkom procesu čini *totalnu grešku*.

1.2.2 Brojni sistemi sa pokretnom tačkom

Kao što je poznato iz standardnih matematičkih kurseva realni brojevi se definišu aksiomatski kao kompletno uređeno polje $\mathbb{R}$. U našem razmatranju ovde, realne brojeve $\mathbb{R}$ ćemo tretirati samo kao skup pozitivnih i negativnih brojeva, reprezentovanih u nekom pogodnom brojnom sistemu, na koje se primenjuju uobičajene aritmetičke operacije. Brojni sistem sa osnovom $b = 10$ (*dekadni* ili *decimalni sistem*) je u svakodnevoj upotrebi. Na primer, u tom sistemu broj 124.257 ima reprezentaciju

$$(124.257)_{10} = 1\cdot 10^2 + 2\cdot 10^1 + 4\cdot 10^0 + 2\cdot 10^{-1} + 5\cdot 10^{-2} + 7\cdot 10^{-3}.$$

[28] Problem opisan diferencijalnim jednačinama.

Brojni sistem sa osnovom $b = 2$ naziva se *binarni sistem*.

U opštem slučaju, u brojnom sistemu sa osnovom b, broj $x \in \mathbb{R}$ ima reprezentaciju

$$\begin{aligned} x &= \pm(c_m c_{m-1} \dots c_1 c_0 \,.\, b_1 b_2 \dots)_b \\ &= \pm\big(\underbrace{c_m b^m + c_{m-1} b^{m-1} + \cdots + c_1 b + c_0}_{\text{celi deo broja}} + \underbrace{b_1 b^{-1} + b_2 b^{-2} + \cdots}_{\text{razlomljeni deo broja}}\big), \end{aligned} \tag{2.2.1}$$

gde je osnova b prirodan broj veći od jedinice, a c_ν i b_ν su cifre brojnog sistema koje pripadaju skupu $\{0,1,\dots,b-1\}$.

U ovom slučaju indeks b, kao i 10 u prethodnom slučaju, označava osnovu upotrebljenog brojnog sistema. Kada ne može doći do zabune indeks b u (2.2.1) izostavljamo. Tačka između c- i b-cifara, tj. između c_0 i b_1, razdvaja celi deo od razlomljenog dela broja x i naziva se decimalna tačka ako je osnova $b = 10$, binarna tačka ako je osnova $b = 2$, itd. U zavisnosti od izabrane osnove (baze) brojnog sistema, odgovarajuće cifre se nazivaju dekadne (decimalne) za $b = 10$, binarne za $b = 2$, itd. Binarne cifre se još nazivaju i *bitovi*. Na primer, za binarni broj 11001.01 imamo

$$(11001.01)_2 = 2\cdot 2^4 + 1\cdot 2^3 + 0\cdot 2^2 + 0\cdot 2^1 + 1\cdot 2^0 + 0\cdot 2^{-1} + 1\cdot 2^{-2} = (41.25)_{10}.$$

U slučaju tzv. *heksadecimalnog sistema* sa osnovom $b = 16$, cifre pripadaju skupu

$$\big\{0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F\big\},$$

sa značenjem $(A)_{16} = (10)_{10}$, $(B)_{16} = (11)_{10}$, $(C)_{16} = (12)_{10}$, $(D)_{16} = (13)_{10}$, $(E)_{16} = (14)_{10}$, $(F)_{16} = (15)_{10}$. Na primer,

$$\begin{aligned} (13AF.C8)_{16} &= 1\cdot 16^3 + 3\cdot 16^2 + 10\cdot 16^1 + 15\cdot 16^0 + 12\cdot 16^{-1} + 8\cdot 16^{-2} \\ &= (5039.78125)_{10}. \end{aligned}$$

Mada je predstavljanje brojeva teorijski ekvivalentno u svim bazama, sa numeričkog stanovišta i načina memorisanja u svakodnevnoj upotrebi je tradicionalno dekadni brojni sistem, a kod računara su to najčešće binarni i ponekad heksadecimalni sistem. Jasno je da se isti broj reprezentuje sa manjim brojem cifara ako je baza veća.

Da bismo izvodili aritmetičke operacije sa brojevima oblika (2.2.1) neophodno je definisati te operacije pomoću sukcesivnih aproksimacija. Naime, za dva takva

realna broja x i y uzmimo njihove konačne aproksimacije x_r i y_r dobijene iz originalnih brojeva x i y odsecanjem svih cifara posle r-te cifre u razlomljenom delu odgovarajućih brojeva. Tada za bilo koju operaciju $* \in \{+, -, \times, \div\}$ rezultat $x_r * y_r$ može biti izračunat na uobičajeni način, a traženi pravi rezultat $x * y$ bi bio granična vrednost niza $x_r * y_r$, kada $r \to +\infty$. Međutim, takav granični proces se ne bi mogao realizovati u konačnom vremenu. Takođe, brojevi sa beskonačnim brojem cifara ne mogu biti predstavljeni (memorisani) u računaru. Dakle, u računaru je moguće reprezentovati samo brojeve sa konačnim brojem cifara i implementirati odgovarajuće operacije sa takvim brojevima. Štaviše, i za taj konačni broj cifara postoji ograničenje određeno brojem memorijskih mesta za pamćenje cifara. Ovo ograničenje nas upućuje na uvođenje tzv. mašinskog broja i aritmetiku konačne dužine, o čemu će biti reči u narednim odeljcima.

S obzirom na to da je u (2.2.1) cifra $c_m \neq 0$, upotrebom multiplikativnog faktora b^{m+1}, moguće je pomeriti tačku tako da je

$$x = \pm\big(0.c_m c_{m-1} \dots c_1 c_0 b_1 b_2 \dots\big) b^{m+1}.$$

Na primer, za brojeve $(3820.104)_{10}$, $(0.003567)_{10}$, $(11010.1011)_2$ imamo redom

$$3820.104 = 0.3820104 \cdot 10^4, \quad 0.003567 = 0.3567 \cdot 10^{-2},$$

$$11010.1011 = 0.110101011 \cdot 2^5.$$

U opštem slučaju, ovakvim načinom pomeranja tačke dolazimo do tzv. *normalizovanog* oblika broja

$$x = \pm(0.a_1 a_2 \dots) b^k \qquad (a_1 \neq 0),$$

gde je k ceo broj, a cifre iza tačke označili smo sada redom sa a_1, a_2, itd.

Broj $0.a_1 a_2 \dots$ $(a_1 \neq 0)$ zovemo *mantisa* broja x i označavamo je sa x^*, dok za broj k kažemo da je *karakteristika* ili *eksponent* broja x. Dakle, normalizovani broj u pokretnoj tački je

$$x = \pm x^* b^k, \tag{2.2.2}$$

koga karakteriše znak koji može biti pozitivan ili negativan, mantisa x^*, osnova brojnog sistema b i karakteristika k. Očigledno, skup normalizovanih brojeva (u pokretnoj tački) ne sadrži nulu. Da bismo obezbedili jedinstvenost (nestandardne) reprezentacije nule, možemo uzeti da je to broj koji ima $\operatorname{sgn}(0) = +$, mantisu sa svim ciframa koje su jednake nuli i, na primer, $k = 0$.

Sada smo spremni da uvedemo (konačni) *brojni sistem sa pokretnom tačkom*[29], za koji ćemo pretpostaviti da egzistira na nekom hipotetičkom računaru. Razmatraćemo pritom i aproksimaciju realnih brojeva $x\ (\in \mathbb{R})$ pomoću mašinskih brojeva iz takvog sistema primenom tzv. *procesa zaokrugljivanja*.

Definicija 2.2.1. Brojni sistem sa pokretnom tačkom $A(b,n,k_{\min},k_{\max}) \subset \mathbb{R}$ je skup realnih brojeva oblika

$$y = \pm(0.a_1a_2\ldots a_n)b^k, \tag{2.2.3}$$

gde su $a_1(\neq 0), a_2, \ldots, a_n$ cifre brojnog sistema sa osnovom b i k karakteristika (eksponent) broja čije su granice $k_{\min}$ i $k_{\max}$, tj. $k_{\min} \leq k \leq k_{\max}$. Broj n se naziva *preciznost* mašinskih brojeva (2.2.3).

Primetimo da je u broju (2.2.3) najmanja vrednost mantise $0.10\ldots 0$, a najveća $0.cc\ldots c$, gde je $c = b-1$ najveća cifra u brojnom sistemu sa osnovom b, tj. da za mantisu $y^* = 0.a_1a_2\ldots a_n$ važi

$$b^{-1} \leq y^* \leq 1 - b^{-n}.$$

To dalje znači da je najmanji pozitivan broj u sistemu $A(b,n,k_{\min},k_{\max})$ upravo broj $b^{k_{\min}-1}$, a najveći $(1-b^{-n})b^{k_{\max}}$. Dakle, opseg svih brojeva u ovom sistemu je

$$b^{k_{\min}-1} \leq |y| \leq (1-b^{-n})b^{k_{\max}}.$$

Broj tih brojeva je, očigledno, konačan.

Primer 2.2.1. Dat je binarni brojni sistem sa pokretnom tačkom $A(2,3,-1,2)$. Svi pozitivni brojevi iz ovog sistema prikazani su u tabeli 2.2.1, a njihove vrednosti su redom

$$\frac{4}{16},\ \frac{5}{16},\ \frac{6}{16},\ \frac{7}{16},\ \frac{8}{16},\ \frac{10}{16},\ \frac{12}{16},\ \frac{14}{16},\ \frac{16}{16},\ \frac{20}{16},\ \frac{24}{16},\ \frac{28}{16},\ \frac{32}{16},\ \frac{40}{16},\ \frac{48}{16},\ \frac{56}{16},$$

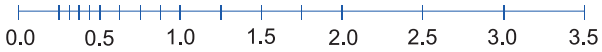
tj. 0.25, 0.3125, 0.375, 0.4375, 0.5, 0.625, 0.75, 0.875, 1.0, 1.25, 1.5, 1.75, 2.0, 2.5, 3.0, 3.5.

Ovi brojevi prikazani su na realnoj pravoj (slika 2.2.1), pri čemu u njihovoj distribuciji primećujemo da se u tačkama 2^k, $k = -1,0,1$, pojavljuju skokovi u rastojanju h između susednih brojeva, tj. h se udvostručava. Na istoj slici, pored pozitivnih brojeva ovog sistema, prikazan je i broj 0, kojim se može proširiti sistem $A(2,3,-1,2)$, tako da je u sistemu ukupno 33 broja. $\triangle$

[29] Na engleskom: *floating-point number system*.

Tabela 2.2.1.

k	mantisa $y^* = 0.a_1a_2a_3$				2^k
-1	0.100	0.101	0.110	0.111	$1/2$
0	0.100	0.101	0.110	0.111	1
1	0.100	0.101	0.110	0.111	2
2	0.100	0.101	0.110	0.111	4



Slika 2.2.1. Skup nenegativnih brojeva sistema $A(2,3,-1,2)$

Sistem $A(b,n,k_{\min},k_{\max})$ može se proširiti *nenormalizovanim* brojevima oblika

$$\pm(0.a_1a_2\ldots a_n)b^{k_{\min}},$$

gde je $a_1 = 0$. Preciznost ovih brojeva je evidentno manja od preciznosti normalizovanih brojeva (2.2.3). Najmanji pozitivan broj ovog tipa je $0.00\ldots1\times b^{k_{\min}} = b^{k_{\min}-n}$, dok je, kao što smo videli, $b^{k_{\min}-1}$ najmanji normalizovani pozitivni broj. Označimo njihov odnos sa

$$e_M = \frac{b^{k_{\min}-n}}{b^{k_{\min}-1}} = b^{-n+1}. \tag{2.2.4}$$

Primetimo da sada za 0 možemo uzeti (jedinstvenu) reprezntaciju

$$+0.00\ldots0\times b^{k_{\min}-1}.$$

Praznina koja se pojavljuje kod normalizovanih brojeva između 0 i broja $b^{k_{\min}-1}$ može se uniformno popuniti prethodno pomenutim nenormalizovanim brojevima, sa korakom koji je isti kao i kod najmanjih pozitivnih normalizovanih brojeva, tj. $h = b^{k_{\min}-n}$. Na ovaj način dobijamo prošireni brojni sistem sa pokretnom tačkom $A^*(b,n,k_{\min},k_{\max})$.

Primer 2.2.2. Sistem $A(2,3,-1,2)$ iz prethodnog primera može se dopuniti nulom i nenormalizovanim brojevima

$$\begin{aligned}
0.001\times2^{-1} &= 2^{-4} = 0.0625,\\
0.010\times2^{-1} &= 2^{-3} = 0.125,\\
0.011\times2^{-1} &= 2^{-3}+2^{-4} = 0.1875.
\end{aligned}$$

Slika 2.2.2. Skup nenegativnih brojeva proširenog sistema $A^*(2,3,-1,2)$

Svi nenegativni brojevi proširenog sistema $A^*(2,3,-1,2)$ prikazani su na slici 2.2.2. $\triangle$

Na osnovu prethodnog vidimo da u proširenom sistemu $A^*(b,n,k_{\min},k_{\max})$ imamo jedinstvenu reprezentaciju brojeva i da je njihov ukupan broj (pozitivnih i negativnih, uz uključenje nule) jednak

$$\operatorname{card} A^*(b,n,k_{\min},k_{\max}) = 2\left\{(b-1)b^{n-1}(k_{\max}-k_{\min}+1)+b^{n-1}-1\right\}+1.$$

Primetimo, da u sistemu $A^*(2,3,-1,2)$ imamo ukupno 39 brojeva.

Takođe, kao što smo videli u primeru 2.2.1, evidentna je nejednaka gustina brojeva, kada se menja eksponent. Na primer, nije teško ustanoviti da za rastojanje između dva susedna broja $y,z \in A(b,n,k_{\min},k_{\max})$, predstavljenih pomoću (2.2.3), važe nejednakosti

$$b^{-1}e_M|y| < \frac{e_M|y|}{b-e_M} \le |y-z| \le e_M|y|,$$

gde je e_M definisano pomoću (2.2.4). Kako je $e_M = b^{1-n}$, vidimo da ta razlika zavisi od preciznosti n, ali i od veličine y. Deobom prethodnih nejednakosti sa $|y|$ $(\neq 0)$ zaključujemo da se odgovarajuća relativna razlika nalazi između $b^{-1}e_M$ i e_M. Za veličinu $e_M = b^{1-n}$ često se koristi termin *mašinska preciznost* ili *mašinski epsilon*[30]. Jednostavno je uveriti se da e_M predstavlja razliku između broja 1 i njemu najbližeg mašinskog broja $(\in A(b,n,k_{\min},k_{\max}))$ koji je veći od jedinice, tj. da je e_M najmanji mašinski broj takav da je $1+e_M > 1$. Ova nejednakost može se iskoristiti za praktično određivanje veličine e_M na konkretnom računaru (videti primer 2.2.3).

U skoro svim modernim računarima imamo binarnu reprezentaciju brojeva u pokretnoj tački, gde se koristi određeni broj bitova za memorisanje znaka, mantise i eksponenta broja.

z	eksponent	mantisa

Za memorisanje znaka broja dovoljan je jedan bit, s obzirom na to da se znak može predstaviti kao $(-1)^z$, gde je $z=0$ za pozitivne brojeve, a $z=1$ za negativne

[30] Na engleskom: *machine epsilon*. Koriste se i nazivi *macheps* i *unit roundoff*.

brojeve. Međutim, izborom broja bitova za mantisu i eksponent broja, moguće je generisati različite formate brojeva. Da bi se izbegle ove razlike, koje su bile prisutne u početnoj fazi razvoja računarstva, 1985. godine uveden je tzv. IEEE standard, a četiri godine kasnije poboljšan međunarodnim IEC standardom[31].

Najčešće su prisutna dva formata brojeva: brojevi u *običnoj* ili *prostoj preciznosti* i brojevi u tzv. *dvostrukoj preciznosti*[32]. Za predstavljanje brojeva u prostoj preciznosti koriste se 32 bita, od toga 8 bitova za eksponent, 23 za mantisu i jedan bit za znak broja, dok se kod brojeva u dvostrukoj preciznosti koriste 64 bita, od toga 11 za eksponent, 52 za mantisu i jedan za znak broja. Napomenimo da se praktično broj bitova za mantisu može povećati za jedan, s obzirom da je prvi bit u mantisi (normalizovanog broja) uvek jednak jedinici i da se on ne mora pamtiti. Dakle, taj tzv. skriveni bit mantise ($a_1 = 1$) uvećava dužinu mantise, tako da je u običnoj preciznosti $n = 24$, a u dvostrukoj preciznosti $n = 53$.

Neka je za eksponent, koji je ceo broj, predviđeno m bitova. Tada je moguće predstaviti najviše 2^m različitih celih brojeva. U pomenutim formatima za eksponent se uzimaju granice $k_{\min} = -125$ i $k_{\max} = 128$ (kod obične preciznosti) i $k_{\min} = -1021$ i $k_{\max} = 1024$ (kod dvostruke preciznosti), čime se predstavlja $|k_{\min}| + k_{\max} + 1$ različitih eksponenata, tj. 254 i 2046, respektivno, što je za dva manje od ukupno raspoloživih mogućnosti 2^m, koje se takođe koriste, o čemu će biti reči u nastavku.

Vratimo se sada na aproksimaciju realnih brojeva x ($\in \mathbb{R}$) pomoću brojeva iz sistema $A^*(b, n, k_{\min}, k_{\max})$. Pretpostavimo da je realan broj x ($\neq 0$) dat u normalizovanom obliku (2.2.2), tj. kao

$$x = \pm(0.a_1a_2 \dots a_na_{n+1} \dots)b^k \qquad (a_1 \neq 0), \tag{2.2.5}$$

gde je $k \leq k_{\max}$. Brojevi kod kojih je $k > k_{\max}$ ne mogu se predstaviti u ovom sistemu i za njih kažemo da postoji tzv. *prekoračenje*[33]. Formalno, svi realni brojevi (pozitivni i negativni) za koje imamo prekoračenje, tretiraju se po pomenutom standardu kao *mašinska beskonačnost* $\widehat{\infty}$, sa jedinstvenom reprezentacijom $k = k_{\max} + 1$ i svi bitovi u mantisi su nule. Vrednost $k = k_{\max} + 1$ i nenula mantisa definišu NaN, tj. tzv. *ne-broj*[34]. S druge strane, s obzirom da je najmanji pozitivni broj u $A^*(b, n, k_{\min}, k_{\max})$ jednak $b^{k_{\min}-n}$, brojevi (2.2.5) kod kojih je $k < k_{\min} - n + 1$ mogu se tretirati kao nula. Ovaj skup brojeva, zajedno sa

[31] IEEE i IEC su engleske skraćenice za *Institute of Electrical and Electronic Engineers* i *International Electronical Commission*, respektivno.

[32] Na engleskom: *single* i *double precision*.

[33] Na engleskom: *overflow*.

[34] Na engleskom: *not a number*.

suprotnim (negativnim) brojevima i nulom, identifikuju se kao jedan broj, tzv. *mašinska nula*, u oznaci $\widehat{0}$ ili jednostavno 0 ako ne dolazi do zabune. Jedinstvena reprezentacija $\overline{0}$ je takva da je $k = k_{\min} - 1$, a svi bitovi u mantisi su nule.

Primer 2.2.3. Odredićemo mašinski epslilon e_M korišćenjem programskog sistema MATLAB, verzija 8.0 (R2012b), na racunaru MacBook Pro (Retina, Mid 2012), pod operativnim sistemom Mac OS X 10.9.2. Test sprovodimo nalaženjem najmanjeg (mašinskog) broja koji zadovoljava nejednakost $1 + e_M > 1$, sledećim nizom naredbi

```
>> a=1; while 1+a > 1; a=a/2; end
>> a
a =
   1.1102e-16
```

Promenljiva a, sa početnom vrednošću $a = 1$, deli se sa dva sve dok važi nejednakost $1 + a > 1$. Ako bismo radili sa skupom realnih brojeva $\mathbb{R}$, proces bi trajao beskonačno. Međutim, ovde se proces završava posle konačnog broja koraka, dajući traženi rezultat $a = \texttt{1.1102e}-\texttt{16} = e_M/2$, tj. $e_M = \texttt{2.2204e}-\texttt{16}$. Napomenimo, da je u MATLABu definisana konstanta `eps`:

```
>> eps
ans =
   2.2204e-16
```

koja je, u stvari, veličina e_M. $\triangle$

Najjednostavniji postupak aproksimacije je *prosto odsecanje*[35], koje se sastoji u odbacivanju cifara a_{n+1}, a_{n+2}, itd. Dakle,

$$\overline{x} = \mathrm{chop}(x) = \pm(0.a_1a_2\ldots a_n)b^k. \tag{2.2.6}$$

Granica apsolutne greške koja se čini na ovaj način može se oceniti na osnovu (2.2.5) i (2.2.6) i činjenice da je $b-1$ najveća cifra brojnog sistema sa osnovom b. Naime,

$$\begin{aligned}|\mathrm{chop}(x) - x| &= \left|\pm\left(a_{n+1}b^{-(n+1)} + a_{n+2}b^{-(n+2)} + \cdots\right)\right| b^k \\ &\leq (b-1)\left(b^{-(n+1)} + b^{-(n+2)} + \cdots\right)b^k \\ &= (b-1)\frac{b^{-(n+1)}}{1-b^{-1}}b^k = b^{-n+k}.\end{aligned}$$

[35] Na engleskom: *chopping*.

Za odgovarajuću granicu relativne greške dobijamo

$$\left|\frac{\text{chop}(x)-x}{x}\right| \leq \frac{b^{-n+k}}{b^{-1}b^{k}} = b^{-n+1},$$

tj. veličinu e_M definisanu ranije (videti (2.2.4)).

Bolja aproksimacija se može dobiti tzv. postupkom *zaokrugljivanja*[36], koji se sastoji u sledećem:

1° ako je $a_{n+1}+a_{n+2}b^{-1}+\cdots < \frac{1}{2}b$, koristi se prosto odsecanje;

2° ako je $a_{n+1}+a_{n+2}b^{-1}+\cdots > \frac{1}{2}b$, cifra a_n se povećava za jedinicu, a cifre $a_{n+1},\ a_{n+2},\ldots$ se odbacuju;

3° ako je $a_{n+1}+a_{n+2}b^{-1}+\cdots = \frac{1}{2}b$, ravnopravno se mogu koristiti oba prethodna pravila (1° ili 2°).

Međutim, na računskim mašinama zaokrugljivanje se najčešće izvodi tako što se broju x dodaje broj $\frac{1}{2}b^{-n+k}$, a zatim se vrši prosto odsecanje, tj.

$$\overline{x} = \text{rd}(x) = \pm(0.\tilde{a}_1\tilde{a}_2\ldots\tilde{a}_n)b^{\tilde{k}} = \text{chop}\left(x+\frac{1}{2}b^{-n+k}\right). \tag{2.2.7}$$

Ovo znači da se u nerešenom slučaju 3° uvek a_n zamenjuje sa a_n+1 (pravilo 2°), tj. $\tilde{a}_n = a_n+1$.

Primer 2.2.4. Izvršićemo zaokrugljivanje binarnog broja $(11111.11101011)_2$ na sedam cifara ($n=7$). Najpre, svodimo broj na normalizovani oblik,

$$x = 0.1111111101011\cdot 2^{k},$$

gde je $k = 5 = (101)_2$. Kako je

$$1+0\cdot 2^{-1}+1\cdot 2^{-2}+0\cdot 2^{-3}+1\cdot 2^{-4}+1\cdot 2^{-5} > \frac{1}{2}\cdot 2,$$

dodaje se jedinica na sedmu cifru posle binarne tačke, a cifre od osmog mesta pa nadalje (101011) se odbacuju. Dakle,

$$\overline{x} = (0.1111111+1\cdot 2^{-7})\cdot 2^{k} = 1.0000000\cdot 2^{k},$$

čiji normalizovani oblik je $\overline{x} = 0.1000000\cdot 2^{\tilde{k}}$, gde je eksponent $\tilde{k} = k+1 = 6 = (110)_2$. $\triangle$

[36] Na engleskom: *rounding*.

Napomenimo da se kod ručnog zaokrugljivanja brojeva u dekadnom sistemu ($b = 10$) u nerešenom slučaju 3° preporučuje sledeće pravilo: ako je cifra a_n paran broj koristiti pravilo 1°, a ako je neparan broj koristiti pravilo 2°.

Primer 2.2.5. Zaokrugljivanje broja $\pi = 3.141592653589793238\ldots$ na 3, 5, 7 i 10 decimala daje približne brojeve: 3.142, 3.14159, 3.1415927, 3.1415926536, respektivno. Međutim, sukcesivno zaokrugljivanje poslednjeg broja daje redom brojeve:

3.141592654,	3.1416,
3.14159265,	3.142,
3.1415926,	3.14,
3.141593,	3.1,
3.14159,	3.

Primetimo da se dobijeni približni brojevi zaokrugljeni na sedam decimala razlikuju na poslednjoj decimali, odakle zaključujemo da zaokrugljivanje ne treba izvoditi sukcesivno. $\triangle$

Nije teško uočiti da se kod procesa zaokrugljivanja brojeva čini greška

$$|\mathrm{rd}(x) - x| \leq \frac{1}{2} b^{-n+k}, \tag{2.2.8}$$

dok je granica relativne greške

$$\left| \frac{\mathrm{rd}(x) - x}{x} \right| \leq \frac{1}{2} b^{-n+1} = \frac{1}{2} e_M,$$

što je dva puta manje nego u slučaju prostog odsecanja.

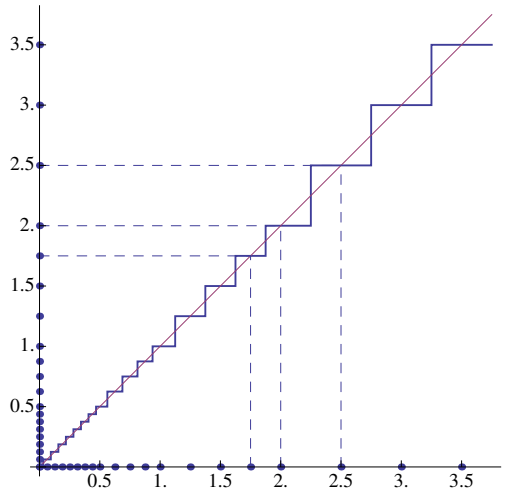
U svakom računaru na raspolaganju su samo brojevi iz konačnog skupa, tzv. skupa mašinskih brojeva, koji ćemo označavati sa $\widehat{\mathbb{R}}$. Dakle, $\widehat{\mathbb{R}} = A^* = A^*(b, n, k_{\min}, k_{\max}) \in \mathbb{R}$, što znači da se realni broj x mora aproksimirati na jedinstven način odgovarajućim mašinskim brojem $\widehat{x} \in \widehat{\mathbb{R}}$. Ovo preslikavanje označavamo[37] sa $f\ell \colon \mathbb{R} \to \widehat{\mathbb{R}}$ i ono se realizuje zaokrugljivanjem pomoću funkcije rd. Neka je realan broj $x \neq 0$ dat u normalizovanom obliku (2.2.5). Tada

$$f\ell(x) = \begin{cases} \mathrm{rd}(x), & k_{\min} - n + 1 \leq k \leq k_{\max}, \\ \widehat{0}, & k < k_{\min} - n + 1, \\ \widehat{\infty}, & k > k_{\max}. \end{cases} \tag{2.2.9}$$

[37] Oznaka $f\ell$ dolazi od engleske reči: *floating*.

Inače, za $x = 0$ imamo $f\ell(0) = \widehat{0}$.

Maksimalni eksponent $k_{\max}$ (i naravno dužina mantise n) određuju maksimalni realni broj $x_{\max}$ koji se može predstaviti mašinskim brojem. Za realne brojeve izvan intervala $(-x_{\max}, x_{\max})$ imamo prekoračenje.



Slika 2.2.3. Grafik funkcije $x \mapsto f\ell(x)$ za prošireni sistem $A^*(2,3,-1,2)$

Ilustracije radi vratimo se na primer 2.2.1, gde smo razmatrali konačni binarni sistem $A(2,3,-1,2)$ u kome je najveći pozitivan broj 0.111×2^2, tj. 3.5 u dekadnom sistemu. Ovim mašinskim brojem biće aproksimairani svi realni brojevi iz intervala $[0.1101 \times 2^2, 0.1111 \times 2^2)$, tj. $[3.25, 3.75)$ u dekadnom sistemu. Za brojeve $x \geq 3.75$ imamo prekoračenje. Grafik funkcije $x \mapsto f\ell(x)$ za prošireni sistem $A^*(2,3,-1,2)$ je prikazan na slici 2.2.3.

Drugi problem koji se pojavljuje ovde su aritmetičke operacije na skupu mašinskih brojeva $\widehat{\mathbb{R}}$. Čak i u slučaju mašinskih brojeva $\widehat{x}$ i $\widehat{y}$, rezultat operacije $\widehat{x} * \widehat{y}$, gde je $* \in \{+, -, \times, \div\}$, ne znači da će pripadati skupu mašinskih brojeva $\widehat{\mathbb{R}}$. Dve osobine funkcije $f\ell$ su evidentne: (1) $f\ell(x) = x$ ako je x mašinski broj; (2) $f\ell(x) \leq f\ell(y)$ ako je $x \leq y$ za svako $x, y \in \mathbb{R}$. Poslednja osobina je poznata kao *osobina monotonosti*.

Na osnovu (2.1.2) i granice relativne greške (2.2.8) zaključujemo da važi

$$r = \frac{f\ell(x) - x}{x}, \quad \text{tj.} \quad f\ell(x) = x(1+r),$$

gde je

$$|r| \leq \frac{1}{2} b^{-n+1} = \frac{1}{2} e_M.$$

1.2.3 Aritmetika konačne dužine i prostiranje grešaka u numeričkim procesima

Kao što smo napomenuli u prethodnom odeljku, kod izvođenja aritmetičkih operacija, čak i sa mašinskim brojevima, u opštem slučaju se ne dobija rezultat koji je mašinski broj. Zato je neophodno uvesti tzv. *aritmetiku konačne dužine*, tj. za svaku aritmetičku operaciju $* : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, gde je $* \in \{+, -, \times, \div\}$, uvesti odgovarajuću pseudo-operaciju na skupu mašinskih brojeva $\widehat{\mathbb{R}}$, $\boxed{*} : \widehat{\mathbb{R}} \times \widehat{\mathbb{R}} \to \widehat{\mathbb{R}}$, pomoću

$$x \boxed{*} y = f\ell(x * y) \quad (x, y \in \widehat{\mathbb{R}}),$$

gde je $\widehat{\mathbb{R}} = A^* = A^*(b, n, k_{\min}, k_{\max}) \in \mathbb{R}$. Ponekad se ova pseudo-operacija definiše i za brojeve iz $\mathbb{R}$, $\boxed{*} : \mathbb{R} \times \mathbb{R} \to \widehat{\mathbb{R}}$, pomoću $x \boxed{*} y = f\ell(f\ell(x) * f\ell(y))$. Dakle, ako su x i y mašinski brojevi, tada je

$$x \boxed{*} y = f\ell(x * y) = (x * y)(1 + r), \quad |r| \leq \frac{1}{2} e_M.$$

Na jednostavnom primeru pokazaćemo kako se kod pseudo-operacije oduzimanja $\boxed{-}$ može načiniti velika greška. Posmatrajmo sistem sa osnovom $b = 10$ i dužinom mantise $n = 2$, pretpostavljajući da nema ograničenja za eksponent. Neka su $x = 1$ i $y = 0.99$, čiji su mašinski reprezenti

$$\widehat{x} = f\ell(x) = 0.10 \times 10^1, \qquad \widehat{y} = f\ell(y) = 0.99 \times 10^0.$$

Dakle, imamo

$$\begin{array}{r} 0.10 \times 10^1 \\ \underline{-0.09 \times 10^1} \\ 0.01 \times 10^1 \end{array}$$

tj. $\widehat{x} \boxed{-} \widehat{y} = 0.10 \times 10^0$. Inače, tačan rezultat je deset puta manji. Da bi se dobio taj rezultat mora se uzeti veća dužina mantise.

Aritmetičke operacije se izvode u aritmetičkom organu (akumulatoru), u kome je za predstavljanje brojeva obezbeđena mantisa najčešće dvostruke dužine[38], a zatim se rezultat vraća na standardnu mantisu dužine n.

[38] Na engleskom: *double precision* accumulator.

Primer 2.3.1. Neka su $b = 10$, $n = 4$ i $\widehat{x} = 0.3947 \times 10^4$ i $\widehat{y} = 0.1372 \times 10^2$ Ako se operacija sabiranja izvodi u akumulatoru, u kome je za predstavljanje mantise brojeva obezbeđeno $2n = 8$ razreda imamo

$$\begin{array}{r} 0.39470000 \times 10^4 \\ \underline{+0.00137200 \times 10^4} \\ 0.39607200 \times 10^4 \end{array}$$

tj. $\widehat{x} \boxed{+} \widehat{y} = fl(\widehat{x}+\widehat{y}) = 0.3961 \times 10^4$, pri čemu je mašinska greška -0.28. Odgovarajuća relativna greška je -0.71×10^{-4}. $\triangle$

Za pseudo-aritmetičke operacije, u opštem slučaju, ne važi asocijativni zakon, tj.

$$fl\,(fl(\widehat{x} * \widehat{y}) * \widehat{z}) \neq fl\,(\widehat{x} * fl(\widehat{y} * \widehat{z}))\,.$$

Na primer, ako $*$ označava operaciju sabiranja, imamo

$$\begin{aligned} fl\,(fl(\widehat{x}+\widehat{y})+\widehat{z}) &= fl\,((\widehat{x}+\widehat{y})(1+r_1)+\widehat{z}) \\ &= ((\widehat{x}+\widehat{y})(1+r_1)+\widehat{z})\,(1+r_2) \end{aligned}$$

i

$$\begin{aligned} fl\,(\widehat{x}+fl(\widehat{y}+\widehat{z})) &= fl\,(\widehat{x}+(\widehat{y}+\widehat{z})(1+r_3)) \\ &= (\widehat{x}+(\widehat{y}+\widehat{z})(1+r_3))\,(1+r_4), \end{aligned}$$

gde su r_k, $k = 1,2,3,4$, relativne greške za koje važi

$$|r_k| \leq \frac{1}{2}\, b^{-n+1} = \frac{1}{2}\, e_M, \qquad k = 1,2,3,4.$$

U konkretnom slučaju, neka se operacija sabiranja izvodi u n-razrednom akumulatoru (*single precision* accumulator). Za $n = 3$ i brojeve $\widehat{x}_1 = 0.100 \times 10^{-2}$, $\widehat{x}_2 = -\widehat{x}_3 = 0.100 \times 10^1$ imamo

$$fl\,(fl(\widehat{x}_1+\widehat{x}_2)+\widehat{x}_3) = 0 \quad \text{i} \quad fl\,(\widehat{x}_1+fl(\widehat{x}_2+\widehat{x}_3)) = \widehat{x}_1.$$

U prethodnom razmatranju smo videli da je kod izvođenja operacije (tačnije rečeno pseudo-operacije) nad mašinskim brojevima, rezultat opterećen (relativnom) greškom čija granica ne prelazi vrednost $e_M/2$. Međutim, postoji dodatna greška, s obzirom na to da se ulazni podaci x i y najpre aproksimiraju odgovarajućim mašinskim brojevima $\widehat{x}$ i $\widehat{y}$, respektivno. Analiziraćemo sada kako se

odgovarajuće greške, $e_x = \widehat{x} - x$ i $e_y = \widehat{y} - y$, manifestuju u rezultatu $x * y$, pod uslovom da se operacija $* \in \{+, -, \times, \div\}$ izvršava tačno, tj. odredićemo grešku $e_{x*y} = \widehat{x} * \widehat{y} - x * y$.

1. Kod sabiranja imamo

$$\widehat{x} + \widehat{y} = (x + e_x) + (y + e_y) = (x + y) + (e_x + e_y),$$

tako da za grešku zbira važi $e_{x+y} = e_x + e_y$.

2. Kod oduzimanja imamo sličnu situaciju

$$e_{x-y} = e_x - e_y.$$

3. Za množenje važi

$$\widehat{x} \cdot \widehat{y} = (x + e_x)(y + e_y) = xy + ye_x + xe_y + e_x e_y,$$

tj.

$$e_{xy} = ye_x + xe_y + e_x e_y \approx ye_x + xe_y,$$

s obzirom da su greške e_x i e_y obično mnogo manje od samih veličina x i y.

4. Kod deljenja imamo

$$\frac{\widehat{x}}{\widehat{y}} = \frac{x + e_x}{y + e_y} = \frac{x + e_x}{y} \cdot \frac{1}{1 + e_y/y}.$$

Kako je $|e_y/y| \ll 1$, iz poslednje jednakosti sleduje

$$\frac{\widehat{x}}{\widehat{y}} = \frac{x + e_x}{y} \left(1 + \frac{e_y}{y} + \left(\frac{e_y}{y}\right)^2 + \cdots\right) \cong \frac{x}{y} + \frac{e_x}{y} - \frac{x}{y^2} e_y,$$

tj.

$$e_{x/y} \cong \frac{1}{y} e_x - \frac{x}{y^2} e_y,$$

pri čemu su uzete pretpostavke kao i kod množenja.

Odgovarajuće relativne greške su redom

$$r_{x+y} = \frac{e_{x+y}}{x+y} = \frac{x}{x+y} r_x + \frac{y}{x+y} r_y, \tag{2.3.1}$$

$$r_{x-y} = \frac{e_{x-y}}{x-y} = \frac{x}{x-y} r_x - \frac{y}{x-y} r_y, \tag{2.3.2}$$

$$r_{xy} = \frac{e_{xy}}{xy} \cong r_x + r_y, \tag{2.3.3}$$

$$r_{x/y} = \frac{e_{x/y}}{x/y} \cong r_x - r_y, \tag{2.3.4}$$

gde su $r_x = e_x/x$ i $r_y = e_y/y$.

Napomena 2.3.1. U slučajevima kada se vrednosti x i y ne znaju tačno, to se umesto njih u prethodno izvedenim izrazima za greške koriste odgovarajući mašinski brojevi $\widehat{x}$ i $\widehat{y}$.

Razmotrićemo sada opšti slučaj, sa operacijom $* \in \{+, -, \times, \div\}$,

$$u = x * y. \tag{2.3.5}$$

Na osnovu jednakosti (2.3.1) – (2.3.4), relativna greška rezultata operacije $*$ može se predstaviti u obliku

$$r_{x*y} = a_x r_x + a_y r_y,$$

gde koeficijenti a_x i a_y zavise od x i y i operacije $*$ (videti tabelu 2.3.1).

Tabela 2.3.1.

$*$	a_x	a_y
$+$	$\frac{x}{x+y}$	$\frac{y}{x+y}$
$-$	$\frac{x}{x-y}$	$\frac{-y}{x-y}$
$\times$	1	1
$\div$	1	-1

Kako je

$$\widehat{x} * \widehat{y} = (x * y)(1 + r_{x*y}),$$

zbog prisustva mašinske greške, imamo

$$\widehat{u} = f\ell(\widehat{x} * \widehat{y}) = (x * y)(1 + r_{x*y})(1 + r),$$

gde je r relativna mašinska greška kod izvođenja operacije $*$, koja je, kao što smo videli na početku ovog odeljka, ograničena sa $e_M/2$. Na osnovu poslednje jednakosti, za totalnu grešku e_u^t, dobijamo

$$e_u^t = \widehat{u} - u = u(r_{x*y} + r + r \cdot r_{x*y}).$$

Kako je, najčešće, $r \cdot r_{x*y}$ mnogo manje od r_{x*y} i r, u daljem razmatranju koristićemo približnu jednakost

$$e_u^t \cong e_u^T = u\,(r_{x*y} + r)\,.$$

Odgovarajuća relativna greška je

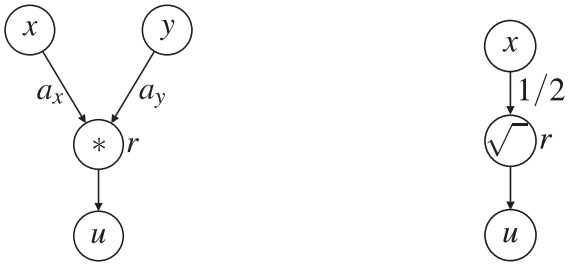
$$r_u^T = \frac{e_u^T}{u} = r_{x*y} + r,$$

tj.

$$r_u^T = a_x r_x + a_y r_y + r. \tag{2.3.6}$$

Radi nalaženja totalne greške nekog numeričkog postupka i odgovarajuće analize, numeričke postupke možemo predstavljati pomoću grafova. Kako se numerički postupak sastoji iz konačnog broja elementarnih operacija, dovoljno je znati kako se elementarna aritmetička operacija predstavlja pomoću grafa.

Graf računske operacije (2.3.5) (videti sl. 2.3.1 (levo)) u simboličkom obliku sadrži jednakost (2.3.6). Relativna greška zaokrugljivanja r rezultata operacije $*$, upisuje se pored temena grafa koje označava operaciju $*$. Smisao koeficijenta grane u grafu je u tome da relativna greška operanda ulazi u rezultat operacije pomnožena koeficijentom grane. Detaljniju analizu prostiranja grešaka u numeričkim procesima korišćenjem grafova opisao je BAUER[39] [4].



Slika 2.3.1. Graf aritmetičke operacije (2.3.5) (levo) i graf unarne operacije $\sqrt{x}$ (desno)

[39] FRIEDRICH LUDWIG BAUER (1924 –), nemački naučnik u oblasti kompjuterskih nauka i dugogodišnji urednik poznatog Springerovog časopisa *Numerische Mathematik*. Sada je profesor emeritus na Tehničkom univerzitetu u Minhenu.

Napomena 2.3.2. Unarne operacije se mogu, takođe, prikazivati pomoću grafova. Pokazaćemo to na primeru $u = \sqrt{x}$, mada, kao što smo videli u odeljku 1.1.1, simbol $\sqrt{\ }$ ne daje numerički postupak za izračunavanje vrednosti $\sqrt{x}$. Međutim, ova činjenica ne utiče na mogućnost analize greške. Naime, koristeći standardna označavanja, imamo

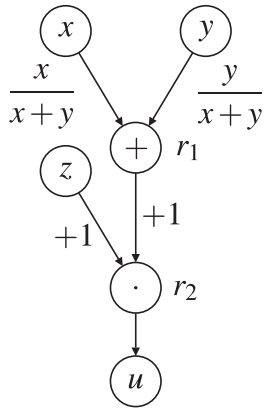
$$\widehat{u} = \sqrt{\widehat{x}}(1+r) = \sqrt{x(1+r_x)}(1+r) \cong \sqrt{x}\left(1+\frac{1}{2}r_x\right)(1+r),$$

tj. $r_u^T = \frac{1}{2}r_x + r$. Na osnovu ove jednakosti dobija se graf unarne operacije $u = \sqrt{x}$ (sl. 2.3.1 (desno)). Relativna greška r, najčešće, nije veća od $b^{-n+1} = e_M$, gde je n broj cifara mantise (videti [44]).

Primer 2.3.2. Ako su poznate približne vrednosti brojeva x, y i z sa relativnim greškama zaokrugljivanja r_x, r_y i r_z respektivno, odredićemo relativnu grešku u rezultatu

$$u = (x+y)z. \tag{2.3.7}$$

Neka su relativne mašinske greške operacija sabiranja i množenja redom r_1 i r_2. Graf odgovarajućeg računskog postupka dat je na sl. 2.3.2. Na osnovu grafa imamo redom



Slika 2.3.2. Graf za (2.3.7)

$$r_{x+y}^T = \frac{x}{x+y}r_x + \frac{y}{x+y}r_y + r_1,$$

$$\begin{aligned} r_u^T &= 1\cdot r_{x+y}^T + 1\cdot r_z + r_2 \\ &= \frac{x}{x+y}r_x + \frac{y}{x+y}r_y + r_z + r_1 + r_2. \end{aligned}$$

Kako su sve relativne greške zaokrugljivanja po apsolutnoj vrednosti manje od $\frac{1}{2}\cdot b^{-n+1} = e_M$, gde je n broj cifara mantise, dobijamo ocenu

$$|r_u^T| \le \left(\left|\frac{x}{x+y}\right| + \left|\frac{y}{x+y}\right| + 3\right)e_M.$$

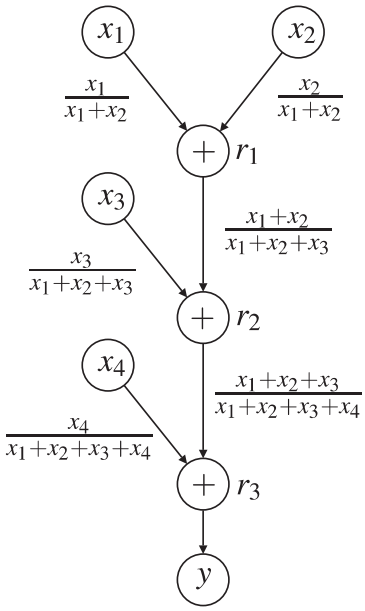
Ako su brojevi x i y istog znaka, prethodna nejednakost se svodi na

$$|r_u^T| \le 2\cdot b^{-n+1} = 4e_M. \quad \triangle$$

Primer 2.3.3. Izvršimo sada analizu greške kod izračunavanja zbira

$$y = x_1 + x_2 + x_3 + x_4, \tag{2.3.8}$$

pri čemu su $0 < x_1 < x_2 < x_3 < x_4$. Jednostavnosti radi, pretpostavimo da su brojevi x_i $(i = 1,2,3,4)$ dati tačno, tj. da su predstavljeni mašinskim brojevima bez zaokrugljivanja. Neka su dalje relativne mašinske greške posle svake operacije sabiranja redom $r_1,\ r_2,\ r_3$. Graf računskog postupka (2.3.8) je dat na sl. 2.3.3. Kako je $r_{x_i} = 0$ $(i = 1,2,3,4)$, na osnovu grafa dobijamo redom



Slika 2.3.3. Graf računskog procesa iz primera 2.3.3

$$r^T_{x_1+x_2} = r_1,$$

$$r^T_{x_1+x_2+x_3} = \frac{x_1+x_2}{x_1+x_2+x_3} r_1 + r_2,$$

$$r^T_y = \frac{x_1+x_2+x_3}{x_1+x_2+x_3+x_4}\left(\frac{x_1+x_2}{x_1+x_2+x_3} r_1 + r_2\right) + r_3,$$

odakle je

$$e_y^T = (x_1+x_2)r_1 + (x_1+x_2+x_3)r_2 + (x_1+x_2+x_3+x_4)r_3. \tag{2.3.9}$$

S obzirom na to da je $|r_i| \le \frac{1}{2}\cdot b^{-n+1} = e_M$ (n – broj cifara mantise), iz (2.3.9) sleduje

$$|e_y^T| \le (3x_1+3x_2+2x_3+x_4)e_M,$$

odakle zaključujemo da je granica apsolutne greške rezultata y minimalna ukoliko se sabiranje izvodi polazeći od najmanjih brojeva.

Slično se može pokazati da kod sabiranja m pozitivnih brojeva $x_1,\dots,x_m$ važi ocena

$$e_y^T = [(m-1)x_1 + (m-1)x_2 + (m-2)x_3 + \cdots + 2x_{m-1} + x_m]\, e_M. \quad \triangle$$

Primer 2.3.4. Izračunajmo zbir brojeva

$$\begin{aligned} x_1 &= 0.1376\cdot 10^0, \quad & x_2 &= 0.4737\cdot 10^0, \quad & x_3 &= 0.742810^1,\\ x_4 &= 0.6439\cdot 10^2, \quad & x_5 &= 0.5763\cdot 10^0, \quad & x_6 &= 0.203410^4, \end{aligned}$$

zaokrugljujući sve međurezultate na četiri značajne cifre.

Sabiranjem brojeva u datom redosledu, imamo redom

$$\begin{aligned} u_1 &= x_1 = 0.1376\cdot 10^0,\\ u_2 &= f\ell(u_1+x_2) = 0.6113\cdot 10^0,\\ u_3 &= f\ell(u_2+x_3) = 0.8039\cdot 10^1,\\ u_4 &= f\ell(u_3+x_4) = 0.7243\cdot 10^2,\\ u_5 &= f\ell(u_4+x_5) = 0.6487\cdot 10^3,\\ u_6 &= f\ell(u_5+x_6) = 0.2683\cdot 10^4, \end{aligned}$$

tj. zbir je $u = u_6 = 0.2683\cdot 10^4$.

Ako sabiranje izvodimo u obrnutom redosledu, imamo

$$\begin{aligned} v_1 &= x_6 = 0.2034\cdot 10^4,\\ v_2 &= f\ell(v_1+x_5) = 0.2610\cdot 10^4,\\ v_3 &= f\ell(v_2+x_4) = 0.2674\cdot 10^4,\\ v_4 &= f\ell(v_3+x_3) = 0.2681\cdot 10^4,\\ v_5 &= f\ell(v_4+x_2) = 0.2681\cdot 10^4,\\ v_6 &= f\ell(v_5+x_1) = 0.2681\cdot 10^4, \end{aligned}$$

tj. zbir je $v = v_6 = 0.2681 \cdot 10^4$.

Tačan zbir je $s = 0.26827293 \cdot 10^4$. Odgovarajuće greške dobijenih zbirova u i v su redom

$$e_1 = u - s \cong 0.27 \qquad \text{i} \qquad e_2 = v - s \cong -1.73. \quad \triangle$$

Primer 2.3.5. Neka su brojevi iz primera 2.3.3 pozitivni i bliski po vrednostima, tj. $x_i = x_0 + \delta_i$, $|\delta_i| \ll x_0$ $(i = 1,2,3,4)$. Korišćenjem rezultata iz pomenutog primera, zaključujemo da je

$$|e_y^T| \leq (9x_0 + 3|\delta_1| + 3|\delta_2| + 2|\delta_3| + |\delta_4|)\, e_M,$$

tj. $|e_y^T| \leq 9x_0 e_M$, s obzirom na pretpostavku $|\delta_i| \ll x_0$ $(i = 1,2,3,4)$.

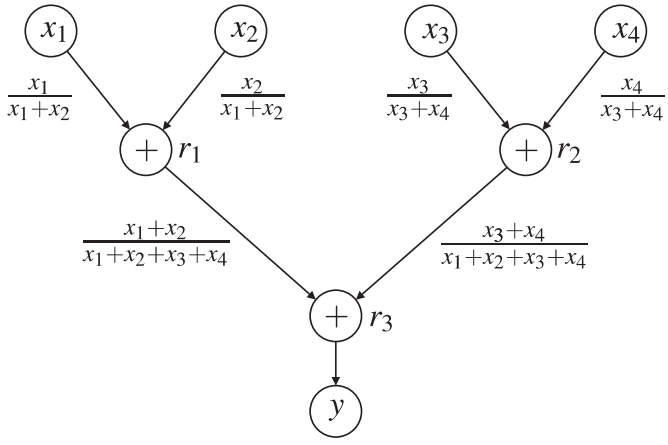
Izmenimo sada redosled izračunavanja. Naime, neka je

$$y = (x_1 + x_2) + (x_3 + x_4)\,.$$

Na osnovu grafa sa sl. 2.3.4 imamo

$$r_y^T = \frac{x_1 + x_2}{x_1 + x_2 + x_3 + x_4} r_1 + \frac{x_3 + x_4}{x_1 + x_2 + x_3 + x_4} r_2 + r_3,$$

odakle je $|e_y^T| \leq (2x_1 + 2x_2 + 2x_3 + 2x_4)\, e_M$, tj. $|e_y^T| \leq 8x_0 e_M$. Dakle, na ovaj način



Slika 2.3.4. Graf računskog procesa iz primera 2.3.5

se smanjuje granica apsolutne greške zbira četiri bliska pozitivna broja.

U opštem slučaju, ako imamo m^2 pozitivnih brojeva, približno jednakih po veličini, koje treba sabrati, granica apsolutne greške biće manja ukoliko brojeve grupišemo u m grupa po m brojeva i sabiramo brojeve u okviru svake grupe, a zatim sabiramo dobijene zbirove. $\triangle$

1.2.4 Uslovljenost i stabilnost numeričkih procesa

Uvodeći pojam algoritma, u odeljku 1.1.1 posmatrali smo transformaciju $P(\boldsymbol{x}) \xrightarrow{A} \boldsymbol{y}$. Pod pretpostavkom da su $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ i $\boldsymbol{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m$ povezani preslikavanjem $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, u oznaci $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$, tj.

$$\text{(2.4.1)} \qquad \begin{cases} y_1 = f_1(x_1, x_2, \ldots, x_n), \\ y_2 = f_2(x_1, x_2, \ldots, x_n), \\ \quad \vdots \\ y_m = f_m(x_1, x_2, \ldots, x_n), \end{cases}$$

gde su $f_k : \mathbb{R}^n \to \mathbb{R}$, $k = 1, 2, \ldots, m$, funkcije od n promenljivih, naš cilj ovde je da razmotrimo kakva je osetljivost preslikavanja $\boldsymbol{f}$ u nekoj tački $\boldsymbol{x}$ pri malim promenama (perturbacijama) u koordinatama tačke $\boldsymbol{x}$, tj. kakve su promene u koordinatama tačke $\boldsymbol{y}$ u poređenju sa perturbacijama u $\boldsymbol{x}$. Za merenje takvog „stepena osetljivosti“ uvodimo tzv. *faktor uslovljenosti* ili *kondicioni broj*[40] preslikavanja $\boldsymbol{f}$ u tački $\boldsymbol{x}$, u oznaci $k(\boldsymbol{f})(\boldsymbol{x})$ ili $(\operatorname{cond} \boldsymbol{f})(\boldsymbol{x})$, pretpostavljajući pri tome da se $\boldsymbol{f}$ izvršava tačno, tj. da se izračunava u aritmetici beskonačne dužine. Ovo znači da je faktor uslovljenosti $\boldsymbol{f}$ lokalno svojstvo preslikavanja koje ne zavisi od algoritma sa kojim se izračunava $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$.

Perturbacije u $\boldsymbol{x}$, po pravilu, nastaju kada se njegove koordinate aproksimiraju mašinskim brojevima (pomoću funkcije (2.2.9)), kako je to objašnjeno u odeljku 1.2.2 (videti, takođe, i odeljak 1.2.3). Dakle, u izračunavanju se umesto $\boldsymbol{x}$ pojavljuje njoj „bliska“ tačka $\overline{\boldsymbol{x}}$, gde je $\overline{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\delta}$, pri čemu se „rastojanje“ između tačaka $\boldsymbol{x}$ i $\overline{\boldsymbol{x}}$, tj. mera odstupanja $\boldsymbol{\delta}$ od nule $(0, 0, \ldots, 0)$, može iskazati u terminima mašinske preciznosti e_M. Tada, čak i tačno izračunavanje funkcije $\boldsymbol{f}$ ne dovodi do vrednosti $\boldsymbol{y}$, već do $\overline{\boldsymbol{y}}$, tj. do vrednosti $\overline{\boldsymbol{y}} = \boldsymbol{f}(\overline{\boldsymbol{x}})$.

Dakle, ako znamo kako preslikvanje $\boldsymbol{f}$ reaguje na male promene kao što je $\boldsymbol{\delta}$, tada možemo reći nešto o greški $\overline{\boldsymbol{y}} - \boldsymbol{y}$ u rešenju $\overline{\boldsymbol{y}}$, koja je uzrokovana tom

[40] Na engleskom: *condition number*.

promenom. Analiziraćemo sada posebno faktor uslovljenosti preslikavanja $\boldsymbol{f}$, a zatim i uslovljenost samog algoritma.

Faktor uslovljenosti preslikavanja $\boldsymbol{f}$. Startovaćemo sa najjednostavnijim slučajem funkcije jedne realne promenljive. Dakle, uzmimo $n = m = 1$, tj. $y = f(x)$.

Pretpostavimo, najpre, da su $x \neq 0$ i $y \neq 0$. Sa Δx označimo male promene od x. Pod pretpostavkom da je funkcija f diferencijabilna u tački x, zamenom priraštaja funkcije njenim diferencijalom, dobijamo

$$\Delta y = f(x+\Delta x) - f(x) \approx \mathrm{d}y = f'(x)\Delta x.$$

S obzirom na to da nas interesuju relativne greške, formulu možemo predstaviti u obliku

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

Ova približna jednakost postaje (tačna) jednakost ako je f linearna funkcija ili u graničnom slučaju kada $\Delta x \to 0$. To sugeriše definisanje uslovljenosti preslikavanja f u tački x pomoću

$$(\mathrm{cond}\, f)(x) = \left| \frac{xf'(x)}{f(x)} \right|. \tag{2.4.2}$$

Ovako definisan faktor uslovljenosti (ili kondicioni broj) pokazuje koliko puta je veća relativna promena y u odnosu na relativnu promenu x. Što je ovaj broj veći kažemo da je preslikavanje f slabije uslovljeno. Obrnuto, što je on manji to je preslikavanje bolje uslovljeno.

U slučaju kada je $x = 0$, a $y \neq 0$, faktor uslovljenosti definišemo sa $|f'(x)/f(x)|$. Slično, za $y = 0$ i $x \neq 0$, faktor uslovljenosti je $|xf'(x)|$. Najzad, ako su $x = y = 0$, faktor uslovljenosti bi bio samo $|f'(x)|$.

Analizirajmo sada opšti slučaj kada su n i m proizvoljni prirodni brojevi, tj. slučaj preslikavanja datog sa (2.4.1), pretpostavljajući da svako od m funkcija f_i, $i = 1,2,\ldots,m$, ima parcijalne izvode u odnosu na n promenljivih u tački $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

Ako imamo promenu u koordinati x_j, na osnovu (2.4.2), promena u funkcijama f_i, $i = 1,2,\ldots,m$, se može okarakterisati sledećim vrednostima

$$\gamma_{i,j}(\boldsymbol{x}) = (\mathrm{cond}_{ij}\, \boldsymbol{f})(\boldsymbol{x}) = \left| \frac{x_j}{f_i(\boldsymbol{x})} \frac{\partial f_i(\boldsymbol{x})}{\partial x_j} \right|. \tag{2.4.3}$$

Ovim dobijamo kompletnu matricu faktora uslovljenosti $\Gamma(\boldsymbol{x}) = [\gamma_{ij}(\boldsymbol{x})]_{m\times n}$. Slično, kao i u jednodimenzionalnom slučaju, zamenom priraštaja svake funkcije (Δy_i) odgovarajućim diferencijalom ($\mathrm{d}y_i$), dobijamo

$$\Delta y_i = f_i(\boldsymbol{x}+\Delta\boldsymbol{x}) - f_i(\boldsymbol{x}) \approx \mathrm{d}y_i = \sum_{j=1}^{n} \frac{\partial f_i(\boldsymbol{x})}{\partial x_j}\Delta x_j, \qquad i = 1,2,\ldots,m,$$

tj.

$$|\Delta y_i| \le \sum_{j=1}^{n} \left|\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}\right| \cdot |\Delta x_j|, \quad i = 1,2,\ldots,m. \tag{2.4.4}$$

Pod pretpostavkama da su sve koordinate x_j i sve vrednosti funkcija $y_i = f_i(\boldsymbol{x})$ različite od nule, imamo

$$\frac{|\Delta y_i|}{|y_i|} \le \sum_{j=1}^{n} \left|\frac{x_j}{f_i(\boldsymbol{x})}\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}\right| \cdot \frac{|\Delta x_j|}{|x_j|} = \sum_{j=1}^{n} \gamma_{i,j}\frac{|\Delta x_j|}{|x_j|}, \quad i = 1,2,\ldots,m, \tag{2.4.5}$$

što se može predstaviti i u matričnom obliku $\boldsymbol{r_y} \le \Gamma(\boldsymbol{x})\boldsymbol{r_x}$, gde su koordinate vektora $\boldsymbol{r_x}$ i $\boldsymbol{r_y}$, u stvari, relative promene (perturbacije) u koordinatama $\boldsymbol{x}$ i $\boldsymbol{y}$, respektivno.

Kako je, na osnovu (2.4.5),

$$\left|\frac{\Delta y_i}{y_i}\right| \le \left(\max_{1\le j\le n}\left|\frac{\Delta x_j}{x_j}\right|\right) \sum_{j=1}^{n}\left|\frac{x_j}{f_i(\boldsymbol{x})}\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}\right|, \quad i = 1,2,\ldots,m,$$

imamo

$$\max_{1\le i\le m}\left|\frac{\Delta y_i}{y_i}\right| \le \left(\max_{1\le j\le n}\left|\frac{\Delta x_j}{x_j}\right|\right)\left\{\max_{1\le i\le m}\sum_{j=1}^{n}\left|\frac{x_j}{f_i(\boldsymbol{x})}\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}\right|\right\},$$

tj. $\|\boldsymbol{r_y}\|_\infty \le \|\Gamma(\boldsymbol{x})\|_\infty \|\boldsymbol{r_x}\|_\infty$, gde su

$$\|\boldsymbol{r_x}\|_\infty = \max_{1\le j\le n}\left|\frac{\Delta x_j}{x_j}\right|, \quad \|\boldsymbol{r_y}\|_\infty = \max_{1\le i\le m}\left|\frac{\Delta y_i}{y_i}\right|. \tag{2.4.6}$$

Ograničenje za $\|\boldsymbol{r_y}\|_\infty/\|\boldsymbol{r_x}\|_\infty$, označeno sa $\|\Gamma(\boldsymbol{x})\|_\infty$, pruža nam mogućnost da uvedemo jedinstveni faktor uslovljenosti preslikavanja (2.4.1) pomoću jedne mere matrice $\Gamma(\boldsymbol{x})$, tzv. norme (za detalje videti odeljak 2.3.6),

$$(\mathrm{cond}\,\boldsymbol{f})(\boldsymbol{x}) = \|\Gamma(\boldsymbol{x})\|, \quad \Gamma(\boldsymbol{x}) = [\gamma_{ij}(\boldsymbol{x})]_{m\times n},$$

što je u našem slučaju

$$(\mathrm{cond}\,\boldsymbol{f})(\boldsymbol{x}) = \|\Gamma(\boldsymbol{x})\|_\infty = \max_{1\le i\le m}\sum_{i=1}^{n}\left|\frac{x_j}{f_i(\boldsymbol{x})}\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}\right|. \tag{2.4.7}$$

Faktor uslovljenosti definisan na ovaj način, zavisi od uvedene norme, ali je njegov red manje-više isti za bilo koju razumnu normu.

Napomena 2.4.1. Ako su koordinate od $\boldsymbol{x}$ ili od $\boldsymbol{y}$ jednake nuli, elementi matrice $\Gamma(\boldsymbol{x})$, dati pomoću (2.4.3), se mogu modifikovati na isti način kako je to prethodno urađeno za jednodimenzionalni slučaj.

Nešto grublja analiza uslovljenosti presikavanja (2.4.1), slična onoj za jednodimenzionalni slučaj, može se izvesti definisanjem relativnih promena za $\boldsymbol{x}$ i $\boldsymbol{y}$, pomoću $\|\Delta\boldsymbol{x}\|_\infty/\|\boldsymbol{x}\|_\infty$ i $\|\Delta\boldsymbol{y}\|_\infty/\|\boldsymbol{y}\|_\infty$, respektivno, što se razlikuje od (2.4.6). U ovom slučaju, direktno iz (2.4.4) nalazimo

$$\max_{1\le i\le m}|\Delta y_i| \le \left(\max_{1\le j\le n}|\Delta x_j|\right)\left\{\max_{1\le i\le m}\sum_{j=1}^{n}\left|\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}\right|\right\},$$

tj.

$$\|\Delta\boldsymbol{y}\|_\infty \le \left\|\frac{\partial\boldsymbol{f}(\boldsymbol{x})}{\partial\boldsymbol{x}}\right\|_\infty \|\Delta\boldsymbol{x}\|_\infty,$$

gde je

$$\frac{\partial\boldsymbol{f}(\boldsymbol{x})}{\partial\boldsymbol{x}} = \begin{bmatrix} \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \cdots & \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \dfrac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial f_2(\boldsymbol{x})}{\partial x_2} & & \dfrac{\partial f_2(\boldsymbol{x})}{\partial x_n} \\ \vdots & & & \\ \dfrac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial f_m(\boldsymbol{x})}{\partial x_2} & & \dfrac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$

JACOBIeva[41] matrica preslikavanja $\boldsymbol{f}$. Na osnovu prethodnog, jednostavno dobijamo

[41] CARL GUSTAV JACOB JACOBI (1804 – 1851), veliki nemački matematičar, poznat po doprinosima u oblasti eliptičkih funkcija, diferencijalnih jednačina i teorije brojeva.

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

tako da možemo uvesti alternativni faktor uslovljenosti pomoću

$$(\mathrm{cond}\, \boldsymbol{f})(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|_\infty \|\partial \boldsymbol{f}/\partial \boldsymbol{x}\|_\infty}{\|\boldsymbol{f}(\boldsymbol{x})\|_\infty} \tag{2.4.8}$$

Napomena 2.4.2. Jasno je da se u slučaju $m = n = 1$, ova definicija svodi na definiciju (2.4.2) (kao i na (2.4.7) datu ranije). Za veće dimenzije (n i/ili m veće od 1), međutim, faktor uslovljenosti u (2.4.8) je mnogo grublji nego onaj u (2.4.7). Ovo je jasno iz činjenice da u slučaju kada su koordinate sa prilično različitim odstupanjima, onda je norma $\|\boldsymbol{x}\|_\infty$ jednaka najvećoj od ovih koordinata uzetih po modulu, dok se sve ostale koordinate ignorišu. Zato rad sa ovako definisanim faktorom uslovljenosti zahteva opreznost!

Faktor uslovljenosti algoritma. Posmatrajmo problem definisan pomoću preslikavanja $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ datog sa $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ u (2.4.1). Pretpostavimo da ovaj problem rešavamo algoritmom A koji, na osnovu $\widehat{\boldsymbol{x}}$, daje rezultat $\widehat{\boldsymbol{y}}_A$, naravno, sve u aritmetici konačne dužine. Drugim rečima, sve koordinate $\boldsymbol{x}$ se najpre aproksimiraju mašinskim brojevima, dajući na taj način $\widehat{\boldsymbol{x}}$, a zatim se drugim preslikavanjem $\boldsymbol{f}_A : \widehat{\mathbb{R}}^n \to \widehat{\mathbb{R}}^m$, koje opisuje algoritam za rešavanje problema, nalazi rezultat $\widehat{\boldsymbol{y}}_A = \boldsymbol{f}_A(\widehat{\boldsymbol{x}})$, pri čemu smo sa $\widehat{\mathbb{R}}$, kao i ranije, označili skup mašinskih brojeva. Preslikavanje $\boldsymbol{f}_A$ je, dakle, jedna aproksimacija osnovnog preslikavanja $\boldsymbol{f}$.

U daljoj analizi pretpostavićemo da za svako $\widehat{\boldsymbol{x}} \in \widehat{\mathbb{R}}^n$ važi $\boldsymbol{f}_A(\widehat{\boldsymbol{x}}) = \boldsymbol{f}(\boldsymbol{x}_A)$ za neko $\widehat{\boldsymbol{x}}_A \in \mathbb{R}^n$, tj. da izračunata vrednost koja odgovara ulaznoj veličini $\widehat{\boldsymbol{x}}$ je tačno rešenje za neku drugu ulaznu veličinu $\boldsymbol{x}_A$, koja ne mora biti jedinstvena i ne mora pripadati $\widehat{\mathbb{R}}^n$.

Faktor uslovljenosti algoritma A se precizno definiše u terminima bliskosti $\boldsymbol{x}_A$ sa $\widehat{\boldsymbol{x}}$ (mereno relativnim odstupanjem sa pogodno odabranom normom), u poređenju sa mašinskom preciznošću e_M upotrebljene aritmetike konačne dužine (videti, na primer, [25, str. 36])

$$(\mathrm{cond}\, A)(\widehat{\boldsymbol{x}}) = \inf_{\boldsymbol{x}_A} \frac{\|\boldsymbol{x}_A - \widehat{\boldsymbol{x}}\|/\|\widehat{\boldsymbol{x}}\|}{e_M},$$

gde je infimum uzet preko svih $\boldsymbol{x}_A$ koji zadovoljavaju $\widehat{\boldsymbol{y}}_A = \boldsymbol{f}_A(\boldsymbol{x}_A)$ (ako $\boldsymbol{x}_A$ nije jedinstveno). Međutim, u primenama se, zbog jednostavnosti, može uzeti bilo koje $\boldsymbol{x}_A$, što daje nešto veću vrednost za faktor uslovljenosti

$$(\mathrm{cond}\, A)(\widehat{\boldsymbol{x}}) \approx \frac{\|\boldsymbol{x}_A - \widehat{\boldsymbol{x}}\|/\|\widehat{\boldsymbol{x}}\|}{e_M}. \tag{2.4.9}$$

Mašinsko (kompjutersko) rešenje problema i totalna greška. Posmatrajmo opet problem definisan pomoću preslikavanja $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, datog sa $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ u (2.4.1), koji se rešava algoritmom A u aritmetici konačne dužine sa mašinskom preciznošću e_M. U početnom problemu, koordinate od $\boldsymbol{x}$ su tačni realni brojevi, kao i koordinate tačnog rešenja $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$.

U procesu rešavanja ovog idealizovanog problema u aritmetici konačne dužine pomoću algoritma A, najpre se početni podaci $(\boldsymbol{x})$ zaokrugljuju na $\widehat{\boldsymbol{x}}$, tako da je relativna greška (u nekoj odabranoj normi) $\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|/\|\boldsymbol{x}\| \leq \varepsilon$, gde ε uključuje greške zaokrugljivanja, kao i sve druge tipove grešaka (na primer, greške uvedene merenjem kod eksperimentalnih podataka). Međutim, na tako dobijene podatke $\widehat{\boldsymbol{x}}$ se ne primenjuje originalno preslikavanje $\boldsymbol{f}$, već preslikavanje $\boldsymbol{f}_A$ koje daje rezultat $\widehat{\boldsymbol{y}}_A = \boldsymbol{f}_A(\widehat{\boldsymbol{x}})$.

Totalna greška koju želimo da ocenimo je tada

$$\frac{\|\widehat{\boldsymbol{y}}_A - \boldsymbol{y}\|}{\|\boldsymbol{y}\|}. \tag{2.4.10}$$

Neka su $\widehat{\boldsymbol{y}} = \boldsymbol{f}(\widehat{\boldsymbol{x}})$ i $\boldsymbol{f}_A(\widehat{\boldsymbol{x}}) = \boldsymbol{f}(\widehat{\boldsymbol{x}}_A)$. Nakon primene relacije trougla (videti definiciju 1.4.1 u odeljku 2.1.4) i aproksimacije $\|\boldsymbol{y}\| \approx \|\widehat{\boldsymbol{y}}\|$, za (2.4.10) se može dati granica u obliku

$$\frac{\|\widehat{\boldsymbol{y}}_A - \boldsymbol{y}\|}{\|\boldsymbol{y}\|} \leq \frac{\|\widehat{\boldsymbol{y}}_A - \widehat{\boldsymbol{y}}\|}{\|\boldsymbol{y}\|} + \frac{\|\widehat{\boldsymbol{y}} - \boldsymbol{y}\|}{\|\boldsymbol{y}\|} \approx \frac{\|\widehat{\boldsymbol{y}}_A - \widehat{\boldsymbol{y}}\|}{\|\widehat{\boldsymbol{y}}\|} + \frac{\|\widehat{\boldsymbol{y}} - \boldsymbol{y}\|}{\|\boldsymbol{y}\|}. \tag{2.4.11}$$

Da bismo ocenili veličine na desnoj strani u (2.4.11), primetimo najpre da je

$$\frac{\|\widehat{\boldsymbol{y}}_A - \widehat{\boldsymbol{y}}\|}{\|\widehat{\boldsymbol{y}}\|} = \frac{\|\boldsymbol{f}_A(\widehat{\boldsymbol{x}}) - \boldsymbol{f}(\widehat{\boldsymbol{x}})\|}{\|\boldsymbol{f}(\widehat{\boldsymbol{x}})\|} = \frac{\|\boldsymbol{f}(\widehat{\boldsymbol{x}}_A) - \boldsymbol{f}(\widehat{\boldsymbol{x}})\|}{\|\boldsymbol{f}(\widehat{\boldsymbol{x}})\|} \leq (\operatorname{cond}\boldsymbol{f})(\widehat{\boldsymbol{x}}) \cdot \frac{\|\widehat{\boldsymbol{x}}_A - \widehat{\boldsymbol{x}}\|}{\|\widehat{\boldsymbol{x}}\|},$$

odakle, korišćenjem (2.4.9), dobijamo

$$\frac{\|\widehat{\boldsymbol{y}}_A - \widehat{\boldsymbol{y}}\|}{\|\widehat{\boldsymbol{y}}\|} \leq (\operatorname{cond}\boldsymbol{f})(\widehat{\boldsymbol{x}}) \cdot (\operatorname{cond}A)(\widehat{\boldsymbol{x}}) \cdot e_M.$$

Slično, za drugi izraz na desnoj strani u (2.4.11) imamo

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

Najzad, pretpostavljajući da je $(\operatorname{cond}\boldsymbol{f})(\widehat{\boldsymbol{x}}) \approx (\operatorname{cond}\boldsymbol{f})(\boldsymbol{x})$, dobijamo granicu za totalnu grešku

$$\frac{\|\widehat{\boldsymbol{y}}_A - \boldsymbol{y}\|}{\|\boldsymbol{y}\|} \leq (\operatorname{cond} \boldsymbol{f})(\boldsymbol{x}) \left[\varepsilon + (\operatorname{cond} A)(\widehat{\boldsymbol{x}}) \cdot e_M\right]. \tag{2.4.12}$$

Dakle, formula (2.4.12) pokazuje koliko greške u ulaznim podacima (ε) i mašinska preciznost (e_M) utiču na totalnu grešku. Kao što vidimo, obe se množe faktorom uslovljenosti preslikavanja $\boldsymbol{f}$, ali se mašinska preciznost dodatno uvećava faktorom uslovljenosti algoritma A.

1.2.5 Statistički prilaz u oceni grešaka

U dosadašnjoj analizi grešaka uvek smo razmatrali tzv. „najgori slučaj“ – slučaj u kome su greške ulaznih podataka i greške zaokrugljivanja istog „smera“, tj. takve da obezbeđuju maksimalnu grešku izlazne informacije. Ovakav slučaj je malo verovatan.

U ovom odeljku ukazaćemo na statistički pristup u oceni greške na jednom jednostavnom primeru. Naime, neka je

$$u = \sum_{i=1}^{n} x_i$$

i neka su brojevi x_i dati svojim približnim vrednostima $\overline{x}_i$, pri čemu su odgovarajuće granice apsolutnih grešaka Δ_i, tj.

$$|e_i| = |\overline{x}_i - x_i| \leq \Delta_i \qquad (i = 1, \ldots, n).$$

Pod pretpostavkom da ne postoje greške zaokrugljivanja u procesu računanja, greška zbira će biti

$$e_u = \sum_{i=1}^{n} e_i,$$

odakle sleduje ocena

$$|e_u| \leq \sum_{i=1}^{n} |e_i| \leq \sum_{i=1}^{n} \Delta_i. \tag{2.5.1}$$

Dakle, ova ocena daje maksimalnu apsolutnu grešku, koja nastupa samo u dva slučaja, tj. kada je $e_i = \Delta_i$ $(i = 1, \ldots, n)$ ili kada je $e_i = -\Delta_i$ $(i = 1, \ldots, n)$. Naravno, sa stanovišta teorije verovatnoće i statistike, ovi slučajevi su malo verovatni. Da bismo mogli dobiti statističku ocenu za grešku e_u potrebno je uvesti pretpostavku o funkcijama raspodele grešaka e_i $(i = 1, \ldots, n)$. Pretpostavimo da su sve granice

apsolutnih grešaka međusobno jednake, tj. $\Delta_i = \Delta$ $(i = 1,\ldots,n)$ i da greške podležu ravnomernom zakonu raspodele u intervalu $(-\Delta,\Delta)$, tj. da je gustina verovatnoće svake od grešaka e_i $(i = 1,\ldots,n)$ jednaka $1/(2\Delta)$. U ovom slučaju, matematičko očekivanje $m(e_i)$ i disperzija $D(e_i)$ su dati sa

$$m(e_i) = \int_{-\Delta}^{+\Delta} \frac{t}{2\Delta}\,\mathrm{d}t = 0 \quad \text{i} \quad D(e_i) = \int_{-\Delta}^{+\Delta} \frac{t^2}{2\Delta}\,\mathrm{d}t = \frac{1}{3}\Delta^2.$$

Pretpostavljajući da su greške e_i nezavisne veličine, za disperziju sume dobijamo

$$D(e_u) = \sum_{i=1}^{n} D(e_i) = \frac{1}{3}n\Delta^2.$$

Da bismo dobili statističku ocenu za e_u potrebno je odrediti funkciju raspodele sume grešaka. S obzirom na glomaznost postupka za konstrukciju ovakve funkcije, pristupićemo izvesnoj aproksimaciji. Naime, pretpostavićemo da se suma grešaka ponaša po normalnom zakonu raspodele sa matematičkim očekivanjem 0 i standardnom devijacijom $\sigma = \Delta\sqrt{n/3}$. Tada je

$$P(|e_u| < 3\sigma) = 0.9973,$$

tj. sa verovatnoćom 0.9973 možemo očekivati da je

$$|e_u| < 3\sigma = \Delta\sqrt{3n}. \tag{2.5.2}$$

Dakle, za statističku granicu apsolutne sume grešaka možemo uzeti veličinu $\Delta\sqrt{3n}$ koja je pri $n > 3$ manja od granice Δ, dobijene na osnovu (2.5.1).

Primetimo da sa verovatnoćom 0.68 možemo očekivati da je $|e_u| < \sigma = \Delta\sqrt{n/3}$.

Primer 2.5.1. Kod sabiranja 75 brojeva, koji su svi dati sa tačnošću $5\cdot 10^{-4}$, granica apsolutne greške je $n\Delta = 75\cdot 5\cdot 10^{-4} = 3.75\cdot 10^{-2}$. Međutim, statistička granica, na osnovu (2.5.2), je

$$|e_u| < 5\cdot 10^{-4}\sqrt{3\cdot 75} = 0.75\cdot 10^{-2},$$

tj. pet puta manja. $\triangle$

1.3 REKURZIVNA IZRAČUNAVANJA I SUMIRANJE

Od izuzetnog značaja u numeričkoj matematici su rekurzivna izračunavanja i postupci sumiranja. Ovo poglavlje započinjemo izlaganjem osnovne teorije linearnih diferencnih jednačina na kojima se zasniva jedna važna klasa tzv. linearnih rekurzivnih postupaka. Jedna od glavnih primena ovih postupaka je u konstrukciji linearnih višekoračnih metoda za rešavanje CAUCHYevog problema kod diferencijalnih jednačina. Posebna pažnja je posvećena numeričkim aspektima tročlane rekurentne relacije koja se sreće u mnogim naučnim problemima.

Poseban tretman je dat izračunavanju vrednosti nekih elementarnih funkcija, koje se sreću kao bibliotečke funkcije kod računara. Imajući u vidu da se aproksimacije funkcija najčešće daju u obliku polinoma, racionalne funkcije ili verižnog razlomka, to se u posebnim odeljcima ovog poglavlja tretiraju algoritmi za izračunavanje vrednosti polinoma, uključujući i tzv. ekonomične šeme, algoritmi za izračunavanje verižnih razlomaka, kao i postupci za transformaciju racionalne funkcije u verižni razlomak i obrnuto. U posebnom odeljku, radi kompletnosti, izloženi su osnovni elementi verižnih razlomaka. Neki postupci za sumiranje redova i ubrzavanje konvergencije dati su u odeljku 1.3.5. Najzad, poslednja dva odeljka su posvećena asimptotskim razvojima koji često nalaze primenu kod izračunavanja vrednosti nekih specijalnih funkcija, posebno gama funkcije.

1.3.1 Diferencne jednačine

Kako se mnogi problemi u numeričkoj matematici svode na rešavanje diferencnih jednačina, u ovom odeljaku izložićemo neke osnovne pojmove i rezultate iz teorije ovih jednačina.

Definicija 3.1.1. Pod konačnom razlikom prvog reda funkcije f podrazumevamo izraz $\Delta f(x) = f(x+h) - f(x)$, gde je $h = \text{const} > 0$.

Definicija 3.1.2. Pod konačnom razlikom reda n funkcije f podrazumevamo izraz $\Delta^n f(x)$ dat rekurzivno pomoću

$$\Delta^n f(x) = \Delta^{n-1} f(x+h) - \Delta^{n-1} f(x) \qquad \left(\Delta^0 f(x) = f(x)\right).$$

Lako se mogu dokazati sledeće formule

$$\Delta^n f(x) = \sum_{i=0}^{n} (-1)^i \binom{n}{i} f(x+(n-i)h) \tag{3.1.1}$$

i

$$f(x+nh)=\sum_{i=0}^{n}\binom{n}{i}\Delta^i f(x).$$

Poslednja formula se može posmatrati i kao diskretni analogon TAYLORovoj formuli. Konstanta h se naziva korak. Često se u prethodnim definicijama uzima $h=1$.

Napomena 3.1.1. Slično prethodnim definicijama mogu se definisati konačne razlike niza $\{y_k\}_{k\in\mathbb{N}}$. Naime, imamo

$$\Delta^0 y_k = y_k, \quad \Delta^n y_k = \Delta^{n-1}y_{k+1} - \Delta^{n-1}y_k \quad (n=1,2,\ldots).$$

Primer 3.1.1. Konačne razlike funkcija f i g, koje su definisane pomoću $f(x)=ax^2+bx+c$ i $g(x)=\mathrm{e}^{ax}$, su redom

$$\Delta f(x)=h(2ax+ah+b), \quad \Delta^2 f(x)=2ah^2, \quad \Delta^n f(x)=0 \quad (n=3,4,\ldots)$$

i

$$\Delta^n g(x)=\mathrm{e}^{ax}\left(\mathrm{e}^{ah}-1\right)^n \qquad (n=1,2,\ldots).$$

△

Jednačina oblika

$$F\left(x;f(x),\Delta f(x),\ldots,\Delta^n f(x)\right)=0,$$

gde je f nepoznata funkcija, predstavlja diferencnu jednačinu reda n, ako posle transformacije pomoću (3.1.1) sadrži $f(x+nh)$ i $f(x)$; u protivnom jednačina je nižeg reda od n. Dakle, diferencna jednačina reda n ima oblik

$$G\left(x;f(x),f(x+h),\ldots,f(x+nh)\right)=0. \tag{3.1.2}$$

Primer 3.1.2. Diferencna jednačina $(h=1)$

$$\Delta^3 f(x)-3\Delta f(x)-2f(x)=x+2$$

je prvog reda jer se pomoću (3.1.1) svodi na

$$f(x+3)-3f(x+2)=x+2,$$

tj.

$$f(z+1)-3f(z)=z,$$

pri čemu smo uveli smenu $z=x+2$. △

Ako uvedemo smenu $x = x_0 + kh$ i označimo $y_k = f(x_0 + kh)$, tada (3.1.2) postaje

$$H(k; y_k, y_{k+1}, \ldots, y_{k+n}) = 0.$$

U našim razmatranjima od interesa su linearne diferencne jednačine, čiji je opšti oblik

$$y_{k+n} + b_1(k)y_{k+n-1} + \cdots + b_n(k)y_k = Q(k). \tag{3.1.3}$$

Ako je $Q(k) \equiv 0$ jednačina je homogena; u protivnom jednačina je nehomogena.

Posmatrajmo, najpre, linearnu diferencnu jednačinu prvog reda

$$y_{k+1} + b(k)y_k = Q(k). \tag{3.1.4}$$

Opšte rešenje homogene jednačine $y_{k+1} + b(k)y_k = 0$ je dato sa

$$y_k = (-1)^k C \prod_{i=0}^{k-1} b(i) \qquad (C \text{ proizvoljna konstanta}). \tag{3.1.5}$$

Ako pretpostavimo da C zavisi od k, tj. $C = C_k$, moguće je polazeći od rešenja (3.1.5) naći rešenje nehomogene jednačine (3.1.4). Naime, u tom slučaju, dobijamo da je rešenje nehomogene jednačine dato sa

$$y_k = (-1)^k \prod_{i=0}^{k-1} b(i) \left(C + \sum_{m=1}^{k} \frac{(-1)^m Q(m-1)}{\prod_{j=0}^{m-1} b(j)} \right),$$

gde je C proizvoljna konstanta.

Opšte rešenje linearne diferencne jednačine n-tog reda (3.1.3) je dato sa

$$y_k = C_1\Phi_1(k) + C_2\Phi_2(k) + \cdots + C_n\Phi_n(k) + f_p(k),$$

gde su $\Phi_1(k), \Phi_2(k), \ldots, \Phi_n(k)$ linearno nezavisna partikularna rešenja odgovarajuće homogene jednačine, a f_p jedno partikularno rešenje nehomogene jednačine. Konstante C_i $(i = 1, \ldots, n)$ su proizvoljne i u konkretnim slučajevima određuju se iz početnih uslova.

Daćemo sada neke napomene koje se odnose na određivanje opšteg rešenja linearne homogene diferencne jednačine sa konstantnim koeficijentima. Pretpostavimo da je rešenje ove jednačine oblika $y_k = \lambda^k$. Tada zamenom u odgovarajućoj jednačini dobijamo

$$\lambda^{k+n} + b_1\lambda^{k+n-1} + \cdots + b_n\lambda^k = 0.$$

Dakle, karakteristična jednačina je

$$\lambda^n + b_1\lambda^{n-1} + \cdots + b_n = 0.$$

U zavisnosti od korena karakteristične jednačine razlikovaćemo slučajeve:

a) ako su svi koreni $\lambda_1, \ldots, \lambda_n$ prosti, opšte rešenje je

$$y_k = C_1\lambda_1^k + \cdots + C_n\lambda_n^k;$$

b) ako su λ_p i λ_q konjugovano-kompleksni koreni, tj.

$$\lambda_p = \rho(\cos\theta + \mathrm{i}\sin\theta) \quad \text{i} \quad \lambda_q = \overline{\lambda}_p,$$

tada njima u opštem rešenju odgovara izraz

$$C_p\lambda_p^k + C_q\lambda_q^k = \rho^k(A_p\cos k\theta + A_q\sin k\theta);$$

c) ako je λ_i višestruki koren reda m, tada njemu u opštem rešenju odgovara izraz

$$\left(C_1 + C_2k + \cdots + C_mk^{m-1}\right)\lambda_i^k.$$

Kod linearnih nehomogenih jednačina sa konstantnim koeficijentima moguće je u izvesnim slučajevima naći partikularno rešenje f_p.

Ako je $Q(k) = P_s(k)$ (P_s polinom stepena s) i $\lambda = 1$ koren karakteristične jednačine reda m, tada se partikularno rešenje može tražiti u obliku

$$f_p(k) = A_0k^m + A_1k^{m+1} + \cdots + A_sk^{m+s}.$$

Ukoliko je $Q(k) = P_s(k)a^k$ i $\lambda = a$ koren karakteristične jednačine reda m, f_p treba tražiti u obliku

$$f_p(k) = a^k\left(A_0k^m + \cdots + A_sk^{m+s}\right).$$

Primer 3.1.3. Rešimo diferencnu jednačinu

$$y_{k+4} + 2y_{k+3} + 3y_{k+2} + 2y_{k+1} + y_k = 0,$$

pod uslovima $y_0 = y_1 = y_3 = 0$ i $y_2 = -1$.

Kako je karakteristična jednačina

$$\lambda^4+2\lambda^3+3\lambda^2+2\lambda+1=0,$$

tj. $\left(\lambda^2+\lambda+1\right)^2=0$, imamo $\lambda_1=\lambda_2=\frac{1}{2}(-1+\mathrm{i}\sqrt{3})$, $\lambda_3=\lambda_4=\frac{1}{2}(-1-\mathrm{i}\sqrt{3})$. Dakle, $\rho=1$, $\theta=2\pi/3$ pa je opšte rešenje dato pomoću

$$y_k=(C_1+C_2k)\cos\frac{2\pi k}{3}+(C_3+C_4k)\sin\frac{2\pi k}{3},$$

gde su C_i $(i=1,\ldots,4)$ proizvoljne konstante. Na osnovu datih uslova nalazimo $C_1=C_2=0$, $C_3=C_4=-2/\sqrt{3}$, tj.

$$y_k=\frac{2(k-1)}{\sqrt{3}}\sin\frac{2\pi k}{3}.$$

Primetimo da se ovo rešenje može predstaviti i u obliku

$$y_k=\begin{cases}0, & k=0\ (\mathrm{mod}\ 3),\\ k-1, & k=1\ (\mathrm{mod}\ 3),\\ 1-k, & k=2\ (\mathrm{mod}\ 3).\end{cases}\quad\triangle$$

Primer 3.1.4. Odredimo opšti član FIBONACCIevog[42] niza: $0,1,1,2,3,5,8,13,21,\ldots$ Kako je $y_{k+2}=y_k+y_{k+1}$, rešenje ove jednačine, pod uslovima $y_0=0$ i $y_1=1$, je

$$y_k=\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^k-\left(\frac{1-\sqrt{5}}{2}\right)^k\right].\quad\triangle$$

1.3.2 Rekurzivna izračunavanja i tročlana rekurentna relacija

Nalaženje opšteg rešenja diferencne jednačine nije uvek neophodno. Naime, daleko češće u numeričkoj matematici se sreće problem određivanja partikularnog rešenja koje zadovoljava određene uslove. Sa zadatim uslovima, diferencnu jednačinu nazivamo rekurentnom relacijom. U numeričkoj matematici od posebnog interesa su tzv. tročlane rekurentne relacije, koje se javljaju kod mnogih klasa specijalnih funkcija matematičke fizike i statistike, zatim u teoriji verižnih razlomaka i teoriji ortogonalnih polinoma, kod rešavanja linearnih diferencijalnih jednačina, kao i u mnogim problemima vezanih za konstrukciju redova. To su relacije oblika (3.1.3), sa $n=2$ i $Q(k)\equiv 0$, tj.

[42] LEONARDO PISANO BIGOLLO ($\approx$ 1170 – 1250), italijanski matematičar poznat, takođe, kao LEONARDO BONACCI, LEONARDO FIBONACCI, ili jednostavno kao FIBONACCI.

$$y_{k+1} + a_k y_k + b_k y_{k-1} = 0, \qquad k = 1,2,\dots, \tag{3.2.1}$$

gde smo indeks k, u odnosu na (3.1.3), smanjili za jedinicu, a koeficijente $b_1(k-1)$ i $b_2(k-1)$, uprošćenja radi, zamenili sa a_k i b_k, respektivno. Da bi relacija bila tročlana mora biti $b_k \neq 0$.

1° Ako poznajemo početne uslove (vrednosti) $y_0 = f_0$ i $y_1 = f_1$, pomoću rekurentne relacije (3.2.1) možemo odrediti niz $y_2, y_3, \dots, y_N$, gde je N proizvoljno izabran prirodan broj $(N > 1)$.

2° Ako poznajemo vrednosti $y_N = f_N$ i $y_{N-1} = f_{N-1}$, pomoću (3.2.1) predstavljene u obliku

$$y_{k-1} = -\frac{1}{b_k}\left(y_{k+1} + a_k y_k\right), \quad k = N-1, N-2, \dots, 1, \tag{3.2.2}$$

za naznačene vrednosti k dobijamo niz $y_{N-2},\ y_{N-3}, \dots,\ y_0$. Ovakvu relaciju nazivamo rekurentnom relacijom unazad.

3° Najsloženiji slučaj je ako su dati $y_0 = f_0$ i $y_N = f_N$, tj. ako imamo tzv. konturne uslove. Tada se ovaj problem može interpretirati kao trodijagonalni sistem linearnih jednačina

$$\begin{bmatrix} a_1 & 1 & & & \mathbf{0} \\ b_2 & a_2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & b_{N-2} & a_{N-2} & 1 \\ \mathbf{0} & & & b_{N-1} & a_{N-1} \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-2} \\ y_{N-1} \end{bmatrix} = \begin{bmatrix} -b_1 f_0 \\ 0 \\ \vdots \\ 0 \\ -f_N \end{bmatrix}. \tag{3.2.3}$$

Dakle, u slučajevima 1° i 2° problem je eksplicitan, dok je u slucaju 3° implicitan jer zahteva rešavanje pomenutog trodijagonalnog sistema linearnih jednačina.

Iako je, na prvi pogled, korišćenje rekurentne relacije u principu jednostavno, ono zaslužuje posebnu pažnju, s obzirom na numeričku nestabilnost koja se može manifestovati. Naime, svaki ciklus primene rekurentne relacije ne samo da generiše nove greške zaokrugljivanja, već i prihvata greške zaokrugljivanja iz prethodnog ciklusa. Tako pod izvesnim – nepovoljnim uslovima, prostiranje grešaka kroz računski proces može „astronomski“ ugroziti rezultat. Ovaj efekat ćemo kasnije ilustrovati na primeru izračunavanja BESSELove[43] funkcije prve vrste i reda k, definisane pomoću integrala

[43] FRIEDRICH WILHELM BESSEL (1784 – 1846), poznati nemački matematičar i astronom.

$$J_k(x) = \frac{1}{\pi}\int_0^{\pi} \cos\left(x\sin\theta - k\theta\right)\,\mathrm{d}\theta.$$

Može se pokazati da $J_k(x)$ zadovoljava diferencnu jednačinu

$$y_{k+1} - \frac{2k}{x}y_k + y_{k-1} = 0, \quad k = 1,2,\ldots, \tag{3.2.4}$$

tj. da važi

$$J_{k+1}(x) = \frac{2k}{x}J_k(x) - J_{k-1}(x), \quad k = 1,2,\ldots\ . \tag{3.2.5}$$

Dakle, za fiksirano x, $J_k(x)$ je jedno partikularno rešenje diferencne jednačine (3.2.4), određeno, na primer, početnim uslovima $y_0 = J_0(x)$ i $y_1 = J_1(x)$. Funkcija $J_k(x)$ može se izraziti u obliku stepenog reda

$$J_k(x) = \left(\frac{x}{2}\right)^k \sum_{m=0}^{+\infty} \frac{(-1)^m}{(m!)^2}\left(\frac{x}{2}\right)^{2m}.$$

Napomena 3.2.1. BESSELova funkcija prve vrste definiše se ne samo za realno, već i za kompleksno x, a pri tome red k ne mora biti nenegativan ceo broj, već može biti realan, ili čak kompleksan broj. Odgovarajuća reprezentacija BESSELove funkcije prve vrste u obliku reda ima oblik

$$J_\nu(z) = \left(\frac{z}{2}\right)^\nu \sum_{m=0}^{+\infty} \frac{(-1)^m}{(m!)\Gamma(\nu+m+1)}\left(\frac{z}{2}\right)^{2m}.$$

Napomena 3.2.2. BESSELova funkcija druge vrste definiše se pomoću

$$Y_\nu(x) = \frac{J_\nu(x)\cos(\nu\pi) - J_{-\nu}(x)}{\sin(\nu\pi)}$$

ako ν nije ceo broj, dok se za $\nu = k$ uzima

$$Y_k(x) = \lim_{\nu\to k} Y_\nu(x).$$

Inače, funkcija $Y_k(x)$ se može izraziti u obliku

$$\begin{aligned} Y_k(x) = &-\frac{1}{\pi}\left(\frac{2}{x}\right)^k \sum_{m=0}^{k-1} \frac{(k-m-1)!}{m!}\left(\frac{x}{2}\right)^{2m} + \frac{2}{\pi}J_k(x)\log\frac{x}{2} \\ &-\frac{1}{\pi}\left(\frac{x}{2}\right)^k \sum_{m=0}^{+\infty} \left(\psi(m+1)+\psi(k+m+1)\right)\frac{(-1)^m}{m!(k+m)!}\left(\frac{x}{2}\right)^{2m}, \end{aligned}$$

gde je funkcija ψ definisana kao logaritamski izvod gama funkcije, tj.

$$\psi(z) = \frac{\mathrm{d}}{\mathrm{d}z}\left(\log\Gamma(z)\right) = \frac{\Gamma'(z)}{\Gamma(z)}.$$

Ako je z ceo broj, imamo

$$\psi(1) = -\gamma, \qquad \psi(n) = -\gamma + \sum_{m=1}^{n-1} \frac{1}{m} \qquad (n \geq 2),$$

gde je γ EULERova[44] konstanta

$$\gamma = \lim_{n\to+\infty}\left(\sum_{m=1}^{n}\frac{1}{m} - \log n\right) = 0.57721566490115328606\ldots \ . \tag{3.2.6}$$

Prethodna formula za $Y_k(x)$ važi za $k \geq 1$. Kada je $k = 0$ imamo

$$Y_0(x) = \frac{2}{\pi}\left(\gamma + \log\frac{x}{2}\right)J_0(x) - \frac{2}{\pi}\sum_{m=1}^{+\infty}\frac{(-1)^m}{(m!)^2}\left(\gamma + \psi(m+1)\right)\left(\frac{x}{2}\right)^{2m}.$$

BESSELova funkcija druge vrste, takođe, je jedno partikularno rešenje diferencne jednačine (3.2.4), tj. važi

$$Y_{k+1}(x) = \frac{2k}{x}Y_k(x) - Y_{k-1}(x), \quad k = 1,2,\ldots \ .$$

S obzirom da su $J_k(x)$ i $Y_k(x)$ linearno nezavisna rešenja, to se opšte rešenje diferencne jednačine (3.2.4) može dati u obliku $y_k = C_1J_k(x) + C_2Y_k(x)$, gde su C_1 i C_2 proizvoljne konstante, a $x = \text{const}$.

Napomena 3.2.3. Funkcija $y(x) = C_1J_k(x) + C_2Y_k(x)$ je opšte rešenje BESSELove diferencijalne jednačine $x^2y'' + xy' + \left(x^2 - k^2\right)y = 0$. Ovde je $k = \text{const}$.

Vratimo se sada problemu određivanja partikularnog rešenja diferencne jednačine (3.2.4) za neke zadate početne uslove. Neka je $x = 1$.

a) S obzirom na to da su

$$J_0(1) \cong 0.7651976866, \qquad J_1(1) \cong 0.4400505857,$$

primenom rekurentne relacije (3.2.5) za $k = 1,2,\ldots,10$ dobijamo niz $\{J_k(1)\}$, koji je dat u tabeli 3.2.1. Broj u zagradi ukazuje na decimalni eksponent. Primetimo da

[44] LÉONHARD EULER (1707 – 1783), veliki švajcarski matematičar i fizičar, poznat po značajnim doprinosima u matematičkoj analizi, teoriji grafova, astronomiji, mehanici, itd.

Tabela 3.2.1.

k	$J_k(1)$	$Y_k(1)$
0	$7.651976866(-1)$	$8.825696420(-2)$
1	$4.400505857(-1)$	$-7.812128213(-1)$
2	$1.149034848(-1)$	$-1.650682607\ (0)$
3	$1.95633535\ (-2)$	$-5.821517606\ (0)$
4	$2.4766362\ (-3)$	$-3.327842303\ (1)$
5	$2.497361\ (-4)$	$-2.604058666\ (2)$
6	$2.07248\ (-5)$	$-2.570780243\ (3)$
7	$-1.0385\ (-6)$	$-3.058895705\ (4)$
8	$-3.52638\ (-5)$	$-4.256746185\ (5)$
9	$-5.631823\ (-4)$	$-6.780204939\ (6)$
10	$-1.01020176\ (-2)$	$-1.216180143\ (8)$

zbog oduzimanja bliskih brojeva nastupa poznati efekat gubitka značajnih cifara, tako da dobijena vrednost za $J_7(1)$ ima samo pet značajnih cifara. Štaviše, ova vrednost je pogrešna do te mere da se razlikuje i u znaku od tačne vrednosti. Tačna vrednost na deset značajnih cifara je

$$J_7(1) = 1.502325817(-6).$$

b) Uzimamo sada za početne vrednosti

$$Y_0(1) \cong 0.08825696420, \qquad Y_1(1) \cong -0.7812128213.$$

Primenom iste rekurentne relacije dobijamo niz $\{Y_k(1)\}$ koji je dat, takođe, u tabeli 3.2.1. Upoređivanjem dobijenih vrednosti sa vrednostima datim u [1, str. 408] može se videti da su dobijeni rezultati korektni u svim ciframa.

Postavlja se sada pitanje koji su to razlozi da se u prvom slučaju generišu pogrešni rezultati, a drugom tačni. Takođe, od interesa je naći postupak za tačno generisanje niza $\{J_k(x)\}$.

U cilju odgovora na prvo pitanje primetimo da se opšte rešenje diferencne jednačine drugog reda (3.2.4), ili uopšte (3.2.1), može predstaviti kao linearna kombinacija bilo koja dva linearno nezavisna rešenja f_k i g_k, tj. $y_k = af_k + bg_k$, gde su a i b proizvoljne konstante. Nas će ovde interesovati specijalan slučaj kada postoji takav par linearno nezavisnih rešenja za koji je

$$\lim_{k\to+\infty} \frac{f_k}{g_k} = 0. \tag{3.2.7}$$

Problemi o kojima je bilo reči kod određivanja niza $\{J_k(1)\}$ nastupaju upravo kada se izračunava rešenje f_k ili cf_k, gde je c konstanta. Da bismo ovo pokazali primetimo da iz (3.2.7) sleduje

$$\lim_{k\to+\infty}\frac{f_k}{y_k}=0 \tag{3.2.8}$$

za svako y_k koje nije proporcionalno sa f_k. Zaista, za $y_k=af_k+bg_k$ $(b\neq 0)$ imamo

$$\lim_{k\to+\infty}\frac{f_k}{y_k}=\lim_{k\to+\infty}\frac{f_k/g_k}{a(f_k/g_k)+b}=0.$$

Pretpostavimo sada da generišemo niz $\{f_k\}$, korišćenjem približnih početnih vrednosti $y_0\cong f_0$ i $y_1\cong f_1$ (na primer, dobijenih zaokrugljivanjem). I pod uslovom da se sve operacije u rekurentnoj relaciji odvijaju tačno (sa beskonačnom preciznošću, tj. sa beskonačnim brojem cifara u mantisi), rešenje koje dobijamo biće, u opštem slučaju, linearno nezavisno od f_k. Imajući u vidu (3.2.8) zaključujemo da će relativna greška dobijenog rešenja težiti u beskonačnost, tj. imaćemo

$$\left|\frac{y_k-f_k}{f_k}\right|=\left|\frac{1-f_k/y_k}{f_k/y_k}\right|\to+\infty\qquad(k\to+\infty).$$

Prema tome, ovakav postupak za određivanje niza $\{f_k\}$ daje pogrešne rezultate. Rešenje f_k sa osobinom (3.2.8) naziva se *minimalno rešenje* diferencne jednačine (3.2.1). Suprotno, neminimalno rešenje zovemo *dominantno rešenje*. Primetimo da je svako dominantno rešenje asimptotski proporcionalno sa g_k.

U predhodno posmatranom primeru imali smo $f_k=J_k(x)$ i $g_k=Y_k(x)$. Na osnovu asimptotskih formula (videti [1, str. 365])

$$J_\nu(x)\sim\frac{1}{\sqrt{2\pi\nu}}\left(\frac{ex}{2\nu}\right)^\nu,\quad Y_\nu(x)\sim-\sqrt{\frac{2}{\pi\mu}}\left(\frac{ex}{2\nu}\right)^{-\nu},$$

pri $\nu\to+\infty$, vidimo da $J_\nu(x)\to 0$ i $Y_\nu(x)\to-\infty$.

Dakle, rezimirajmo da se minimalno rešenje navedenim postupkom ne može izračunati, za razliku od dominantnog rešenja koje se veoma tačno određuje.

Razne metode za određivanje minimalnog rešenja tročlane rekurentne relacije, uključujući i odgovarajuće primene, razmatrao je W. GAUTSCHI[45] u radu [22].

[45] WALTER GAUTSCHI (1927 –), američki matematičar, rođen u Švajcarskoj. Jedan je od vodećih naučnika u oblasti numeričke analize (videti: *The History of Numerical Analysis and Scientific Computing*: `http://history.siam.org/oralhistories/gautschi.htm`). Punih dvanaest godina (1984 – 1995) bio je glavni urednik naučnog časopisa *Mathematics of Computation*, koji izdaje AMS (Američko matematičko društvo). Sada je profesor emeritus na Purdue univerzitetu (SAD).

Na kraju ovog odeljka navešćemo jedan postupak za stabilno određivanje niza $\{J_k(x)\}$, poznat kao MILLERov[46] postupak. Na osnovu prethodno pomenute asimptotske formule za $J_k(x)$ $(x = \text{const})$ videli smo da $J_k(x)$ teži nuli kada k teži beskonačnosti. Zato izaberimo dovoljno veliko n i stavimo

$$\overline{J}_n(x) = 0 \qquad \text{i} \qquad \overline{J}_{n-1}(x) = 1. \tag{3.2.9}$$

Nešto bolji izbor je uzeti, umesto 0 i 1, vrednosti koje se dobijaju iz pomenute asimptotske relacije za $\nu = n$ i $\nu = n-1$. Jednostavnosti radi, opredelićemo se za vrednosti date pomoću (3.2.9), a zatim ćemo primenom rekurentne relacije unazad (3.2.2), koja u našem slučaju ima oblik

$$\overline{J}_{k-1}(x) = \frac{2k}{x}\overline{J}_k(x) - \overline{J}_{k+1}(x), \tag{3.2.10}$$

odrediti redom $\overline{J}_{n-2}(x), \ldots, \overline{J}_1(x), \overline{J}_0(x)$. Naravno, dobijene vrednosti ne odgovaraju tačnim vrednostima BESSELovih funkcija, ali se, zbog linearnosti relacije (3.2.10) od njih razlikuju za istu multiplikativnu konstantu. S obzirom na indenitet

$$J_0(x) + 2\left(J_2(x) + J_4(x) + \cdots\right) = 1$$

moguće je odrediti ovu multiplikativnu konstantu. Naime, ako odredimo sumu

$$S = \overline{J}_0(x) + 2\left(\overline{J}_2(x) + \overline{J}_4(x) + \cdots\right)$$

iz proporcije $1 : S = J_k(x) : \overline{J}_k(x)$ nalazimo

$$J_k(x) = \frac{1}{S}\overline{J}_k(x), \quad k = 0, 1, \ldots\ . \tag{3.2.11}$$

Obično za n uzimamo dovoljno veliki neparan broj i tada prilikom izračunavanja sume S polazimo od najmanjeg sabirka $\overline{J}_{n-1}(x)$ $(= 1)$. Za dobijanje rezultata sa većom tačnošću treba poći od većeg broja n.

Ako uzmemo $n = 11$, tj. $\overline{J}_{11}(1) = 0$ i $\overline{J}_{10}(1) = 1$, dobijamo rezultate koji su dati u tabeli 3.2.2. Ovde je $S = 3810092281$, a $1/S = 2.624608346(-10)$. Pomoću (3.2.11) dobijamo kolonu tabele sa vrednostima za $J_k(1)$. Upoređivanjem sa tačnim vrednostima može se videti da su pogrešne vrednosti samo za $J_{10}(1)$, $J_9(1)$ i $J_8(1)$ (podvučene cifre su pogrešne).

Ako uzmemo $n = 17$, dobićemo sledeće vrednosti (sa deset značajnih cifara):

[46] JEFFREY CHARLES PERCY MILLER (1906 – 1981), engleski matematičar.

Tabela 3.2.2.

k	$\overline{J}_k(1)$	$J_k(1)$
10	1	$2.6\underline{24608346}(-10)$
9	20	$5.249\underline{16692}\ (-9)$
8	359	$9.42234\underline{3962}\ (-8)$
7	5724	$1.502325817\ (-6)$
6	79777	$2.093833800\ (-5)$
5	951600	$2.497577302\ (-4)$
4	9436223	$2.476638964\ (-3)$
3	74538184	$1.956335398\ (-2)$
2	437792881	$1.149034849\ (-1)$
1	1676633340	$4.400505857\ (-1)$
0	2915473799	$7.651976866\ (-1)$

$$J_8(1) = 9.422344172\ (-8),$$
$$J_9(1) = 5.249250180\ (-9),$$
$$J_{10}(1) = 2.630615124(-10).$$

Upoređivanjem vidimo da su sve cifre tačne.

MILLERov postupak se vrlo često korisri u numeričkoj matematici u više oblika koji su slični navedenom.

1.3.3 Izračunavanje vrednosti elementarnih funkcija

Pri izračunavanju vrednosti funkcija na računskim mašinama, vrlo je važno u kom su obliku date odgovarajuće formule, s obzirom na aritmetiku konačne dužine. Naime, često vrednost ekvivalentnih matematičkih izraza u numeričkom smislu nije ista (zbog grešaka o kojima je bilo reči u prethodnom poglavlju).

U ovom odeljku ukazaćemo na neke standardne načine za izračunavanje vrednosti elementarnih funkcija, pri čemu ćemo posebnu pažnju posvetiti izračunavanju funkcije $x \mapsto \sqrt{x}$, eksponencijalne funkcije $x \mapsto \mathrm{e}^x$, trigonometrijske funkcije $x \mapsto \sin x$, logaritamske funkcije $x \mapsto \log x$ i inverzne trigonometrijske funkcije $x \mapsto \arctan x$. Sem funkcija $x \mapsto \sqrt{x}$ i $x \mapsto \log x$ koje su definisane za $x \geq 0$ i $x > 0$ respektivno, ostale funkcije su definisane za svako realno x. Kao što ćemo videti, u svim slučajevima neophodno je transformisati x na neki novi argument koji pripada tzv. osnovnom intervalu na kome je poznata aproksimativna formula ili neki postupak za jednostavno izračunavanje date funkcije sa unapred određenom tačnošću. Najčešće aproksimativne formule (aproksimacije) su polinomski razvoji

ili racionalne funkcije. Sledeći odeljak je zato posvećen izračunavanju vrednosti polinoma. Primenu verižnih razlomaka i asimptotskih razvoja razmatraćemo na kraju ove glave, dok će opšti metodi za aproksimaciju funkcija biti tretirani u jednoj od narednih knjiga iz ove serije.

1. Funkcija $x \mapsto \sqrt{x}$. Na samom početku ove knjige, u odeljku 1.1.1, naveli smo način za nalaženje kvadratnog korena iz pozitivnog broja. Kod implementacije ovog iterativnog postupka, za datu konstantu $b > 1$, najpre se vrši transformacija argumenta pomoću $x = b^m z$, gde je m ceo broj takav da z pripada osnovnom intervalu $[1/b, 1)$. Tada,

$$\sqrt{x} = \sqrt{b^m}\sqrt{z},$$

pri čemu, u slučaju $m > 0$, faktor $\sqrt{b^m}$ računamo pomoću

$$\sqrt{b^m} = \underbrace{b \cdots b}_{k \text{ puta}} \cdot \begin{cases} 1, & m = 2k, \\ \sqrt{b}, & m = 2k+1. \end{cases}$$

Naravno, za $m < 0$ imamo $\sqrt{b^m} = 1/\sqrt{b^{-m}}$. Primetimo da su, u ovom postupku, neophodne konstante b i $\sqrt{b}$.

Sada za $z \in [1/b, 1)$ konstruišemo niz $\{z_k\}$ pomoću

$$z_{k+1} = \frac{1}{2}\left(z_k + \frac{z}{z_k}\right) = z_k + \frac{1}{2}\left(\frac{z}{z_k} - z_k\right), \qquad k = 0, 1, \ldots, \tag{3.3.1}$$

startujući sa $z_0 = z$. Ako uvedemo relativnu grešku aproksimacije z_k pomoću $r_k := (z_k - \sqrt{z})/\sqrt{z}$, tj. $z_k = \sqrt{z}(1 + r_k)$, jednostavno nalazimo da je

$$r_{k+1} = \frac{r_k^2}{2(1 + r_k)}, \quad k = 0, 1, \ldots .$$

S obzirom na to da je $r_0 = (z_0 - \sqrt{z})/\sqrt{z} = \sqrt{z} - 1$ i $|r_0| < 1$, zaključujemo da greška r_k teži nuli kvadratno, tj. $r_{k+1} = O(r_k^2)$, kada $k \to +\infty$. Dakle, z_k teži kvadratnom korenu $\sqrt{z}$, tako da, u zavisnosti od željene tačnosti, možemo za neko $n > 0$ uzeti $z_n \cong \sqrt{z}$.

Primer 3.3.1. Odredićemo $\sqrt{10}$. Ako uzmemo $b = 2$, osnovni interval je $[1/2, 1)$, pa je $10 = 2^4 z$, gde je $z = 10/16 = 0.625$.

Primenom postupka (3.3.1) dobijamo redom vrednosti koje su navedene u tabeli 3.3.1.

Tabela 3.3.1.

k	z_k
0	0.625
1	0.8125
2	0.790865385
3	0.790569470
4	0.790569415
5	0.790569415

Prema tome,

$$\sqrt{10} \cong 4 \cdot z_4 = 4 \cdot 0.790569415 = 3.16227766. \quad \triangle$$

Umesto primene iterativnog postupka (3.3.1), za određivanje vrednosti $\sqrt{z}$ moguće je koristiti neku aproksimativnu polinomsku formulu ili formulu u obliku racionalne funkcije.

Primer 3.3.2. Neka je $b = \sqrt{2} \cong 1.41421$ ($\sqrt{b} \cong 1.18921$). Na osnovnom intervalu $[1/\sqrt{2}, 1)$ može se uzeti jednostavna formula

$$\sqrt{z} \cong 0.3433929 + 0.8174949z - 0.1609689z^2,$$

sa relativnom greškom (po apsolutnoj vrednosti) manjom od 10^{-4}.

Za $x = 10$, na osnovu $x = b^m z$, nalazimo $m = 7$ i $z = 10/b^7 \cong 0.883883$. Na osnovu prethodne formule imamo $\sqrt{z} \cong 0.940206$, pa je najzad

$$\sqrt{10} = b^3\sqrt{b}\sqrt{z} \cong 1.41421^3 \cdot 1.18921 \cdot 0.940206 \cong 3.16245,$$

sa relativnom greškom $5.45 \cdot 10^{-5}$.

Bolja aproksimacija se može dati pomoću racionalne funkcije.

Neka je $b = \sqrt[3]{4} \cong 1.587401$ ($\sqrt{b} \cong 1.259921$). Osnovni interval je, u ovom slučaju, $[1/\sqrt[3]{4}, 1) = [0.6299605, 1)$ za koji važi formula

$$\sqrt{z} \cong \frac{P(z)}{Q(z)} = \frac{0.5612310 + 7.1065058z + 4.4693601z^2}{3.1603148 + 7.9767830z + z^2}, \tag{3.3.2}$$

sa relativnom greškom (po apsolutnoj vrednosti) manjom od 10^{-7}.

Za $x = 10$, na osnovu $x = b^m z$, nalazimo $m = 5$ i $z = 10/b^5 \cong 0.9921257$. Tada imamo $\sqrt{10} = b^2\sqrt{b}\sqrt{z}$. Primenom formule (3.3.2) dobijamo

$$\sqrt{10} \cong 1.587401^2 \cdot 1.259921 \cdot 0.9960551 \cong 3.1622775,$$

sa relativnom greškom $-5.06 \cdot 10^{-8}$. $\triangle$

2. Eksponencijalna funkcija. Za izračunavanje vrednosti funkcije $x \mapsto \mathrm{e}^x$ može se koristiti TAYLORov razvoj

$$\mathrm{e}^x = \sum_{k=0}^{n} \frac{x^k}{k!} + R_n(x), \tag{3.3.3}$$

gde je

$$R_n(x) = \frac{x^{n+1}}{(n+1)!}\mathrm{e}^{\xi x} \qquad (0 \le \xi \le 1).$$

Ako je $0 \le x \le 1$ moguće je ostatak $R_n(x)$ oceniti na sledeći način. Naime, kako je

$$\begin{aligned} R_n(x) = \sum_{k=n+1}^{+\infty} \frac{x^k}{k!} &= \frac{x^{n+1}}{(n+1)!}\left(1 + \frac{x}{n+2} + \frac{x^2}{(n+2)(n+3)} + \cdots\right) \\ &< \frac{x^{n+1}}{(n+1)!}\left(1 + \frac{x}{n+2} + \left(\frac{x}{n+2}\right)^2 + \cdots\right), \end{aligned}$$

sumiranjem dobijene geometrijske progresije dobijamo

$$R_n(x) < \frac{x^{n+1}}{(n+1)!} \cdot \frac{n+2}{n+2-x}.$$

Kako je $0 \le x \le 1$, iz prethodne nejednakosti sleduje

$$R_n(x) < \frac{x^{n+1}}{(n+1)!} \cdot \frac{n+2}{n+1} \le \frac{n+2}{(n+1)(n+1)!}. \tag{3.3.4}$$

Broj članova n u razvoju (3.3.3), koje treba uzeti, zavisi od tačnosti koja se zahteva za rezultat. Ukoliko je ovaj broj članova veliki, to će konačan rezultat ipak, biti pogrešan usled grešaka zaokrugljivanja tokom računskog procesa.

U konkretnom slučaju, kada je $0 \le x \le 1$, uslov $R_n(x) < 10^{-7}$, na osnovu (3.3.4), ispunjen je za najmanje $n = 10$. Kada se svi brojevi koji učestvuju u izračunavanjima zaokrugljuju i predstavljaju u normalizovanom obliku sa sedmorazrednom mantisom, ovaj broj članova razvoja nije veliki, s obzirom na to da je relativna greška zaokrugljivanja, takođe, reda 10^{-7}.

Primer 3.3.3. Izračunajmo broj e sa relativnom greškom koja je manja od 10^{-7}. Na osnovu prethodnog razmatranja, u razvoju (3.3.3) treba uzeti $n = 10$ i $x = 1$. Dakle,

$$\mathrm{e} = \sum_{k=0}^{10} u_k + R_{10}(1),$$

gde je

$$u_0 = 1, \qquad u_k = \frac{1}{k} u_{k-1} \qquad (k = 1, \ldots, 10).$$

Zaokrugljeni brojevi u_k $(k = 0, 1, \ldots, 10)$, predstavljeni u normalizovanom obliku sa sedmorazrednom mantisom, su redom

$$\begin{aligned} u_0 = u_1 &= 0.1000000 \cdot 10^{1}, & u_2 &= 0.5000000 \cdot 10^{0}, \\ u_3 &= 0.1666667 \cdot 10^{0}, & u_4 &= 0.4166667 \cdot 10^{-1}, \\ u_5 &= 0.8333333 \cdot 10^{-2}, & u_6 &= 0.1388889 \cdot 10^{-2}, \\ u_7 &= 0.1984127 \cdot 10^{-3}, & u_7 &= 0.2480158 \cdot 10^{-4}, \\ u_9 &= 0.2755731 \cdot 10^{-5}, & u_{10} &= 0.2755731 \cdot 10^{-6}. \end{aligned}$$

Iz odeljka 1.2.3 poznato je da će greška biti najmanja ako sabiramo brojeve polazeći od najmanjeg, tj. od u_{10} u ovom slučaju. Tada za zbir dobijamo

$$S_0 = 0.2718282 \cdot 10^{1}.$$

Broj S_0 aproksimira broj $\mathrm{e}\,(= 0.27182818248\ldots \cdot 10^{1})$ sa sedam značajnih cifara. Relativna greška je manja od 10^{-7} (tačnije, manja je od $0.67 \cdot 10^{-7}$). $\triangle$

U opštem slučaju, za izračunavanje vrednosti funkcije $x \mapsto \mathrm{e}^x$ poželjno je koristiti se formulom

$$\mathrm{e}^x = \mathrm{e}^{[x]} \mathrm{e}^z,$$

gde je $[x]$ najveće celo[47] od x i $z = x - [x]$. Pri ovome vrednost $\mathrm{e}^{[x]}$ izračunava se prostim množenjem, tj.

$$\mathrm{e}^{[x]} = \begin{cases} 1 & ([x] = 0), \\ \underbrace{\mathrm{e} \cdots \mathrm{e}}_{[x] \text{ puta}} & ([x] > 0), \\ \underbrace{(1/\mathrm{e}) \cdots (1/\mathrm{e})}_{-[x] \text{ puta}} & ([x] < 0), \end{cases}$$

[47] Najveći ceo broj koji nije veći od x; na primer, $[3.14] = 3$, $[-3.14] = -4$.

gde su

$$\mathrm{e} = 2.71828182845904\ldots \quad \text{i} \quad \frac{1}{\mathrm{e}} = 0.36787944117144\ldots\ .$$

Izračunavanje vrednosti e^z $(0 \le z < 1)$, kao što smo ranije videli, ne predstavlja posebnu teškoću.

3. Trigonometrijske funkcije. Za izračunavanje vrednosti trigonometrijskih funkcija dovoljno je poznavati postupak za izračunavanje, na primer, funkcije $x \mapsto \sin x$. Vrednosti ostalih trigonometrijskih funkcija izračunavaju se jednostavno pomoću sinusne funkcije korišćenjem poznatih trigonometrijskih identiteta. Na primer, $\cos x = \sin(x+\pi/2)$ i $\tan x = \sin x/\sin(x+\pi/2)$.

Pokazaćemo da je za izračunavanje funkcije $x \mapsto \sin x$, za proizvoljnu vrednost argumenta x, dovoljno znati kako se može izračunati vrednost $\sin(\pi z/2)$ za $z \in [-1,1]$. Naime, ako za dato x izračunamo redom veličine

$$u = \frac{2}{\pi}x, \quad v = u - 4\left[\frac{1}{4}(u+1)\right], \quad z = \begin{cases} v & (v \le 1), \\ 2-v & (v > 1), \end{cases}$$

gde je $[p]$ oznaka za najveće celo od p, tada je $\sin x = \sin(\pi z/2)$. Da bismo ovo dokazali primetimo da je

$$\sin x = \sin\frac{\pi u}{2} = \sin\frac{\pi}{2}\left(v + 4\left[\frac{1}{4}(u+1)\right]\right),$$

tj.

$$\sin x = \sin\left(\frac{\pi v}{2} + 2\pi\left[\frac{1}{4}(u+1)\right]\right) = \sin\frac{\pi v}{2}.$$

Kako je

$$v = 4\left(\frac{1}{4}(u+1) - \left[\frac{1}{4}(u+1)\right]\right) - 1$$

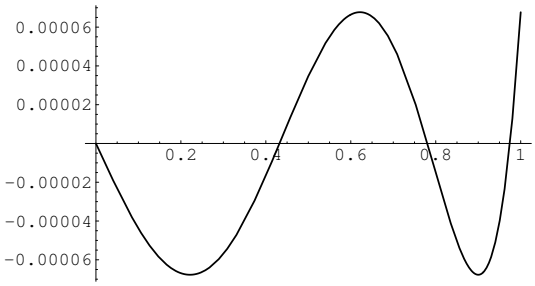
i

$$0 \le \frac{1}{4}(u+1) - \left[\frac{1}{4}(u+1)\right] < 1$$

zaključujemo da je $-1 \le v < 3$. Razlikujemo sada dva slučaja:

Slučaj $v \le 1$. Tada je $z = v$ $(\in [-1,1])$ i $\sin x = \sin(\pi v/2) = \sin(\pi z/2)$.

Slučaj $v > 1$. Kako je $z = 2 - v$ i $1 < v < 3$ zaključujemo da $z \in (-1,1)$ i da je pri tome



Slika 3.3.1. Greška aproksimacije (3.3.5) za $0 \le z \le 1$

$$\sin x = \sin\frac{\pi v}{2} = \sin\frac{\pi}{2}(2-z) = \sin\frac{\pi z}{2}\,.$$

Prema tome, dovoljno je znati neku približnu formulu (aproksimaciju) za $\sin(\pi z/2)$, kada $z \in [-1,1]$. Jedna takva formula je, na primer,

$$\sin\frac{\pi z}{2} \simeq 1.57032002z - 0.64211317z^3 + 0.07186085z^5, \tag{3.3.5}$$

za koju apsolutna greška ne prelazi vrednost $6.8 \cdot 10^{-5}$, kada $z \in [-1,1]$ (videti sliku 3.3.1). U jednoj od narednih knjiga iz ove serije razmatraćemo razne postupke za aproksimaciju funkcija.

4. Logaritamska funkcija. Za izračunavanje vrednosti logaritamske funkcije[48] $x \mapsto \log x$ $(0 < x < +\infty)$ koristimo transformaciju $x = 2^m z$, gde su z i m takvi da je $1/2 \le z < 1$ i $m \in \mathbb{Z}$. Tada je

$$\log x = m\log 2 + \log z \qquad \big(\log 2 = 0.69314718\ldots\big).$$

Ako uvedemo bilinearnu transformaciju $z = (1-y)/(1+y)$, imamo

$$\log z = \log\frac{1-y}{1+y} = -2\left(y + \frac{y^3}{3} + \frac{y^5}{5} + \cdots + \frac{y^{2n-1}}{2n-1}\right) - R_n(y),$$

gde su

$$y = \frac{1-z}{1+z}$$

[48] Logaritamsku funkciju od x za osnovu $b\,(>0)$ označavamo sa $\log_b x$; na primer $\log_{10} x$ za osnovu 10. Ako je osnova e u oznaci izostavljamo osnovu, tj. $\log_{\mathrm{e}} x = \log x$. Često se za ovaj logaritam koristi stara oznaka $\ln x$.

i

$$R_n(y) = 2\left(\frac{y^{2n+1}}{2n+1} + \frac{y^{2n+3}}{2n+3} + \cdots\right) < \frac{2}{1-y^2}\cdot\frac{y^{2n+1}}{2n+1}.$$

Kako je $0 < y \le 1/3$ ($\Leftarrow 1/2 \le z < 1$), za ostatak važe nejednakosti

$$0 < R_n(y) < \frac{1}{4\cdot 3^{2n-1}(2n+1)}.$$

Uvođenjem oznake $u_k = \dfrac{y^{2k-1}}{2k-1}$ $(k = 1,\ldots,n)$, dobijamo

$$\log x = m\log 2 - 2\,(u_1 + \cdots + u_n) - R_n,$$

gde se n određuje iz uslova $u_n < 4\varepsilon$ (ε zadata tačnost), s obzirom na činjenicu da je tada

$$R_n \equiv R_n(y) \le \frac{1}{4}u_n < \varepsilon.$$

5. Inverzne trigonometrijske funkcije. Kod izračunavanja inverznih trigonometrijskih funkcija najpogodnije je poznavati algoritam za izračunavanje funkcije $\arctan x$ kada $x \in \mathbb{R}$, s obzirom na to da za ostale funkcije, pri $|x| \le 1$, važe sledeće formule

$$\arcsin x = \arctan\frac{x}{\sqrt{1-x^2}} = \operatorname{sgn}(x)\left[\frac{\pi}{2} - \arctan\frac{\sqrt{1-x^2}}{|x|}\right],$$

$$\arccos x = \frac{\pi}{2} - \arcsin x, \quad \operatorname{arccot} x = \arctan\frac{1}{x} = \frac{\pi}{2} - \arctan x,$$

pri čemu važe nejednakosti $0 \le \arccos x \le \pi$ i $0 \le \operatorname{arccot} x \le \pi$.

Kako je $\arctan(-x) = -\arctan x$, dovoljno je poznavati postupak za izračunavanje ove funkcije na $(0,+\infty)$.

Za dato $n \in \mathbb{N}$ na intervalu $(0,+\infty)$ definišimo deone tačke X_ν i čvorove x_ν pomoću

$$X_0 = 0, \quad X_\nu = \tan\frac{(2\nu-1)\pi}{4n}, \quad \nu = 1,\ldots,n, \quad X_{n+1} = +\infty;$$
$$x_\nu = \tan\frac{(2\nu-2)\pi}{4n} = \tan\frac{(\nu-1)\pi}{2n}, \quad \nu = 2,\ldots,n+1,$$

za koje važe nejednakosti

$$0 = X_0 < X_1 < x_2 < X_2 < x_3 < X_3 < \cdots < x_n < X_n < x_{n+1} = X_{n+1} = +\infty.$$

Svaki od segmenata $[X_{\nu-1}, X_\nu]$, $\nu = 2, \ldots, n+1$, pomoću

$$t = \frac{1}{x_\nu} - \frac{1/x_\nu^2 + 1}{1/x_\nu + x} = \frac{x - x_\nu}{1 + x x_\nu}$$

transformiše se na osnovni segment $[-X_1, X_1]$, tako da je

$$\arctan x = \arctan t + \arctan x_\nu = \arctan t + \frac{(\nu-1)\pi}{2n}.$$

Dakle, za izračunavanje vrednosti $\arctan x$ potrebno je imati valjanu aproksimaciju samo na segmentu $[0, X_1]$. Takve aproksimacije su date u [31, str. 120–130] za $n = 1, 2, 3, 4$ i 8.

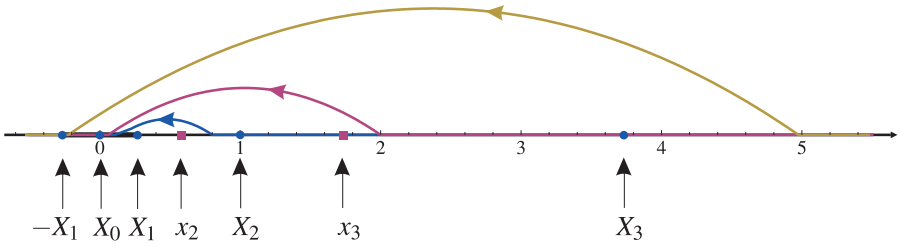
Primer 3.3.4. Ako izaberemo $n = 3$, na osnovu prethodnog, deone tačke i čvorovi su dati sa

$$X_0 = 0, \quad X_1 = \tan\frac{\pi}{12} = 2 - \sqrt{3} \cong 0.2679491924311227, \quad X_2 = \tan\frac{\pi}{4} = 1,$$
$$X_3 = \tan\frac{5\pi}{12} = 2 + \sqrt{3} \cong 3.732050807568877, \quad X_4 = +\infty$$

i

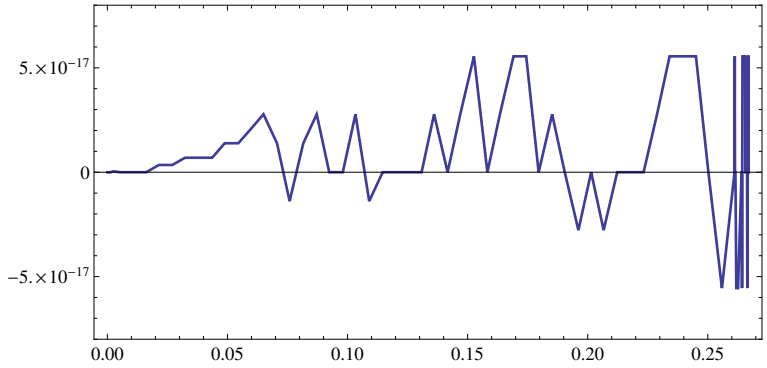
$$x_2 = \tan\frac{\pi}{6} = \frac{\sqrt{3}}{3}, \quad x_3 = \tan\frac{\pi}{3} = \sqrt{3}, \quad x_4 = +\infty,$$

respektivno (videti sliku 3.3.2). Na slici je prikazana i transformacija tačaka sa



Slika 3.3.2. Transformacija argumenta na osnovni segment $[-X_1, X_1]$

segmenta $[X_{\nu-1}, X_\nu]$, $\nu = 2, 3, 4$, na osnovni segment $[-X_1, X_1]$. Konkretno su uzete tačke $0.8 \in [X_1, X_2]$, $2.0 \in [X_2, X_3]$ i $5.0 \in [X_3, +\infty)$, koje se preslikavaju redom na tačke 0.1523036760962039, 0.06002309434948968 i -0.2.



Slika 3.3.3. Greška u aproksimaciji (3.3.6) na $[0, X_1]$

Na kraju navodimo aproksimaciju

$$\arctan t \cong t\,\frac{P(t^2)}{Q(t^2)}, \quad |t| \in \left[0, \tan\frac{\pi}{12}\right], \tag{3.3.6}$$

gde su P i Q polinomi trećeg stepena dati sa

$$\begin{aligned} P(t) &= 12.822976800612627093 35 + 16.242807815134221366 2t \\ &\quad +4.923443373635526899t^2 + 0.205608723964456163t^3, \\ Q(t) &= 12.822976800612628214 74 + 20.517133415336848884t \\ &\quad +9.197892485655692716t^2 + t^3, \end{aligned}$$

Odgovarajuća apsolutna greška u ovoj aproksimaciji ne prelazi vrednost 10^{-16} (slika 3.3.3). $\triangle$

1.3.4 Izračunavanje vrednosti polinoma

Jedan elementaran, ali važan problem je izračunavanje vrednosti polinoma

$$P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n. \tag{3.4.1}$$

Ako bismo izračunavali vrednost polinoma $P(x)$, na osnovu (3.4.1), bilo bi potrebno $2n-1$ množenja i n sabiranja. Međutim, ukoliko $P(x)$ predstavimo u obliku

$$P(x) = (\cdots(((a_0x + a_1)x + a_2)x + a_3)x + \cdots + a_{n-1})x + a_n \tag{3.4.2}$$

potrebno je samo n množenja i n sabiranja.

Na osnovu (3.4.2) može se formirati rekurzivni postupak za izračunavanje vrednosti polinoma za $x = x_0$

$$b_0 = a_0, \quad b_k = a_k + x_0 b_{k-1}, \quad k = 1, \ldots, n, \tag{3.4.3}$$

koji posle n koraka daje vrednost polinoma, tj. $P(x_0) = b_n$. Izloženi postupak je poznat kao HORNERova[49] šema i može se interpretirati kroz sledeću šemu:

	a_0	a_1	a_2	a_3	$\ldots$	a_{n-1}	a_n
x_0		$x_0 b_0$	$x_0 b_1$	$x_0 b_2$		$x_0 b_{n-2}$	$x_0 b_{n-1}$
	b_0	b_1	b_2	b_3		b_{n-1}	$b_n = P(x_0)$

Naime, u prvoj vrsti šeme pišemo koeficijente polinoma (3.4.1), počev od najstarijeg koeficijenta, a drugu vrstu započinjemo sa vrednošću argumenta x_0 za koji izračunavamo vrednost polinoma. U trećoj vrsti pišemo koeficijente b_k, koje izračunavamo sabiranjem odgovarajućih elemenata prve i druge vrste, pri čemu je $b_0 = a_0$. Elemente druge vrste formiramo množenjem vrednosti x_0 sa prethodnim elementom iz treće vrste.

Primetimo da su b_k $(k = 0, 1, \ldots, n-1)$ koeficijenti polinoma $P_1(x)$ koji se dobija deljenjem $P(x)$ sa linearnim faktorom $x - x_0$. Zaista, upoređivanjem koeficijenata uz odgovarajuće stepene u sledećoj jednakosti

$$a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n = (x - x_0)\left(b_0 x^{n-1} + b_1 x^{n-2} + \cdots + b_{n-1}\right) + b_n$$

dobijamo (3.4.3).

Kao što smo prethodno pomenuli, HORNERova šema zahteva n množenja i n sabiranja za izračunavanje vrednosti polinoma (3.4.1). OSTROWSKI[50] je dokazao da je $2n$ minimalan broj operacija za izračunavanje vrednosti polinoma ako je $n \leq 4$. Ako je $4 \leq n \leq 6$ može se dokazati ([55], [69], [36]) da se vrednost polinoma može izračunati sa $\left[\frac{1}{2}(n+3)\right]$ množenja i ne više od $n+1$ sabiranja. Oznaka $[t]$ predstavlja najveće celo od t.

Tako u slučaju $n = 4$ imamo sledeći ekvivalentni oblik polinoma (3.4.1)

[49] WILLIAM GEORGE HORNER (1786 – 1837), britanski matematičar.

[50] ALEXANDER MARKOWICH OSTROWSKI (1893 – 1986), poznati matematičar rođen u Kijevu (današnja Ukrajina). Posle I svetskog rata živeo je i radio u Nemačkoj i Engleskoj do 1926. godine kada dobija poziciju šefa departmana na Univerzitetu u Bazelu (Švajcarska), gde ostaje sve do penzionisanja 1958. godine.

$$P(x) = \left((Ax+B)^2 + Ax + C\right)\left((Ax+B)^2 + D\right) + E, \tag{3.4.4}$$

koji zahteva 3 množenja i 5 sabiranja. Vrednosti parametra u (3.4.4) mogu se odrediti pomoću

$$A = \sqrt[4]{a_0}, \quad B = \frac{a_1 - A^3}{4A^3}, \quad D = 3B^2 + 8B^3 + \frac{a_3 A - 2a_2 B}{A^2},$$

$$C = \frac{a_2}{A^2} - 2B - 2B^2 - D, \quad E = a_4 - (C+D)B^2 - B^4 - CD,$$

gde smo, ne umanjujući opštost, pretpostavili da je $a_0 > 0$.

U slučaju $n = 5$ ekonomična formula je oblika

$$P(x) = \left(\left((Ax+B)^2 + C\right)\ (Ax+B)^2 + D\right)\ (Ax+E) + F. \tag{3.4.5}$$

Upoređivanjem koeficijenata uz odgovarajuće stepene u (3.4.5) i (3.4.1) (za $n = 5$) dobijamo sledeće jednakosti

$$\begin{aligned}
A^5 &= a_0,\\
A^4(4B+E) &= a_1,\\
A^3\left(6B^2 + 4BE + C\right) &= a_2,\\
A^2\left(4B^3 + 6B^2E + 2BC + CE\right) &= a_3,\\
A\left(B^4 + 4B^3E + B^2C + 2BCE + D\right) &= a_4,\\
B^4E + B^2CE + DE + F &= a_5.
\end{aligned}$$

Ako stavimo $c_k = (a_k/a_0)A^k$ $(k = 1,2,3,4,5)$, iz prethodnih jednakosti dobijamo

$$A = \sqrt[5]{a_0}, \quad C = c_2 + 10B^2 - 4c_1B, \quad E = c_1 - 4B,$$

$$D = c_4 - B^4 - 4B^3E - B^2C - 2BCE, \quad F = c_5 - B^4E - B^2CE - DE,$$

gde je B koren jednačine

$$40B^3 - 24c_1B^2 + 2\left(2c_1^2 + c_2\right)B + (c_3 - c_1c_2) = 0. \tag{3.4.6}$$

Jednačina (3.4.6) može imati jedno ili tri realna rešenja, tako da formula (3.4.5), u opštem slučaju, nije jedinstvena.

Za izračunavanje vrednosti polinoma (3.4.5) potrebna su 4 množenja i 5 sabiranja. Napomenimo da računske operacije koje obezbeđuju potrebne transformacije koeficijenata u cilju svođenja na ekonomični oblik (3.4.5) ne ubrajamo u

potrebne operacije za izračunavanje vrednosti polinoma, s obzirom na to da se one za konkretan polinom izvode samo jednom.

Slično, za $n = 6$ postoji ekonomična formula sa 4 množenja i 7 sabiranja (videti [20, str. 57–59]).

Primer 3.4.1. Za aproksimaciju funkcije $x \mapsto 2^x$ za $-1/4 \leq x \leq 0$, sa relativnom greškom manjom od 2^{-28} ($\cong 3.7 \times 10^{-9}$), može se koristiti sledeći polinom četvrtog stepena

$$\begin{aligned} P(x) &= 0.9999999975 + 0.6931466849x \\ &\quad + 0.2402108680x^2 + 0.0553299147x^3 + 0.0088171049x^4. \end{aligned}$$

Saglasno prethodnom imamo

$$A = \sqrt[4]{a_0} = 0.3064301558, \quad B = \frac{1}{4}\left(\frac{a_1}{A^3} - 1\right) = 0.2307347358,$$

$$D = 1.3394747869, \quad C = 0.4377960868, \quad E = 3161295697,$$

pri čemu su rezultati zaokrugljeni na deset decimala.

Algoritam, koji zahteva 3 množenja i 5 sabiranja, se može iskazati u obliku

$$\begin{aligned} \alpha &:= A \cdot x, \\ \beta &:= \alpha + B, \qquad \beta := \beta \cdot \beta, \\ P &:= (\alpha + \beta + C) \cdot (\beta + D) + E. \quad \triangle \end{aligned}$$

Primer 3.4.2. Posmatrajmo polinom

$$P(x) = 32x^5 + 48x^4 + 8x^3 - 12x^2 - 8x + 7,$$

za koji imamo

$$A = \sqrt[5]{a_0} = 2, \quad c_1 = \frac{48}{32} \cdot 2 = 3, \quad c_2 = \frac{8}{32} \cdot 4 = 1, \quad c_3 = \frac{-12}{32} \cdot 8 = -3,$$

$$c_4 = \frac{-8}{32} \cdot 16 = -4, \quad c_5 = \frac{7}{32} \cdot 32 = 7.$$

Jednačina (3.4.6) postaje

$$40B^3 - 72B^2 + 38B - 6 = 0,$$

tj.

$$(B-1)(2B-1)(10B-3)=0,$$

odakle zaključujemo da postoje tri različite ekonomične formule oblika (3.4.5):

1° za $B=1$, na osnovu prethodnog, imamo

$$C=-1,\quad E=-1,\quad D=-1,\quad F=5,$$

tj.

$$P(x)=\left(\left((2x+1)^2-1\right)(2x+1)^2-2\right)(2x-1)+5;$$

2° za $B=0.5$ imamo

$$C=-2.5,\quad E=1,\quad D=-1.4375,\quad F=9,$$

tj.

$$P(x)=\left(\left((2x+0.5)^2-2.5\right)(2x+0.5)^2-1.4375\right)(2x+1)+9;$$

3° za $B=0.3$ imamo

$$C=-1.7,\quad E=1.8,\quad D=-2.2135,\quad F=11.24512,$$

tj.

$$P(x)=\left(\left((2x+0.3)^2-1.7\right)(2x+0.3)^2-2.2135\right)(2x+1.8)+11.24512.$$

Algoritam, kao što je ranije rečeno, zahteva 4 množenja i 5 sabiranja i može se iskazati u obliku:

$$\begin{aligned}&\alpha :=A\cdot x,\quad \beta :=\alpha+B,\quad \beta :=\beta\cdot\beta,\\ &\gamma :=(\beta+C)\cdot\beta+D,\\ &P :=\gamma\cdot(\alpha+E)+F.\quad \triangle\end{aligned}$$

Posmatrajmo sada polinom proizvoljnog stepena n i izrazimo ga u obliku

$$P(x)=Q_0(x)+Q_1(x), \tag{3.4.7}$$

gde su

$$\begin{aligned}Q_0(x)&=a_0x^n+a_2x^{n-2}+a_4x^{n-4}+\cdots,\\ Q_1(x)&=a_1x^{n-1}+a_3x^{n-3}+a_5x^{n-5}+\cdots.\end{aligned}$$

U zavisnosti od toga da li je n parno ili neparno definisaćemo polinome S_0 i T pomoću:

$$S_0(x^2) = Q_0(x), \quad T(x^2) = \frac{Q_1(x)}{x} \quad (n = 2m+2)$$

i

$$S_0(x^2) = Q_1(x), \quad T(x^2) = \frac{Q_0(x)}{x} \quad (n = 2m+1).$$

Primetimo da je $T(t)$ polinom stepena m. Neka su $\alpha_1, \alpha_2, \ldots, \alpha_m$ nule polinoma $T(t)$, koje su nam poznate.

Deljenjem polinoma $S_0(t)$ sa faktorom $t - \alpha_m$ dobijamo polinom $S_1(t)$ i odgovarajući ostatak β_m, tj.

$$S_0(t) = (t - \alpha_m)S_1(t) + \beta_m.$$

Nastavljajući proces deljenja $S_1(t)$ sa $t - \alpha_{m-1}$, itd. dobijamo rekurzivni postupak za određivanje niza $\beta_m,\ \beta_{m-1}, \ldots, \beta_1$:

$$S_{i-1}(t) = (t - \alpha_{m-i+1})\, S_i(t) + \beta_{m-i+1}, \quad i = 1, 2, \ldots, \tag{3.4.8}$$

pri čemu je

$$S_m(t) = \begin{cases} a_0 t + \beta_0 & (n = 2m+2), \\ a_1 & (n = 2m+1). \end{cases}$$

Dakle, za svako $i = 1, \ldots, m$ primenjujemo HORNERovu šemu.

Poznavajući nizove $(\alpha_1, \alpha_2, \ldots, \alpha_m)$ i $(\beta_1, \beta_2, \ldots, \beta_m)$, vrednost polinoma $P(x)$ se može odrediti pomoću $\left[\frac{1}{2}(n+4)\right]$ množenja i n sabiranja.

Na osnovu prethodnog iz (3.4.7) sleduje

$$P(x) = S_0(x^2) + xT(x^2).$$

Tada korišćenjem (3.4.8) za $t = x^2$, jednostavno dobijamo

$$P(x) = \left(\cdots \left(P_0 \cdot \left(x^2 - \alpha_1\right) + \beta_1\right)\left(x^2 - \alpha_2\right) + \cdots + \beta_{m-1}\right)\left(x^2 - \alpha_m\right) + \beta_m,$$

gde je

$$P_0 = \begin{cases} a_0 x^2 + a_1 x + \beta_0 & (n = 2m+2), \\ a_0 x + a_1 & (n = 2m+1). \end{cases} \tag{3.4.9}$$

Primetimo da je u $P(x)$ uključen i sabirak

$$xT(x^2) = Ax\left(x^2-\alpha_1\right)\cdot\left(x^2-\alpha_2\right)\cdots\left(x^2-\alpha_m\right),$$

gde je $A = a_1$ ako je $n = 2m+2$ i $A = a_0$ ako je $n = 2m+1$. Početna vrednost P_0 je, evidentno, jednaka $P_0 = S_m(x^2) + Ax$. Prema tome, startujući sa (3.4.9), vrednost polinoma $P(x)$ možemo odrediti pomoću

$$P_i = P_{i-1}\cdot\left(x^2-\alpha_i\right)+\beta_i, \quad i = 1,\ldots,m,$$

pri čemu je $P(x) = P_m$.

Navedeni metod za izračunavanje vrednosti polinoma dao je KNUTH [36]. Jedna modifikacija ovog metoda je data u [19].

1.3.5 Sumiranje redova i ubrzavanje konvergencije

Izlaganje u ovom odeljku započećemo jednim konkretnim primerom. Naime, razmotrićemo problem sumiranja reda

$$S = \sum_{k=0}^{+\infty}(-1)^k\frac{4}{2k+1} \tag{3.5.1}$$

čija tačna suma $S = \pi$ $(= 3.1415926535\ldots)$.

Da bismo postigli tačnost od 10^{-4} direktnim sumiranjem reda „član po član“, na osnovu nejednakosti $4/(2n+1) < 10^{-4}$ zaključujemo da je potrebno uzeti oko 20000 članova reda. Praktično ovakav način za sumiranje je neizvodljiv. Za red (3.5.1) kažemo da pripada klasi tzv. sporokonvergentnih redova.

Na osnovu prethodnog primera može se zaključiti da kod sporokonvergentnih redova, direktno sumiranje članova reda ne dovodi do željenih rezultata (zbog velikog broja članova reda koje treba sabrati i grešaka zaokrugljivanja koje pri tome nastaju). U ovakvim slučajevima treba naći neke načine za tzv. brzo sumiranje ili kako se često kaže za ubrzavanje konvergencije redova. U literaturi postoji veliki broj metoda koji ovaj problem rešavaju. Navešćemo nekoliko od njih.

Posmatrajmo konvergentni alternativni red

$$\sum_{k=0}^{+\infty}(-1)^k a_k = a_0 - a_1 + a_2 - a_3 + \cdots \qquad (a_k \geq 0), \tag{3.5.2}$$

čija je suma A. Pokazaćemo da je red

$$\frac{1}{2}a_0 - \frac{1}{2}(a_1 - a_0) + \frac{1}{2}(a_2 - a_1) - \cdots, \tag{3.5.3}$$

takođe, konvergentan i da ima istu sumu kao i red (3.5.2).

Pretpostavimo da su $\{A_n\}$ i $\{B_n\}$ nizovi parcijalnih suma redova (3.5.2) i (3.5.3) respektivno. Tada je

$$A_n - B_n = \frac{1}{2}(-1)^{n-1} a_{n-1} \qquad (n = 1,2,\ldots). \tag{3.5.4}$$

Kako je prvi red konvergentan, imamo da je $\lim a_n = 0$ pa iz (3.5.4) sledi

$$\lim B_n = \lim A_n = A.$$

Primenimo izloženu transformaciju na red (3.5.1). Tada je

$$S = 2 + \sum_{k=1}^{+\infty} (-1)^k \frac{1}{2} \left(\frac{4}{2k+1} - \frac{4}{2k-1} \right),$$

tj.

$$S = 2 + \sum_{k=1}^{+\infty} \frac{4(-1)^{k-1}}{4k^2 - 1}. \tag{3.5.5}$$

Ocenimo koliko je sada članova reda (3.5.5) potrebno sabrati da bi se dobila suma sa istom tačnošću kao i ranije. Iz $4/\left(4n^2-1\right) < 10^{-4}$ sledi $n > 100$. Dakle, sada se ista tačnost može postići ako se uzme prvih sto članova reda.

Izložena transformacija za sumiranje alternativnih numeričkih redova predstavlja specijalan slučaj EULER-ABELove[51] transformacije koja se primenjuje kod stepenih redova.

Razmotrimo stepeni red

$$f(x) = \sum_{k=0}^{+\infty} a_k x^k, \tag{3.5.6}$$

čiji je poluprečnik konvergencije R konačan. Ne umanjujući opštost razmatranja možemo uzeti $R = 1$.

Iz $f(x) = a_0 + xg(x)$, gde je

[51] NIELS HENRIK ABEL (1802 – 1829), poznati norveški matematičar sa značajnim naučnim doprinosima. Umro je jako mlad od tuberkuloze. Povodom 200. godišnjice njegovog rođenja, norveška vlada je ustanovila ABEL*ovu nagradu* za izuzetan naučni doprinos u oblasti matematike. Od 2003. godine Norveška akademija nauka svake godine dodeljuje ovu prestižnu nagradu.

$$g(x)=\sum_{k=0}^{+\infty}a_{k+1}x^k,$$

sleduje

$$\begin{aligned}(1-x)g(x)&=(1-x)\sum_{k=0}^{+\infty}a_{k+1}x^k\\&=\sum_{k=0}^{+\infty}a_{k+1}x^k-\sum_{k=0}^{+\infty}a_{k+1}x^{k+1}\\&=a_0+\sum_{k=0}^{+\infty}(a_{k+1}-a_k)x^k,\end{aligned}$$

odakle je

$$f(x)=a_0+\frac{x}{1-x}\left(a_0+\sum_{k=0}^{+\infty}(a_{k+1}-a_k)x^k\right),$$

tj.

$$f(x)=\frac{a_0}{1-x}+\frac{x}{1-x}\sum_{k=0}^{+\infty}\Delta a_k x^k.$$

Iz poslednje formule za $x=-1$ i $a_k\geq 0$, sleduje

$$\sum_{k=0}^{+\infty}(-1)^k a_k=\frac{1}{2}a_0-\frac{1}{2}\sum_{k=0}^{+\infty}(-1)^k\Delta a_k,$$

što je ekvivalentno sa (3.5.3).

Sukcesivnom primenom izložene EULER-ABELove transformacije m puta na red (3.5.6) dobijamo

$$f(x)=\sum_{k=0}^{m-1}\frac{\Delta^k a_0 x^k}{(1-x)^{k+1}}+\left(\frac{x}{1-x}\right)^m\sum_{k=0}^{+\infty}\Delta^m a_k x^k.$$

Posebno je interesantan slučaj kada $m\to+\infty$. Tada imamo

$$f(x)=\frac{1}{1-x}\sum_{k=0}^{+\infty}\Delta^k a_0\left(\frac{x}{1-x}\right)^k. \tag{3.5.7}$$

Primer 3.5.1. Sukcesivnom primenom EULER-ABELove transformacije dva puta na red

$$f(x)=\sum_{k=0}^{+\infty}\frac{x^k}{k+1},$$

dobijamo

$$f(x)=\frac{1}{1-x}+\frac{\left(\frac{1}{2}-1\right)x}{(1-x)^2}+\left(\frac{x}{1-x}\right)^2\sum_{k=0}^{+\infty}\left(\frac{1}{k+3}-\frac{2}{k+2}+\frac{1}{k+1}\right)x^k$$

tj.

$$f(x)=\frac{1}{1-x}-\frac{x}{2(1-x)^2}+\left(\frac{x}{1-x}\right)^2\sum_{k=0}^{+\infty}\frac{2x^k}{(k+1)(k+2)(k+3)}.$$

Primenimo dobijenu formulu na izračunavanje vrednosti $\log 2$. S obzirom na to da je $\log 2=f(-1)$ imamo

$$\log 2\cong\frac{5}{8}+\frac{1}{2}\sum_{k=0}^{n}\frac{(-1)^k}{(k+1)(k+2)(k+3)},$$

odakle, za $n=2$, dobijamo $\log 2\cong 0.6958333$. Dobijeni rezultat je tačan na dve decimale.

Na ubrzavanje datog stepenog reda primenićemo sada formulu (3.5.7). Kako je $a_m=1/(m+1)$ za k-tu razliku imamo

$$\Delta^k a_m=\sum_{i=0}^{k}(-1)^i\binom{k}{i}a_{m+k-i}=-\sum_{i=0}^{k}\frac{(-1)^i}{i-(m+k+1)}\binom{k}{i}.$$

S obzirom na indentitet

$$\sum_{i=0}^{k}\frac{(-1)^i}{i+a}\binom{k}{i}=\frac{k!}{a(a+1)\cdots(a+k)}\quad(a\neq 0,-1,\ldots,-k),\tag{3.5.8}$$

zaključujemo da je

$$\Delta^k a_m=\frac{(-1)^k k!m!}{(m+k+1)!}\qquad\text{i}\qquad\Delta^k a_0=\frac{(-1)^k}{k+1}.$$

Primenom (3.5.7) dobijamo da je u ovom slučaju

$$f(x)=\frac{1}{1-x}\sum_{k=0}^{+\infty}\frac{(-1)^k}{k+1}\left(\frac{x}{1-x}\right)^k.$$

Najzad, za $x=-1$ imamo

$$\log 2=\sum_{k=0}^{+\infty}\frac{(-1)^k}{k+1}=\sum_{k=1}^{+\infty}\frac{1}{2^k k}.$$

△

Primer 3.5.2. Posmatrajmo alternativni red

$$S_r=\sum_{k=1}^{+\infty}\frac{(-1)^{k-1}}{(2k+2r+1)\pi^{2k}}.$$

Primetimo da se za $r=0$ dobija

$$S_0=\sum_{k=1}^{+\infty}\frac{(-1)^{k-1}}{(2k+1)\pi^{2k}}=1-\pi\arctan\frac{1}{\pi}. \tag{3.5.9}$$

Neka je potrebno odrediti narednih n članova niza $\{S_n\}$. Nije teško videti da važi rekurentna relacija

$$S_r=\frac{1}{2r-1}-\pi^2 S_{r-1}\qquad(r=1,2,\ldots,n). \tag{3.5.10}$$

Startujući sa (3.5.9) i korišćenjem rekurentne relacije (3.5.10) dobijamo rezultate koji su dati u tabeli 3.5.1, pri čemu je uzeto $n=12$. Broj u zagradi ukazuje na decimalni eksponent. Sva izračunavanja su sprovedena sa aritmetikom od 12 dekadnih cifara u mantisi. Dobijeni članovi niza $\{S_r\}$, su evidentno pogrešni, jer je $S_r>0$ za svako r. U ovoj tabeli, rezultati su zaokrugljeni na 10 značajnih cifara.

Može se postaviti pitanje zbog čega su ovi rezultati pogrešni. Odgovor leži u tzv. slaboj uslovljenosti rekurentne relacije (3.5.10). Stavljajući $r=1,2,\ldots,n$ na osnovu (3.5.10) dobijamo sistem od n linearnih jednačina

$$\begin{bmatrix} 1 & 0 & & 0 & 0\\ \pi^2 & 1 & & 0 & 0\\ \vdots & & \ddots & & \\ 0 & 0 & & 1 & 0\\ 0 & 0 & & \pi^2 & 1 \end{bmatrix}\cdot\begin{bmatrix} S_1\\ S_2\\ \vdots\\ S_{n-1}\\ S_n \end{bmatrix}=\begin{bmatrix} 1/3-\pi^2 S_0\\ 1/5\\ \vdots\\ 1/(2n-1)\\ 1/(2n+1) \end{bmatrix},$$

čija je matrica dvodijagonalna. Slično sistemu jednačina iz primera 2.1.2, i ovde imamo slabouslovljeni sistem, kod koga male promene (perturbacije) u sistemu,

Tabela 3.5.1.

r	S_r pomoću (3.5.10)	S_r pomoću (3.5.11)
0	$0.3185831012(-1)$	$0.3185831012(-1)$
1	$0.1890441551(-1)$	$0.1890441558(-1)$
2	$0.1342089717(-1)$	$0.1342089682(-1)$
3	$0.1039819711(-1)$	$0.1039820054(-1)$
4	$0.8485019141(-2)$	$0.8484985269(-2)$
5	$0.7165308652(-2)$	$0.7165642950(-2)$
6	$0.6204315120(-2)$	$0.6201015723(-2)$
7	$0.5432530853(-2)$	$0.5465094591(-2)$
8	$0.5206599001(-2)$	$0.4885207787(-2)$
9	$0.1244506534(-2)$	$0.4416510673(-2)$
10	$0.3533626045(-1)$	$0.4029834444(-2)$
11	$-0.3052766508(\ 0)$	$0.3705389102(-2)$
12	$0.3052959776(\ 1)$	$0.3429275407(-2)$

nastale kao posledica zaokrugljivanja, prouzrokuju velike promene u rešenju, što vodi pogrešnom rezultatu.

Ako se, međutim, najpre izračuna, sa potrebnom tačnošću, poslednji član niza S_n, a rekurentna relacija predstavi u obliku

$$S_{r-1} = \frac{1}{\pi^2}\left(\frac{1}{2r-1} - S_r\right), \quad r = n, \ldots, 1, \tag{3.5.11}$$

dobijamo dobro uslovljeni proces. Ekvivalentni sistem linearnih jednačina ima oblik

$$\begin{bmatrix} \pi^2 & 0 & \ldots & 0 & 0 \\ 1 & \pi^2 & & 0 & 0 \\ \vdots & & \ddots & & \\ 0 & 0 & & \pi^2 & 0 \\ 0 & 0 & & 1 & \pi^2 \end{bmatrix} \cdot \begin{bmatrix} S_{n-1} \\ S_{n-2} \\ \vdots \\ S_1 \\ S_0 \end{bmatrix} = \begin{bmatrix} 1/(2n+1) - S_n \\ 1/(2n-1) \\ \vdots \\ 1/5 \\ 1/3 \end{bmatrix}.$$

Da bismo odredili S_n, posmatraćemo stepeni red

$$f(x) = \sum_{k=1}^{+\infty} \frac{x^{k-1}}{2k+2n+1} = \sum_{k=0}^{+\infty} \frac{x^k}{2k+2n+3}, \tag{3.5.12}$$

čiji je poluprečnik korvergencije jednak jedinici, i na čije sumiranje ćemo primeniti EULER-ABELovu transformaciju (3.5.7) (slučaj kada $m \to +\infty$). Kako je $a_j = 1/(2j+2n+3)$, imamo

$$\Delta^k a_j = \sum_{i=0}^{k} (-1)^i \binom{k}{i} a_{j+k-i} = \sum_{i=0}^{k} \frac{(-1)^i}{2(j+k-i)+2n+3} \binom{k}{i},$$

tj.

$$\Delta^k a_j = -\frac{1}{2} \sum_{i=0}^{k} \frac{(-1)^k}{i-\left(j+n+k+\frac{3}{2}\right)} \binom{k}{i}.$$

Korišćenjem identiteta (3.5.8), sa $a = -(n+k+3/2) = -b$ i $j = 0$ dobijamo

$$\Delta^k a_0 = -\frac{1}{2} \cdot \frac{(-1)^{k+1} k!}{b^{(k+1)}},$$

gde je $p^{(s)} = p(p-1)\cdots(p-s+1)$. Dakle, na osnovu (3.5.7) imamo

$$f(x) = \frac{1}{2(1-x)} \sum_{k=0}^{+\infty} \frac{(-1)^k k!}{b^{(k+1)}} \left(\frac{x}{1-x}\right)^k. \tag{3.5.13}$$

Ako stavimo $x = -1/\pi^2$, na osnovu (3.5.12) i (3.5.13) zaključujemo da je

$$S_n = \frac{1}{\pi^2} f\left(-1/\pi^2\right) = \frac{1}{2} \sum_{k=0}^{+\infty} \frac{k!}{\left(\pi^2+1\right)^{k+1} b^{(k+1)}}.$$

Numerički red u poslednjoj jednakosti brzo konvergira. Parcijalne sume ovog reda za $n = 12$ su redom

$$\begin{aligned}
S_{12}^{(0)} &= 0.3407395124 \cdot 10^{-2}, \\
S_{12}^{(1)} &= 0.3429014381 \cdot 10^{-2}, \\
S_{12}^{(2)} &= 0.3429271021 \cdot 10^{-2}, \\
S_{12}^{(3)} &= 0.3429275314 \cdot 10^{-2}, \\
S_{12}^{(4)} &= 0.3429275404 \cdot 10^{-2}, \\
S_{12}^{(5)} &= 0.3429275407 \cdot 10^{-2}.
\end{aligned}$$

Uzimajući $S_{12} \cong S_{12}^{(5)}$ i primenom rekurentne relacije (3.5.11) dobićemo rezultate koji su dati, takođe, u tabeli 3.5.1. Primetimo da se dobijena vrednost za S_0 poklapa sa tačnom vrednošću (3.5.9) na svih 10 značajnih cifara. Ovo ukazuje da su sve vrednosti niza $\{S_n\}$ dobijene sa visokom tačnošću. $\triangle$

Za ubrzavanje konvergencije redova postoje i transformacije koje se zasnivaju na ubrzanju konvergencije nizova. Među njima postoje kako linearne, tako i nelinearne transformacije.

Neka je $\{S_n\}$ niz parcijalnih suma alternativnog reda (3.5.2), tj.

$$S_n = \sum_{k=0}^{n-1} (-1)^k a_k \qquad (n = 1,2,\ldots). \tag{3.5.14}$$

Najjednostavnija linearna transformacija je definisana aritmetičkom sredinom sukcesivnih parcijalnih suma, poznata kao CESÀROova[52] transformacija,

$$S'_n = C(S_n) = \frac{1}{2}(S_n + S_{n+1}) \qquad (n = 1,2,\ldots).$$

Niz $\{S'_n\}$ brže konvergira od niza $\{S_n\}$ ka istoj granici S ($\lim S'_n = \lim S_n = S$). Daljom primenom iste transformacije na niz $\{S'_n\}$ dobija se ubrzani niz $\{S''_n\}$, gde je

$$S''_n = C(S'_n) = \frac{1}{2}\left(S'_n + S'_{n+1}\right) \qquad (n = 1,2,\ldots).$$

Uopšte, polazeći od niza parcijalnih suma (3.5.14) mogu se sukcesivno konstruisati nizovi $\{S_n^{(m)}\}$ $(m = 1,2,\ldots)$ pomoću

$$S_n^{(m)} = C\left(S_n^{(m-1)}\right) = \frac{1}{2}\left(S_n^{(m-1)} + S_{n+1}^{(m-1)}\right) \qquad (n = 1,2,\ldots), \tag{3.5.15}$$

gde smo stavili $S^{(0)} \equiv S_n$. Šematski postupak (3.5.15) se može interpretirati kroz konstrukciju tzv. S-tabele:

$$\begin{array}{ccccccc}
S_1^{(0)} & \longrightarrow & S_1^{(1)} & \longrightarrow & S_1^{(2)} & \longrightarrow & S_1^{(3)} \longrightarrow \\
\searrow & & \searrow & & \searrow & & \\
S_2^{(0)} & \longrightarrow & S_2^{(1)} & \longrightarrow & S_2^{(2)} & \longrightarrow & \vdots \\
\searrow & & \searrow & & & & \\
S_3^{(0)} & \longrightarrow & S_3^{(1)} & \longrightarrow & \vdots & & \\
\searrow & & & & & & \\
S_4^{(0)} & \longrightarrow & \vdots & & & & \\
\vdots & & & & & &
\end{array}$$

[52] ERNESTO CESÀRO (1859 – 1906), italijanski matematičar poznat po radovima iz diferencijalne geometrije i konceptu „sumiranja“ divergentnih redova.

Ona se konstruiše izračunavajući elemente sledećim redom

$$S_1^{(0)};\ S_2^{(0)},\ S_1^{(1)};\ S_3^{(0)},\ S_2^{(1)},\ S_1^{(2)};\ \text{itd.}$$

Proces se prekida kada je, na primer, ispunjen uslov $|S_1^{(m)} - S_2^{(m-1)}| \leq \varepsilon$ gde je ε unapred data dozvoljena greška, i tada se uzima $S \cong S_1^{(m)}$.

Algoritamski se ovaj postupak može iskazati kroz sledeća četiri koraka:

1° $s_1 := a_0, \quad z := 1, \quad n := 0;$

2° $n := n+1, \quad z := -z, \quad s_{n+1} := s_n + z a_n;$

3° za $k = n, \ldots, 1, \quad s_k := \dfrac{1}{2}(s_{k+1} + s_k);$

4° ako je

$$\begin{aligned} |s_1 - s_2| &> \varepsilon \ \text{ preći na } 2^\circ, \\ &\leq \varepsilon \ \text{ kraj izračunavanja } S := s_1. \end{aligned}$$

Primetimo da ovaj postupak ne zahteva memorijski prostor za pamćenje čitave S-tabele. Naime, u n-tom koraku primene ovog postupka pamte se samo dijagonalni elementi $S_{n+1}^{(0)}$, $S_n^{(1)}, \ldots, S_2^{(n-2)}$, $S_1^{(n)}$, koji su označeni redom sa s_{n+1}, s_n, …, s_2, s_1. Sa $z := -z$ obezbeđuje se neophodna promena znaka kod izračunavanja parcijalnih suma alternativnog reda.

Može se pokazati da nizovi $\{S_n^{(m)}\}_{n\in\mathbb{N}}$ i $\{S_n^{(m)}\}_{n\in\mathbb{N}_0}$ (po kolonama i vrstama u S-tabeli), kao i nizovi po dijagonalama konvergiraju ka S.

Napomena 3.5.1. Kako je $S_{n+1} = S_n + (-1)^{n+1} a_{n+1}$ imamo da je

$$\begin{aligned} S_n^{(1)} &= \frac{1}{2}(S_n + S_{n+1}) = S_n + \frac{1}{2}(-1)^{n+1} a_{n+1}, \\ S_n^{(2)} &= \frac{1}{2}\left(S_n^{(1)} + S_{n+1}^{(1)}\right) = S_n + \frac{1}{4}\left(3(-1)^{n+1} a_{n+1} + (-1)^{n+2} a_{n+2}\right). \end{aligned}$$

Matematičkom indukcijom jednostavno dokazujemo da važi sledeća reprezentacija

$$S_n^{(m)} = S_n + \frac{1}{2^m} \sum_{k=1}^{m} \alpha_{mk} (-1)^{n+k} a_{n+k}, \tag{3.5.16}$$

gde su koeficijenti α_{mk} dati rekurzivno pomoću

$$\begin{aligned} \alpha_{m1} &= 2^{m-1} + \alpha_{m-1,i}, \\ \alpha_{mk} &= \alpha_{m-1,k-1} + \alpha_{m-1,k} \qquad (k = 2, \ldots, m-1), \\ \alpha_{mm} &= \alpha_{m-1,m-1}. \end{aligned}$$

Primetimo da je $\alpha_{m1} = 2^m - 1$ i $\alpha_{mm} = 1$.

Na osnovu prethodnog vidimo da se elementi S-tabele mogu izračunavati i pomoću (3.5.16). Međutim, ovakav postupak bi bio numerički neefikasan zbog velikog broja računskih operacija.

Primer 3.5.3. Formirajmo S-tabelu za određivanje sume alternativnog reda

$$S = \sum_{k=0}^{\infty} \frac{(-1)^k}{k+1} = \log 2 = 0.693147\ldots$$

sa tri tačne decimale. Ovde je

$$a_k = \frac{1}{k+1}, \quad S^{(0)} = 1, \quad S_{k+1}^{(0)} = S_k^{(0)} + (-1)^{k+1} a_{k+1}.$$

Primenom (3.5.15), saglasno prethodno navedenim algoritamskim koracima 1° – 4° dobijamo S-tabelu:

1.0000	0.7500	0.7083	0.6978	0.6947	0.6936	0.6932
0.5000	0.6666	0.6874	0.6916	0.6926	0.6929	
0.8333	0.7083	0.6958	0.6937	0.6932		
0.5833	0.6833	0.6916	0.6928			
0.7833	0.7000	0.6940				
0.6167	0.6881					
0.7595						

Pri ovome smo sva izračunavanja sproveli tako što smo sve međurezultate zaokrugljivali na četiri decimale. Kako je

$$|S_1^{(6)} - S_2^{(5)}| = |0.6932 - 0.6929| < 0.5 \cdot 10^{-3},$$

zaključujemo da je $S \cong 0.693$. Ako se želi postići veća tačnost, izračunavanja mogu da se sprovode u aritmetici sa većim brojem cifara u mantisi. Čitaocu prepuštamo da izvrši analizu prostiranja grešaka kroz kolone S-tabele, znajući relativne mašinske greške elemenata iz prve kolone r_n ($|r_n| \leq 0.5 \cdot 10^{-\ell+1}$, gde je ℓ broj cifara mantise).

Ako izračunavanja sprovodemo, kao u primeru 3.5.2, sa aritmetikom od 12 cifara u mantisi, u zavisnosti od zahtevne tačnosti ε, dobijamo rezultate koji su navedeni u tabeli 3.5.2.

U drugoj koloni dat je broj m koji ukazuje na broj sukcesivnih primena transformacije za postizanje zadate tačnosti ε. Drugim rečima, m je najmanji prirodan broj za koji je $|S_1^{(m)} - S_2^{(m-1)}| \leq \varepsilon$. Prva pogrešna cifra u rezultatu $S_1^{(m)}$

Tabela 3.5.2.

ε	m	$S_1^{(m)}$
10^{-2}	4	0.69<u>4</u>7016667
10^{-3}	6	0.693<u>3</u>779762
10^{-4}	8	0.6931<u>8</u>42138
10^{-5}	10	0.6931<u>5</u>36346
10^{-6}	13	0.693147<u>6</u>997
10^{-7}	16	0.693147<u>2</u>258
10^{-8}	19	0.6931471<u>8</u>47

je podvučena. Primetimo da direktno sumiranje reda „član po član“ zahteva, prema LEIBNIZovom[53] kriterijumu, približno $[1/\varepsilon]$ članova. Na primer, za tačnost $\varepsilon = 10^{-8}$ potrebno je sabrati fantastičnih 100 miliona članova. $\triangle$

Na kraju ovog odeljka napomenimo da postoji čitava klasa nelinearnih transformacija koje se primenjuju kod sumiranja redova (videti posebno BREZINSKI[54] [5]–[9], SHANKS[55] [60] i DELAHAYE[56] [14]–[16]). Jedna od najjednostavnijih je Δ^2-transformacija definisana pomoću

$$S'_n = T(S_n) = S_n - \frac{(\Delta S_n)^2}{\Delta^2 S_n} \qquad (n = 1,2,\ldots)$$

i o njoj će biti reči u opštoj teoriji iterativnih procesa (odeljak 3.2.2).

Napomena 3.5.2. Ako se pretpostavi asimptotsko ponašanje parcijalne sume S_k u obliku $S_k \sim A + B\mathrm{e}^{-\alpha k}$ $(\alpha > 0)$, tada eliminacijom B i α iz ove asimptotske relacije za $k = n,\ n+1,\ n+2$ dobijamo

$$A \sim S_n - \frac{(S_{n+1} - S_n)^2}{S_{n+2} - 2S_{n+1} + S_n},$$

[53] GOTTFRIED WILHELM VON LEIBNIZ (1646 – 1716), poznati nemački matematičar i filozof. Nezavisno od NEWTONa zasnovao je infinitezimalni račun, uveo simboličko označavanje za diferenciranje i dao niz doprinosa u mnogim oblastima matematike, filozofije, fizike, itd.

[54] CLAUDE BREZINSKI (1941 –), francuski matematičar sa doprinosima u ekstrapolacionim metodima, aproksimaciji racionalnih funkcija i formalnoj teoriji ortogonalnosti.

[55] DANIEL SHANKS (1917 – 1996), američki matematičar sa doprinosima u numeričkoj analizi i teoriji brojeva.

[56] JEAN-PAUL DELAHAYE (1952 –), francuski informatičar i matematičar.

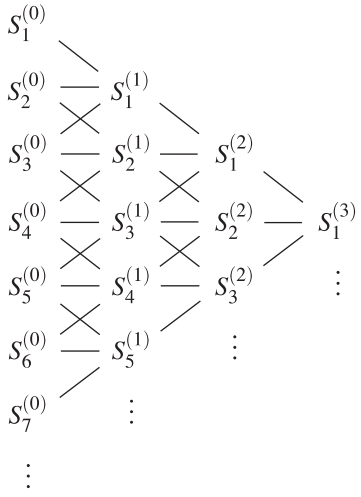
tj. $A \sim T(S_n)$. Primetimo da je

$$S = \lim_{k\to+\infty} S_k = \lim_{k\to+\infty} \left(A + B\mathrm{e}^{-\alpha k}\right) = A.$$

Kao i CESÀROova transformacija i Δ^2-transformacija se može sukcesivno primenjivati

$$S_n^{(m)} = T\left(S_n^{(m-1)}\right) = S_n^{(m-1)} - \frac{\left(\Delta S_n^{(m-1)}\right)^2}{\Delta^2 S_n^{(m-1)}} \qquad (n = 1,2,\ldots),$$

gde smo, kao i ranije, stavili $S_n^{(0)} \equiv S_n$. Naravno, konstrukcija S-tabele, u ovom slučaju zbog nelinearnosti, je nešto složenija i može se šematski predstaviti u obliku:



Primer 3.5.4. Na red iz primera 3.5.3 primenimo Δ^2-transformaciju. Odgovarajuća S-tabela ima oblik

1.000000			
0.500000	0.700000		
0.833333	0.690476	0.693277	
0.583333	0.694444	0.693106	0.693149
0.783333	0.692424	0.693163	
0.616667	0.693590		
0.759524			

pri čemu smo koristili aritmetiku sa šestorazrednom mantisom. Dobijeni rezultat u četvrtoj koloni ima tačnih pet decimala. △

1.3.6 EULER–MACLAURINova sumaciona formula i RIEMANNova zeta funkcija

Za razliku od metoda za sumiranje redova koji su razmatrani u prethodnom odeljku, postoje metodi sumiranja koji su povezani sa integracijom. Takvi metodi će biti detaljno analizirani u jednoj od narednih knjiga iz ove serije. Ovde ćemo dati samo osnovnu EULER–MACLAURINovu[57] sumacionu formulu, koja predstavlja osnov za razvoj tzv. sumaciono/integracionih metoda.

EULER–MACLAURINova sumaciona formula igra važnu ulogu u mnogim oblastima numeričke analize, analitičke teorije brojeva, u teoriji asimptotskih razvoja, kao i u mnogim primenama u drugim oblastima. Formulu je otkrio L. EULER 1735. godine u vezi sa određivanjem sume reda[58]

$$1+\frac{1}{2^2}+\frac{1}{3^2}+\cdots=\frac{\pi^2}{6}.$$

U modernoj terminologiji radi se o nalaženju vrednosti RIEMANNove[59] zeta funkcije $\zeta(2)$.

Za $\operatorname{Re} s>1$, RIEMANNova zeta funkcija je definisana pomoću reda ili pomoću proizvoda preko svih prostih brojeva p, tj.

$$\zeta(s)=\sum_{n=1}^{+\infty}\frac{1}{n^s}=1+\frac{1}{2^s}+\frac{1}{3^s}+\cdots=\prod_{p}\frac{1}{1-p^{-s}}\,. \tag{3.6.1}$$

dok je za ostale tačke u kompleksnoj ravni definisana analitičkim produženjem koje zadovoljava funkcionalnu jednačinu

$$\zeta(s)=2^s\pi^{s-1}\sin\frac{\pi s}{2}\Gamma(1-s)\zeta(1-s), \tag{3.6.2}$$

gde je Γ gama funkcija (videti odeljak 1.3.11).

[57] COLIN MACLAURIN (1698 – 1746), poznati škotski matematičar.

[58] Problem (poznat i kao „Basel problem“) je postavio još 1644. godine italijanski matematičar PIETRO MENGOLI (1626 – 1686). Inače, MENGOLI je, u jednom svom članku iz 1650. godine, dokazao da je $1-\frac{1}{2}+\frac{1}{3}-\cdots=\log 2$.

[59] GEORG FRIEDRICH BERNHARD RIEMANN (1826 – 1866), veliki nemački matematičar.

Funkcija $s \mapsto \zeta(s)$ je regularna u celoj kompleksnoj ravni $\mathbb{C}$ izuzimajući prost pol u tački $s = 1$, sa LAURENTovim[60] razvojem

$$\zeta(s) = \frac{1}{s-1} + \sum_{\nu=1}^{+\infty} \gamma_\nu (s-1)^k,$$

gde su γ_ν takozvane STIELTJESove[61] konstante date sa

$$\gamma_\nu = \frac{(-1)^\nu}{\nu!} \lim_{n\to+\infty} \left(\sum_{k\le n} \frac{\log^\nu k}{k} - \frac{\log^{\nu+1} n}{\nu+1} \right).$$

Ovde, γ_0 je EULERova konstanta definisana ranije u odeljku 1.3.2 (formula (3.2.6)). Detaljna analiza RIEMANNove zeta funkcije može se naći u poznatim monografijama IVIĆa[62] [33] i [34].

RIEMANNova ζ funkcija daje vezu sa raspodelom prostih brojeva i najznačajnija je funkcija u *Teoriji brojeva*. Na osnovu produktne formule u (3.6.1) jasno je da se $\zeta(s)$ ne može anulirati u poluravni $\operatorname{Re} s > 1$, a takođe ni u levoj poluravni $\operatorname{Re} s < 0$, na osnovu (3.6.2), sem u tačkama $s_k = -2k$, $k \in \mathbb{N}$, na negativnom delu realne ose (trivijalne nule zeta funkcije). Dakle, kompleksne (netrivijalne) nule RIEMANNove ζ funkcije mogu se nalaziti samo u traci $0 < \operatorname{Re} s < 1$, a one su bitne jer kontrolišu raspodelu prostih brojeva! RIEMANN je 1859. godine formulisao hipotezu da *sve netrivijalne nule funkcije $\zeta(s)$ leže na kritičnoj pravoj $s = 1/2 + \mathrm{i}y$ u kompleksnoj ravni*.

Grafik (realne) funkcije $y \mapsto \zeta(\frac{1}{2} + \mathrm{i}y)$ na kritičnoj pravoj, kada $0 \le y \le 60$, je prikazan na slici 3.6.1.

RIEMANNova hipoteza (RH) je do sada numerički potvrđena na nekoliko milijardi prvih nula.

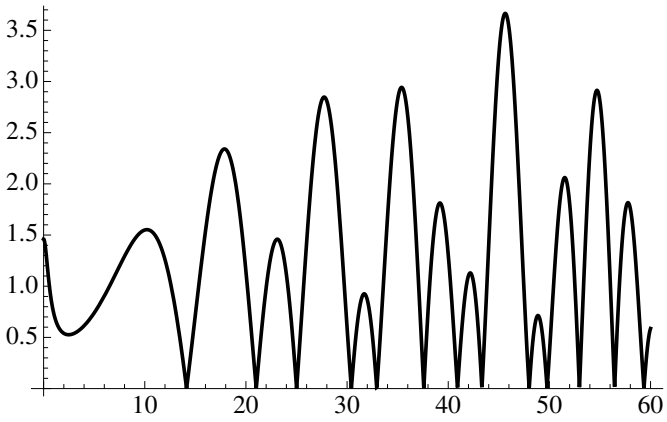
HILBERT[63] jednom prilikom izjavio: *ako me za* 1000 *godina neko probudi iz mrtvih, moje prvo pitanje će biti – da li je dokazana RH?*

[60] PIERRE ALPHONSE LAURENT (1813 – 1854), francuski matematičar.

[61] THOMAS JOANNES STIELTJES (1856 – 1894), holandski matematičar sa doprinosima u matematičkoj analizi.

[62] ALEKSANDAR IVIĆ (1949 –), srpski matematičar, redovni profesor Rudarsko-geološkog fakulteta Univerziteta u Beogradu i redovni član Srpske akademije nauka i umetnosti (SANU). Bavi se teorijom brojeva.

[63] DAVID HILBERT (1862 – 1943), veliki nemački matematičar, sa fundamentalnim doprinosima u skoro svim oblastima matematike (algebra, teorija brojeva, geometrija, varijacioni račun, integralne jednačine, itd.). Na Svetskom kongresu matematičara u Parizu 1900. godine HILBERT je izložio 23 čuvena matematička problema, među kojima i RH i najavio nekoliko pravaca za razvoj matematike u 20. veku.



Slika 3.6.1. Grafik funkcije $y \mapsto \zeta(\frac{1}{2}+\mathrm{i}y)$ za $0 \le y \le 60$

RIEMANNova hipoteza je danas svakako najizazovniji matematički problem[64]. Izračunavanje vrednosti RIEMANNove zeta funkcije razmatraćemo na kraju ovog odeljka primenom EULER–MACLAURINove sumacione formule, a neke osobine gama funkcije, kao i izračunavanje vrednosti $\Gamma(z)$ za $z \in \mathbb{C}$, biće dato u odeljku 1.3.11.

Pre formulisanja EULER–MACLAURINove sumacione formule neophodne su neke definicije.

Definicija 3.6.1. Koeficijenti B_k, $k = 0, 1, \dots$, u razvoju

$$\frac{t}{\mathrm{e}^t - 1} = \sum_{k=0}^{+\infty} \frac{B_k}{k!} t^k \qquad (|t| < 2\pi) \tag{3.6.3}$$

nazivaju se BERNOULLIjevi[65] brojevi.

Na osnovu (3.6.3), tj. jednakosti

$$t = \Big(t + \frac{1}{2!}t^2 + \frac{1}{3!}t^3 + \cdots\Big)\Big(B_0 + B_1 t + \frac{1}{2!}B_2 t^2 + \cdots\Big),$$

[64] RH spada među sedam tzv. *Millennium Prize Problems*, koje je CMI (Clay Mathematics Institute) objavio 2000. godine i za korektno rešenje svakog problema ponudio po milion američkih dolara. Jedan od tih milenijumskih problema – POINCARÉovu hipotezu – dokazao je mladi ruski matematičar GRIGORI PERELMAN (1966 –). Na Svetskom Kongresu u Madridu 2006. godine on je odbio da primi FIELDSovu medalju, a 2010. i milenijumsku nagradu!

[65] JACOB BERNOULLI (1655 – 1705), jedan od znamenitih švajcarskih matematičara, koji potiče iz poznate familije BERNOULLIjevih.

zaključujemo da su brojevi $B_{2k+1} = 0$ $(k = 1, 2, \ldots)$ i

$$B_0 = 1,\ B_1 = -\frac{1}{2},\ B_2 = \frac{1}{6},\ B_4 = -\frac{1}{30},\ B_6 = \frac{1}{42},$$
$$B_8 = -\frac{1}{30},\ B_{10} = \frac{5}{66},\ B_{12} = -\frac{691}{2730},\ B_{14} = \frac{7}{6},\ \text{itd.}$$

BERNOULLIjevi brojevi parnog reda mogu se izraziti pomoću EULERove formule (videti, na primer, [54, str. 48])

$$B_{2k} = (-1)^{k+1} \frac{2(2k)!}{(2\pi)^{2k}} \zeta(2k), \qquad k \geq 1. \tag{3.6.4}$$

Na osnovu (3.6.4) zaključujemo da za svako $k \geq 1$ važi nejednakost

$$(-1)^{k+1} B_{2k} > 0.$$

BERNOULLIjevi brojevi se mogu izraziti kao vrednosti odgovarajućih BERNOULLIjevih polinoma $B_k(x)$ u tački $x = 0$, tj. $B_k = B_k(0)$.

Definicija 3.6.2. Polinomi $B_k(x)$, $k = 0, 1, \ldots$, u razvoju

$$\frac{t\mathrm{e}^{xt}}{\mathrm{e}^t - 1} = \sum_{k=0}^{+\infty} \frac{B_k(x)}{k!} t^k. \tag{3.6.5}$$

nazivaju se BERNOULLIjevi polinomi.

Prvih nekoliko BERNOULLIjevih polinoma su

$$B_0(x) = 1,\ B_1(x) = x - \frac{1}{2},\ B_2(x) = x^2 - x + \frac{1}{6},\ B_3(x) = x^3 - \frac{3x^2}{2} + \frac{x}{2},$$
$$B_4(x) = x^4 - 2x^3 + x^2 - \frac{1}{30},\ B_5(x) = x^5 - \frac{5x^4}{2} + \frac{5x^3}{3} - \frac{x}{6},\ \text{itd.}$$

Neke interesantne osobine ovih polinoma su

$$B_k'(x) = kB_{k-1}(x),\ B_k(1-x) = (-1)^k B_k(x),\ \int_0^1 B_k(x)\,\mathrm{d}x = 0 \quad (k \in \mathbb{N}).$$

Za $0 \leq x < 1$ i $r \geq 1$ važe sledeći FOURIERovi razvoji za BERNOULLIjeve polinome (videti, na primer, [54, str. 49])

$$B_{2r}(x) = (-1)^{r-1}\frac{2(2r)!}{(2\pi)^{2r}}\sum_{k=1}^{+\infty}\frac{\cos(2k\pi x)}{k^{2r}}, \tag{3.6.6}$$

$$B_{2r-1}(x) = (-1)^{r}\frac{2(2r-1)!}{(2\pi)^{2r-1}}\sum_{k=1}^{+\infty}\frac{\sin(2k\pi x)}{k^{2r-1}}.$$

Sledeća teorema daje EULER–MACLAURINovu sumacionu formulu sa ocenom ostatka.

Teorema 3.6.1. *Za svako $n, r \in \mathbb{N}$ i $f \in C^{2r}[0,n]$ važi formula*

$$\sum_{k=0}^{n} f(k) = \int_0^n f(x)\,\mathrm{d}x + \frac{1}{2}(f(0)+f(n)) + \sum_{j=1}^{r}\frac{B_{2j}}{(2j)!}\left[f^{(2j-1)}(n) - f^{(2j-1)}(0)\right] + E_r(f). \tag{3.6.7}$$

Ostatak $E_r(f)$ se može izraziti u obliku

$$E_r(f) = (-1)^r \sum_{k=1}^{+\infty}\int_0^n \frac{\mathrm{e}^{\mathrm{i}2\pi kx} + \mathrm{e}^{-\mathrm{i}2\pi kx}}{(2\pi k)^{2r}} f^{(2r)}(x)\,\mathrm{d}x, \tag{3.6.8}$$

ili u obliku

$$E_r(f) = -\int_0^n \frac{B_{2r}(x-[x])}{(2r)!} f^{(2r)}(x)\,\mathrm{d}x, \tag{3.6.9}$$

gde $[x]$ označava najveći ceo broj ne veći od x.

Formulu (3.6.7) je nezavisno dokazao MACLAURIN, ali je on, za razliku od EULERa, koristio formulu za izračunavanje integrala. Istorijski podaci u vezi sa EULER–MACLAURINovom formulom mogu se naći u radu [3], a niz drugih detalja u radovima [48], [2], [26], [27], [12], [52].

U dokazu teoreme 3.6.1 koristićemo red

$$L_j(z) = (-1)^j \sum_{k=1}^{+\infty}\frac{\mathrm{e}^{-2k\pi z}}{(2k\pi)^j} \quad (j \in \mathbb{N}_0), \tag{3.6.10}$$

koji za $j=0$ i $j=1$ konvergira u otvorenoj desnoj poluravni $\mathrm{Re}\,z > 0$, a za $j \geq 2$ u zatvorenoj oblasti $\mathrm{Re}\,z \geq 0$, pri čemu u tim oblastima L_j definiše analitičku funkciju. Napomenimo, takođe, da za $\mathrm{Re}\,z > 0$ i svako $j \in \mathbb{N}_0$ važe jednakosti

$$L_j(z+\mathrm{i}) = L_j(z), \quad L'_{j+1}(z) = L_j(z). \tag{3.6.11}$$

Takođe, na osvnovu (3.6.6), zaključujemo da se vrednost $L_{2j}(0)$ za $j \geq 1$ može izraziti pomoću BERNOULLIjevih brojeva, tj.

$$L_{2j}(0) = \sum_{k=1}^{+\infty} \frac{1}{(2k\pi)^{2j}} = (-1)^{j-1} \frac{B_{2j}}{2(2j)!}, \quad j \geq 1. \tag{3.6.12}$$

Dokaz teoreme 3.6.1. Neka $f \in C^{2r}[0,n]$. Definišimo polinom drugog stepena na segmentu $[0,1]$ kao $p(x) = \frac{1}{2}B_2(x) = \frac{1}{2}x^2 - \frac{1}{2}x + \frac{1}{12}$ i sa $\widetilde{p}(x)$ označimo njegovo periodično produženje na $\mathbb{R}$ (sa periodom 1). Ako na integral $\int_0^1 p(x)f''(x)\,\mathrm{d}x$ primenimo dva puta parcijalnu integraciju imamo

$$\begin{aligned} \int_0^1 p(x)f''(x)\,\mathrm{d}x &= \big[p(x)f'(x) - p'(x)f(x)\big]_0^1 + \int_0^1 f(x)\,\mathrm{d}x \\ &= -\frac{1}{2}\big(f(0)+f(1)\big) + \frac{1}{12}\big(f'(1)-f'(0)\big) + \int_0^1 f(x)\,\mathrm{d}x. \end{aligned} \tag{3.6.13}$$

Posmatrajmo sada integral od $\widetilde{p}(x)f''(x)$ na $[0,n]$ u obliku

$$\int_0^n \widetilde{p}(x)f''(x)\,\mathrm{d}x = \sum_{k=0}^{n-1} \int_k^{k+1} \widetilde{p}(x)f''(x)\,\mathrm{d}x = \sum_{k=0}^{n-1} \int_0^1 p(x)f''(x+k)\,\mathrm{d}x,$$

a zatim na integrale pod sumom na desnoj strani ove formule primenimo formulu (3.6.13), tako da dobijamo

$$\begin{aligned} \int_0^n \widetilde{p}(x)f''(x)\,\mathrm{d}x = -\frac{1}{2}f(0) - \sum_{k=1}^{n-1} f(k) - \frac{1}{2}f(n) \\ + \frac{1}{12}\big(f'(n)-f'(0)\big) + \int_0^n f(x)\,\mathrm{d}x. \end{aligned} \tag{3.6.14}$$

S obzirom na činjenicu da je $\widetilde{p}$ apsolutno neprekidna parna funkcija, njen FOURIERov red je apsolutno neprekidan kosinusni red,

$$\widetilde{p}(x) = \frac{1}{2}B_2(x-[x]) = \sum_{k=0}^{+\infty} a_k \cos(2k\pi x) = \sum_{k=0}^{+\infty} a_k \frac{\mathrm{e}^{\mathrm{i}2k\pi x} + \mathrm{e}^{-\mathrm{i}2k\pi x}}{2},$$

gde su

$$a_0 = \int_0^1 p(x)\,\mathrm{d}x = 0, \quad a_k = 2\int_0^1 p(x)\cos(2k\pi x)\,\mathrm{d}x = \frac{2}{(2k\pi)^2} \quad (k \in \mathbb{N}).$$

Napomenimo da se vrednost za a_k $(k \in \mathbb{N})$ dobija jednostavno pomoću formule (3.6.13), stavljajući $f(x) = -(2k\pi)^{-2}\cos(2k\pi x)$. Dakle, imamo

$$\widetilde{p}(x) = \frac{1}{2}B_2(x-[x]) = \sum_{k=1}^{+\infty} \frac{2}{(2k\pi)^2}\cos(2k\pi x),$$

što zamenom u (3.6.14), nakon preuređenja članova, daje EULER-MACLAURIN-ovu formulu (3.6.7) za $r = 1$, sa ostatkom u obliku (3.6.8). Opšta formula za proizvoljno $r \in \mathbb{N}$ može se dobiti iz ovog specijalnog slučaja, ponavljajući parcijalnu integraciju na levoj strani u (3.6.14), kao i korišćenjem uvedene funkcije $L_j(z)$ pomoću (3.6.10) i njenih osobina (3.6.11) i (3.6.12). Na primer, nakon dve parcijalne integracije, za $r = 2$ dobijamo

$$\begin{aligned}
\int_0^n \widetilde{p}(x)f''(x)\,\mathrm{d}x &= \int_0^n [L_2(\mathrm{i}x) + L_2(-\mathrm{i}x)]f''(x)\,\mathrm{d}x \\
&= \Big[\frac{1}{\mathrm{i}}[L_3(\mathrm{i}x) - L_3(-\mathrm{i}x)]f''(x) + [L_4(\mathrm{i}x) + L_4(-\mathrm{i}x)]f'''(x)\Big]_{x=0}^{x=n} \\
&\qquad - \int_0^n [L_4(\mathrm{i}x) + L_4(-\mathrm{i}x)]f^{(4)}(x)\,\mathrm{d}x \\
&= -\frac{B_4}{4!}[f'''(n) - f'''(0)] - \sum_{k=1}^{+\infty}\int_0^n \frac{\mathrm{e}^{\mathrm{i}2\pi kx} + \mathrm{e}^{-\mathrm{i}2\pi kx}}{(2\pi k)^4} f^{(4)}(x)\,\mathrm{d}x,
\end{aligned}$$

što zamenom u (3.6.14) daje formulu (3.6.7) za $r = 2$. Inače, oblik ostatka (3.6.9) se dobija iz (3.6.8) i činjenice da je $B_{2r}(x-[x])$ dato pomoću (3.6.6). □

Pod pretpostavkom da $f \in C^{2r+1}[0,n]$, nakon pacijalne integracije u (3.6.9) i činjenice da su BERNOULLIjevi brojevi neparnog reda (≥ 3) jednaki nuli, dobijamo (videti, na primer, [32, p. 455])

$$E_r(f) = \int_0^n \frac{B_{2r+1}(x-[x])}{(2r+1)!} f^{(2r+1)}(x)\,\mathrm{d}x. \tag{3.6.15}$$

Ako $f \in C^{2r+2}[0,n]$, korišćenjem DARBOUXove[66] formule možemo dobiti (3.6.7), sa

[66] JEAN-GASTON DARBOUX (1842 – 1917), francuski matematičar sa značajnim doprinosima u matematičkoj analizi i geometriji.

$$E_r(f) = \frac{1}{(2r+2)!}\int_0^1 [B_{2r+2} - B_{2r+2}(x)] \left(\sum_{k=0}^{n-1} f^{(2r+2)}(k+x)\right) \mathrm{d}x \tag{3.6.16}$$

(videti, na primer, [72, p. 128]). Ovaj izraz za $E_r(f)$ se može, takođe, izvesti iz izraza (3.6.15), predstavljenog u obliku

$$\int_0^1 \frac{B_{2r+1}(x)}{(2r+1)!}\left(\sum_{k=0}^{n-1} f^{(2r+1)}(k+x)\right) \mathrm{d}x = \int_0^1 \frac{B'_{2r+2}(x)}{(2r+2)!}\left(\sum_{k=0}^{n-1} f^{(2r+1)}(k+x)\right) \mathrm{d}x,$$

i tada primenom parcijalne integracije, izraz na desnoj strani prethodne jednakosti postaje

$$\left[\frac{B_{2r+2}(x)}{(2r+2)!}\left(\sum_{k=0}^{n-1} f^{(2r+1)}(k+x)\right)\right]_0^1 - \int_0^1 \frac{B_{2r+2}(x)}{(2r+2)!}\left(\sum_{k=0}^{n-1} f^{(2r+2)}(k+x)\right) \mathrm{d}x.$$

Imajući u vidu da je $B_{2r+2}(1) = B_{2r+2}(0) = B_{2r+2}$, ostatak $E_r(f)$ se može predstaviti u obliku (3.6.16). Kako je $(-1)^r [B_{2r+2} - B_{2r+2}(x)] \geq 0$ na $[0,1]$ i

$$\int_0^1 [B_{2r+2} - B_{2r+2}(x)] \,\mathrm{d}t = B_{2r+2},$$

na osnovu poznate teoreme o srednjoj vrednosti integrala, postoji $\eta \in (0,1)$ takvo da je

$$\begin{aligned} E_r(f) &= \frac{B_{2r+2}}{(2r+2)!}\left(\sum_{k=0}^{n-1} f^{(2r+2)}(k+\eta)\right) \\ &= n\,\frac{B_{2r+2}}{(2r+2)!} f^{(2r+2)}(\xi), \qquad 0 < \xi < n. \end{aligned} \tag{3.6.17}$$

EULER–MACLAURINova sumaciona formula (3.6.7) se može transformisati sa segmenta $[0,n]$ na $[m,n]$, tako da imamo

$$\begin{aligned} \sum_{k=m}^{n} f(k) = \int_m^n f(x)\,\mathrm{d}x + \frac{1}{2}(f(m)+f(n)) \\ + \sum_{j=1}^{r} \frac{B_{2j}}{(2j)!}\left[f^{(2j-1)}(n) - f^{(2j-1)}(m)\right] + E_r^{m,n}(f), \end{aligned} \tag{3.6.18}$$

sa ostatakom, na primer, u obliku

$$E_r^{m,n}(f) = -\int_m^n \frac{B_{2r}(x-[x])}{(2r)!} f^{(2r)}(x)\,\mathrm{d}x. \tag{3.6.19}$$

Kada $n \to +\infty$, formula (3.6.18) se svodi na

$$\sum_{k=m}^{+\infty} f(k) = \int_m^{+\infty} f(x)\,\mathrm{d}x + \frac{1}{2}f(m) - \sum_{j=1}^{r} \frac{B_{2j}}{(2j)!} f^{(2j-1)}(m) + E_r^{m,\infty}(f). \tag{3.6.20}$$

Standardna primena EULER–MACLAURINove sumacione formule na sumiranje redova oblika $T = \sum\limits_{k=1}^{+\infty} f(k)$ se sastoji u direktnom izračunavanju sume prvih $m-1$ članova reda i primeni formule (3.6.20) na preostali deo reda, tj.

$$\begin{aligned} T = \sum_{k=1}^{+\infty} f(k) = \sum_{k=1}^{m-1} f(k) + \frac{1}{2}f(m) + \int_m^{+\infty} f(x)\,\mathrm{d}x \\ - \sum_{j=1}^{r} \frac{B_{2j}}{(2j)!} f^{(2j-1)}(m) + E_{m,r}(f), \end{aligned} \tag{3.6.21}$$

gde je $E_{m,r}$ odgovarajući ostatak koji zavisi od brojeve m i r.

U daljem tekstu razmatramo izračunavanje vrednosti RIEMANNove zeta funkcije $\zeta(s)$ $(s \in \mathbb{C})$ primenom formule (3.6.21). Za ocenu ostatka $E_{m,r}$ može se koristiti nejednakost

$$|E_{m,r}| \le \frac{m|B_{2r+2}|}{(2r+2)!} \sup_{x \ge m} |f^{(2r+2)}(x)|, \tag{3.6.22}$$

kada $f^{(2r+2)}(x)$ i $f^{(2r+4)}(x)$ ne menjaju znak za $x \in (c, +\infty)$ (videti [18, str. 51]).

Sledeći MATHEMATICA kôd izračunava vrednosti funkcije $\zeta(s)$ sa `dig` cifara, kada su u EULER-MACLAURINovoj formuli (3.6.21) uzete konkretne vrednosti za m i r.

```
f[x_] := 1/x^s;
kor[m_,j_] :=
  BernoulliB[2j]/(2j)!*(D[f[x], {x,2j-1}] /. x->m);
  EMf[m_,r_,dig_] := Module[{k,zbir,integ,x,j},
   zbir = Sum[f[k], {k, 1, m - 1}] + 1/2 f[m];
   integ = m^(1 - s)/(s - 1);
   zbir = zbir + integ - Sum[kor[m, j], {j, 1, r}];
   {N[zbir, dig], N[Abs[m*kor[m, r + 1]], 3]}
   ];
```

Radi kontrole tačnosti, pored vrednosti $\zeta(s)$, funkcija `EMf[m,r,dig]` daje i granicu za grešku (3.6.22). Program se može testirati za različite vrednosti m i r u (3.6.21).

U našem slučaju, uzeli smo $m = 20$, $r = 14$ i `dig=30`, tako da funkcija `EMf[20,14,30]` izračunava $\zeta(s)$ za $s = 2(1)10$, sa 30 cifara. Za parne vrednosti s, tačne vrednosti $\zeta(z)$ se mogu izraziti pomoću BERNOULLIjevih brojeva, saglasno formuli (3.6.4),

$$\zeta(2) = \frac{\pi^2}{6},\ \ \zeta(4) = \frac{\pi^4}{90},\ \ \zeta(6) = \frac{\pi^6}{945},\ \ \zeta(8) = \frac{\pi^8}{9450},\ \ \zeta(10) = \frac{\pi^{10}}{93555}.$$

Dobijeni su sledeći rezultati:

```
In[4]:= For[s = 2, s < 11, s++, Print[EMf[20, 14, 30]]]

{1.64493406684822643647241516665, 5.60*10^-31}

{1.20205690315959428539973816151, 4.34*10^-31}

{1.08232323371113819151600369654, 2.32*10^-31}

{1.03692775514336992633136548646, 9.55*10^-32}

{1.01734306198444913971451792979, 3.25*10^-32}

{1.00834927738192282683979754985, 9.47*10^-33}

{1.00407735619794433937868523851, 2.44*10^-33}

{1.00200839282608221441785276923, 5.63*10^-34}

{1.00099457512781808533714595890, 1.19*10^-34}
```

Funkcija se može primeniti i za izračunavanje vrednosti $\zeta(s)$ na kritičnoj pravoj. Na primer, za $s = 1/2 + 12$i, sa $m = r = 20$ i `dig=30`, dobijamo:

```
In[5]:= s = 1 / 2 + 12 I; EMf[20, 20, 30]

Out[5]= {1.01593665062277459549720673710 9 -
          0.74511247223013278206909021568 7 i, 6.88*10^-31}
```

1.3.7 Elementi teorije verižnih razlomaka

Izraz oblika

$$a_0+\cfrac{b_1}{a_1+\cfrac{b_2}{a_2+\cfrac{b_3}{a_3+\cdots}}} \tag{3.7.1}$$

naziva se verižni razlomak i predstavlja se u jednom od sledećih oblika

$$\left[a_0;\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \frac{b_3}{a_3},\ldots\right],\quad a_0+\frac{b_1|}{|a_1}+\frac{b_2|}{|a_2}+\frac{b_3|}{|a_3}+\cdots, \tag{3.7.2}$$

$$a_0+\frac{b_1}{a_1+}\frac{b_2}{a_2+}\frac{b_3}{a_3+}\ldots,$$

pri čemu su a_k, b_k, u opštem slučaju, promenljive. U specijalnom slučaju oni mogu biti brojevi (realni ili kompleksni), matrice, operatori, itd. Element a_k se naziva k-ti parcijalni imenilac, b_k k-ti parcijalni brojilac, a a_0 slobodni član. U našem izlaganju a_k i b_k su realne ili kompleksne veličine.

Ako je koeficijent $a_0=0$, tada se u ovim notacijama, obično, on izostavlja. Na primer, pišemo samo

$$\left[\frac{b_1}{a_1},\ \frac{b_2}{a_2},\ \frac{b_3}{a_3},\ \ldots\right].$$

Verižni razlomak može biti konačan. Na primer,

$$1+\cfrac{2}{2+\cfrac{5}{3}}=\left[1;\ \frac{2}{2},\ \frac{5}{3}\right]=1+\frac{2|}{|2}+\frac{5|}{|3}=1+\frac{2}{2+}\frac{5}{3}.$$

Razmotrimo sada niz konačnih verižnih razlomaka koji se dobijaju iz (3.7.1), tj. (3.7.2), uzimajući samo konačan broj članova. Tako, za $k\in\mathbb{N}$, stavimo

$$R_k=\frac{P_k}{Q_k}\equiv\left[a_0,\ \frac{b_1}{a_1},\ \frac{b_2}{a_2},\ldots,\ \frac{b_k}{a_k}\right]. \tag{3.7.3}$$

Ako egzistira vrednost verižnog razlomka (3.7.1), tada se ona definiše kao

$$R=\lim_{k\to+\infty}R_k. \tag{3.7.4}$$

Razlomak (3.7.3) nazivaćemo k-tom aproksimacijom, ili k-tim konvergentom, verižnog razlomka (3.7.1).

Ako uzmemo

$$P_0 = a+0, \quad Q_0 = 1, \quad P_{-1} = 1, \quad Q_{-1} = 0, \tag{3.7.5}$$

indukcijom se jednostavno dokazuju rekurentne relacije za P_k i Q_k:

$$\begin{aligned} P_k &= a_k P_{k-1} + b_k P_{k-2} \\ Q_k &= a_k Q_{k-1} + b_k Q_{k-2} \end{aligned} \qquad (n = 1,2,\ldots). \tag{3.7.6}$$

Primetimo da su P_k i Q_k dva rešenja diferencne jednačine

$$y_k - a_k y_{k-1} - b_k y_{k-2} = 0.$$

Na osnovu (3.7.5) i (3.7.6) jednostavno se dokazuju jednakosti

$$R_k - R_{k-1} = (-1)^{k+1} \frac{b_1 b_2 \cdots b_k}{Q_{k-1} Q_k} \qquad (k = 1,2,\ldots) \tag{3.7.7}$$

i

$$R_k - R_{k-2} = (-1)^k \frac{b_1 b_2 \cdots b_{k-1} a_k}{Q_{k-2} Q_k} \qquad (k = 2,3,\ldots). \tag{3.7.8}$$

Neka su w_k proizvoljni brojevi različiti od nule. Nad verižnim razlomkom (3.7.1) možemo izvršiti ekvivalentnu transformaciju odgovarajućim množenjem sa w_k, tako da, u notaciji (3.7.2), imamo

$$\left[a_0;\ \frac{w_1 b_1}{w_1 a_1},\ \frac{w_1 w_2 b_2}{w_2 a_2},\ \frac{w_2 w_3 b_3}{w_3 a_3},\ \ldots \right]. \tag{3.7.9}$$

Ako sa S_k i T_k označimo brojilac i imenilac u k-tom konvergentu verižnog razlomka (3.7.9) imamo

$$S_k = w_1 w_2 \cdots w_k P_k \qquad \text{i} \qquad T_k = w_1 w_2 \cdots w_k Q_k.$$

Primetimo da je pri ekvivalentnoj transformaciji verižnog razlomka, vrednost k-tog konvergenta invarijantna, tj.

$$\frac{S_k}{T_k} = \frac{P_k}{Q_k} = R_k.$$

Ekvivalentnom transformacijom, sa specijalnim izborom brojeva w_k, verižni razlomak se može uprostiti.

Neka su $b_k \neq 0$ $(k = 1,2,\ldots)$. Izaberimo konstante w_k tako da su parcijalni brojioci u (3.7.9) jednaki jedinici, tj.

$$w_1 b_1 = 1, \quad w_1 w_2 b_2 = 1, \ \ldots, \ w_{k-1} w_k b_k = 1, \ \ldots \ .$$

Tada imamo

$$w_1 = \frac{1}{b_1}, \quad w_2 = \frac{b_1}{b_2}, \quad w_3 = \frac{b_2}{b_1 b_3}, \ldots,$$

ili u opštem slučaju,

$$w_{2k} = \frac{b_1 b_3 \cdots b_{2k-1}}{b_2 b_4 \cdots b_{2k}} \quad \text{i} \quad w_{2k+1} = \frac{b_2 b_4 \cdots b_{2k}}{b_1 b_3 \cdots b_{2k+1}}.$$

Zamenom ovih vrednosti u (3.7.9) dobijamo ekvivalentni verižni razlomak kod koga su parcijalni brojioci jednaki jedinici.

Slično, ako je $a_k \neq 0$ $(k = 1,2,\ldots)$, iz (3.7.9) za $w_k = 1/a_k$ dobijamo ekvivalentan verižni razlomak

$$\left[a_0; \ \frac{c_1}{1}, \ \frac{c_2}{1}, \ \frac{c_3}{1}, \ \ldots\right], \tag{3.7.10}$$

gde su

$$c_1 = \frac{b_1}{a_1}, \quad c_2 \frac{b_2}{a_1 a_2}, \quad c_3 = \frac{b_3}{a_2 a_3}, \ \text{itd.}$$

Ako brojeve w_k izaberemo tako da je

$$w_1 a_1 = 1, \quad w_{k-1} w_k b_k + w_k a_k = 1 \quad (k = 2,3,\ldots),$$

tada se (3.7.9) svodi na oblik

$$\left[a_0; \ \frac{\alpha_1}{1}, \ \frac{\alpha_2}{1-\alpha_2}, \ \ldots \ , \frac{\alpha_k}{1-\alpha_k}, \ \ldots\right], \tag{3.7.11}$$

koji je poznat kao EULERov verižni razlomak.

Druga rekurentna relacija u (3.7.6) za razlomak (3.7.11) (umesto Q_k uzimamo ranije uvedenu oznaku T_k) postaje

$$T_1 = 1 \cdot T_0 + \alpha_1 T_{-1}, \quad T_K = (1-\alpha_k) T_{k-1} + \alpha_k T_{k-2} \quad (k \geq 2),$$

sa $T_{-1} = 0$ i $T_0 = 1$. Primetimo da se, za svako $k \geq 1$, dobija $T_k = 1$.

Verižni razlomak (3.7.1) se može svesti na EULERov oblik (3.7.11) ako i samo ako su $Q_i \neq 1$ $(i = 1,2,\ldots)$. Pri tome su

$$\alpha_1 = \frac{b_1}{Q_1}, \quad \alpha_i = Q_{i-2}Q_i b_i \quad (i = 2,3,\ldots) \tag{3.7.12}$$

i

$$R_k = a_0 + \sum_{i=1}^{k} (-1)^{i+1} \alpha_1 \alpha_2 \cdots \alpha_i \quad (k = 1,2,\ldots). \tag{3.7.13}$$

Pretpostavimo na dalje da je $Q_i \neq 0$ $(i = 1,2,\ldots)$. Na osnovu (3.7.13) imamo

$$\lim_{k\to\infty} R_k = a_0 + \sum_{i=1}^{k} (-1)^{i+1} \alpha_1 \alpha_2 \cdots \alpha_i, \tag{3.7.14}$$

tj. red na desnoj strani u poslednjoj jednakosti i verižni razlomak (3.7.1) su ekvikonvergentni. Ako je razlomak (3.7.1) konvergentan, tj. ako važi (3.7.4), na osnovu (3.7.13) i (3.7.14) zaključujemo da je

$$|R_k - R| = \left| \sum_{i=k+1}^{+\infty} (-1)^{i+1} \alpha_1 \alpha_2 \cdots \alpha_i \right|. \tag{3.7.15}$$

Na osnovu (3.7.7) i (3.7.8) važe sledeći rezultati.

Teorema 3.7.1. *Ako su a_k, $b_k > 0$, tada važe sledeće nejednakosti*

$$R_1 > R_3 > \cdots > R_{2k-1} > \cdots, \quad R_0 < R_2 < \cdots < R_{2k} < \cdots,$$

$$R_{2m-1} > R_{2k} \qquad (za\ svako\ m\ i\ k).$$

Teorema 3.7.2. *Ako su a_k i b_k pozitivni i takvi da je $b_k \leq a_k$ i $a_k \geq \varepsilon$ $(k = 1,2,\ldots)$, gde je ε neka konstanta, tada je verižni razlomak* (3.7.1) *konvergentan.*

Ako su $a_k b_k > 0$, na osnovu (3.7.12), (3.7.5), (3.7.6) zaključujemo da je

$$\alpha_1 \alpha_2 \cdots \alpha_i = \frac{b_1 b_2 \cdots b_i}{Q_{i-1} Q_i} > 0.$$

Tada iz (3.7.15), na osnovu LEIBNIZovog kriterijuma za alternativne redove, dobijamo sledeću ocenu

$$|R_k - R| \leq \frac{b_1 b_2 \cdots b_{k+1}}{Q_k Q_{k+1}}.$$

Za slučaj kada su a_k i b_k kompleksni, važi sledeća teorema.

Teorema 3.7.3. *Ako je* $|a_k| - |b_k| \geq 1$ $(k = 1,2,\ldots)$*, verižni razlomak* (3.7.1) *konvergira.*

Kao što smo videli, ako je $a_k \neq 0$ $(k = 1,2,\ldots)$, ekvivalentni oblik verižnog razlomka (3.7.1) može se dati u obliku (3.7.10). Ne umanjujući opštost, sa stanovišta konvergencije, umesto (3.7.10) može se posmatrati verižni razlomak

$$\left[\frac{1}{1}, \frac{c_2}{1}, \frac{c_3}{1}, \ldots\right]. \tag{3.7.16}$$

Definicija 3.7.1. Za verižni razlomak (3.7.16) kažemo da zadovoljava fundamentalne nejednakosti ako postoji niz nenegativnih brojeva $\{r_i\}_{i\in\mathbb{N}}$ takav da za parcijalne brojioce, za svako $i \in \mathbb{N}$, važe nejednakosti

$$r_{i-2}r_i|c_i| + |c_{i+1}| \leq r_i|1 + c_i + c_{i+1}|,$$

pri čemu su $c_1 = 0$, $r_0 = r_{-1} = 0$.

Ako verižni razlomak (3.7.16) zadovoljava fundamentalne nejednakosti, može se dokazati (videti, na primer, [61]) da je tada $Q_k \neq 0$ $(k = 1,2,\ldots)$ i da za brojeve

$$\alpha_i = \frac{Q_{i-2}}{Q_i}c_i$$

važe nejednakosti

$$|\alpha_i| \leq r_{i-1} \qquad (i = 2,3,\ldots). \tag{3.7.17}$$

Dakle, u tom slučaju (3.7.16) može da se ekvivalentno transformiše na EULERov razlomak, kod koga je $a_0 = 0$ i $\alpha_1 = 1$. Ako je red $\sum_{i=2}^{+\infty} r_1 r_2 \cdots r_{i-1}$ konvergentan, na osnovu (3.7.13) i (3.7.17), zaključujemo da je verižni razlomak (3.7.16) konvergentan.

Teorema 3.7.4. *Neka su* c_i $(i = 2,3,\ldots)$ *funkcije jedne promenljive* (*ili više promenljivih*)*, definisane u nekoj oblasti D, u kojoj je*

$$|c_i| \leq \frac{1}{4} \qquad (i = 2,3,\ldots).$$

Tada važe tvrđenja:

a) *verižni razlomak* (3.7.16) *ravnomerno konvergira u D;*

b) *vrednosti svih konvergenata verižnog razlomka* (3.7.16), *kao i njegova vrednost, pripadaju oblasti kruga*

$$K = \left\{ z \;\middle|\; \left| z - \frac{4}{3} \right| \le \frac{2}{3} \right\};$$

c) *konstanta* $1/4$ *u oblasti kruga K je najbolja mogućnost, tj. konstanta se ne može povećati, a krug K se ne može smanjiti.*

1.3.8 Razvoj racionalne funkcije u verižni razlomak

Od interesa je proučiti razlaganje racionalne funkcije u verižni razlomak.
Neka je

$$f(x) = \frac{c_{10} + c_{11}x + c_{12}x^2 + \cdots}{c_{00} + c_{01}x + c_{02}x^2 + \cdots} \qquad (c_{10} \neq 0).$$

Tada je

$$f(x) = \frac{1}{\dfrac{c_{00}}{c_{10}} + \dfrac{c_{00} + c_{01}x + c_{02}x^2 + \cdots}{c_{10} + c_{11}x + c_{12}x^2 + \cdots} - \dfrac{c_{00}}{c_{10}}} = \frac{c_{10}}{c_{00} + x f_1(x)},$$

gde su

$$f_1(x) = \frac{c_{20} + c_{21}x + c_{22}x^2 + \cdots}{c_{10} + c_{11}x + c_{12}x^2 + \cdots}, \quad c_{2j} = c_{10}c_{0,j+1} - c_{00}c_{1,j+1} \quad (j = 0, 1, \ldots).$$

Nastavljajući ovaj proces dobija se razvoj u verižni razlomak

$$f(x) = \left[\frac{c_{10}}{c_{00}},\ \frac{c_{20}x}{c_{10}},\ \frac{c_{30}x}{c_{20}}, \ldots \right],$$

gde su

$$c_{ij} = -\begin{vmatrix} c_{i-2,0} & c_{i-2,j+1} \\ c_{i-1,0} & c_{i-1,j+1} \end{vmatrix} \qquad (i = 2, 3, \ldots).$$

Primer 3.8.1. Za funkciju $x \mapsto f(x) = \dfrac{1-x}{1-5x+6x^2}$ imamo

$$f(x) = \left[\frac{1}{1},\ \frac{-4x}{1},\ \frac{-2x}{-4},\ \frac{-12x}{-2} \right] = \frac{1}{1 - \dfrac{4x}{1 - \dfrac{2x}{-4+6x}}}. \quad \triangle$$

Obrnuto, verižni razlomak

$$f(x)=\cfrac{b_1}{x+\cfrac{b_2}{x+\cfrac{b_3}{x+\cdots}}}=\left[\frac{b_1}{x},\ \frac{b_2}{x},\ \frac{b_3}{x},\ \ldots,\ \frac{b_n}{x}\right]$$

se može predstaviti u obliku racionalne funkcije (videti: SLAVIĆ[67] [62])

$$f(x)=\left(\frac{c_{2n}x+c_{4n}x^3+c_{6n}x^5+\cdots}{c_{1n}+c_{3n}x^2+c_{5n}x^4+\cdots}\right)^{(-1)^n},$$

gde se koeficijenti c_{kn} $(k=1,\ldots,n+1)$ mogu izračunati po sledećem algoritmu:

$$c_{11}=b_n;$$

$$c_{i+1,i}=1 \qquad (i=1,\ldots,n);$$

$$c_{ki}=\begin{cases} c_{k,i-1}b_{n-i+1} & (k+i \text{ parno}),\\ c_{k,i-1}+c_{k-1,i-1} & (k+i \text{ neparno},\ k>1),\\ c_{1,i-1} & (i \text{ parno}),\end{cases}$$

pri čemu k uzima vrednost $k=i,i-1,\ldots,1$ za svako $i=2,\ldots,n$.

Ilustracije radi koeficijenti c_{ki} su dati u tabeli 3.8.1.

Tabela 3.8.1.

i	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
1	b_n	1	0	0	0
2	b_n	b_{n-1}	1	0	0
3	b_nb_{n-2}	b_n+b_{n-1}	b_{n-2}	1	0
4	b_nb_{n-2}	$(b_n+b_{n-1})b_{n-3}$	$b_n+b_{n-1}+b_{n-2}$	b_{n-3}	1

Algoritam se može uprostiti zamenom matrice $[c_{ki}]_{(n+1)\times n}$ jednodimenzionalnim nizom $c(k)=c_{ki}$ $(k=1,\ldots,n+1)$, čime se vrši ušteda memorijskog prostora. Modifikovani algoritam je:

[67] DUŠAN V. SLAVIĆ (1942 – 1990), srpski matematičar.

$$c(k) := 1 \qquad (k = 1, \ldots, n+1);$$

$$c(k) := \begin{cases} c(k)b_{n-i+1} & (k+i \text{ parno}), \\ c(k)+c(k-1) & (k+i \text{ neparno},\ k>1), \end{cases}$$

pri čemu $k = i, i-1, \ldots, 1$ za svako $i = 1, \ldots, n$.

Primer 3.8.2. Neka je

$$f(x) = \left[\frac{1}{x}, \frac{2}{x}, \frac{3}{x}, \frac{4}{x}\right],$$

tj. $b_k = k$ $(k = 1, \ldots, 4)$. Kako je $n = 4$, iz četvrte vrste tabele 3.8.1 $(i = 4)$ nalazimo

$$c_{14} = b_4 b_2 = 8, \quad c_{24} = (b_4 + b_3)b_1 = 7,$$

$$c_{34} = b_4 + b_3 + b_2 = 9, \quad c_{44} = b_1 = 1, \quad c_{54} = 1,$$

tj.

$$f(x) = \frac{7x + x^3}{8 + 9x^2 + x^4}. \quad \triangle$$

1.3.9 Algoritmi za izračunavanje verižnih razlomaka

U ovom odeljku razmotrićemo nekoliko algoritama za izračunavanje n-tog konvergenta verižnog razlomka, datog pomoću (3.7.3). Pretpostavimo da je $a_0 = 0$, tj. da R_n ima oblik

$$R_n = \frac{P_n}{Q_n} = \left[\frac{b_1}{a_1}, \frac{b_2}{a_2}, \ldots, \frac{b_n}{a_n}\right]. \tag{3.9.1}$$

ALGORITAM A1. Za (3.9.1) definišemo rekurentnu relaciju unazad pomoću

$$y_{n,k} = \frac{b_k}{a_k + y_{n,k+1}} \qquad (k = n, n-1, \ldots, 1), \tag{3.9.2}$$

pri čemu je $y_{n,n+1} = 0$. Tada je $R_n y_{n,1}$.

ALGORITAM A2. Algoritam se zasniva na primeni rekurentnih formula

$$c_k = \frac{b_k}{f_{k-1}}, \quad f_k = a_k + c_k, \quad q_k = -\frac{c_k q_{k-1}}{f_k}, \quad R_k = R_{k-1} + q_k \quad (k = 2, \ldots, n),$$

startujući sa $f_1 = a_1$, $q_1 = R_1 = b_1/a_1$.

ALGORITAM A3. Ovaj algoritam se zasniva na primeni rekurentnih relacija (3.7.6), tj.

$$\begin{aligned} P_k &= a_k P_{k-1} + b_k P_{k-2} \\ Q_k &= a_k Q_{k-1} + b_k Q_{k-2} \end{aligned} \qquad (n = 1,2,\ldots,n). \tag{3.9.3}$$

sa startnim vrednostima $P_0 = 0,\ P_{-1} = 1;\ Q_0 = 1,\ Q_{-1} = 0$. Tada je $R_n = P_n/Q_n$. Ovo je poznati EULERov algoritam.

ALGORITAM A4. Startujući sa $y_1 = 1$ i $y_2 = a_n$ određujemo

$$y_k = a_{n+2-k} y_{k-1} + b_{n+3-k} y_{k-2} \qquad (k = 3,\ldots,n). \tag{3.9.4}$$

Tada je

$$R_n = \frac{b_1 y_n}{a_1 y_1 + b_2 y_{n-1}}. \tag{3.9.5}$$

Poslednji algoritam je matematički ekvivalentan EULERovom algoritmu A3, ali su oni u numeričkom smislu različiti ([47]).

Primetimo da EULERov algoritam zahteva dvostruku primenu rekurentne relacije (3.9.3) (za P_k i Q_k). Ta relacija je oblika

$$S_k = a_k S_{k-1} + b_k S_{k-2} \qquad (k = 1,2,\ldots,n),$$

što se može predstaviti u matričnom obliku kao trougaoni sistem linearnih jednačina $A\boldsymbol{s} = \boldsymbol{f}$, tj.

$$\begin{bmatrix} 1 & & & & & \mathbf{0} \\ -a_2 & 1 & & & & \\ -b_3 & -a_3 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & a_{n-1} & 1 & \\ \mathbf{0} & & & -b_n & -a_n & 1 \end{bmatrix} \cdot \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ \vdots \\ S_{n-1} \\ S_n \end{bmatrix} = \begin{bmatrix} a_1 S_0 + b_1 S_{-1} \\ b_2 S_0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}.$$

Kako je $\boldsymbol{s} = A^{-1}\boldsymbol{f}$, a nama je potrebna samo poslednja koordinata u vektoru $\boldsymbol{s}$, tj. S_n, to je za njeno određivanje dovoljno poznavati samo poslednju vrstu u matrici A^{-1}. Označimo elemente ove vrste redom sa $y_n, y_{n-1}, \ldots, y_1$. Iz jednakosti $A^{-1}A = I$ dobijamo (3.9.4). Kako su u vektoru $\boldsymbol{f}$ samo prve dve koordinate različite od nule imamo

$$S_n = y_n\,(a_1 S_0 + b_1 S_{-1}) + y_{n-1} b_2 S_0,$$

tj.

$$S_n = b_1 y_n S_{-1} + (a_1 y_n + b_2 y_{n-1})\, S_0.$$

Da bismo dobili P_n, u poslednjoj jednakosti treba staviti $S_{-1} = P_{-1} = 1$ i $S_0 = P_0 = 0$, dok za Q_n treba staviti $S_{-1} = Q_{-1} = 0$ i $S_0 = Q_0 = 1$. Na taj način dobijamo (3.9.5).

Tabela 3.9.1.

ALGORITAM	Broj operacija			Ukupan broj operacija
	sabiranja	množenja	deljenja	
A1	$n-1$	0	n	$2n-1$
A2	$2n-2$	$n-1$	$2n-1$	$5n-4$
A3	$2n-3$	$4n-6$	1	$6n-8$
A4	$n-1$	$2n-2$	1	$3n-2$

U tabeli 3.9.1 je dat pregled broja operacija za napred navedena četiri algoritma, iz koje vidimo da algoritam $A1$, dat pomoću (3.9.2), zahteva najmanji broj operacija. Takođe, njegova programska realizacija je jednostavna. Međutim, u nekim slučajevima algoritam $A4$ može biti efikasniji od $A1$, ako je „cena deljenja“ znatno veća od „cene množenja“. Za cenu operacije se može uzeti, na primer, procesorsko vreme potrebno za realizaciju odgovarajuće operacije.

Za izračunavanje vrednosti nekih funkcija umesto TAYLORovog razvoja, može se koristiti razvoj u verižni razlomak. Zato ćemo sada navesti, bez dokaza, razvoje sa funkcije $x \mapsto \mathrm{e}^x$ i $x \mapsto \tan x$.

EULER je dokazao razvoj

$$\mathrm{e}^x = \left[\frac{1}{1}, \frac{-2x}{2+x}, \frac{x^2}{6}, \frac{x^2}{10}, \ldots, \frac{x^2}{4n+2}, \ldots\right],$$

koji konvergira za svako realno ili kompleksno x, dok je razvoj

$$\tan x = \left[\frac{x}{1}, \frac{-x^2}{3}, \frac{-x^2}{5}, \ldots, \frac{-x^2}{2n+1}, \ldots\right]$$

dokazao LAMBERT[68]. Poslednji razvoj konvergira za svako x, za koje je funkcija $x \mapsto \tan x$ neprekidna.

[68] JOHANN HEINRICH LAMBERT (1728 – 1777), švajcarski matematičar, fizičar, astronom i filozof.

Napomena 3.9.1. Za izračunavanje vrednosti $\tan x$ pogodno je uvesti smenu $\tan x = x/y$. Tada se najpre izračunava vrednost

$$y = \left[1;\ \frac{-x^2}{3},\ \frac{-x^2}{5}, \dots\right],$$

a zatim $\tan x = x/y$.

Primer 3.9.1. Na osnovu EULERovog razvoja za e^x naći ćemo prvih pet aproksimacija. Ovde je $b_1 = a_1 = 1$, $b_2 = -2x$, $a_2 = 2+x$, $b_3 = b_4 = b_5 = x^2$, $a_3 = 6$, $a_4 = 10$, $a_5 = 14$. Primenom, na primer, EULERovog algoritma dobijamo

$$\begin{aligned}
R_1 &= R_1(x) = \frac{1}{1},\\
R_2 &= R_2(x) = \frac{2+x}{2-x},\\
R_3 &= R_3(x) = \frac{12+6x+x^2}{12-6x+x^2},\\
R_4 &= R_4(x) = \frac{120+60x+12x^2+x^3}{120-60x+12x^2-x^3},\\
R_5 &= R_5(x) = \frac{1680+840x+180x^2+20x^3+x^4}{1680-840x+180x^2-20x^3+x^4}.
\end{aligned}$$

Primetimo da dobijene aproksimacije $R_k(x)$ zadovoljavaju uslov

$$R_k(x) = R_k(-x) = 1.$$

Može se pokazati da racionalna funkcija $R(x)$ ispunjava ovaj uslov ako i samo ako se ona može predstaviti u obliku

$$R(x) = 1 - \frac{2x}{T(x^2)+x},$$

gde je $T(x^2)$ racionalna funkcija po x^2. Za dokaz ovog tvrđenja treba, najpre, dokazati da takva racionalna funkcija mora imati reprezentaciju u obliku $R(x) = P(x)/P(-x)$, gde je P algebarski polinom. Tada je

$$T(x^2) = x\frac{P(-x)+P(x)}{P(-x)-P(x)}.$$

Na primer, za funkciju $R_4(x)$ imamo $P(x) = 120+60x+12x^2+x^3$. Odgovarajuća funkcija $T(x^2)$ je data sa

$$T(x^2) = -12\frac{x^2+10}{x^2+60}.$$

Dakle, imamo

$$R_4(x) = 1 - \frac{2x}{x - 12\dfrac{x^2+10}{x^2+60}}.$$

Ako stavimo, na primer, $x = 0.5$ imamo

$$\mathrm{e}^{0.5} \cong R_4(0.5) = 1.6487214,$$

što predstavlja tačnu vrednost na šest decimala. $\triangle$

1.3.10 Asimptotski razvoji

Polinomski razvoji i razvoji u verižne razlomke obično se koriste za izračunavanje vrednosti (aproksimaciju) funkcija na konačnim intervalima realne ose. Izračunavanje vrednosti funkcija za velike vrednosti argumenta moguće je kod nekih funkcija svesti na prethodni slučaj i to transformacijom argumenta, ili pak korišćenjem izvesnih svojstava funkcije, koja dozvoljavaju korišćenje aproksimacija sa konačnog segmenta. Na primer, izračunavanje gama funkcije $\Gamma(x)$ za $x \in (0,+\infty)$ moguće je svesti, korišćenjem funkcionalne relacije $\Gamma(x+1) = x\Gamma(x)$, na izračunavanje vrednosti ove funkcije na konačnom segmentu $[1,2]$, na primer, pomoću polinomskog razvoja

$$\begin{aligned}\Gamma(z) \cong\ & 1 - 0.57710166t + 0.98585399t^2 - 0.87642182t^3 \\ & + 0.83282120t^4 - 0.56847290t^5 + 0.25482049t^6 - 0.05149930t^7,\end{aligned}$$

gde je $t = z - 1$ $(0 \le t \le 1)$[69]. Međutim, kada ovi pristupi nisu mogućni, bilo bi dobro imati izvesne razvoje funkcija, tzv. asimptotske razvoje, koji bi važili za velike vrednosti argumenta.

Pretpostavimo da su funkcije f i g definisane na $(0,+\infty)$ i da je $g(x) \neq 0$ za $x > 0$.

Definicija 3.10.1. Razvoj

$$g(x)\sum_{k=0}^{+\infty}\frac{a_k}{x^k} \tag{3.10.1}$$

[69] Neke osobine gama funkcije i jedan efikasan algoritam za izračunavanje vrednosti $\Gamma(z)$ sa proizvoljnom tačnošću, kada $z \in \mathbb{C}$, biće dat u narednom odeljku.

naziva se asimptotski razvoj za $f(x)$, ako za svako $n = 0, 1, 2, \ldots$, važi

$$\left(\frac{f(x)}{g(x)} - \sum_{k=0}^{n} \frac{a_k}{x^k} \right) x^n \to 0, \quad \text{kada } x \to \infty. \tag{3.10.2}$$

Ovu činjenicu označavamo sa

$$f(x) \sim g(x) \sum_{k=0}^{+\infty} \frac{a_k}{x^k}.$$

Asimtotski razvoj (3.10.1) može biti, u opštem slučaju, divergentan red.

Napomena 3.10.1. Ako je

$$f_1(x) - f_2(x) \sim g(x) \sum_{k=0}^{+\infty} a_k x^{-k},$$

tada ćemo

$$\Phi(x) = f_2(x) + g(x) \sum_{k=0}^{+\infty} a_k x^{-k}$$

tretirati kao asimptotski razvoj funkcije f_1, tj. $f_1(x) \sim \Phi(x)$. Na primer, za $\log \Gamma(x)$ imamo

$$\begin{aligned} \log \Gamma(x) \sim{}& \left(x - \frac{1}{2} \right) \log x - x + \frac{1}{2} \log 2\pi \\ &+ \frac{1}{12x} \left(1 - \frac{1}{30x^2} + \frac{1}{105x^4} - \frac{1}{140x^6} + \frac{1}{99x^8} - \cdots \right). \end{aligned}$$

Primer 3.10.1. Odredimo asimptotski razvoj za funkciju $E_1(x) = \displaystyle\int_x^{+\infty} \frac{\mathrm{e}^{-t}}{t}\, \mathrm{d}t$.

Sukcesivnom primenom parcijalne integracije nalazimo

$$\begin{aligned} E_1(x) &= \frac{\mathrm{e}^{-x}}{x} - \int_x^{+\infty} \frac{\mathrm{e}^{-t}}{t^2}\, \mathrm{d}t \\ &= \frac{\mathrm{e}^{-x}}{x} - \frac{\mathrm{e}^{-x}}{x^2} + 2 \int_x^{+\infty} \frac{\mathrm{e}^{-t}}{t^3}\, \mathrm{d}t \\ &\vdots \\ &= \mathrm{e}^{-x} \left(\frac{1}{x} - \frac{1}{x^2} + \frac{2!}{x^3} - \frac{3!}{x^4} + \cdots + (1)^{n-1} \frac{(n-1)!}{x^n} \right) + R_n, \end{aligned}$$

gde je

$$R_n = (-1)^n n! \int_x^{+\infty} \frac{\mathrm{e}^{-t}}{t^{n+1}}\,\mathrm{d}t.$$

Kako je za $x > 0$

$$|R_n| = n! \int_x^{+\infty} \frac{\mathrm{e}^{-t}}{t^{n+1}}\,\mathrm{d}t = n!\frac{\mathrm{e}^{-x}}{x^{n+1}} - (n+1)! \int_x^{+\infty} \frac{\mathrm{e}^{-t}}{t^{n+2}}\,\mathrm{d}t < n!\frac{\mathrm{e}^{-x}}{x^{n+1}},$$

imamo

$$\left| \left(\mathrm{e}^x E_1(x) - \sum_{k=1}^{n} (-1)^{k-1} \frac{(k-1)!}{x^k} \right) x^n \right| < \frac{n!}{x},$$

odakle zaključujemo da je uslov (3.10.2) ispunjen. Prema tome, dobili smo asimptotski razvoj

$$E_1(x) \sim \mathrm{e}^{-x} \sum_{k=1}^{+\infty} (-1)^{k-1} \frac{(k-1)!}{x^k}. \tag{3.10.3}$$

Primetimo da je ovaj red divergentan za svako x. $\triangle$

Primer 3.10.2. Slično se može naći asimptotski razvoj i za komplementarnu funkciju greške

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t. \tag{3.10.4}$$

Tako, primenom parcijalne integracije, nalazimo

$$2\int_x^{\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t = \int_x^{\infty} \left(-\frac{1}{t}\right) \left(-2t\mathrm{e}^{-t^2}\right) \mathrm{d}t = \frac{1}{x}\mathrm{e}^{-x^2} - \int_x^{\infty} \frac{1}{t^2}\,\mathrm{e}^{-t^2}\,\mathrm{d}t.$$

Nastavljajući ovaj proces dobijamo

$$\frac{2}{\sqrt{\pi}} \int_x^{\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t = \frac{\mathrm{e}^{-x^2}}{\sqrt{\pi}} \left(\frac{1}{x} - \frac{1}{2x^3} + \frac{1\cdot 3}{2^2 x^5} - \cdots + (-1)^n \frac{1\cdot 3\cdots(2n-1)}{2^n x^{2n+1}} \right) + R_n,$$

gde je

$$R_n = (-1)^{n+1} \frac{(2n+1)!!}{2^n\sqrt{\pi}} \int_x^{\infty} \frac{1}{t^{2n+2}} \mathrm{e}^{-t^2}\,\mathrm{d}t.$$

Primetimo da je

$$R_n = (-1)^{n+1}\frac{(2n+1)!!}{2^n\sqrt{\pi}}\left(\frac{\mathrm{e}^{-x^2}}{2x^{2n+3}} - \frac{2n+3}{2}\int_x^\infty \frac{\mathrm{e}^{-t^2}}{t^{2n+4}}\,\mathrm{d}t\right),$$

tj.

$$R_n = (-1)^{n+1}\frac{(2n+1)!!}{2^{n+1}\sqrt{\pi}}\cdot\frac{\mathrm{e}^{-x^2}}{x^{2n+3}} + R_{n+1},$$

i da su R_n i R_{n+1} suprotnog znaka ($R_nR_{n+1} < 0$). Tada se može zaključiti da je

$$|R_n| < \frac{(2n+1)!!}{2^{n+1}\sqrt{\pi}}\cdot\frac{\mathrm{e}^{-x^2}}{x^{2n+3}},$$

što pokazuje da apsolutna greška odsecanja nije veća od modula prvog odbačenog člana. Takođe, korišćenjem ove nejednakosti zaključujemo da je

$$\lim_{x\to+\infty}\left(\sqrt{\pi}\,\mathrm{e}^{x^2}R_n x^{2n+1}\right) = 0,$$

za svako $n = 0,1,\ldots$, što znači da je uslov (3.10.2) ispunjen. Tako smo dobili asimptotski razvoj za komplementarnu funkciju greške

$$\operatorname{erfc}(x) \sim \frac{\mathrm{e}^{-x^2}}{x\sqrt{\pi}}\left(1 + \sum_{k=1}^{+\infty}(-1)^k\frac{(2k-1)!!}{(2x^2)^k}\right).$$

△

Napomena 3.10.2. U radu [13] izvedena je interesantna formula za komplementarnu funkciju greške (3.10.4) u obliku

$$\operatorname{erfc}(x) = \frac{\mathrm{e}^{-x^2}\varepsilon x}{\pi}\left(\frac{1}{x^2} + 2\sum_{k=1}^{+\infty}\frac{\mathrm{e}^{-k^2\varepsilon^2}}{k^2\varepsilon^2+x^2}\right) + \frac{2}{1-\mathrm{e}^{2\pi x/\varepsilon}} + O(\mathrm{e}^{-\pi^2/\varepsilon^2}),$$

gde je ε mali parametar izabran tako da je $0 < \varepsilon < \pi/x$.

I pored toga što su asimptotski razvoji divergentni, oni mogu biti korišćeni za izračunavanje vrednosti funkcija za velike vrednosti argumenta x. Naime, odsecanjem asimptotskg razvoja, tj. uzimanjem samo konačnog broja članova, dobijamo funkciju

$$\Phi_n(x) = g(x)S_n(x),$$

gde je

$$S_n(x) = \sum_{k=0}^{n} \frac{a_k}{x^k}, \tag{3.10.5}$$

koja sa proizvoljnom tačnošću može da aproksimira $f(x)$ za dovoljno velike vrednosti argumenta x. Dakle, kada x raste, tačnost u aproksimaciji

$$f(x) \simeq \Phi_n(x)$$

se povečava. Geometrijski Φ_n predstavlja asimptotu od f sa dodirom najmanje reda n u beskonačnosti. Međutim, kada je x fiksno postoji jedna bitna razlika asimptotskih razvoja u odnosu na konvergentne, na primer, stepene redove. Kod stepenih redova za datu vrednost argumenta x može se povećati tačnost uzimanjem većeg broja članova razvoja. Ova osobina ne važi kod divergentnih asimptotskih razvoja, tj. uzimanje većeg broja članova (veće n u $S_n(x)$) ne dovodi uvek do povećanja tačnosti za dato x (> 0).

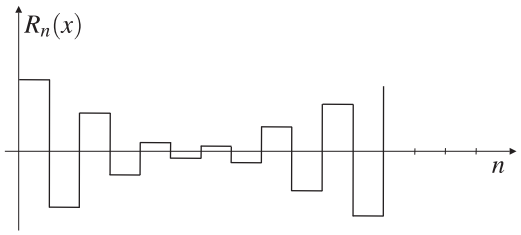
U vezi sa (3.10.5) neka je

$$R_n(x) = S(x) - S_n(x) \qquad (S(x) = f(x)/g(x)).$$

Pretpostavimo da je asimptotski razvoj alternativan tako da su $R_n(x)$ i $R_{n+1}(x)$ suprotnog znaka. S obzirom na to da je razvoj divergentan imaćemo sistuaciju kao na sl. 3.10.1. Dakle, za fiksno $x > 0$ i svako $n = 0, 1, 2, \ldots$, imamo da je $S_n(x) + R_n(x) = S(x)$, pri čemu i $S_n(x)$ i $R_n(x)$ divergiraju kada $n \to +\infty$. Međutim, kao i kod konvergentnih alternativnih redova, imaćemo da je

$$\left| S_n(x) - \frac{1}{2}\big(S_n(x) + S_{n+1}(x)\big) \right| \leq \frac{1}{2}\,|u_{n+1}|, \tag{3.10.6}$$

gde je $u_{n+1} = S_{n+1}(x) - S_n(x)$. U našem slučaju je $u_{n+1} = a_{n+1}/x^{n+1}$.



Slika 3.10.1. Greška kod asimptotskog razvoja

Na osnovu (3.10.6) vidimo da možemo uzeti

$$S_n(x) \cong \frac{1}{2}\big(S_n(x) + S_{n+1}(x)\big) = S_n(x) + \frac{1}{2}\cdot\frac{a_{n+1}}{x^{n+1}} \tag{3.10.7}$$

sa apsolutnom greškom koja ne prelazi $|a_{n+1}|/(2x^{n+1})$.

Primetimo da se u (3.10.7) na desnoj strani pojavljuje CESÀROova transformacija od $S_n(x)$. Jasno je da kod konvergentnih redova sa povećavanjem n u (3.10.7) imamo bolju aproksimaciju (veću tačnost). Međutim, kod asimptotskih razvoja postoji optimalni broj n za dato x, koji se određuje iz uslova da granica apsolutne greške u (3.10.7) bude najmanja. Dakle, n određujemo iz uslova

$$\min_k \frac{|a_{k+1}|}{x^{k+1}} = \frac{|a_{n+1}|}{x^{n+1}}. \tag{3.10.8}$$

Primer 3.10.3. Korišćenjem asimptotskog razvoja (3.10.3) odredićemo vrednosti za $E_1(5)$ i $E_1(9.5)$. U ovom slučaju imamo $|a_{k+1}| = k!$. Da bismo odredili optimalan broj članova koje treba uzeti u razvoju (3.10.3), posmatrajmo niz $\{q_k\}$, gde je $q_k = |a_k|/x^k = (k-1)!/x^k$.

Kako je $q_{k+1}/q_k = k/x$, zaključujemo da je $q_{k+1} < q_k$ za $k < x$, dok je za $k > x$, $q_{k+1} > q_k$. Prema tome, minimum u (3.10.8) nastupa za dve vrednosti k, tj. za $k = n_1 = x-1$ i za $k = n_2 = x$.

Za $x = 5$ uzimamo $n = n_1 = 4$. Na osnovu (3.10.7) imamo

$$\mathrm{e}^5 E_1(5) \cong S_4(5) + \frac{1}{2}\cdot\frac{4!}{5^5} = \frac{1}{5} - \frac{1}{5^2} + \frac{2}{5^3} - \frac{6}{5^4} + \frac{12}{5^5} \cong 0.17024,$$

sa apsolutnom greškom manjom od $12/5^5 = 3.84\cdot 10^{-3}$. Primetimo da se isti rezultat dobija i za $n = n_2 = 5$:

$$\mathrm{e}^5 E_1(5) \cong S_5(5) - \frac{1}{2}\cdot\frac{5!}{5^6} \cong 0.17024.$$

Sada imamo

$$E_1(5) \cong 0.001147,$$

pri čemu apsolutna greška nije veća od $2.6\cdot 10^{-5}$. Tačna vrednost je $E_1(5) = 0.00114829\ldots$.

Za $x = 9.5$ imamo $n = [9.5] = 9$, pa je

$$\mathrm{e}^{9.5} E_1(9.5) \cong S_9(9.5) - \frac{1}{2}\cdot\frac{9!}{9.5^{10}} \cong 0.0959866,$$

sa apsolutnom greškom manjom od $3\cdot 10^{-5}$. Najzad, imamo

$$E_1(9.5) \cong 7.184773\cdot 10^{-6},$$

pri čemu je apsolutna greška manja od $2.3\cdot 10^{-9}$. Tačna vrednost je, inače, $E_1(9.5) = 7.1847746\ldots\cdot 10^{-6}$. $\triangle$

1.3.11 Izračunavanje vrednosti gama funkcije

Jedna od najznačajnijih analitičkih funkcija je gama funkcija $\Gamma(z)$ koja predstavlja proširenje faktorijelne funkcije sa skupa $\mathbb{N}_0$ na kompleksnu ravan $\mathbb{C}$, tako da je $\Gamma(n+1) = n!$. Integralnu reprezentaciju

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} \mathrm{e}^{-t}\, \mathrm{d}t, \tag{3.11.1}$$

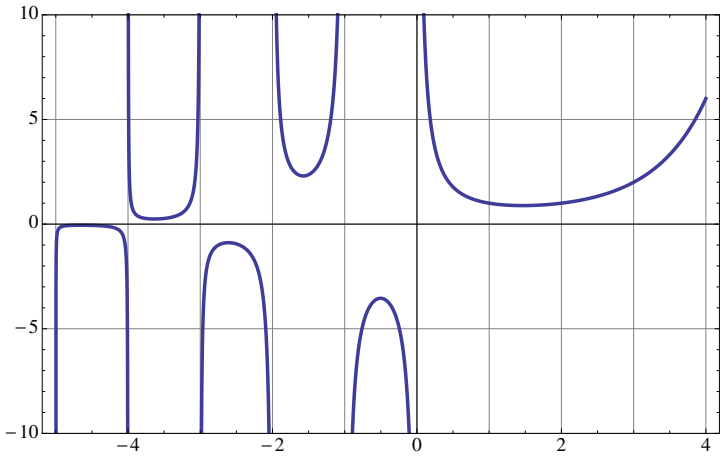
koja važi za svako z u desnoj poluravni kompleksne ravni, tj. za svako $\mathrm{Re}\, z > 0$, uveo je EULER, koji je, takođe, dokazao i osnovnu funkcionalnu jednačinu

$$\Gamma(z+1) = z\Gamma(z),$$

koja se koristi za analitičko produženje gama funkcije na čitavu kompleksnu ravan. Na primer,

$$\Gamma(z) = \frac{\Gamma(z+1)}{z} = \frac{\Gamma(z+2)}{z(z+1)} = \frac{\Gamma(z+3)}{z(z+1)(z+2)} = \cdots .$$

Očigledno, ovako produžena gama funkcija ima samo proste polove u tačakama $z = 0, -1, -2, \ldots$. Za realno z, grafik gama funkcije je prikazan na slici 3.11.1. Umesto $\Gamma(z+1)$ koristi se i prirodnija (originalna) oznaka $z!$. Inače, oznaku $\Gamma(z)$



Slika 3.11.1. Grafik funkcije $\Gamma(z)$ za $-5 < z < 4$

uveo je LEGENDRE[70] On je, takođe, dokazao tzv. *duplikacionu formulu*

$$\Gamma(2z) = \frac{2^{2z-1}}{\sqrt{\pi}}\Gamma(z)\Gamma\left(z+\frac{1}{2}\right),$$

koja je specijalan slučaj opšte *multiplikacione formule*

$$\Gamma(nz) = \frac{n^{nz-1/2}}{(2\pi)^{(n-1)/2}}\Gamma(z)\Gamma\left(z+\frac{1}{n}\right)\Gamma\left(z+\frac{2}{n}\right)\cdots\Gamma\left(z+\frac{n-1}{n}\right).$$

Gama funkcija se može definisati i WEIERSTRASSovim[71] proizvodom koji važi za svako $z \in \mathbb{C}$,

$$\frac{1}{z!} = \frac{1}{\Gamma(z+1)} = \mathrm{e}^{\gamma z}\prod_{k=1}^{+\infty}\left[\left(1+\frac{z}{k}\right)\mathrm{e}^{-z/k}\right], \tag{3.11.2}$$

gde je γ EULERova konstanta, koja se u programskom paketu MATHEMATICA može izračunati sa proizvoljnom preciznošću. Na primer, `N[EulerGamma, 50]` daje $\gamma \approx 0.57721566490153286060651209008240243104215933593992$. Za definiciju EULERove konstante videti odeljak 1.3.2 (formula (3.2.6)).

Imajući na umu tzv. *refleksionu formulu*

$$(-z)!z! = \Gamma(1-z)\Gamma(1+z) = \frac{\pi z}{\sin \pi z}, \tag{3.11.3}$$

za izračunavanje $z!$ dovoljno je znati postupak, na primer, kada je $\mathrm{Re}\, z \geq c > 0$. Tada za $\mathrm{Re}\, z < c$, imamo

$$\Gamma(z) = \frac{\pi \csc \pi z}{\Gamma(1-z)}. \tag{3.11.4}$$

Refleksiona formula (3.11.3) se može predstaviti i u sledećim simetričnim oblicima

$$\Gamma\left(\frac{1}{2}+z\right)\Gamma\left(\frac{1}{2}-z\right) = \frac{\pi}{\cos \pi z} \quad \text{i} \quad \Gamma(z)\Gamma(-z) = -\frac{\pi}{z \sin \pi z}.$$

Za velike vrednosti $|z|$ važi STIRLINGova[72] aproksimaciona formula

[70] ADRIEN-MARIE LEGENDRE (1752 – 1833), poznati francuski matematičar.

[71] KARL THEODOR WILHELM WEIERSTRASS (1815 – 1897), veliki nemački matematičar, sa značajnim doprinosima u matematičkoj analizi.

[72] JAMES STIRLING (1692 – 1770), poznati škotski matematičar.

$$\Gamma(z) = \sqrt{2\pi}\, z^{z-1/2} \mathrm{e}^{-z} \left(1 + \frac{1}{12z} + \frac{1}{288z^2} - \frac{139}{51840z^3} - \frac{571}{2488320z^4} + O\left(\frac{1}{z^5}\right) \right).$$

Korišćenjem EULERovog integrala (3.11.1), LANCZOS[73] [43] je korigovao STIRLINGovu aproksimaciju u obliku

$$z! = \left(z + a + \frac{1}{2} \right)^{z+1/2} \mathrm{e}^{-(z+a+1/2)} \sqrt{2\pi} \left(\frac{\rho_0}{2} + \frac{\rho_1 z}{z+1} + \frac{\rho_2 z(z-1)}{(z+1)(z+2)} + \cdots \right),$$

gde je a proizvoljni parametar takav da je $\mathrm{Re}(z+a+1/2) > 0$. Izrazi za koeficijente $\rho_0, \rho_1, \rho_2, \ldots$, koji zavise od parametra a, su dosta komplikovani. Stavljajući $a := a + 1/2$ i rastavljanjem racionalnih funkcija na desnoj strani prethodne formule u parcijalne razlomke, dobijamo

$$z! = (z+a)^{z+1/2} \mathrm{e}^{-(z+a)} \sqrt{2\pi} \left[c_0 + \sum_{k=1}^{N} \frac{c_k}{z+k} + \varepsilon(z) \right], \tag{3.11.5}$$

gde optimalna vrednost za broj članova N u prethodnoj sumi zavisi od parametra a. LANCZOSev razvoj brzo konvergira i zahteva relativno mali broj članova u razvoju da bi se dobila vrednost gama funkcije u standardnoj dvostrukoj preciznosti. Taj razvoj je, na primer, korišćen za izračunavanje gama funkcije u poznatoj kolekciji algoritama *Numerical Recipes*, sa odgovarajućim programskim implementacijama u raznim programskim jezicima (videti, na primer, [59, str. 213–216]).

J. L. SPOUGE[74] [66] je, međutim, odredio u prethodnom razvoju koeficijente c_k, u funkciji realnog parametra $a > 0$, tako da (3.11.5) važi za $\mathrm{Re}\,(z+a) > 0$ i $N = \lceil a \rceil - 1$, gde je sa $\lceil a \rceil$ označen ceo broj koji zadovoljava $\lceil a \rceil - 1 < a \leq \lceil a \rceil$. Taj razvoj je nešto sporiji, nego LANCZOSev, ali su koeficijenti c_k jednostavniji za izračunavanje,

$$c_0 = 1, \quad c_k = \frac{1}{\sqrt{2\pi}} \frac{(-1)^{k-1}}{(k-1)!} (-k+a)^{k-1/2} \mathrm{e}^{a-k}, \quad k = 1, 2, \ldots, N.$$

Očigledno, $c_k = \operatorname*{Res}_{z=-k} \Gamma(z+1)(z+a)^{-(z+1/2)} \mathrm{e}^{z+a} (2\pi)^{-1/2}$, $k \geq 1$. Ono što je najvažnije, uzimajući da je $a \geq 3$, SPOUGE je dokazao jednostavnu strogu ocenu za relativnu grešku

[73] CORNELIUS (CORNEL) LANCZOS (1893 – 1974), mađarski matematičar i fizičar jevrejskog porekla.

[74] JOHN L. SPOUGE (1955 –), američki matematičar rođen u Engleskoj.

$$|r(a,z)| = \left| \frac{\varepsilon(z)}{\Gamma(z+1)(z+a)^{-(z+1/2)}\mathrm{e}^{z+a}(2\pi)^{-1/2}} \right| < \frac{\sqrt{a}}{(2\pi)^{a+1/2}\mathrm{Re}\,(z+a)}.$$

Na primer, ako je z realan pozitivan broj, očigledno je da važi uniformna granica

$$|r(a,z)| < \frac{1}{\sqrt{a}(2\pi)^{a+1/2}} < \frac{1}{(2\pi)^a}.$$

Međutim, ako se uzme da je $\mathrm{Re}\,z \geq 1/2$, može se uvek odrediti donja granica za parametar a, koja će obezbediti zahtevanu relativnu grešku, na primer, $|r(a,z)| < 10^{-d}$, gde je d broj tačnih decimalnih cifara u mantisi. U sledećem MATHEMATICA kodu realizovan je rekurzivni postupak u kome se za izračunavanje gama funkcije za $\mathrm{Re}\,z \geq c = 1/2$ koristi prethodni SPOUGEov razvoj, a za $\mathrm{Re}\,z < c = 1/2$ formula (3.11.4). Kod izračunavanja se zahteva $d = 20$ tačnih decimalnih cifara u mantisi. U cilju obezbeđenja ove tačnosti za svako $z \in \mathbb{C}$, uzet je tzv. *faktor sigurnosti* $5/4$ (videti, na primer, [18, str. 83]).

```
In[1]:= d = 20; a = Ceiling[N[5 / 4 d / Log[10, 2 Pi]]];
        c[k_] := 1 / Sqrt[2 Pi] (-1) ^ (k - 1) / (k - 1) ! (a - k) ^ (k - 1 / 2) Exp[a - k];
        gam[z_] := If[Re[z] < 1 / 2, Pi Csc[Pi z] / gam[1 - z],
            (z + a - 1) ^ (z - 1 / 2) Exp[1 - z - a] Sqrt[2 Pi] (1 + Sum[c[k] / (z + k - 1),
                  {k, 1, Ceiling[a] - 1}])];
```

Navodimo sada neka izračunavanja sprovedena pomoću ovog programa.

```
In[4]:= N[gam[1 + I], 20]
Out[4]= 0.49801566811835604271 - 0.15494982830181068512 i

In[5]:= 20 !
Out[5]= 2 432 902 008 176 640 000

In[6]:= N[gam[21], 20]
Out[6]= 2.4329020081766400000 × 10^18

In[7]:= N[gam[21], 40] / 20 ! - 1
Out[7]= -7.71 × 10^-37

In[8]:= N[gam[21], 40]
Out[8]= 2.4329020081766399999999999999999999998123 × 10^18
```

Interesantno je primetiti da je dobijena tačnost znatno veća od 20 cifara. Na primer, kod izračunavanja 20!, vidimo da je relativna greška 7.71×10^{-37}, što znači da se dobija rezultat sa oko 36 tačnih decimalnih cifara u mantisi (za ovaj, inače, celi broj sa 19 dekadnih cifara).

Najzad, kao test koristimo identitet

$$|\Gamma(1+\mathrm{i}y)| = \sqrt{\frac{\pi y}{\sinh(\pi y)}}, \quad -\infty < y < +\infty,$$

i izračunavamo $\Gamma(1+10\mathrm{i})$:

```
In[9]:= Abs[N[gam[1 + 10 I], 50]] - N[Sqrt[10 Pi / Sinh[10 Pi]], 50]
Out[9]= -5.644897194 × 10^-46
In[10]:= Abs[N[gam[1 + 10 I], 50]] / N[Sqrt[10 Pi / Sinh[10 Pi]], 50] - 1
Out[10]= -4.725501136 × 10^-40
```

Primetimo da se za $y = 10$, izračunata vrednost $|\Gamma(1+10\mathrm{i})|$ razlikuje od tačne vrednosti za manje od 10^{-45}. Odgovarajuća relativna greška ukazuje da je skoro 40 cifara tačno, tj. dva puta više nego što je zahtevano ($d = 20$).

Na slici 3.11.2 prikazan je 3D-grafik funkcije $(x,y) \mapsto |\Gamma(x+\mathrm{i}y)|$ za (x,y) u pravougaoniku $[-4,3] \times [-2,2]$. Sa grafika vidimo polove funkcije u tačkama $z = -k$, $k = 0,1,2,3,4$ (na realnoj osi).

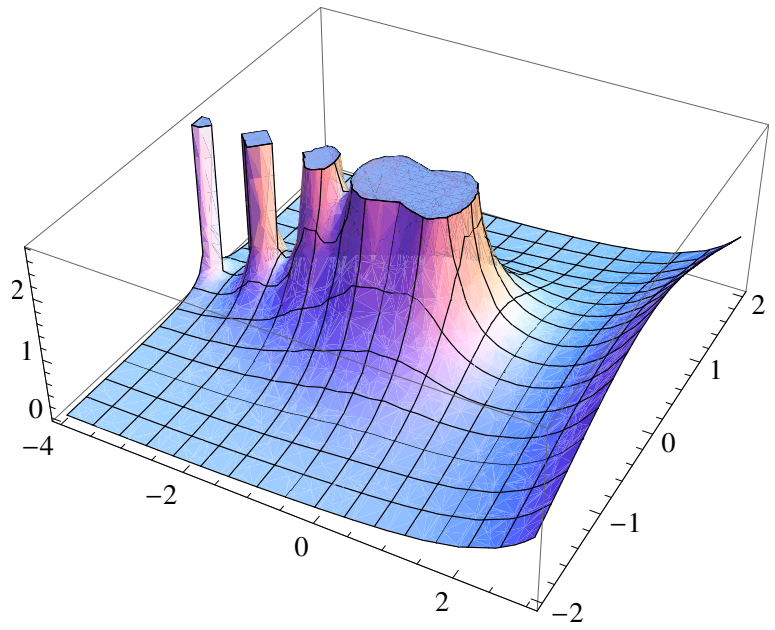
Napomena 3.11.1. U vezi sa gama funkcijom postoji obimna literatura[75]. Pomenimo samo dva rezultata u vezi sa izračunavanjem vrednosti gama funkcije. U [21] su određene numeričke vrednosti koeficijenata u TAYLORovom razvoju

$$\Gamma(a+x)^m = \sum_{k=0}^{+\infty} g_k(m,a)x^k$$

za izvesne vrednosti m i a, a zatim je to korišćeno za izračunavanje $\Gamma(p/q)$ $(p,q = 1(1)10;\ p < q)$ sa visokom preciznošću, dok su u radu [10], BOHMAN[76] i

[75] Na sajtu `http://www.milanmerkle.com`, profesor M. MERKLE je postavio bibliografiju sa nekih 900 referenci o gama funkciji i srodnim problemima. MILAN MERKLE (1954 –), srpski matematičar, redovni profesor Elektrotehničkog fakulteta Univerziteta u Beogradu.

[76] JAN BOHMAN (nemamo podatke o ovom autoru).



Slika 3.11.2. 3D-grafik funkcije $(x,y)\mapsto|\Gamma(x+\mathrm{i}y)|$ za $-4<x<3$, $-2<y<2$

FRÖBERG[77] izveli stepeni razvoj

$$\Gamma(n+1+z)=n!(1+d_1z+d_2z^2+\cdots)\qquad(n=2,3,4,10),$$

kao i

$$(-1)^n n!\Gamma(-n+z)=\frac{n}{1-z}-\frac{1}{(n+1)(1+z)}+\frac{1}{z}\big(1+f_1z+f_2z^2+\cdots\big)$$

za $n=0$ 1, 2, 10.

Na kraju ovog odeljka navešćemo i neke rezultate koji se odnose na druge faktorijelne funkcije i njihovo analitičko produženje na $\mathbb{C}$.

[77] CARL-ERIK FRÖBERG (1918 – 2007), poznati švedski teorijski fizičar i matematičar. Prvi je uveo program numeričke matematke na Univerzitetu u Lundu i pokrenuo poznati časopis *BIT Numerical Mathematics* (ISSN 0006–3835), čiji je urednik bio u periodu 1961–1992. Pod njegovim imenom uvedena je nagrada, koja se dodeljuje svake druge godine mladim autorima (do 35 godina) za najbolji rad u časopisu *BIT*.

Pored faktorijelne funkcije $n!$ koja definiše ukupan broj permutacija od n objekata, poznata je i tzv. *subfaktorijelna* funkcija[78] $n \mapsto S(n)$, koja daje broj permutacija od n objekata, u kojima nema objekta koji se pojavljuje na njegovom prirodnom mestu. EULER je još 1819. godine izračunao prvih deset članova, koji su ovde navedeni kao rezultat funkcije koja postoji u paketu MATHEMATICA:

```
In[1]:= Table[Subfactorial[n], {n, 10}]
Out[1]= {0, 1, 2, 9, 44, 265, 1854, 14 833, 133 496, 1 334 961}
```

Inače,

$$S(n) = n! \sum_{k=0}^{n} \frac{(-1)^k}{k!} = \sum_{k=0}^{n} (-1)^{n-k} k! \binom{n}{k}. \tag{3.11.6}$$

U daljem tekstu analiziraćemo i funkciju $n \mapsto K(n)$, tzv. „*levi faktorijel*“, definisanu pomoću

$$K(0) = 0, \qquad K(n) = 0! + 1! + \cdots + (n-1)!,$$

za koju se koriste i oznake $!n$ ili $L!n$. KUREPA[79], je razmatrajući neke probleme u teoriji brojeva 1971. godine postavio hipotezu (KH) da *za svako* $n \geq 2$*, najveći zajednički delilac* (NZD) *za levi i desni faktorijel od* n *je broj* 2, tj.

$$(\forall n \geq 2) \qquad \text{NZD}(!n, n!) = 2$$

(videti [37], kao i članak IVIĆa i MIJAJLOVIĆa[80] [35]). Hipoteza je navedena kao problem B44 u knjizi [30] i do sada je numerički verifikovana za $n < 2^{31} = 2147483648$.

U radu [38] KUREPA je definisao funkciju $z \mapsto K(z)$ pomoću

[78] Termin „*subfactorial*“ je uveo WILLIAM ALLEN WHITWORTH (1840 – 1905), engleski matematičar i sveštenik. U literaturi je definisan i izučavan veliki broj faktorijelnih funkcija, kao na primer, dvostruki faktorijel, faktorijelni stepen, hiperfaktorijel, super faktorijel, binomni koeficijenti, multimonomijalni koeficijenti, CATALANovi brojevi, STIRLINGovi brojevi, itd.

[79] ĐURO KUREPA (1907 – 1993), poznati srpski matematičar, redovni profesor Sveučilišta u Zagrebu do 1965. godine, a zatim sve do penzionisanja (1977) redovni profesor Univerziteta u Beogradu. Bio je redovni član Srpske akademije nauka i umetnosti (SANU) i Jugoslovenske akademije znanosti i umjetnosti (JAZU), koja je od 1991. promenila ime u Hrvatska akademija znanosti i umjetnosti (HAZU).

[80] ŽARKO MIJAJLOVIĆ (1948 –), srpski matematičar, redovni profesor Matematičkog fakulteta Univerziteta u Beogradu.

$$K(z)=\int_0^{+\infty}\frac{t^z-1}{t-1}\mathrm{e}^{-t}\,\mathrm{d}t,\quad \mathrm{Re}\,z>0, \tag{3.11.7}$$

i proširio analitičkim produženjem na celu kompleksnu ravan pomoću

$$K(z)=K(z+1)-\Gamma(z+1),$$

gde je $\Gamma(z)$ gama funkcija. Ovako definisana KUREPINA funkcija $K(z)$ je regularna sa prostim polovima u tačkama $z_k=-k$ $(k\in\mathbb{N}\setminus\{2\})$. Za $z=n(\in\mathbb{N})$ (3.11.7) se svodi na

$$K(n)=\int_0^{+\infty}\frac{t^n-1}{t-1}\mathrm{e}^{-t}\,\mathrm{d}t=\int_0^{+\infty}\left(\sum_{k=0}^{n-1}t^k\right)\mathrm{e}^{-t}\,\mathrm{d}t=\sum_{k=0}^{n-1}k!=K(n).$$

SLAVIĆ [63] je dokazao reprezentaciju

$$K(z)=-\frac{\pi}{\mathrm{e}}\cot\pi z+\frac{1}{\mathrm{e}}\left(\sum_{n=1}^{+\infty}\frac{1}{n!n}+\gamma\right)+\sum_{n=0}^{+\infty}\Gamma(z-n),$$

gde je γ EULERova konstanta.

U radu [50] MILOVANOVIĆ[81] je dokazao da za $|z|<a+1$ važi,

$$K(z)=\frac{1}{a+1+z}\sum_{\nu=1}^{+\infty}\beta_\nu(a)z^\nu,$$

gde su $\beta_0(a)=(a+1)b_0(a)$, $\beta_\nu(a)=(a+1)b_{\nu+1}(a)+b_\nu(a)$, $\nu\geq 1$, sa

$$b_0(a)=K(a),\quad b_\nu(a)=\frac{1}{\nu!}K^{(\nu)}(a)\quad(\nu\geq 1).$$

U slučaju $a=0$, imamo $\beta_\nu(0)=b_{\nu+1}(0)+b_\nu(0)=(-1)^\nu\Delta\varepsilon_\nu$, pri čemu $b_\nu(0)=(-1)^{\nu+1}(1+\varepsilon_\nu)$ i $\lim\limits_{\nu\to+\infty}\varepsilon_\nu=0$. Koeficijenti $\beta_\nu(0)$ i $\beta_\nu(1)$ za $\nu=0(1)60$, sa trideset tačnih cifara, određeni su u [50], kao i ČEBIŠEVljevi razvoji za $K(1+z)$ i $1/K(1+z)$.

MILOVANOVIĆ [51] je, takođe, proučavao niz funkcija $\left\{K_m(z)\right\}_{m=-1}^{+\infty}$ definisanih pomoću

$$K_m(z)=\int_0^{+\infty}\frac{t^{z+m}-Q_m(t;z)}{(t-1)^{m+1}}\mathrm{e}^{-t}\,\mathrm{d}t\quad(\mathrm{Re}\,z>0),$$

[81] GRADIMIR V. MILOVANOVIĆ (1948 –), srpski matematičar, autor ove knjige. Redovni je član Srpske akademije nauka i umetnosti (SANU).

gde su polinomi $Q_m(t;z)$ dati sa

$$Q_{-1}(t;z)=0,\quad Q_m(t;z)=\sum_{\nu=0}^{m}\binom{m+z}{\nu}(t-1)^\nu$$

i za svako $m\in\mathbb{N}_0$ zadovoljavaju jednakost

$$Q_m(t;z)=Q_{m-1}(t;z+1)+\frac{1}{m!}(z+1)(z+2)\cdots(z+m)(t-1)^m.$$

Na primer,

$$\begin{aligned}
Q_0(t;z)&=1,\\
Q_1(t;z)&=1+(z+1)(t-1),\\
Q_2(t;z)&=1+(z+2)(t-1)+\frac{1}{2}(z^2+3z+2)(t-1)^2,\\
Q_3(t;z)&=1+(z+3)(t-1)+\frac{1}{2}(z^2+5z+6)(t-1)^2\\
&\qquad+\frac{1}{6}(z^3+6z^2+11z+6)(t-1)^3,\ \text{itd.}
\end{aligned}$$

Primetimo da su $K_{-1}(z)=\Gamma(z)$ i $K_0(z)=K(z)$, kao i da se funkcije K_m mogu analitički produžiti na celu kompleksnu ravan, pomoću funkcionalne jednakosti

$$K_m(z)=K_m(z+1)-K_{m-1}(z+1).$$

Vrednosti funkcije K_m u tačakma $z=n\,(\in\mathbb{N})$ se mogu izraziti u obliku

$$K_m(n)=\sum_{i=0}^{n-1}\frac{(-1)^i}{i!}\sum_{\nu=i}^{n-1}\nu!\binom{m+n}{\nu+m+1},\quad K_m(0)=0,$$

ili pomoću subfaktorijela $S(\nu)$ (videti (3.11.6)) kao

$$K_m(n)=\sum_{\nu=0}^{n-1}\binom{m+n}{\nu+m+1}S(\nu).$$

Funkcije $K_m(z)$, $m\geq 1$, su regularne funkcije u $\mathbb{C}$, sa samo prostim polovima u tačkama $z=-(m+1),-(m+2),\ldots$.

Literatura

1. M. ABRAMOWITZ, I.A. STEGUN, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables,* Dover, New York, 1965.
2. T. M. APOSTOL, *An elementary view of Euler's summation formula*, Amer. Math. Monthly **106** (1999), 409–418.
3. E. W. BARNES, *The Maclaurin sum-formula*, Proc. London Math. Soc. **2-3** (1905), 253–272.
4. F.L. BAUER, *Computational graphs and rounding error,* SIAM J. Numer. Anal. **11** (1974), 87–96.
5. C. BREZINSKI, *Méthode d'accélération de la convergence en analyse numérique* (These), Grenoble, 1971.
6. C. BREZINSKI, *Accélération de suites à convergence logarithmique,* C. R. Acad. Sci. Paris Sér. A-B **273** (1971), A727–A730.
7. C. BREZINSKI, *Accélération de la convergence en analyse numérique,* Lecture Notes in Mathematics, Vol. 584, Springer-Verlag, Berlin–New York, 1977.
8. C. BREZINSKI, *A general extrapolation algorithm,* Numer. Math. **35** (1980), 175–187.
9. C. BREZINSKI, *Some new convergence acceleration methods,* Math. Comp. **39** (1982), 133–145.
10. J. BOHMAN, C.-E. FRÖBERG, *The Γ-function revisited: power series expansions and real-imaginary zero lines*, Math. Comp. **58** (1992), 315–322.
11. A. BULTHEEL, R. COOLS, eds., *The Birth of Numerical Analysis,* World Scientific, New Yersey – London – Singapore, 2010.
12. P. L. BUTZER, P. J. S. G. FERREIRA, G. SCHMEISSER, R. L. STENS, *The summation formulae of Euler-Maclaurin, Abel-Plana, Poisson, and their interconnections with the approximate sampling formula of signal analysis*, Results. Math. **59** (2011), 359–400.
13. C. CHIARELLA, A. REICHEL, *On the evaluation of integrals related to the error function,* Math. Comp. **22** (1968), 137–143.
14. J.-P. DELAHAYE, *Algorithmes pour suites non convergentes,* Numer. Math. **34** (1980), 333–347.
15. J.-P. DELAHAYE, *Optimalité du procédé Δ*sp2 *d'Aitken pour l'accélération de la convergence linéaire,* RAIRO Anal. Numér. **15** (1981), 321–330.
16. J.-P. DELAHAYE, *Automatic selection of sequence transformations,* Math. Comp. **37** (1981), 197–204.
17. L. BLUM, F. CUCKER, M. SHUB, S. SMALE, *Complexity and Real Computation,* Springer-Verlag, New York, 1998.
18. R. E. CRANDALL, *Topics in Advanced Sccientific Computation,* Springer-Verlag, New York; TELOS. The Electronic Library of Science, Santa Clara, CA, 1996.
19. J. EVE, *The evaluation of polynomials,* Numer. Math. **6** (1964), 17–21.
20. C. T. FIKE, *Computer Evaluation of Mathematical Functions,* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1968.
21. A. FRANSÉN, S. WRIGGE, *High-precision values of the gamma function and some related coefficients*, Math. Comp. **34** (1980), 553–566.
22. W. GAUTSCHI, *Computational aspects of three-term recurrence relations,* SIAM Review **9** (1967), 24–82.
23. W. GAUTSCHI, *Minimal solutions of three-term recurrence relations and orthogonal polynomials,* Math. Comp. **36** (1981), 547–554.
24. W. GAUTSCHI, *On the convergence behavior of continued fractions with real elements,* Math. Comp. **40** (1983), 337–342.
25. W. GAUTSCHI, *Numerical Analysis. An Introduction,* Birkhäuser Boston, Inc., Boston, MA, 1997.

26. W. GAUTSCHI, *Leonhard Eulers Umgang mit langsam konvergenten Reihen* (*Leonhard Euler's handling of slowly convergent series*), Elem. Math. **62** (2007), 174–183.
27. W. GAUTSCHI, *Leonhard Euler: his life, the man, and his works*, SIAM Rev. **50** (2008), 3–33.
28. A. O. GEL'FOND, *The Calculus of Finite Differences,* Izdat. "Nauka", Moskva, 1967 (na ruskom).
29. W. B. GRAGG, *Truncation error bounds for g-fractions,* Numer. Math. **11** (1968), 370–379.
30. R.K. GUY, *Unsolved Problems in Number Theory,* Third edition, Problem Books in Mathematics, Springer–Verlag, New York, 2004 (First edition 1981).
31. J. F. HART at al., *Computer Approximations*, Wiley, New York – London – Sydney, 1968.
32. P. HENRICI, *Applied and Computational Complex Analysis. Volume 2. Special functions – Integral transforms – Asymptotics – Continued Fractions*, Pure and Applied Mathematics, John Wiley & Sons Inc., New York – London – Sydney, 1977.
33. A. IVIĆ, *The Riemann Zeta-Function,* John Wiley & Sons, New York, 1985 [Dover, Mineola, New York, 2003].
34. A. IVIĆ, *The Theory of Hardy's Z-Function,* Cambridge University Press, Cambridge, 2013.
35. A. IVIĆ, Ž. MIJAJLOVIĆ, *On Kurepa problems in number theory,* Publ. Inst. Math. (Beograd) (N.S.) **57** (71) (1995), 19–28.
36. D. E. KNUTH, *Evaluation of polynomials by computer,* Comm. ACM **5** (1962), 595–599.
37. Đ. KUREPA, *On the left factorial function* !*n,* Math. Balkanica **1** (1971), 147–153.
38. Đ. KUREPA, *Left factorial function in complex domain,* Math. Balkanica **3** (1973), 297–307.
39. D. LEVIN, A. SIDI, *Two new classes of nonlinear transformations for accelerating the convergence of infinite integrals and series,* Appl. Math. Comput. **9** (1981), 175–215.
40. H. LEVY, F. LESSMAN, *Finite Difference Equations,* Dover Publications, Inc., New York, 1992.
41. LJ. M. KOCIĆ, *Geometrijsko modeliranje,* Elektronski fakultet u Nišu, Niš, 2009.
42. D. LEVIN, A. SIDI, *Two new classes of nonlinear transformations for accelerating the convergence of infinite integrals and series,* Appl. Math. Comput. **9** (1981), 175–215.
43. C. LANCZOS, *A precision approximation of the gamma function,* J. SIAM Numer. Anal. Ser. B**1** (1964), 86–96.
44. D. D. MCCRACKEN, W. S. DORN, *Numerical Methods and FORTRAN Programming,* Wiley, New York, 1964.
45. J. MIKLOŠKO, *Investigation of algorithms for the numerical computation of continued fractions,* Ž. Vyčisl. Mat. i Mat. Fiz. **16** (1976), 827–837.
46. J. MIKLOŠKO, *A fast algorithm for repeated computation of linear recurrence relations,* BIT **17** (1977), 430–436.
47. J. MIKLOŠKO, *An algorithn for calculating continued fractions,* J. Comput. Appl. Math. **3** (1977), 273–275.
48. S. MILLS, *The independent derivations by Euler, Leonhard and Maclaurin, Colin of the Euler-Maclaurin summation formula*, Arch. Hist. Exact Sci. **33** (1985), 1–13.
49. G.V. MILOVANOVIĆ, *Numerička analiza, I deo,* Naučna knjiga, Beograd, 1985.
50. G.V. MILOVANOVIĆ, *Expansions of the Kurepa function,* Publ. Inst. Math. (Beograd) (N.S.) **57** (71) (1995), 81–90.
51. G.V. MILOVANOVIĆ, *A sequence of Kurepa's functions,* Symposium Dedicated to the Memory of Đuro Kurepa (Belgrade, 1996). Sci. Rev. Ser. Sci. Eng. No. 19-20 (1996), 137–146.
52. G.V. MILOVANOVIĆ, *Families of Euler-Maclaurin formulae for composite Gauss-Legendre and Lobatto quadratures,* Bull. Cl. Sci. Math. Nat. Sci. Math. **38** (2013), 63–81.
53. G.V. MILOVANOVIĆ, R.Ž. ĐORĐEVIĆ, *Matematička analiza I,* Elektronski fakultet u Nišu, Niš, 2005.
54. G.V. MILOVANOVIĆ, D.S. MITRINOVIĆ, TH.M. RASSIAS, *Topics in Polynomials: Extremal Problems, Inequalities, Zeros,* World Scientific Publ. Co., Singapore – New Jersey – London – Hong Kong, 1994.

55. T. S. MOTZKIN, *Evaluation of polynomials,* Bull. Amer. Math. Soc. **61** (1955), 163 [Report of the sixty-first Annual Meeting of the American Mathematical Society was held at the University of Pittsburgh, December 27-29, 1954].
56. K. S. MILLER, *An Introduction to the Calculus of Finite Differences and Difference Equations,* Henry Holt and Co., New York, 1960.
57. W. NIETHAMMER, *Numerical application of Euler's series transformation and its generalizations,* Numer. Math. **34** (1980), 271–283.
58. N. E. NÖRLUND, *Vorlesungen über Differenzrechnung,* Springer, Berlin, 1924.
59. W. H. PRESS, S. A. TEUKOLSKY, W. T. VETTERLING, B. P. FLANNERY, *Numerical Recipes in C: The Art of Scientific Computing,* Cambridge University Press, Cambridge, 2002.
60. D. SHANKS, *Non-linear transformations of divergent and slowly convergent sequences,* J. Math. and Phys. **34** (1955), 1–42.
61. V. YA. SKOROBOGAT'KO, *The Theory of Branching Continued Fractions and its Application in Numerical Mathematics,* Nauka, Moskva, 1983 (na ruskom).
62. D. SLAVIĆ, *Transformation of the continued fraction into a rational function,* Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fiz. No. 577–598 (1977), 49–52.
63. D. V. SLAVIĆ, *On the left factorial function of the complex argument*, Math. Balkanica **3** (1973), 472–477.
64. D. A. SMITH, W. F. FORD, *Acceleration of linear and logarithmic convergence,* SIAM J. Numer. Anal. **16** (1979), 223–240.
65. D. A. SMITH, W. F. FORD, *Numerical comparisons of nonlinear convergence accelerators,* Math. Comp. **38** (1982), 481–499.
66. J. L. SPOUGE, *Computation of the gamma, digamma, and trigamma functions,* SIAM J. Numer. Anal. **31** (1994), 931–944.
67. J. STOER, *Einführung in die Numerische Mathematik I,* Springer Verlag, Berlin – Heidelberg – New York, 1972.
68. W. J. THRON, H. WAADELAND, *Truncation error bounds for limit periodic continued fractions,* Math. Comp. **40** (1983), 589–597.
69. J. TODD, *Motivation for working in numerical analysis,* Comm. Pure Appl. Math. **8** (1955), 97–116.
70. D. TOŠIĆ, *Uvod u numeričku analizu,* Naučna knjiga, Beograd, 1982.
71. J. H. WILKINSON, *Error analysis of floating-point computation,* Numer. Math. **2** (1960), 319–340.
72. E. T. WHITTAKER, G. N. WATSON, *A Course of Modern Analysis*, 4th edition, Series: Cambridge Mathematical Library, Cambridge University Press, 1996.
73. S. WOLFRAM, *The Future of Computation,* The Mathematica Journal **10** (2) (2006), 329–362.

2. ELEMENTI FUNKCIONALNE ANALIZE I LINEARNE ALGEBRE

2.1 PROSTORI

2.1.1 Linearni prostor

Izlaganje u ovoj glavi započinjemo osnovnim pojmovima i definicijama linearne algebre, a nastavljamo elementima funkcionalne analize, koji su ključni u rešavanju linearnih problema, ali i mnogih nelinearnih problema koji su na određeni način povezani sa linearnim ili se na njih svode nekim aproksimacionim pristupima. Razvoj modernih numeričkih i aproksimacionih metoda su u ogromnoj meri zasnovani na primeni koncepata linearne i nelinearne funkcionalne analize. Drugo poglavlje se odnosi na teoriju operatora, dok se u trećem poglavlju daju osnovni elementi matričnog računa. Izložena materija u okviru ove glave biće dovoljna za praćenje i razumevanje svega onoga što se razmatra u ostalim glavama ove knjige.

U našem izlaganju sa $\mathbb{K}$ ćemo označavati polje realnih ili kompleksnih brojeva, $\mathbb{R}$ ili $\mathbb{C}$, dok ćemo sa X označavati proizvoljni skup elemenata $\{u, v, w, \dots\}$, koji mogu biti tačke, vektori, funkcije, itd.

Definicija 1.1.1. Skup X se naziva *linearni* (*vektorski*) *prostor* nad poljem $\mathbb{K}$ ako je:

1° u skupu X definisana binarna operacija $+$ u odnosu na koju skup X čini ABELovu grupu;

2° svakom paru (u, c) $(u \in X;\ c \in \mathbb{K})$ dodeljen po jedan element cu skupa X tako da su ispunjeni uslovi:

$$
\begin{aligned}
c_1(c_2u) &= (c_1c_2)u, & (c_1+c_2)u &= c_1u + c_2u,\\
c(u_1+u_2) &= cu_1 + cu_2, & 1\,u &= u
\end{aligned}
\tag{1.1.1}
$$

za sve elemente $u, u_i \in X$ i $c, c_i \in \mathbb{K}$ $(i = 1, 2)$.

Jedinični element polja $\mathbb{K}$ je označen sa 1. Elementi skupa X nazivaju se vektori (tačke), elementi polja $\mathbb{K}$ skalari, operacija $+$ u skupu X vektorsko sabiranje (unutrašnja kompozicija) i operacija $(u,c) \to cu$ množenje vektora skalarom (spoljašnja kompozicija).

Treća jednakost iz (1.1.1), pri $c_1 = 1$ i $c_2 = -1$, daje

$$(\forall u \in X) \qquad 0u = u + (-u) = \theta,$$

gde je θ neutralni element skupa X za operaciju vektorskog sabiranja i ovaj element se naziva *nula-vektor* prostora X.

Osim toga, za $u_1 + u_2 = \theta$, iz druge jednakosti u (1.1.1) zaključujemo da je za svako $c \in \mathbb{K}$,

$$c\theta = \theta.$$

Definicija 1.1.2. Za vektore u_i $(i = 1,\dots,n)$ linearnog prostora X kaže se da su *linearno zavisni* ako u polju $\mathbb{K}$ postoje skalari c_i $(i = 1,\dots,n)$, koji istovremeno nisu svi jednaki nuli, tako da je

$$c_1 u_1 + \cdots + c_n u_n = \theta. \tag{1.1.2}$$

Vektori u_i $(i = 1,\dots,n)$ su *linearno nezavisni*, ako je jednakost (1.1.2) tačna samo za $c_i = 0$ $(i = 1,\dots,n)$.

Definicija 1.1.3. Za beskonačno mnogo vektora kažemo da su linearno nezavisni, ako je svaki sistem konačnog broja tih vektora linearno nezavisan.

Definicija 1.1.4. Ako u vektorskom prostoru postoji n linearno nezavisnih vektora i ako je svaki skup od $n+1$ vektora linearno zavisan, kažemo da je prostor *konačno–dimenzionalan*, tačnije rečeno *n–dimenzionalan*. Za broj n kažemo da je *dimenzija prostora*.

Ako u vektorskom prostoru postoji beskonačno mnogo linearno nezavisnih vektora, za taj vektorski prostor kažemo da je *beskonačno–dimenzionalan*.

Definicija 1.1.5. Neka je $A = \{u_1,\dots,u_m\}$, gde su u_k $(k = 1,\dots,m)$ vektori prostora X. Skup svih linearnih kombinacija ovih vektora naziva se *linearni omotač* ili *lineal* nad A i označava se sa $L(A)$.

Definicija 1.1.6. Skup B linearno nezavisnih vektora prostora X obrazuje *algebarsku* ili *Hamelovu*[82] bazu prostora X, ako je $L(B) = X$.

[82] GEORG KARL WILHELM HAMEL (1877 – 1954), nemački mehaničar i matematičar.

Teorema 1.1.1. *Svaki vektor linearnog prostora X može se na jedinstven način izraziti kao linearna kombinacija vektora algebarske baze tog prostora.*

Napomenimo da vektorski prostor ima beskonačno mnogo različitih baza, međutim, sve one su iste kardinalnosti. Kod n-dimenzionalnog prostora sve baze sadrže n vektora. Na primer, ako je u n-dimenzionalnom prostoru X skup vektora $\{u_1,\ldots,u_n\}$ linearno nezavisan, onda je $L(\{u_1,\ldots,u_n\}) = X$ i taj skup vektora je jedna baza u X.

Definicija 1.1.7. Svaka baza prostora naziva se *koordinatni sistem* tog prostora.

Neka je $B = \{u_1,\ldots,u_n\}$ jedna baza konačno–dimenzionalnog prostora X. Tada se, na osnovu teoreme 1.1.1, vektor $u \in X$ može predstaviti u obliku

$$u = x_1u_1 + \cdots + x_nu_n.$$

Dakle, ako je zadata baza B, vektor u je potpuno određen skalarima x_1, …, x_n i može se korišćenjem matrične notacije opisati pomoću tzv. koordinatne reprezentacije

$$\boldsymbol{x} = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}^{\mathrm{T}}.$$

Skalari $x_1,\ldots,x_n$ nazivaju se *koordinate vektora*. Često se, ako to ne dovodi do zabune, u i $\boldsymbol{x}$ poistovećuju.

Primer 1.1.1. Neka je $X = \mathbb{R}^n = \left\{(x_1,\ldots,x_n) \mid x_i \in \mathbb{R}\ (i = 1,\ldots,n)\right\}$ i $\mathbb{K} = \mathbb{R}$. Ako u ovaj skup uvedemo unutrašnju i spoljašnju kompoziciju pomoću

$$(x_1,\ldots,x_n) + (y_1,\ldots,y_n) = (x_1+y_1,\ldots,x_n+y_n),$$
$$c(x_1,\ldots,x_n) = (cx_1,\ldots,cx_n),$$

on postaje vektorski prostor nad poljem $\mathbb{R}$, sa dimenzijom $\dim \mathbb{R}^n = n$. Jedna baza ovog prostora je skup $\{e_1,e_2,\ldots,e_n\}$, gde su

$$e_1 = (1,0,\ldots,0),\ \ e_2 = (0,1,\ldots,0),\ \ldots,\ \ e_n = (0,0,\ldots,1)\,.$$

Ukoliko drugačije nije rečeno, uvek ćemo u daljem tekstu podrazumevati da je u prostoru $\mathbb{R}^n$ zadata pomenuta baza, koja se naziva i *prirodna baza*. Saglasno prethodnom, za tačke ovog prostora, pored oznake $u = (x_1,\ldots,x_n)$, koristićemo i koordinatnu reprezentaciju

$$\boldsymbol{x} = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}^{\mathrm{T}}.$$

Primetimo da je prirodna baza privilegovana u smislu da se tačka $u = (x_1, \ldots, x_n)$ i njena koordinatna reprezentacija opisuju pomoću istih skalara $x_1, \ldots, x_n \in \mathbb{R}$. Zato ćemo koristiti i notaciju $\boldsymbol{x} = (x_1, \ldots, x_n)$.

Slično se može razmatrati i konačno–dimenzionalni vektorski prostor $\mathbb{C}^n$ nad poljem $\mathbb{C}$, čija je dimenzija, takođe, n. Međutim, ako se razmatra $\mathbb{C}^n$ nad poljem $\mathbb{R}$, tada je dimenzija takvog prostora $2n$. Ovo proizilazi iz činjenice da se $\mathbb{C}^n$ može tretirati kao $\mathbb{R}^{2n}$, s obzirom da se $\mathbb{C}$ može tretirati kao $\mathbb{R}^2$ (videti, na primer, [20, str. 66–82]). $\triangle$

Primer 1.1.2. Skup ℓ^p $(p \geq 1)$ realnih (ili kompleksnih) nizova $u = \{x_k\}_{k\in\mathbb{N}}$ za koje važi $\sum_{k=1}^{+\infty} |x_k|^p < +\infty$, obrazuje vektorski prostor ako su unutrašnja i spoljašnja kompozicija uvedene sa

$$u + v = \{x_k\}_{k\in\mathbb{N}} + \{y_k\}_{k\in\mathbb{N}} = \{x_k + y_k\}_{k\in\mathbb{N}}, \quad cu = c\{x_k\}_{k\in\mathbb{N}}. \quad \triangle$$

Primer 1.1.3. Posmatrajmo skup svih realnih funkcija koje su neprekidne na konačnom segmentu $[a,b]$, $-\infty < a < b < +\infty$, i u taj skup uvedimo unutrašnju i spoljašnju kompoziciju sa

$$(u+v)(t) = u(t) + v(t) \qquad \text{i} \qquad (cu)(t) = cu(t), \quad c \in \mathbb{R}.$$

Jednostavno se može pokazati da ovaj skup na taj način postaje vektorski prostor nad poljem realnih brojeva. Nula–vektor ovog prostora je funkcija koja je identički jednaka nuli na $[a,b]$. Taj beskonačno–dimenzionalni prostor označavamo sa $C[a,b]$, mada neki insistiraju na oznaci $C([a,b])$, s obzirom da se $C(D)$ koristi za skup (ili prostor) neprekidnih funkcija definisanih na znatno opštijem skupu D.

Na potpuno isti način može se razmatrati i vektorski prostor m puta neprekidno-diferencijabilnih funkcija, u oznaci $C^m[a,b]$. $\triangle$

Primer 1.1.4. Posmatrajmo stepene funkcije (monome) $t \mapsto t^k$ $(k = 0, 1, \ldots, n)$ definisane na $\mathbb{R}$. Kako je

$$(\forall t \in \mathbb{R}) \qquad c_0 1 + c_1 t + c_2 t^2 + \cdots + c_n t^n = 0$$

samo ako je $c_0 = c_1 = \cdots = c_n = 0$, zaključujemo da je skup (sistem) funkcija $\{1, t, t^2, \ldots, t^n\}$ linearno nezavisan. Lineal nad njim je skup svih polinoma stepena ne višeg od n, u oznaci $\mathcal{P}_n$,

$$\mathcal{P}_n = \left\{ u \mid u(t) = c_0 + c_1 t + \cdots + c_n t^n,\ c_i \in \mathbb{R}\ (i = 0, 1, \ldots, n) \right\}.$$

Ako u skup $\mathcal{P}_n$ uvedemo unutrašnju i spoljašnju kompoziciju (sabiranje dva polinoma i množenje polinoma skalarom) na uobičajeni način kao i u prostoru neprekidnih funkcija (videti primer 1.1.3), tada $\mathcal{P}_n$ postaje vektorski prostor nad poljem $\mathbb{R}$. Nula-vektor ovog prostora je polinom koji je identički jednak nuli. Slično se $\mathcal{P}_n$ može tretirati i kao vektorski prostor nad poljem kompleksnih brojeva.

Kako baza prostora $\{1, t, t^2, \ldots, t^n\}$ sadrži $n+1$ elemenata (vektora, funkcija), prostor $\mathcal{P}_n$ je dimenzije $n+1$. $\triangle$

Primer 1.1.5. Prostor svih polinoma stepena ne višeg od dva, tj. prostor svih kvadratnih trinoma

$$\mathcal{P}_2 = \{u \mid u(t) = c_0 + c_1 t + c_2 t^2,\ c_0, c_1, c_2 \in \mathbb{R}\}$$

je trodimenzionalan.

Umesto baze $B = \{1, t, t^2\}$ moguće je uzeti i neku drugu bazu, na primer, $\widetilde{B} = \{1, t-1, t^2+t\}$. Trinom $u(t) = 5 + 3t - 2t^2$ u novoj bazi $\widetilde{B}$ ima reprezentaciju

$$u(t) = 10 + 5(t-1) - 2(t^2+t).$$

Dakle, odgovarajuće koordinatne reprezentacije ovog elementa u u bazama B i $\widetilde{B}$ su $[5\ \ 3\ \ -2]^{\mathrm{T}}$ i $[10\ \ 5\ \ -2]^{\mathrm{T}}$, respektivno. $\triangle$

Primer 1.1.6. Neka je $X = L^1(a,b)$ skup svih LEBESGUE–integrabilnih[83] funkcija definisanih na $[a,b]$, tj. takvih za koje je $\int_a^b |u(t)|\,dt < +\infty$. Takav skup, takođe, obrazuje vektorski prostor ako je unutrašnja i spoljašnja kompozicija uvedena kao u primeru 1.1.3. Njegovi vektori su, u stvari, *klase ekvivalencije* funkcija u odnosu na relaciju „jednakost *skoro svuda*“[84]. Dakle, dve funkcije pripadaju istoj klasi ekvivalencije ako se međusobno razlikuju samo na skupu mere nula.

Pored prostora $L^1(a,b)$, za dato $p \geq 1$, razmatraju se i vektorski prostori p–integrabilnih funkcija $L^p(a,b)$, za koje je $\int_a^b |u(t)|^p\,dt < +\infty$. Posebno su od interesa prostori $L^2(a,b)$, koji igraju značajnu ulogu u tzv. srednje–kvadratnim aproksimacijama, naravno uz dodatno obogaćivanje njihove topološke i metričke strukture. $\triangle$

[83] HENRI LUI LEBESGUE (1875 – 1941), poznati francuski matematičar.

[84] Ako je neko svojstvo ispunjeno u svim tačkama skupa $A \setminus E$, gde je E skup mere nula, kažemo da je to svojstvo ispunjeno *skoro svuda* na A, ili da važi u skoro svim tačkama skupa A. Pojam skoro svuda je veoma važan u matematičkoj analizi. Može se, na primer, govoriti o ispunjenju jednakosti $u(t) = v(t)$ skoro svuda na $[a,b]$, kao u posmatranom slučaju, o neprekidnosti funkcije skoro svuda na $[a,b]$, itd. Inače, za skup $E \subset \mathbb{R}$ kažemo da ima LEBESGUEovu meru nula ako za svako $\varepsilon > 0$ postoji konačan ili prebrojiv sistem prekrivajućih intervala za E, čija ukupna dužina nije veća od ε.

Definicija 1.1.8. Neprazan skup $Y \subset X$ je *potprostor vektorskog prostora* X nad poljem $\mathbb{K}$ ako je Y vektorski prostor nad istim poljem $\mathbb{K}$.

Da bi prostor Y bio potprostor od X potrebno je i dovoljno da Y sadrži vektor $\alpha u + \beta v$ $(\alpha, \beta \in \mathbb{K})$ kad god on sadrži vektore u i v.

Za detalje o linearnim prostorima videti, na primer, [20].

2.1.2 Metrički i topološki prostori

Pored linearnih prostora kod kojih je bitna algebarska struktura moguće je uvesti znatno opštiji koncept poput topološkog prostora koji omogućava razmatranje granica beskonačnih nizova tačaka i neprekidnost funkcija, što je neophodno u tretiranju mnogih aproksimacionih i numeričkih problema. U daljem izlaganju razmatraćemo specijalne klase topoloških prostora, kakvi su metrički prostori, normirani prostori, prostori sa unutrašnjim proizvodom, itd.

Neka je X neprazan skup čiji su elementi $u, v, w, \ldots$. Najpre uvodimo definiciju metričkog prostora koja će obezbediti merenje rastojanja između njegovih elemenata. Dakle, bitno je definisati funkciju $d\colon X^2 \to [0, +\infty)$ koja daje nenegativno rastojanje između proizvoljne dve tačke $u, v \in X$, tj. $d(u,v) \geq 0$, ali, takođe, treba obezbediti da je rastojanje između tačaka u i v isto kao i rastojanje između v i u, tj. da je $d(u,v) = d(v,u)$. Formalna definicija metričkog prostora se može iskazati u sledećem obliku.

Definicija 1.2.1. Ako preslikavanje $d\colon X^2 \to [0, +\infty)$ ispunjava uslove

1° $d(u,v) = 0 \Leftrightarrow u = v$ $(u, v \in X)$,

2° $d(u,v) = d(v,u)$ za svako $u, v \in X$,

3° $d(u,w) \leq d(u,v) + d(v,w)$ za svako $u, v, w \in X$.

za njega kažemo da je *funkcija rastojanja* ili *metrika* u skupu X, a za sâm skup X snabdeven metrikom d da je *metrički prostor* i to označavamo sa (X,d).

Primetimo da je $d(u,v) > 0$ kad god je $u \neq v$. Osobine 2° i 3° su poznate kao *simetričnost* i *relacija trougla*, respektivno. Osobina 3° predstavlja analogon dobro poznate nejednakosti za trougao da jedna stranica trougla nikad nije veća od zbira druge dve stranice.

Dakle, metrički prostor čine skup X i uvedena metrika d. Ako su d_1 i d_2 dve različite metrike definisane na istom skupu X, tada su (X,d_1) i (X,d_2) različiti metrički prostori. Međutim, uvek kada je poznata metrika d i kada ne može doći do zabune, umesto (X,d) koristimo samo oznaku X.

Definicija 1.2.2. Za dva metrička prostora (X,d) i (Y,ρ) kažemo da su *izometrična* ako između njih postoji biunivoko preslikavanje $f: X \to Y$ takvo da je

$$\big(\forall (u,v) \in X^2\big) \qquad \rho(f(u),f(v)) = d(u,v).$$

Za preslikavanje f kažemo da je *izometrija* između ova dva prostora.

Izometrične prostore, po pravilu, ćemo poistovećivati.

Primer 1.2.1. Neka je $X = \mathbb{R}$ i $d(x,y) = |x-y|$, $x,y \in \mathbb{R}$. Jednostavno se dokazuju sve tri osobine iz definicije 1.2.1, tako da je $(\mathbb{R},d)$ metrički prostor. $\triangle$

Primer 1.2.2. Neka je $X = \mathbb{R}^2$. Za dve tačke $u = \boldsymbol{x} = (x_1,x_2)$ i $v = \boldsymbol{y} = (y_1,y_2)$, funkcija rastojanja se može uvesti na više načina. Najčešće, ako nije drugačije naglašeno, pod rastojanjem između dve tačke u ravni podrazumevamo tzv. *euklidsko rastojanje*,

$$d_2(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2}.$$

Nije teško proveriti da je $(\mathbb{R}^2,d_2)$ metrički prostor, s obzirom na važenje svih osobina iz definicije 1.2.1. Osobina 3° je klasična nejednakost trougla.

Metrika se, međutim, u $\mathbb{R}^2$ može uvesti i na sledeći način, za svako $p \geq 1$,

$$d_p(\boldsymbol{x},\boldsymbol{y}) = \left(|x_1-y_1|^p + |x_2-y_2|^p\right)^{1/p}, \tag{1.2.1}$$

tako da su $(\mathbb{R}^2,d_p)$ međusobno različiti metrički prostori, iako svi oni koriste isti skup tačaka $\mathbb{R}^2$. Na primer, za $p = 1$ imamo

$$d_1(\boldsymbol{x},\boldsymbol{y}) = |x_1-y_1| + |x_2-y_2|,$$

dok se za $p \to +\infty$ metrika (1.2.1) svodi na

$$d_\infty(\boldsymbol{x},\boldsymbol{y}) = \max\big\{|x_1-y_1|, |x_2-y_2|\big\}.$$

U opštem slučaju, $(\mathbb{R}^n,d_p)$, $p \geq 1$, je n–dimenzionalni metrički prostor sa metrikom

$$d_p(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_{k=1}^{n} |x_k-y_k|^p\right)^{1/p}, \quad \boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n. \tag{1.2.2}$$

Osobine 1° i 2° su očigledne, dok je osobina 3° posledica nejednakosti MINKOWSKOG[85]

[85] HERMANN MINKOWSKI (1864 – 1909), nemački matematičar i fizičar.

$$\left(\sum_{k=1}^{n}|a_k+b_k|^p\right)^{1/p}\leq\left(\sum_{k=1}^{n}|a_k|^p\right)^{1/p}+\left(\sum_{k=1}^{n}|b_k|^p\right)^{1/p}, \tag{1.2.3}$$

koja važi za svako $p\geq 1$ i proizvoljne nizove realnih ili kompleksnih brojeva a_k i b_k $(k=1,2,\ldots,n)$. Zaista, ako za tačke u $\mathbb{R}^n$,

$$\boldsymbol{x}=(x_1,x_2,\ldots,x_n),\quad \boldsymbol{y}=(y_1,y_2,\ldots,y_n),\quad \boldsymbol{z}=(z_1,z_2,\ldots,z_n),$$

stavimo $a_k:=x_k-z_k$, $b_k:=z_k-y_k$, $k\geq 1$, tada je $a_k+b_k=x_k-y_k$, $k=1,2,\ldots,n$, i nejednakost (1.2.3) se svodi na

$$d_p(\boldsymbol{x},\boldsymbol{y})\leq d_p(\boldsymbol{x},\boldsymbol{z})+d_p(\boldsymbol{z},\boldsymbol{y}),\quad p\geq 1.$$

Metrički prostor $(\mathbb{R}^n,d_2)$ je poznat kao *Euklidov*[86] *prostor*. Kada $p\to+\infty$, metrika (1.2.2) postaje

$$d_\infty(\boldsymbol{x},\boldsymbol{y})=\max_{1\leq k\leq n}|x_k-y_k|.$$

Za dokaz nejednakosti (1.2.3) videti, na primer, [20, str. 28]. $\triangle$

Primer 1.2.3. Neka je $X=\ell^p$ $(p\geq 1)$ skup realnih (ili kompleksnih) nizova $u=\{x_k\}_{k\in\mathbb{N}}$ za koje važi $\sum\limits_{k=1}^{+\infty}|x_k|^p<+\infty$ (videti primer 1.1.2). Ako za dva niza $u=\{x_k\}_{k\in\mathbb{N}}$ i $v=\{y_k\}_{k\in\mathbb{N}}$ iz ℓ^p definišimo funkciju rastojanja pomoću

$$d(u,v)=\left(\sum_{k=1}^{+\infty}|x_k-y_k|^p\right)^{1/p},$$

tada je (ℓ^p,d) metrički prostor. $\triangle$

Primer 1.2.4. Neka je $X=C[a,b]$ skup funkcija neprekidnih na konačnom segmentu $[a,b]$, $-\infty<a<b<+\infty$. Standardna metrika se uvodi pomoću

$$d(u,v)=\max_{a\leq t\leq b}|u(t)-v(t)|,\quad u,v\in C[a,b]. \tag{1.2.4}$$

Osobina 3° sleduje iz činjenice da je za svako $u,v,w\in C[a,b]$ i $a\leq t\leq b$,

$$\begin{aligned}|u(t)-v(t)|&\leq|u(t)-w(t)|+|w(t)-v(t)|\\&\leq\max_{a\leq t\leq b}|u(t)-w(t)|+\max_{a\leq t\leq b}|w(t)-v(t)|,\end{aligned}$$

[86] EUKLID (IV–III vek pre naše ere), starogrčki matematičar.

tj.

$$\max_{a\le t\le b}|u(t)-v(t)| \le \max_{a\le t\le b}|u(t)-w(t)| + \max_{a\le t\le b}|w(t)-v(t)|.$$

Za ovaj metrički prostor koristićemo oznaku $C[a,b]$, podrazumevajući prethodno uvedenu tzv. *uniformnu metriku* (1.2.4). Inače, skup $C[a,b]$ se može snabdeti i integralnom metrikom

$$d_1(u,v) = \int_a^b |u(t)-v(t)|\,\mathrm{d}t, \quad u,v\in C[a,b]. \quad \triangle \tag{1.2.5}$$

Definicija 1.2.3. Ako je (X,d) metrički prostor i $A\subset X$, odstojanje tačke $a\in X$ od skupa A, u oznaci $d(a,A)$, određeno je pomoću $d(a,A)=\inf\{d(a,u)\mid u\in A\}$.

Ako $a\in A$, tada je $d(a,A)=0$. Međutim, obrnuto ne mora da važi, tj. ako je $d(a,A)=0$ ne znači da $a\in A$. Zaista, ako u prostoru $(\mathbb{R},d)$ iz primera 1.2.1, za skup A uzmemo $A=(0,+\infty)\subset\mathbb{R}$, tada $d(0,A)=0$, ali $0\notin A$.

Definicija 1.2.4. Ako je (X,d) metrički prostor i ako su A i B podskupovi skupa X, rastojanje između skupova A i B, u oznaci $d(A,B)$, određeno je sa

$$d(A,B)=\inf\{d(u,v)\mid u\in A,\, v\in B\}.$$

Definicija 1.2.5. Ako je (X,d) metrički prostor i $A\subset X$, tada za veličinu

$$\operatorname{diam} A=\sup\{d(u,v)\mid u,v\in A\}$$

kažemo da je *dijametar skupa* A.

Definicija 1.2.6. Skup $A\subset X$ je *ograničen* ako ima konačan dijametar.

U metričke prostore moguće je uvoditi topološku strukturu. Osnovni pojam koji u tome igra značajnu ulogu je pojam otvorene kugle. U daljem tekstu neka je (X,d) metrički prostor, tačka $a\in X$ i $r\in\mathbb{R}^+$.

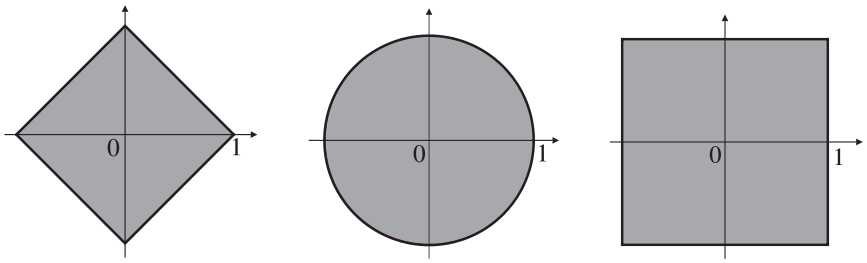
Definicija 1.2.7. Skup $K(a,r)=\{u\in X\mid d(u,a)<r\}$ nazivamo *otvorena kugla* poluprečnika r, sa centrom u tački a.

Definicija 1.2.8. Skup $K[a,r]=\{u\in X\mid d(u,a)\le r\}$ nazivamo *zatvorena kugla* poluprečnika r, sa centrom u tački a.

Kugla $K(a,r)$ se često naziva i r–okolina tačke a. Naravno, šta kugla $K(a,r)$ predstavlja u metričkom prostoru (X,d) zavisi od prirode skupa X, ali i od uvedene metrike d. Ilustrovaćemo to na nekim primerima.

Primer 1.2.5. U metričkom prostoru $(\mathbb{R}, d)$, gde je $d(x,y) = |x-y|$, kugla $K(a,r)$ je interval $(a-r, a+r)$. $\triangle$

Primer 1.2.6. Neka je $(\mathbb{R}^2, d_p)$, $p \geq 1$, metrički prostor sa metrikom datom pomoću (1.2.1).

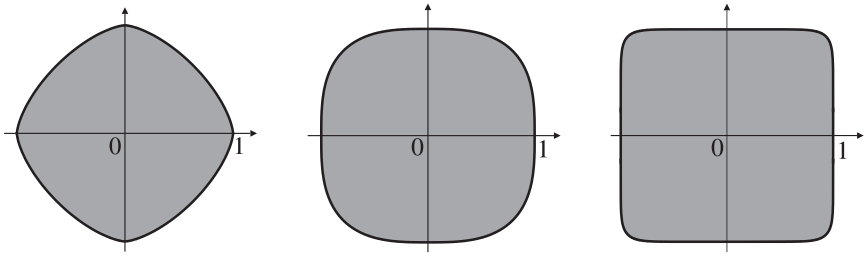


Slika 1.2.1. Kugle $K_p(\mathbf{0}, 1)$ za $p = 1$, $p = 2$, $p = +\infty$

Ako su $\boldsymbol{a} = (a_1, a_2)$ i $u = \boldsymbol{x} = (x_1, x_2)$, kugla $K(\boldsymbol{a}, r)$ je, saglasno definiciji 1.2.7,

$$K(\boldsymbol{a}, r) = K_p(\boldsymbol{a}, r) = \left\{ \boldsymbol{x} \in \mathbb{R}^2 \mid (x_1 - a_1)^p + (y_1 - a_2)^p < r^p \right\}.$$

Na primer, za $\boldsymbol{a} = \mathbf{0} = (0,0)$, $r = 1$, i za tri različite vrednosti $p = 1$, $p = 2$, $p = +\infty$, odgovarajuće kugle $K_p(\mathbf{0}, 1)$ su prikazane na slici 1.2.1. Primetimo da je kugla u euklidskoj normi ($p = 2$) jedinični krug, dok je za $p = 1$ i $p = +\infty$ kvadrat sa stranicama $\sqrt{2}$ i 2, respektivno. Na slici 1.2.2 prikazane su kugle za $p = 3/2$, $p = 3$, $p = 10$. $\triangle$



Slika 1.2.2. Kugle $K_p(\mathbf{0}, 1)$ za $p = 3/2$, $p = 3$, $p = 10$

Definicija 1.2.9. Skup $A \subset X$ je *otvoren* skup ako je prazan ili ako za svaku tačku $u \in A$ postoji neka njena ε-okolina koja je podskup od A, tj. postoji $\varepsilon > 0$ takvo da $K(u,\varepsilon) \subset A$.

Naravno, otvorena kugla je otvoren skup u metričkom prostoru. Nije teško zaključiti da je svaki neprazan otvoren skup A u metričkom prostoru X unija familije otvorenih kugli metričkog prostora.

Definicija 1.2.10. Okolinom skupa $A \subset X$ nazivamo svaki skup iz X koji sadrži neki otvoren skup u kome leži skup A.

Definicija 1.2.11. Za tačku $u \in A$ kažemo da je *unutrašnja* tačka skupa $A \subset X$ ako postoji okolina tačke u koja je podskup skupa A. Skup svih unutrašnjih tačaka skupa A čini njegovu *unutrašnjost*, u oznaci $\operatorname{int} A$.

Prema tome, tačka u je unutrašnja tačka skupa $A \subset X$ ako je i sâm skup A jedna njena okolina. Očigledno je $\operatorname{int} A \subseteq A$.

Definicija 1.2.12. Tačka $a \in X$ je *tačka nagomilavanja* skupa $A \subset X$ ako svaka okolina tačke a sadrži bar jednu tačku iz A, različitu od tačke a.

Dakle, tačka nagomilavanja skupa A ne mora pripadati skupu A.

Definicija 1.2.13. Tačka $a \in A \subset X$ je *izolovana tačka* skupa A ako postoji okolina tačke a u kojoj, sem tačke a, nema drugih tačaka iz A.

Definicija 1.2.14. Tačka $u \in X$ je *adherentna tačka* skupa A ako u svakoj njenoj okolini ima tačaka iz skupa A. Skup svih adherentnih tačaka skupa A čini *adherenciju* skupa A, u oznaci $\tilde{A}$.

Napomenimo da adherentna tačka skupa A ne mora da pripada skupu A. Treba uočiti i da su izolovana tačka i tačka nagomilavanja skupa A njegove adherentne tačke. Takođe, očigledno je da važi $A \subseteq \tilde{A}$.

Definicija 1.2.15. Za familiju $\mathcal{T}$ svih otvorenih skupova metričkog prostora (X,d) kažemo da je *topologija* tog prostora, koja je indukovana metrikom d.

U smislu ove definicije, sâm skup X i prazan skup su otvoreni skupovi. Na osnovu prethodnog zaključujemo sledeće dve osobine otvorenih skupova: (a) unija konačno ili beskonačno mnogo otvorenih skupova iz $\mathcal{T}$ pripada $\mathcal{T}$; (b) presek konačno mnogo otvorenih skupova iz $\mathcal{T}$ pripada $\mathcal{T}$.

Definicija 1.2.16. Neka je X metrički prostor. Skup $A \subset X$ je *zatvoren* skup, ako je njegov komplement A'_X, u odnosu na skup X, otvoren skup.

Iz prethodnih definicija i činjenice da su prazan skup i sâm skup X otvoreni skupovi, sleduje da su prazan skup i čitav skup X takođe i zatvoreni skupovi.

Topologija se može uvesti nezavisno od metrike.

Definicija 1.2.17. *Topološki prostor* je skup snabdeven *topologijom* $\mathcal{T}$, koja je familija podskupova – *otvorenih skupova* – sa osobinama:

1° presek bilo koja dva otvorena skupa je otvoren skup;

2° unija bilo koje kolekcije otvorenih skupova je otvoren skup;

3° prazan skup i ceo prostor su otvoreni skupovi.

Element topološkog prostora se naziva *tačka.*

Na kraju ovog odeljka navodimo definicije koneksnog i separabilnog prostora.

Definicija 1.2.18. Za skup A kažemo da je *svuda gust* u skupu B ako je svaka tačka skupa B adherentna tačka skupa A, tj. ako važi $B \subset \tilde{A}$.

Definicija 1.2.19. Za prostor X kažemo da je *koneksan* ili *povezan* ako se ne može predstaviti kao unija dva neprazna disjunktna otvorena skupa iz X.

Definicija 1.2.20. Za metrički prostor X kažemo da je *separabilan* ako u njemu postoji najviše prebrojivi svuda gust skup tačaka.

2.1.3 Konvergencija nizova u metričkom prostoru

Niz je jedan od osnovnih pojmova koji se uvodi već u početnim kursevima matematičke analize. U numeričkoj analizi i teoriji aproksimacija uloga nizova je ogromna, posebno u konstrukciji iterativnih procesa. U ovom odeljku dajemo osnovne karakteristike nizova u metričkom prostoru.

Definicija 1.3.1. Za niz $\{u_n\}_{n\in\mathbb{N}}$ u metričkom prostoru (X,d) kažemo da je *konvergentan* ka elementu $a \in X$ ako za svako $\varepsilon > 0$ postoji prirodan broj n_0 takav da je $d(u_n,a) < \varepsilon$ kad god je $n > n_0$. Tada pišemo $\lim\limits_{n\to+\infty} u_n = a$ i za tačku a kažemo da je *granica niza*.

Dakle, kod konvergentnog niza rastojanje $d(u_n,a) \to 0$ kada $n \to +\infty$. Jednostavno je dokazati da je kod konvergentnih nizova granica jedinstvena.

Uočimo sada jedan niz $\{n_k\}_{k\in\mathbb{N}}$ prirodnih brojeva n_1, n_2, ..., takav da je

$$n_1 < n_2 < \cdots .$$

Definicija 1.3.2. Za svaki niz $\{u_{n_k}\}_{k\in\mathbb{N}}$ kažemo da je *delimični niz* ili *podniz* niza $\{u_n\}_{n\in\mathbb{N}}$.

Definicija 1.3.3. Ako je delimični niz $\{u_{n_k}\}_{k\in\mathbb{N}}$ konvergentan, za njegovu granicu

$$\lim_{k\to+\infty} u_{n_k} = c$$

kažemo da je *delimična granica* niza $\{u_n\}_{n\in\mathbb{N}}$.

U skladu sa definicijom 1.2.12, delimična granica c je *tačka nagomilavanja* niza $\{u_n\}_{n\in\mathbb{N}}$. Nije teško pokazati da je skup delimičnih granica niza $\{u_n\}_{n\in\mathbb{N}}$ u metričkom prostoru (X,d) zatvoren skup (videti, na primer, [21, str. 62]). Inače, niz $\{u_n\}_{n\in\mathbb{N}}$ konvergira ka a ako i samo ako svi njegovi delimični nizovi konvergiraju ka istoj tački a.

Definicija 1.3.4. Za niz $\{u_n\}_{n\in\mathbb{N}}$ u metričkom prostoru (X,d) kažemo da je CAUCHY*ev niz* ako za svako $\varepsilon > 0$ postoji prirodan broj n_0 takav da je $d(u_n,u_m) < \varepsilon$ kad god su $n,m > n_0$.

Primetimo da je svaki konvergentan niz CAUCHYev niz. Zaista, ako niz $\{u_n\}_{n\in\mathbb{N}}$ konvergira ka granici a, tj. $\lim_{n\to+\infty} u_n = a$, tada je, na osnovu relacije trougla,

$$d(u_n,u_m) \le d(u_n,a) + d(a,u_m),$$

odakle zaključujemo da $d(u_n,u_m) \to 0$ kada $n,m \to +\infty$.

U prostoru $(\mathbb{R},d)$, sa $d(x,y) = |x-y|$ (videti primer 1.2.1) važi i obrnuto tvrđenje. Dakle, niz $\{x_n\}_{n\in\mathbb{N}}$ u $\mathbb{R}$ je konvergentan *ako i samo ako* je CAUCHYev niz (za dokaz videti, na primer, [21, str. 66]). Međutim, u opštem slučaju, ako je niz CAUCHYev ne znači da je i konvergentan. Na primer, ako u prostoru racionalnih brojeva $\mathbb{Q}$, sa uobičajenom metrikom $d(x,y) = |x-y|$, posmatramo niz racionalnih brojeva $\{x_n\}_{n\in\mathbb{N}}$, dobijenih iz decimalnog razvoja broja $\sqrt{2}$ zadržavanjem prvih n decimala (dakle, $x_1 = 1.4$, $x_2 = 1.41$, $x_3 = 1.414$, $x_4 = 1.4142$, itd.), jednostavno zaključujemo da je ovaj niz CAUCHYev, s obzirom da je za svako $m > n > 0$,

$$|x_n - x_m| < \varepsilon = 10^{-n},$$

kao i da je njegova granica broj $\sqrt{2} \notin \mathbb{Q}$.

Za mnoge primene, posebno u iterativnim procesima, pogodno je raditi sa prostorima u kojima svaki CAUCHYev niz konvergira.

Definicija 1.3.5. Ako je svaki CAUCHYev niz u metričkom prostoru (X,d) konvergentan, tada za prostor (X,d) kažemo da je *kompletan*.

Primer 1.3.1. Razmotrićemo metričke prostore $(C[a,b],d)$ i $(C[a,b],d_1)$, sa uniformnom i integralnom metrikom d i d_1, datim sa (1.2.4) i (1.2.5), respektivno.

1° *Prostor* $(C[a,b],d)$ *je kompletan*. Pretpostavimo da je niz neprekidnih funkcija $\{u_n(t)\}_{n\in\mathbb{N}}$ CAUCHYev niz u prostoru $(C[a,b],d)$, tj. da za svako $\varepsilon>0$ postoji $n_0\in\mathbb{N}$, tako da je za svako $m>n>n_0$,

$$|u_m(t)-u_n(t)|\leq \max_{a\leq t\leq b}|u_m(t)-u_n(t)| = d(u_m,u_n)<\varepsilon \quad (t\in[a,b]). \tag{1.3.1}$$

Istovremeno (1.3.1) znači da je za svako dato $t\in[a,b]$, niz $\{u_n(t)\}_{n\in\mathbb{N}}$ CAUCHYev niz u (kompletnom) prostoru $\mathbb{R}$ sa standardnom metrikom (videti primer 1.2.1), tako da postoji jedinstvena granična vrednost

$$\lim_{n\to+\infty} u_n(t)=u(t)\in\mathbb{R}$$

za svako $t\in[a,b]$, tj. da je konvergencija uniformna na $[a,b]$. Iz činjenice da je granica niza neprekidnih funkcija sa uniformnom konvergencijom na $[a,b]$ neprekidna funkcija, zaključujemo da $u\in C[a,b]$. Dakle,

$$\lim_{n\to+\infty} d(u_n,u)=0,$$

tj. prostor $(C[a,b],d)$ je kompletan.

2° *Prostor* $(C[a,b],d_1)$ *nije kompletan*. Ne umanjujući opštost razmatraćemo prostor $C[0,1]$ sa odgovarajućom integralnom normom. Definišimo niz funkcija $\{u_n(t)\}_{n\in\mathbb{N}}$ pomoću

$$u_n(t)=\begin{cases} 0, & 0\leq t<\dfrac{1}{n},\\[2mm] \dfrac{1}{\sqrt{t}}, & \dfrac{1}{n}\leq t\leq 1,\end{cases}$$

i dokažimo da je u datom prostoru ovaj niz CAUCHYev, ali, što je evidentno, njegova granica

$$u(t)=\lim_{n\to+\infty} u_n(t)=\begin{cases} 0, & t=0,\\[2mm] \dfrac{1}{\sqrt{t}}, & 0<t\leq 1,\end{cases}$$

ne pripada $C[0,1]$, što znači da prostor nije kompletan.

Zaista, ako pretpostavimo da je $m > n > n_0$ (tj. $0 < 1/m < 1/n < 1$), imamo

$$d_1(u_n,u_m) = \int_0^1 |u_m(t) - u_n(t)|\,\mathrm{d}t = \int_{1/m}^{1/n} \frac{1}{\sqrt{t}}\,\mathrm{d}t = 2\left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{m}}\right) < \frac{2}{\sqrt{n}}.$$

Dakle, za dato $\varepsilon > 0$, izborom $n_0 = [4/\varepsilon^2] + 1$, zaključujemo da je $d_1(u_n,u_m) < 2/\sqrt{n} < 2/\sqrt{n_0} < \varepsilon$ za svako $m > n > n_0$, tj. da je niz funkcija $\{u_n(t)\}_{n\in\mathbb{N}}$ CAUCHYev niz. $\triangle$

2.1.4 Normirani prostor i BANACHov prostor

Veoma značajna klasa metričkih prostora, koja pored topološke strukture ima i strukturu linearnog prostora, su normirani prostori.

Definicija 1.4.1. Linearni prostor X (nad poljem $\mathbb{K}$) je normiran ako postoji nenegativna funkcija $u \mapsto \|u\|$ definisana za svako $u \in X$, koju nazivamo norma od u, takva da je

1° $\|u\| = 0 \Leftrightarrow u = \theta$ (definisanost);

2° $\|cu\| = |c| \cdot \|u\|$ (homogenost);

3° $\|u+v\| \le \|u\| + \|v\|$ (relacija trougla),

gde su $u, v \in X$ i $c \in \mathbb{K}$.

Normirani prostor je metrički ako metriku uvedemo pomoću

$$d(u,v) = \|u - v\|. \tag{1.4.1}$$

Primer 1.4.1. Vektorski prostor $\mathbb{R}^n$ (videti primer 1.1.1) se može normirati uvođenjem norme pomoću

$$\|\boldsymbol{x}\|_p = \begin{cases} \left(\sum_{k=1}^n |x_k|^p\right)^{1/p} & (1 \le p < +\infty), \\ \max\limits_k |x_k| & (p = +\infty). \end{cases} \tag{1.4.2}$$

Od normi (1.4.2) najčešće se koriste norme za $p = 1$ i $p = 2$, tj.

$$\|\boldsymbol{x}\|_1 = \sum_{k=1}^n |x_k| \qquad \text{i} \qquad \|\boldsymbol{x}\|_2 = \left(\sum_{k=1}^n |x_k|^2\right)^{1/2},$$

kao i norma $\|\boldsymbol{x}\|_\infty$. Norma $\|\boldsymbol{x}\|_2$ je poznata kao euklidska norma i često se označava sa $\|\boldsymbol{x}\|_E$. $\triangle$

Primer 1.4.2. Prostor $C[a,b]$ se može normirati, na primer, uvođenjem norme pomoću

$$\|u\| = \|u\|_{[a,b]} = \max_{a\le t\le b} |u(t)| \tag{1.4.3}$$

ili pomoću

$$\|u\| = \|u\|_1 = \int_a^b |u(t)|\,\mathrm{d}t. \tag{1.4.4}$$

△

Primer 1.4.3. Prostor $L^p(a,b)$, $p \ge 1$, se normira uvođenjem norme pomoću

$$\|u\| = \|u\|_p = \left(\int_a^b |u(t)|^p\,\mathrm{d}t\right)^{1/p}. \tag{1.4.5}$$

Slučaj $p = +\infty$, tj. prostor $L^\infty(a,b)$ se normira uzimanjem tzv. esencijalnog supremuma,

$$\|u\| = \|u\|_\infty = \operatorname*{ess\,sup}_{a\le t\le b} |u(t)|.$$

△

Primer 1.4.4. Prostor nizova ℓ^p se normira pomoću

$$\|u\| = \left(\sum_{k=1}^{+\infty} |x_k|^p\right)^{1/p}, \tag{1.4.6}$$

gde je $u = \{x_k\}_{k\in\mathbb{N}}$. △

U skladu sa metrikom (1.4.1), u normiranom prostoru se govori o tzv. *konvergenciji po normi*.

Definicija 1.4.2. Neka je $\{u_n\}_{n\in\mathbb{N}}$ niz tačaka u normiranom prostoru X i neka je $u \in X$ takvo da je $\lim_{n\to+\infty} \|u_n - u\| = 0$. Tada kažemo da ovaj niz konvergira po normi ka tački u.

U tom slučaju, niz $\{u_n\}_{n\in\mathbb{N}}$ se karakteriše kao CAUCHYev niz ako je

$$\lim_{n,m\to+\infty} \|u_n - u_m\| = 0$$

i normirani vektorski prostor je kompletan ako u njemu svaki CAUCHYev niz konvergira.

Definicija 1.4.3. Kompletan normirani prostor naziva se BANACHov[87] prostor.

Da li je normirani prostor kompletan ili nije, zavisi od uvedene norme. Tako, na primer, prostori ℓ^p i $L^p(a,b)$ su kompletni u odnosu na norme (1.4.5) i (1.4.6) respektivno, dok je prostor $C[a,b]$ kompletan u odnosu na normu (1.4.3), a nije kompletan u odnosu na (1.4.4) (videti primer 1.3.1).

2.1.5 HILBERTov prostor

Definicija 1.5.1. Vektorski prostor X nad poljem kompleksnih brojeva ($\mathbb{K}=\mathbb{C}$) naziva se prostor sa unutrašnjim (skalarnim) proizvodom ili unitaran prostor, ako postoji funkcija $(u,v)\colon\ X^2\to\mathbb{C}$ koja zadovoljava sledeće uslove

$1^\circ\ \ (u,u)\geq 0$;

$2^\circ\ \ (u,u)=0\ \Leftrightarrow\ u=\theta$;

$3^\circ\ \ (u+v,w)=(u,w)+(v,w)$;

$4^\circ\ \ (cu,v)=c(u,v)$

$5^\circ\ \ (u,v)=\overline{(v,u)}$

za svako $u,v,w\in X$ i $c\in\mathbb{C}$.

Funkcija (u,v) se naziva *skalarni* ili *unutrašnji proizvod*.

Ako je X realni vektorski prostor, tada je skalarni proizvod $(u,v)\colon X^2\to\mathbb{R}$ takav da uslov 5° u prethodnoj definiciji postaje tzv. *uslov simetrije*

$5^\#\ \ (u,v)=(v,u)$.

Inače, uslov 5° je poznat kao *hermitska simetrija*.

Teorema 1.5.1. *Za skalarni proizvod važi*

(a) $(u,cv)=\overline{c}(u,v)$;

(b) $(u,v_1+v_2)=(u,v_1)+(u,v_2)$;

(c) $|(u,v)|^2\leq(u,u)(v,v)$.

Dokaz. Tvrđenja (a) i (b) se jednostavno dokazuju. Da bismo dokazali tvrđenje (c), koje je poznato kao CAUCHY-SCHWARZ[88]-BUNIAKOWSKYeva[89] nejednakost, uzmimo tačku $w=u+t(u,v)v$, gde je t realno i $u,v\in X$. Kako na osnovu 1° (iz definicije 1.5.1) važi

[87] STEFAN BANACH (1892 – 1945), poznati poljski matematičar.
[88] KARL HERMANN AMANDUS SCHWARZ (1843 – 1921), nemački matematičar.
[89] VIKTOR JAKOVLEVIČ BUNIAKOWSKY (1804 – 1889), ruski matematičar.

$$(w,w) = (u+t(u,v)v, u+t(u,v)v) \geq 0,$$

korišćenjem osobina 3°, 4°, 5° i (a) zaključujemo da je

$$(u,u) + 2|(u,v)|^2 t + |(u,v)|^2 (u,u)(v,v)t^2 \geq 0\,,$$

odakle sleduje da diskriminanta D dobijenog kvadratnog trinoma mora biti manja ili jednaka nuli, tj.

$$\frac{D}{4} = |(u,v)|^4 - |(u,v)|^2 (u,u)(v,v) \leq 0.$$

Iz poslednje nejednakosti sleduje nejednakost (c). □

Unitaran vektorski prostor može se normirati uvođenjem

$$\|u\| = \sqrt{(u,u)}, \tag{1.5.1}$$

s obzirom na to da funkcija $u \mapsto \sqrt{(u,u)}$ ispunjava sve uslove definicije 1.4.1.

Definicija 1.5.2. Unitaran vektorski prostor sa normom (1.5.1) naziva se pred-HILBERTov prostor. Ukoliko je ovaj prostor kompletan naziva se HILBERTov.

Primer 1.5.1. Vektorski prostor $\mathbb{R}^n$ postaje HILBERTov prostor ako se skalarni proizvod uvede pomoću

$$(\boldsymbol{x},\boldsymbol{y}) = \sum_{k=1}^{n} x_k y_k = \boldsymbol{y}^T \boldsymbol{x}.$$

Slično, kompleksan vektorski prostor $\mathbb{C}^n$ postaje HILBERTov sa skalarnim proizvodom

$$(\boldsymbol{x},\boldsymbol{y}) = \sum_{k=1}^{n} x_k \overline{y}_k = \boldsymbol{y}^* \boldsymbol{x},$$

pri čemu $\boldsymbol{y}^*$ označava vektor koji se dobija transponovanjem vektora $\boldsymbol{y}$ i konjugovanjem njegovih komponenata. Norma koja izvire iz skalarnog proizvoda, u ovom slučaju je data sa

$$\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x},\boldsymbol{x})} = \left(\sum_{k=1}^{n} |x_k|^2 \right)^{1/2},$$

što u stvari predstavlja euklidsku normu $\|\boldsymbol{x}\|_E$ (videti primer 1.4.1).

CAUCHY-SCHWARZ-BUNIAKOWSKYeva nejednakost u $\mathbb{C}^n$ ima oblik

$$\left| \sum_{k=1}^{n} x_k \overline{y}_k \right|^2 \leq \left(\sum_{k=1}^{n} |x_k|^2 \right) \left(\sum_{k=1}^{n} |y_k|^2 \right).$$

$\triangle$

Primer 1.5.2. Ako se u prostor ℓ^2 uvede skalarni proizvod pomoću

$$(u,v) = \sum_{k=1}^{+\infty} x_k \overline{y}_k \qquad \Big(u = \{x_k\}_{k\in\mathbb{N}},\ v = \{y_k\}_{k\in\mathbb{N}} \Big),$$

dobija se HILBERTov prostor. $\triangle$

Primer 1.5.3. Sa skalarnim proizvodom

$$(u,v) = \int_a^b u(t)\overline{v(t)}\,\mathrm{d}t, \tag{1.5.2}$$

prostor $L^2(a,b)$ postaje HILBERTov. U slučaju kada su elementi iz $L^2(a,b)$ samo realne funkcije, (1.5.2) se svodi na

$$(u,v) = \int_a^b u(t)v(t)\,\mathrm{d}t.$$

Umesto skalarnog proizvoda (1.5.2) može se uzeti opštiji skalarni proizvod, na primer,

$$(u,v) = \int_a^b w(t)u(t)\overline{v(t)}\,\mathrm{d}t, \tag{1.5.3}$$

gde je $w\colon (a,b) \to \mathbb{R}^+$ data težinska funkcija[90]. Odgovarajući HILBERTov prostor označavamo sa $L^2_w(a,b)$. Norma koja izvire iz skalarnog proizvoda uvedenog na ovaj način je data sa

$$\|u\| = \|u\|_{2,w} = \left(\int_a^b w(t)|u(t)|^2\,\mathrm{d}t \right)^{1/2}.$$

U ovom slučaju nejednakost (c) postaje

$$\left| \int_a^b w(t)u(t)v(t)\,\mathrm{d}t \right|^2 \leq \left(\int_a^b w(t)|u(t)|^2\,\mathrm{d}t \right) \left(\int_a^b w(t)|v(t)|^2\,\mathrm{d}t \right). \quad \triangle$$

[90] Precizno definisanje težinske funkcije biće dato kod razmatranja ortogonalnih polinoma u posebnoj knjizi iz ove serije.

2.1.6 Ortogonalni sistemi u HILBERTovom prostoru

Orogonalni sistemi, a posebno ortogonalni polinomi, zauzimaju značajno mesto u rešavanju mnogih problema u različitim naučnim oblastima. Njihova primena je posebno izražena u teoriji aproksimacija, interpolacionim procesima, kvadraturnim procesima, itd. (za detalje videti, na primer, našu monografiju pisanu zajedno sa G. MASTROIANNIjem[91] [17]).

Definicija 1.6.1. Skup vektora $\{u_k\}_{k\in I}$ u HILBERTovom prostoru X obrazuje *ortogonalan sistem* ako je

$$(\forall n,k \in I) \qquad (u_n,u_k) = \delta_{n,k}\|u_k\|^2.$$

gde je $\delta_{n,k}$ – KRONECKERova[92] delta i $\|u_k\| = \sqrt{(u_k,u_k)}$.

Skup indeksa I može biti konačan, prebrojiv ili neprebrojiv.

Ukoliko za svako $k \in I$ imamo $\|u_k\| = 1$, kažemo da skup vektora $\{u_k\}_{k\in I}$ u X obrazuje *ortonomiran sistem* i takav sistem označavaćemo sa $\{u_k^*\}_{k\in I}$.

Definicija 1.6.2. Ortonomiran sistem $\{u_k^*\}_{k\in I}$ je *potpun* u X ako nije pravi deo nekog ortonomiranog sistema.

Teorema 1.6.1. *U svakom* HILBERT*ovom prostoru* X $(\neq \{0\})$ *postoji ortonomirani sistem.*

Definicija 1.6.3. Koeficijenti a_k u razvoju $u = \sum_{k\in I} a_k u_k^*$ se nazivaju FOURIERovi koeficijenti, a sâm razvoj FOURIERov razvoj ili FOURIERov red.

Teorema 1.6.2. *Neka su* a_k FOURIER*ovi koeficijenti vektora* $u \in X$ *u odnosu na ortonormirani sistem* $\{u_k^*\}_{k\in I}$. *Tada su iskazi*

1° $\{u_k^*\}_{k\in I}$ *je potpun sistem u* X*;*

2° $u = \sum_{k\in I} a_k u_k^*$ *za svako* $u \in X$*,*

međusobno ekvivalentni.

Zbog ekvivalentnosti iskaza 1° i 2° u prethodnoj teoremi, potpun ortonormirani sistem nazivamo i *ortonormirana baza* prostora. Slično, potpun ortogonalni sistem nazivamo *ortogonalna baza* prostora.

U daljem tekstu navodimo dva jednostavna i veoma važna ortogonalna sistema.

[91] GIUSEPPE MASTROIANNI (1939 –), italijanski matematičar, poznat u teoriji interpolacionih procesa i numeričkih kvadratura. Sada je profesor emeritus na Univerzitetu Basilicata u Potenci (Department of Mathematics, Computer Sciences and Economics).

[92] LEOPOLD KRONECKER (1823 – 1891), nemački matematičar.

Trigonometrijski sistem. Najjednostavniji ortogonalni sistem je *trigonometrijski sistem*,

$$T = \{1, \cos t, \sin t, \cos 2t, \sin 2t, \ldots, \cos nt, \sin nt, \ldots\}. \tag{1.6.1}$$

Korišćenjem formula

$$\begin{cases} \sin nt \cos mt = \dfrac{1}{2}\big[\sin(n+m)t + \sin(n-m)t\big], \\ \sin nt \sin mt = \dfrac{1}{2}\big[\cos(n-m)t - \cos(n+m)t\big], \\ \cos nt \cos mt = \dfrac{1}{2}\big[\cos(n+m)t + \cos(n-m)t\big], \end{cases} \tag{1.6.2}$$

lako je dokazati sledeće relacije ortogonalnosti

$$\begin{cases} \displaystyle\int_0^{2\pi} \sin nt \cos mt \,\mathrm{d}t = 0, \\ \displaystyle\int_0^{2\pi} \cos nt \cos mt \,\mathrm{d}t = \begin{cases} 0, & n \neq m, \\ \pi, & n = m \neq 0, \\ 2\pi, & n = m = 0, \end{cases} \\ \displaystyle\int_0^{2\pi} \sin nt \sin mt \,\mathrm{d}t = \begin{cases} 0, & n \neq m, \\ \pi, & n = m \neq 0, \\ 0, & n = m = 0. \end{cases} \end{cases} \tag{1.6.3}$$

Ovo znači da je trigonometrijski sistem (1.6.1) ortogonalan u HILBERTovom prostoru $L^2(\mathbb{T})$ sa unutrašnjim proizvodom

$$(u, v) = \int_{\mathbb{T}} u(t)v(t)\,\mathrm{d}t = \int_0^{2\pi} u(t)v(t)\,\mathrm{d}t, \tag{1.6.4}$$

gde je $\mathbb{T}$ jedinična kružnica. Sistem T je potpun u $L^2(\mathbb{T})$, tako da se svaka funkcija $f \in L^2(\mathbb{T})$ može predstaviti pomoću FOURIERovog reda

$$\frac{1}{2}a_0 + \sum_{k=1}^{+\infty} (a_k \cos kt + b_k \sin kt), \tag{1.6.5}$$

sa FOURIERovim koeficijentima

$$a_k = \frac{1}{\pi}\int_0^{2\pi} f(t)\cos kt \,\mathrm{d}t, \quad b_k = \frac{1}{\pi}\int_0^{2\pi} f(t)\sin kt \,\mathrm{d}t. \tag{1.6.6}$$

Napomenimo da se parcijalne sume FOURIERovog reda (1.6.5), tj.

$$S_n f(x) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx), \tag{1.6.7}$$

često koriste u aproksimaciji 2π-periodičnih integrabilnih funkcija $f \in L^1(\mathbb{T})$ pomoću trigonometrijskih polinoma. Korišćenjem formula (1.6.6) za FOURIERove koeficijente a_k i b_k, parcijalne sume (1.6.7) se mogu predstaviti u obliku

$$\begin{aligned} S_n f(x) &= \frac{1}{2} a_0 + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx) \\ &= \frac{1}{2\pi} \int_0^{2\pi} f(t)\,\mathrm{d}t + \frac{1}{\pi} \sum_{k=1}^{n} \int_0^{2\pi} f(t) \big(\cos kt \cos kx + \sin kt \sin kx\big)\,\mathrm{d}t, \end{aligned}$$

tj.

$$S_n f(x) = \frac{1}{\pi} \int_0^{2\pi} f(t) \Big[\frac{1}{2} + \sum_{k=1}^{n} \cos k(x-t)\Big]\,\mathrm{d}t = \frac{1}{\pi} \int_0^{2\pi} D_n(x-t) f(t)\,\mathrm{d}t, \tag{1.6.8}$$

gde je D_n takozvano DIRICHLETovo[93] jezgro definisano pomoću

$$D_n(\theta) = \frac{1}{2} + \sum_{k=1}^{n} \cos k\theta. \tag{1.6.9}$$

Kao što možemo videti DIRICHLETovo jezgro je paran trigonometrijski polinom stepena n i za $\theta \neq 2\nu\pi$ $(\nu \in \mathbb{Z})$ jezgro se može predstaviti u obliku

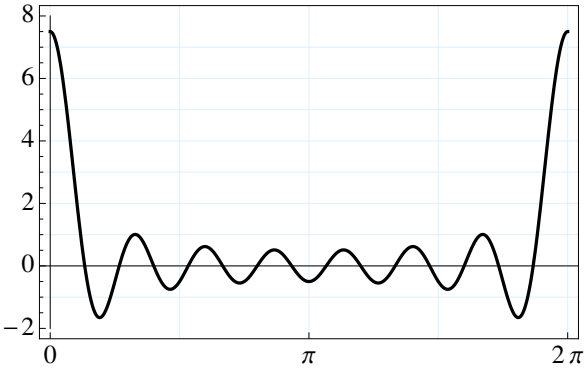
$$D_n(\theta) = \mathrm{Re}\left\{ \frac{1}{2} + \sum_{k=1}^{n} \mathrm{e}^{\mathrm{i}k\theta} \right\} = \frac{\sin\big((2n+1)\theta/2\big)}{2\sin(\theta/2)}.$$

Kako su njegove nule u intervalu $[0, 2\pi)$ date sa $\theta_k = 2k\pi/(2n+1)$, $k = 1, \ldots, 2n$, DIRICHLETovo jezgro se može izraziti i u obliku (videti [17, str. 25])

$$D_n(\theta) = \frac{2n+1}{2} \prod_{k=1}^{2n} \frac{\sin\dfrac{\theta - \theta_k}{2}}{\sin\dfrac{\theta_k}{2}}.$$

Grafik DIRICHLETovog jezgra za $n = 7$ na intervalu $(0, 2\pi)$ je prikazan na slici 1.6.1.

[93] PETER GUSTAV LEJEUNE DIRICHLET (1805 – 1859), francuski matematičar.



Slika 1.6.1. DIRICHLETovo jezgro $D_7(\theta)$

S obzirom na osobine 2π-periodičnih integrabilnih funkcija $f, g \in L^1$, da za svako $x \in \mathbb{R}$, važe jednakosti

$$\int_x^{x+2\pi} f(t)\,\mathrm{d}t = \int_0^{2\pi} f(t)\,\mathrm{d}t = \int_0^{2\pi} f(t+x)\,\mathrm{d}t$$

i

$$\int_0^{2\pi} g(x-t)f(t)\,\mathrm{d}t = \int_0^{2\pi} g(t)f(x-t)\,\mathrm{d}t,$$

parcijalne sume (1.6.8) se mogu predstaviti u ekvivalentnim oblicima

$$S_n f(x) = \frac{1}{\pi}\int_0^{2\pi} D_n(t)f(x-t)\,\mathrm{d}t = \frac{1}{\pi}\int_0^{2\pi} D_n(t)f(x+t)\,\mathrm{d}t$$

i

$$S_n f(x) = \frac{1}{\pi}\int_0^{\pi} D_n(t)\big[f(x+t)+f(x-t)\big]\,dt.$$

ČEBIŠEVljevi polinomi. Dva jednostavna, ali veoma važna trigonometrijska polinoma su

$$\theta \mapsto \cos n\theta \qquad \text{i} \qquad \theta \mapsto \frac{\sin(n+1)\theta}{\sin\theta}.$$

Oni se mogu izraziti u obliku algebarskih polinoma stepena n u zavisnosti od $\cos\theta$. Naime, ako stavimo $x = \cos\theta$, dobijamo dobro poznate ČEBIŠEVljeve polinome prve i druge vrste,

$$T_n(x) = T_n(\cos\theta) = \cos n\theta \quad \text{i} \quad U_n(x) = U_n(\cos\theta) = \frac{\sin(n+1)\theta}{\sin\theta},$$

respektivno. Njihove algebarske reprezentacije za $-1 \le x \le 1$ su

$$T_n(x) = \cos(n\arccos x) \quad \text{i} \quad U_n(x) = \frac{\sin\big((n+1)\arccos x\big)}{\sqrt{1-x^2}}. \tag{1.6.10}$$

Lako je videti da su $T_0(x) = 1$, $T_1(x) = x$ i $U_0(x) = 1$, $U_1(x) = 2x$. Takođe, korišćenjem treće trigonometrijske jednakosti iz (1.6.2), sa $m = 1$ i $t = \theta = \arccos x$, zaključujemo da polinomi T_n zadovoljavaju rekurentnu relaciju

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n = 1,2,\dots. \tag{1.6.11}$$

Ista rekurentna relacija važi i za polinome druge vrste, tj.

$$U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x), \qquad n = 1,2,\dots.$$

Startujući sa $T_0(x) = 1$ i $T_1(x) = x$ ili sa $U_0(x) = 1$ i $U_1(x) = 2x$, oba niza polinoma $\{T_n(x)\}_{n=0}^{+\infty}$ i $\{U_n(x)\}_{n=0}^{+\infty}$ mogu se jednostavno generisati. Na primer, za $n = 0,1,\dots,6$, dobijamo

$$\begin{aligned}
&T_0(x) = 1, && U_0(x) = 1,\\
&T_1(x) = x, && U_1(x) = 2x,\\
&T_2(x) = 2x^2 - 1, && U_2(x) = 4x^2 - 1,\\
&T_3(x) = 4x^3 - 3x, && U_3(x) = 8x^3 - 4x,\\
&T_4(x) = 8x^4 - 8x^2 + 1, && U_4(x) = 16x^4 - 12x^2 + 1,\\
&T_5(x) = 16x^5 - 20x^3 + 5x, && U_5(x) = 32x^5 - 32x^3 + 6x,\\
&T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1, && U_6(x) = 64x^6 - 80x^4 + 24x^2 - 1.
\end{aligned}$$

Relacija (1.6.11) je poznata kao *troclana rekurentna relacija*. Za detalje u vezi sa takvim relacijama videti odeljak 1.3.2.

Nizovi polinoma $\{T_n(x)\}_{n=0}^{+\infty}$ i $\{U_n(x)\}_{n=0}^{+\infty}$ su *ortogonalni* na $(-1,1)$ u odnosu na skalarni proizvod oblika (1.5.3), sa težinskim funkcijama

$$w_1(x) = \frac{1}{\sqrt{1-x^2}} \qquad \text{i} \qquad w_2(x) = \sqrt{1-x^2},$$

respektivno.

Da bismo dokazali ortogonalnost niza polinoma $\{T_n(x)\}_{n=0}^{+\infty}$ pođimo od druge relacije u (1.6.3), koja se može napisati u obliku

$$\int_0^{\pi} \cos nt \, \cos mt \, \mathrm{d}t = \begin{cases} 0, & n \neq m, \\ \pi/2, & n = m \neq 0, \\ \pi, & n = m = 0. \end{cases}$$

Posle uvođenja nove promenljive $x = \cos t$, ovaj integral se svodi na

$$\int_{-1}^{1} \cos(n \arccos x) \cos(m \arccos x) \frac{\mathrm{d}x}{\sqrt{1-x^2}} = \begin{cases} 0, & n \neq m, \\ \pi/2, & n = m \neq 0, \\ \pi, & n = m = 0, \end{cases}$$

tj.

$$\int_{-1}^{1} T_n(x) T_m(x) \frac{\mathrm{d}x}{\sqrt{1-x^2}} = \delta_{n,m} \|T_n\|^2,$$

gde je $\delta_{n,m}$ – KRONECKERova delta i $\|T_n\|^2$ je kvadrat norme ČEBIŠEVljevog polinoma T_n u HILBERTovom prostoru $L^2_{w_1}(-1,1)$, sa skalarnim proizvodom

$$(u,v) = (u,v)_{w_1} = \int_{-1}^{1} u(t) v(t) w_1(t) \, \mathrm{d}t.$$

Napomenimo da su $\|T_0\|^2 = \pi$ i $\|T_n\|^2 = \pi/2$, $n \geq 1$.

Ortogonalnost niza ČEBIŠEVljevih polinoma druge vrste $\{U_n(x)\}_{n=0}^{+\infty}$ može se dokazati na sličan način polazeći od treće relacije u (1.6.3). Tada dobijamo

$$\int_{-1}^{1} U_n(x) U_m(x) \sqrt{1-x^2} \, \mathrm{d}x = \delta_{n,m} \|U_n\|^2,$$

gde je $\|U_n\|^2 = \pi/2$ $(n \geq 0)$ kvadrat norme ČEBIŠEVljevog polinoma druge vrste U_n u HILBERTovom prostoru $L^2_{w_2}(-1,1)$.

Napomena 1.6.1. ČEBIŠEVljevi polinomi $T_n(x)$ i $U_n(x)$ su specijalni slučajevi takozvanih JACOBIevih polinoma $\{P_n^{(\alpha,\beta)}(x)\}_{n=0}^{+\infty}$, sa parametrima $\alpha, \beta > -1$, koji su ortogonalni u prostoru $L^2_{v^{\alpha,\beta}}(-1,1)$, sa skalarnim proizvodom

$$(u,v) = (u,v)_{v^{\alpha,\beta}} = \int_{-1}^{1} u(t) v(t) v^{\alpha,\beta}(t) \, \mathrm{d}t,$$

gde je težinska funkcija data sa[94]

$$w(t) = v^{\alpha,\beta}(t) = (1-t)^{\alpha}(1+t)^{\beta}, \quad \alpha, \beta > -1.$$

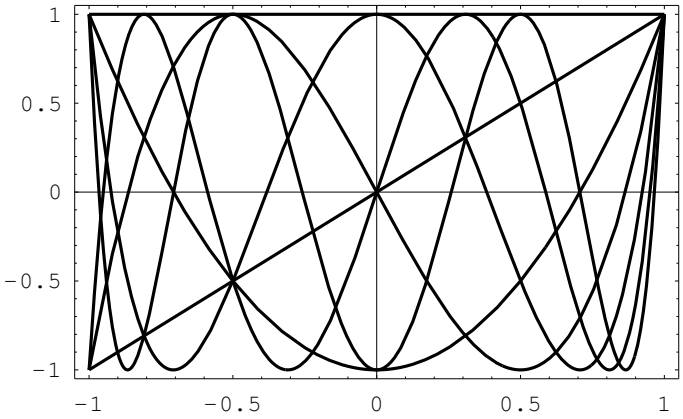
[94] Detaljna teorija ortogonalnosti biće razmatrana u posebnoj knjizi iz ove serije.

Eksplicitni izrazi za ČEBIŠEVljeve polinome prve i druge vrste su redom

$$T_n(x) = \frac{n}{2} \sum_{k=0}^{[n/2]} \frac{(-1)^k (n-k-1)!}{k!(n-2k)!} (2x)^{n-2k} \qquad (n \geq 1)$$

i

$$U_n(x) = \sum_{k=0}^{[n/2]} \frac{(-1)^k (n-k)!}{k!(n-2k)!} (2x)^{n-2k}.$$



Slika 1.6.2. Grafici polinoma $y = T_n(x)$, $-1 \leq x \leq 1$, za $n = 0, 1, \ldots, 6$

Grafici ČEBIŠEVljevih polinoma prve vrste $T_n(x)$, $n = 0, 1, \ldots, 6$, su prikazani na slici 1.6.2. S obzirom na to da je $T_n(x) = \cos(n \arccos x)$ za $-1 \leq x \leq 1$, zaključujemo da je $|T_n(x)| \leq 1$, $-1 \leq x \leq 1$, za svako $n \geq 0$ (videti sliku 1.6.2).

Na osnovu (1.6.10) nule polinoma $T_n(x)$ se mogu izraziti u eksplicitnom obliku

$$x_k = x_{n,k} = \cos \frac{(2k-1)\pi}{2n} \qquad (k = 1, \ldots, n). \tag{1.6.12}$$

Nule x_k date pomoću (1.6.12) su realne, različite i leže u $(-1, 1)$.

Drugi skup interesantnih tačaka su one gde je $T_n(x) = \pm 1$,

$$\xi_k = \xi_{n,k} = -\cos \frac{k\pi}{n} \qquad (k = 0, 1, \ldots, n). \tag{1.6.13}$$

Za tačke (1.6.13) kažemo da su *ekstremalne tačke* polinoma $T_n(x)$. Neke interesantne vrednosti za ČEBIŠEVljeve polinome prve vrste su:

$$T_n(\pm 1) = (\pm 1)^n, \qquad T_{2n}(0) = (-1)^n, \quad T_{2n+1}(0) = 0,$$

$$T_n'(\pm 1) = (\pm 1)^n n^2, \qquad T_n^{(k)}(1) = \frac{n^2(n^2-1)\cdots(n^2-(k-1)^2)}{(2k-1)!!}.$$

Za ČEBIŠEVljeve polinome druge vrste važe sledeće nejednakosti

$$|U_n(x)| \le n+1 \qquad \text{i} \qquad |\sqrt{1-x^2}\,U_n(x)| \le 1 \quad (-1 \le x \le 1).$$

Interesantne veze između ove dve vrste ČEBIŠEVljevih polinoma su

$$\frac{1}{2} + \sum_{k=1}^{n} T_{2k}(x) = \frac{1}{2} U_{2n}(x) \qquad \text{i} \qquad \sum_{k=1}^{n} T_{2k-1}(x) = \frac{1}{2} U_{2n-1}(x).$$

Na kraju ovog odeljka napomenimo da ČEBIŠEVljevi polinomi imaju veliku primenu u teoriji aproksimacija, ali i u mnogim drugim oblastima matematike, fizike, itd. Postoji veliki broj tzv. ekstremalnih problema u kojima se kao rešenja upravo pojavljuju ČEBIŠEVljevi polinomi (videti [19], kao i monografiju MILOVANOVIĆa, MITRINOVIĆa[95] i RASSIASa[96] [22], pisanu na engleskom jeziku). Navešćemo samo jedan od tih ekstremalnih problema koji se odnosi na skup svih realnih moničnih polinoma stepena n,

$$\widehat{\mathcal{P}}_n = \big\{ p \mid p(x) = x^n + \text{članovi nižeg stepena} \big\},$$

u prostoru $C[-1,1]$ sa uniformnom normom (1.4.3) (videti primer 1.4.2),

$$\|u\|_{[-1,1]} = \max_{-1 \le t \le 1} |u(t)|.$$

Teorema 1.6.3. *Među svim moničnim polinomima stepena $n \ge 1$, najmanju uniformnu normu na $[-1,1]$ ima monični* ČEBIŠEV*ljev polinom prve vrste $p^*(x) = \widehat{T}_n(x) = 2^{1-n} T_n(x)$, tj.*

$$\min_{p \in \widehat{\mathcal{P}}_n} \|p\|_{[-1,1]} = \|\widehat{T}_n\|_{[-1,1]} = 2^{1-n} \|T_n\|_{[-1,1]} = 2^{1-n}. \tag{1.6.14}$$

[95] DRAGOSLAV S. MITRINOVIĆ (1908 – 1995), poznati srpski matematičar, sa doprinosima u teoriji običnih diferencijalnih jednačina, teoriji nejednakosti, kompleksnoj analizi i specijalnim funkcijama. Autor je velikog broja knjiga i monografija objavljenih na srpskom i engleskom jeziku. Pod njegovim rukovodstvom autor ove knjige je radio doktorsku disertaciju koju je odbranio 1976. godine na Univerzitetu u Nišu.

[96] THEMISTOCLES M. RASSIAS (1951 –), grčki matematičar sa doprinosima u više oblasti matematike.

Dokaz. Primetimo najpre da su ekstremalne tačke ξ_k, date sa (1.6.13), uređene kao $-1 = \xi_0 < \xi_1 < \cdots < \xi_n = 1$, a zatim da je polinom $r(x) = \widehat{T}_n(x) - p(x)$ ($p \in \mathcal{P}_n$) stepena ne višeg od $n-1$. Primetimo da je $r(\xi_k) = (-1)^{n+k}2^{1-n} - p(\xi_k)$, $k = 0, 1, \ldots, n$.

Da bismo dokazali (1.6.14), tj. da je $\|p\|_{[-1,1]} \geq 2^{1-n}$ za svako $p \in \mathcal{P}_n$, pretpostavićemo suprotno, tj. da važi

$$Q = \|p\|_{[-1,1]} < \|\widehat{T}_n\|_\infty = 2^{1-n}. \tag{1.6.15}$$

Neka $x \in [-1, 1]$. Tada, prema (1.6.15), važi $-2^{1-n} < -Q \leq \pm p(x) \leq Q < 2^{1-n}$, odakle dobijamo dvostruku nejednakost $0 < 2^{1-n} - Q \leq 2^{1-n} \pm p(x)$, iz koje sledi da je

$$-2^{1-n} - p(x) \leq -2^{1-n} + Q < 0$$

i

$$2^{1-n} - p(x) \geq 2^{1-n} - Q > 0.$$

Ako za x izaberemo ekstremalne tačke ξ_k, leva strana u ovim nejednakostima postaje vrednost polinoma r u tim tačkama, tako da možemo zaključiti da polinom r ima naizmenično pozitivne i negativne vrednosti u ekstremalnim tačkama ξ_k, $k = 0, 1, \ldots, n$. Prema tome, polinom r mora imati bar n nula, što dovodi do kontradikcije, s obzirom da $r \in \mathcal{P}_{n-1}$. □

Proučavajući osobine polinoma sa najmanjim odstupanjem od date neprekidne funkcije, ČEBIŠEV [3] je praktično uveo polinome $T_n(x)$ i on se smatra začetnikom teorije aproksimacija ili konstruktivne teorije funkcija kako se to često naziva u ruskoj literaturi. Oznaku $T_n(x) = 2^{1-n}\cos(n \arccos x)$ je uveo BERNSTEIN[97] i ona je izvedena iz francuske transkripcije prezimena ČEBIŠEV, u obliku TCHEBYCHEFF. O ČEBIŠEVljevim polinomima je objavljeno dosta radova i knjiga (videti, na primer, [6], [16], [24], [25], [27]).

2.1.7 GRAM–SCHMIDTov postupak ortogonalizacije

Zbog jednostavnosti u primerima, ortogonalna baza je u prednosti nad algebarskom bazom u HILBERTovom prostoru X sa datim skalarnim proizvodom $(\cdot, \cdot)$. Zbog toga je interesantno proučiti postupak za konstrukciju ortogonalne baze.

[97] SERGEI NATANOVICH BERNSTEIN (1880 – 1968), poznati ruski matematičar, sa značajnim rezultatima u oblasti teorije aproksimacija, diferencijalnih jednačina, teoriji verovatnoće, itd.

Neka je dat najviše prebrojiv skup linearno nezavisnih vektora $U = \{g_0, g_1, \ldots\}$. Postupak kojim se ovom skupu vektora može pridružiti ortogonalni sistem vektora $\{u_0, u_1, \ldots\}$, tako da se lineali nad ovim skupovima poklapaju, poznat je kao GRAM[98]–SCHMIDTov[99] postupak ortogonalizacije i on se može iskazati na sledeći način.

Uzmimo najpre $u_0 = g_0$, a zatim u_1 predstavimo u obliku

$$u_1 = g_1 + \lambda_{1,0} u_0,$$

gde je $\lambda_{1,0}$ nepoznati parametar koji određujemo iz uslova da je vektor u_1 ortogonalan na u_0. Tada je

$$(u_1, u_0) = (g_1, u_0) + \lambda_{1,0}(u_0, u_0) = 0,$$

odakle sleduje

$$\lambda_{1,0} = -\frac{(g_1, u_0)}{(u_0, u_0)}.$$

Pretpostavimo sada da smo već konstruisali vektore $u_0, u_1, \ldots, u_{n-1}$. Vektor u_n predstavimo u obliku

$$u_n = g_n + \lambda_{n,0} u_0 + \lambda_{n,1} u_1 + \cdots + \lambda_{n,n-1} u_{n-1}.$$

Nepoznate parametre $\lambda_{n,k}$, $k = 0, 1, \ldots, n-1$, određujemo iz uslova ortogonalnosti vektora u_n prema svim prethodnim vektorima u_k, $k = 0, 1, \ldots, n-1$, tj.

$$(u_n, u_k) = (g_n, u_k) + \sum_{i=0}^{n-1} \lambda_{n,i}(u_i, u_k) = 0, \qquad k = 0, 1, \ldots, n-1.$$

Dakle, imamo

$$\lambda_{n,k} = -\frac{(g_n, u_k)}{(u_k, u_k)}, \qquad k = 0, 1, \ldots, n-1,$$

tj.

$$u_n = g_n - \sum_{k=0}^{n-1} \frac{(g_n, u_k)}{(u_k, u_k)} u_k, \qquad n = 1, 2, \ldots \ .$$

Odgovarajući ortonomirani sistem vektora je $S = \{\phi_0, \phi_1, \ldots\}$, gde su

[98] JØRGEN PEDERSEN GRAM (1850 – 1916), danski matematičar.
[99] ERHARD SCHMIDT (1876 – 1959), poznati nemački matematičar.

$$\phi_n = u_n^* = \frac{u_n}{\|u_n\|}, \qquad n = 0, 1, \dots .$$

Evidentno, ϕ_n je linearna kombinacija vektora (funkcija) g_0, g_1, …, g_n, takva da je $(\phi_n, \phi_k) = 0$ for $n \neq k$.

Korišćenjem vektora g_0, g_1, …, g_n i GRAMove matrice reda $n+1$,

$$G_{n+1} = \begin{bmatrix} (g_0,g_0) & (g_0,g_1) & \cdots & (g_0,g_n) \\ (g_1,g_0) & (g_1,g_1) & & (g_1,g_n) \\ \vdots & & & \\ (g_n,g_0) & (g_n,g_1) & & (g_n,g_n) \end{bmatrix},$$

za ortonormirane vektore ϕ_n moguće je dobiti eksplicitne formule (videti [17, str. 76]).

Teorema 1.7.1. *Ortonormirane funkcije ϕ_n su date sa*

$$\phi_n(z) = \frac{1}{\sqrt{\Delta_n \Delta_{n+1}}} \begin{vmatrix} (g_0,g_0) & (g_0,g_1) & \cdots & (g_0,g_{n-1}) & g_0(z) \\ (g_1,g_0) & (g_1,g_1) & & (g_1,g_{n-1}) & g_1(z) \\ \vdots & & & & \\ (g_n,g_0) & (g_n,g_1) & & (g_n,g_{n-1}) & g_n(z) \end{vmatrix}, \tag{1.7.1}$$

gde su $\Delta_n = \det G_n$ i $\Delta_0 = 1$.

Primetimo da je GRAMova matrica nesingularna. Naime, poznato je da je $\Delta_{n+1} = \det G_{n+1} \neq 0$ ako i samo ako je sistem vektora $\{g_0, g_1, g_2, \dots, g_n\}$ linearno nezavisan. Štaviše, može se dokazati da je $\det G_{n+1} > 0$.

Na kraju ovog odeljka primenućemo GRAM–SCHMIDTov postupak ortogonalizacije na konstrukciju ortogonalnih polinoma u prostoru $L^2_w(a,b)$, sa skalarnim proizvodom

$$(f,g) = \int_a^b w(x) f(x) g(x)\,\mathrm{d}x \qquad (f, g \in L^2_w(a,b)),$$

gde je $x \mapsto w(x)$ nenegativna težinska funkcija na (a,b).

Polazeći od monomialnog bazisa $U = \{1, x, x^2, \dots\}$ konstruisaćemo ortogonalni (polinomski) bazis $\{Q_n\}_{n\in\mathbb{N}_0}$. Lineal nad ovim bazisom je skup svih algebarskih polinoma, koji je svuda gust u $L^2_w(a,b)$. U zavisnosti od težinske funkcije w dobijaju se različite klase ortogonalnih polinoma.

Primer 1.7.1. U prostoru $L^2_w(-1,1)$, sa $w(x)=(1-x^2)^{3/2}$, odredićemo prvih pet članova ortogonalnog sistema $\{Q_n\}_{n\in\mathbb{N}_0}$.

Izračunaćemo najpre integral

$$N_k=\int_{-1}^{+1}x^{2k}(1-x^2)^{3/2}\,\mathrm{d}x\qquad(k\in\mathbb{N}_0).$$

Primenom parcijalne integracije na integral

$$N_{k-1}-N_k=\int_{-1}^{+1}x^{2(k-1)}(1-x^2)^{5/2}\,\mathrm{d}x\qquad(k\in\mathbb{N}),$$

dobijamo

$$N_{k-1}-N_k=\frac{5}{2k-1}N_k,\quad\text{tj.}\quad N_k=\frac{2k-1}{2k+4}N_{k-1}\quad(k\in\mathbb{N}).$$

Kako je $N_0=3\pi/8$, imamo $N_k=3\pi\,(2k+1)!!/(2k+4)!!$ $(k\in\mathbb{N})$.

Polazeći od prirodnog bazisa $U=\{1,x,x^2,\ldots\}$, GRAM–SCHMIDTovim postupkom ortogonalizacije dobijamo redom

$$\begin{aligned}
&Q_0(x)=1,\quad Q_1(x)=x-\frac{(x,Q_0)}{(Q_0,Q_0)}\,Q_0(x)=x,\\
&Q_2(x)=x^2-N_1N_0^{-1}=x^2-\frac{1}{6},\quad Q_3(x)=x^3-N_2N_1^{-1}x=x^3-\frac{3}{8}x,\\
&Q_4(x)=x^4-N_2N_0^{-1}-\left(N_3-\frac{1}{6}N_2\right)\left(N_2-\frac{1}{3}N_1+\frac{1}{36}N_0\right)^{-1}\left(x^2-\frac{1}{6}\right)\\
&\qquad=x^4-\frac{3}{5}x^2+\frac{3}{80},\quad\text{itd.}
\end{aligned}$$

Prva četiri člana odgovarajućeg ortonormiranog sistema $\{Q_k^*\}_{k\in\mathbb{N}_0}$ su redom

$$Q_0^*(x)=\sqrt{\frac{8}{3\pi}},\qquad Q_1^*(x)=\frac{4x}{\sqrt{\pi}},\qquad Q_2^*(x)=8\sqrt{\frac{6}{5\pi}}\left(x^2-\frac{1}{6}\right),$$

$$Q_3^*(x)=\frac{32}{\sqrt{3\pi}}\left(x^3-\frac{3}{8}x\right),\quad Q_4^*(x)=\frac{3}{20}\sqrt{\frac{7}{10\pi}}\left(x^4-\frac{3}{5}x^2+\frac{3}{80}\right).$$

Ovi polinomi su poznati kao ultrasferni ili GEGENBAUERovi[100] polinomi i oni su specijalni slučaj JACOBIjevih polinoma (videti napomenu 1.6.1) za $\alpha=$

[100] LEOPOLD GEGENBAUER (1849 – 1903) austrijski matematičar.

$\beta = 3/2$. Opšti GEGENBAUERovi polinomi se obično označavaju sa $C_n^{\lambda}(x)$, gde je $\lambda = \alpha - 1/2$ (videti [17, str. 133–135]). U našem primeru, polinomi $Q_n(x)$ su GEGENBAUERovi polinomi $C_n^2(x)$ ($\lambda = 2$). $\triangle$

U paketu MATHEMATICA postoji funkcija `Orthogonalize` kojom je realizovan GRAM–SCHMIDTov postupak ortogonalizacije. Sledeći MATHEMATICA kôd daje konstrukciju prvih $m+1$ ortonormiranih GEGENBAUERovih polinoma za težinsku funkciju $w(x) = v^{\alpha,\alpha}(x) = (1-x^2)^{\alpha}$:

```
pol[m_,t_] := Table[t^k, {k,0,m}];
ort[m_,x_,al_] := (Orthogonalize[pol[m,t],
     Integrate[Times[##](1-t^2)^al, {t,-1,1}] &] //
ort[9,x,3/2]
```

U konkretnom slučaju (za $\alpha = 3/2$ i $m = 9$) dobijamo prvih deset članova ortonormiranog niza GEGENBAUERovih polinoma $C_n^2(x)$, $n = 0,1,\ldots,9$,

$$\left\{2\sqrt{\frac{2}{3\pi}},\ \frac{4x}{\sqrt{\pi}},\ 8\sqrt{\frac{6}{5\pi}}\left(-\frac{1}{6}+x^2\right),\ \frac{4x\left(-3+8x^2\right)}{\sqrt{3\pi}},\ 2\sqrt{\frac{2}{35\pi}}\left(3-48x^2+80x^4\right),\right.$$
$$4\sqrt{\frac{2}{3\pi}}\,x\left(3-20x^2+24x^4\right),\ \frac{8}{3}\sqrt{\frac{2}{7\pi}}\left(-1+30x^2-120x^4+112x^6\right),$$
$$4\sqrt{\frac{2}{5\pi}}\,x\left(-5+60x^2-168x^4+128x^6\right),\ \frac{2}{3}\sqrt{\frac{2}{11\pi}}\left(5-240x^2+1680x^4-3584x^6+2304x^8\right),$$
$$\left.\frac{4x\left(15-280x^2+1344x^4-2304x^6+1280x^8\right)}{\sqrt{15\pi}}\right\}$$

U slučaju kada je $\alpha = 0$, tj. za (konstantnu) težinsku funkciju $w(x) = 1$ na $[-1,1]$, odgovarajući ortogonalni polinomi su poznati kao LEGENDREovi[101] polinomi. Označavaju se sa $P_n(x)$ i mogu se izraziti u eksplicitnom obliku pomoću tzv. RODRIGUESove[102] formule

$$P_n(x) = \frac{(-1)^n}{2^n n!} \cdot \frac{\mathrm{d}^n}{\mathrm{d}x^n}(1-x^2)^n, \quad n \geq 0.$$

[101] ADRIEN-MARIE LEGENDRE (1752 – 1833), poznati francuski matematičar.

[102] BENJAMIN OLINDE RODRIGUES (1795 – 1851), francuski bankar, matematičar i društveni reformator.

Kvadrat norme ovih polinoma je

$$\|P_n\|^2 = \frac{2}{2n+1}, \qquad n \geq 0.$$

Korišćenjem prethodnog MATHEMATICA kôda sa naredbom `ort[11,x,0]` dobijamo ortonormirane LEGENDREove polinome $P_k^*(x)$, $k = 0, 1, \ldots, 11$:

$$\left\{\frac{1}{\sqrt{2}},\ \sqrt{\frac{3}{2}}\ x,\ \frac{1}{2}\sqrt{\frac{5}{2}}\ \left(-1+3\,x^2\right),\ \frac{1}{2}\sqrt{\frac{7}{2}}\ x\left(-3+5\,x^2\right),\right.$$

$$\frac{3\left(3-30\,x^2+35\,x^4\right)}{8\sqrt{2}},\ \frac{1}{8}\sqrt{\frac{11}{2}}\ x\left(15-70\,x^2+63\,x^4\right),$$

$$\frac{1}{16}\sqrt{\frac{13}{2}}\ \left(-5+105\,x^2-315\,x^4+231\,x^6\right),\ \frac{1}{16}\sqrt{\frac{15}{2}}\ x\left(-35+315\,x^2-693\,x^4+429\,x^6\right),$$

$$\frac{1}{128}\sqrt{\frac{17}{2}}\ \left(35-1260\,x^2+6930\,x^4-12\,012\,x^6+6435\,x^8\right),$$

$$\frac{1}{128}\sqrt{\frac{19}{2}}\ x\left(315-4620\,x^2+18\,018\,x^4-25\,740\,x^6+12\,155\,x^8\right),$$

$$\frac{1}{256}\sqrt{\frac{21}{2}}\ \left(-63+3465\,x^2-30\,030\,x^4+90\,090\,x^6-109\,395\,x^8+46\,189\,x^{10}\right),$$

$$\left.\frac{1}{256}\sqrt{\frac{23}{2}}\ x\left(-693+15\,015\,x^2-90\,090\,x^4+218\,790\,x^6-230\,945\,x^8+88\,179\,x^{10}\right)\right\}$$

LEGENDREovi polinomi P_n se mogu dobiti i kao koeficijenti u TAYLORovom razvoju (videti [17, str. 128–131])

$$\Phi(x,t) = \frac{1}{\sqrt{1-2xt+t^2}} = \sum_{n=0}^{+\infty} P_n(x)t^n,$$

gde se funkcija $\Phi(x,t)$ naziva *generatrisa* ili *generatorska funkcija*.

2.2 UVOD U TEORIJU OPERATORA

U ovom poglavlju dajemo osnovne elemente teorije operatora koji su neophodni u mnogim problemima numeričke analize i teorije aproksimacija (na primer, kod rešavanja operatorskih jednačina, numeričke integracije, itd.). Pored linearnih operatora razmatraćemo i neke klase nelinearnih operatora. Znatno kompletnija teorija operatora može se naći, na primer, u [1], [4], [11], [13], [14].

2.2.1 Linearni operatori

Neka su X i Y linearni prostori nad istim poljem skalara $\mathbb{K} = \mathbb{R}$ (ili $\mathbb{K} = \mathbb{C}$). Pod operatorom $A\colon X \to Y$ podrazumeva se preslikavanje

$$u \mapsto g = Au \tag{2.1.1}$$

koje svakom elementu $u \in X$ pridružuje samo jedan element $g \in Y$. Ako je $Y = \mathbb{R}$ za operator $A\colon X \to \mathbb{R}$ se često koristi termin funkcionela.

Prostor X se naziva *oblast definisanosti operatora* A. Element g iz (2.1.1) naziva se slika elementa u, a sâm element u original. Skup svih slika, tj. skup $\{Au \,|\, u \in X\}$, naziva se oblast vrednosti operatora A i označava se sa $A(X)$. U slučaju, kada svakom elementu $g \in A(X)$ odgovara samo jedan original, za operator se kaže da je *obostrano jednoznačan*, tj. da je preslikavanje A bijekcija prostora X na $A(X)$.

U daljem razmatranju interesuju nas tzv. *linearni operatori*, koji imaju značajnu ulogu u opštoj teoriji operatora. Posebno su interesantni linearni operatori na konačno-dimenzionalnim prostorima jer su oni u uskoj vezi sa teorijom matrica.

Pre nego što definišemo linearni operator, definisaćemo aditivni operator i homogeni operator.

Definicija 2.1.1. Operator $A\colon X \to Y$ je *aditivan* ako je

$$A(u+v) = Au + Av$$

za svako $u, v \in X$.

Definicija 2.1.2. Operator $A\colon X \to Y$ je *homogen* ako je

$$A(\lambda u) = \lambda Au$$

za svako $u \in X$ i svako $\lambda \in \mathbb{K}$.

Definicija 2.1.3. Operator $A\colon X \to Y$ je *linearan* ako je istovremeno aditivan i homogen, tj. ako je za svako $u, v \in X$ i svako $\lambda, \mu \in \mathbb{K}$

$$A(\lambda u + \mu v) = \lambda Au + \mu Av.$$

Definicija 2.1.4. Za operator $I\colon X \to X$, za koji je $Iu = u$ za svako $u \in X$, kažemo da je *identički operator*.

Primetimo da *nula-operator* $O\colon X \to Y$ preslikava svako $u \in X$ na neutralni element $\theta \in Y$, tj. da je $Ou = \theta$.

Skup svih linearnih operatora koji preslikavaju prostor X u prostor Y, označavaćemo sa $L(X,Y)$. Primetimo da svaki operator $A \in L(X,Y)$ preslikava $\theta \in X$ u $\theta \in Y$, tj. $A\theta = \theta$.

Na samom početku ovog odeljka pomenuli smo $A(X)$ kao oblast vrednosti operatora A. Za linearni operator A sa T_A označimo ovu oblast. Nije teško utvrditi da je T_A potprostor prostora Y. Zaista, ako su $g = Au$ i $h = Av$, vektor $\alpha g + \beta h$ je slika vektora $\alpha u + \beta v$, za svako $\alpha, \beta \in \mathbb{K}$. Dakle, $\alpha g + \beta h \in T_A$.

Definicija 2.1.5. *Rang operatora* $A \in L(X,Y)$, u oznaci r_A ili $\operatorname{rang} A$, je dimenzija potprostora T_A, tj.

$$r_A = \operatorname{rang} A = \dim(T_A).$$

U vezi sa potprostorom T_A može se razmatrati i skup vektora $u \in X$ koji zadovoljavaju jednakost $Au = \theta$.

Definicija 2.1.6. *Jezgro operatora* $A \in L(X,Y)$, u oznaci N_A ili $\ker A$, je skup

$$N_A = \ker A = \{u \in X \,|\, Au = \theta\}.$$

Dimenzija jezgra naziva se *defekt operatora* A i označava se sa n_A ili $\operatorname{def} A$.

Jezgro operatora $\ker A \subset X$ je potprostor prostora X, s obzirom na implikaciju

$$(\forall \alpha, \beta \in \mathbb{K}) \qquad u, v \in \ker A \quad \Rightarrow \quad \alpha u + \beta v \in \ker A\,.$$

Rang operatora i defekt operatora nisu nezavisne karakteristike linearnog operatora. O tome govori sledeća teorema:

Teorema 2.1.1. *Neka je* $\dim X = n$ *i neka* $A \in L(X,Y)$. *Tada je*

$$\operatorname{rang} A + \operatorname{def} A = n.$$

Na osnovu prethodne teoreme imamo

$$\operatorname{rang} A = \dim T_A \leq \dim X = n,$$

što znači da dimenzija oblasti vrednosti operatora ne može biti veća od dimenzije oblasti definisanosti operatora.

U daljem tekstu neka su $A, B \in L(X,Y)$.

Definicija 2.1.7. Zbir operatora A i B, u oznaci $A+B$, je operator C određen pomoću

$$(\forall u \in X) \qquad Cu = (A+B)u = Au + Bu.$$

Definicija 2.1.8. Proizvod operatora A i skalara λ iz polja $\mathbb{K}$ je operator C određen pomoću

$$(\forall u \in X) \qquad Cu = (\lambda A)u = \lambda(Au).$$

Interesantan slučaj je za $\lambda = -1$. Odgovarajući operator $-A$ nazivamo *suprotan operator* operatoru A. Primetimo da je $A + (-A) = O$ nula-operator. Naime, za svako $u \in X$, imamo $Ou = \theta$.

Definicija 2.1.9. Proizvod operatora $A\colon Y \to Z$ i $B\colon X \to Y$ je operator $C = AB\colon X \to Z$, definisan pomoću

$$(\forall u \in X) \qquad Cu = ABu = A(Bu).$$

Pitanje odnosa ranga operatora AB i ranga operatora A i B daje sledeća teorema (za dokaz videti, na primer, [20, str. 129–130]).

Teorema 2.1.2. *Neka su X, Y, Z konačno-dimenzionalni prostori. Za operatore $A\colon Y \to Z$ i $B\colon X \to Y$ važe nejednakosti*

$$\operatorname{rang} A + \operatorname{rang} B - \dim Y \leq \operatorname{rang}(AB) \leq \min\bigl(\operatorname{rang} A, \operatorname{rang} B\bigr).$$

Ako za svako $u \in X$ važi implikacija $u \in X \;\Rightarrow\; Au \in X$, tada se može definisati iterirani operator A^n (n-ti stepen operatora A) kao

$$A^n = A(A^{n-1}) \qquad (n \in \mathbb{N}),$$

pri čemu je $A^0 = I$ identički operator.

Za operatore A^n i A^m $(n, m \in \mathbb{N}_0)$ važi jednakost $A^n A^m = A^{n+m}$.

Nije teško pokazati da skup $L(X,Y)$, snabdeven operacijom sabiranja kao unutrašnjom kompozicijom i množenjem operatora skalarom kao spoljašnjom kompozicijom, obrazuje linearni prostor nad poljem skalara $\mathbb{K}$.

U skupu linearnih operatora koji preslikavaju X u X, za koje obično kažemo da deluju u X, moguće je odrediti neki podskup operatora koji ima strukturu grupe u odnosu na operaciju množenja operatora. Da bi se odredio takav podskup potrebno je uvesti pojam regularnog operatora.

Definicija 2.1.10. Za linearni operator $A: X \to X$ kažemo da je *regularan* ili da je *nesingularan* ako se njegovo jezgro sastoji samo od nula-vektora θ.

Za operator koji nije regularan kažemo da je *singularan* ili da je *neregularan operator*.

Na primer, identički (jedinični) operator I je regularan, ali je nula-operator O singularan.

Regularni operatori poseduju više interesantnih osobina:

1° Defekt regularnog operatora $A: X \to X$ jednak je nuli, odakle sleduje

$$\text{rang}\,(A) = \dim X\,;$$

2° $T_A = X$;

3° za svako $g \in X$ postoji jedinstveno $u \in X$ takvo da je $Au = g$ (jedinstvenost originala);

4° proizvod konačnog broja regularnih operatora je regularan operator.

Osobina 3° je izuzetno važna. Da bismo je dokazali pretpostavimo da za neko $g \in X$ postoje dva vektora $u, u' \in X$ takva da je $Au = g$ i $Au' = g$. Tada je

$$A(u - u') = \theta.$$

Kako se, s druge strane, jezgro regularnog operatora A sastoji samo od nula-vektora, zaključujemo da je $u - u' = \theta$, tj. $u = u'$. Napomenimo da se često za definiciju regularnog operatora uzima upravo osobina 3°.

Osobine 2° i 3° kazuju da je regularan operator A bijekcija prostora X na samog sebe.

Saglasno osobinama 2° i 3° da svakom vektoru $g \in X$ odgovara jedan i samo jedan vektor $u \in X$, može se za svaki regularan operator A definisati inverzan operator A^{-1}.

Definicija 2.1.11. Neka je $A\colon X \to X$ regularan operator. Za preslikavanje A^{-1}, za koje važi

$$(\forall u \in X) \qquad A^{-1}(Au) = u,$$

kažemo da je *inverzan operator* od A.

Teorema 2.1.3. *Neka je $A\colon X \to X$ regularan linearni operator. Tada je inverzan operator A^{-1}, takođe, regularan linearni operator.*

Dokaz. Neka su $Au_1 = g_1$ i $Au_2 = g_2$, tj. $A^{-1}g_1 = u_1$ i $A^{-1}g_2 = u_2$, i neka su c_1 i c_2 proizvoljni skalari iz polja $\mathbb{K}$. S obzirom na to da je A linearan operator, imamo

$$\begin{aligned} A^{-1}(c_1g_1 + c_2g_2) &= A^{-1}(c_1Au_1 + c_2Au_2) \\ &= A^{-1}A(c_1u_1 + c_2u_2) \\ &= c_1u_1 + c_2u_2 \\ &= c_1A^{-1}g_1 + c_2A^{-1}g_2. \end{aligned}$$

Dokažimo još da je i A^{-1} regularan operator.

Za svako $g \in \ker A^{-1}$ imamo $A^{-1}g = \theta$. Primenom operatora A na levu i desnu stranu poslednje jednakosti dobijamo

$$A(A^{-1}g) = A\theta,$$

tj. $g = \theta$, jer je $A\theta = \theta$. Dakle, jezgro operatora A^{-1} sastoji se samo od nula-vektora, tj. A^{-1} je regularan operator. □

Primer 2.1.1. Odredimo operator inverzan linearnom integralnom operatoru

$$(Af)(x) = \int_a^b \mathrm{e}^{-|\alpha(x)-\alpha(t)|} f(t)\,\mathrm{d}t\,,$$

gde je α data monotono neopadajuća funkcija na $[a,b]$ i dvaput neprekidno-diferencijabilna.

Neka je $\alpha(x) - \alpha(t) = u$. Tada je

$$g(x) = (Af)(x) = \int_a^x \mathrm{e}^{-u} f(t)\,\mathrm{d}t + \int_x^b \mathrm{e}^{u} f(t)\,\mathrm{d}t,$$

odakle sleduje

$$g'(x) = f(x) - \int_a^x \mathrm{e}^{-u}\alpha'(x) f(t)\,\mathrm{d}t - f(x) + \int_x^b \mathrm{e}^{u}\alpha'(x) f(t)\,\mathrm{d}t,$$

tj.

$$g'(x) = \alpha'(x)\left[-\int_a^x \mathrm{e}^{-u} f(t)\,\mathrm{d}t + \int_x^b \mathrm{e}^{u} f(t)\,\mathrm{d}t\right]. \tag{2.1.2}$$

Na dalje, imamo

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{g'(x)}{\alpha'(x)}\right) = -2f(x) + \alpha'(x)g(x),$$

odakle je

$$f(x) = (A^{-1}g)(x) = \frac{1}{2}\alpha'(x)g(x) - \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{g'(x)}{\alpha'(x)}\right).$$

Kako iz (2.1.2) sleduje

$$g'(a) = \alpha'(a)g(a) \qquad \text{i} \qquad g'(b) = -\alpha'(b)g(b), \tag{2.1.3}$$

zaključujemo da je inverzan operator A^{-1} u prostoru $C^2[a,b]$ definisan za funkcije koje zadovoljavaju granične uslove (2.1.3). $\triangle$

U daljem tekstu pretpostavićemo da su X i Y BANACHovi prostori.

Definicija 2.1.12. Operator $A : X \to Y$ je neprekidan ako za svaki niz $\{u_n\}_{n\in\mathbb{N}}$ iz X važi

$$u_n \to u \;\Rightarrow\; Au_n \to Au \qquad (u \in X).$$

Definicija 2.1.13. Linearni operator $A : X \to Y$ je ograničen ako postoji nenegativan broj M takav da je

$$(\forall u \in X) \qquad \|Au\| \le M\|u\|. \tag{2.1.4}$$

Infimum brojeva M za koje važi (2.1.4) označava se sa $\|A\|$ i naziva norma operatora A.

Teorema 2.1.4. *Linearni operator $A : X \to Y$ je ograničen ako i samo ako je neprekidan.*

Dokaz. Ako je operator A ograničen, tada važi

$$\|Au_n - Au\| = \|A(u_n - u)\| \le \|A\| \cdot \|u_n - u\|,$$

odakle pri $u_n \to u$ sleduje njegova neprekidnost u proizvoljnoj tački u.

Pretpostavimo sada da operator A nije ograničen. Tada za proizvoljno $n > 0$ postoji par elemenata $u_n,\ v_n \in X$, takav da je

$$u_n \neq v_n \qquad \text{i} \qquad \|Au_n - Av_n\| > 2n\|u_n - v_n\|.$$

Neka je dalje $\{c_n\}$ niz racionalnih brojeva sa osobinom

$$\|u_n - v_n\| < c_n < 2\|u_n - v_n\|.$$

Tada za $k_n = \dfrac{1}{nc_n}(u_n - v_n)$, imamo

$$\|Ak_n\| = \frac{1}{nc_n}\|Au_n - Av_n\| > \frac{2n}{nc_n}\|u_n - v_n\| = \frac{2}{c_n}\|u_n - v_n\| > 1$$

i

$$\|k_n\| = \frac{1}{nc_n}\|u_n - v_n\| < \frac{1}{n}\,.$$

Kako $\|k_n\| \to 0$ i $\|Ak_n\| > 1$ $(n \to +\infty)$ protivureči neprekidnosti operatora A, sleduje da operator A mora biti ograničen. □

Primer 2.1.2. Neka su $X = C[0,1]$ i $Y = \mathbb{R}$. Operator $A : X \to Y$ definisan pomoću

$$Au = \int_0^1 u(t)\ \mathrm{d}t$$

je ograničena linearna funkcionela, s obzirom na to da je, za svako $u_1, u_2 \in X$ i svako $c_1, c_2 \in \mathbb{R}$,

$$\int_0^1 (c_1u_1(t) + c_2u_2(t))\ \mathrm{d}t = c_1\int_0^1 u_1(t)\ \mathrm{d}t + c_2\int_0^1 u_2(t)\ \mathrm{d}t,$$

kao i

$$(\forall u \in C[0,1]) \qquad |Au| = \left|\int_0^1 u(t)\ \mathrm{d}t\right| \leq \max_{0\leq t\leq 1}|u(t)| \int_0^1 \mathrm{d}t = \|u\|_{[0,1]}. \quad \triangle$$

Primer 2.1.3. Neka je $X = C^n[a,b]$ i neka su $a_k : [a,b] \to \mathbb{R}$, $k = 1,\ldots,n$, date funkcije koje pripadaju X. Tada je pomoću

$$Du = \sum_{k=0}^{n} a_k u^{(k)}$$

definisan linearni operator $D : X \to X$. $\triangle$

Primer 2.1.4. Neka je $X = Y = C[a,b]$ i $K : [a,b]^2 \to \mathbb{R}$ data neprekidna funkcija na kvadratu $[a,b] \times [a,b]$. Tada je pomoću

$$v(x) = (Au)(x) = \lambda \int_a^b K(x,t)u(t)\,\mathrm{d}t, \tag{2.1.5}$$

gde je λ realan parametar različit od nule, definisan linearni integralni operator $A : X \to X$, što je jednostavno dokazati. Funkcija dve promenljive $(x,t) \mapsto K(x,t)$ se naziva jezgro integralnog operatora. Može se, takođe, dokazati da je operator A ograničen. Zaista, ako je M konstanta takva da je $|K(x,t)| \leq M$ za svako $(x,t) \in [a,b]^2$, tada je

$$\begin{aligned} \|Au\| = \|v\| = \max_{a\leq x\leq b} |v(x)| &= \max_{a\leq x\leq b} \Big| \lambda \int_a^b K(x,t)u(t)\,\mathrm{d}t \Big| \\ &\leq |\lambda| \max_{a\leq x\leq b} \int_a^b |K(x,t)||u(t)|\,\mathrm{d}t \\ &\leq |\lambda| M \max_{a\leq x\leq b} |u(t)|(b-a) \\ &\leq |\lambda| M(b-a)\|u\|. \end{aligned}$$

Dakle, ako uvedemo konstantu $C = |\lambda|M(b-a)$, vidimo da je $\|Au\| \leq C\|u\|$ za svaku neprekidnu funkciju $u \in C[a,b]$, što znači da je linearni operator A, definisan pomoću (2.1.5), ograničen.

Međutim, operator $T : X \to X$ definisan sa

$$(Tu)(x) = \int_a^b K(x,t,u(t))\,\mathrm{d}t,$$

gde je $K : [a,b]^2 \times \mathbb{R} \to \mathbb{R}$ data funkcija, u opštem slučaju, je nelinearni integralni operator. $\triangle$

Primer 2.1.5. Neka je $X = L^1(\mathbb{R})$ i

$$(\mathcal{F}u)(\xi) = \widehat{u}(\xi) = \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}2\pi\xi t} u(t)\,\mathrm{d}t.$$

Ovaj linearni operator se naziva FOURIER*ova transformacija*. Jednostavno je primetiti da je funkcija $\xi \mapsto \mathrm{e}^{-\mathrm{i}2\pi\xi t}u(t)$ neprekidna na $\mathbb{R}$ i ograničena sa $|u(t)|$, koja pripada $L^1(\mathbb{R})$. Zato je funkcija $\xi \mapsto \widehat{u}(\xi)$ neprekidna i ograničena na $\mathbb{R}$.

Linearni operator $\mathcal{F}: L^1(\mathbb{R}) \to L^\infty(\mathbb{R})$ je neprekidan, pri čemi je $\|\widehat{u}\|_\infty \leq \|u\|_1$, kao i $\lim\limits_{|\xi|\to+\infty} |\widehat{u}(\xi)| = 0$.

Za L^1-funkciju definisanu sa $u(t) = \dfrac{t^k}{k!}\,\mathrm{e}^{-at}h(t)$, gde su $\operatorname{Re} a > 0$, $k \in \mathbb{N}_0$ i h funkcija skoka ili tzv. HEAVISIDEova[103] funkcija, definisana pomoću

$$h(t) = \begin{cases} 1, & t > 0, \\ 0, & t \leq 0, \end{cases}$$

FOURIERova transformacija je

$$(\mathcal{F}u)(\xi) = \widehat{u}(\xi) = \frac{1}{(a + 2\pi\xi\,\mathrm{i})^{k+1}}.$$

U slučaju eksponencijalne funkcije $u(t) = \mathrm{e}^{-at^2}$, $a > 0$, za FOURIERovu transformaciju dobijamo

$$(\mathcal{F}u)(\xi) = \widehat{u}(\xi) = \sqrt{\frac{\pi}{a}}\,\mathrm{e}^{-(\pi\xi)^2/a}. \quad \triangle$$

Neka su A i B operatori koji preslikavaju prostor X u prostor Y.

Ako su A i B linearni ograničeni operatori, koji preslikavaju prostor X u prostor Y, tada važi

$$\|Au + Bu\| \leq \|Au\| + \|Bu\| \leq (\|A\| + \|B\|)\,\|u\|,$$

tj.

$$\|A + B\| \leq \|A\| + \|B\|.$$

Neka je T_2u tačka prostora X.

Ako su T_1 i T_2 ograničeni operatori, tada važi

$$\|Tu\| = \|T_1\,(T_2u)\| \leq \|T_1\| \cdot \|T_2u\| \leq \|T_1\| \cdot \|T_2\| \cdot \|u\|,$$

tj.

$$\|T_1T_2\| \leq \|T_1\| \cdot \|T_2\|.$$

Ako za svako $u \in X$ važi implikacija $u \in X \;\Rightarrow\; Tu \in X$, tada se može definisati iterirani operator T^n (n-ti stepen operatora T) kao

[103] OLIVER HEAVISIDE (1850 – 1925), engleski elektroinženjer, matematičar i fizičar, koji je napustio školovanje sa 16 godina i posvetio se samostalnom učenju.

$$T^n = T\left(T^{n-1}\right) \qquad (n \in \mathbb{N}),$$

pri čemu je $T^0 = I$ identički operator ($Iu = u$ za svako $u \in X$).

Za operatore T^n i T^m $(n, m \in \mathbb{N}_0)$ važi jednakost

$$T^n T^m = T^{n+m}.$$

Ako je T ograničen operator tada je

$$\|T^n\| \leq \|T\|^n \qquad (n \in \mathbb{N}_0).$$

Teorema 2.1.5. *Neka je X* BANACH*ov prostor, I identički operator u X i $T : X \to X$ ograničen linearni operator kod koga je $\|T\| \leq q < 1$. Tada postoji operator $(I-T)^{-1}$ za koji važi:*

1° $\quad (I-T)^{-1} = \sum_{k=0}^{+\infty} T^k$ $\quad$ (NEUMANN*ov razvoj*);

2° $\quad \|(I-T)^{-1}\| \leq \dfrac{1}{1-q}.$

Dokaz. S obzirom na to da je $\|T\| \leq q < 1$ imamo

$$\sum_{k=0}^{+\infty} \|T^k\| \leq \sum_{k=0}^{+\infty} \|T\|^k \leq \frac{1}{1-q} < +\infty. \tag{2.1.6}$$

Kako je, dalje, prostor X kompletan, to iz konvergencije reda $\sum_{k=0}^{+\infty} \|T^k\|$ sledi da je $\sum_{k=0}^{+\infty} T^k$ ograničen linearan operator.

Iz jednakosti

$$(I-T)\sum_{k=0}^{n} T^k = \sum_{k=0}^{n} T^k (I-T) = I - T^{n+1},$$

koja važi za svako $n \in \mathbb{N}$, a s obzirom na

$$\|T^{n+1}\| \leq \|T\|^{n+1} \to 0 \;\Rightarrow\; T^{n+1} \to 0 \quad (n \to +\infty),$$

sledi

$$(I-T)\sum_{k=0}^{+\infty} T^k = \sum_{k=0}^{+\infty} T^k (I-T) = I,$$

tj.

$$(I-T)^{-1} = \sum_{k=0}^{+\infty} T^k.$$

Korišćenjem (2.1.6) neposredno dobijamo

$$\|(I-T)^{-1}\| \leq \sum_{k=0}^{+\infty} \|T^k\| \leq \frac{1}{1-q}. \quad \square$$

Teorema 2.1.6. *Neka su $A : X \to X$ i $B : X \to X$ linearni operatori za koje postoje neprekidni inverzni operatori A^{-1} i B^{-1}. Tada je*

$$\|A^{-1} - B^{-1}\| \leq \frac{\|B^{-1}\| \cdot \|I - B^{-1}A\|}{1 - \|I - B^{-1}A\|},$$

gde je I identički operator u X.

Definicija 2.1.14. Neka je $T : X \to X$ linearan operator. Kompleksan broj λ je sopstvena ili karakteristična vrednost operatora T ako postoji vektor u različit od nula–vektora takav da je

$$Tu = \lambda u.$$

Vektor u koji odgovara sopstvenoj vrednosti λ naziva se sopstveni ili karakteristični vektor operatora T.

Teorema 2.1.7. *Neka su X i Y* BANACH*ovi prostori, $A : X \to Y$ ograničen linearni operator i $\{A_n\}_{n\in\mathbb{N}}$ niz ograničenih linearnih operatora koji preslikavaju X u Y. Potrebni i dovoljni uslovi za konvergenciju niza $\{A_n u\}$ ka Au za svako $u \in X$ su:*

1° *niz $\{\|A_n\|\}_{n\in\mathbb{N}}$ je ograničen;*

2° *postoji $\lim_{n\to+\infty} A_n u$ na nekom u X svuda gustom skupu.*

Dokaz ovog tvrđenja, koja je poznato kao BANACH–STEINHAUSova[104] teorema, može se naći, na primer, u [1].

2.2.2 Matrica linearnog operatora na konačno-dimenzionalnim prostorima

Neka su X i Y konačno-dimenzionalni prostori sa bazama $B_u = \{u_1, \ldots, u_n\}$ i $B_v = \{v_1, \ldots, v_m\}$, respektivno, i neka je $A : X \to Y$ linearan operator. Nije teško

[104] WŁADYSŁAW HUGO DIONIZY STEINHAUS (1887–1972), poznati poljski matematičar.

dokazati da je operator A potpuno određen ako su poznate slike vektora baze B_u, tj. ako su poznati vektori Au_i $(i = 1, \ldots, n)$. Razložimo ove vektore po vektorima baze B_v. Tada imamo

$$\begin{aligned} Au_1 &= a_{11}v_1 + a_{21}v_2 + \cdots + a_{m1}v_m, \\ Au_2 &= a_{12}v_1 + a_{22}v_2 + \cdots + a_{m2}v_m, \\ &\vdots \\ Au_n &= a_{1n}v_1 + a_{2n}v_2 + \cdots + a_{mn}v_m, \end{aligned} \tag{2.2.1}$$

Na osnovu (2.2.1) formirajmo matricu

$$A_{vu} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & & a_{mn} \end{bmatrix}.$$

Definicija 2.2.1. Za A_{vu} kažemo da je matrica operatora $A : X \to Y$ u odnosu na baze B_u i B_v.

Posmatrajmo proizvoljne vektore $u \in X$ i $v \in Y$, čije su koordinatne reprezentacije, u bazama B_u i B_v, date sa

$$u = \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \text{i} \qquad v = \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

respektivno. Veza između koordinata vektora u i vektora $v = Au$ može se iskazati pomoću jednakosti

$$\boldsymbol{y} = A_{vu}\boldsymbol{x}\,.$$

Formalno, linearni operator A možemo zameniti njegovom matricom A_{vu}.

Da bismo opisali operator $A : X \to X$ dovoljno je fiksirati jednu bazu $B_u = \{u_1, \ldots, u_n\}$. Naime, prethodno razmatranje ostaje u važnosti ako stavimo $Y = X$ i $B_v = B_u$. Matricu A_{uu}, u tom slučaju, označavamo sa A_u.

Na kraju napomenimo da će elementi matričnog računa biti posebno razmatrani u poglavlju 2.3.

2.2.3 Bilinearni i n-linearni operatori

Skup ograničenih linearnih operatora $L(X,Y)$ koji preslikavaju BANACHov prostor X u BANACHov prostor Y ima strukturu BANACHovog prostora ako su u njega uvedene unutrašnja i spoljašnja kompozicija pomoću

$$(T+S)u = Tu+Su, \quad (cT)u = c(Tu) \quad (T,S \in L(X,Y);\ c \in \mathbb{K}),$$

dok se pod normom elementa $T \in L(X,Y)$ podrazumeva norma ograničenog linearnog operatora T u smislu definicije 2.1.13.

Neka su X i Y BANACHovi prostori i operator $B : X^2 \to Y$.

Definicija 2.3.1. Operator B je bilinearan ako svakom uređenom paru elemenata $(u,u') \in X^2$ odgovara element $g = B(u,u') \in Y$, pri čemu za sve $u_1,u_2,u_1',u_2' \in X$ i sve $c_1,c_2 \in \mathbb{K}$, važe jednakosti:

$$B(c_1u_1+c_2u_2,u') = c_1B(u_1,u')+c_2B(u_2,u'),$$

$$B(u,c_1u_1'+c_2u_2') = c_1B(u,u_1')+c_2B(u,u_2').$$

Ako postoji pozitivan broj M takav da je

$$\|B(u,u')\| \le M\|u\| \cdot \|u'\| \tag{2.3.1}$$

za svako $u,u' \in X$, operator B je ograničen.

Infimum brojeva M za koje važi (2.3.1) naziva se normom bilinearnog operatora B i označava se sa $\|B\|$.

Shodno prethodnoj definiciji, bilinearna preslikavanja prostora X u prostor Y, obrazuju linearni normirani prostor, koji označavamo sa $B\left(X^2,Y\right)$. Može se pokazati da su prostori $L\big(X,L(X,Y)\big)$ i $B\left(X^2,Y\right)$ izometrični.

Definicija 2.3.2. Operator $N : X^n \to Y$ je n-linearan ako svakom uređenom sistemu elemenata $(u_1,u_2,\ldots,u_n)$ iz X odgovara element $g = N(u_1,u_2,\ldots,u_n) \in Y$, pri čemu je on linearan po svakom u_i, pri fiksiranim ostalim elementima $u_1,\ldots,u_{i-1},u_{i+1},\ldots,u_n$.

Ako postoji pozitivan broj M takav da je

$$\|N(u_1,u_2,\ldots,u_n)\| \le M\|u_1\| \cdot \|u_2\| \cdots \|u_n\|,$$

operator N je ograničen.

Sva n-linearna preslikavanja prostora X u prostor Y obrazuju linearni normiran prostor, koji označavamo sa $N\left(X^n,Y\right)$.

2.2.4 FRÉCHETova diferenciranja

Neka su X i Y BANACHovi prostori i neka je nelinearni operator $T : X \to Y$ definisan na skupu $D \subset X$.

Definicija 2.4.1. Operator T je FRÉCHET[105]–diferencijabilan u tački $u \in D$, ako postoji takav ograničeni linearan operator $A \in L(X,Y)$ da je

$$\lim_{\|h\|\to 0} \frac{\|T(u+h) - Tu - Ah\|}{\|h\|} = 0.$$

Operator A se naziva izvod operatora T u tački u i označava sa $T'_{(u)}$.

Teorema 2.4.1. *Ako je operator T* FRÉCHET*–diferencijabilan u tački u, onda je njegov izvod jedinstven.*

Dokaz. Neka su A_1 i A_2 linearni ograničeni operatori, takvi da je

$$L_i = \lim_{\|h\|\to 0} \frac{\|T(u+h) - Tu - A_i h\|}{\|h\|} = 0, \quad i = 1,2,$$

i neka je $A = A_1 - A_2$. Korišćenjem

$$\|Ah\| = \|A_1 h - A_2 h\| \le \|T(u+h) - Tu - A_1 h\| + \|T(u+h) - Tu - A_2 h\|$$

nalazimo

$$\lim_{\|h\|\to 0} \frac{\|Ah\|}{\|h\|} = L_1 + L_2 = 0. \tag{2.4.1}$$

Međutim, ako je za neko h

$$\frac{\|Ah\|}{\|h\|} = \lambda \neq 0,$$

tada će i za svako $\varepsilon \neq 0$ biti

$$\frac{\|A(\varepsilon h)\|}{\|\varepsilon h\|} = \lambda,$$

odakle zaključujemo da je jednakost (2.4.1) nemoguća kada je A nenula–operator.

Dakle, (2.4.1) važi ako je A nula–operator, tj. $A_1 = A_2$. □

[105] MAURICE RENÉ FRÉCHET (1878 – 1973), poznati francuski matematičar.

Teorema 2.4.2. *Svaki linearan ograničeni operator je* FRÉCHET*–diferencijabilan, pri čemu je* $T'_{(u)}\cdot = T\cdot$ *za svako* $u \in D$.

Dokaz. Kako je $T \in L(X,Y)$, na osnovu definicije 2.4.1 i prethodne teoreme neposredno sledi je $T'_{(u)}h = Th$ za svako $u \in D$. $\square$

Sada ćemo dati neka pravila diferencijalnog računa.

1° Neka su operatori $T : X \to Y$ i $H : X \to Y$ neprekidni i diferencijabilni u tački u. Tada je operator $\alpha T + \beta H$ $(\alpha, \beta \in K)$, takođe, diferencijabilan u toj tački, pri čemu je

$$(\alpha T + \beta H)'_{(u)} = \alpha T'_{(u)} + \beta H'_{(u)}\,.$$

2° Neka su X, Y, Z BANACHovi prostori, a $T : X \to Y$ i $S : Y \to Z$ FRÉCHET–diferencijabilni operatori. Ako su $Tu = g$ i $Sg = h$, izvod složenog operatora ST $(h = S(Tu) = (ST)u)$ u tački u je

$$(ST)'_{(u)} = S'_{(Tu)}T'_{(u)}\,.$$

3° Ako je A proizvoljan linearan ograničeni operator, a T FRÉCHET-diferencijabilan operator, tada važi

$$(AT)'_{(u)} = AT'_{(u)} \quad \text{i} \quad (TA)'_{(u)} = T'_{(Au)}A.$$

Ako je operator T dva puta FRÉCHET–diferencijabilan, tada je njegov drugi izvod $T''_{(u)} \in L(X, L(X,Y))$, tj. $T''_{(u)} \in B\left(X^2, Y\right)$. Slično, ako je T n puta FRÉCHET–diferencijabilan operator, tada je $T^{(n)}_{(u)} \in N\left(X^n, y\right)$.

Primer 2.4.1. Neka su $X = \mathbb{R}^n$, $Y = \mathbb{R}$, $T\boldsymbol{x} = f(\boldsymbol{x}) = f(x_1, \ldots, x_n)$, $\boldsymbol{x} = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}^T$, $\boldsymbol{h} = \begin{bmatrix} h_1 & \ldots & h_n \end{bmatrix}^T$, $\|\boldsymbol{x}\|_\infty = \max_{1\le k\le n} \|x_k\|$. Pod uslovom da f ima neprekidne parcijalne izvode drugog reda, naći ćemo $T'_{(\boldsymbol{x})}$.

Na osnovu TAYLORove formule imamo

$$f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \sum_{k=1}^{n} \left.\frac{\partial f}{\partial x_k}\right|_{\boldsymbol{x}} h_k + \frac{1}{2}\sum_{k,j} \left.\frac{\partial^2 f}{\partial x_k \partial x_j}\right|_{\boldsymbol{c}} h_k h_j\,,$$

tj.

$$\left| f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x}) - (\operatorname{grad} f)^T \boldsymbol{h} \right| \le M(f,\boldsymbol{c}) \left(\max_k \|h_k\|\right)^2, \tag{2.4.2}$$

gde je $\boldsymbol{c} = \boldsymbol{x} + \xi\boldsymbol{h}$ $(0 < \xi < 1)$,

$$M(f,\boldsymbol{c}) = \frac{1}{2}\left|\sum_{k,j}\frac{\partial^2 f}{\partial x_k \partial x_j}\Big|_{\boldsymbol{c}}\right| \qquad \text{i} \qquad \operatorname{grad} f = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \vdots \\ \dfrac{\partial f}{\partial x_n} \end{bmatrix}.$$

S obzirom na neprekidnost parcijalnih izvoda drugog reda funkcije f, veličina $M(f,\boldsymbol{c})$ je konačna, pa na osnovu (2.4.2) imamo

$$\frac{\|f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x}) - (\operatorname{grad} f)^T \boldsymbol{h}\|}{\|\boldsymbol{h}\|} \to 0$$

kada $\|\boldsymbol{h}\| \to 0$. Dakle,

$$T'_{(\boldsymbol{x})} = (\operatorname{grad} f)^T .$$

$\triangle$

Primer 2.4.2. Neka su $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, $\boldsymbol{x} \in X$, $\boldsymbol{y} \in Y$, $\|\boldsymbol{x}\|_\infty = \max_{1\le k\le n}\|x_k\|$, $\|\boldsymbol{y}\|_\infty = \max_{1\le k\le n}\|y_k\|$, $\boldsymbol{h} = [h_1 \ldots h_n]^T$ $(\in X)$ i neka je operator $T : X \to Y$ definisan pomoću

$$T\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(x_1,\ldots,x_n) \\ \vdots \\ f_m(x_1,\ldots,x_n) \end{bmatrix}.$$

Ako pretpostavimo da funkcije f_k $(k = 1,\ldots,m)$ imaju neprekidne izvode drugog reda, primenjujući postupak iz prethodnog primera na svaku od funkcija f_k, dobijamo

$$T'_{(\boldsymbol{x})} = W(\boldsymbol{f}),$$

gde je W tzv. JACOBIeva matrica za $\boldsymbol{f}$, tj.

$$W(\boldsymbol{f}) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & & \\ \dfrac{\partial f_m}{\partial x_1} & & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

Primetimo da je matrica $W(\boldsymbol{f})$, u ovom slučaju, tipa $m \times n$. $\triangle$

2.2.5 TAYLORova formula

Najpre uvodimo pojam L-metričkog porostora.

Definicija 2.5.1. *Metrički prostor (X,d) naziva se L-metrički, ako za svako $u \in X$ postoji linearna ograničena funkcionela $L : X \to \mathbb{R}$, takva da je*

$$\|L\| = 1 \qquad i \qquad Lu = d(u,\theta). \tag{2.5.1}$$

Može se pokazati da BANACHov prostor poseduje svojstvo (2.5.1) (videti [4]).

Neka je D konveksan[106] podskup BANACHovog prostora X i neka je T proizvoljan $n+1$ puta FRÉCHET–diferencijabilni operator u oblasti D. Tada važi sledeće tvrđenje.

Teorema 2.5.1. *Ako su u i $u+h$ zadate tačke iz D, tada je*

$$T(u+h) = \sum_{k=0}^{n} \frac{1}{k!} T^{(k)}_{(u)} (\underbrace{h,h,\ldots,h}_{k \text{ puta}}) + W(u,h), \tag{2.5.2}$$

gde je

$$\|W(u,h)\| \le \frac{1}{(n+1)!} \sup_{t\in[0,1]} \|T^{(n+1)}_{(u+th)}\| \cdot \|h\|^{n+1}. \tag{2.5.3}$$

Dokaz. S obzirom na to da X ima osobinu L-metričkog prostora, to za element $W(u,h)$ postoji linearna ograničena funkcionela L, takva da je

$$\|L\| = 1 \qquad \text{i} \qquad LW = d(W,\theta) = \|W(u,h)\|.$$

Uvedimo sada pomoćnu funkciju

$$t \mapsto F(t) = LT(u+th), \tag{2.5.4}$$

čiji su izvodi redom jednaki

$$\begin{aligned} F'(t) &= LT'_{(u+th)}h, \\ &\vdots \\ F^{(n+1)}(t) &= LT^{(n+1)}_{(u+th)}(\underbrace{h,h,\ldots,h}_{n+1 \text{ puta}}), \end{aligned}$$

[106] Skup D je konveksan ako važi implikacija $u,\ u+h \in D \Rightarrow u+th \in D$ $(0 \le t \le 1)$.

s obzirom na činjenicu da su $(LT)'_{(a)} = LT'_{(a)}$ i $\dfrac{\mathrm{d}}{\mathrm{d}t}(u+th) = h$.

Sada imamo

$$\|W(u,h)\| = LW = LT(u+h) - LTu - \frac{1}{1!}LT'_{(u)}h - \cdots - \frac{1}{n!}LT^{(n)}_{(u)}(\underbrace{h,\ldots,h}_{n \text{ puta}}),$$

tj.

$$\|W(u,h)\| = F(1) - F(0) - \frac{1}{1!}F'(0) - \cdots - \frac{1}{n!}F^{(n)}(0).$$

Kako za funkciju (2.5.4) važi klasična ocena ostataka u TAYLORovoj formuli

$$\left\| F(1) - F(0) - \frac{1}{1!}F'(0) - \cdots - \frac{1}{n!}F^{(n)}(0) \right\| \leq \frac{1}{(n+1)!} \max_{t\in[0,1]} \|F^{(n+1)}(t)\|,$$

dobijamo

$$\|W(u,h)\| \leq \frac{1}{(n+1)!} \sup_{t\in[0,1]} \left(\|L\| \cdot \|T^{(n+1)}_{(u+th)}\| \right) \|h\|^{n+1},$$

odakle, na osnovu (2.5.1), zaključujemo da važi (2.5.3). □

Formula (2.5.2) se naziva TAYLORova formula za operatore.

Za $n = 0$ iz (2.5.3) dobija se sledeća važna nejednakost

$$\|T(u+h) - Tu\| \leq \sup_{t\in[0,1]} \|T'_{(u+th)}\| \cdot \|h\|,$$

koja predstavlja teoremu o srednjoj vrednosti.

2.3 ELEMENTI MATRIČNOG RAČUNA

2.3.1 Operacije sa matricama razbijenim na blokove

U ovom i narednim odeljcima ovog poglavlju dajemo osnovne elemente matričnog računa koji su neophodni za praćenje izlaganja u narednim glavama, a posebno u četvrtoj glavi.

Definicija 3.1.1. Ako se matrica A tipa $n \times m$ mrežom horizontalnih i vertikalnih pravih razloži na više matrica, kaže se da je matrica razbijena na blokove.

Blokovi su matrice A_{ij} tipa $m_i \times n_j$, gde su

$$\sum_{i=1}^{p} m_i = m \qquad \text{i} \qquad \sum_{j=1}^{q} n_j = n.$$

Operacije sa matricama razbijenom na blokove su formalno iste sa operacijama kod običnih matrica. Jednostavno se dokazuju sledeći rezultati.

Teorema 3.1.1. *Neka su matrice A i B razbijene na blokove, tj.*

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1q} \\ A_{21} & A_{22} & & A_{2q} \\ \vdots & & & \\ A_{p1} & A_{p2} & & A_{pq} \end{bmatrix} \qquad i \qquad B = \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1q} \\ B_{21} & B_{22} & & B_{2q} \\ \vdots & & & \\ B_{p1} & B_{p2} & & B_{pq} \end{bmatrix},$$

gde su A_{ij} i B_{ij} matrice istog tipa. Tada su

$$cA = \begin{bmatrix} cA_{11} & cA_{12} & \dots & cA_{1q} \\ cA_{21} & cA_{22} & & cA_{2q} \\ \vdots & & & \\ cA_{p1} & cA_{p2} & & cA_{pq} \end{bmatrix} \quad i \quad A+B = \begin{bmatrix} A_{11}+B_{11} & A_{12}+B_{12} & \dots & A_{1q}+B_{1q} \\ A_{21}+B_{21} & A_{22}+B_{22} & & A_{2q}+B_{2q} \\ & \vdots & & \\ A_{p1}+B_{p1} & A_{p2}+B_{p2} & & A_{pq}+B_{pq} \end{bmatrix},$$

gde je c skalar.

Teorema 3.1.2. *Neka su*

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1q} \\ A_{21} & A_{22} & & A_{2q} \\ \vdots & & & \\ A_{p1} & A_{p2} & & A_{pq} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1q} \\ B_{21} & B_{22} & & B_{2q} \\ \vdots & & & \\ B_{p1} & B_{p2} & & B_{pq} \end{bmatrix},$$

i neka su blokovi takvi da je broj kolona bloka A_{ij} jednak broju vrsta bloka B_{jk} $(i = 1,\dots,p;\ j = 1,\dots,q;\ k = 1,\dots,s)$. Tada je

$$AB = \begin{bmatrix} C_{11} & C_{12} & \dots & C_{1s} \\ C_{21} & C_{22} & & C_{2s} \\ \vdots & & & \\ C_{p1} & C_{p2} & & C_{ps} \end{bmatrix},$$

gde je $C_{ik} = \sum_{j=1}^{q} A_{ij}B_{jk}$ $(i = 1,\dots,p;\ k = 1,\dots,s)$.

Primer 3.1.1. Neka su

$$A=\begin{bmatrix}5&2&0&0\\2&1&0&0\\0&0&8&3\\0&0&5&2\end{bmatrix},\qquad B=\begin{bmatrix}1&-2&-1&0&0\\-2&5&2&0&0\\0&0&0&2&-3\\0&0&0&-5&8\end{bmatrix}.$$

Ako stavimo

$$A_{11}=\begin{bmatrix}5&2\\2&1\end{bmatrix},\ A_{22}=\begin{bmatrix}8&3\\5&2\end{bmatrix},\ B_{11}=\begin{bmatrix}1&-2&-1\\-2&5&2\end{bmatrix},\ B_{22}=\begin{bmatrix}2&-3\\-5&8\end{bmatrix},$$

imamo

$$AB=\begin{bmatrix}A_{11}B_{11}&0\\0&A_{22}B_{22}\end{bmatrix}.$$

Kako je

$$A_{11}B_{11}=\begin{bmatrix}5&2\\2&1\end{bmatrix}\cdot\begin{bmatrix}1&-2&-1\\-2&5&2\end{bmatrix}=\begin{bmatrix}1&0&-1\\0&1&0\end{bmatrix}$$

i

$$A_{22}B_{22}=\begin{bmatrix}8&3\\5&2\end{bmatrix}\cdot\begin{bmatrix}2&-3\\-5&8\end{bmatrix}=\begin{bmatrix}1&0\\0&1\end{bmatrix},$$

dobijamo

$$AB=\left[\begin{array}{ccc|cc}1&0&-1&0&0\\0&1&0&0&0\\\hline0&0&0&1&0\\0&0&0&0&1\end{array}\right].\quad\triangle$$

Teorema 3.1.3. *Neka je*

$$A=\begin{bmatrix}A_{11}&A_{12}\\A_{21}&A_{22}\end{bmatrix}$$

regularna matrica, gde su A_{11} i A_{22} kvadratne matrice. Ako je matrica A_{22} regularna, tada je

$$A^{-1}=\begin{bmatrix}X_{11}&X_{12}\\X_{21}&X_{22}\end{bmatrix},$$

gde su

$$X_{11}=\left(A_{11}-A_{12}A_{22}^{-1}A_{21}^{-1}\right)^{-1},\qquad X_{12}=-X_{11}A_{12}A_{22}^{-1},$$
$$X_{21}=-A_{22}^{-1}A_{21}X_{11},\qquad X_{22}=A_{22}^{-1}(I-A_{21}X_{12}).$$

Primer 3.1.2. Za matricu

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

odredićemo A^{-1}. Neka su

$$A_{11} = \begin{bmatrix} 0 & 0 \\ 0 & 3 \end{bmatrix}, \quad A_{12} = \begin{bmatrix} 1 & -1 \\ 1 & 4 \end{bmatrix}, \quad A_{21} = \begin{bmatrix} 2 & 7 \\ 1 & 2 \end{bmatrix}, \quad A_{22} = \begin{bmatrix} 6 & -1 \\ 2 & -1 \end{bmatrix}.$$

Na osnovu teoreme 3.1.3 imamo

$$X_{11} = \left(A_{11} - A_{12}A_{22}^{-1}A_{21}^{-1}\right)^{-1} = \frac{1}{6}\begin{bmatrix} -1 & 3 \\ -3 & -7 \end{bmatrix},$$

$$X_{12} = -X_{11}A_{12}A_{22}^{-1} = \frac{1}{6}\begin{bmatrix} -7 & 20 \\ 5 & -10 \end{bmatrix},$$

$$X_{21} = -A_{22}^{-1}A_{21}X_{11} = \frac{1}{6}\begin{bmatrix} 9 & 3 \\ 3 & 3 \end{bmatrix},$$

$$X_{22} = A_{22}^{-1}(I - A_{21}X_{12}) = \frac{1}{6}\begin{bmatrix} -3 & 6 \\ -3 & 6 \end{bmatrix}.$$

Dakle,

$$A^{-1} = \left[\begin{array}{cc|cc} -1 & 3 & -7 & 20 \\ -7 & -3 & 5 & -10 \\ \hline 9 & 3 & -3 & 6 \\ 3 & 3 & -3 & 6 \end{array}\right]. \quad \triangle$$

Definicija 3.1.2. Kvadratna matrica A ima svojstvo (A) ako se permutacijom vrsta i kolona može svesti na oblik

$$A^{-1} = \begin{bmatrix} D_1 & F_1 & 0 & 0 & \cdots & 0 & 0 \\ E_1 & D_2 & F_2 & 0 & & 0 & 0 \\ 0 & E_2 & D_3 & F_3 & & 0 & 0 \\ \vdots & & & & & & \\ 0 & 0 & 0 & 0 & & E_{m-1} & D_m \end{bmatrix},$$

gde su D_j $(j = 1, \ldots, m)$ dijagonalne, a E_j i F_j $(j = 1, \ldots, m-1)$ su pravougaone matrice.

2.3.2 LR faktorizacija kvadratne matrice

Često se kod rešavanja sistema linearnih jednačina javlja problem predstavljanja kvadratne matrice u obliku proizvoda dve trougaone matrice. Ovaj odeljak posvećen je rešavanju tog problemu.

Teorema 3.2.1. *Ako su sve determinante*

$$\Delta_k = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & & a_{2k} \\ \vdots & & & \\ a_{k1} & a_{k2} & & a_{kk} \end{vmatrix}, \quad k = 1,2,\dots,n-1,$$

različite od nule, matrica $A = [a_{ij}]_{n\times n}$ *se može predstaviti u obliku*

$$A = LR, \tag{3.2.1}$$

gde je L donja i R gornja trougaona matrica.

Trougaone matrice L i R reda n imaju oblike

$$L = [\ell_{ij}]_{n\times n} \quad (\ell_{ij} = 0 \Leftarrow i < j) \tag{3.2.2}$$

i

$$R = [r_{ij}]_{n\times n} \quad (r_{ij} = 0 \Leftarrow i > j). \tag{3.2.3}$$

Razlaganje (3.2.1), poznato kao LR faktorizacija (dekompozicija), nije jedinstveno, s obzirom na jednakost

$$(\forall c \neq 0) \qquad LR = (cL)\Big(\frac{1}{c}R\Big).$$

Međutim, ako se dijagonalnim elementima matrice R (ili matrice L) fiksiraju vrednosti od kojih nijedna nije jednaka nuli, razlaganje je jedinstveno.

S obzirom na (3.2.2) i (3.2.3) i imajući u vidu da je

$$a_{ij} = \sum_{k=1}^{\min(i,j)} \ell_{ik} r_{kj} \quad (i,j = 1,\dots,n),$$

elementi matrica L i R mogu se jednostavno odrediti rekurzivnim postupkom, ukoliko se unapred zadaju elementi dijagonalni elementi r_{ii} $(\neq 0)$ ili ℓ_{ii} $(\neq 0)$, $i = 1,\dots,n$.

Tako, na primer, neka su dati brojevi r_{ii} $(\neq 0)$, $i = 1,\dots,n$. Tada važi

$$(1)\qquad \left.\begin{aligned} \ell_{11} &= \frac{a_{11}}{r_{11}},\\ r_{1i} &= \frac{a_{1i}}{\ell_{11}}\\ \ell_{i1} &= \frac{a_{i1}}{r_{11}} \end{aligned}\right\} \quad (i = 2,\dots,n);$$

$$(\mathrm{i})\qquad \left.\begin{aligned} \ell_{11} &= \frac{1}{r_{ii}}\left(a_{ii} - \sum_{k=1}^{i-1} \ell_{ik} r_{ki}\right),\\ r_{ij} &= \frac{1}{\ell_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} \ell_{jk} r_{ki}\right)\\ \ell_{ji} &= \frac{1}{r_{ii}}\left(a_{ji} - \sum_{k=1}^{i-1} \ell_{jk} r_{ki}\right) \end{aligned}\right\} \ (j = i+1,\dots,n) \Bigg\} \ (i = 2,\dots,n).$$

Slično bismo mogli iskazati i rekurzivni postupak za određivanje elemenata matrica L i R ako su unapred dati brojevi ℓ_{ii} $(\neq 0)$, $i = 1,\dots,n$.

U primenama, najčešće se uzima $r_{ii} = 1$ (ili $\ell_{ii} = 1$), $i = 1,\dots,n$.

Primer 3.2.1. Razložimo matricu

$$A = \begin{bmatrix} 1 & 4 & 1 & 3\\ 0 & -1 & 2 & -1\\ 3 & 14 & 4 & 1\\ 1 & 2 & 2 & 9 \end{bmatrix}$$

u obliku (3.2.2), tako da jediničnu dijagonalu ima (a) matrica R; (b) matrica L.

(a) Kako je $r_{ii} = 1$, $i = 1,\dots,4$, na osnovu izloženog rekurzivnog postupka, imamo redom:

(1) $\ell_{11} = 1,$

$r_{12} = 4, \quad \ell_{21} = 0, \quad r_{13} = 1, \quad \ell_{31} = 3, \quad r_{14} = 3, \quad \ell_{41} = 1;$

(2) $\ell_{22} = -1,$

$r_{23} = -2, \quad \ell_{32} = 2, \quad r_{24} = 1, \quad \ell_{42} = -2;$

(3) $\ell_{33} = 5,$

$r_{34} = -2, \quad \ell_{43} = -3;$

(4) $\ell_{44} = 2.$

Dakle, dobili smo

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 3 & 2 & 5 & 0 \\ 1 & -2 & -3 & 2 \end{bmatrix} \quad \text{i} \quad R = \begin{bmatrix} 1 & 4 & 1 & 3 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

(b) Polazeći od $\ell_{ii} = 1$, $i = 1, \ldots, 4$, imamo redom

(1) $r_{11} = 1,$

$r_{12} = 4, \quad \ell_{21} = 0, \quad r_{13} = 1, \quad \ell_{31} = 3, \quad r_{14} = 3, \quad \ell_{41} = 1;$

(2) $r_{22} = -1,$

$r_{23} = 2, \quad \ell_{32} = -2, \quad r_{24} = -1, \quad \ell_{42} = 2;$

(3) $r_{33} = 5,$

$r_{34} = -10, \quad \ell_{43} = -3/5;$

(4) $r_{44} = 2,$

tj.

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 3 & -2 & 1 & 0 \\ 1 & 2 & -3/5 & 1 \end{bmatrix} \quad \text{i} \quad R = \begin{bmatrix} 1 & 4 & 1 & 3 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & 5 & -10 \\ 0 & 0 & 0 & 2 \end{bmatrix}. \quad \triangle$$

U primenama vrlo često se javljaju višedijagonalne matrice, tj. matrice čiji su elementi različiti od nule samo na glavnoj dijagonali i oko glavne dijagonale. Na primer, ako je $a_{ij} \neq 0$ za $|i-j| \leq 1$ i $a_{ij} = 0$ za $|i-j| > 1$, matrica je *trodijagonalna* ili *tridijagonalna*. Obično elemente ovakve matrice predstavljamo vektorima $(a_2, \ldots, a_n)$, $(b_1, \ldots, b_n)$, $(c_1, \ldots, c_{n-1})$, tj.

$$A = \begin{bmatrix} b_1 & c_1 & 0 & \ldots & 0 & 0 \\ a_2 & b_2 & c_2 & & 0 & 0 \\ 0 & a_3 & b_3 & & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & & b_{n-1} & c_{n-1} \\ 0 & 0 & 0 & & a_n & b_n \end{bmatrix}. \tag{3.2.4}$$

Ako je $a_{ij} \neq 0$ $(|i-j| \leq 2)$ i $a_{ij} = 0$ $(|i-j| > 2)$, imamo slučaj *petodijagonalne* matrice.

Pretpostavimo sada da trodijagonalna matrica (3.2.4) ispunjava uslove teoreme 3.2.1. Za dekompoziciju ovakve matrice dovoljno je pretpostaviti da su

$$L = \begin{bmatrix} \beta_1 & 0 & 0 & \ldots & 0 & 0 \\ \alpha_2 & \beta_2 & 0 & & 0 & 0 \\ 0 & \alpha_3 & \beta_3 & & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & & \alpha_n & \beta_n \end{bmatrix} \quad (\beta_1\beta_2\cdots\beta_n \neq 0)$$

i

$$R = \begin{bmatrix} 1 & \gamma_1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \gamma_2 & & 0 & 0 \\ 0 & 0 & 1 & & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & & 0 & 1 \end{bmatrix}.$$

Upoređivanjem odgovarajućih elemenata matrice A i matrice

$$LR = \begin{bmatrix} \beta_1 & \beta_1\gamma_1 & 0 & \ldots & 0 & 0 \\ \alpha_2 & \alpha_2\gamma_1 + \beta_2 & \beta_2\gamma_2 & & 0 & 0 \\ 0 & \alpha_3 & \alpha_3\gamma_2 + \beta_3 & & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & & \alpha_n & \alpha_n\gamma_{n-1} + \beta_n \end{bmatrix}$$

dobijamo sledeće rekurzivne formule za određivanje elemenata $\alpha_i, \beta_i, \gamma_i$:

$$\begin{aligned} &\beta_1 = b_1, \quad \gamma_1 = \frac{c_1}{\beta_1}, \\ &\alpha_i = a_i, \quad \beta_i = b_i - \alpha_i\gamma_{i-1}, \quad \gamma_i = \frac{c_i}{\beta_i}, \qquad i = 2,\dots,n-1, \\ &\alpha_n = a_n, \quad \beta_n = b_n - \alpha_n\gamma_{n-1}. \end{aligned} \tag{3.2.5}$$

2.3.3 Sopstveni vektori i sopstvene vrednosti matrica

Definicija 3.3.1. Neka je A kompleksna kvadratna matrica reda n. Svaki vektor $\boldsymbol{x} \in \mathbb{C}^n$, koji je različit od nula–vektora, naziva se *sopstveni vektor* matrice A ako postoji skalar $\lambda \in \mathbb{C}$ takav da je

$$A\boldsymbol{x} = \lambda\boldsymbol{x}. \tag{3.3.1}$$

Skalar λ naziva se odgovarajuća *sopstvena vrednost*.

S obzirom na činjenicu da se (3.3.1) može predstaviti u obliku

$$(A - \lambda I)\boldsymbol{x} = \boldsymbol{0},$$

zaključujemo da (3.3.1) ima netrivijalna rešenja (po $\boldsymbol{x}$) ako i samo ako je

$$\det(A - \lambda I) = 0.$$

Definicija 3.3.2. Ako je A kvadratna matrica, za polinom $P(\lambda) = \det(A - \lambda I)$ se kaže da je njen *karakteristični polinom*, a za odgovarajuću jednačinu $P(\lambda) = 0$ da je njena *karakteristična jednačina*.

Neka je $A = [a_{ij}]_{n\times n}$. Karakteristični polinom matrice A se može predstaviti u determinantnom obliku

$$P(\lambda) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & & a_{nn} - \lambda \end{vmatrix}$$

ili

$$P(\lambda) = (-1)^n \left[\lambda^n - p_1\lambda^{n-1} + p_2\lambda^{n-2} - \dots + (-1)^{n-1}p_{n-1}\lambda + (-1)^n p_n\right],$$

gde je p_k zbir svih glavnih minora reda k determinante matrice A, tj.

$$p_k = \sum_{1\le i_1<i_2<\cdots<i_k\le n} \det\left(A^{i_1 i_2 \ldots i_k}_{i_1 i_2 \ldots i_k}\right).$$

Primetimo da je

$$p_1 = \sum_{i=1}^{n} a_{ii} = \operatorname{tr} A \qquad \text{i} \qquad p_n = \det(A).$$

Često se umesto karakterističnog polinoma P koristi tzv. *normirani karakteristični polinom* H, definisan pomoću

$$H(\lambda) = (-1)^n P(\lambda) = \lambda^n - p_1\lambda^{n-1} + p_2\lambda^{n-2} - \cdots + (-1)^n p_n.$$

Sopstvene vrednosti λ_i matrice A, tj. nule polinoma P, označavaćemo sa $\lambda_i(A)$, $i = 1,\ldots,n$.

Definicija 3.3.3. Skup svih sopstvenih vrednosti kvadratne matrice A naziva se *spektar matrice* i označava sa $\operatorname{Sp}(A)$.

Definicija 3.3.4. *Spektralni radijus* $\rho(A)$ kvadratne matrice A je broj

$$\rho(A) = \max_i |\lambda_i(A)|.$$

Teorema 3.3.1. *Svaka matrica je, u matričnom smislu, nula svog karakterističnog polinoma.*

Ova teorema je poznata kao CAYLEY[107]–HAMILTONOVA[108] teorema.

Teorema 3.3.2. *Neka su $\lambda_1, \ldots, \lambda_n$ sopstvene vrednosti matrice A reda n i neka je $x \mapsto Q(x)$ skalarni polinom. Tada su*

$$Q(\lambda_1),\ \ldots\ ,\ Q(\lambda_n)$$

sopstvene vrednosti matrice $Q(A)$.

Teorema 3.3.3. *Neka su $\lambda_1, \ldots, \lambda_n$ sopstvene vrednosti regularne matrice A reda n. Tada su*

$$\lambda_1^{-1},\ \ldots\ ,\ \lambda_n^{-1}$$

sopstvene vrednosti matrice A^{-1}.

[107] ARTHUR CAYLEY (1821 – 1895), engleski matematičar.
[108] WILLIAM ROWAN HAMILTON (1805 – 1865), irski matematičar i astronom.

Teorema 3.3.4. *Sopstvene vrednosti trougaone matrice jednake su njenim dijagonalnim elementima.*

Sledeća teorema daje rekurzivni postupak za nalaženje karakterističnog polinoma trodijagonalne matrice (3.2.4).

Teorema 3.3.5. *Neka su*

$$A_k = \begin{bmatrix} b_1 & c_1 & 0 & \dots & 0 \\ a_2 & b_2 & c_2 & & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & b_k \end{bmatrix} \qquad i \qquad H_k(\lambda) = (-1)^k \det(A_k - \lambda I).$$

Normiran karakteristični polinom $\lambda \mapsto H(\lambda)$ $(= H_n(\lambda))$ matrice A $(= A_n)$ dobija se rekurzivnim postupkom

$$H_k(\lambda) = (\lambda - b_k)H_{k-1}(\lambda) - a_{k-1}c_{k-1}H_{k-2}(\lambda), \quad k = 2, \dots, n,$$

gde su $H_0(\lambda) = 1$ i $H_1(\lambda) = \lambda - b_1$.

Definicija 3.3.5. Za matricu B se kaže da je *slična* matrici A ako postoji bar jedna regularna matrica C, takva da je

$$B = C^{-1}AC.$$

Teorema 3.3.6. *Slične matrice imaju identične karakteristične polinome, a samim tim i iste sopstvene vrednosti.*

2.3.4 Specijalne matrice i njihove osobine

Neka je $A = [a_{ij}]_{m\times n}$ $(a_{ij} \in \mathbb{C})$. Sa $\overline{A}$ označavaćemo konjugovanu matricu za A, tj. $\overline{A} = [\overline{a}_{ij}]_{m\times n}$, dok ćemo sa A^T označavati transponovanu matricu od A. Nadalje, matricu $\overline{A}^T$ označavaćemo sa A^*.

Sledeće osobine se jednostavno dokazuju

$$(A+B)^* = A^* + B^*, \quad (AB)^* = B^*A^*, \quad (A^*)^* = A,$$

$$(A^*A)^* = A^*A, \quad \det A^* = \det A, \quad (A^{-1})^* = (A^*)^{-1}.$$

Definicija 3.4.1. Ako za kvadratnu matricu A važi jednakost $A = A^T$, matrica A se naziva *simetrična matrica*.

Definicija 3.4.2. Ako za kvadratnu matricu A važi $A = A^*$, matrica A se naziva *hermitska matrica*.

Definicija 3.4.3. Ako za kvadratnu matricu A važi $A = -A^*$, matrica A se naziva *kosohermitska matrica*.

Definicija 3.4.4. Ako za kvadratnu matricu A važi $A^*A = I$ (I jedinična matrica), matrica A se naziva *unitarna matrica*.

Definicija 3.4.5. Ako za regularnu matricu važi $A^T = A^{-1}$, matrica A se naziva *ortogonalna matrica*.

Korišćenjem skalarnog proizvoda dva vektora $\boldsymbol{x} = [x_1 \ \dots \ x_n]^T$ i $\boldsymbol{y} = [y_1 \ \dots \ y_n]^T$, datog sa

$$(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{y}^*\boldsymbol{x} = \sum_{i=1}^{n} x_i\overline{y}_i,$$

uslov za hermitsku matricu $(A = A^*)$ se može predstaviti u ekvivalentnoj formi

$$(\forall\, \boldsymbol{x},\boldsymbol{y} \in \mathbb{C}^n) \quad (A\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x},A\boldsymbol{y}).$$

Slično se uslov za kosohermitsku matricu $(A = -A^*)$ može predstaviti u obliku

$$(\forall\, \boldsymbol{x},\boldsymbol{y} \in \mathbb{C}^n) \quad (A\boldsymbol{x},\boldsymbol{y}) + (\boldsymbol{x},A\boldsymbol{y}) = 0.$$

Definicija 3.4.6. Hermitska matrica A se naziva *pozitivno definitna* ako je za svako $\boldsymbol{x} \neq \boldsymbol{0}$ ispunjen uslov

$$(A\boldsymbol{x},\boldsymbol{x}) = \boldsymbol{x}^*A\boldsymbol{x} > 0. \tag{3.4.1}$$

Ako su $A = [a_{ij}]_{n\times n}$ i $\boldsymbol{x} = [x_1 \ \cdots \ x_n]^T$, uslov (3.4.1) se može predstaviti u obliku

$$(A\boldsymbol{x},\boldsymbol{x}) = \sum_{i,j=1}^{n} a_{ij}\overline{x}_i x_j > 0,$$

odakle zaključujemo da je $a_{ii} > 0$ $(i = 1,\dots,n)$.

Definicija 3.4.7. Simetrčna pozitivno definitna matrica se naziva *normalna matrica*.

Teorema 3.4.1. *Matrica $A = [a_{ij}]_{n\times n}$ je pozitivno definitna ako i samo ako su ispunjeni* SYLVESTER*ovi*[109] *uslovi*

$$\Delta_k = \begin{vmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \\ a_{k1} & & a_{kk} \end{vmatrix} > 0 \quad (k = 1, \dots, n).$$

Teorema 3.4.2. *Ako matrica $C = [a_{ij}]_{m\times n}$ $(m \geq n)$ ima rang n, tada je matrica C^*C pozitivno definitna.*

Teorema 3.4.3. *Sve sopstvene vrednosti hermitske matrice su realni brojevi.*

Teorema 3.4.4. *Sve sopstvene vrednosti pozitivno definitne matrice su pozitivni brojevi.*

Teorema 3.4.5. *Sve sopstvene vrednosti kosohermitske matrice su čisto imaginarni brojevi.*

Teorema 3.4.6. *Ako je A hermitska matrica, tada je*

1° $\lambda(A^*A) = \lambda(A)^2$;

2° $\lambda_{\min}(A)(\boldsymbol{x},\boldsymbol{x}) \leq (A\boldsymbol{x},\boldsymbol{x}) \leq \lambda_{\max}(A)(\boldsymbol{x},\boldsymbol{x})$.

Napomenimo da se jednakosti u 2° postižu za sopstvene vektore koji odgovaraju sopstvenim vrednostima $\lambda_{\min}(A)$ i $\lambda_{\max}(A)$.

2.3.5 JORDANov kanonički oblik

Struktura linearnog operatora na konačno dimenzionalnim prostorima se detaljno proučava u linearnoj algebri, gde se razmatraju *invarijantni potprostori* karakterisani u odnosu na linearni operator, kao i razlaganje prostora X na direktnu sumu invarijantnih potprostora (za detalje videti [20, str. 317–326]). Tada se, pogodnom konstrukcijom bazisa u ovim potprostorima, može dobiti najprostiji oblik za matricu operatora[110], takozvani JORDANOV[111] KANONIČKI OBLIK. Za operatore proste strukture problem konstrukcije baze u kojoj matrica operatora ima najprostiji mogući oblik je veoma jednostavan. Baza se sastoji od sopstvenih

[109] JAMES JOSEPH SYLVESTER (1814 – 1897), engleski matematičar.
[110] Matrica lineanog operatora je razmatrana u odeljku 2.2.2.
[111] MARIE ENNEMOND CAMILLE JORDAN (1838 – 1922), francuski matematičar.

vektora, a matrica operatora je dijagonalna sa sopstvenim vrednostima na glavnoj dijagonali. Drugim rečima, matrica operatora proste strukture uvek je slična nekoj dijagonalnoj matrici. Međutim, klasa operatora proste strukture ne iscrpljuje skup svih linearnih operatora $L(X,X)$. U ovom odeljku dajemo samo neke osnovne elemente u vezi sa JORDANovim kanoničkim oblikom matrica (za detalje videti [20, str. 326–335]).

Definicija 3.5.1. *Kvadratna matrica reda r*

$$J_1(\lambda) = [\lambda] \qquad (r = 1),$$

$$J_r(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix} \qquad (r \geq 2), \tag{3.5.1}$$

naziva se JORDAN*ov blok.*

Definicija 3.5.2. *Kvazidijagonalna matrica oblika*

$$J = \begin{bmatrix} J_{r_1}(\lambda) & 0 & \cdots & 0 \\ 0 & J_{r_2}(\lambda) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_{r_m}(\lambda) \end{bmatrix} \tag{3.5.2}$$

naziva se JORDAN*ova matrica.*

JORDANova matrica često se predstavlja i u obliku

$$J = J_{r_1}(\lambda_1) \dotplus J_{r_2}(\lambda_2) \dotplus \cdots \dotplus J_{r_m}(\lambda_m), \tag{3.5.3}$$

kao *direktna suma* JORDANovih blokova.

Pod JORDANovim kanoničkim oblikom matrice A podrazumeva se ona JORDANova matrica koja je slična matrici A i njena egzistencija sledi iz sledeće teoreme.

Teorema 3.5.1. *Svaka klasa sličnih matrica sadrži bar jednu* JORDAN*ovu matricu.*

Napomenimo da je broj JORDANovih blokova u (3.5.2), tj. (3.5.3), jednak broju linearno nezavisnih vektora matrice A. Primetimo da je λ_i, takođe, sopstvena vrednost matrica $J_{r_i}(\lambda_i)$, J i A.

Napomena 3.5.1. Neka je $x \mapsto Q(x) = a_0x^k + a_1x^{k-1} + \cdots + a_k$ skalarni polinom i $J_r(\lambda)$ JORDANov blok (3.5.1). Tada je

$$Q(J_r(\lambda)) = \begin{bmatrix} Q(\lambda) & Q'(\lambda) & \frac{1}{2!}Q''(\lambda) & \cdots & \frac{1}{(r-1)!}Q^{(r-1)}(\lambda) \\ 0 & Q(\lambda) & Q'(\lambda) & \cdots & \frac{1}{(r-2)!}Q^{(r-2)}(\lambda) \\ 0 & 0 & Q(\lambda) & \cdots & \frac{1}{(r-3)!}Q^{(r-3)}(\lambda) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & Q(\lambda) \end{bmatrix}.$$

Štaviše, ako je J JORDANova matrica, imamo

$$Q(J) = Q(J_{r_1}(\lambda_1)) \dot{+} Q(J_{r_2}(\lambda_2)) \dot{+} \cdots \dot{+} Q(J_{r_m}(\lambda_m)).$$

Teorema 3.5.2. *Za proizvoljnu hermitsku matricu A, postoji unitarna matrica S takva da je S^*AS dijagonalna matrica, tj.*

$$S^*AS = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}.$$

Dve važne posledice ove teoreme su sledeće.

Posledica 3.5.1. *Svaka hermitska matrica A reda n ima n linearno nezavisnih sopstvenih vektora, koji obrazuju ortogonalni sistem.*

Posledica 3.5.2. *Potrebni i dovoljni uslovi da je hermitska matrica A reda n pozitivno definitna su $\lambda_i(A) > 0$ $(i = 1, \ldots, n)$.*

2.3.6 Norme vektora i matrica

Norme vektora i norme matrica igraju značajnu ulogu u mnogim problemima numeričke analize i teorije aproksimacija. Na primer, u teoriji iterativnih procesa njihova uloga je neizbežna pri ispitivanju konvergencije.

Neka je X realan prostor $\mathbb{R}^n$ (kompleksan prostor $\mathbb{C}^n$), sa nula-vektorom $\mathbf{0}$ $(= [0 \ \cdots \ 0]^T)$. Za vektor $\boldsymbol{x}$ $(= [x_1 \ \cdots \ x_n]^T)$ uvodi se norma saglasno definiciji 1.4.1 (odeljak 2.1.4). Od svih mogućih normi vektora od interesa su norme oblika

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \qquad (1 \leq p < +\infty). \tag{3.6.1}$$

U graničnom slučaju, kada $p \to +\infty$, iz (3.6.1) sleduje

$$\|\boldsymbol{x}\|_\infty = \max_i |x_i|. \tag{3.6.2}$$

U primenama, pored norme (3.6.2), koriste se i norme koje se iz (3.6.1) dobijaju za $p = 1$ i $p = 2$, tj.

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i| \tag{3.6.3}$$

i

$$\|\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_E = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2} \tag{3.6.4}$$

Norma (3.6.4) poznata je kao *euklidska norma*. Primetimo da je

$$\|\boldsymbol{x}\|_E = (\boldsymbol{x}^* \boldsymbol{x})^{1/2} = \sqrt{(\boldsymbol{x}, \boldsymbol{x})}.$$

Nije teško dokazati da za proizvoljno $\boldsymbol{x} \in X$ važe nejednakosti

$$\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_1 \leq n \|\boldsymbol{x}\|_\infty,$$
$$\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2 \leq \sqrt{n}\, \|\boldsymbol{x}\|_\infty,$$
$$\frac{1}{\sqrt{n}} \|\boldsymbol{x}\|_1 \leq \|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1.$$

Neka je X_M linearan realan ili kompleksan prostor kvadratnih matrica reda n, sa nula–matricom 0 $(\in X_M)$.

Definicija 3.6.1. Pod normom matrice $A \in X_M$ se podrazumeva nenegativan broj $\|A\|$, takav da je

1° $\|A\| = 0 \Leftrightarrow A = 0$ (definisanost),

2° $\|cA\| = |c| \cdot \|A\|$ (homogenost),

3° $\|A + B\| \leq \|A\| + \|B\|$ (relacija trougla),

gde su $A, B \in X_M$ i $c \in \mathbb{C}$.

Napomenimo da se norma $\|A\|$ može razmatrati i kao norma operatora A koji se primenjuje na vektore iz prostora X.

Sada dajemo definicije *saglasnosti* i *potčinjenosti* norme matrica sa normom vektora.

Definicija 3.6.2. Norma $\|A\|$ je *saglasna* sa normom $\|\boldsymbol{x}\|$, ako su ispunjeni uslovi

$$(\forall A \in X_M)\ (\forall \boldsymbol{x} \in X) \quad \|A\boldsymbol{x}\| \leq \|A\| \cdot \|\boldsymbol{x}\|$$

i

$$(\forall A, B \in X_M) \quad \|AB\| \leq \|A\| \cdot \|B\|.$$

Definicija 3.6.3. Neka je norma $\|A\|$ saglasna sa normom $\|\boldsymbol{x}\|$. Ako se za svako $A \in X_M$ može naći $\boldsymbol{x}$ $(\neq \mathbf{0})$, takvo da je $\|A\boldsymbol{x}\| \leq \|A\| \cdot \|\boldsymbol{x}\|$, za normu $\|A\|$ kažemo da je *potčinjena* normi $\|\boldsymbol{x}\|$.

Za svaku normu matrica, potčinjenu normi vektora, važi $\|I\| = 1$, gde je I jedinična matrica.

Može se dokazati (videti, na primer, [4]) da za svaku normu vektora $\|\boldsymbol{x}\|$ postoji bar jedna potčinjena norma $\|A\|$. Na primer, norma matrice A, data sa

$$\|A\| = \max_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\| = \max_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|}, \tag{3.6.5}$$

potčinjena je upotrebljenoj normi vektora.

Na osnovu (3.6.5) možemo, na primer, naći potčinjenu normu matrice, za euklidsku normu vektora. Tako imamo

$$\|A\| = \max_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|A\boldsymbol{x}\|_E}{\|\boldsymbol{x}\|_E}.$$

Kako je $\|A\boldsymbol{x}\|_E^2 = (A\boldsymbol{x})^*(A\boldsymbol{x}) = \boldsymbol{x}^*A^*A\boldsymbol{x}$ i matrica A^*A hermitska, imamo

$$\|A\boldsymbol{x}\|_E^2 \leq \lambda_{\max}\|\boldsymbol{x}\|_E^2. \tag{3.6.6}$$

gde je $\lambda_{\max}$ najveća sopstvena vrednost matrice A^*A (videti teoremu 3.4.6). Nejednakost (3.6.6) sugeriše sledeću definiciju.

Definicija 3.6.4. Pod spektralnom normom kvadratne matrice A podrazumeva se broj $\sigma(A)$, dat pomoću

$$\sigma(A) = \|A\|_{\text{sp}} = \sqrt{\max \lambda(A^*A)}.$$

Za spektralnu normu važe sledeći rezultati.

Teorema 3.6.1. *Ako je matrica A regularna, tada je*

$$\sigma(A^{-1}) = \sqrt{\frac{1}{\min \lambda(A^*A)}}.$$

Teorema 3.6.2. *Spektralni radijus $\rho(A)$ nije veći od njene spektralne norme, tj. važi $\rho(A) \leq \sigma(A)$. Ako je matrica hermitska, tada je $\rho(A) = \sigma(A)$.*

Teorema 3.6.3. *Spektralna norma matrice je potčinjena euklidskoj normi vektora.*

Teorema 3.6.4. *Neka je A hermitska matrica. Tada spektralna norma $\sigma(A)$ ima najmanju vrednost od svih mogućih normi $\|A\|$, saglasnih sa nekom normom vektora.*

Pored spektralne norme u upotrebi su i sledeće norme matrice $A = [a_{ij}]_{n\times n}$:

1° $\quad \|A\|_1 = \max\limits_j \left(\sum\limits_{i=1}^{n} |a_{ij}| \right)$;

2° $\quad \|A\|_2 = \|A\|_F = \varepsilon(A) = \left(\sum\limits_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}$;

3° $\quad \|A\|_\infty = \max\limits_i \left(\sum\limits_{j=1}^{n} |a_{ij}| \right)$.

Norma $\|A\|_F$ je poznata kao FROBENIUSova[112] norma. Često se koristi i termin HILBERT-SCHMIDTova norma i u tom slučaju se koristi oznaka $\varepsilon(A)$.

Teorema 3.6.5. *Norma $\|A\|_1$ potčinjena je normi vektora $\|\boldsymbol{x}\|_1$.*

Teorema 3.6.6. HILBERT-SCHMIDT*ova norma $\varepsilon(A)$ za matricu reda n je saglasna sa euklidskom normom* (3.6.4)*, ali joj nije potčinjena za $n > 1$.*

Teorema 3.6.7. *Norma $\|A\|_\infty$ je potčinjena normi vektora $\|\boldsymbol{x}\|_\infty$.*

[112] FERDINAND GEORG FROBENIUS (1849 – 1917), poznati nemački matematičar.

Lako se može dokazati da za proizvoljnu kvadratnu matricu reda n važe sledeće nejednakosti:

$$\frac{1}{n}m(A) \le \|A\|_p \le m(A) \qquad (p = 1, 2, \infty),$$

$$\frac{1}{\sqrt{n}}\varepsilon(A) \le \|A\|_p \le \sqrt{n}\,\varepsilon(A) \quad (p = 1, \infty),$$

$$\frac{1}{\sqrt{n}}\sigma(A) \le \|A\|_p \le \sqrt{n}\,\sigma(A) \quad (p = 1, \infty),$$

$$m(A) \le n\sigma(A), \quad \varepsilon(A) \le \sqrt{n}\,\sigma(A), \quad \frac{1}{n}\|A\|_\infty \le \|A\|_1 \le n\|A\|_\infty,$$

gde je $m(A) = n\max\limits_{i,j}|a_{ij}|$.

Neka je $A = [a_{ij}]_{n\times n}$ data kvadratna matrica i $\boldsymbol{b} = [b_1 \ \cdots \ b_n]^T$ dati n–dimenzionalni vektor. Sledeća teorema se odnosi na sistem linearnih jednačina $A\boldsymbol{x} = \boldsymbol{b}$.

Teorema 3.6.8. *Ako je*

$$\min_{1\le i\le n}\left\{|a_{ii}| - \sum_{j\ne i}|a_{ij}|\right\} = q > 0, \tag{3.6.7}$$

tada sistem jednačina $A\boldsymbol{x} = \boldsymbol{b}$ ima jedinstveno rešenje i važi nejednakost

$$\|\boldsymbol{x}\|_\infty \le \frac{1}{q}\|\boldsymbol{b}\|_\infty. \tag{3.6.8}$$

Dokaz. Neka je ispunjen uslov (3.6.7) i neka je

$$\|\boldsymbol{x}\|_\infty = \max_{1\le i\le n}|x_i| = |x_m|.$$

Tada imamo

$$\|\boldsymbol{b}\|_\infty \ge |a_{mm}|\cdot|x_m| - \sum_{j\ne m}|a_{mj}|\cdot|x_m| \ge q\|\boldsymbol{x}\|_\infty,$$

što daje (3.6.8). Da je rešenje $\boldsymbol{x}$ jedinstveno sledi iz (3.6.8). Zaista, ako je $\boldsymbol{b}$ nula vektor, na osnovu (3.6.8) zaključujemo da je $\|\boldsymbol{x}\|_\infty = 0$, tj. $\boldsymbol{x} = \boldsymbol{0}$, što znači da sistem ima samo trivijalno rešenje. Odavde, dalje, zaključujemo da je determinanta sistema različita od nule, što znači da sistem ima jedinstveno rešenje za bilo koji vektor $\boldsymbol{b}$. □

Ako je za matricu A ispunjen uslov (3.6.7) kažemo da matrica ima *dominantnu* dijagonalu.

U nejednakosti (3.6.8) norma $\|\cdot\|_\infty$ može biti zamenjena normom $\|\cdot\|_1$ ili $\|\cdot\|_2$, ali u tim slučajevima i uslov (3.6.7) treba zameniti sa

$$\min_{1\le i\le n}|a_{ii}| - \max_{1\le i\le n}\left\{\sum_{i\neq j}|a_{ij}|\right\} = q > 0$$

ili

$$\min_{1\le i\le n}|a_{ii}| - \max_{1\le j\le n}\left(\sum_{i\neq j}|a_{ij}|\right)^{1/2}\min_{1\le i\le n}\left(\sum_{j\neq i}|a_{ij}|\right)^{1/2} = q > 0,$$

respektivno.

Pomoću regularne matrice H, moguće je date norme $\|\boldsymbol{x}\|$ i $\|A\|$ transformisati u $\|\boldsymbol{x}\|^H$ i $\|A\|^H$, tako da ove poslednje zadovoljavaju uslove za normu. Lako je proveriti da $\|\boldsymbol{x}\|^H$ i $\|A\|^H$ predstavljaju norme, gde su

$$\|\boldsymbol{x}\|^H = \|H^{-1}\boldsymbol{x}\| \quad \text{i} \quad \|A\|^H = \|H^{-1}AH\|.$$

Štaviše, svojstva saglasnosti i potčinjenosti ostaju u važnosti i za transformisane norme $\|\boldsymbol{x}\|^H$ i $\|A\|^H$.

Kao transformaciona matrica H, najčešće se koristi dijagonalna matrica sa pozitivnim elementima, tj.

$$H = \begin{bmatrix} h_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & h_n \end{bmatrix} = \operatorname{diag}(h_1,\ldots,h_n) \qquad (h_i > 0 \ \ (i = 1,\ldots,n)).$$

Na ovaj način, norme $\|\boldsymbol{x}\|_1$, $\|A\|_1$ i norme $\|\boldsymbol{x}\|_\infty$, $\|A\|_\infty$ transformišu se u

$$\|\boldsymbol{x}\|_1^H = \sum_{i=1}^n h_i^{-1}|x_i|, \quad \|A\|_1^H = \max h_j\left(\sum_{i=1}^n h_i^{-1}|a_{ij}|\right)$$

i

$$\|\boldsymbol{x}\|_\infty^H = \max_i h_i^{-1}|x_i|, \quad \|A\|_\infty^H = \max_i h_i^{-1}\left(\sum_{i=1}^n h_j|a_{ij}|\right)$$

respektivno.

2.3.7 Konvergencija matričnih nizova i redova

Posmatrajmo niz vektora $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$, gde je $\boldsymbol{x}^{(k)} = [x_1^{(k)} \ \cdots \ x_n^{(k)}]^T$.

Definicija 3.7.1. Ako postoje granične vrednosti

$$a_i = \lim_{k\to+\infty} x_i^{(k)} \quad (i = 1,\ldots,n),$$

kažemo da niz $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ konvergira ka vektoru $\boldsymbol{a} = [a_1 \ \cdots \ a_n]^T$, za koji kažemo da je granična vrednost niza $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$.

Slično se može definisati i konvergencija niza matrica $\{A^{(k)}\}_{k\in\mathbb{N}}$, gde su članovi ovog niza matrice oblika $A^{(k)} = [a_{ij}^{(k)}]_{n\times n}$.

Definicija 3.7.2. Ako postoje granične vrednosti

$$a_{ij} = \lim_{k\to+\infty} a_{ij}^{(k)} \quad (i,j = 1,\ldots,n),$$

kažemo da niz $\{A^{(k)}\}_{k\in\mathbb{N}}$ konvergira ka matrici $A = [a_{ij}]_{n\times n}$. Za matricu A kažemo da je granična vrednost niza matrica $\{A^{(k)}\}_{k\in\mathbb{N}}$.

Mogu se dati i druge definicije konvergencije niza vektora i niza matrica zasnovane na ranije uvedenom pojmu norme (konvergencija po normi). Naime, kažemo da $\boldsymbol{x}^{(k)} \to \boldsymbol{a}$ $(k \to +\infty)$ ako $\|\boldsymbol{x}^{(k)} - \boldsymbol{a}\| \to 0$, kada $k \to +\infty$. Slično, kažemo da $A^{(k)} \to A$ $(k \to +\infty)$, ako $\|A^{(k)} - A\| \to 0$, kada $k \to +\infty$. Može se, međutim, pokazati da je definicija konvergencije po normi ekvivalentna prethodno datoj definiciji po koordinatama (videti, na primer [12]).

Napomena 3.7.1. Na osnovu nejednakosti

$$\big|\, \|\boldsymbol{x}^{(k)}\| - \|\boldsymbol{a}\| \,\big| \le \|\boldsymbol{x}^{(k)} - \boldsymbol{a}\|$$

zaključujemo da iz $\boldsymbol{x}^{(k)} - \boldsymbol{a}$ $(k \to +\infty)$ sledi $\|\boldsymbol{x}^{(k)}\| \to \|\boldsymbol{a}\|$ $(k \to +\infty)$. Takođe, važi implikacija

$$A^{(k)} \to A \ (k \to +\infty) \ \Rightarrow \ \|A^{(k)}\| \to \|A\| \ (k \to +\infty).$$

Teorema 3.7.1. *Niz matrica $\{A^{(k)}\}_{k\in\mathbb{N}}$ konvergira ka nula matrici ako i samo ako su sve sopstvene vrednosti matrice A po modulu manje od jedinice.*

Dokaz. S obzirom na to da se matrica A može transformisati na JORDANov oblik, tj. da postoji regularna matrica C, takva da je

$$J = C^{-1}AC,$$

gde je J JORDANova matrica (videti odeljak 2.3.5), za svako $k \in \mathbb{N}$ imamo

$$A^k = CJ^kC^{-1},$$

odakle zaključujemo da niz $\{A^{(k)}\}_{k\in\mathbb{N}}$ konvergira (ne konvergira) ka nula matrici ako i samo ako niz $\{J^{(k)}\}_{k\in\mathbb{N}}$ konvergira (ne konvergira) ka nula matrici.

Neka su $\lambda_1, \ldots, \lambda_m$ $(m \leq n)$ međusobno različite sopstvene vrednosti matrice A (matrice J). Kako, na osnovu napomene 3.5.1 (sa $Q(x) = x^k$), važe ekvivalencije

$$\lim_{k\to+\infty} J_r(\lambda)^k = 0 \;\Leftrightarrow\; |\lambda| < 1$$

i

$$(\forall i) \quad \lim_{k\to+\infty} J^k = 0 \;\Leftrightarrow\; \lim_{k\to+\infty} J_{r_i}(\lambda_i)^k = 0,$$

dokaz je završen. □

Sledeća teorema daje dovoljan uslov za konvergenciju niza $\{A^{(k)}\}_{k\in\mathbb{N}}$ ka nula matrici.

Teorema 3.7.2. *Ako je bilo koja norma matrice A manja od jedinice, tada važi $A^k \to 0$, kada $k \to +\infty$.*

Dokaz. Kako je

$$\|A^k - 0\| = \|A^k\| = \|AA^{k-1}\| \leq \|A\| \cdot \|A^{k-1}\| \leq \cdots \leq \|A\|^k$$

i kako je po pretpostavci $\|A\| < 1$, imamo $\|A^k\| \to 0$, kada $k \to +\infty$, tj. $A^k \to 0$ $(k \to +\infty)$. □

Sada ćemo dokazati jedan važan rezultat koji je u vezi sa teoremama 3.6.2 i 3.6.4.

Teorema 3.7.3. *Spektralni radijus $\rho(A)$ matrice A nije veći od bilo koje njene norme.*

Dokaz. Za proizvoljan pozitivan broj ε definišimo matricu B pomoću

$$B = \frac{1}{\|A\| + \varepsilon} A.$$

Kako je $\|B\| = \|A\|/(\|A\| + \varepsilon) < 1$, iz teoreme 3.7.2 sledi $B^k \to 0 \ \ (k \to +\infty)$. S druge strane, na osnovu teoreme 3.7.1 zaključujemo da su sve sopstvene vrednosti matrice B po modulu manje od jedinice, tj. $|\lambda_i(B)| < 1$. Ako sa $\lambda_i(A)$ označimo proizvoljnu sopstvenu vrednost matrice A, tada je

$$|\lambda_i(B)| = \frac{1}{\|A\| + \varepsilon} |\lambda_i(A)| < 1,$$

tj. $|\lambda_i(A)| < \|A\| + \varepsilon$. S obzirom na to da se ε može uzeti dovoljno malo zaključujemo da je $|\lambda_i(A)| \leq \|A\|$, tj. $\rho(A) \leq \|A\|$. $\square$

Oslanjajući se na koncept konvergencije matričnog niza, moguće je definisati matrični red pomoću

$$\sum_{m=0}^{+\infty} B^{(m)} = \lim_{k \to +\infty} \sum_{m=0}^{k} B^{(m)} \tag{3.7.1}$$

gde su $B^{(m)}$, $m = 0, 1, \ldots$, matrice istog reda.

Teorema 3.7.4. *Ako matrični red* (3.7.1) *konvergira, tada je*

$$\lim_{k \to +\infty} B^{(k)} = 0.$$

Dokaz. Neka su parcijalne sume reda (3.7.1)

$$S^{(k)} = \sum_{m=0}^{k} B^{(m)}, \quad k = 0, 1, \ldots,$$

i neka je njegova suma jednaka S, tj. $\lim\limits_{k \to +\infty} S^{(k)} = S$. Tada je

$$S^{(k)} - S^{(k-1)} = B^{(k)},$$

tj.

$$\lim_{k \to +\infty} S^{(k)} - \lim_{k \to +\infty} S^{(k-1)} = \lim_{k \to +\infty} B^{(k)},$$

odakle neposredno sledi tvrđenje teoreme. $\square$

U daljem tekstu daćemo potrebne i dovoljne uslove za konvergenciju matrične geometrijske progresije

$$\sum_{m=0}^{+\infty} A^m = I + A + A^2 + \cdots, \tag{3.7.2}$$

gde je A kvadratna matrica reda n. Ovde je $B^{(m)} = A^m$ $(m = 0, 1, \dots)$.

Teorema 3.7.5. *Matrični red* (3.7.2) *konvergira ako i samo ako je*

$$\lim_{k \to +\infty} A^k = 0. \tag{3.7.3}$$

Štaviše, tada je

$$I + A + A^2 + \cdots = (I - A)^{-1}. \tag{3.7.4}$$

Dokaz. Pretpostavimo da matrični red (3.7.2) konvergira. Tada, na osnovu teoreme 3.7.4, važi (3.7.3).

Obrnuto, pretpostavimo da je ispunjen uslov (3.7.3). Tada su, na osnovu teoreme 3.7.4, sve sopstvene vrednosti matrice A po modulu manje od jedinice, tj. $|\lambda_i(A)| < 1$ $(i = 1, \dots, n)$. Kako je

$$\det(I - A) = \prod_{i=1}^{n} (1 - \lambda_i(A)) \neq 0,$$

zaključujemo da postoji matrica $(I - A)^{-1}$. Množenjem identiteta

$$(I + A + A^2 + \cdots + A^k)(I - A) = I - A^{k+1}$$

matricom $(I - A)^{-1}$ s desne strane, dobijamo

$$I + A + A^2 + \cdots + A^k = (I - A)^{-1} - A^{k+1}(I - A)^{-1},$$

odakle sleduje

$$\begin{aligned} \lim_{k \to +\infty} (I + A + A^2 + \cdots + A^k) &= (I - A)^{-1} - \lim_{k \to +\infty} A^{k+1}(I - A)^{-1} \\ &= (I - A)^{-1}, \end{aligned}$$

tj. (3.7.4), čime je dokaz teoreme završen. □

S obzirom na tvrđenje teoreme 3.7.1, prvi deo teoreme 3.7.5 se može formulisati i na sledeći način.

Teorema 3.7.6. *Matrični red* (3.7.2) *konvergira ako i samo ako su sve sopstvene vrednosti matrice A po modulu manje od jedinice.*

Literatura

1. S. ALJANČIĆ, *Uvod u realnu i funkcionalnu analizu,* Građevinska knjiga, Beograd, 1979 (novo izdanje: Zavod za udžbenike, Beograd, 2011).
2. R. G. BERTLE, *A Modern Theory of Integration*, Graduate Studies in Mathematics, Vol. 32, American Mathematical Society, Providence, RI, 2001.
3. P. L. CHEBYSHEV, *Théorie des mécanismes connus sous le nom de parallélogrammes*. Mém. Acad. sci. St.-Pétersbourg **7** (1854), 539–564 [Œuvres, vol. 2, AN SSSR, Moscow – Leningrad, 1948, pp. 23–51].
4. L. COLLATZ, *Funktionanalysis und Numerische Mathematik,* Springer-Verlag, Berlin-Heidelberg-New York, 1964.
5. J. DIEUDONNÉ, *Foundations of Modern Analysis,* Academic Press, New York 1969.
6. L. FOX, I. B. PARKER, *Chebyshev Polynomials in Numerical Analysis*. Oxford University Press, London, 1968.
7. W. GAUTSCHI, *Numerical Analysis. An Introduction,* Birkhäuser Boston, Inc., Boston, MA, 1997.
8. W. GAUTSCHI, *Orthogonal Polynomials: Computation and Approximation,* Numerical Mathematics and Scientific Computation, Oxford Science Publications, Oxford University Press, New York, 2004.
9. F. R. GANTMACHER, *Teorija matrica,* Nauka, Moskva, 1988 (na ruskom).
10. G. H. GOLUB, CH. F. VAN LOAN, *Matrix Computations,* Fourth edition, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 2013.
11. L. V. KANTOROVICH, G. P. AKILOV, *Functional Analysis*, Pergamon Press, Oxford-Elmsford, N.Y., 1982 (originalno izdanje na ruskom: Nauka, Moskva, 1977).
12. V. I. KRYLOV, V. V. BOBKOV, P. I. MONASTYRNYIĬ, *Numerical Methods of Higher Mathematics. Vol. 1,* Izdat. "Vyšéĭšaja Škola", Minsk, 1972.
13. S. KUREPA, *Konačno-dimenzionalni vektorski prostori i primjene,* Tehnička knjiga, Zagreb, 1967.
14. S. KUREPA, *Funkcionalna analiza – Elementi teorije operatora,* Školska knjiga, Zagreb, 1981.
15. A. J. LAUB, *Computational Matrix Analysis,* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2012.
16. J. C. MASON, D. C. HANDSCOMB, *Chebyshev Polynomials,* Chapman & Hall/CRC, Boca Raton, FL, 2003.
17. G. MASTROIANNI, G.V. MILOVANOVIĆ, *Interpolation Processes – Basic Theory and Applications,* Springer Monographs in Mathematics, Springer – Verlag, Berlin –Heidelberg, 2008.
18. G.V. MILOVANOVIĆ, *Numerička analiza, I deo,* Naučna knjiga, Beograd, 1985.
19. G.V. MILOVANOVIĆ, *Ekstremalni problemi i nejednakosti za polinome,* Zavod za udžbenike, Beograd, 2012.
20. G.V. MILOVANOVIĆ, R.Ž. ĐORĐEVIĆ, *Linearna algebra,* Elektronski fakultet u Nišu, Niš, 2004.
21. G.V. MILOVANOVIĆ, R.Ž. ĐORĐEVIĆ, *Matematička analiza I,* Elektronski fakultet u Nišu, Niš, 2005.
22. G.V. MILOVANOVIĆ, D.S. MITRINOVIĆ, TH.M. RASSIAS, *Topics in Polynomials: Extremal Problems, Inequalities, Zeros,* World Scientific Publ. Co., Singapore – New Jersey – London – Hong Kong, 1994.
23. D. S. MITRINOVIĆ, D. Ž. ĐOKOVIĆ, *Polinomi i matrice,* ICS, Beograd 1975.
24. S. PASZKOWSKI, *Numerical Applications of the Chebyshev Polynomials and Series*. Nauka, Moscow, 1983 (na ruskom).
25. T. J. RIVLIN, *The Chebyshev Polynomials*. John Wiley & Sons, New York, 1974.

26. N. TEOFANOV, *Predavanja iz primenjene analize,* Zavod za udžbenike, Beograd, 2011.
27. N. I. VASIL'EV, YU. A. KLOKOV, A. YA. SHKERSTENA, *Primena Čebiševljevih polinoma u numeričkoj analizi,* Zinatne, Riga, 1984 (na ruskom).
28. V. V. VOEVODIN, *Linearna algebra,* Nauka, Moskva, 1980 (na ruskom).
29. X. ZHAN, *Matrix Theory,* Graduate Studies in Mathematics, Vol. 147, American Mathematical Society, Providence, RI, 2013.

3. OPŠTA TEORIJA ITERATIVNIH PROCESA

3.1 REŠAVANJE OPERATORSKIH JEDNAČINA

3.1.1 Osnovne napomene o rešavanju operatorskih jednačina

Ovaj i naredni odeljci u ovom poglavlju su posvećeni problemu egzistencije rešenja operatorskih jednačina u BANACHovom prostoru i definiciji iterativnih procesa.

Neka su X i Y BANACHovi prostori, D konveksan podskup prostora X i $F : D \to Y$. Posmatrajmo operatorsku jednačinu

$$Fu = \theta, \tag{1.1.1}$$

gde je θ nula–vektor prostora Y.

Veliki broj problema u nauci i tehnici svodi se na rešavanje jednačine oblika (1.1.1). Navešćemo nekoliko primera.

Primer 1.1.1. Ako su $X = Y = \mathbb{R}$, $u = x$, $F = f$, nelinearna jednačina

$$f(x) = x - \cos x = 0,$$

kao i algebarska jednačina

$$f(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n = 0$$

su oblika (1.1.1). $\triangle$

Primer 1.1.2. Ako su $X = Y = \mathbb{R}^n$, $u = \boldsymbol{x} = [x_1 \ \cdots \ x_n]^T$ i

$$Fu = F(\boldsymbol{x}) = \begin{bmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_n(x_1, \ldots, x_n) \end{bmatrix},$$

gde su $f_i : \mathbb{R}^n \to \mathbb{R}$ date funkcije, jednačina (1.1.1) predstavlja sistem nelinearnih jednačina

$$f_i(x_1,\dots,x_n) = 0, \qquad i = 1,\dots,n.$$

Ako je F linearan operator, na primer, $F(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b}$, gde su matrica A i vektor $\boldsymbol{b}$ dati sa

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & & a_{nn} \end{bmatrix} \quad \text{i} \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix},$$

jednačina (1.1.1) predstavlja sistem linearnih algebarskih jednačina

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = b_i, \qquad i = 1,\dots,n. \quad \triangle$$

Primer 1.1.3. Neka su $X = C^2[a,b]$, $Y = C[a,b] \times \mathbb{R}$, $u \equiv u(t)$,

$$Fu = \begin{bmatrix} f_1(u) \\ f_2(u) \end{bmatrix}$$

i

$$f_1(u)(t) = u''(t) - f(t,u(t),u'(t)) \quad (t \in [a,b]), \quad f_2(u) = g(u(a),u(b)),$$

gde su $F : \mathbb{R}^3 \to \mathbb{R}$ i $g : \mathbb{R}^2 \to \mathbb{R}$ date funkcije. Tada rešenje operatorske jednačine (1.1.1) predstavlja rešenje konturnog problema

$$u''(t) = f(t,u(t),u'(t)) \quad (t \in [a,b]),$$
$$g(u(a),u(b)) = 0. \quad \triangle$$

Sve jednačine navedene u prethodnim primerima, kao i niz drugih, mogu se na jedinstven način tretirati. Zato je predmet našeg razmatranja ovde rešavanje operatorske jednačine (1.1.1), tj. nalaženje takve tačke $u \in D$, koja zadovoljava (1.1.1). U tom cilju ovu jednačinu predstavimo u ekvivalentnom obliku

$$u = Tu \tag{1.1.2}$$

tako da operator T preslikava D u D, tj. da je $Tu = H(u,Fu)$, gde operator H preslikava $D \times Y$ u D. Za jednačinu (1.1.1), oblik (1.1.2) očigledno nije jedinstven, što pokazuje sledeći primer.

Primer 1.1.4. Jednačina $f(x) = 0$ se može predstaviti u ekvivalentnom obliku

$$x = x + \lambda f(x) \tag{1.1.3}$$

za svako λ različito od nule, ali postoje i mnogi drugi ekvivalentni oblici različiti od (1.1.3). $\triangle$

Jedan od načina za rešavanje jednačine (1.1.2), tj. jednačine (1.1.1), zasniva se na konstrukciji niza $\{u_k\}_{k\in\mathbb{N}_0}$ pomoću

$$u_{k+1} = Tu_k, \quad k = 0, 1, \ldots, \tag{1.1.4}$$

polazeći od neke tačke $u_0 \in D$. Pod izvesnim uslovima za T, niz $\{u_k\}_{k\in\mathbb{N}_0}$ može konvergirati ka traženom rešenju jednačine (1.1.2), o čemu će biti reči u sledećem odeljku.

Formulu (1.1.4), pomoću koje se generiše niz $\{u_k\}_{k\in\mathbb{N}}$ zvaćemo *iterativnim procesom*.

Pored iterativnih procesa oblika (1.1.4) mogu se razmatrati i opštiji iterativni procesi oblika

$$u_{k+1} = S(u_k, u_{k-1}, \ldots, u_{k-m+1}), \quad k = m-1, m, \ldots,$$

pri čemu $S : X^m \to X$. Ovakvi procesi, za koje kažemo da su procesi sa memorijom dužine m, za startovanje zahtevaju m početnih vrednosti $u_0, u_1, \ldots, u_{m-1} \in D$.

Naredni odeljak posvećujemo najjednostavnijem slučaju rešavanja obične jednačine $f(x) = 0$, da bismo zatim razmatrali opšti slučaj operatorske jednačine u BANACHovom prostoru.

3.1.2 Iterativni procesi za rešavanje običnih jednačina

Neka je data jednačina

$$f(x) = 0, \tag{1.2.1}$$

gde $f : [\alpha, \beta] \to \mathbb{R}$, i neka je

$$x = \phi(x) \tag{1.2.2}$$

njen ekvivalentni oblik.

Teorema 1.2.1. *Neka je ϕ realna funcija definisana i neprekidna na konačnom segmentu $[\alpha,\beta] \subset \mathbb{R}$ i neka njene vrednosti $\phi(x) \in [\alpha,\beta]$ za svako $x \in [\alpha,\beta]$. Tada postoji tačka a u $[\alpha,\beta]$ takva da je $a = \phi(a)$.*

Dokaz. Neka je $f(x) = x - \phi(x)$. Kako je, za svako $x \in [\alpha,\beta]$, $\alpha \leq \phi(x) \leq \beta$, tj. $\alpha - \phi(x) \leq 0$ i $\beta - \phi(x) \geq 0$, zaključujemo da je $f(\alpha) = \alpha - \phi(\alpha) \leq 0$ i $f(\beta) = \beta - \phi(\beta) \geq 0$, tj. $f(\alpha)f(\beta) \leq 0$.

Razlikovaćemo dva slučaja, kada je $f(\alpha)f(\beta) = 0$ i kada je $f(\alpha)f(\beta) < 0$.

Prvi slučaj $f(\alpha)f(\beta) = 0$ je trivijalan. Tada očigledno postoji tačka $a = \alpha$ i/ili $a = \beta$ takva da je $\phi(a) = a$.

U slučaju kada su vrednosti $f(\alpha)$ i $f(\beta)$ suprotnog znaka, na osnovu neprekidnosti funkcije $f : [\alpha,\beta] \to \mathbb{R}$, zaključujemo da postoji tačka $a \in (\alpha,\beta)$ takva da je $f(a) = 0$ (videti, na primer, [10, teorema 2.3.3, str. 93], tj. $\phi(a) = a$. □

Tačka a sa osobinom da je $\phi(a) = a$ naziva se *fiksna* ili *nepokretna tačka* funkcije ϕ. Teorema 1.2.1, poznata kao BROUWERova teorema o fiksnoj tački[113], garantuje egzistenciju fiksne tačke, ali ne i njenu jedinstvenost.

Primer 1.2.1. Neka je na segmentu $[\alpha,\beta] = [1,4]$ definisana funkcija

$$\phi(x) = \frac{3}{2}\cos(2x) + \frac{5}{2}.$$

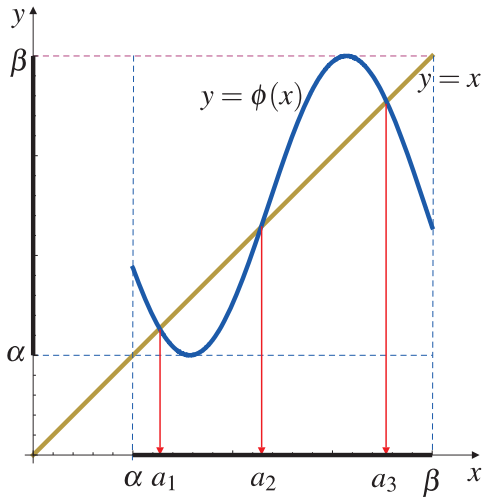
Kako $\phi(x) \in [1,4]$ kada $x \in [1,4]$ (videti grafik na slici 1.2.1), na osnovu teoreme 1.2.1, zaključujemo da postoji fiksna tačka a takva da je $\phi(a) = a$. Štaviše, sa slike vidimo da postoje tri takve tačke $a_1 \approx 1.26754$, $a_2 \approx 2.28391$ i $a_3 \approx 3.54269$. △

Kao što je ranije rečeno (videti primer 1.1.4), postoji beskonačno mnogo načina za ekvivalentno predstavljanje jednačine (1.2.1) u obliku (1.2.2). Međutim, shodno teoremi 1.2.1, neohodno je da taj ekvivalentan oblik obezbeđuje takvu funkciju ϕ koja preslikava segment $[\alpha,\beta]$ u samog sebe, ali kao što ćemo videti u daljem izlaganju to još uvek nije dovoljno za konvergenciju niza

$$x_{k+1} = \phi(x_k), \quad k = 0,1,\ldots, \tag{1.2.3}$$

koji se konstruiše pomoću funkcije ϕ, kako je rečeno na kraju prethodnog odeljka. Da bismo ovo ilustrovali razmotrićemo sledeći numerički primer.

[113] LUITZEN EGBERTUS JAN BROUWER (1881–1966), poznati holandski matematičar i filozof. U literaturi postoji na stotine teorema o fiksnoj tački. Ovde pomenuta BROUWERova teorema je najjednostavniji slučaj njegove mnogo opštije teoreme o fiksnoj tački u topologiji koja govori o egzistenciji fiksne tačke neprekidne funkcije koja slika konveksni kompaktni podskup EUKLIDovog prostora na samog sebe. Jedno proširenje tog rezultata na topološke vektorske prostore je poznata SCHAUDERova teorema. JULIUSZ PAWEŁSCHAUDER (1899 – 1943) je bio poznati poljski matematičar jevrejskog porekla.



Slika 1.2.1. Grafik funkcije ϕ na segmentu $[\alpha,\beta]=[1,4]$

Primer 1.2.2. Za funkciju ϕ iz primera 1.2.1 konstruišimo odgovarajuće nizove

$$x_{k+1}=\phi(x_k)=\frac{3}{2}\cos(2x_k)+\frac{5}{2},\quad k=0,1,\dots\ ,$$

Tabela 1.2.1.

k	$x_k^{(1)}$	$x_k^{(2)}$	$x_k^{(3)}$	$\widehat{x}_k^{(1)}$	$\widehat{x}_k^{(2)}$	$\widehat{x}_k^{(3)}$
0	1.3	2.3	3.5	1.3	2.3	3.5
1	1.21467	2.33177	3.63085	1.24905	2.28933	3.56213
2	1.36467	2.42676	3.33737	1.26117	2.28933	3.53360
3	1.21467	2.71099	3.88647	1.24905	2.28453	3.54689
4	1.55617	3.47729	2.62143	1.27128	2.28412	3.54074
5	1.00064	3.67443	3.25891	1.26536	2.28398	3.54360
6	1.87403	3.22587	3.95890	1.26882	2.28394	3.54227
7	1.26750	3.97874	2.40434	1.26679	2.28392	3.54289
8	1.26760	2.34502	2.64421	1.26798	2.28392	3.54260
9	1.26743	2.46648	3.31705	1.26728	2.28392	3.54273
10	1.26773	2.82818	3.90858	1.26769	2.28391	3.54267

sa startnim vrednostima 1.3, 2.3, 3.5, koje su bliske vrednostima nula (fiksnim tačkama) a_1, a_2, a_3, respektivno. Dobijene nizove označili smo redom sa

$\{x_k^{(\nu)}\}$, $\nu = 1,2,3$, i prikazali u tabeli 1.2.1. Sva izračunavanja su obavljena u dvostrukoj preciznosti (mašinski epsilon $e_M \approx 2.22 \times 10^{-16}$), dok su u tabeli rezultati prikazani sa 6 značajnih cifara zbog uštede prostora.

Na osnovu dobijenih numeričkih rezultata može se zaključiti da nizovi divergiraju, sa izvesnom sumnjom kod prvog niza $\{x_k^{(1)}\}$. Međutim, narednih deset članova tog niza: 1.26721, 1.26810, 1.26659, 1.26917, 1.26476, 1.27231, 1.25943, 1.28156, 1.24405, 1.30905, ipak pokazuju njegovu divergenciju. $\triangle$

Da bi se obezbedila konvergencija niza (1.2.3), tj. *metoda proste iteracije*, kako se često naziva, dovoljno je pretpostaviti da funkcija ϕ iz BROUWERove teoreme 1.2.1 bude *kontrakcija* na segmentu $[\alpha,\beta]$, tj. da zadovoljava LIPSCHITZov[114] uslov, sa konstantom L manjom od jedinice. To znači da postoji konstanta L takva da je $0 < L < 1$ i da je

$$(\forall x,y \in [\alpha,\beta]) \quad |\phi(x)-\phi(y)| \le L|x-y|. \tag{1.2.4}$$

Drugi rečima, funkcija $\phi : [\alpha,\beta] \to [\alpha,\beta]$ treba da bude takva da je, za bilo koje dve tačke $x,y \in [\alpha,\beta]$, rastojanje između slika $|\phi(x)-\phi(y)|$ uvek manje od rastojanja između originala $|x-y|$. Za takvo preslikavanje kažemo da je *kontrabilno* ili da je *kontrakcija*. Konstanta L se naziva LIPSCHITZ*ova konstanta*.

U slučaju kontrakcije, konstantu L obično označavamo sa q $(0 < q < 1)$, a za funkciju ϕ kažemo da je *iterativna funkcija*.

Ako iterativna funkcija ϕ ima izvod u svakoj tački $x \in [\alpha,\beta]$, takav da je

$$|\phi'(x)| \le q < 1, \tag{1.2.5}$$

na osnovu LAGRANGEove[115] formule važi

$$\phi(x)-\phi(y) = \phi'(\xi)(x-y),$$

gde su x i y proizvoljne tačke iz $[\alpha,\beta]$ i $\xi = y+\theta(x-y)$ $(0 < \theta < 1)$. Kako je

$$|\phi(x)-\phi(y)| = |\phi'(\xi)| \cdot |x-y| \le q|x-y|,$$

zaključujemo da ϕ zadovoljava LIPSCHITZov uslov (1.2.4), sa konstantom manjom od jedinice, tj. ϕ je kontrakcija na $[\alpha,\beta]$.

[114] RUDOLF OTTO SIGISMUND LIPSCHITZ (1832 – 1903), poznati nemački matematičar.
[115] JOSEPH-LOUIS LAGRANGE (1736 – 1813), poznati francuski matematičar i astronom.

Teorema 1.2.2. *Pretpostavimo da je funkcija* $\phi : [\alpha,\beta] \to [\alpha,\beta]$ *neprekidna i da zadovoljava* LIPSCHITZ*ov uslov* (1.2.4) *sa konstantom manjom od jedinice. Tada jednačina* (1.2.2) *ima jedinstveno rešenje* $a \in [\alpha,\beta]$ *i ono se može odrediti iterativnim procesom*

$$x_{k+1} = \phi(x_k), \quad k = 0,1,\ldots, \tag{1.2.6}$$

sa proizvoljnim $x_0 \in [\alpha,\beta]$.

Dokaz. Na osnovu BROUWERove teoreme 1.2.1, funkcija ϕ ima fiksnu tačku $a \in [\alpha,\beta]$ takvu da je $a = \phi(a)$. Uslov kontrakcije obezbeđuje jedinstvenost takve tačke. Zaista, ako pretpostavimo da postoji još jedna fiksna tačka $b \in [\alpha,\beta]$, različita od a, tada

$$0 < |a-b| = |\phi(a) - \phi(b)| \leq q|a-b|$$

tj. $(1-q)|a-b| \leq 0$. S obzirom na to da je $1-q > 0$, zaključujemo da mora biti $a = b$, tj. da je fiksna tačka jedinstvena.

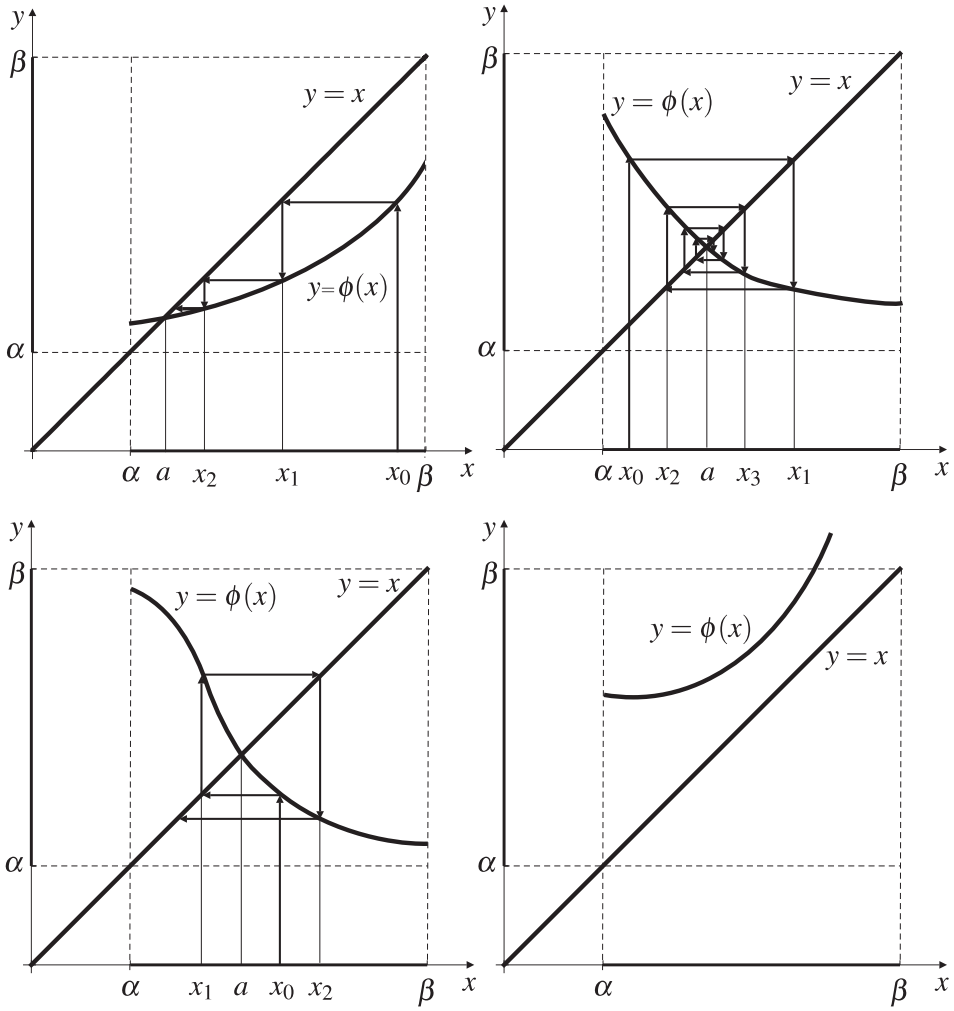
Da bismo dokazali konvergenciju iterativnog procesa (1.2.6) ka jedinstvenoj fiksnoj tački $a \in [\alpha,\beta]$, sa proizvoljnim $x_0 \in [\alpha,\beta]$, razmotrimo rastojanje proizvoljnog člana x_k (iteracije) od fiksne tačke, $|x_k - a|$, uz korišćenje kontrabilnosti funkcije ϕ. Tada, imamo

$$|x_k - a| = |\phi(x_{k-1}) - \phi(a)| \leq q|x_{k-1} - a|, \quad k \geq 1,$$

odakle zaključujemo da je, za svako $k \geq 1$, $|x_k - a| \leq q^k |x_0 - a|$. Kako je $0 < q < 1$ i $\lim\limits_{k\to+\infty} q^k = 0$, iz poslednje nejednakosti sledi $\lim\limits_{k\to+\infty} |x_k - a| = 0$, tj.

$$\lim_{k\to+\infty} x_k = a. \quad \square$$

Dakle, da bi iterativni proces (1.2.6) konvergirao ka jedinstvenoj nepokretnoj tački a, iterativna funkcija ϕ mora ispunjavati određene uslove. Uslov kontrakcije se može zameniti nešto strožijim uslovom (1.2.5), ali jednostavnijim za proveru. Na slikama 1.2.2 je data geometrijska interpretacija iterativnih procesa oblika (1.2.6). Prva dva procesa (slike iznad) su konvergentna. Interesantno je primetiti da je kod drugog od njih $\phi'(x) < 0$ i da u tom slučaju niz $\{x_k\}$ konvergira oscilatorno, tj. greška $e_k = x_k - a$ alternativno menja znak. Druga dva procesa (slike ispod) su divergentna. Kod prvog od njih nije ispunjen uslov (1.2.5), a kod drugog $\phi(x) \notin [\alpha,\beta]$, tj. funkcija ϕ nema fiksnu tačku.



Slika 1.2.2. Geometrijska interpretacija iterativnih procesa oblika (1.2.6)

Primer 1.2.3. Neka je data jednačina

$$x^3 + x - 60 = 0. \tag{1.2.7}$$

Nije teško utvrditi (na primer, grafički) da ova jednačina ima koren $x = a$ koji leži u intervalu $(3,4)$. Da bismo rešili datu jednačinu, treba je prethodno svesti na oblik (1.2.2). Na primer, neki od tih oblika su

$$x = \phi_1(x) = 60 - x^3,$$
$$x = \phi_2(x) = \sqrt[3]{60 - x},$$
$$x = \phi_3(x) = \frac{60}{x^2} - \frac{1}{x}.$$

Neposrednim proveravanjem zaključujemo da od navedenih iterativnih funkcija samo ϕ_2 zadovoljava uslove teoreme 1.2.2, pri čemu je

$$|\phi_2'(x)| = \left| \frac{1}{3\sqrt[3]{(60-x)^2}} \right| \leq \frac{1}{3(56)^{2/3}} \cong 0.022$$

kada $x \in [3,4]$.

Koren jednačine (1.2.7) može se odrediti iterativnim procesom

$$x_{k+1} = \sqrt[3]{60 - x_k}, \quad k = 0, 1, \ldots \ . \tag{1.2.8}$$

Polazeći od $x_0 = 4$, pomoću (1.2.8) dobijamo

k	x_k
0	4.
1	3.8258623
2	3.8298239
3	3.8297338
4	3.8297359
5	3.8297358

tj. $a \cong 3.8297358$. $\triangle$

Napomena 1.2.1. Ako je $|\phi'(x)| \geq p > 1$, kada $x \in [\alpha, \beta]$, iterativni proces (1.2.6) divergira. Međutim, ako se jednačina (1.2.2) napiše u obliku

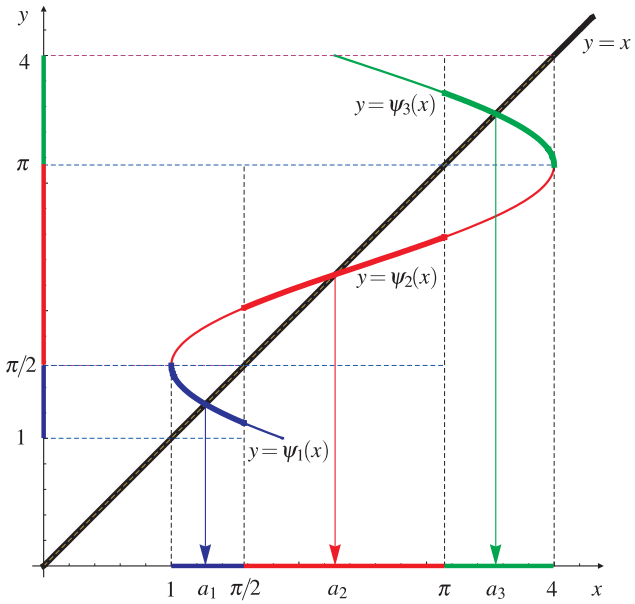
$$x = \psi(x),$$

gde je ψ inverzna funkcija od ϕ, iterativni proces

$$x_{k+1} = \psi(x_k) \qquad (k = 0, 1, \ldots)$$

biće konvergentan, s obzirom na činjenicu da je

$$|\psi'(x)| = \left| \frac{1}{\phi'(\psi(x))} \right| \leq \frac{1}{p} = q < 1.$$



Slika 1.2.3. Grafik funkcije ϕ na segmentu $[\alpha,\beta]=[1,4]$

Primer 1.2.4. Razmotrimo opet problem iz primera 1.2.1, gde smo pokazali da funkcija ϕ ima tri fiksne tačke na segmentu $[1,4]$, ali da se one ne mogu odrediti metodom proste iteracije, s obzirom na to da odgovarajući nizovi divergiraju (videti primer 1.2.2).

Na osnovu prethodnog razmatranja, za funkciju ϕ je neophodno odrediti inverznu funkciju, ali nažalost ona ne postoji u ovom slučaju. Međutim, na osnovu grafika sa slike 1.2.1, vidimo da se funkcija ϕ na $[1,4]$ može razmatrati posebno na svakom od njenih intervala monotonosti, tj. kao tri funkcije $\phi_\nu(x)=\phi(x)$, $x\in I_\nu$, $\nu=1,2,3$, gde su $I_1=[1,\pi/2]$, $I_2=[\pi/2,\pi]$, $I_3=[\pi,4]$. Svaka od tih funkcija ima odgovarajuću inverznu funkciju ψ_ν, $\nu=1,2,3$, čiji su grafici prikazani na slici 1.2.3, redom u plavoj, crvenoj i zelenoj boji.

Posmatrajmo sada tri nove iterativne funkcije, uzete kao restrikcije prethodno pomenutih inverznih funkcija, $\psi_\nu : I_\nu \to I_\nu$, $\nu=1,2,3$, tj.

$$\psi_1(x)=\psi(x),\ x\in I_1;\quad \psi_2(x)=\pi-\psi(x),\ x\in I_2;\quad \psi_3(x)=\pi+\psi(x),\ x\in I_3,$$

gde je

$$\psi(x)=\frac{1}{2}\arccos\frac{2x-5}{3}.$$

Grafici ovih funkcija su predstavljeni podebljanim linijama na istoj slici 1.2.3. Primetimo da je funkcija ψ neprekidna na $[1,4]$, kao i da ima neprekidan izvod u intervalu $(1,4)$, pri čemu je

$$\psi'(x) = \frac{-1}{2\sqrt{-x^2+5x-4}} \quad (1 < x < 4).$$

Jasno je da su i funkcije ψ_ν, $\nu = 1,2,3$, takođe neprekidne na odgovarajućim segmentma I_ν, $\nu = 1,2,3$. Štaviše, one su i diferencijabilne sa neprekidnim izvodom u unutrašnjosti ovih segmenata.

Na osnovu teoreme o fiksnoj tački, svaka od funkcija ψ_ν, $\nu = 1,2,3$, ima fiksnu tačku. Odgovarajući nizovi, sa istim početnim uslovima kao i u primeru 1.2.2,

$$\widehat{x}_{k+1}^{(\nu)} = \psi_\nu(\widehat{x}_k^{(\nu)}), \quad k = 0,1,\ldots,$$

pokazuju konvergenciju (videti poslednje tri kolone u tabeli 1.2.1, što se zaista može jednostavno proveriti, u ovom jednostavnom primeru, nalaženjem izvoda iterativne funkcije u odgovarajućim fiksnim tačkama $\psi_1'(a_1) \approx -0.585$, $\psi_2'(a_2) \approx 0.337$, $\psi_3'(a_3) \approx -0.464$. Kako su sve ove vrednosti po modulu manje od jedinice, zbog neprekidnosti prvih izvoda ψ_ν' postoje okoline tačaka a_ν u kojima su $|\psi_\nu'(x)| \le q < 1$, $\nu = 1,2,3$, tako da su funkcije ψ_ν kontrakcije i odgovarajući metodi proste iteracije konvergiraju ka odgovarajućim nepokretnim tačkama. $\triangle$

Na kraju ovog odeljka pretpostavimo da jednačina (1.2.1) ima u $[\alpha,\beta]$ jedinstveno rešenje $x = a$, gde je segment $[\alpha,\beta]$ tako odabran da f ne menja znak na njemu. Postavlja se pitanje kako preći na ekvivalentan oblik (1.2.2), a da pri tome funkcija ϕ zadovoljava uslove teoreme 1.2.2. Izložićemo sada jedan od načina za rešavanje ovog problema.

Neka je

$$0 < m \le f'(x) \le M \qquad (\alpha \le x \le \beta). \tag{1.2.9}$$

Ukoliko je $f'(x) < 0$, umesto jednačine (1.2.1), može se uzeti jednačina sa suprotnim znakom, tj. $-f(x) = 0$.

Na osnovu primera 1.1.4, izaberimo

$$\phi(x) = x - \lambda f(x)$$

i pozitivan parametar λ takav da je

$$0 \le \phi'(x) = 1 - \lambda f'(x) \le q < 1 \tag{1.2.10}$$

za svako $x \in [\alpha, \beta]$. Tada, na osnovu (1.2.9), imamo

$$0 \leq 1 - \lambda M \leq 1 - \lambda m \leq q,$$

odakle zaključujemo da će (1.2.10) biti ispunjeno ako je $\lambda = 1/M$. U tom slučaju imamo

$$q = \max_{\alpha \leq x \leq \beta} (1 - \lambda f'(x)) = 1 - \frac{m}{M} < 1,$$

što znači da se može uzeti $\phi(x) = x - \dfrac{1}{M} f(x)$.

3.1.3 BANACHov stav o nepokretnoj tački

U ovom odeljku dokazaćemo jedan vrlo važan rezultat o egzistenciji i jedinstvenosti rešenja operatorskih jednačina koji je poznat kao BANACHov stav o nepokretnoj tački i koji predstavlja uopštenje teoreme 1.2.2.

Neka je T operator u BANACHovom prostoru X. Tačke u za koje je

$$u = Tu \tag{1.3.1}$$

zovu se *nepokretne tačke* operatora T. Dovoljan uslov za egzistenciju jedinstvene nepokretne tačke operatora T dao je BANACH.

Definicija 1.3.1. Za operator $T : X \to X$ kažemo da je *kontrakcija* ako postoji pozitivan broj $q < 1$ takav da je za bilo koje dve tačke $u, v \in X$,

$$\|Tu - Tv\| \leq q\|u - v\|. \tag{1.3.2}$$

Nejednakost (1.3.2) iskazuje činjenicu da je kod kontraktivnog preslikavanja (kontrakcije) rastojanje između slika manje od rastojanja između originala.

Sledeća teorema je poznata kao BANACHov stav o nepokretnoj tački.

Teorema 1.3.1. *Kontrakcija T BANACHovog prostora X u samog sebe ima jednu i samo jednu nepokretnu tačku.*

Dokaz. Polazeći od proizvoljne tačke $u_0 \in X$ konstruišimo niz $\{u_k\}_{k \in \mathbb{N}}$ pomoću

$$u_{k+1} = Tu_k, \quad k = 0, 1, \ldots .$$

Kako je za $k \geq 1$,

$$\|u_{k+1} - u_k\| = \|Tu_k - Tu_{k-1}\| \leq q\|u_k - u_{k-1}\|,$$

imamo

$$\|u_{k+1}-u_k\| \le q^k \|u_1-u_0\|.$$

Ako je $m > k$, tada je

$$\begin{aligned}(1.3.3)\quad \|u_m-u_k\| &= \|(u_{k+1}-u_k)+(u_{k+2}-u_{k+1})+\cdots+(u_m-u_{m-1})\| \\ &\le \|u_{k+1}-u_k)\|+\|u_{k+2}-u_{k+1}\|+\cdots+\|u_m-u_{m-1}\| \\ &\le (q^k+q^{k+1}+\cdots+q^{m-1})\|u_1-u_0\| \\ &< \frac{q^k}{1-q}\|u_1-u_0\| \qquad (q<1),\end{aligned}$$

odakle je $\lim\limits_{k\to+\infty}\|u_m-u_k\|=0$, što znači da je $\{u_k\}_{k\in\mathbb{N}_0}$ CAUCHYev niz.

S obzirom na to da je X Banachov prostor, to je on kompletan pa postoji

$$(1.3.4)\qquad \lim_{k\to+\infty} u_k = a \qquad (a\in X).$$

S druge strane, kako je

$$\|u_{k+1}-Ta\| = \|Tu_k-Ta\| \le q\|u_k-a\|,$$

imamo $\lim\limits_{k\to+\infty}\|u_{k+1}-Ta\|=0$, tj. $\lim\limits_{k\to+\infty} u_{k+1}=Ta$.

Iz poslednje jednakosti i jednakosti (1.3.4) sledi $a=Ta$, tj. a je nepokretna tačka operatora T. Ovim je dokazana egzistencija nepokretne tačke.

Za dokaz jedinstvenosti pretpostavimo da postoje dve nepokretne tačke a i b, tj. neka su

$$a=Ta, \quad b=Tb, \quad \|a-b\|\neq 0.$$

Kako je T kontrakcija imamo

$$\|Ta-Tb\| = \|a-b\| \le q\|a-b\|,$$

sa konstantom q, $0<q<1$, odakle neposredno sledi $\|a-b\|=0$, tj. $a=b$. □

Dakle, ako je T kontrakcija X u X, jednačina $Tu=u$ ima jedno i samo jedno rešenje i ono može biti dobijeno kao granična vrednost niza, koji se generiše pomoću

$$(1.3.5)\qquad u_{k+1}=Tu_k, \quad k=0,1,\ldots,$$

gde je u_0 proizvoljna tačka iz X.

Ako se zadržimo na k-toj aproksimaciji u_k, koja je određena pomoću (1.3.5), iz (1.3.3), pri $m \to +\infty$, sleduje

$$\|a - u_k\| < \frac{q^k}{1-q}\|u_1 - u_0\|,$$

što znači da se ova aproksimacija nalazi u kugli sa centrom u tački a i poluprečnikom $r_k = \frac{q^k}{1-q}\|u_1 - u_0\|$. Očigledno je, da sa porastom k, poluprečnik kugle teži nuli.

Konstrukcija iterativnih procesa oblika (1.2.6) za rešavanje običnih jednačina, kao i iterativni procesi za rešavanje sistema nelinearnih jednačina biće tretirani u posebnoj glavi.

Zbog posebnog značaja koje zauzimaju sistemi linearnih jednačina, kao i drugi relevantni problemi u linearnoj algebri, posebna glava biće posvećena metodima linearne algebre, uključujući pri tome i iterativne procese u linearnoj algebri za rešavanje sistema linearnih jednačina, inverziju matrica, rešavanje problema sopstvenih vrednosti, itd.

Na kraju ovog odeljka daćemo ukratko neke napomene o rešavanju sistema linearnih jednačina

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{1.3.6}$$

gde su matrica A i vektori $\boldsymbol{x}$ i $\boldsymbol{b}$ dati kao u primeru 1.1.2. Kako se sistem jednačina (1.3.6) može predstaviti (na beskonačno mnogo načina) u obliku

$$\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{c}, \tag{1.3.7}$$

gde su $B = [b_{ij}]_{n\times n}$ i $\boldsymbol{c} = [c_1 \ \cdots \ c_n]^T$, može se konstruisati iterativni proces sa operatorom $\boldsymbol{x} \mapsto T(\boldsymbol{x}) = B\boldsymbol{x} + \boldsymbol{c}$.

S obzirom na činjenicu da je T linearan operator, na osnovu

$$\|T\boldsymbol{x} - T\boldsymbol{y}\| = \|B(\boldsymbol{x} - \boldsymbol{y})\| \le \|B\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\|$$

zaključujemo da je $\|B\| \le q < 1$ dovoljan uslov da je T kontrakcija. Dakle, ako je $\|B\| \le q < 1$, sistem jednačina (1.3.7), tj. (1.3.6), ima jedinstveno rešenje i odgovarajući iterativni proces

$$\boldsymbol{x}_{k+1} = T\boldsymbol{x}_k, \quad k = 0, 1, \ldots,$$

konvergira ka $\boldsymbol{a} = A^{-1}\boldsymbol{b}$. Napomenimo da vrednost $\|B\|$ zavisi od izbora metrike u $\mathbb{R}^n$.

3.2 KARAKTERISTIKE ITERATIVNIH PROCESA

U ovom poglavlju se definišu osnovne karakteristike iterativnih procesa – *red konvergencije* i *asimptotska konstanta greške*, a zatim se daju neki postupci za ubrzavanje konvergencije. U poslednjem odeljku uveden je pojam *R-reda konvergencije*.

3.2.1 Red konvergencije iterativnih procesa

Neka je X BANACHov prostor i operator $T : X \to X$ i neka niz

$$u_{k+1} = Tu_k, \quad k = 0, 1, \ldots, \tag{2.1.1}$$

konvergira ka tački $a \in X$.

Niz (2.1.1) definiše iterativni proces za rešavanje operatorske jednačine

$$u = Tu.$$

Kao što je rečeno u uvodnom odeljku 3.1.1, osim iterativnih procesa oblika (2.1.1), mogu se posmatrati i opštiji iterativni procesi oblika

$$u_{k+1} = S(u_k, u_{k-1}, \ldots, u_{k-m+1}), \quad k = m-1, m, \ldots, \tag{2.1.2}$$

pri čemi $S : X^m \to X$.

Definicija 2.1.1. Za iterativni proces koji konvergira ka a, kaže se da ima *red konvergencije* r ako je

$$\|u_{k+1} - a\| = O(\|u_k - a\|^r), \tag{2.1.3}$$

tj. ako postoji konstanta A takva da je za dovoljno veliko k

$$\|u_{k+1} - a\| \leq A\|u_k - a\|^r.$$

Označimo sa $U(a)$ konveksnu okolinu tačke a. Tada, za iterativni proces (2.1.1), umesto (2.1.3) u prethodnoj definiciji, možemo uzeti

$$\|Tu - a\| = O(\|u - a\|^r) \qquad (u \in U(a)).$$

Teorema 2.1.1. *Ako je operator $T : X \to X$ r-puta* FRÉCHET*-diferencijabilan u konveksnoj okolini $U(a)$, iterativni proces* (2.1.1) *je reda r ako su ispunjeni sledeći uslovi:*

1° $Ta = a$,

2° $T'_{(a)}, T''_{(a)}, \ldots, T^{(r-1)}_{(a)}$ *su nula–operatori,*

3° $T^{(r)}_{(a)}$ *je nenula–operator sa normom koja je ograničena na* $U(a)$.

Dokaz. Neka $u \in U(a)$, $\|T^{(r)}_{(u)}\| \leq M_r$ i

$$q = Ta + \frac{1}{1!}T'_{(a)}(u-a) + \cdots + \frac{1}{(r-1)!}T^{(r-1)}_{(a)}(\underbrace{u-a, \ldots, u-a}_{r-1 \text{ puta}}).$$

Tada na osnovu TAYLORove formule imamo

$$\|Tu - q\| \leq \frac{1}{r!} \sup_{t\in[0,1]} \|T^{(r)}_{(a+t(u-a))}\| \cdot \|u-a\|^r,$$

tj.

$$\|Tu - q\| \leq \frac{M_r}{r!}\|u-a\|^r.$$

Kako je, na osnovu pretpostavki teoreme, $q = Ta = a$, imamo

$$\|Tu - a\| \leq A\|u-a\|^r \qquad \left(A = \frac{M_r}{r!}\right). \quad \square$$

U daljem tekstu, kada govorimo o iterativnom procesu reda r, pretpostavljaćemo uvek da su ispunjeni uslovi teoreme 2.1.1.

Napomenimo da je red konvergencije iterativnih procesa tipa (2.1.1) uvek prirodan broj, a da je kod procesa tipa (2.1.2) realan broj (≥ 1).

U slučaju $r = 1$, dovoljan uslov za konvergenciju iterativnog procesa je dat sa $0 \leq A < 1$.

Posmatrajmo sada slučaj kada je $X = \mathbb{R}$.

Teorema 2.1.2. *Neka je*

$$x_{k+1} = \phi(x_k), \quad k = 0, 1, \ldots,$$

iterativni proces reda r, gde je $x \mapsto \phi(x)$ r puta neprekidno–diferencijabilna funkcija u okolini tačke $a = \lim_{k\to+\infty} x_k$. Tada je

$$\lim_{k\to+\infty}\frac{x_{k+1}-a}{(x_k-a)^r}=\frac{\phi^{(r)}(a)}{r!}. \tag{2.1.4}$$

Dokaz. Kako je $\phi(a)=a$ i $\phi^{(i)}(a)=0$ $(i=1,\dots,r-1)$, na osnovu TAYLORove formule dobijamo

$$x_{k+1}=\phi(x_k)=a+\frac{1}{r!}\phi^{(r)}(a+\theta(x_k-a))(x_k-a)^r, \tag{2.1.5}$$

gde je $0<\theta<1$. Kako je $a=\lim\limits_{k\to+\infty}x_k$ i $\phi^{(r)}$ neprekidna funkcija, deobom jednakosti (2.1.5) sa $(x_k-a)^r$ i prelaskom na graničnu vrednost zaključujemo da važi (2.1.4). □

Definicija 2.1.2. Veličina

$$C_r=\lim_{k\to+\infty}\frac{|x_{k+1}-a|}{|x_k-a|^r}$$

naziva se *faktor konvergencije* ili *asimptotska konstanta greške*.

Napomena 2.1.1. Iterativni proces kod koga je $r=1$ naziva se *proces sa linearnom konvergencijom*. Njegov faktor konvergencije C_1 mora biti manji od jedinice.

Primer 2.1.1. Neka se za izračunavanje vrednosti $a=\sqrt[m]{N}$ $(N>0)$ koristi iterativni proces

$$x_{k+1}=\phi(x_k),\quad k=0,1,\dots,$$

gde je

$$\phi(x)=x\frac{(m-1)x^m+(m+1)N}{(m+1)x^m+(m-1)N}.$$

Sukcesivnim diferenciranjem jednakosti

$$g(x)\phi(x)=(m-1)x^{m+1}+(m+1)Nx, \tag{2.1.6}$$

gde je

$$g(x)=(m+1)x^m+(m-1)N,$$

nalazimo redom

$$\begin{aligned}
g(x)\phi'(x) &= -m(m+1)x^{m-1}\phi(x)+(m-1)(m+1)x^m+(m+1)N,\\
g(x)\phi''(x) &= -2m(m+1)x^{m-1}\phi'(x)-m(m-1)(m+1)x^{m-2}\phi(x)\\
&\quad +m(m-1)(m+1)x^{m-1},\\
g(x)\phi'''(x) &= -3m(m+1)x^{m-1}\phi''(x)-3m(m-1)(m+1)x^{m-2}\phi'(x)\\
&\quad -m(m-2)(m-1)(m+1)x^{m-3}\phi(x)+m(m-1)^2(m+1)x^{m-2}.
\end{aligned}$$

Kako je $a^m = N$, iz (2.1.6) i prethodnih jednakosti sleduje

$$\phi(a)=a,\quad \phi'(a)=0,\quad \phi''(a)=0,\quad \phi'''(a)=\frac{m^2-1}{2a^2}. \tag{2.1.7}$$

Ako je $|m| \neq 1$, na osnovu (2.1.7) zaključujemo da je dati iterativni proces trećeg reda. U ovom slučaju (2.1.4) se svodi na

$$\lim_{k\to+\infty}\frac{x_{k+1}-a}{(x_k-a)^3}=\frac{\phi'''(a)}{3!}=\frac{m^2-1}{12a^2}.$$

Primetimo da se za $m=\pm 1$ iterativna funkcija ϕ svodi na konstantu $\phi(x)=N$ ili $\phi(x)=1/N$. $\triangle$

3.2.2 Aitkenov Δ^2 metod

U ovom odeljku razmatraćemo problem ubrzavanja konvergencije realnog niza $\{x_k\}_{k\in\mathbb{N}_0}$, koji se generiše pomoću

$$x_{k+1}=\phi(x_k),\quad k=0,1,\ldots, \tag{2.2.1}$$

pri čemu je $\lim\limits_{k\to+\infty} x_k = a$.

Pretpostavimo da je iterativni proces (2.2.1) sa linearnom konvergencijom, tj. da je

$$x_{k+1}-a=C_k(x_k-a),\quad k=0,1,\ldots, \tag{2.2.2}$$

gde su $C_k = C+\delta_k$ $(k=0,1,\ldots)$, C konstanta takva da je $|C|<1$ i $\delta_k\to 0$, kada $k\to+\infty$.

Iz asimptotskih relacija

$$x_{k+2}-a\sim C(x_{k+1}-a)\quad \text{i}\quad x_{k+1}-a\sim C(x_k-a)\qquad (k\to+\infty),$$

koje se dobijaju na osnovu (2.2.2), eliminacijom konstante C, sleduje

$$a \sim \frac{x_{k+2}x_k - x_{k+1}^2}{x_{k+2} - 2x_{k+1} + x_k} \qquad (k \to +\infty).$$

Dobijena asimptotska relacija sugeriše konstrukciju niza $\{x_k^*\}_{k\in\mathbb{N}_0}$ pomoću

$$x_k^* = \frac{x_{k+2}x_k - x_{k+1}^2}{x_{k+2} - 2x_{k+1} + x_k}, \quad k = 0, 1, \ldots. \tag{2.2.3}$$

Formula (2.2.3), međutim, nije pogodna za primenu sa numeričkog stanovišta, pa se zato koriste njeni ekvivalentni matematički oblici

$$x_k^* = x_k - \frac{(x_{k+1} - x_k)^2}{x_{k+2} - 2x_{k+1} + x_k} \tag{2.2.4}$$

$$x_k^* = x_{k+1} - \frac{(x_{k+1} - x_k)(x_{k+2} - x_{k+1})}{x_{k+2} - 2x_{k+1} + x_k} \tag{2.2.5}$$

$$x_k^* = x_{k+2} - \frac{(x_{k+2} - x_{k+1})^2}{x_{k+2} - 2x_{k+1} + x_k}, \tag{2.2.6}$$

gde je $k = 0, 1, \ldots$.

Poslednja formula se najčešće koristi u primenama, s obzirom na izvesne prednosti nad formulama (2.2.4) i (2.2.5). Naime, glavni deo u vrednosti x_k^*, po formulama (2.2.4), (2.2.5) i (2.2.6) je x_k, x_{k+1} i x_{k+2}, respektivno. S druge strane, kako je glavni deo u formuli (2.2.6) najpribližniji vrednosti a (u poređenju sa x_k i x_{k+1}), to izlazi da ova formula, sa numeričkog stanovišta, ima prednosti nad formulama (2.2.4) i (2.2.5).

Navedene formule mogu se predstaviti pomoću operatora Δ. Na primer, formula (2.2.4) dobija oblik

$$x_k^* = x_k - \frac{(\Delta x_k)^2}{\Delta^2 x_k}, \quad k = 0, 1, \ldots. \tag{2.2.7}$$

Sledeća teorema ukazuje na to da niz $\{x_k^*\}_{k\in\mathbb{N}_0}$ brže konvergira ka a, nego niz $\{x_k\}_{k\in\mathbb{N}_0}$.

Teorema 2.2.1. *Neka je za niz* $\{x_k\}_{k\in\mathbb{N}_0}$, $x_k \neq a$ *i*

$$x_{k+1} - a = (C + \delta_k)(x_k - a) \qquad \left(|C| < 1, \lim_{k\to+\infty} \delta_k = 0\right).$$

Tada je

$$\lim_{k\to+\infty}\frac{x_k^*-a}{x_k-a}=0, \tag{2.2.8}$$

gde je $a=\lim_{k\to+\infty}x_k$.

Dokaz. Neka je $e_k=x_k-a$. Tada je $e_{k+1}=(C+\delta_k)e_k$ i

$$\Delta x_k=\Delta e_k=e_k[(C-1)+\delta_k],$$

$$\Delta^2 x_k=\Delta^2 e_k=e_k[(C+\delta_{k+1})(C+\delta_k)-2(C+\delta_k)+1]=e_k[(C-1)^2+\gamma_k],$$

gde je $\gamma_k=C(\delta_{k+1}+\delta_k)-2\delta_k+\delta_k\delta_{k+1}$. Primetimo da uslov $\lim_{k\to+\infty}\delta_k=0$ implicira uslov $\lim_{k\to+\infty}\gamma_k=0$.

Da bismo dokazali (2.2.8) pođimo od (2.2.7).

Kako je $e_k\neq 0$, $C\neq 1$ i $\gamma_k\to 0$ $(k\to+\infty)$, imamo $\Delta^2 x_k=0$, što znači da je niz $\{x_k^*\}_{k\in\mathbb{N}_0}$ dobro definisan pomoću (2.2.7) za svako k. Tada je, na osnovu prethodnog,

$$x_k^*-a=e_k-\frac{((C-1)+\delta_k)^2}{(C-1)^2+\gamma_k}e_k,$$

tj.

$$\lim_{k\to+\infty}\frac{x_k^*-a}{x_k-a}=\lim_{k\to+\infty}\frac{\gamma_k-2(C-1)\delta_k-\delta_k^2}{(C-1)^2+\gamma_k}=0. \quad \square$$

Nakon rada AITKENa[116] [1] iz 1926. godine, navedeni postupak ubrzavanja konvergencije niza $\{x_k\}_{k\in\mathbb{N}_0}$ poznat je kao AITKEN*ov* Δ^2 *metod*. Inače, metod je otkrio i koristio T. SEKI[117] još pre 1680. godine. Za detalje videti pregledni rad [12].

U daljem razmatranju ovog metoda uvešćemo dodatnu pretpostavku za iterativni proces (2.2.1). Naime, neka je

$$\delta_k=\omega_k e_k,$$

pri čemu $\omega_k\to\omega$, kada $k\to+\infty$. Tada (2.2.2) postaje

$$e_{k+1}=Ce_k+\omega_k e_k^2,$$

gde smo stavili $e_k=x_k-a$. Jasno je, međutim, da svi iterativni procesi sa linearnom konvergencijom ne poseduju ovu osobinu.

[116] ALEXANDER CRAIG AITKEN (1895 – 1967), poznati matematičar sa Novog Zelanda.
[117] TAKAKAZU SEKI (? – 1708), japanski matematičar.

Teorema 2.2.2. *Neka su ispunjeni uslovi teoreme* 2.2.1 *sa* $\delta_k = \omega_k e_k$ ($\omega_k \to \omega$, *kada* $k \to +\infty$). *Tada je*

$$\lim_{k\to+\infty} \frac{x_k^* - a}{(x_k - a)^2} = \frac{C\omega}{C-1}. \tag{2.2.9}$$

Dokaz. S obzirom na to da je

$$\gamma_k - 2(C-1)\delta_k - \delta_k^2 = (C+\delta_k)(\delta_{k+1} - \delta_k) = e_k(C+\delta_k)[\omega_{k+1}(C+\delta_k) - \omega_k],$$

imamo $x_k^* - a = \alpha_k e_k^2$, gde smo stavili

$$\alpha_k = \frac{(C+\delta_k)[\omega_{k+1}(C+\delta_k) - \omega_k]}{(C-1)^2 + \gamma_k}.$$

Kako je, dalje, $\lim_{k\to+\infty} \alpha_k = \dfrac{C\omega}{C-1}$, iz prethodnog neposredno sleduje (2.2.9). □

Teorema 2.2.3. *Ako su ispunjeni uslovi prethodne teoreme važi*

$$\lim_{k\to+\infty} \frac{x_{k+1}^* - a}{(x_k^* - a)^2} = C^2. \tag{2.2.10}$$

Dokaz. Na osnovu dokaza prethodne teoreme i jednakosti (2.2.2), imamo

$$\frac{x_{k+1}^* - a}{(x_k^* - a)^2} = \frac{\alpha_{k+1} e_{k+1}^2}{\alpha_k e_k^2} = \frac{\alpha_{k+1}}{\alpha_k}(C+\delta_k)^2,$$

odakle sleduje (2.2.10). □

Napomena 2.2.1. Iz teoreme 2.2.3 sleduje da je i transformisani niz $\{x_k^*\}_{k\in\mathbb{N}_0}$ sa linearnom konvergencijom, ali je njegov faktor konvergencije manji nego kod niza $\{x_k\}_{k\in\mathbb{N}_0}$ ($C^2 < |C| < 1$).

Napomena 2.2.2. Posmatrajmo iterativni proces $x_{k+1} = \phi(x_k)$, $k = 0, 1, \ldots$, gde $x_0 \in [\alpha, \beta]$. Ako za iterativnu funkciju $\phi : [\alpha, \beta] \to [\alpha, \beta]$ pretpostavimo uslove:

1° $\phi \in C^2[\alpha, \beta]$,

2° $|\phi'(x)| < 1$ za svako $x \in [\alpha, \beta]$,

na osnovu TAYLORove formule, imamo

$$x_{k+1} = \phi(x_k) = a + \phi'(a)e_k + \frac{1}{2}\phi''(\xi_k)e_k^2,$$

gde su $a = \lim_{k\to+\infty} x_k$, $e_k = x_k - a$, $\xi_k = a + \theta_k(x_k - a)$ $(0 \le \theta \le 1)$.

Kako iterativni proces, definisan na ovaj način, zadovoljava uslove teoreme 2.2.2 sa

$$C = \phi'(a) \quad \text{i} \quad \omega_k = \frac{1}{2}\phi''(\xi_k),$$

to na osnovu (2.2.9) imamo

$$\lim_{k\to+\infty} \frac{x_k^* - a}{(x_k - a)^2} = \frac{1}{2}\frac{\phi''(a)\phi'(a)}{\phi'(a) - 1}.$$

Primer 2.2.1. Primenom formule (2.2.6) na niz koji se generiše pomoću procesa $x_{k+1} = \cos x_k$, $k = 0, 1, \ldots$, sa $x_0 = 1$, dobijamo

k	x_k	x_k^*
0	1.000000	0.728010
1	0.540302	0.733665
2	0.857553	0.736906
3	0.654290	0.738050
4	0.793480	0.738636
5	0.701369	0.738876
6	0.763960	0.738992
7	0.722102	0.739042
8	0.750418	0.739066
9	0.731404	
10	0.744237	

Primetimo da x_7^* aproksimira koren jednačine $\cos x - x = 0$ sa četiri tačne decimale. $\triangle$

3.2.3 O metodima bliskim AITKENovom metodu

Posmatrajmo iterativni proces

$$x_{k+1} = \phi(x_k), \quad k = 0, 1, \ldots. \tag{2.3.1}$$

Kako su $x_{k+1} = \phi(x_k)$ i $x_{k+2} = \phi(x_{k+1}) = \phi(\phi(x_k)) = \phi^2(x_k)$, formula (2.2.3) iz prethodnog odeljka može se predstaviti u obliku

$$x_k^* = \psi(x_k), \quad k = 0, 1, \ldots, \tag{2.3.2}$$

gde je

$$\psi(x) = \frac{x\phi^2(x) - \phi^2(x)}{\phi^2(x) - 2\phi(x) + x}. \tag{2.3.3}$$

Funkciju ψ, definisanu sa (2.3.3), STEFFENSEN[118] (videti na primer [13, str. 241–246]) je dobio primenom metoda sečice (videti odeljak 5.1.4) na rešavanje jednačine

$$F(x) \equiv x - \phi(x) = 0,$$

ne uvodeći pretpostavku o redu konvergencije procesa (2.3.1).

Za razliku od AITKENovog Δ^2 metoda, definisanog sa (2.3.2), STEFFENSEN je došao do procesa

$$x_{k+1} = \psi(x_k), \quad k = 0, 1, \ldots .$$

Na osnovu dva iterativna procesa $(\nu = 1, 2)$ istog reda r,

$$x_{k+1}^{(\nu)} = \phi_\nu(x_k^{(\nu)}), \quad k = 0, 1, \ldots,$$

koji konvergiraju ka istoj tački $x = a$, HAUSEHOLDER[119] [4] je dobio opštiji ubrzani metod

$$x_{k+1} = \widetilde{\psi}(x_k), \quad k = 0, 1, \ldots, \tag{2.3.4}$$

gde je nova iterativna funkcija $\widetilde{\psi}$ definisana pomoću

$$\widetilde{\psi}(x) = \frac{x\phi_1(\phi_2(x)) - \phi_1(x)\phi_2(x)}{x - \phi_1(x) - \phi_2(x) + \phi_1(\phi_2(x))}, \tag{2.3.5}$$

i pri tome dokazao sledeći rezultat.

Teorema 2.3.1. *Ako je*

$$(\phi_1'(a) - 1)(\phi_2'(a) - 1) \neq 0$$

iterativni proces (2.3.4) *ima red konvergencije* $2r - 1$ *ako je* $r > 1$ *ili dva ako je* $r = 1$.

Napomena 2.3.1. Za $\phi_1 = \phi_2 = \phi$, (2.3.5) se svodi na (2.3.3).

[118] JOHAN FREDERIK STEFFENSEN (1873 – 1961), danski matematičar.

[119] ALSTON SCOTT HOUSEHOLDER (1904 – 1993), američki matematičar poznat u numeričkoj analizi i matematičkoj biologiji.

Pod pretpostavkom da za $\nu = 1,\dots,m$ dato m iterativnih procesa,

$$x_{k+1}^{(\nu)} = \phi_\nu(x_k^{(\nu)}), \quad k = 0,1,\dots,$$

čiji je red konvergencije r, u radu [14] dato je sledeće uopštenje prethodnih rezultata.

Teorema 2.3.2. *Ako su* $\phi_{p,q,\dots,s,t}(x) = \phi_p(\phi_q(\cdots\phi_s(\phi_t(x))\cdots))$,

$$D_n(x) = \begin{vmatrix} x & \phi_{i_1,i_2,\dots,i_n}(x) \\ \phi_j(x) & \phi_{j_1,j_2,\dots,j_{n+1}}(x) \end{vmatrix}, \quad \Delta_n(x) = \begin{vmatrix} x-\phi_{i_1,i_2,\dots,i_n}(x) & 1 \\ \phi_j(x)-\phi_{j_1,j_2,\dots,j_{n+1}}(x) & 1 \end{vmatrix},$$

$(j,i_\alpha,j_\beta \in \{1,2,\dots,m\},\ \alpha = 1,2,\dots,n,\ \beta = 1,2,\dots,n+1)$, *i*

$$1-\phi_j'(a)-\phi_{i_1}'(a)\phi_{i_2}'(a)\cdots\phi_{i_n}'(a)+\phi_{j_1}'(a)\phi_{j_2}'(a)\cdots\phi_{j_{n+1}}'(a) \neq 0,$$

iterativni proces definisan pomoću

$$x_{k+1} = \frac{D_n(x_k)}{\Delta_n(x_k)}, \quad k = 0,1,\dots,$$

ima red konvergencije $r^n + r - 1$, *ako je* $r > 1$, *ili dva, ako je* $r = 1$.

3.2.4 Metodi za ubrzavanje konvergencije iterativnih procesa

U ovom odeljku daćemo jedan opšti prilaz ubrzavanju konvergencije iterativnih procesa. Naime, metodi koji će biti izloženi omogućavaju dobijanje iterativnih procesa višeg reda, ako se pođe od nekog poznatog iterativnog procesa.

Neka se niz $\{x_k\}_{k\in\mathbb{N}_0}$ koji konvergira ka tački a generiše pomoću

$$x_{k+1} = \phi(x_k), \quad k = 0,1,\dots\ . \tag{2.4.1}$$

U radu [5] JOVANOVIĆ[120] je dokazao sledeću teoremu.

Teorema 2.4.1. *Neka je iterativni proces* (2.4.1) *sa redom konvergencije r, definisan funkcijom* ϕ *koja je* $(r+1)$*-puta diferencijabilna u okolini granične tačke a i neka je* $\phi'(a) \neq r$. *Tada je*

[120] BOŠKO S. JOVANOVIĆ (1946 –), srpski matematičar, redovni profesor Matematičkog fakulteta Univerziteta u Beogradu. Bavi se numeričkim metodima za parcijalne diferencijalne jednačine.

$$x_{k+1} = \phi(x_k) + \frac{1}{r}\phi'(x_k)(x_{k+1} - x_k),$$

tj.

$$x_{k+1} = \frac{\phi(x_k) - \frac{1}{r}\phi'(x_k)x_k}{1 - \frac{1}{r}\phi'(x_k)} = x_k - \frac{x_k - \phi(x_k)}{1 - \frac{1}{r}\phi'(x_k)}$$

iterativni proces najmanje reda $r+1$.

U daljem tekstu, posmatraćemo iterativne procese u BANACHovom prostoru. Neka je X BANACHov prostor, operator $T : X \to X$ i neka niz koji se generiše pomoću

$$u_{k+1} = Tu_k, \quad k = 0, 1, \ldots, \tag{2.4.2}$$

konvergira ka $a \in X$. Jednakost (2.4.2) definiše iterativni proces za rešavanje operatorske jednačine $u = Tu$. Sa $U(a)$ označićemo konveksnu okolinu granične tačke a.

U pomenutom radu [5] navedena je sledeća teorema, koja predstavlja generalizaciju teoreme 2.4.1.

Teorema 2.4.2. *Neka je iterativni proces* (2.4.2) *reda r, sa operatorom* $T : X \to X$ *koji je* $(r+1)$*-puta* FRÉCHET*–diferencijabilan u okolini* $U(a)$ *i neka postoji inverzan operator* $\left[I - \frac{1}{r}T'_{(u)}\right]^{-1}$ *kada* $u \in U(a)$. *Tada je*

$$u_{k+1} = Gu_k = u_k - \left[I - \frac{1}{r}T'_{(u)}\right]^{-1}(u_k - Tu_k) \tag{2.4.3}$$

iterativni proces najmanje reda $r+1$.

Osnovni nedostatak metoda, definisanog poslednjom teoremom, je što zahteva nalaženje inverznog operatora $\left[I - \frac{1}{r}T'_{(u)}\right]^{-1}$, što je u većini slučajeva vrlo složeno.

Sada ćemo izložiti jedan metod za ubrzavanje konvergencije iterativnih procesa, koji ne zahteva nalaženje inverznog operatora (videti: MILOVANOVIĆ [7]). Nažalost, ovaj metod se ne može primeniti za ubrzavanje procesa sa redom konvergencije jedan.

Teorema 2.4.3. *Neka je iterativni proces* (2.4.2) *sa konvergencijom reda* $r(\geq 2)$ *i neka je operator* $T : X \to X$ FRÉCHET*–diferencijabilan* $(r+1)$*-puta u okolini* $U(a)$. *Tada je*

$$u_{k+1} = Fu_k = Tu_k - \frac{1}{r}T'_{(u_k)}(u_k - Tu_k), \quad k = 0,1,\ldots, \tag{2.4.4}$$

iterativni proces najmanje reda $r+1$.

Dokaz. Neka $u \in U(a)$. S obzirom da je iterativni proces (2.4.2) sa konvergencijom reda r i operator T FRÉCHET–diferencijabilan $(r+1)$-puta u $U(a)$, to su ispunjeni uslovi teoreme 2.1.1.

Na osnovu TAYLORove formule imamo

$$T'_{(u)} = \frac{1}{(r-1)!}T^{(r)}_{(a)}(\underbrace{u-a,\ldots,u-a}_{r-1 \text{ puta}}) + W(a,u-a), \tag{2.4.5}$$

$$Tu = a + \frac{1}{r!}T^{(r)}_{(a)}(\underbrace{u-a,\ldots,u-a}_{r \text{ puta}}) + w(a,u-a), \tag{2.4.6}$$

$$\|W(a,u-a)\| = O(\|u-a\|^r) \quad \text{i} \quad \|w(a,u-a)\| = O(\|u-a\|^{r+1}).$$

S obzirom na to da je $T'_{(u)}$ linearan operator, to se (2.4.4) može predstaviti u obliku

$$Fu - a = Tu - a - \frac{1}{r!}T'_{(u)}(u-a) + \frac{1}{r!}T'_{(u)}(Tu-a).$$

Korišćenjem (2.4.5) i (2.4.6) dobijamo

$$Fu - a = w(a,u-a) - \frac{1}{r}W(a,u-a)(u-a) + \frac{1}{r}T'_{(u)}(Tu-a),$$

odakle sleduje

$$\|Fu-a\| \leq \|w(a,u-a)\| + \frac{1}{r}\|W(a,u-a)\| \cdot \|u-a\| + \frac{1}{r}\|T'_{(u)}\| \cdot \|Tu-a\|.$$

Kako je $\|T'_{(u)}\| = O(\|u-a\|^{r-1})$ i $\|Tu-a\| = O(\|u-a\|^r)$ pri $r \geq 2$, najzad dobijamo

$$\|Fu-a\| = O(\|u-a\|^{r+1}). \quad \square$$

Sada ćemo dati dokaz teoreme 2.4.2.

Dokaz teoreme 2.4.2. Neka $u \in U(a)$. Formuli (2.4.3) može se dati sledeći oblik

$$\left[I - \frac{1}{r}T'_{(u)}\right](u - Gu) = u - Tu,$$

tj.

$$Gu = Tu - \frac{1}{r}T'_{(u)}(u - Gu).$$

S obzirom na činjenicu da je $T'_{(u)}$ linearan operator, (2.4.3) se može predstaviti u ekvivalentnom obliku

$$Gu - a = Tu - a - \frac{1}{r}T'_{(u)}(u - a) + \frac{1}{r}T'_{(u)}(Gu - a)$$

odakle, korišćenjem (2.4.5) i (2.4.6), dobijamo

$$Gu - a = w(a, u - a) - \frac{1}{r}W(a, u - a)(u - a) + \frac{1}{r}T'_{(u)}(Gu - a).$$

Iz poslednje jednakosti sleduje

$$\left(1 - \frac{1}{r}\|T'_{(u)}\|\right)\|Gu - a\| \leq \|w(a, u - a)\| + \frac{1}{r}\|W(a, u - a)\| \cdot \|u - a\|,$$

tj.

$$\|Gu - a\| = O(\|u - a\|^{r+1}). \quad \square$$

Teorema 2.4.4. *Neka je iterativni proces* (2.4.1) *sa konvergencijom reda* $r\,(\geq 2)$ *i funkcija* ϕ *diferencijabilna* $(r+1)$*-puta u okolini granične tačke a. Tada je*

$$x_{k+1} = \phi(x_k) - \frac{1}{r}\phi'(x_k)(x_k - \phi(x_k)), \quad k = 0, 1, \ldots,$$

iterativni proces najmanje reda $r + 1$.

Ova teorema je očigledno posledica teoreme 2.4.3.

3.2.5 *R*-red konvergencije iterativnih procesa

U ovom odeljku ukazaćemo na jedan poseban tretman konvergencije iterativnih procesa u prostoru $X = \mathbb{R}^n$, koji su uveli ORTEGA[121] i RHEINBOLDT[122] [11].

Definicija 2.5.1. Neka je $\{\boldsymbol{x}^{(k)}\}$ proizvoljan niz u $\mathbb{R}^n$ koji konvergira ka $\boldsymbol{a}$. Tada se brojevi

$$R_p\{\boldsymbol{x}^{(k)}\} = \begin{cases} \limsup\limits_{k\to+\infty} \|\boldsymbol{x}^{(k)} - \boldsymbol{a}\|^{1/k}, & \text{ako je } p = 1, \\ \limsup\limits_{k\to+\infty} \|\boldsymbol{x}^{(k)} - \boldsymbol{a}\|^{1/p^k}, & \text{ako je } p > 1, \end{cases}$$

nazivaju *faktori konvergencije po korenu*, ili kraće *R–faktori* niza $\{\boldsymbol{x}^{(k)}\}$. Ako je IP iterativni proces sa graničnom tačkom $\boldsymbol{a}$ i $C(IP,\boldsymbol{a})$ skup svih nizova generisanih pomoću IP, koji konvergiraju ka $\boldsymbol{a}$, tada se veličina

$$R_p(IP,\boldsymbol{a}) = \sup\big\{R_p\{\boldsymbol{x}^{(k)}\} \mid \{\boldsymbol{x}^{(k)}\} \in C(IP,\boldsymbol{a})\big\} \qquad (1 \le p < +\infty)$$

naziva *R–faktorom iterativnog procesa* u tački $\boldsymbol{a}$.

Napomena 2.5.1. Ako niz $\{\boldsymbol{x}^{(k)}\}$ konvergira ka $\boldsymbol{a}$, tada postoji k_0 takvo da je

$$0 \le \|\boldsymbol{x}^{(k)} - \boldsymbol{a}\| \le 1 \quad \text{za svako } k \ge k_0,$$

odakle zaključujemo da je $0 \le R_p\{\boldsymbol{x}^{(k)}\} \le 1$ za svako $p \ge 1$.

Sada navodimo bez dokaza dve teoreme.

Teorema 2.5.1. *Neka je $\{\boldsymbol{x}^{(k)}\}$ proizvoljan niz u $\mathbb{R}^n$ koji konvergira ka $\boldsymbol{a}$. Faktor $R_p\{\boldsymbol{x}^{(k)}\}$ ne zavisi od izbora norme u $\mathbb{R}^n$ ni za jedno $p \in [1,+\infty)$. Takođe, R–faktor $R_p(IP,\boldsymbol{a})$ iterativnog procesa je nezavisan od izbora norme.*

Teorema 2.5.2. *Neka je IP iterativni proces sa graničnom tačkom $\boldsymbol{a}$. Tada važi jedan od sledećih uslova:*

a) $R_p(IP,\boldsymbol{a}) = 0$ *za svako* $p \in [1,+\infty)$*;*

b) $R_p(IP,\boldsymbol{a}) = 1$ *za svako* $p \in [1,+\infty)$*;*

c) *postoji* $p_0 \in [1,+\infty)$ *takvo da je* $R_p(IP,\boldsymbol{a}) = 0$ *za svako* $p \in [1,p_0)$ *i* $R_p(IP,\boldsymbol{a}) = 1$ *za svako* $p \in [p_0,+\infty)$.

[121] JAMES M. ORTEGA (1932 – 2008), američki matematičar.

[122] WERNER C. RHEINBOLDT, američki matematičar nemačkog porekla.

Na osnovu prethodnog može se uvesti tzv. *R–red konvergencije* iterativnih procesa.

Definicija 2.5.2. *R-red konvergencije iterativnog procesa* u tački $\boldsymbol{a}$ je veličina

$$O_R(IP,\boldsymbol{a}) = \begin{cases} +\infty & (\text{ako je } R_p(IP,\boldsymbol{a}) = 0 \text{ za svako } p \in [1,+\infty)), \\ \inf\{p \in [1,+\infty) \mid R_p(IP,\boldsymbol{a}) = 1\} & (\text{u ostalim slučajevima}). \end{cases}$$

Primetimo da ovako definisani *R*–red konvergencije iterativnog procesa ne zavisi od izbora norme u $\mathbb{R}^n$. Takođe, primetimo sledeće činjenice:

– ako je $R_p(IP,\boldsymbol{a}) < 1$ za neko $p \in [1,+\infty)$, tada je *R*–red ne manji od p, tj. važi $O_R(IP,\boldsymbol{a}) \geq p$;

– ako je $R_q(IP,\boldsymbol{a}) > 0$ za neko $q \in [1,+\infty)$, tada za *R*–red važi $O_R(IP,\boldsymbol{a}) \leq q$;

– ako je $0 < R_p(IP,\boldsymbol{a}) < 1$ za neko $p \in [1,+\infty)$, tada je $O_R(IP,\boldsymbol{a}) = p$.

U slučaju kada je $0 < R_1(IP,\boldsymbol{a}) < 1$ kažemo da je konvergencija procesa u tački $\boldsymbol{a}$ *R–linearna*. Ako je, međutim, $R_1(IP,\boldsymbol{a}) = 0$ ili $R_1(IP,\boldsymbol{a}) = 1$, za konvergenciju kažemo da je *R–podlinearna*, odnosno *R–nadlinearna*.

Kada upoređujemo konvergenciju dva iterativna procesa IP_1 i IP_2 postupamo na sledeći način: najpre uporedimo *R*–red, tj. veličine $O_R(IP_1,\boldsymbol{a})$ i $O_R(IP_2,\boldsymbol{a})$, pri čemu je brži onaj proces koji ima veći *R*–red. Međutim, ako je $O_R(IP_1,\boldsymbol{a}) = O_R(IP_2,\boldsymbol{a}) = \overline{p}$, tada upoređujemo *R*–faktore za $p = \overline{p}$. Ako je pri tome, na primer, $R_{\overline{p}}(IP_1,\boldsymbol{a}) < R_{\overline{p}}(IP_2,\boldsymbol{a})$, tada je IP_2 brži od IP_1.

Literatura

1. A.C. AITKEN, *On Bernoulli's numerical solution of algebraic equations*, Proc. R. Soc. Edinb. Ser. A **46** (1926), 289–305.
2. L. COLLATZ, *Functionalanalysis und Numerische Mathematik*, Springer Verlag, Berlin – Heidelberg – New York, 168.
3. O. HADŽIĆ, *Fixed Point Theory in Topological Vector Spaces,* Univ. Movi Sad, Faculty of Science Institute of Mathematics, Novi Sad, 1984.
4. A. S. HAUSEHOLDER, *Principles of Numerical Analysis,* Dover, New York, 1953.
5. B. JOVANOVIĆ, *A method for obtaining iterative formulae of higher order*, Mat. Vesnik **9** (24) (1972), 365–369.
6. L. V. KANTOROVICH, G. P. AKILOV, *Functional Analysis*, Pergamon Press, Oxford-Elmsford, N.Y., 1982 (originalno izdanje na ruskom: Nauka, Moskva, 1977).
7. G.V. MILOVANOVIĆ, *A method to accelerate iterative processes in Banach space*, Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fiz. No 661 – No 497 (1974), 67–71.
8. G.V. MILOVANOVIĆ, *Numerička analiza, I deo,* Naučna knjiga, Beograd, 1985.
9. G.V. MILOVANOVIĆ, R.Ž. ĐORĐEVIĆ, *Linearna algebra,* Elektronski fakultet u Nišu, Niš, 2004.
10. G.V. MILOVANOVIĆ, R.Ž. ĐORĐEVIĆ, *Matematička analiza I,* Elektronski fakultet u Nišu, Niš, 2005.
11. J. M. ORTEGA, W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York – London, 1970.
12. N. OSADA, *The early history of convergence acceleration methods,* Numer. Algorithms **60** (2012), 205–221.
13. A. OSTROWSKI, *Solution of Equations and Systems of Equations*, Academic Press, New York, 1966.
14. A. V. PROKOPČENKO, *Iteration processes of higher orders*, Zh. Vychisl. Mat. i Mat. Fiz. **14** (1974), 230–233 (na ruskom).
15. J. STOER, *Einführung in die Numerische Mathematik I,* Springer Verlag, Berlin – Heidelberg – New York, 1972.
16. D. TOŠIĆ, *Uvod u numeričku analizu,* Naučna knjiga, Beograd, 1982.
17. J. F. TRAUB, *Iterative Methods for the Solution of Equations,* Prentice-Hall, New Jersey, 1964.

4. NUMERIČKI METODI U LINEARNOJ ALGEBRI

4.1 DIREKTNI METODI

4.1.1 Uvodne napomene

Numerički problemi u linearnoj algebri mogu se klasifikovati u nekoliko grupa:

1° rešavanje sistema linearnih algebarskih jednačina

$$A\boldsymbol{x} = \boldsymbol{b}$$

sa regularnom matricom A, izračunavanje determinante od A i inverzija matrice A;

2° rešavanje proizvoljnog sistema linearnih jednačina metodom najmanjih kvadrata;

3° određivanje sopstvenih vrednosti i sopstvenih vektora date kvadratne matrice;

4° rešavanje zadatka linearnog programiranja.

Za rešavanje ovih problema razvijen je čitav niz metoda, koji se mogu podeliti na direktne i iterativne. Ovo poglavlje posvećujemo direktnim metodima za rešavanje problema iz tačke 1°.

Osnovna karakteristika direktnih metoda, ili tačnih metoda, kako se ponekad nazivaju, je ta da se posle konačnog broja transformacija (koraka) dolazi do rezultata. Ukoliko bi se sve računske operacije izvodile tačno, dobijeni rezultat bi bio apsolutno tačan. Naravno, kako se proces računanja izvodi sa zaokrugljivanjem međurezultata, krajnji rezultat je ograničene tačnosti. Zbog toga je od velikog značaja analiza grešaka kod ovih metoda. Jasno je da će uticaj grešaka biti veći ukoliko je broj operacija kod primenjenog metoda veći.

Posmatrajmo sistem linearnih algebarskih jednačina

$$A\boldsymbol{x} = \boldsymbol{b},$$

gde su

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & & a_{nn} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \tag{1.1.1}$$

Pretpostavimo da sistem ima jedinstveno rešenje, tj. da je $\det A \neq 0$. Tada se rešenja mogu izraziti pomoću CRAMEROvih[123] formula

$$x_i = \frac{\det A_i}{\det A} \quad (i = 1, \ldots, n),$$

gde je A_i matrica dobijena iz matrice A zamenom i-te kolone vektorom $\boldsymbol{b}$. Međutim, ove formule su nepogodne za praktična izračunavanja, s obzirom da je za izračunavanje $n+1$ determinanata potreban veliki broj aritmetičkih operacija. Naime, ako bismo vrednost determinante n-tog reda izračunavali po definiciji

$$\det A = \sum (-1)^i a_{1i_1} a_{2i_2} \ldots a_{ni_n},$$

gde se sumiranje obavlja preko svih permutacija $(i_1, i_2, \ldots, i_n)$ osnovnog skupa $(1, 2, \ldots, n)$ (i je broj inverzija u permutaciji $(i_1, i_2, \ldots, i_n)$), potrebno je izvršiti $S_n = n! - 1$ sabiranja i $M_n = (n-1)n!$ množenja, što ukupno iznosi

$$P_n = S_n + M_n \cong n \cdot n! \text{ operacija.}$$

Pod pretpostavkom da je za obavljanje jedne aritmetičke operacije potrebno $10\,ns$ ($100\,MFLOPS$[124]), to bi za izračunavanje vrednosti jedne determinante tridesetog reda ($n = 30$) bilo potrebno oko

$$\frac{30 \cdot 30! \cdot 10 \cdot 10^{-9}}{3600 \cdot 24 \cdot 365} \cong 2.5 \cdot 10^{18} \text{ godina.}$$

Uopšteno govoreći ovakav postupak je praktično neprimenljiv već za determinante reda $n = 13$, za koje je potrebno vreme oko 13.5 minuta.

[123] GABRIEL CRAMER (1704 – 1752), švajcarski matematičar.

[124] 100 miliona operacija u sekundi u aritmetici sa pokretnom tačkom. *FLOPS* je skraćenica od engleskog izraza *FLoating point Operations Per Second*.

4.1.2 GAUSSov metod eliminacije i GAUSS–JORDANov metod

Neka je dat sistem linearnih jednačina

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n, \end{aligned} \tag{1.2.1}$$

koji ima jedinstveno rešenje. Za rešavanje ovog sistema, tj.

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{1.2.2}$$

gde su A, $\boldsymbol{x}$, $\boldsymbol{b}$ dati pomoću (1.1.1), postoji veliki broj direktnih metoda. Jedan od najpogodnijih je svakako GAUSSov[125] metod eliminacije, koji ima više varijanata. U suštini GAUSSov metod se zasniva na redukciji sistema (1.2.2), primenom elementarnih transformacija, na trougaoni sistem jednačina

$$R\boldsymbol{x} = \boldsymbol{c}, \tag{1.2.3}$$

gde su

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix} \quad \text{i} \quad \boldsymbol{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

Sistem (1.2.3) se rešava sukcesivno polazeći od poslednje jednačine. Naime,

$$\begin{aligned} x_n &= \frac{c_n}{r_{nn}}, \\ x_i &= \frac{1}{r_{ii}}\Big(c_i - \sum_{k=i+1}^{n} r_{ik}x_k\Big), \quad i = n-1, \ldots, 1. \end{aligned}$$

Napomenimo da su koeficijenti $r_{ii} \neq 0$, jer po pretpostavci sistem (1.2.2), tj. (1.2.3), ima jedinstveno rešenje.

[125] JOHANN CARL FRIEDRICH GAUSS (1777–1855), veliki nemački matematičar, sa značajnim doprinosima u matematici (teorija brojeva, algebra, analiza, statistika, diferencijalna geometrija, itd.), ali i u drugim oblastima nauke (geodezija, fizika, astronomija, itd.).

Pokazaćemo sada kako se sistem (1.2.1) može redukovati na ekvivalentan sistem sa trougaonom matricom.

Pod pretpostavkom da je $a_{11} \neq 0$, izračunajmo najpre faktore

$$m_{i1} = \frac{a_{i1}}{a_{11}} \quad (i = 2, \ldots, n),$$

a zatim množenjem prve jednačine u sistemu (1.2.1) sa m_{i1} i oduzimanjem od i-te jednačine, dobijamo sistem od $n-1$ jednačina

$$\begin{aligned} a_{22}^{(2)} x_2 + \cdots + a_{2n}^{(2)} x_n &= b_2^{(2)}, \\ &\vdots \\ a_{n2}^{(2)} x_2 + \cdots + a_{nn}^{(2)} x_n &= b_2^{(n)}, \end{aligned} \tag{1.2.4}$$

gde su

$$a_{ij}^{(2)} = a_{ij} - m_{i1} a_{1j}, \quad b_i^{(2)} = b_i - m_{i1} b_{1j} \quad (i, j = 2, \ldots, n).$$

Pod pretpostavkom da je $a_{22}^{(2)} \neq 0$, primenjujući isti postupak na (1.2.4) sa $m_{i2} = a_{i2}^{(2)} / a_{22}^{(2)}$ $(i = 3, \ldots, n)$ dobijamo sistem od $n-2$ jednačine

$$\begin{aligned} a_{33}^{(3)} x_3 + \cdots + a_{3n}^{(3)} x_n &= b_3^{(3)}, \\ &\vdots \\ a_{n3}^{(3)} x_3 + \cdots + a_{nn}^{(3)} x_n &= b_3^{(n)}, \end{aligned}$$

gde su

$$a_{ij}^{(3)} = a_{ij}^{(2)} - m_{i2} a_{2j}^{(2)}, \quad b_i^{(3)} = b_i^{(2)} - m_{i2} b_{2j}^{(2)} \quad (i, j = 3, \ldots, n).$$

Nastavljajući ovaj postupak, posle $n-1$ koraka dolazimo do jednačine

$$a_{nn}^{(n)} x_{nn} = b_n^{(n)}.$$

Najzad, iz dobijenih sistema, uzimanjem prvih jednačina, dobijamo trougaoni sistema jednačina

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \cdots + a_{1n}^{(1)}x_n &= b_1^{(1)}, \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \cdots + a_{2n}^{(2)}x_n &= b_2^{(2)}, \\ a_{33}^{(3)}x_3 + \cdots + a_{3n}^{(3)}x_n &= b_3^{(3)}, \\ &\vdots \\ a_{nn}^{(n)}x_n &= b_n^{(n)}, \end{aligned}$$

pri čemu smo stavili $a_{ij}^{(1)} = a_{ij},\ b_i^{(1)} = b_i$.

Navedena trougaona redukcija ili kako se često kaže GAUSSova eliminacija, se svodi na izračunavanje koeficijenata

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

za $k = 1,2,\ldots,n-1$. Primetimo da su elementi matrice R i vektora $\boldsymbol{c}$ dati sa

$$r_{ij} = a_{ij}^{(i)}, \quad c_i = b_i^{(i)} \quad (i = 1,\ldots,n;\ j = 1,\ldots,n).$$

Da bi navedena redukcija egzistirala, potrebno je obezbediti uslov $a_{kk}^{(k)} \neq 0$. Elementi $a_{kk}^{(k)}$ su poznati kao glavni elementi ili stožerski elementi[126]. Pod pretpostavkom da je matrica A sistema jednačina (1.2.2) regularna, uslove $a_{kk}^{(k)} \neq 0$ moguće je obezbediti permutacijom jednačina u sistemu (videti [56]).

Primer 1.2.1. Neka je dat sistem jednačina

$$\begin{aligned} 2.304x_1 - 1.213x_2 + 2.441x_3 &= 7.201, \\ 8.752x_1 - 5.608x_2 + 3.916x_3 &= 9.284, \\ 1.527x_1 + 4.333x_2 - 2.214x_3 &= 3.551. \end{aligned}$$

Zaokrugljujući sve međurezultate na četiri značajne cifre, GAUSSovom eliminacijom dobijamo trougaoni sistem jednačina

$$\begin{aligned} 2.304x_1 \quad -1.213x_2 \quad +2.441x_3 &= \quad 7.201, \\ -0.9998x_2 \quad +5.357x_3 &= -18.07, \\ -31.36x_3 &= -94.07, \end{aligned}$$

[126] Na engleskom jeziku *pivotal element*, ili kraće *pivot*.

odakle dobijamo redom $x_3 = 3.000$, $x_2 = 1.999$, $x_1 = 0.9995$. Faktori m_{ij}, u ovom slučaju, su

$$m_{21} = 3.799, \quad m_{31} = 6.628, \quad m_{32} = -5.138.$$

Napomenimo, da je tačno rešenje datog sistema $x_1 = 1$, $x_2 = 2$, $x_3 = 3$. $\triangle$

S obzirom na numerički postupak za trougaonu redukciju i proces zaokrugljivanja međurezultata, javiće se greška u elementima matrice R i vektora $\boldsymbol{c}$. Naime, umesto sistema (1.2.3) dobijamo sistem jednačina $R_0\boldsymbol{x} = \boldsymbol{c}_0$, gde su $R_0 = R + \Delta R$ i $\boldsymbol{c}_0 = \boldsymbol{c} + \Delta\boldsymbol{c}$. Rešenje ovog sistema biće $\boldsymbol{x}_0 = \boldsymbol{x} + \Delta\boldsymbol{x}$, gde je $\boldsymbol{x}$ tačno rešenje sistema (1.2.3). Nije teško ustanoviti da će greška biti veća, što je glavni element $a_{kk}^{(k)}$ manji po modulu od preostalih elemenata matrice. U vezi sa ovim navodimo jedan interesantan primer koji potiče od WILKINSONa[127] ([73, str. 205]).

Primer 1.2.2. GAUSSovim metodom eliminacije rešićemo sistem jednačina

$$\begin{bmatrix} 0.000003 & 0.213472 & 0.332147 \\ 0.215512 & 0.375623 & 0.476625 \\ 0.173257 & 0.663257 & 0.625675 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.235262 \\ 0.127653 \\ 0.285321 \end{bmatrix},$$

uzimajući sve međurezultate na šest značajnih cifara. S obzirom da su faktori u ovom slučaju

$$m_{21} = 71837.3, \quad m_{31} = 57752.3, \quad m_{32} = 0.803905,$$

odgovarajući trougaoni sistem postaje

$$\begin{bmatrix} 0.000003 & 0.213472 & 0.332147 \\ & -15334.9 & -23860.0 \\ & & -0.500000 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.235262 \\ -16900.5 \\ -0.20000 \end{bmatrix},$$

odakle dobijamo $x_3 = 0.400000$, $x_2 = 0.479723$, $x_1 = -1.33333$.

Posmatrajmo sada dati sistem jednačina u kome su prva i druga jednačina permutovane. S obzirom da su sada faktori

[127] JAMES HARDY WILKINSON (1919 – 1986), čuveni engleski matematičar u oblasti numeričke analize i kompjuterskih nauka. U njegovu čast ustanovljena je nagrada za dostignuća u oblasti numeričkog softvera, koja se svake četvte godine (počev od 1991. god.) dodeljuje od strane *Argonne National Laboratory* (SAD), *National Physical Laboratory* (Velika Britanija) i poznate softverske kompanije *Numerical Algorithms Group* (*NAG*).

$$m_{21} = 0.0000139203 \quad \text{i} \quad m_{31} = 0.803932,$$

posle prvog eliminacionog koraka dolazimo do sistema jednačina

$$\begin{bmatrix} 0.215512 & 0.375623 & 0.476625 \\ & 0.213467 & 0.332140 \\ & 0.361282 & 0.242501 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.127653 \\ 0.235260 \\ 0.182697 \end{bmatrix}.$$

Nadalje, kako je element na poziciji $(2,2)$ manji od elementa na poziciji $(3,2)$, u matrici poslednjeg sistema jednačina izvršimo permutaciju druge i treće vrste. Tada, s obzirom na faktor $m_{32} = 0.590860$, posle drugog eliminacionog koraka dobijamo

$$\begin{bmatrix} 0.215512 & 0.375623 & 0.476625 \\ & 0.361282 & 0.242501 \\ & & 0.188856 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.127653 \\ 0.182697 \\ 0.127312 \end{bmatrix},$$

odakle sleduje $x_3 = 0.674122$, $x_2 = 0.0532050$, $x_1 = -0.991291$.

Napomenimo, da je tačno rešenje datog sistema, sa deset značajnih cifara,

$$x_3 = 0.6741214694, \quad x_2 = 0.05320393391, \quad x_1 = -0.9912894252. \quad \triangle$$

Na osnovu poslednjeg primera vidimo da strategija izbora glavnog elementa bitno utiče na tačnost rezultata. Modifikacija GAUSSovog eliminacionog metoda u ovom smislu, naziva se GAUSS*ov metod sa izborom glavnog elementa*. Dakle, prema ovom metodu, za glavni element u k-tom eliminacionom koraku uzimamo element $a_{rk}^{(k)}$, za koji je

$$|a_{rk}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}|,$$

uz permutaciju k-te i r-te vrste.

Ako dozvolimo i permutaciju nepoznatih najbolje je za glavni element u k-tom eliminacionom koraku uzeti element $a_{rs}^{(k)}$, za koji je

$$|a_{rs}^{(k)}| = \max_{k \le i,j \le n} |a_{ij}^{(k)}|,$$

uz permutaciju k-te i r-te vrste i k-te i s-te kolone. Ovakav postupak se naziva *metod sa totalnim izborom glavnog elementa*.

Odredimo sada broj aritmetičkih operacija u GAUSSovom metodu pri rešavanju sistema od n jednačina sa n nepoznatih.

Kod redukcije sistema na trougaoni oblik, u prvom eliminacionom koraku, potrebno je $n-1$ deljenja, $n(n-1)$ množenja i $n(n-1)$ oduzimanja, što iznosi $(n-1)(2n+1) = 2(n-1)^2 + 3(n-1)$. Na osnovu ovoga, može se zaključiti da je potreban broj aritmetičkih operacija u k-tom eliminacionom koraku $2(n-k)^2 + 3(n-k)$, pa je ukupan broj operacija za trougaonu redukciju

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

Pri rešavanju sistema sa trougaonom matricom potrebno je n deljenja, $n(n-1)/2$ množenja i $n(n-1)/2$ oduzimanja, što iznosi ukupno n^2 operacija. Dakle, ukupan broj računskih operacija u GAUSSovom metodu iznosi

$$N(n) = \frac{1}{6}(4n^3+9n^2-7n).$$

Za dovoljno veliko n imamo $N(n) \cong 2n^3/3$. Dugo vremena se mislilo da je GAUSSov metod eliminacije optimalan u pogledu broja aritmetičkih operacija. V. STRASSEN[128] [70] je 1969. godine, uvodeći iterativni algoritam za množenje i inverziju matrica, dao metod za rešavanje sistema linearnih jednačina, kod koga je broj računskih operacija reda $n^{\log_2 7}$. STRASSENov metod je, dakle, efikasniji od GAUSSovog metoda ($\log_2 7 < 3$).

Napomena 1.2.1. Prema STRASSENovom algoritmu množenje matrica drugog reda

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

izvodi se pomoću

$$s_1 = (a_{11}+a_{22})(b_{11}+b_{22}), \quad s_2 = (a_{21}+a_{22})b_{11}, \quad s_3 = a_{22}(b_{12}-b_{11}),$$

$$s_4 = a_{22}(b_{21}-b_{11}), \quad s_5 = (a_{11}+a_{12})b_{22},$$

$$s_6 = (a_{21}-a_{11})(b_{11}+b_{12}), \quad s_7 = (a_{12}-a_{22})(b_{21}+b_{22})$$

i

$$c_{11} = s_1+s_4-s_5+s_7, \quad c_{21} = s_2+s_4, \quad c_{12} = s_3+s_5, \quad c_{22} = s_1+s_3-s_2+s_6.$$

[128] VOLKER STRASSEN (1936 –), nemački matematičar. Sada je profesor emeritus na Univerzitetu u Konstanci (Nemačka).

Trougaona redukcija obezbeđuje lako izračunavanje determinante sistema. Naime, važi

$$\det A = a_{11}^{(1)} a_{22}^{(2)} \cdots a_{nn}^{(n)}.$$

Ukoliko je korišćen GAUSSov metod sa izborom glavnog elementa treba samo voditi računa o broju permutacija vrsta (i kolona kod metoda sa totalnim izborom glavnog elementa), koje utiču na znak determinante. Ovakav način za izračunavanje determinante je veoma efikasan. Na primer, za izračunavanje determinante reda $n = 30$, potrebno je 0.18 ms, ako se jedna aritmetička operacija obavlja za 10 ns.

Pored izloženih varijanti GAUSSovog eliminacionog metoda za rešavanje sistema linearnih jednačina postoje i drugi direktni metodi koji transformišu sistem jednačina (1.2.2) na neki sistem, poput (1.2.3), koji je pogodan za rešavanje. Na primer, ako je transformisani sistem dijagonalan, tj.

$$D\boldsymbol{x} = \boldsymbol{c} \qquad \big(D = \operatorname{diag}(d_{11}, \ldots, d_{nn})\big), \tag{1.2.5}$$

njegovo rešavanje je veoma jednostavno

$$x_k = \frac{c_k}{d_{kk}} \quad (k = 1, 2, \ldots, n).$$

Štaviše, ako je D jedinična matrica, onda je već postupkom eliminacije završeno i rešavanje sistema jednačina (1.2.5), dakle, $\boldsymbol{x} = \boldsymbol{c}$.

Izložićemo sada postupak za transformaciju sistema (1.2.2) na oblik (1.2.5), koji je poznat ako GAUSS–JORDANov metod. Prvi korak u ovom metodu isti je kao u GAUSSovom metodu. Neka je, kao i ranije, $a_{ij}^{(1)} = a_{ij}$, $b_i^{(1)} = b_i$. Pod pretpostavkom da je $a_{11}^{(1)} \neq 0$, posle prvog eliminacionog koraka, dobijamo $n-1$ jednačina koje pridružujemo prvoj jednačini. Tako imamo

$$\begin{aligned}
a_{11}^{(1)} x_1 + a_{12}^{(2)} x_2 + a_{13}^{(2)} x_3 + \cdots + a_{1n}^{(2)} x_n &= b_1^{(2)}, \\
a_{22}^{(2)} x_2 + a_{23}^{(2)} x_3 + \cdots + a_{2n}^{(2)} x_n &= b_2^{(2)}, \\
a_{32}^{(2)} x_2 + a_{33}^{(2)} x_3 + \cdots + a_{3n}^{(2)} x_n &= b_3^{(2)}, \\
&\vdots \\
a_{n2}^{(2)} x_2 + a_{n3}^{(2)} x_3 + \cdots + a_{nn}^{(2)} x_n &= b_n^{(2)},
\end{aligned}$$

gde smo dodatno izvršili prenumeraciju $a_{1j}^{(2)} = a_{1j}^{(1)}$ $(j = 2, \ldots, n)$ i $b_1^{(2)} = b_1^{(1)}$.

Pod pretpostavkom da je $a_{22}^{(2)} \neq 0$, u drugom eliminacionom koraku eliminišemo x_2 (tj. anuliramo koeficijente uz x_2) iz svih jednačina sem iz druge. Naime, kod GAUSSovog metoda smo eliminaciju sproveli samo u jednačinama koje se nalaze posle druge jednačine. Ovde, dakle, eliminaciju sprovodimo i u prvoj jednačini. Tako, sa faktorima $m_{i2} = a_{i2}^{(2)}/a_{22}^{(2)}$ $(i \neq 2)$, dobijamo

$$\begin{aligned} a_{11}^{(1)}x_1 \quad & +a_{13}^{(3)}x_3 + \cdots + a_{1n}^{(3)}x_n = b_1^{(3)}, \\ a_{22}^{(2)}x_2 \; & +a_{23}^{(3)}x_3 + \cdots + a_{2n}^{(3)}x_n = b_2^{(3)}, \\ & a_{33}^{(3)}x_3 + \cdots + a_{3n}^{(3)}x_n = b_3^{(3)}, \\ & \qquad\qquad\qquad\quad \vdots \\ & a_{n3}^{(3)}x_3 + \cdots + a_{nn}^{(3)}x_n = b_n^{(3)}, \end{aligned}$$

gde su $a_{2j}^{(3)} = a_{2j}^{(2)}$ $(j = 3, \ldots, n)$, $b_2^{(3)} = b_2^{(2)}$ i

$$a_{ij}^{(3)} = a_{ij}^{(2)} - m_{i2}a_{2j}^{(2)}, \quad b_i^{(3)} = b_i^{(2)} - m_{i2}b_2^{(2)} \quad (i \neq 2;\ j = 3, \ldots, n).$$

Nastavljajući ovaj postupak posle n koraka dobijamo ekvivalentni dijagonalni sistem jednačina

$$\begin{aligned} a_{11}^{(1)}x_1 \quad & = b_1^{(n+1)}, \\ a_{22}^{(2)}x_2 \quad & = b_2^{(n+1)}, \\ & \quad \vdots \\ a_{nn}^{(n)}x_n & = b_n^{(n+1)}. \end{aligned} \tag{1.2.6}$$

Primetimo da je ovde neophodno n koraka, za razliku od GAUSSovog metoda gde je bilo dovoljno $n-1$ koraka. Naime, ovde se posle $n-1$ koraka dobija sistem jednačina oblika

$$\begin{aligned} a_{11}^{(1)}x_1 \quad & +a_{1n}^{(n)}x_n = b_1^{(n)}, \\ a_{22}^{(2)}x_2 \quad & +a_{2n}^{(n)}x_n = b_2^{(n)}, \\ & \quad \vdots \\ & a_{nn}^{(n)}x_n = b_n^{(n)}. \end{aligned}$$

U poslednjem n-tom koraku treba eliminisati x_n iz prvih $n-1$ jednačina.

GAUSS–JORDANov metod zahteva više operacija od GAUSSovog metoda. Naime, ovde je u k-tom eliminacionom koraku potrebno $n-1$ deljenja, zatim $(n-1)(n-k+1)$ množenja i isto toliko oduzimanja, što znači da je za redukciju sistema na dijagonalni oblik potrebno $n(n-1)$ deljenja i po $\frac{1}{2}n(n^2-1)$ množenja i oduzimanja. Rešavanje dijagonalnog sistema zahteva još n deljenja. Prema tome, ukupno je potrebno n^2 deljenja i po $\frac{1}{2}n(n^2-1)$ množenja i oduzimanja, što sve iznosi

$$\overline{N}(n) = n^3 + n^2 - n.$$

Ovo je jasno veće u poređenju sa GAUSSovim metodom. Naime, za dovoljno veliko n imamo da je $\overline{N}(n) \cong \frac{3}{2}N(n)$, gde je $N(n)$ broj operacija kod GAUSSovog metoda.

Na osnovu (1.2.6) vidimo da su dijagonalni elementi u matričnom sistemu jednačina (1.2.5) dati sa $d_{kk} = a_{kk}^{(k)}$ i da su $c_k = b_k^{(n+1)}$.

I ovde, kao i kod GAUSSovog metoda, na potpuno istovetan način se može izvesti izbor glavnih elemenata i pomoću njih sprovesti eliminacioni postupak.

Ako se pre svakog eliminacionog koraka izvrši deljenje odgovarajuće jednačine sa glavnim elementom, tj. pre prvog koraka podeli prva jednačina, pre drugog koraka podeli druga jednačina, itd., transformisani sistem imaće jediničnu matricu $D = I$. Naravno, ovakav postupak zahteva veći broj operacija.

4.1.3 Inverzija matrica

Neka je $A = [a_{ij}]_{n\times n}$ regularna matrica i neka je

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & & x_{2n} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \ldots\ \boldsymbol{x}_n]$$

njena inverzna matrica. Vektori $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_n$ su redom prva, druga, ..., n-ta kolona matrice X. Definišimo vektore $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ pomoću

$$\boldsymbol{e}_1 = [1\ 0\ \cdots\ 0]^T, \quad \boldsymbol{e}_2 = [0\ 1\ \cdots\ 0]^T, \ \ldots, \ \boldsymbol{e}_n = [0\ 0\ \cdots\ 1]^T.$$

S obzirom na jednakost $AX = [A\boldsymbol{x}_1\ A\boldsymbol{x}_2\ \cdots\ A\boldsymbol{x}_n] = I = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \ldots\ \boldsymbol{e}_n]$, problem određivanja inverzne matrice može se svesti na rešavanje n sistema linearnih jednačina

$$A\boldsymbol{x}_i = \boldsymbol{e}_i \quad (i = 1, \ldots, n). \tag{1.3.1}$$

Za rešavanje sistema (1.3.1) pogodno je koristiti GAUSSov metod, s obzirom da se matrica A pojavljuje kao matrica svih sistema, pa njenu trougaonu redukciju treba izvršiti samo jednom. Pri ovome sve elementarne transformacije koje su potrebne za trougaonu redukciju matrice A treba primeniti i na jediničnu matricu $I = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_n]$. Na taj način se matrica A transformiše u trougaonu matricu R, a matrica I u matricu $C = [\boldsymbol{c}_1\ \boldsymbol{c}_2\ \cdots\ \boldsymbol{c}_n]$. Najzad, ostaje da se reše trougaoni sistemi $R\boldsymbol{x}_i = \boldsymbol{c}_i$ $(i = 1, \ldots, n)$. Dakle, primena GAUSSovog metoda može se iskazati kao transformacija

$$\left[A \mid I\right] \ \rightarrow\ \left[R \mid C\right].$$

Za inverziju matrica vrlo često se koristi GAUSS–JORDANov metod i to u varijanti sa jediničnom dijagonalom. U ovom slučaju, odgovarajuća transformacija je

$$\left[A \mid I\right] \ \rightarrow\ \left[I \mid A^{-1}\right],$$

što znači da se na mesto jedinične matrice pojavljuje inverzna matrica.

Primer 1.3.1. Neka je data matrica

$$A = \begin{bmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix}.$$

Primenom GAUSS–JORDANovog metoda na $\left[A \mid I\right]$ imamo redom

$$\left[\begin{array}{ccc|ccc} 1 & 1/3 & 2 & 1/3 & 0 & 0 \\ 2 & 1 & 3 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 1/3 & 2 & 1/3 & 0 & 0 \\ 0 & 1/3 & -1 & -2/3 & 1 & 0 \\ 0 & 2/3 & -1 & -1/3 & 0 & 1 \end{array}\right] \rightarrow$$

$$\rightarrow \left[\begin{array}{ccc|ccc} 1 & 1/3 & 2 & 1/3 & 0 & 0 \\ 0 & 1 & -3 & -2 & 3 & 0 \\ 0 & 2/3 & -1 & -1/3 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 3 & 1 & -1 & 0 \\ 0 & 1 & -3 & -2 & 3 & 0 \\ 0 & 0 & 1 & 1 & 2 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -2 & 5 & -3 \\ 0 & 1 & 0 & 1 & -3 & 3 \\ 0 & 0 & 1 & 1 & -2 & 1 \end{array}\right].$$

Dakle,

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

4.1.4 Faktorizacioni metodi

Faktorizacioni metodi za rešavanje sistema linearnih jednačina zasnivaju se na razlaganju matrice sistema na proizvod dve matrice čiji je oblik takav da omogućava svođenje sistema na dva sistema jednačina koji se jednostavno sukcesivno rešavaju. U ovom odeljku ukazaćemo na metode zasnovane na LR faktorizaciji matrice (videti odeljak 2.3.2).

Neka je dat sistem jednačina

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{1.4.1}$$

sa kvadratnom matricom A, čiji su svi glavni dijagonalni minori različiti od nule. Tada, na osnovu teoreme 3.2.1 (glava 2), postoji faktorizacija matrice $A = LR$, gde je L donja i R gornja trougaona matrica. Faktorizacija je jednoznačno određena, ako se, na primer, usvoji da matrica L ima jediničnu dijagonalu. U tom slučaju, sistem (1.4.1), tj. $LR\boldsymbol{x} = \boldsymbol{b}$, se može predstaviti u ekvivalentnom obliku

$$L\boldsymbol{y} = \boldsymbol{b}, \quad R\boldsymbol{x} = \boldsymbol{y}. \tag{1.4.2}$$

Na osnovu prethodnog, za rešavanje sistema jednačina (1.4.1), može se formulisati sledeći metod:

1° stavimo $\ell_{ii} = 1 \quad (i = 1, \ldots, n)$,

2° odredimo ostale elemente matrice $L = [\ell_{ij}]_{n\times n}$ i matrice $R = [r_{ij}]_{n\times n}$ (videti odeljak 2.3.2),

3° rešimo prvi sistem jednačina u (1.4.2),

4° rešimo drugi sistem jednačina u (1.4.2),

Koraci 3° i 4° se jednostavno izvode. Naime, neka su

$$\boldsymbol{b} = [b_1\ b_2\ \cdots\ b_n]^T, \quad \boldsymbol{y} = [y_1\ y_2\ \cdots\ y_n]^T, \quad \boldsymbol{x} = [x_1\ x_2\ \cdots\ x_n]^T.$$

Tada je

$$y_1 = b_1, \quad y_i = b_i - \sum_{k=1}^{i-1} \ell_{ik} y_k \quad (i = 2, \ldots, n)$$

i

$$x_n = \frac{y_n}{r_{nn}}, \quad x_i = \frac{1}{r_{ii}}\left(y_i - \sum_{k=i+1}^{n} r_{ik} x_k\right) \quad (i = n-1, \ldots, 1).$$

Izloženi metod se u literaturi sreće kao CHOLESKYev[129] metod. U slučaju kada je matrica A normalna, tj. kada je simetrična i pozitivno definitna, CHOLESKYev metod se može uprostiti. Naime, tada se može uzeti da je $L = R^T$. Dakle, treba odrediti faktorizaciju matrice A u obliku $A = R^T R$. Na osnovu formula iz odeljka 2.3.2, za elemente matrice R važe formule

$$\left.\begin{aligned} & r_{11} = \sqrt{a_{11}}, \quad r_{1j} = \frac{a_{1j}}{r_{11}} \quad (j = 2, \ldots, n) \\ & r_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2} \\ & r_{ij} = \frac{1}{r_{ii}} \Big(a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj} \Big) \quad (j = i+1, \ldots, n) \end{aligned} \right\} \quad (i = 2, \ldots, n).$$

U ovom slučaju, sistemi (1.4.2) postaju

$$R^T \boldsymbol{y} = \boldsymbol{b}, \quad R\boldsymbol{x} = \boldsymbol{y}.$$

Navedena modifikacija CHOLESKYevog metoda se naziva *metod kvadratnog korena*.

Napomena 1.4.1. Determinanta normalne matrice se može izračunati metodom kvadratnog korena

$$\det A = (r_{11} r_{22} \cdots r_{nn})^2.$$

Faktorizacioni metodi su naročito pogodni za rešavanje sistema linearnih jednačina, kod kojih se matrica sistema ne menja, već samo vektor slobodnih članova $\boldsymbol{b}$. Ovakvi sistemi se često javljaju u tehnici[130].

Sada ćemo pokazati da se GAUSSov metod eliminacije može interpretirati kao LR faktorizacija matrice A. Uzmimo matricu A takvu, da prilikom eliminacije ne treba vršiti permutaciju vrsta i kolona. Polazni sistem označimo sa $A^{(1)}\boldsymbol{x} = \boldsymbol{b}^{(1)}$. GAUSSov eliminacioni postupak daje $n-1$ ekvivalentnih sistema $A^{(2)}\boldsymbol{x} = \boldsymbol{b}^{(2)}$, …, $A^{(n)}\boldsymbol{x} = \boldsymbol{b}^{(n)}$, pri čemu matrica $A^{(k)}$ ima oblik takav da su svi njeni elementi ispod glavne dijagonale i ispred k-te kolone jednaki nuli, tj.

[129] ANDRÉ-LOUIS CHOLESKY (1875 – 1918), oficir francuske vojske i matematičar.
[130] Na primer, u elektronici kod izračunavanja prenosnih karakteristika elektronskih kola.

$$A^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1k}^{(1)} & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & & a_{2k}^{(2)} & & a_{2n}^{(2)} \\ & & \ddots & \vdots & & \\ & & & a_{kk}^{(k)} & & a_{kn}^{(k)} \\ & & & \vdots & & \\ & & & a_{nk}^{(k)} & & a_{nn}^{(k)} \end{bmatrix}.$$

Analizirajmo modifikaciju elementa $a_{ij} = a_{ij}^{(1)}$ u procesu trougaone redukcije. Kako je, za $k = 1,2,\ldots,n-1$,

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)} \qquad (i,j = k+1,\ldots,n)$$

i

$$a_{i1}^{(k+1)} = a_{i2}^{(k+1)} = \cdots = a_{ik}^{(k+1)} = 0 \quad (i = k+1,\ldots,n),$$

sumiranjem dobijamo

$$\sum_{k=1}^{i-1} a_{ij}^{(k+1)} = \sum_{k=1}^{i-1} a_{ij}^{(k)} - \sum_{k=1}^{i-1} m_{ik}a_{kj}^{(k)} \quad (i \le j)$$

i

$$\sum_{k=1}^{j} a_{ij}^{(k+1)} = \sum_{k=1}^{j} a_{ij}^{(k)} - \sum_{k=1}^{j} m_{ik}a_{kj}^{(k)} \quad (i > j),$$

tj.

$$a_{ij} = a_{ij}^{(1)} = a_{ij}^{(i)} + \sum_{k=1}^{i-1} m_{ik}a_{kj}^{(k)} \quad (i \le j),$$

i

$$a_{ij} = a_{ij}^{(1)} = 0 + \sum_{k=1}^{j} m_{ik}a_{kj}^{(k)} \quad (i > j).$$

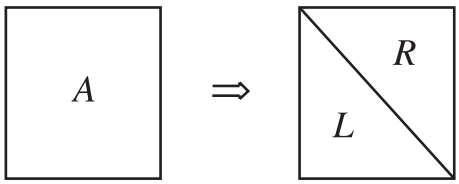
Definišući $m_{ii} = 1$ $(i = 1,\ldots,n)$, poslednje dve jednakosti se mogu predstaviti u obliku

$$a_{ij} = \sum_{k=1}^{p} m_{ik}\, a_{kj}^{(k)} \qquad (i,j = 1,\ldots,n), \tag{1.4.3}$$

gde je $p = \min\{i,j\}$. Jednakost (1.4.3) ukazuje da GAUSSova eliminacija daje LR faktorizaciju matrice A, gde su

$$L = \begin{bmatrix} 1 & & & \\ m_{21} & 1 & & \\ \vdots & & \ddots & \\ m_{n1} & m_{n2} & \cdots & 1 \end{bmatrix}, \quad R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & & r_{2n} \\ & & \ddots & \\ & & & r_{nn} \end{bmatrix}$$

i $r_{kj} = a_{kj}^{(k)}$. Pri programskoj realizaciji GAUSSovog metoda, u cilju dobijanja LR faktorizacije matrice A, nije potrebno koristiti nove memorijske elemente za pamćenje matrice L, već je pogodno faktore m_{ik} smeštati na mesto koeficijenata matrice A koji se anuliraju u procesu trougaone redukcije. Na taj način, posle završene trougaone redukcije, na mesto matrice A biće memorisane matrice L i R, prema sledećoj šemi



Uočimo da se dijagonalni elementi matrice L, koji su svi jednaki jedinici, ne moraju memorisati.

CHOLESKYev metod, zasnovan na LR faktorizaciji, primenjuje se u slučajevima kada matrica A ispunjava uslove teoreme 3.2.1 (glava 2). Međutim, primenljivost ovog metoda može se proširiti i na druge sisteme sa regularnom matricom, uzimajući u obzir permutaciju jednačina u sistemu. Za faktorizaciju iskoristimo GAUSSov eliminacioni metod sa izborom glavnog elementa. Pri ovome biće $LR = A'$, gde se matrica A' dobija iz matrice A konačnim brojem razmena vrsta. Ovo znači da u procesu eliminacije treba memorisati niz indeksa glavnih elemenata $I = (p_1, \ldots, p_{n-1})$, pri čemu je p_k broj vrste iz koje se uzima glavni element u k-tom eliminacionom koraku. Kod rešavanja sistema $A\boldsymbol{x} = \boldsymbol{b}$, neposredno posle faktorizacije treba, u skladu sa nizom indeksa I, permutovati koordinate vektora $\boldsymbol{b}$. Na taj način se dobija transformisani vektor $\boldsymbol{b}'$, pa se rešavanje datog sistema svodi na sukcesivno rešavanje trougaonih sistema

$$L\boldsymbol{y} = \boldsymbol{b}' \quad \text{i} \quad R\boldsymbol{x} = \boldsymbol{y}.$$

Primetimo da se pomoću indeksnog niza I može konstruisati permutaciona matrica P, takva da je $A' = PA$ i $\boldsymbol{b}' = P\boldsymbol{b}$, što znači da se, u ovom slučaju, radi o faktorizaciji $LR = PA$.

Ilustrujmo ovo jednim primerom.

Primer 1.4.1. Primenom GAUSSovog metoda sa izborom glavnog elementa na

$$A = \begin{bmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix}$$

dobijamo redom

$$\begin{bmatrix} \textcircled{3} & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 3 & 1 & 6 \\ 2/3 & 1/3 & -1 \\ 1/3 & \textcircled{2/3} & -1 \end{bmatrix} \Rightarrow \begin{bmatrix} 3 & 1 & 6 \\ 1/3 & 2/3 & -1 \\ 2/3 & 1/2 & -1/2 \end{bmatrix},$$

pri čemu je indeksni niz glavnih elemenata $I = (1,3)$. Glavni elementi su u navedenim matricama zaokruženi. Ovde je

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 1/2 & 1 \end{bmatrix} \quad \text{i} \quad L = \begin{bmatrix} 3 & 1 & 6 \\ 0 & 2/3 & -1 \\ 0 & 0 & -1/2 \end{bmatrix}.$$

Rešimo sada sistem $A\boldsymbol{x} = \boldsymbol{b} = [2 \quad 7 \quad 4]^T$. Kako je $\boldsymbol{b}' = [2 \quad 4 \quad 7]^T$ iz $L\boldsymbol{y} = \boldsymbol{b}'$ sleduje

$$y_1 = 2, \quad y_2 = 4 - \frac{1}{3}y_1 = \frac{10}{3}, \quad y_3 = 7 - \frac{2}{3}y_1 - \frac{1}{2}y_2 = 4.$$

Najzad, na osnovu $R\boldsymbol{x} = \boldsymbol{y}$, nalazimo

$$x_3 = -8, \quad x_2 = \frac{3}{2}\left(\frac{10}{3} + x_3\right) = -7, \quad x_1 = \frac{1}{3}(2 - x_2 - 6x_3) = 19.$$

Permutaciona matrica, u ovom slučaju, ima oblik

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \quad \triangle$$

4.1.5 Metod ortogonalizacije

Posmatrajmo sistem jednačina (1.2.1) sa regularnom matricom. Ako definišemo $(n+1)$-dimenzionalne vektore

$$\boldsymbol{y} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ 1 \end{bmatrix}, \quad \boldsymbol{a}_i = \begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{in} \\ -b_i \end{bmatrix} \quad (i = 1, \ldots, n),$$

tada se ovaj sistem može predstaviti u obliku

$$(\boldsymbol{a}_i, \boldsymbol{y}) = \boldsymbol{y}^T \boldsymbol{a}_i = 0 \quad (i = 1, \ldots, n). \tag{1.5.1}$$

Jednačine (1.5.1) ukazuju na mogućnost rešavanja sistema (1.2.1), korišćenjem uslova ortogonalnosti vektora $\boldsymbol{y}$ sa vektorima $\boldsymbol{a}_i$ $(i = 1, \ldots, n)$. Pomenuta ortogonalnost je ekvivalentna ortogonalnosti vektora $\boldsymbol{y}$ sa linearnim potprostorom H_n, koji je generisan vektorima $\boldsymbol{a}_i$ $(i = 1, \ldots, n)$, tj. $H_n = L(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$. Polazeći od baze $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}$, korišćenjem GRAM-SCHMIDTovog postupka, konstruišimo ortonormiranu bazu $B_n = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ potprostora H_n. Vektor $\boldsymbol{y}$ je, očigledno, ortogonalan sa svim vektorima ortonormirane baze B_n.

Kako je vektor $\boldsymbol{a}_{n+1} = [0\ 0\ \cdots\ 0\ 1]^T$ linearno nezavisan u odnosu na vektore baze B_n, izvršićemo ortogonalizaciju i ovog vektora. Dakle,

$$\boldsymbol{u} = \boldsymbol{a}_{n+1} - \sum_{i=1}^{n} (\boldsymbol{a}_{n+1}, \boldsymbol{v}_i) \boldsymbol{v}_i.$$

Neka su koordinate vektora $\boldsymbol{u}$ redom $u_1, \ldots, u_n, u_{n+1}$, tj. $\boldsymbol{u} = [u_1\ \cdots\ u_n\ u_{n+1}]^T$. Kako vektor $\boldsymbol{u}$ ispunjava uslove ortogonalnosti $(\boldsymbol{u}, \boldsymbol{v}_i) = 0$ $(i = 1, \ldots, n)$, na osnovu prethodnog, zaključujemo da je, takođe, $(\boldsymbol{u}, \boldsymbol{a}_i) = 0$ $(i = 1, \ldots, n)$, tj.

$$\begin{aligned}
(\boldsymbol{u}, \boldsymbol{a}_1) &= a_{11}u_1 + a_{12}u_2 + \cdots + a_{1n}u_n - b_1 u_{n+1} = 0, \\
(\boldsymbol{u}, \boldsymbol{a}_2) &= a_{21}u_1 + a_{22}u_2 + \cdots + a_{2n}u_n - b_2 u_{n+1} = 0, \\
&\vdots \\
(\boldsymbol{u}, \boldsymbol{a}_n) &= a_{n1}u_1 + a_{n2}u_2 + \cdots + a_{nn}u_n - b_n u_{n+1} = 0.
\end{aligned}$$

Primetimo da je $u_{n+1} \neq 0$. Naime, ako bi bilo $u_{n+1} = 0$, tada bi n-torka $(u_1, u_2, \ldots, u_n)$ bila rešenje homogenog sistema jednačina sa matricom $A = [a_{ij}]$.

Međutim, kako homogeni sistem ($\det A \neq 0$) ima samo trivijalna rešenja, to bismo imali $u_i = 0$ $(i = 1,\ldots,n)$, pa bi vektor $\boldsymbol{a}_{n+1}$ bio linearna kombinacija vektora baze B_n, što je u kontradikciji sa izborom ovog vektora.

Ako sve jednačine poslednjeg sistema podelimo sa u_{n+1}, lako se zaključuje da je vektor $\boldsymbol{u}$, sa

$$x_i = \frac{u_i}{u_{n+1}} \quad (i = 1,\ldots,n),$$

rešenje sistema jednačina (1.5.1). Tada je n-torka $(x_1, x_2, \ldots, x_n)$ rešenje sistema (1.2.1).

Na kraju, primetimo da navedeni metod ortogonalizacije zahteva veći broj operacija množenja i deljenja nego GAUSSov metod eliminacije.

4.1.6 Analiza greške i slabouslovljeni sistemi

U ovom odeljku posmatraćemo uticaj promene vektora $\boldsymbol{b}$ na rešenje $\boldsymbol{x}$ u sistemu jednačina

$$A\boldsymbol{x} = \boldsymbol{b},$$

sa regularnom matricom A. Neka se vektor $\boldsymbol{b}$ promeni za $\Delta\boldsymbol{b}$. Tada će se rešenje promeniti za $\Delta\boldsymbol{x}$, tj. imaćemo

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

odakle je $A\,\Delta\boldsymbol{x} = \Delta\boldsymbol{b}$, tj. $\Delta\boldsymbol{x} = A^{-1}\Delta\boldsymbol{b}$. Iz ove jednakosti sleduje

$$\|\Delta\boldsymbol{x}\| \leq \|A^{-1}\| \cdot \|\Delta\boldsymbol{b}\|. \tag{1.6.1}$$

Kako je $\|\boldsymbol{b}\| = \|A\boldsymbol{x}\| \leq \|A\| \cdot \|\boldsymbol{x}\|$, imamo

$$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|} = k(A)\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}, \tag{1.6.2}$$

gde je $k(A) = \mathrm{cond}(A) = \|A\| \cdot \|A^{-1}\|$. Za broj $k(A)$ kažemo da je *faktor uslovljenosti* ili *kondicioni broj* matrice A. Faktor uslovljenosti zavisi od upotrebljene norme matrice, ali je uvek $k(A) \geq 1$, što sleduje iz činjenice da je

$$\|\boldsymbol{x}\| = \|I\boldsymbol{x}\| = \|AA^{-1}\boldsymbol{x}\| \leq \|A\| \cdot \|A^{-1}\| \cdot \|\boldsymbol{x}\|.$$

Što je faktor $k(A)$ veći od jedinice kažemo da je matrica A slabije uslovljena. Za sistem sa slabouslovljenom matricom kažemo da je *slabouslovljeni sistem*[131].

[131] U anglo-saksonskoj literaturi: *ill-conditioned systems*, a u ruskoj: *плохо обусловленные системы*.

Kada se koristi spektralna norma, faktor uslovljenosti je dat sa

$$k(A) = \sigma(A)\sigma(A^{-1}) = \sqrt{\frac{\max|\lambda(A^*A)|}{\min|\lambda(A^*A)|}}.$$

Ukoliko je matrica A hermitska, prethodni izraz se pojednostavljuje, tj. postaje

$$k(A) = \frac{\max|\lambda(A)|}{\min|\lambda(A)|}.$$

Zbog svojstva minimalnosti spektralne norme, ova vrednost broja $k(A)$ za hermitsku matricu A je najmanja u odnosu na vrednosti koje se dobijaju korišćenjem drugih normi.

Nejednakost (1.6.1) se može interpretirati i na sledeći način. Neka je $\boldsymbol{x}_p$ približno rešenje jednačine $A\boldsymbol{x} = \boldsymbol{b}$. Sa $\boldsymbol{r}(\boldsymbol{x}_p)$ označimo odgovarajući *„vektor ostatak“*, tj.

$$\boldsymbol{r}(\boldsymbol{x}_p) = A\boldsymbol{x}_p - \boldsymbol{b} = A(\boldsymbol{x}_p - \boldsymbol{x}).$$

Dakle, $\boldsymbol{x}_p$ je tačno rešenje jednačine $A\boldsymbol{x}_p = \boldsymbol{b} + \boldsymbol{r}(\boldsymbol{x}_p)$ i za grešku važi

$$\|\Delta\boldsymbol{x}\| \leq \|A^{-1}\| \cdot \|\boldsymbol{r}(\boldsymbol{x}_p)\|. \tag{1.6.3}$$

Na osnovu ove nejednakosti može se zaključiti da vrednost norme $\|\Delta\boldsymbol{x}\|$ može biti velika, čak i u slučajevma kada je veličina $\|\boldsymbol{r}(\boldsymbol{x}_p)\|$ dovoljno mala.

Teorema 1.6.1. *Neka je A regularna matrica reda n, $B = A(I+F)$, $\|F\| < 1$ i vektori $\boldsymbol{x}$ i $\Delta\boldsymbol{x}$ definisani pomoću $A\boldsymbol{x} = \boldsymbol{b}$, $B(\boldsymbol{x}+\Delta\boldsymbol{x}) = \boldsymbol{b}$. Tada važe sledeće ocene:*

1° $\dfrac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \dfrac{\|F\|}{1-\|F\|}$;

2° $\dfrac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \dfrac{k(A)}{1-k(A)\dfrac{\|B-A\|}{\|A\|}} \cdot \dfrac{\|B-A\|}{\|A\|}$ *ako je* $k(A)\dfrac{\|B-A\|}{\|A\|} < 1$.

Dokaz. Kako, na osnovu učinjenih pretpostavki, B^{-1} egzistira, imamo

$$\Delta\boldsymbol{x} = B^{-1}\boldsymbol{b} - A^{-1}\boldsymbol{b} = B^{-1}(A-B)A^{-1}\boldsymbol{b}, \quad \boldsymbol{x} = A^{-1}\boldsymbol{b},$$

$$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \|B^{-1}(A-B)\| = \|-(I+F)^{-1}A^{-1}AF\| \leq \|(I+F)^{-1}\| \cdot \|F\| \leq \frac{\|F\|}{1-\|F\|}.$$

Imajući u vidu da je $F = A^{-1}(B-A)$ i $\|F\| \leq k(A)\|B-A\|/\|A\|$, na osnovu poslednje nejednakosti, dobijamo nejednakost 2°. □

Pomoću teoreme 1.6.1 moguće je oceniti grešku u rešenju $\boldsymbol{x}$, pri zameni matrice A nekom drugom matricom. Naime, ako stavimo $C = (I+F)^{-1} = B^{-1}A$, $F = A^{-1}B - I$, imamo

$$\|B^{-1}A\| \leq \frac{1}{1 - \|A^{-1}B - I\|}.$$

S druge strane, iz $A^{-1} = A^{-1}BB^{-1}$ sleduje

$$\|A^{-1}\| \leq \|A^{-1}B\| \cdot \|B^{-1}\| \leq \frac{\|B^{-1}\|}{1 - \|I - B^{-1}A\|},$$

pri čemu smo iskoristili jednakost $A^{-1}B = (B^{-1}A)^{-1} = (I - (I - B^{-1}A))^{-1}$.

Najzad, na osnovu (1.6.3), dobijamo

$$\|\boldsymbol{x}_p - \boldsymbol{x}\| \leq \frac{\|B^{-1}\|}{1 - \|I - B^{-1}A\|} \|\boldsymbol{r}(\boldsymbol{x}_p)\| \qquad (\|I - B^{-1}A\| < 1).$$

Primer 1.6.1. Jedan tipičan primer slabouslovljenog sistema je sledeći sistem jednačina

$$\begin{bmatrix} 121734 & 169217 & 176624 & 166662 \\ 169217 & 235222 & 245505 & 231653 \\ 176624 & 245505 & 256423 & 242029 \\ 166662 & 231653 & 242029 & 228474 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 634237 \\ 881597 \\ 920581 \\ 868818 \end{bmatrix},$$

čije je tačno rešenje $x_1 = x_2 = x_3 = x_4 = 1$. Međutim, ako koordinate vektora slobodnih članova variraju samo za ± 1 i recimo budu

$$b_1 = 634238, \quad b_2 = 881596, \quad b_3 = 920580, \quad b_4 = 868819,$$

tačno rešenje sistema postaje

$$x_1^* = 130214370, \quad x_2^* = -78645876, \quad x_3^* = -32701403, \quad x_4^* = 19395881.$$

Uzimajući normu $\|\cdot\|_\infty$, u ovom primeru imamo

$$\|\boldsymbol{b}\|_\infty = \max_{1 \leq i \leq 4} |b_i| = 920580, \quad \|\Delta\boldsymbol{b}\|_\infty = \max_{1 \leq i \leq 4} |\Delta b_i| = 1,$$

tj. $\|\Delta\boldsymbol{b}\|_\infty / \|\boldsymbol{b}\|_\infty \cong 1.1 \times 10^{-6}$, dok je odgovarajuća promena u rešenju

$$\|\boldsymbol{x}\|_\infty = \max_{1\le i\le 4} |x_i| = 1, \quad \|\Delta\boldsymbol{x}\|_\infty = \max_{1\le i\le 4} |x_i^* - x_i| = 130214369,$$

tj. $\|\Delta\boldsymbol{x}\|_\infty/\|\boldsymbol{x}\|_\infty \cong 1.3\times 10^8$. Dakle, relativna promena od samo 10^{-6} u slobodnom članu izaziva ogromnu promenu u rešenju (reda 10^8). Na osnovu (1.6.2), dobijamo

$$k(A) \ge \frac{\|\Delta\boldsymbol{x}\|_\infty/\|\boldsymbol{x}\|_\infty}{\|\Delta\boldsymbol{b}\|_\infty/\|\boldsymbol{b}\|_\infty} \cong \frac{1.3\times 10^8}{1.1\times 10^{-6}} \cong 1.18\times 10^{14}.$$

Kako je

$$A^{-1} = \begin{bmatrix} 64975255 & -39243257 & -16317569 & 9678288 \\ -39243257 & 23701842 & 9855360 & -5845418 \\ -16317569 & 9855360 & 4097916 & -2430559 \\ 9678288 & -5845418 & -2430559 & 1441615 \end{bmatrix},$$

zaista, za faktor uslovljenosti, direktnim izračunavanjem, dobijamo

$$\|A\|_\infty = \max_i \left(\sum_{j=1}^{4} |a_{ij}| \right) = 920581, \quad \|A^{-1}\|_\infty = 130214369,$$

tj.

$$k(A) = \|A\|_\infty \|A^{-1}\|_\infty \cong 1.20\times 10^{14}.$$

Napomenimo da smo kod rešavanja prethodnog sistema jednačina sva izračunavanja izveli potpuno tačno radeći na skupu svih racionalnih brojeva, što je ekvivalentno korišćenju aritmetike beskonačne dužine. Kod korišćenja aritmetike konačne dužine (što je inače standardni slučaj kod računara), visok faktor uslovljenosti „jede“ cifre u rezultatu. Na primer, ako je faktor uslovljenosti reda 10^m, a želimo rezultat sa relativnom greškom manjom od 10^{-d}, tj. aproksimativno sa d korektnih značajnih cifara u mantisi, neophodno je koristiti aritmetiku sa dužinom najmanje $d+m$.

Ovaj dobro konstruisan primer slabouslovljene matrice je naveden u doktorskoj disertaciji P. MADIĆa[132] [52]. △

Iz datog primera se može videti da je problem rešavanja slabouslovljenih sistema veoma složen i njemu treba pristupati obazrivo.

[132] PETAR B. MADIĆ (1922 – 2009), srpski matematičar i informatičar. Bio je jedan od pionira uvođenja računarstva, programiranja i numeričke analize u univerzitetske programe krajem šezdesetih godina prošlog veka na prostorima Srbije. Autor ove knjige bio je njegov student, a kasnije i saradnik nekoliko godina na predmetima *Obrada informacija* i *Programiranje* na Elektronskom fakultetu u Nišu.

4.2 ITERATIVNI METODI

Drugu važnu klasu metoda u linearnoj algebri čine iterativni metodi, kod kojih se teorijski rezultat dobija posle beskonačnog broja koraka. Praktično, međutim, za nalaženje rešenja sa dovljnom tačnošću potrebno je izvršiti konačan broj koraka. Broj koraka, jasno, zavisi od zahtevane tačnosti. Sa porastom dimenzije matrice, iterativni metodi postaju ozbiljna konkurencija direktnim metodima, imajući na umu ukupan broj neophodnih aritmetičkih operacija.

Važnost iterativnih metoda u linearnoj algebri proizilazi iz činjenice da se u primenama pojavljuju sistemi jednačina sa sve većom dimenzijom i da direktni metodi koji zahtevaju $O(n^3)$ aritmetičkih operacija postaju neupotrebljivi sa porastom n i pored značajnog progresa kako u teorijskom smislu, tako i u računarskoj tehnologiji.

U matričnim izračunavanjima pojam „veliki sistem jednačina“ istorijski posmatrano se značajno menjao (videti, na primer, klasifikaciju iz [71, str. 243–249]). Na svakih petnaestak godina dimenzija tzv. „velikog sistema jednačina“ se uvećavala 10 puta, počev od pedesetih godina prethodnog stoleća (period prepoznatljiv po radovima WILKINSONa), kada se za takav sistem smatrao svaki onaj koji je imao dimenziju veću od $n = 20$. Pojava knjige [29] sredinom šezdesetih godina (tzv. „FORSYTHE[133]–MOLERova[134] era“) pomera tu granicu na $n = 200$, a već osamdesetih godina, sa pojavom paketa LINPACK, dimenzija se povećava na $n = 2000$. Paket LAPACK[135] [6] pisan orginalno[136] na jeziku FORTRAN 77 pojavio se sredinom devedesetih pomerio je granicu „velikih sistema“ na $n = 20000$. Dakle, za nepunih pedeset godina dimenzije matrica sa kojima možemo jednostavno operisati povećale su se za faktor 10^3. Ovaj impresivni progres je, međutim, u velikoj meri uzrokovan mnogo većim progresom koji je postignut u istom periodu u računarskom hardveru, podizanjem brzine sa faktorom od 10^9 (od sekunde do nano sekunde po operaciji), što znači da je broj operacija $O(n^3)$ velika prepreka. Očigledno je da se metodi, kod kojih je broj operacija redukovan na

[133] GEORGE E. FORSYTHE (1917 – 1972), poznati američki naučnik u oblasti kompjuterskih nauka.

[134] CLEVE BARRY MOLER (1939 –), poznati američki matematičar, programer i ekspert u numeričkoj analizi. Jedan je od autora u razvoju FORTRAN programskih paketa za linearnu algebru LINPACK and EISPACK, kreator programskog sistema MATLAB i jedan od osnivača kompanije MathWorks za razvoj i komercijalizaciju ovog sistema.

[135] Skraćenica izvedena iz engleskog naziva *Linear Algebra PACKage*. Paket je nastao kao naslednik prethodnih paketa LINPACK i EISPACK (za probleme sopstvenih vrednosti) i on obezbeđuje rutine za rešavanje sistema linearnih jednačina, problema sopstvenih vredosti i faktorizaciju proizvoljnih realnih ili kompleksnih matrica, poput *SVD* (na engleskom: *Singular Value Decomposition*).

[136] LAPACK verzija 3.2 pojavila se 2008. godine u FORTRANu 90.

$O(n^p)$, gde je $p < 3$, mogu primeniti na matrice znatno veće dimenzije. Za neke klase matrica $O(n^2)$ je postignuto sa iterativnim metodima.

U ovom poglavlju razmatraćemo samo iterativne metode za rešavanje sistema linearnih jednačina i metode za inverziju matrica. Iterativni metodi za rešavanje problema sopstvenih vrednosti (kao uostalom i direktni metodi) biće razmatrani u posebnom poglavlju. Napomenimo još jednom da se za rešavanje sistema sa velikim brojem jednačina, kakvi se javljaju pri rešavanju parcijalnih diferencijalnih jednačina, uglavnom koriste iterativni metodi.

4.2.1 Načini formiranja iterativnih metoda

Posmatrajmo sistem linearnih jednačina

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n, \end{aligned} \tag{2.1.1}$$

koji se može predstaviti i u matričnom obliku

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{2.1.2}$$

gde su

$$A = \begin{bmatrix} a_{11} & a_{12} \cdots a_{1n} \\ a_{21} & a_{22} \quad a_{2n} \\ \vdots & \\ a_{n1} & a_{n2} \quad a_{nn} \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Uvek u ovom poglavlju, pretpostavljamo da sistem (2.1.1), tj. (2.1.2) ima jedinstveno rešenje.

Iterativni metodi za rešavanje sistema (2.1.2) imaju za cilj određivanje rešenja $\boldsymbol{x}$ sa unapred zadatom tačnošću. Naime, polazeći od proizvoljnog vektora $\boldsymbol{x}^{(0)}$ $(= [\boldsymbol{x}_1^{(0)} \ \cdots \ \boldsymbol{x}_n^{(0)}]^T)$, iterativnim metodom se određuje niz $\{\boldsymbol{x}^{(k)}\}$, gde su $\boldsymbol{x}^{(k)} = [\boldsymbol{x}_1^{(k)} \ \cdots \ \boldsymbol{x}_n^{(k)}]^T$, $k = 1, 2, \ldots,$ tako da je

$$\lim_{k \to +\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}.$$

Jedan opšti iterativni metod se može predstaviti u obliku

$$\boldsymbol{x}^{(k)} = F_k\big(\boldsymbol{x}^{(0)},\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(k-1)}\big) \qquad (k=1,2,\dots),$$

gde funkcija F_k, u opštem slučaju, zavisi od A, $\boldsymbol{b}$, k.

Sa stanovišta primene, najinteresantniji su iterativni metodi oblika

$$\boldsymbol{x}^{(k)} = F_k(\boldsymbol{x}^{(k-1)}) \quad (k=1,2,\dots). \tag{2.1.3}$$

Ako F_k ne zavisi od k, metod (2.1.3) je stacioniran.

Mi ćemo razmatrati slučajeve kada je F_k linearna funkcija po $\boldsymbol{x}$, tj.

$$F_k(\boldsymbol{x}) = B_k\boldsymbol{x} + \boldsymbol{c}_k, \tag{2.1.4}$$

gde je B_k kvadratna matrica i $\boldsymbol{c}_k$ vektor.

Da bi iterativni metod, definisan funkcijom (2.1.4) bio konvergentan, potreban uslov je da F_k ima nepokretnu tačku $\boldsymbol{x} = A^{-1}\boldsymbol{b}$, tj. da je

$$A^{-1}\boldsymbol{b} = B_kA^{-1}\boldsymbol{b} + \boldsymbol{c}_k. \tag{2.1.5}$$

Iz (2.1.5) sleduje

$$\boldsymbol{c}_k = (I - B_k)A^{-1}\boldsymbol{b} = C_k\boldsymbol{b},$$

gde smo stavili $C_k = (I-B_k)A^{-1}$, tj. $C_kA + B_k = I$.

Na taj način (2.1.4) postaje

$$F_k(\boldsymbol{x}) = B_k\boldsymbol{x} + C_k\boldsymbol{b}, \tag{2.1.6}$$

ili

$$F_k(\boldsymbol{x}) = \boldsymbol{x} - C_k(A\boldsymbol{x} - \boldsymbol{b}). \tag{2.1.7}$$

Često se za C_k uzima dijagonalna matrica sa jednakim elementima na dijagonali, tj. $C_k = \operatorname{diag}(c_k,\dots,c_k) = c_kI$ $(c_k \in \mathbb{R})$. Tada se (2.1.7) svodi na

$$F_k(\boldsymbol{x}) = \boldsymbol{x} - c_k(A\boldsymbol{x} - \boldsymbol{b}). \tag{2.1.8}$$

Ako se matrica A predstavi u obliku

$$A = D_k + E_k,$$

gde je D_k regularna matrica, tada se F_k može zadati u implicitnom obliku kao

$$D_kF_k(\boldsymbol{x}) + E_k\boldsymbol{x} = \boldsymbol{b}. \tag{2.1.9}$$

U praktičnim primenama, za D_k se najčešće uzima dijagonalna ili trougaona matrica.

Formule (2.1.6), (2.1.7), (2.1.8), (2.1.9) se koriste za formiranje različitih iterativnih metoda za rešavanje sistema jednačina (2.1.1).

Na kraju, napomenimo da se metodom najmanjih kvadrata može dobiti niz iterativnih procesa. Ovaj metod se zasniva na minimizaciji funkcionele f definisane pomoću

$$f(\boldsymbol{x}) = \| A\boldsymbol{x} - \boldsymbol{b} \|^2 .$$

U slučaju da je A realna normalna matrica, za rešavanje sistema (2.1.1) može se koristiti minimizacija funkcionele određene sa

$$f(\boldsymbol{x}) = (A\boldsymbol{x}, \boldsymbol{x}) - 2(\boldsymbol{b}, \boldsymbol{x}).$$

4.2.2 Metod proste iteracije

Jedan od najprostijih stacionarnih metoda za rešavanje sistema linearnih jednačina, tzv. metod proste iteracije, zasnovan je na primeni funkcije date pomoću (2.1.6), tj.

$$F(\boldsymbol{x}) = B\boldsymbol{x} + C\boldsymbol{b}. \tag{2.2.1}$$

Ako stavimo $C\boldsymbol{b} = \beta$, iz (2.2.1) sleduje

$$\boldsymbol{x}^{(k)} = B\boldsymbol{x}^{(k-1)} + \boldsymbol{\beta} \quad (k = 1, 2, \dots). \tag{2.2.2}$$

Napomenimo, da je jednačina

$$\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{\beta} \tag{2.2.3}$$

ekvivalentna sa

$$A\boldsymbol{x} = \boldsymbol{b}. \tag{2.2.4}$$

Matrica B se naziva *iterativna matrica*.

Ako se pođe od proizvoljnog vektora $\boldsymbol{x}^{(0)}$, pomoću (2.2.2) generiše se niz $\{\boldsymbol{x}^{(k)}\}$. Razmatraćemo uslove pod kojima generisani niz konvergira ka tačnom rešenju sistema jednačina (2.2.3).

Ako je

$$B = \begin{bmatrix} b_{11} & b_{12} \cdots & b_{1n} \\ b_{21} & b_{22} & b_{2n} \\ \vdots & & \\ b_{n1} & b_{n2} & b_{nn} \end{bmatrix} \quad \text{i} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix},$$

iterativni metod (2.2.2) može se predstaviti skalarno

$$\begin{aligned} x_1^{(k)} &= b_{11}x_1^{(k-1)} + b_{12}x_2^{(k-1)} + \cdots + b_{1n}x_n^{(k-1)} + \beta_1, \\ x_2^{(k)} &= b_{21}x_1^{(k-1)} + b_{22}x_2^{(k-1)} + \cdots + b_{2n}x_n^{(k-1)} + \beta_2, \\ &\vdots \\ x_n^{(k)} &= b_{n1}x_1^{(k-1)} + b_{n2}x_2^{(k-1)} + \cdots + b_{nn}x_n^{(k-1)} + \beta_n, \end{aligned}$$

gde je $k = 1, 2, \ldots$.

Dokazaćemo sledeće dve teoreme.

Teorema 2.2.1. *Ako je $\boldsymbol{x}^{(0)}$ proizvoljan vektor, $\|B\| < 1$ je dovoljan uslov za konvergenciju procesa* (2.2.2) *ka tačnom rešenju $\boldsymbol{x}$ sistema* (2.2.3).

Teorema 2.2.2. *Ako je $\boldsymbol{x}^{(0)}$ proizvoljan vektor i $\|B\| < 1$ tada za svako $k \in \mathbb{N}$, važe nejednakosti*

$$\|(I-B)^{-1} - (I + B + \cdots + B^{k-1})\| \leq \frac{\|B\|^k}{1 - \|B\|} \tag{2.2.5}$$

i

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \|B\|^k \|\boldsymbol{x}^{(0)}\| + \frac{\|\boldsymbol{\beta}\| \cdot \|B\|^k}{1 - \|B\|}; \tag{2.2.6}$$

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \|B\| \cdot \|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}\|; \tag{2.2.7}$$

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \|B\|^k \|\boldsymbol{x}^{(0)} - \boldsymbol{x}\|. \tag{2.2.8}$$

Dokaz teoreme 2.2.1. Na osnovu (2.2.2), matematičkom indukcijom lako dokazujemo jednakost

$$\boldsymbol{x}^{(k)} = B^k \boldsymbol{x}^{(0)} + (I + B + \cdots + B^{k-1})\boldsymbol{\beta} \quad (k \in \mathbb{N}). \tag{2.2.9}$$

Kako je $\|B\| < 1$, imamo

$$\lim_{k\to+\infty}\|B^k\|=0, \quad \text{tj.} \quad \lim_{k\to+\infty}B^k=0,$$

i

$$\lim_{k\to+\infty}(I+B+\cdots+B^{k-1})=(I-B)^{-1}.$$

Tada iz (2.2.9) sleduje

$$\lim_{k\to+\infty}\boldsymbol{x}^{(k)}=\lim_{k\to+\infty}B^k\boldsymbol{x}^{(0)}+\lim_{k\to+\infty}(I+B+\cdots+B^{k-1})\boldsymbol{\beta}=(I-B)^{-1}\boldsymbol{\beta},$$

odakle zaključujemo da niz $\{\boldsymbol{x}^{(k)}\}$ konvergira ka rešenju jednačine (2.2.3). □

Dokaz teoreme 2.2.2. Iz jednakosti

$$(I-B)^{-1}=I+B+B^2+\cdots \quad (\|B\|<1)$$

sleduje

$$(I-B)^{-1}-(I+B+\cdots+B^{k-1})=B^k+B^{k+1}+\cdots,$$

tj.

$$\|(I-B)^{-1}-(I+B+\cdots+B^{k-1})\|\le\|B\|^k+\|B\|^{k+1}+\cdots=\frac{\|B\|^k}{1-\|B\|}.$$

Na osnovu (2.2.9), imamo

$$\begin{aligned}\boldsymbol{x}^{(k)}-\boldsymbol{x}&=B^k\boldsymbol{x}^{(0)}+(I+B+\cdots+B^{k-1})\boldsymbol{\beta}-(I-B)^{-1}\boldsymbol{\beta}\\&=B^k\boldsymbol{x}^{(0)}-[(I-B)^{-1}-(I+B+\cdots+B^{k-1})]\boldsymbol{\beta},\end{aligned}$$

odakle je

$$\begin{aligned}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}\|&\le\|B^k\boldsymbol{x}^{(0)}\|+\|(I-B)^{-1}-(I+B+\cdots+B^{k-1})\|\cdot\|\boldsymbol{\beta}\|\\&\le\|B\|^k\|\boldsymbol{x}^{(0)}\|+\frac{\|\boldsymbol{\beta}\|\cdot\|B\|^k}{1-\|B\|},\end{aligned}$$

pri čemu smo iskoristili nejednakost (2.2.5).

Napomenimo da je korišćena norma matrice saglasna sa izabranom normom vektora.

Kako je $\boldsymbol{x}=B\boldsymbol{x}+\boldsymbol{\beta}$ i $\boldsymbol{x}^{(k)}=B\boldsymbol{x}^{(k-1)}+\boldsymbol{\beta}$, dobijamo

$$\boldsymbol{x}^{(k)}-\boldsymbol{x}=B(\boldsymbol{x}^{(k-1)}-\boldsymbol{x}),$$

tj.

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \|B\| \cdot \|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}\|.$$

Najzad, iteriranjem poslednje nejednakosti, dobijamo

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \|B\|^k \cdot \|\boldsymbol{x}^{(0)} - \boldsymbol{x}\|.$$

Ovim je teorema 2.2.2 dokazana. □

Nejednakosti (2.2.7) i (2.2.8) ukazuju da iterativni proces (2.2.2) ima red konvergencije jedan (linearna konvergencija ili konvergencija tipa geometrijske progresije).

Napomena 2.2.1. Najčešće se uzima $\boldsymbol{x}^{(0)} = \boldsymbol{\beta}$. Tada se nejednakost (2.2.6), iz teoreme 2.2.1, može pooštriti. Naime, važi nejednakost

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| = \|((I-B)^{-1} - (I + B + \cdots + B^{k-1}))\boldsymbol{\beta}\| \leq \frac{\|B\|^{k+1}\|\boldsymbol{\beta}\|}{1 - \|B\|}.$$

Posledica 2.2.1. *Iterativni proces* (2.2.2) *konvergira ako je ispunjen bilo koji od uslova*

$1^\circ \quad \|B\|_\infty = \max_i \sum_{j=1}^{n} |b_{ij}| < 1;$

$2^\circ \quad \|B\|_1 = \max_j \sum_{i=1}^{n} |b_{ij}| < 1;$

$3^\circ \quad \|B\|_2 = \varepsilon(B) = \left(\sum_{i,j} |b_{ij}|^2 \right)^{1/2} < 1.$

Napomena 2.2.2. Ako je $\|B\| < 1$, važi

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \frac{\|B\|}{1 - \|B\|} \, \| \, \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \, \| \quad (k \in \mathbb{N}).$$

Ova nejednakost je posledica jedne opštije nejednakosti, koja će biti dokazana u sledećem odeljku.

Do sada smo razmatrali dovoljne uslove za konvergenciju iterativnog procesa (2.2.2). Sledeća teorema daje potrebne i dovoljne uslove.

Teorema 2.2.3. *Neka je $\boldsymbol{x}^{(0)}$ proizvoljan vektor. Potreban i dovoljan uslov za konvergenciju iterativnog procesa* (2.2.2) *je da su sve sopstvene vrednosti matrice B po modulu manje od jedinice.*

Dokaz. Kako za iterativni proces (2.2.2) važi jednakost (2.2.9), tj.

$$\boldsymbol{x}^{(k)} = B^k\boldsymbol{x}^{(0)} + (I + B + \cdots + B^{k-1})\boldsymbol{\beta} \quad (k = 1,2,\dots), \tag{2.2.10}$$

zaključujemo da je proces (2.2.2) ekvikonvergentan sa matričnim redom

$$I + B + B^2 + \cdots = \sum_{m=0}^{+\infty} B^m. \tag{2.2.11}$$

S druge strane, kako su potrebni i dovoljni uslovi za konvergenciju reda (2.2.11) (videti odeljak 2.3.7)

$$|\lambda_i(B)| < 1 \quad (i = 1,\dots,n), \tag{2.2.12}$$

dokaz teoreme 2.2.3 je završen. □

Napomena 2.2.3. Pod uslovima (2.2.12) imamo

$$\lim_{k\to+\infty} B^k = 0 \quad \text{i} \quad \lim_{k\to+\infty} (I + B + \cdots + B^{k-1}) = (I - B)^{-1}.$$

Tada iz (2.2.10) sleduje

$$\begin{aligned}\lim_{k\to+\infty} \boldsymbol{x}^{(k)} &= \lim_{k\to+\infty} [B^k\boldsymbol{x}^{(0)} + (I + B + \cdots + B^{k-1})\boldsymbol{\beta}] \\ &= \lim_{k\to+\infty} B^k\boldsymbol{x}^{(0)} + \lim_{k\to+\infty} (I + B + \cdots + B^{k-1})\boldsymbol{\beta} \\ &= (I - B)^{-1}\boldsymbol{\beta},\end{aligned}$$

što predstavlja tačno rešenje jednačine (2.2.3).

Dakle, iz teoreme 2.2.3 sleduje da iterativni proces (2.2.2) konvergira ako i samo ako su sve nule polinoma

$$\lambda \mapsto \det(B - \lambda I) = \begin{vmatrix} b_{11} - \lambda & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} - \lambda & & b_{2n} \\ \vdots & & & \\ b_{n1} & b_{n2} & & b_{nn} - \lambda \end{vmatrix}$$

po modulu manje od jedinice.

Značaj uslova (2.2.12) u teorijskim razmatranjima je vrlo veliki. Međutim, za praktičnu primenu oni nisu pogodni, s obzirom da je problem nalaženja sopstvenih vrednosti matrice dosta težak. Uslov $\|B\| < 1$, u teoremi 2.2.1, je sa stanovišta praktične primene vrlo pogodan za ispitivanje konvergencije. Nažalost, ovaj uslov je samo dovoljan, ali ne i potreban.

Primer 2.2.1. Posmatrajmo metod proste iteracije

$$\begin{aligned} x^{(k)} &= \frac{1}{3}x^{(k-1)} - \frac{1}{9}y^{(k-1)} + \frac{1}{9}, \\ y^{(k)} &= 2x^{(k-1)} - \frac{1}{3}y^{(k-1)} + \frac{4}{3}, \end{aligned} \tag{2.2.13}$$

gde je $k = 1, 2, \ldots$.

Na osnovu posledice 2.2.1, ništa ne možemo zaključiti o konvergenciji procesa (2.2.13), s obzirom na to da nijedan od uslova 1°, 2°, 3° za matricu

$$B = \begin{bmatrix} 1/3 & -1/9 \\ 2 & 1/3 \end{bmatrix}$$

nije ispunjen. Naime,

$$\begin{aligned} \|B\|_\infty &= \max\left\{\frac{1}{3}+\frac{1}{9}, 2+\frac{1}{3}\right\} = \frac{7}{3} > 1, \\ \|B\|_1 &= \max\left\{\frac{1}{3}+2, \frac{1}{9}+\frac{1}{3}\right\} = \frac{7}{3} > 1, \\ \|B\|_2 &= \left(\frac{1}{9}+\frac{1}{81}+4+\frac{1}{9}\right)^{1/2} = 2.057 > 1. \end{aligned}$$

Međutim, sopstvene vrednosti matrice B su $\lambda_{1,2} = (1 \pm i\sqrt{2})/3$. Kako je $|\lambda_1| = |\lambda_2| = \sqrt{3}/3 < 1$, iterativni proces (2.2.13) je konvergentan za proizvoljne vrednosti $x^{(0)}$ i $y^{(0)}$. $\triangle$

Uslovi (2.2.12) se mogu zameniti uslovom

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

gde je $\rho(B)$ spektralni radijus matrice B. Za iterativnu matricu B se, u ovom slučaju, kaže da je konvergentna (videti [40]).

Kao kriterijum za brzinu konvergencije iterativnog procesa uzima se veličina $\rho(B)$. Naime, iterativni proces konvergira brže, ukoliko je $\rho(B)$ bliže nuli. Broj koji karakteriše *brzinu konvergencije* definisan je kao

$$v(B) = -\log \rho(B).$$

Pri ovome, za ispunjenje uslova $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| < \varepsilon \|\boldsymbol{x}^{(0)} - \boldsymbol{x}\|$, gde je ε dovoljno mali pozitivan broj i $0 < \rho(B) < 1$, potreban broj iteracija je približno dat sa

$$k \cong -\frac{\log \varepsilon}{v(B)}.$$

Pored brzine konvergencije iterativnog procesa vrlo je bitan i broj aritmetičkih i logičkih operacija neophodnih za obavljanje jednog iterativnog koraka (iteracije). Ovaj broj često se naziva *cena iteracije* i označava se sa $C(B)$, gde je B odgovarajuća iterativna matrica.

Tako je ukupan broj operacija za postizanje tačnosti (u prethodno navedenom smislu) približno dat sa

$$N(B, \varepsilon) = kC(B) \cong -\frac{C(B) \log \varepsilon}{v(B)}.$$

Iterativni proces je efikasniji, ukoliko je ovaj broj manji.

Pokazaćemo sada kako se može dobiti jedan tzv. *optimalni iterativni proces* za rešavanje jednačine $A\boldsymbol{x} = \boldsymbol{b}$ u slučaju kada je A hermitska pozitivno definitna matrica. Ovaj metod je baziran na primeni formule (2.1.8).

Ovde je

$$0 < m \leq \lambda_1(A) \leq M < +\infty \quad (i = 1, \ldots, n). \tag{2.2.14}$$

S obzirom na to da se sopstvene vrednosti $\lambda_1(A)$ obično ne znaju to se za m i M u (2.2.14) mogu uzeti neke ocene za $\lambda_{\min}(A)$ i $\lambda_{\max}(A)$, respektivno.

Posmatrajmo iterativni proces

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - c(A\boldsymbol{x}^{(k-1)} - \boldsymbol{b}), \tag{2.2.15}$$

tj.

$$\boldsymbol{x}^{(k)} = (I - cA)\boldsymbol{x}^{(k-1)} + c\boldsymbol{b},$$

gde je c realan broj.

Kako su sopstvene vrednosti matrice $B = I - cA$,

$$\lambda_i(B) = 1 - c\lambda_i(A),$$

spektralni radijus matrice B je

$$\rho(B) = \max |1 - c\lambda_i(A)|. \tag{2.2.16}$$

Za iterativni proces (2.2.15) kaže se da je *optimalan*, ako se parametar c odredi tako da $\rho(B)$ ima minimalnu vrednost.

S obzirom na (2.2.14), ako se uvede smena

$$\mu = \frac{2}{M-m}\lambda_i(A) - \frac{M+m}{M-m},$$

(2.2.16) se svodi na

$$\rho(B) = \frac{1}{2}|c|(M-m) \max_{-1\le\mu\le 1} \left| \mu - \frac{2-c(M+m)}{c(M-m)} \right|.$$

Minimalna vrednost za $\rho(B)$ se dobija ako je $c = 2/(M+m)$ i ona iznosi

$$\min_c \rho(B) = \frac{M-m}{M+m}.$$

Dakle, dobili smo optimalni iterativni proces

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - \frac{2}{M+m}(A\boldsymbol{x}^{(k-1)} - \boldsymbol{b}). \tag{2.2.17}$$

Teorema 2.2.4. *Neka je A hermitska pozitivno definitna matrica čije sopstvene vrednosti zadovoljavaju uslov* (2.2.14). *Za optimalni iterativni proces* (2.2.17) *važi ocena greške*

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|_E \le \left(\frac{M-m}{M+m}\right)^k \|\boldsymbol{x}^{(0)} - \boldsymbol{x}\|_E \quad (k \in \mathbb{N}), \tag{2.2.18}$$

gde je $\boldsymbol{x}$ tačno rešenje sistema jednačina $A\boldsymbol{x} = \boldsymbol{b}$.

Dokaz. S obzirom na pretpostavke za matricu A, matrica $B = I - 2A/(M+m)$ je hermitska. Tada je, na osnovu teoreme 3.6.2,

$$\sigma(B) = \rho(B) = \frac{M-m}{M+m}.$$

Uzimajući euklidsku normu za normu vektora i spektralnu normu za normu matrica, nejednakost (2.2.8), iz teoreme 2.2.2, se svodi na (2.2.18). □

O nekim opštijim iterativnim procesima tipa (2.2.17) može se naći u [9, str. 352–363].

Na kraju ovog odeljka, izložićemo jedan praktičan način za formiranje metoda proste iteracije.

Neka je dat sistem jednačina $A\boldsymbol{x} = \boldsymbol{b}$, gde su $A = [a_{ij}]_{n\times n}$ i $\boldsymbol{b} = [b_1 \ \cdots \ b_n]^T$. Neka je

$$D = \operatorname{diag}(A) = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & a_{nn} \end{bmatrix}$$

regularna matrica. Tada se ovaj sistem može predstaviti u ekvivalentnom obliku

$$\boldsymbol{x} = D^{-1}(D-A)\boldsymbol{x} + D^{-1}\boldsymbol{b}.$$

Napomenimo da je odgovarajući skalarni oblik

$$\begin{aligned} x_1 &= -\frac{a_{12}}{a_{11}}x_2 - \frac{a_{13}}{a_{11}}x_3 - \cdots - \frac{a_{1n}}{a_{11}}x_n + \frac{b_1}{a_{11}}, \\ x_2 &= -\frac{a_{21}}{a_{22}}x_1 - \frac{a_{23}}{a_{22}}x_3 - \cdots - \frac{a_{2n}}{a_{22}}x_n + \frac{b_2}{a_{22}}, \\ &\vdots \\ x_n &= -\frac{a_{n1}}{a_{nn}}x_1 - \frac{a_{n2}}{a_{nn}}x_2 - \cdots - \frac{a_{n,n-1}}{a_{nn}}x_{n-1} + \frac{b_n}{a_{nn}}. \end{aligned}$$

Na osnovu prethodnog, može se formirati metod proste iteracije

$$\boldsymbol{x}^{(k)} = D^{-1}(D-A)\boldsymbol{x}^{(k-1)} + D^{-1}\boldsymbol{b} \quad (k = 1,2,\dots). \tag{2.2.19}$$

koji je u literaturi poznat kao JACOBI*ev metod*.

Kako je karakteristični polinom matrice $D^{-1}(D-A)$ dat sa

$$P(\lambda) = \det[D^{-1}(D-A) - \lambda I] = -\det(D^{-1})\det[\lambda D + (A-D)],$$

iz teoreme 2.2.3 sleduje da JACOBIev iterativni proces konvergira ako i samo ako su svi koreni jednačine

$$\begin{vmatrix} \lambda a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & \lambda a_{22} & & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & & \lambda a_{nn} \end{vmatrix} = 0$$

po modulu manji od jedinice.

4.2.3 GAUSS–SEIDELov metod

GAUSS–SEIDELov[137] metod se dobija modifikacijom metoda proste iteracije. Kao što smo ranije videli, kod metoda proste iteracije, vrednost i-te komponente $x_i^{(k)}$ vektora $\boldsymbol{x}^{(k)}$ izračunava se na osnovu vrednosti $x_1^{(k-1)}, \ldots, x_n^{(k-1)}$, tj.

$$x_i^{(k)} = \sum_{j=1}^{n} b_{ij} x_j^{(k-1)} + \beta_i \qquad (i = 1, \ldots, n;\ k = 1, 2, \ldots).$$

Ovaj metod se može modifikovati na taj način što bi se za izračunavanje vrednosti $x_i^{(k)}$ koristile vrednosti $x_i^{(k)}, \ldots, x_{i-1}^{(k)}, x_i^{(k-1)}, \ldots, x_n^{(k-1)}$, tj.

$$x_i^{(k)} = \sum_{j=1}^{i-1} b_{ij} x_j^{(k)} + \sum_{j=i}^{n} b_{ij} x_j^{(k-1)} + \beta_i \qquad (i = 1, \ldots, n;\ k = 1, 2, \ldots). \tag{2.3.1}$$

Ova modifikacija metoda proste iteracije poznata je kao GAUSS–SEIDELov metod.

Iterativni proces (2.3.1) se može predstaviti i u matričnoj formi. Naime, neka je matrica B predstavljena kao zbir dve matrice

$$B = B_1 + B_2,$$

gde su

$$B_1 = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ b_{21} & 0 & & 0 & 0 \\ \vdots & & & & \\ b_{n1} & b_{n2} & & b_{n,n-1} & 0 \end{bmatrix} \quad \text{i} \quad B_2 = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ 0 & b_{22} & & b_{2n} \\ \vdots & & & \\ 0 & 0 & & b_{nn} \end{bmatrix}$$

Tada (2.3.1) postaje

$$\boldsymbol{x}^{(k)} = B_1 \boldsymbol{x}^{(k)} + B_2 \boldsymbol{x}^{(k-1)} + \boldsymbol{\beta} \quad (k = 1, 2, \ldots). \tag{2.3.2}$$

[137] PHILIPP LUDWIG VON SEIDEL (1821 – 1896), nemački matematičar.

Teorema 2.3.1. *Pri proizvoljnom vektoru $\boldsymbol{x}^{(0)}$, iterativni proces* (2.3.2) *konvergira ako i samo ako su svi koreni jednačine*

$$\det\left[B_2-(I-B_1)\lambda\right]\equiv\begin{vmatrix} b_{11}-\lambda & b_{12} & \cdots & b_{1n} \\ b_{21}\lambda & b_{22}-\lambda & & b_{2n} \\ \vdots & & & \\ b_{n1}\lambda & b_{n2}\lambda & & b_{nn}-\lambda \end{vmatrix}=0 \tag{2.3.3}$$

po modulu manji od jedinice.

Dokaz. Kako je $\det(I-B_1)=1$, tj. matrica $I-B_1$ regularna, za (2.3.2) može se dobiti ekvivalentan metod proste iteracije. Naime, iz (2.3.2) sledi

$$(I-B_1)\boldsymbol{x}^{(k)}=B_2\boldsymbol{x}^{(k-1)}+\boldsymbol{\beta}\quad(k=1,2,\dots),$$

tj.

$$\boldsymbol{x}^{(k)}=(I-B_1)^{-1}B_2\boldsymbol{x}^{(k-1)}+(I-B_1)^{-1}\boldsymbol{\beta}\quad(k=1,2,\dots)$$

Na osnovu teoreme 2.2.3, ovaj iterativni proces konvergira, pri proizvoljnom vektoru $\boldsymbol{x}^{(0)}$, ako i samo ako su svi koreni jednačine $\det\left[(I-B)^{-1}B_2-\lambda I\right]=0$ po modulu manji od jedinice.

Iz poslednje jednačine sledi

$$\det\left[(I-B_1)^{-1}(B_2-(I-B_1)\lambda)\right]=0,$$

tj.

$$\det(I-B_1)^{-1}\det\left[(B_2-(I-B_1)\lambda)\right]=0.$$

Kako je $\det(I-B_1)^{-1}=1$, poslednja jednačina se svodi na jednačinu (2.3.3), čime je dokazana teorema 2.3.1. □

Posmatrajmo sada sistem jednačina $A\boldsymbol{x}=\boldsymbol{b}$ u obliku (2.2.19). Ako stavimo

$$A-D=C_1+C_2,$$

gde su

$$C_1=\begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & & 0 & 0 \\ \vdots & & & & \\ a_{n1} & a_{n2} & & a_{n,n-1} & 0 \end{bmatrix}\quad\text{i}\quad C_2=\begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & 0 & a_{23} & & a_{2n} \\ \vdots & & & & \\ 0 & 0 & & & 0 \end{bmatrix},$$

može se konstruisati GAUSS–SEIDELov proces kao

$$\boldsymbol{x}^{(k)} = -D^{-1}C_1\boldsymbol{x}^{(k)} - D^{-1}C_2\boldsymbol{x}^{(k-1)} + D^{-1}\boldsymbol{b} \quad (k = 1,2,\dots). \tag{2.3.4}$$

Ova varijanta GAUSS–SEIDELovog metoda ponekad se naziva metod NEKRASOVa[138] (videti [59]).

Iz teoreme 2.3.1 sleduje sledeća teorema.

Teorema 2.3.2. *Pri proizvoljnom vektoru $\boldsymbol{x}^{(0)}$, iterativni proces* (2.3.4) *konvergira ako i samo ako su svi koreni jednačine*

$$\det\left[C_2 + (D + C_1)\lambda\right] = \begin{vmatrix} a_{11}\lambda & a_{12} & \cdots & a_{1n} \\ a_{21}\lambda & a_{22}\lambda & & a_{2n} \\ \vdots & & & \\ a_{n1}\lambda & a_{n2}\lambda & & a_{nn}\lambda \end{vmatrix} = 0$$

po modulu manji od jedinice.

Kao što je i ranije napomenuto, ovi spektralni uslovi za konvergenciju iterativnih procesa, nažalost, nemaju veliki praktični značaj.

Za sistem jednačina sa simetričnom matricom važi sledeći rezultat ([61]).

Teorema 2.3.3. *Neka je matrica A realna i simetrična i neka su joj svi dijagonalni elementi pozitivni. Iterativni proces* (2.3.4) *konvergira ako i samo ako su sve veličine*

$$a_{11}, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix}, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{vmatrix}, \dots, \quad \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & & a_{2n} \\ \vdots & & & \\ a_{1n} & a_{2n} & & a_{nn} \end{vmatrix}$$

pozitivne.

L. COLLATZ[139] ([15], [16]) dokazao je sledeći rezultat.

[138] PAVEL ALEKSEEVICH NEKRASOV (1858 – 1924), ruski matematičar, poznat posebno po radovima iz teorije verovatnoće.

[139] LOTHAR COLLATZ (1910 –1 990), poznati nemački matematičar.

Teorema 2.3.4. JACOBI*ev iterativni proces* (2.2.19) *i* GAUSS–SEIDEL*ov proces* (2.3.4) *konvergiraju, ako matrica A reda n ispunjava sledeća dva uslova:*

1° *matrica A ne sadrži nula-submatricu tipa* $m \times (n-m) \quad (1 \leq m \leq n-1)$;

2° *za svako* $i \in I = \{1, \ldots, n\}$ *su ispunjeni uslovi* $|a_{ii}| \geq s_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$, *i bar za jedno* $i \in I$ *je* $|a_{ii}| > s_i$.

Primer 2.3.1. Ispitaćemo primenljivost iterativnog procesa (2.3.4) na rešavanje sistema jednačina

$$\begin{aligned} 10x_1 + 3x_2 - \quad x_3 &= 12, \\ -x_1 + 5x_2 - \quad x_3 &= \ 3, \\ x_1 + 2x_2 + 10x_3 &= 13. \end{aligned} \tag{2.3.5}$$

Kako za elemente matrice datog sistema

$$A = \begin{bmatrix} 10 & 3 & -1 \\ -1 & 5 & -1 \\ 1 & 2 & 10 \end{bmatrix}$$

važe nejednakosti

$$\begin{aligned} |a_{11}| &= 10 > s_1 = |a_{12}| + |a_{13}| = 4, \\ |a_{22}| &= \ 5 > s_2 = |a_{21}| + |a_{23}| = 2, \\ |a_{33}| &= 10 > s_3 = |a_{31}| + |a_{32}| = 3, \end{aligned}$$

i kako A ne sadrži nula-submatricu tipa 1×2 ili tipa 2×1, zaključujemo da su uslovi 1° i 2° u teoremi 2.3.4 ispunjeni. Dakle, iterativni proces (2.3.4) primenjen na rešavanje sistema jednačina (2.3.5) konvergira.

Polazeći od $\boldsymbol{x}^{(0)} = \boldsymbol{\beta} = [1.2 \quad 0.6 \quad 1.3]^T$, na osnovu

$$\begin{aligned} x_1^{(k)} &= -0.3\, x_2^{(k-1)} + 0.1\, x_3^{(k-1)} + 1.2, \\ x_2^{(k)} &= \ \ 0.2\, x_1^{(k)} \quad + 0.2\, x_3^{(k-1)} + 0.6, \\ x_3^{(k)} &= -0.1\, x_1^{(k)} \quad - 0.2\, x_2^{(k)} \quad + 1.3, \end{aligned}$$

gde je $k = 1, 2, \ldots$, dobijamo niz iteracija

$$\boldsymbol{x}^{(1)} = [1.1500000 \quad 1.0900000 \quad 0.9670000]^T,$$
$$\boldsymbol{x}^{(2)} = [0.9697000 \quad 0.9873400 \quad 1.0055620]^T,$$
$$\boldsymbol{x}^{(3)} = [1.0043542 \quad 1.0019832 \quad 0.9991680]^T,$$
$$\boldsymbol{x}^{(4)} = [0.9993219 \quad 0.9996979 \quad 1.0001283]^T,$$

itd. Primetimo da je tačno rešenje sistema (2.3.5) dato sa $\boldsymbol{x} = [1 \quad 1 \quad 1]^T$. $\triangle$

Još jedno interesantno pitanje koje se može postaviti u vezi sa razmatranim iterativnim metodima je da li GAUSS–SEIDELov metod uvek konvergira kada konvergira odgovarajući metod proste iteracije? Odgovor na ovo pitanje nije potvrdan. Naime, u dosta velikom broju slučajeva, ako metod proste iteracije konvergira, konvergiraće i GAUSS–SEIDELov metod i to brže, međutim, postoje slučajevi kada ovaj poslednji ne konvergira.

Štaviše, postoje i slučajevi kada GAUSS–SEIDELov metod konvergira, a metod proste iteracije divergira. Sledeći prost primer ovo lepo ilustruje.

Primer 2.3.2. Neka je $\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{\beta}$, gde su

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad B = \begin{bmatrix} p & q \\ -q & p \end{bmatrix} \quad (p, q \in \mathbb{R}).$$

Odredićemo parametre p i q tako da konvergira

1° metod proste iteracije;

2° GAUSS–SEIDELov metod.

Iz uslova

$$\begin{vmatrix} p-\lambda & q \\ -q & p-\lambda \end{vmatrix} = 0$$

sleduje $\lambda_{1,2} = p \pm \mathrm{i}q$, odakle, za uslov konvergencije metoda proste iteracije, dobijamo

$$|\lambda_1| = |\lambda_2| = \sqrt{p^2+q^2} < 1,$$

tj.

$$p^2 + q^2 < 1.$$

Odgovarajuća karakteristična jednačina za GAUSS–SEIDELov metod je

$$\begin{vmatrix} p-\lambda & q \\ -q\lambda & p-\lambda \end{vmatrix} = 0,$$

tj.

$$\lambda^2 - (2p - q^2)\lambda + p^2 = 0.$$

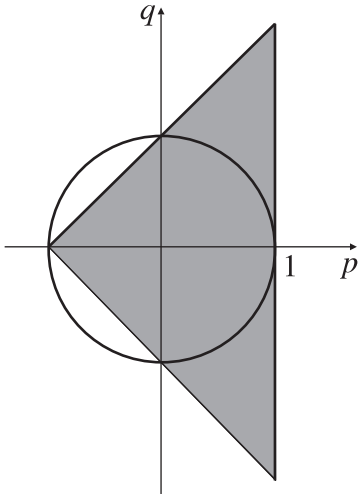
Potrebni i dovoljni uslovi da poslednja jednačina ima korene po modulu manje od jedinice su (videti [19], [77])

$$|2p - q^2| < p^2 + 1 \quad \text{i} \quad |p^2| < 1,$$

odakle, nakon elementarnih algebarskih transformacija dobijamo

$$|q| < 1 + p \quad \text{i} \quad |p| < 1. \tag{2.3.6}$$

U ravni $p0q$, nejednakosti (2.3.6) definišu oblast prikazanu šrafirano na slici 2.3.1. Ako (p,q) pripada ovoj oblasti, GAUSS–SEIDELov metod konvergira. S druge strane, metod proste iteracije konvergira, kao što je dokazano, ako se (p,q) nalazi u unutrašnjosti jediničnog kruga. $\triangle$



Slika 2.3.1. Oblast konvergencije iterativnih procesa u primeru 2.3.2

Pređimo sada na određivanje greške približnog rešenja koje se dobija primenom iterativnog procesa (2.3.2). Kako je tačno rešenje $\boldsymbol{x}$ nepoznato, greška $\boldsymbol{\varepsilon}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}$, se ne može tačno odrediti. Međutim, kako se greška $\boldsymbol{\varepsilon}^{(k)}$ može izraziti pomoću priraštaja $\boldsymbol{\delta}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}$, moguće je dati ocenu za $\|\boldsymbol{\varepsilon}^{(k)}\|$.

Teorema 2.3.5. *Ako je* $\boldsymbol{x}$ *rešenje jednačine*

$$\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{\beta} \quad (B = B_1 + B_2)$$

i ako je $\|B\| \leq q < 1$*, važi nejednakost*

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \frac{\|B_2\|}{1 - \|B\|} \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\| \quad (k \in \mathbb{N}), \tag{2.3.7}$$

gde se niz $\{\boldsymbol{x}^{(k)}\}$ *generiše pomoću* (2.3.2).

Dokaz. Na osnovu (2.3.2), za svako $k \in \mathbb{N}$, imamo

$$\boldsymbol{\varepsilon}^{(k)} = (B_1\boldsymbol{x}^{(k)} + B_2\boldsymbol{x}^{(k-1)} + \boldsymbol{\beta}) - (B_1\boldsymbol{x} + B_2\boldsymbol{x} + \boldsymbol{\beta}) = B_1\boldsymbol{\varepsilon}^{(k)} + B_2\boldsymbol{\varepsilon}^{(k-1)}.$$

Kako je $\boldsymbol{\varepsilon}^{(k-1)} = \boldsymbol{x}^{(k-1)} - \boldsymbol{x} = \boldsymbol{\varepsilon}^{(k)} - \boldsymbol{\delta}^{(k)}$, iz prethodne jednakosti sleduje

$$\boldsymbol{\varepsilon}^{(k)} = B_1\boldsymbol{\varepsilon}^{(k)} + B_2(\boldsymbol{\varepsilon}^{(k)} - \boldsymbol{\delta}^{(k)}) \quad (k \in \mathbb{N}),$$

tj.

$$\boldsymbol{\varepsilon}^{(k)} = -(I - B)^{-1}B_2\boldsymbol{\delta}^{(k)} \quad (k \in \mathbb{N}),$$

s obzirom da egzistira $(I - B)^{-1}$, na osnovu pretpostavke $\|B\| \leq q < 1$.

Ako koristimo normu matrice saglasnu sa normom vektora, iz poslednje jednakosti dobijamo

$$\|\boldsymbol{\varepsilon}^{(k)}\| \leq \|(I - B)^{-1}B_2\| \cdot \|\boldsymbol{\delta}^{(k)}\| \leq \frac{\|B_2\|}{1 - \|B\|} \|\boldsymbol{\delta}^{(k)}\| \quad (k \in \mathbb{N}),$$

čime smo dokazali teoremu 2.3.5. □

Napomena 2.3.1. Ako je $B_1 = 0$ ($\Rightarrow B_2 = B$), (2.3.7) se svodi na nejednakost datu u napomeni 2.2.2.

4.2.4 Opšte napomene o relaksacionim metodima

Neka je $\boldsymbol{x}_p$ približno rešenje sistema jednačina

$$A\boldsymbol{x} = \boldsymbol{b}. \tag{2.4.1}$$

Kako je tačno rešenje $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ nepoznato, postavlja se pitanje u kojoj meri $\boldsymbol{x}_p$ zadovoljava dati sistem jednačina. Kao mera ovoga, najčešće se koristi norma vektora ostatka

$$\boldsymbol{r} = A\boldsymbol{x}_p - \boldsymbol{b}. \tag{2.4.2}$$

Ako je $\boldsymbol{r}$ nula-vektor, lako se uočava da je $\boldsymbol{x}_p$ tačno rešenje sistema (2.4.1). Na osnovu prethodnog, može se zaključiti da je sistem (2.4.1) skoro zadovoljen ako su komponente vektora $\boldsymbol{r}$ bliske nuli. Poslednje tvrđenje, međutim, nije uvek u važnosti. Naime, kod slabo uslovljenih sistema, norma vektora $\boldsymbol{x}_p - A^{-1}\boldsymbol{b}$ može biti velika, i u slučajevima kada je norma vektora ostatka mala, s obzirom da je $\boldsymbol{x}_p - A^{-1}\boldsymbol{b} = A^{-1}\boldsymbol{r}$.

I pored ovog nedostatka, vektor $\boldsymbol{r}$ igra važnu ulogu u širokoj klasi iterativnih metoda, tzv. *relaksacionih metoda*.

Pod relaksacionim metodom podrazumeva se svaki metod kod koga se sledeća aproksimacija rešenja dobija na osnovu prethodne aproksimacije i vektora ostatka (u opštem slučaju zavisi od više prethodnih aproksimacija), koji se koristi kao indikator za veličinu korekcije.

Prve ideje o relaksacionim modelima potiču još od GAUSSa. Međutim, sistematska teorijska istraživanja na ovim metodima datiraju iz poslednjih pedestak godina. S obzirom da su ovi metodi pogodni za rešavanje sistema sa velikim brojem jednačina, to oni sve više nalaze primenu kod rešavanja parcijalnih jednačina. O relaksacionim metodima danas postoji vrlo obimna literatura (videti posebno [2], [8], [36], [65], [72], [76]).

U sledećim odeljcima obradićemo nekoliko klasa relaksacionih metoda.

4.2.5 Metod sukcesivne gornje relaksacije

U ovom odeljku razmotrićemo jedan relaksacioni metod, koji predstavlja generalizaciju jedne varijante GAUSS–SEIDELovog metoda za rešavanje sistema $A\boldsymbol{x} = \boldsymbol{b}$.

Neka je matrica $D = \operatorname{diag} A$ regularna. Razložićemo matricu A u obliku

$$A = C_1 + D + C_2,$$

gde su

$$C_1 = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & & 0 & 0 \\ \vdots & & & & \\ a_{n1} & a_{n2} & & a_{n,n-1} & 0 \end{bmatrix} \quad \text{i} \quad C_2 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & & a_{2n} \\ \vdots & & & \\ 0 & 0 & & a_{nn} \end{bmatrix}.$$

Ako za izračunavanje vektora ostatka $\boldsymbol{r}$ uzmemo za x_i poslednju izračunatu vrednost (kao kod GAUSS–SEIDELovog metoda), tj.

$$\boldsymbol{r} = C_1\boldsymbol{x}^{(k)} + (D+C_2)\boldsymbol{x}^{(k-1)} - \boldsymbol{b}$$

i ako stavimo $\boldsymbol{\delta}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}$, možemo formirati iterativni proces u obliku

$$B\boldsymbol{\delta}^{(k)} = -\omega\boldsymbol{r} \quad (k = 1,2,\dots),$$

tj.

$$D\boldsymbol{x}^{(k)} = D\boldsymbol{x}^{(k-1)} + \omega[\boldsymbol{b} - C_1\boldsymbol{x}^{(k)} - (D+C_2)\boldsymbol{x}^{(k-1)}], \tag{2.5.1}$$

gde je $k = 1,2,\dots$ i ω realan parametar. Parametar ω nazivamo *relaksacioni množilac* i njega u opštem slučaju možemo menjati u toku računanja.

Iterativni proces (2.5.1) može se predstaviti i u skalarnom obliku

$$a_{ii}x_i^{*(k)} = -\sum_{j<i} a_{ij}x_j^{(k)} - \sum_{j>i} a_{ij}x_j^{(k-1)} + b_i,$$

$$x_i^{(k)} = x_i^{(k-1)} + \omega\left(x_i^{*(k)} - x_i^{(k-1)}\right)$$

gde su $i = 1,\dots,n$; $k = 1,2,\dots$.

Primetimo da se za $\omega = 1$, (2.5.1) svodi na GAUSS–SEIDELov proces (2.3.7).

Na osnovu (2.5.1) imamo

$$(D+\omega C_1)\boldsymbol{x}^{(k)} = [(1-\omega)D - \omega\zeta]\boldsymbol{x}^{(k-1)} + \omega\boldsymbol{b} \quad (k = 1,2,\dots), \tag{2.5.2}$$

tj.

$$\boldsymbol{x}^{(k)} = K(\omega)\boldsymbol{x}^{(k-1)} + \boldsymbol{f}(\omega) \quad (k = 1,2,\dots), \tag{2.5.3}$$

gde su

$$K(\omega) = (D+\omega C_1)^{-1}[(1-\omega)D - \omega\zeta] \quad \text{i} \quad \boldsymbol{f}(\omega) = \omega(D+\omega C_1)^{-1}\boldsymbol{b}.$$

Teorema 2.5.1. *Iterativni proces* (2.5.1) *konvergira ako i samo ako su sve sopstvene vrednosti matrice* $K(\omega)$ *po modulu manje od jedinice.*

Dokaz. Kako je relaksacioni metod (2.5.1) ekvivalentan sa metodom proste iteracije (2.5.3), teorema 2.5.1 se dobija kao posledica teoreme 2.2.3. □

Za jednu specijalnu klasu sistema, koja se vrlo često javlja u praksi, važi sledeći kriterijum ([16]).

Teorema 2.5.2. *Ako je A hermitska pozitivno definitna matrica, relaksacioni iterativni proces* (2.5.1) *konvergira ka tačnom rešenju sistema $A\boldsymbol{x} = \boldsymbol{b}$ kada je* $0 < \omega < 2$.

Dokaz. Na osnovu učinjenih pretpostavki za matricu A $(= C_1 + D + C_2)$ imamo da je $a_{11} > 0$ $(i = 1, \ldots, n)$ i $C_1 = C_2^*$. Kako je, dalje,

$$(C_1 - C_2)^* = C_1^* - C_2^* = C_2 - C_1 = -(C_1 - C_2),$$

zaključujemo da je matrica $C_1 - C_2$ kosohermitska, pa je za svako $\boldsymbol{y}$

$$\boldsymbol{y}^*(C_1 - C_2)\boldsymbol{y} = \mathrm{i}c \quad (c \in \mathbb{R}). \tag{2.5.4}$$

Pretpostavimo sada da je λ bilo koja sopstvena vrednost matrice $K(\omega)$ i $\boldsymbol{y}$ odgovarajući sopstveni vektor. Tada je $K(\omega)\boldsymbol{y} = \lambda\boldsymbol{y}$, tj.

$$[(1-\omega)D - \omega C_2]\boldsymbol{y} = \lambda(D + \omega C_1)\boldsymbol{y}. \tag{2.5.5}$$

Množenjem jednakosti (2.5.5) sleva sa $2\boldsymbol{y}^*$ dobijamo

$$\boldsymbol{y}^*[(2-\omega)D - \omega C_2]\boldsymbol{y} = \lambda\boldsymbol{y}^*[(2-\omega)D + \omega(D + 2C_1)]\boldsymbol{y},$$

tj.

$$\begin{aligned}\boldsymbol{y}^*[(2-\omega)D - \omega A + \omega(C_1 - C_2)]\boldsymbol{y} \\ = \lambda\boldsymbol{y}^*[(2-\omega)D + \omega A + \omega(C_1 - C_2)]\boldsymbol{y},\end{aligned} \tag{2.5.6}$$

s obzirom na to da je $A = C_1 + D + C_2$.

Kako su matrice A i D pozitivno definitne imamo

$$(\forall \boldsymbol{y} \neq \boldsymbol{0}) \quad \boldsymbol{y}^*A\boldsymbol{y} = a > 0 \quad \text{i} \quad \boldsymbol{y}^*D\boldsymbol{y} = d > 0. \tag{2.5.7}$$

Na osnovu (2.5.6), (2.5.4), (2.5.7) dobijamo

$$\lambda = \frac{\xi d - a + c_i}{\xi d + a + c_i} \quad \left(\xi = \frac{2}{\omega} - 1\right),$$

odakle, pod pretpostavkom da je

$$0 < \xi < +\infty, \text{ tj.} \quad 2 > \omega > 0,$$

sleduje $|\lambda| < 1$, čime je teorema 2.5.2 dokazana. $\square$

U daljem izlaganju dajemo ocenu greške za razmatrani relaksacioni metod. Kako je $A\boldsymbol{x} = \boldsymbol{b}$, na osnovu (2.5.2) redom dobijamo

$$\omega A\boldsymbol{x}^{(k)} = [(1-\omega)D - \omega C_2](\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k)}) + \omega\boldsymbol{b},$$

$$\omega A(\boldsymbol{x}^{(k)} - \boldsymbol{x}) = [(1-\omega)D - \omega C_2](\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k)}),$$

$$\boldsymbol{x}^{(k)} - \boldsymbol{x} = A^{-1}(\gamma D + C_2)(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}), \tag{2.5.8}$$

gde je $\gamma = 1 - 1/\omega$. Iz (2.5.8) sleduje:

Teorema 2.5.3. *Kod iterativnog procesa* (2.5.1) *važi ocena za grešku*

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \|A^{-1}(\gamma D + C_2)\| \cdot \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\| \quad (k \in \mathbb{N}). \tag{2.5.9}$$

Izraz (2.5.9), koji dodaje ocenu greške, najopštiji je za iterativni proces (2.5.1), s obzirom da ne zahteva nikakva ograničenja za matricu sistema A (sem da je regularna). Međutim, zbog komplikovanosti izraza $\|A^{-1}(\gamma D + C_2)\|$, ocena (2.5.9) nije pogodna. Zato ćemo, nadalje, za matricu A pretpostaviti dodatne uslove (kao u teoremi 2.5.2).

Dakle, neka je matrica A hermitska pozitivno definitna. Tada je D, takođe, pozitivno definitna, pa postoji matrica $D^{1/2}$ i njena inverzna matrica $D^{-1/2}$.

Uvođenjem oznaka i transformacija

$$C = C_1 + C_2, \quad D^{-1/2}AD^{-1/2} = B, \quad D^{-1/2}CD^{-1/2} = -T,$$

$$D^{-1/2}C_1D^{-1/2} = -T_1, \quad D^{-1/2}C_2D^{-1/2} = -T_2, \quad D^{1/2}\boldsymbol{x} = \boldsymbol{y}, \quad D^{1/2}\boldsymbol{b} = \boldsymbol{\beta},$$

sistem $A\boldsymbol{x} = \boldsymbol{b}$ se svodi na

$$B\boldsymbol{y} = \boldsymbol{\beta}, \quad \text{tj.} \quad \boldsymbol{y} = T\boldsymbol{y} + \boldsymbol{\beta}. \tag{2.5.10}$$

Primetimo da je matrica B pozitivno definitna i T hermitska.

Greška kod relaksacionog metoda primenjenog na (2.5.10), tj. kod metoda

$$(I - \omega T_1)\boldsymbol{y}^{(k)} = [(1-\omega)I + \omega T_2]\boldsymbol{y}^{(k-1)} + \omega\boldsymbol{\beta} \quad (k = 1,2,\dots), \tag{2.5.11}$$

je

$$\|\boldsymbol{y}^{(k)} - \boldsymbol{y}\| \leq \|B^{-1}(\gamma I - T_2)\| \cdot \|\boldsymbol{y}^{(k)} - \boldsymbol{y}^{(k-1)}\| \quad (k \in \mathbb{N}), \tag{2.5.12}$$

gde je $B = I - T$.

Kod primene relaksacionog metoda veoma je važan izbor vrednosti relaksacionog parametra ω. U literaturi ([2], [8], [36]) je definisan optimalni parametar ω_{opt} pomoću

$$\rho(K(\omega_{\text{opt}})) = \min_{\omega} \rho(K(\omega)),$$

gde je sa $\rho(S)$ označen spektralni radijus matrice S. Primetimo da je kod (2.5.11)

$$K(\omega) = (I - \omega T_1)^{-1}[(1-\omega)I + \omega T_2]. \tag{2.5.13}$$

J. ALBRECHT[140] ([2]) je razmatrao iterativni proces (2.5.11), pri uslovu da je hermitska matrica T (u sistemu (2.5.10) ciklična indeksa 2 (videti [72]), tj. da je oblika

$$T = \begin{bmatrix} O & | & F \\ - & - & - \\ E & | & O \end{bmatrix} \begin{matrix} n_1 \text{ vrsta} \\ \\ n_2 \text{ vrsta} \end{matrix} \qquad (n_1 + n_2 = n) \tag{2.5.14}$$

$$\begin{matrix} n_1 & n_2 \\ \text{kolona} & \text{kolona} \end{matrix}$$

i da je $\rho(T) < 1$. Matrica T ima svojstvo (A) (videti definiciju 3.1.2, odeljak 2.3.1).

Na osnovu prethodnog imamo

$$T_1 = \begin{bmatrix} O & | & O \\ - & - & - \\ E & | & O \end{bmatrix} \quad \text{i} \quad T = \begin{bmatrix} O & | & F \\ - & - & - \\ O & | & O \end{bmatrix}.$$

Teorema 2.5.4. *Neka su T i $K(\omega)$ matrice date pomoću* (2.5.14) *i* (2.5.13)*, respektivno. Ako je $\begin{pmatrix} \tau \\ \lambda \end{pmatrix}$ sopstvena vrednost od $\begin{pmatrix} T \\ K(\omega) \end{pmatrix}$ i važi*

$$(\lambda + \omega - 1)^2 = \lambda \omega^2 \tau^2, \tag{2.5.15}$$

tada je $\begin{pmatrix} \tau \\ \lambda \end{pmatrix}$ sopstvena vrednost od $\begin{pmatrix} T \\ K(\omega) \end{pmatrix}$.

Na osnovu teoreme 2.5.4, odredićemo optimalnu vrednost relaksacionog množioca ω u iterativnom procesu (2.5.11).

[140] JULIUS ALBRECHT (1926 – 2012), nemački matematičar.

Neka je hermitska matrica T ciklična indeksa 2 i neka je njen spektralni radijus $\rho = \rho(T) = \max|\tau(T)| < 1$.

Ako je τ sopstvena vrednost matrice T, tada je $-\tau$ takođe njena sopstvena vrednost. Na osnovu teoreme 2.5.4, ovim sopstvenim vrednostima odgovaraju dve sopstvene vrednosti matrice $K(\omega)$. Naime, iz (2.5.15) sleduje

$$\lambda_{\pm} = f_{\pm}(\omega,\tau) = \frac{1}{4}\left(\omega|\tau| \pm \sqrt{\omega^2\tau^2 - 4\omega + 4}\right)^2.$$

Kako je $f_+(\omega,\tau)f_-(\omega,\tau) = (\omega-1)^2$, zaključujemo da iterativni proces (2.5.11) divergira ako je $|\omega - 1| \geq 1$, tj. $\omega \leq 0$ ili $\omega \geq 2$. Zato ćemo razmotriti samo slučaj kada je $0 < \omega < 2$.

Ako je $0 < \omega \leq \omega_1 = \omega_1(\tau) = \dfrac{1}{1+\sqrt{1-\tau^2}}$, koreni $\lambda_{\pm}$ su realni i pozitivni i veći od njih je

$$\lambda_+ = f_+(\omega,\tau) = \frac{1}{4}\left(\omega|\tau| + \sqrt{\omega^2\tau^2 - 4\omega + 4}\right)^2.$$

Ako je $\omega_1 < \omega < 2$, koreni $\lambda_{\pm}$ su kompleksni sa modulom $\omega - 1$. Kako je funkcija $|\tau| \mapsto f_+(\omega,\tau)$ rastuća na $(0,\rho)$, zaključujemo da je

$$\rho(K(\omega)) = \begin{cases} f_+(\omega,\tau), & 0 < \omega < \dfrac{2}{1+\sqrt{1-\rho^2}}, \\ \omega - 1, & \dfrac{2}{1+\sqrt{1-\rho^2}} < \omega < 2. \end{cases}$$

Grafik funkcije $\omega \mapsto \rho(K(\omega))$ $(0 < \omega < 2)$ prikazan je na slici 2.5.1.

Optimalna vrednost parametra ω je

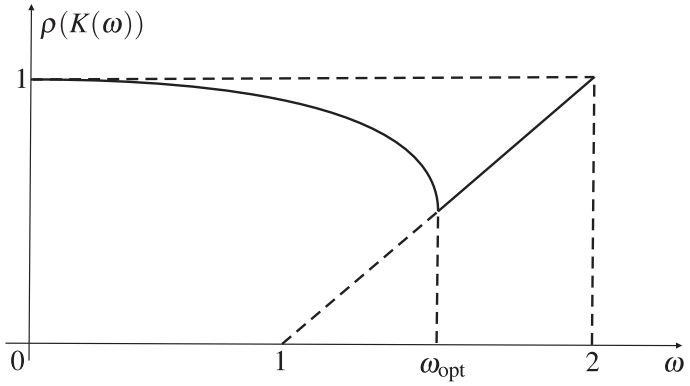
$$\omega_{\text{opt}} = \omega_1(\rho) = \frac{2}{1+\sqrt{1-\rho^2}},$$

s obzirom na to da funkcija $\omega \mapsto \rho(K(\omega))$ ima minimum za $\omega = \omega_{\text{opt}}$ na segmentu $(0,2)$. Odgovarajuća vrednost za $\rho(K(\omega))$ je

$$\rho(K(\omega_{\text{opt}})) = \omega_{\text{opt}} - 1 = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1+\rho^2}}.$$

Uvedimo oznaku

$$M(\omega) = [(I-T)^{-1}(\gamma I - T_2)]^* \cdot [(I-T)^{-1}(\gamma I - T_2)].$$



Slika 2.5.1. Grafik funkcije $\omega \mapsto \rho(K(\omega))$ kada $\omega \in [0,2]$

Teorema 2.5.5. *Ako je* $\begin{pmatrix}\tau\\ \mu\end{pmatrix}$ *sopstvena vrednost od* $\begin{pmatrix}T\\ M(\omega)\end{pmatrix}$ *i važi*

$$\frac{1}{2}\left\{\frac{1-\tau^2}{\gamma^2}\mu+\frac{1}{\dfrac{1-\tau^2}{\gamma^2}\mu}\right\}=1+\frac{1}{2}\cdot\frac{\xi^2\tau^2+\tau^4}{\gamma^2(1-\tau^2)},$$

gde su $\gamma=1-\dfrac{1}{\omega}$ *i* $\xi=\dfrac{2}{\omega}-1$, *tada je* $\begin{pmatrix}\mu\\ \tau\end{pmatrix}$ *sopstvena vrednost od* $\begin{pmatrix}M(\omega)\\ T\end{pmatrix}$.

Na osnovu teoreme 2.5.5, nalazimo

$$\begin{aligned}H(\omega)&=\|(I-T)^{-1}(\gamma I-T_2)\|_{\text{sp}}\\&=\frac{1}{2}\cdot\frac{\sqrt{\rho^2(\xi^2+\rho^2)+4\gamma^2(1-\rho^2)}+\sqrt{\rho^2(\xi^2+\rho^2)}}{1-\rho^2},\end{aligned}$$

gde je $\rho=\rho(T)=\sigma(T)<1$.

Teorema 2.5.6. *Neka je hermitska matrica* T *ciklična indeksa* 2 *i neka je njen spektralni radijus* $\rho=\rho(T)<1$.

Ako je $0<\omega<2$, *za iterativni proces* (2.5.11) *važi ocena za grešku*

$$\|\boldsymbol{y}^{(k)}-\boldsymbol{y}\|_E\le H(\omega)\cdot\|\boldsymbol{y}^{(k)}-\boldsymbol{y}^{(k-1)}\|_E\quad(k\in\mathbb{N}).$$

Napomenimo da je kod GAUSS–SEIDELovog metoda, tj. kada je relaksacioni parametar $\omega = 1$,

$$H = H(1) = \frac{\rho\sqrt{1-\rho^2}}{1-\rho^2}.$$

4.2.6 ČEBIŠEVljev semi-iterativni metod

U ovom odeljku obradićemo ČEBIŠEVljev semi-iterativni metod i ukazaćemo na njegovu vezu sa metodom sukcesivne gornje relaksacije.

Neka je dat sistem jednačina

$$\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{\beta} \tag{2.6.1}$$

sa hermitskom matricom B reda n. Ako je $\rho(B) < 1$, iterativni proces

$$\boldsymbol{x}^{(k)} = B\boldsymbol{x}^{(k-1)} + \boldsymbol{\beta} \quad (k = 1, 2, \dots) \tag{2.6.2}$$

konvergira ka rešenju $\boldsymbol{x}$ sistema (2.6.1). U cilju ubrzanja konvergencije procesa (2.6.2), posmatrajmo linearnu kombinaciju vektora $\boldsymbol{x}^{(k)}$, tj.

$$\boldsymbol{y}^{(k)} = \sum_{i=0}^{k} c_{ki}\boldsymbol{x}^{(i)} \quad (k = 0, 1, \dots), \tag{2.6.3}$$

pod uslovom $\sum\limits_{i=0}^{k} c_{ki} = 1$. Često se ovakav postupak za ubrzavanje konvergencije naziva linearno ubrzavanje (videti [28]).

Ako stavimo $\boldsymbol{\varepsilon}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}$ i $\boldsymbol{\zeta}(k) = \boldsymbol{y}^{(k)} - \boldsymbol{x}$, $k = 0, 1, \dots$, imamo

$$\boldsymbol{\zeta}(k) = \sum_{i=0}^{k} c_{ki}\boldsymbol{\varepsilon}^{(i)} = \left(\sum_{i=0}^{k} c_{ki}B^i\right)\boldsymbol{\varepsilon}^{(0)},$$

tj.

$$\boldsymbol{\zeta}^{(k)} = Q_k(B)\boldsymbol{\varepsilon}^{(0)}, \tag{2.6.4}$$

gde je

$$Q_k(t) = \sum_{i=0}^{k} c_{ki}t^i \quad \text{i} \quad Q_k(1) = 1. \tag{2.6.5}$$

Neka je $\{\boldsymbol{e}_i\}$ ortonormiran sistem sopstvenih vektora hermitske matrice B, gde je $B\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i \ \ (\lambda_i \equiv \lambda_i(B)),\ \ i = 1, \ldots, n$. Ako vektor $\varepsilon^{(0)}$ razvijemo po sopstvenim vektorima $\boldsymbol{e}_i$, tj. $\varepsilon^{(0)} = p_1\boldsymbol{e}_1 + \cdots + p_n\boldsymbol{e}_n$, jednakost (2.6.4) postaje

$$\zeta(k) = \sum_{i=1}^{n} p_i Q_k(\lambda_i)\boldsymbol{e}_i.$$

Označimo sa $\mathcal{P}_k$ skup svih polinoma Q_k oblika (2.6.5), tj. skup svih polinoma stepena k sa normalizacijom $Q_k(1) = 1$. Može se pokazati da opšta tročlana rekurentna formula kojom se generišu polinomi $Q_k \in P_k$ ima oblik

$$Q_{k+1}(t) = (\alpha_k t + 1 - \alpha_k - \beta_k)Q_k(t) + \beta_k Q_{k-1}(t) \quad (k = 0, 1, \ldots), \tag{2.6.6}$$

gde su $Q_0(t) = 1$, $\beta_0 = 0$, α_k i β_k realni brojevi.

Kako je

$$\|\zeta^{(k)}\|_E = \left(\sum_{i=1}^{n} p_i^2 Q_k(\lambda_i)^2\right)^{1/2} \leq \max_i |Q_k(\lambda_i)| \left(\sum_{i=1}^{n} p_i^2\right)^{1/2},$$

tj.

$$\|\zeta^{(k)}\|_E \leq \|\varepsilon^{(k)}\|_E \max_i |Q_k(\lambda_i)|$$

za izbor polinoma $\{\widetilde{Q_k}\}$ usvojimo kriterijum

$$\min_{Q_k \in \mathcal{P}_k} \left(\max_{-\rho \leq t \leq \rho} |Q_k(t)|\right) = \max_{-\rho \leq t \leq \rho} |\widetilde{Q}_k(t)|,$$

gde je $\rho \equiv \rho(B)$. S obzirom na ČEBIŠEVljevu teoremu 1.6.3 (glava 2) nalazimo (videti, takođe, [2] i [36])

$$\widetilde{Q}_k(t) = \frac{T_k(t/\rho)}{T_k(1/\rho)} \quad (k = 0, 1, \ldots), \tag{2.6.7}$$

gde je T_k ČEBIŠEVljev polinom prve vrste stepena k. Lako je pokazati da se, u ovom slučaju, rekurentna relacija (2.6.6) svodi na

$$\widetilde{Q}_{k+1}(t) = \alpha_k(t)\widetilde{Q}_k(t) + (1 - \alpha_k)\widetilde{Q}_{k-1}(t) \quad (k = 0, 1, \ldots), \tag{2.6.8}$$

gde su

$$\alpha_0 = 1,\ \beta_0 = 0,\ \alpha_k = \frac{2}{\rho}\cdot\frac{T_k(1/\rho)}{T_{k+1}(1/\rho)},\ \beta_k = 1-\alpha_k \quad (k=1,2,\dots). \tag{2.6.9}$$

Dakle, na osnovu (2.6.8), generiše se niz polinoma $\{\widetilde{Q}_k\}$, a zatim, na osnovu (2.6.3), i niz $\{\boldsymbol{y}^{(k)}\}$. Ovaj postupak je poznat kao ČEBIŠEVljev semi–iterativni metod.

Pokazaćemo sada kako se izloženi metod može predstaviti i u eksplicitnom obliku (videti [36]).

Iz (2.6.8) sleduje

$$\widetilde{Q}_{k+1}(t) - \widetilde{Q}_{k-1}(t) = \alpha_k(t\widetilde{Q}_k(t) - \widetilde{Q}_{k-1}(t)).$$

S druge strane, kako je $\boldsymbol{\zeta}^{(k+1)} - \boldsymbol{\zeta}^{(k-1)} = (\widetilde{Q}_{k+1}(B) - \widetilde{Q}_{k-1}(B))\boldsymbol{\varepsilon}^{(0)}$ imamo

$$\boldsymbol{\zeta}^{(k+1)} - \boldsymbol{\zeta}^{(k-1)} = \alpha_k(B\widetilde{Q}_k(B) - \widetilde{Q}_{k-1}(B))\boldsymbol{\varepsilon}^{(0)} = \alpha_k(B\boldsymbol{\zeta}^{(k)} - \boldsymbol{\zeta}^{(k-1)}),$$

tj.

$$\boldsymbol{y}^{(k+1)} = \boldsymbol{y}^{(k-1)} + \alpha_k(B\boldsymbol{y}^{(k)} + \boldsymbol{\beta} - \boldsymbol{y}^{(k-1)}) \quad (k=1,2,\dots), \tag{2.6.10}$$

gde je niz $\{\alpha_k\}$ definisan pomoću (2.6.9) i

$$\boldsymbol{y}^{(0)} = \boldsymbol{x}^{(0)} \quad \text{i} \quad \boldsymbol{y}^{(1)} = B\boldsymbol{x}^{(0)} + \boldsymbol{\beta}.$$

Na osnovu prethodnog, članovi niza $\{\alpha_k\}$ su

$$\alpha_0 = 1, \quad \alpha_1 = \frac{2}{2-\rho^2}, \quad \alpha_k = \frac{1}{1-\dfrac{1}{4}\rho^2\alpha_{k-1}} \quad (k=2,3,\dots),$$

pri čemu je

$$2 \geq \alpha_1 \geq \alpha_2 \geq \cdots \geq \lim_{k\to+\infty}\alpha_k = \omega_{\text{opt}} = \frac{2}{1+\sqrt{1-\rho^2}} > 1.$$

Dakle, α_k teži optimalnom relaksacionom faktoru, koji je dobijen kod metoda gornje relaksacije.

Napomena 2.6.1. Ako za niz polinoma $\{\widetilde{Q}_k\}$ izaberemo

$$\widetilde{Q}_k(t) = \frac{1-T_{k+1}(t)}{(k+1)^2(1-t)} = \frac{\sin^2\dfrac{k+1}{2}\theta}{(k+1)^2\sin^2\dfrac{\theta}{2}} \quad (t=\cos\theta),$$

dobijamo metod LANCZOSa ([51]).

Napomena 2.6.2. E. STIEFEL[141] ([67]) je razmatrao jedan relaksacioni metod, tzv. hipergeometrijsku relaksaciju, koristeći umesto ČEBIŠEVljevih polinoma ultrasferne (GEGENBAUERove) polinome.

4.2.7 Gradijetni metodi

U klasi metoda koji se koriste kod minimizacije funkcionela posebnu ulogu igraju gradijentni metodi. U ovom odeljku obradićemo dva gradijentna metoda za rešavanje sistema linearnih jednačina

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{2.7.1}$$

gde je A normalna matrica (videti definiciju 3.4.7, odeljak 2.3.4). Prvi od njih je metod najbržeg pada, a drugi metod konjugovanih gradijenata. Napomenimo da metod konjugovanih gradijenata u suštini nije iterativni metod. Oba navedena metoda zasnivaju se na minimizaciji funkcionele $F : \mathbb{R}^n \to \mathbb{R}$, definisane pomoću

$$F(\boldsymbol{x}) = (A\boldsymbol{x},\boldsymbol{x}) - 2(\boldsymbol{b},\boldsymbol{x}). \tag{2.7.2}$$

Kako je A normalna matrica imamo

$$F(\boldsymbol{x}) - F(A^{-1}\boldsymbol{b}) = (A(\boldsymbol{x} - A^{-1}\boldsymbol{b}),\boldsymbol{x} - A^{-1}\boldsymbol{b}) \geq 0,$$

odakle zaključujemo da funkcionela F postiže minimum za $\boldsymbol{x} = A^{-1}\boldsymbol{b}$, što predstavlja rešenje sistema (2.7.1).

1. Metod najbržeg pada. Svaki metod oblika

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - \alpha_{k-1} \operatorname{grad} F(\boldsymbol{x}^{(k-1)}) \quad (k = 1,2,\dots) \tag{2.7.3}$$

naziva se *metod gradijentnog pada*.

Ako se parametar α_p određuje iz uslova da je za svako $p \in \mathbb{N}_0$ veličina

$$F\left(\boldsymbol{x}^{(p)} - \alpha_p \operatorname{grad} F(\boldsymbol{x}^{(p)})\right) \tag{2.7.4}$$

minimalna, metod (2.7.3) se naziva *metod najbržeg pada*.

Iz (2.7.2) sleduje $\operatorname{grad} F(\boldsymbol{x}) = 2(A\boldsymbol{x} - \boldsymbol{b})$. Tada je

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - 2\alpha_{k-1}(A\boldsymbol{x}^{(k-1)} - \boldsymbol{b}) \quad (k = 1,2,\dots). \tag{2.7.5}$$

[141] EDUARD L. STIEFEL (1909 – 1978), švajcarski matematičar.

Ako u jednakosti

$$F(\boldsymbol{x}+t\boldsymbol{r}) = F(\boldsymbol{x}) + 2t(A\boldsymbol{x}-\boldsymbol{b},\boldsymbol{r}) + t^2(A\boldsymbol{r},\boldsymbol{r}) \quad (t \in \mathbb{R})$$

izvršimo supstituciju

$$\begin{pmatrix} \boldsymbol{x} & \boldsymbol{r} & t \\ \boldsymbol{x}^{(p)} & A\boldsymbol{x}^{(p)}-\boldsymbol{b} & -2\alpha_p \end{pmatrix} \quad (p \in \mathbb{N}_0).$$

dobijamo da je vrednost izraza (2.7.4) data sa

$$F(\boldsymbol{x}^{(p+1)} = F(\boldsymbol{x}^{(p)}) - 4\alpha_p(\boldsymbol{r}^{(p)},\boldsymbol{r}^{(p)}) + 4\alpha_p^2(A\boldsymbol{r}^{(p)},\boldsymbol{r}^{(p)}),$$

gde je $\boldsymbol{r}^{(p)} = A\boldsymbol{x}^{(p)} - \boldsymbol{b}$.

Kvadratni trinom $\alpha_p \mapsto \phi(\alpha_p) = F(\boldsymbol{x}^{(p+1)})$ ima minimalnu vrednost ako je

$$2\alpha_p = \frac{(\boldsymbol{r}^{(p)},\boldsymbol{r}^{(p)})}{(A\boldsymbol{r}^{(p)},\boldsymbol{r}^{(p)})}. \tag{2.7.6}$$

Na osnovu (2.7.5) i (2.7.6) dobijamo metod najvećeg pada

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - \frac{\|A\boldsymbol{x}^{(k-1)}-\boldsymbol{b}\|_E^2}{(A(A\boldsymbol{x}^{(k-1)}-\boldsymbol{b}),A\boldsymbol{x}^{(k-1)}-\boldsymbol{b})}(A\boldsymbol{x}^{(k-1)}-\boldsymbol{b}), \tag{2.7.7}$$

gde je $k = 1,2,\dots$.

Teorema 2.7.1. *Neka je $\boldsymbol{x}$ tačno rešenje sistema (2.7.1) čija je matrica A normalna. Za metod najvećeg pada važe nejednakosti*

$$F(\boldsymbol{x}^{(k)}) - F(\boldsymbol{x}) \leq \left(\frac{M-m}{M+m}\right)^2 (F(\boldsymbol{x}^{(k-1)}) - F(\boldsymbol{x})) \tag{2.7.8}$$

i

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|_E \leq \left(\frac{M-m}{M+m}\right)^k \sqrt{\frac{M}{m}}\, \| \boldsymbol{x}^{(0)} - \boldsymbol{x} \|_E, \tag{2.7.9}$$

gde je $\boldsymbol{x}^{(0)}$ proizvoljan početni vektor i $0 < m \leq \lambda_i(A) \leq M$.

Dokaz. Neka je $\boldsymbol{x}^{(k-1)}$ dobijeno pomoću (2.7.7), polazeći od proizvoljnog vektora $\boldsymbol{x}^{(0)}$. Stavimo $\boldsymbol{y}^{(k-1)} = \boldsymbol{x}^{(k-1)}$ i odredimo $\boldsymbol{y}^{(k)}$ koristeći optimalni metod (2.2.14), tj.

$$\boldsymbol{y}^{(k)} = \boldsymbol{y}^{(k-1)} - \frac{2}{M+m}(A\boldsymbol{y}^{(k-1)} - \boldsymbol{b}). \tag{2.7.10}$$

Kako je (2.7.10) oblika (2.7.5) sa parametrom $2/(M+m)$ i kako je $F(\boldsymbol{x}^{(k)})$ minimalno kada je $2\alpha_{k-1}$ određeno pomoću (2.7.6), zaključujemo da važi nejednakost

$$F(\boldsymbol{x}^{(k)}) \leq F(\boldsymbol{y}^{(k)}). \tag{2.7.11}$$

Neka je $\{\boldsymbol{e}_i\}$ ortonormiran sistem sopstvenih vektora matrice A. Tada vektor $\boldsymbol{\zeta}^{(k-1)} = \boldsymbol{y}^{(k-1)} - \boldsymbol{x}$ možemo predstaviti u obliku $\boldsymbol{\zeta}^{(k-1)} = c_1\boldsymbol{e}_1 + \cdots + c_n\boldsymbol{e}_n$. Kako je $A\boldsymbol{e}_i = \lambda_i\boldsymbol{e}_i$ $(\lambda_i \equiv \lambda_i(A))$, imamo

$$(A\boldsymbol{\zeta}^{(k-1)}, \boldsymbol{\zeta}^{(k-1)}) = \sum_{i=1}^{n} c_i^2\lambda_i. \tag{2.7.12}$$

Na osnovu (2.7.10) imamo $B = I - \dfrac{2}{M+m}A$ i

$$\boldsymbol{\zeta}^{(k)} = \boldsymbol{y}^{(k)} - \boldsymbol{x} = B(\boldsymbol{y}^{(k-1)} - \boldsymbol{x}) = B\boldsymbol{\zeta}^{(k-1)}.$$

Nadalje, kako je matrica B simetrična sa sopstvenim vrednostima $1 - \dfrac{2}{M+m}\lambda_i$ $(i = 1, \ldots, n)$ imamo

$$(A\boldsymbol{\zeta}^{(k)}, \boldsymbol{\zeta}^{(k)}) = (AB\boldsymbol{\zeta}^{(k-1)}, B\boldsymbol{\zeta}^{(k-1)}) = (BAB\boldsymbol{\zeta}^{(k-1)}, \boldsymbol{\zeta}^{(k-1)}),$$

tj.

$$(A\boldsymbol{\zeta}^{(k)}, \boldsymbol{\zeta}^{(k)}) = \sum_{i=1}^{n} c_i^2\lambda_i\Big(1 - \frac{2}{M+m}\lambda_i\Big)^2, \tag{2.7.13}$$

s obzirom da su sopstvene vrednosti matrice BAB određene sa

$$\lambda_i\Big(1 - \frac{2}{M+m}\lambda_i\Big)^2 \quad (i = 1, \ldots, n).$$

S druge strane, kako je

$$\left|1-\frac{2}{M+m}\lambda_i\right| \le \frac{M-m}{M+m} \quad (i=1,\dots,n),$$

na osnovu (2.7.12) i (2.7.13), zaključujemo da je

$$(A\zeta^{(k)},\zeta^{(k)}) \le \left(\frac{M-m}{M+m}\right)^2 (A\zeta^{(k-1)},\zeta^{(k-1)}), \tag{2.7.14}$$

tj.

$$F(y^{(k)})-F(x) \le \left(\frac{M-m}{M+m}\right)^2 (F(y^{(k-1)})-F(x)). \tag{2.7.15}$$

Da bismo dokazali nejednakost (2.7.9) stavimo $\varepsilon^{(p)}=x^{(p)}-x$ $(p=0,1,\dots)$. Primetimo da je $\varepsilon^{(k-1)}=\zeta^{(k-1)}$ i

$$(A\varepsilon^{(k)},\varepsilon^{(k)}) = F(x^{(k)})-F(x) \le F(y^{(k)})-F(x) = (A\zeta^{(k)},\zeta^{(k)}).$$

Tada iz (2.7.14) sleduje

$$(A\varepsilon^{(k)},\varepsilon^{(k)}) \le \left(\frac{M-m}{M+m}\right)^2 (A\varepsilon^{(k-1)},\varepsilon^{(k-1)}).$$

Iteriranjem poslednje nejednakosti dobijamo

$$(A\varepsilon^{(k)},\varepsilon^{(k)}) \le \left(\frac{M-m}{M+m}\right)^{2k} (A\varepsilon^{(0)},\varepsilon^{(0)}),$$

odakle je, s obzirom na teoremu 3.4.6 (odeljak 2.3.4),

$$m\|\varepsilon^{(k)}\|_E^2 \le \left(\frac{M-m}{M+m}\right)^{2k} M\|\varepsilon^{(0)}\|_E^2,$$

tj. (2.7.9). Ovim je dokazana teorema 2.7.1. □

Na kraju napomenimo da su metod najbržeg pada i optimalni metod (2.7.15) dosta slični. Principijelna razlika ovih metoda je u tome što metod najbržeg pada ne zahteva informaciju o granicama spektra matrice A, kao što je slučaj kod optimalnog metoda.

2. Metod konjugovanih gradijenata. U radovima [43] i [66] predložen je metod za rešavanje jednačine (2.7.1) koji teorijski (ne uzimajući u obzir greške zaokruglji-vanja) konvergira ka tačnom rešenju u najviše n iteracija, gde je n red matrice. Metod se sastoji u konstrukciji niza $\boldsymbol{x}^{(k)}$ pomoću

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(0)} + \sum_{i=0}^{k-1} \alpha_i A^i \boldsymbol{r}^{(0)} \qquad (k = 1, \ldots, n),$$

polazeći od proizvoljnog početnog vektora $\boldsymbol{x}^{(0)}$, pri čemu je $\boldsymbol{r}^{(0)} = A\boldsymbol{x}^{(0)} - \boldsymbol{b}^{(0)}$, dok se koeficijenti α_i određuju iz uslova da funkcija $(\alpha_0, \ldots, \alpha_{k-1}) \mapsto F(\boldsymbol{x}^{(k)})$ postiže minimum. Iz praktičnih razloga, kod primene ovog metoda dobro je konstruisati niz vektora $\{\boldsymbol{p}^{(k)}\}$, koji su međusobno konjugovani u smislu

$$(\boldsymbol{p}^{(i)}, A\boldsymbol{p}^{(j)}) = 0 \quad (i \neq j).$$

Tada se navedeni metod može iskazati rekurzivno pomoću sledećih formula:

$$\begin{aligned}
\boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} - c_k \boldsymbol{p}^{(k)}) \qquad (k = 0, 1, \ldots, n-1), \\
\boldsymbol{r}^{(k)} &= A\boldsymbol{x}^{(k)} - \boldsymbol{b} \qquad (k = 0, 1, \ldots, n), \\
\boldsymbol{p}^{(0)} &= -\boldsymbol{r}^{(0)}, \quad \boldsymbol{p}^{(k)} = -\boldsymbol{r}^{(k)} + q_k \boldsymbol{p}^{(k-1)} \qquad (k = 1, \ldots, n-1), \\
c_k &= \frac{(\boldsymbol{p}^{(k)}, \boldsymbol{r}^{(k)})}{(\boldsymbol{p}^{(k)}, A\boldsymbol{p}^{(k)})}, \quad q_k = \frac{(\boldsymbol{r}^{(k)}, A\boldsymbol{p}^{(k-1)})}{(\boldsymbol{p}^{(k-1)}, A\boldsymbol{p}^{(k-1)})} \qquad (k = 0, 1, \ldots, n-1).
\end{aligned}$$

Kao što je napred rečeno, za neko $m \leq n$ teorijski biće $\boldsymbol{r}^{(m)} = 0$, što znači da je određeno tačno rešenje sistema (2.7.1).

4.2.8 Iterativni metodi za inverziju matrica

S obzirom da veliki broj metoda u numeričkoj analizi zahteva inverziju ma-trica, ili u opštem slučaju inverziju linearnih ograničenih operatora, ovaj odeljak posvećujemo ovom problemu. Sve rezultate koje ćemo ovde izneti, odnose se na inverziju matrica, ali se mogu formalno preneti i na inverziju linearnih ograničenih operatora (videti [5], [60]).

Pretpostavimo da je A regularna matrica reda n.

Teorema 2.8.1. *Neka je r prirodan broj veći od jedinice i neka je*

$$F_k = I - AX_k \quad (k = 0, 1, \dots),$$

gde je X_0 data matrica takva da je

$$\|F_0\| = \|I - AX_0\| \leq q < 1. \tag{2.8.1}$$

Tada niz matrica $\{X_k\}$ definisan sa

$$X_k = X_{k-1}(I + F_{k-1} + \cdots + F_{k-1}^{r-1}) \quad (k = 1, 2, \dots), \tag{2.8.2}$$

konvergira ka A^{-1}.

Dokaz. Na osnovu

$$\begin{aligned} F_k &= I - AX_k \\ &= I - AX_{k-1}(I + F_{k-1} + \cdots + F_{k-1}^{r-1}) \\ &= I - (I - F_{k-1})(I + F_{k-1} + \cdots + F_{k-1}^{r-1}) \\ &= F_{k-1}^r, \end{aligned}$$

dobijamo

$$F_k = F_o^{r^k} \quad (k = 0, 1, \dots).$$

Iz poslednje jednakosti i uslova (2.8.1) redom sleduje

$$\|F_k\| \leq \|F_0\|^{r^k} \leq q^{r^k} \quad (k = 0, 1, \dots)$$

i

$$\lim_{k \to +\infty} F_k = \lim_{k \to +\infty} (I - AX_k) = 0, \quad \text{tj.} \quad \lim_{k \to +\infty} X_k = A^{-1}. \quad \square$$

Teorema 2.8.2. *Ako su ispunjeni uslovi teoreme* 2.8.1*, za svako $k \in \mathbb{N}$ važe nejednakosti*

$$\|X_k - A^{-1}\| \leq \frac{\|X_k F_k\|}{1 - \|F_k\|}, \tag{2.8.3}$$

$$\|X_k - A^{-1}\| \leq \|F_{k-1}\|^{r-1} \frac{\|X_{k-1} F_{k-1}\|}{1 - \|F_{k-1}\|}, \tag{2.8.4}$$

$$\|X_k - A^{-1}\| \leq \|F_0\|^{r^k} \frac{\|X_0\|}{1 - \|F_0\|}. \tag{2.8.5}$$

Dokaz. Ako stavimo, $E_k = A^{-1} - X_k$, imamo

$$E_k - E_k F_k = (A^{-1} - X_k)AX_k = X_k(I - AX_k) = X_k F_k.$$

Kako je $\|F_k\| < 1$, važi nejednakost

$$\|E_k\|(1 - \|F_k\|) \leq \|E_k - E_k F_k\| = \|X_k F_k\|,$$

odakle neposredno sleduje (2.8.3).

Nejednakosti (2.8.4) i (2.8.5) dokazuju se na sličan način (videti [60]). □

Teorema 2.8.3. *Ako su ispunjeni uslovi teoreme* 2.8.1*, iterativni proces* (2.8.2) *ima red konvergencije r.*

Dokaz. Ako u nejednakosti (2.8.3), tj. nejednakosti

$$\|X_{k+1} - A^{-1}\| \leq \frac{\|X_{k+1}F_{k+1}\|}{1 - \|F_{k+1}\|},$$

uvedemo smenu

$$F_{k+1} = F_k^r = (I - AX_k)^r = A^r(A^{-1} - X_k)^r$$

dobijamo

$$\|X_{k+1} - A^{-1}\| \leq \frac{\|X_{k+1}A^r(A^{-1} - X_k)^r\|}{1 - \|F_{k+1}\|} \leq \frac{\|X_{k+1}\| \cdot \|A\|^r}{1 - \|F_{k+1}\|}\|X_k - A^{-1}\|^r.$$

Kako je

$$\lim_{k \to +\infty} \frac{\|X_{k+1}\| \cdot \|A\|^r}{1 - \|F_{k+1}\|} = \|A^{-1}\| \cdot \|A\|^r < +\infty,$$

imamo

$$\|X_{k+1} - A^{-1}\| = O(\|X_k - A^{-1}\|^r) \quad (k \to +\infty),$$

što znači da iterativni proces (2.8.2) za inverziju matrice A, ima red konvergencije r, čime je dokaz teoreme 2.8.3 završen. □

U specijalnom slučaju za $r = 2$, (2.8.2) se svodi na

$$X_k = X_{k-1}(2I - AX_{k-1}) \quad (k = 1, 2, \dots).$$

Ovaj iterativni proces sa kvadratnom konvergencijom potiče od G. SCHULZa[142] ([64]), dok se ocene za grešku, analogne onim u teoremi 2.8.2, mogu naći u radovima [4], [7], [21], [46]. (videti, takođe, [22], [1]).

U slučaju kada je $r = 3$, odgovarajući iterativni proces sa kubnom konvergencijom je

$$X_k = X_{k-1}(3I - 3AX_{k-1} + (AX_{k-1})^2) \quad (k = 1, 2, \dots).$$

U radu [4], J. ALBRECHT je ukazao na vezu ovog procesa sa NEWTONovim metodom za rešavanje nelinearnih jednačina (videti poglavlje 5.1).

4.3 PROBLEM SOPSTVENIH VREDNOSTI

4.3.1 Lokalizacija sopstvenih vrednosti

U odeljku 2.3.3 su date osnovne definicije i stavovi koji se odnose na sopstvene vrednosti i sopstvene vektore kvadratne matrice $A = [a_{ij}]_{n\times n}$. Ovo poglavlje biće posvećeno numeričkim metodima za njihovo odredjivanje, s obzirom na veliki značaj koji oni imaju u numeričkoj matematici. Naime, veliki broj problema se svodi na rešavanje problema sopstvenih vrednosti. U ovom odeljku daćemo neke rezultate koji se odnose na lokalizaciju sopstvenih vrednosti u kompleksnoj ravni, dok se u narednim odeljcima daju metodi za određivanje karakterističnog polinoma matrice, metodi za određivanje dominantnih i subdominantnih sopstvenih vrednosti i odgovarajućih sopstvenih vektora, kao i metodi za rešavanje kompletnog problema sopstvenih vrednosti. Posebna pažnja je posvećena simetričnim trodijagonalnim matricama zbog važnosti koje one imaju u mnogim primenama.

Sledeći rezultat je poznat kao GERSHGORINova[143] teorema.

Teorema 3.1.1. *Neka je $A = [a_{ij}]_{n\times n}$ kvadratna matrica reda n i neka su $C_1, \dots, C_n$ diskovi u kompleksnoj ravni određeni sa*

$$C_i = \left\{ z \in \mathbb{C} \;\middle|\; |z - a_{ii}| \le r_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \right\} \qquad (i = 1, \dots, n).$$

Ako sa C označimo uniju ovih diskova, tada se sve sopstvene vrednosti matrice A nalaze u C.

[142] GÜNTHER SCHULZ (1903 – 1962), nemački matematičar.
[143] SEMYON ARANOVICH GERSHGORIN (1901 – 1933), beloruski matematičar.

Dokaz. Neka je λ sopstvena vrednost matrice A, a $\boldsymbol{x}$ odgovarajući sopstveni vektor normalizovan tako da je $\|\boldsymbol{x}\|_\infty = \max\limits_i |x_i| = x_m = 1$. Tada je $\lambda\boldsymbol{x} = A\boldsymbol{x}$, tj.

$$(\lambda - a_{ii})x_i = \sum_{\substack{j=1 \\ j\neq m}}^{n} a_{ij}x_j \quad (i = 1,\ldots,n),$$

odakle, za $i = m$ imamo

$$|\lambda - a_{mm}| \leq \sum_{\substack{j=1 \\ j\neq m}}^{n} |a_{ij}| \cdot |x_j| \leq \sum_{\substack{j=1 \\ j\neq m}}^{n} |a_{ij}| = r_m.$$

Dakle, sopstvena vrednost λ leži u disku C_m. Kako je λ proizvoljna sopstvena vrednost matrice A, zaključujemo da se sve njene sopstvene vrednosti nalaze u uniji diskova, tj. u C. $\square$

Napomena 3.1.1. S obzirom da matrica A^T ima iste sopstvene vrednosti kao i matrica A, na osnovu prethodne teoreme može se zaključiti da se sve sopstvene vrednosti matrice A nalaze i u uniji D diskova

$$D_j = \left\{ z \in \mathbb{C} \;\middle|\; |z - a_{jj}| \leq s_j = \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}| \right\} \qquad (j = 1,\ldots,n).$$

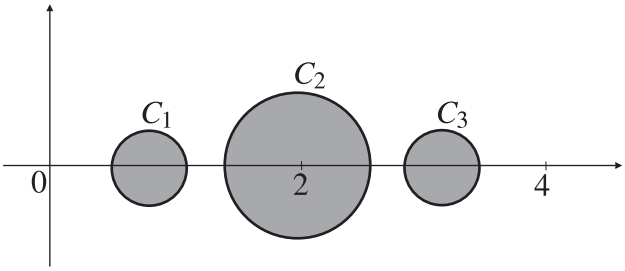
Na osnovu prethodnog zaključujemo da sve sopstvene vrednosti matrice A leže u preseku skupova C i D.

Teorema 3.1.2. *Ako m diskova iz teoreme* 3.1.1 *čini povezanu oblast, koja je izolovana od ostalih diskova, tada se u toj povezanoj oblasti nalazi tačno m sopstvenih vrednosti matrice A.*

Dokaz ove teoreme može se naći, na primer, u izvanrednoj monografiji WILKINSONa ([73]).

Primer 3.1.1. Neka je

$$A = \begin{bmatrix} 1 & 0.1 & -0.1 \\ 0 & 2 & 0.4 \\ -0.2 & 0 & 3 \end{bmatrix}.$$



Slika 3.1.1. Lokalizacija sopstvenih vrednosti matrice A

Na osnovu teoreme 3.1.1 sopstvene vrednosti se nalaze u sledećim diskovima (videti sliku 3.1.1): $C_1 = \{z \in \mathbb{C} \mid |z-1| \leq 0.2\}$, $C_2 = \{z \in \mathbb{C} \mid |z-2| \leq 0.4\}$, $C_3 = \{z \in \mathbb{C} \mid |z-3| \leq 0.2\}$.

Primetimo da na osnovu prethodne napomene 3.1.1 sleduje da diskovi D_1, D_2, D_3 imaju redom poluprečnike 0.2, 0.1, 0.5. Inače, tačne vrednosti sopstvenih vrednosti matrice A, na sedam decimala, su $\lambda_1 = 0.9861505$, $\lambda_2 = 2.0078436$, $\lambda_3 = 3.0060058$, dok je normirani karakteristični polinom

$$H(\lambda) = \lambda^3 - 6\lambda^2 + 10.98\lambda - 5.952. \quad \triangle$$

Teorema o lokalizaciji sopstvenih vrednosti ima i teorijski i praktični značaj (na primer, za određivanje početnih vrednosti kod iterativnih metoda, za analizu kod perturbacionih problema, itd.).

Za određivanje sopstvenih vrednosti postoji veliki broj metoda, pri čemu neki od njih omogućavaju nalaženje svih sopstvenih vrednosti, dok se neki mogu primeniti samo na određivanje nekih od sopstvenih vrenosti, na primer, određivanje dominantnih, tj. onih sa maksimalnim modulom. Neki od metoda omogućavaju samo nalaženje koeficijenata karakterističnog polinoma, tako da se za određivanje sopstvenih vrednosti mora primenjivati neki od metoda za rešavanje algebarskih jednačina (videti poglavlje 5.3). Ovakav pristup u određivanju sopstvenih vrednosti se ne preporučuje jer je u većini slučajeva numerički nestabilan, tj. slabo uslovljen. Naime, kako su koeficijenti karakterističnog polinoma opterećeni, u opštem slučaju, greškama zaokrugljivanja, usled slabe uslovljenosti karakterističnog polinoma dolazi do velikih grešaka u sopstvenim vrednostima. O uticaju promene koeficijenata na promenu nula kod slabo uslovljenih polinoma videti odeljak 5.3.1.

4.3.2 Metodi za određivanje karakterističnog polinoma

U ovom odeljku ukratko ćemo navesti nekoliko metoda za određivanje karakterističnog polinoma matrice $A = [a_{ij}]_{n\times n}$,

$$P(\lambda) = \det(A - \lambda I). \tag{3.2.1}$$

Kao što je rečeno na kraju prethodnog odeljka nije preporučljivo koristiti ovako dobijeni polinom za odredjivanje sopstvenih vrednosti i sopstvenih vektora matrice A, sem u slučajevima kada je karakteristični polinom dobro uslovljen.

1. Metod KRILOVa. Umesto (3.2.1) posmatraćemo normalizovani karakteristični polinom

$$H(\lambda) = (-1)^n p(\lambda) = \lambda^n - p_1\lambda^{n-1} + p_2\lambda^{n-2} - \cdots + (-1)^n p_n. \tag{3.2.2}$$

Na osnovu CAYLEY-HAMILTONove teoreme 3.3.1 (odeljak 2.3.3) imamo

$$H(A) = A^n - p_1A^{n-1} + p_2A^{n-2} - \cdots + (-1)^n p_nI = 0.$$

Neka je sada $\boldsymbol{y}^{(0)}$ proizvoljan n-dimenzionalni vektor, sa kojim pomnožimo prethodnu jednakost s desne strane. Tada dobijamo

$$p_1A^{n-1}\boldsymbol{y}^{(0)} - p_2A^{n-2}\boldsymbol{y}^{(0)} + \cdots + (-1)^{n-1}p_n\boldsymbol{y}^{(0)} = A^n\boldsymbol{y}^{(0)},$$

odakle, korišćenjem koordinatne reprezentacije, dolazimo da sistema linearnih jednačina

$$\begin{bmatrix} y_1^{(n-1)} & -y_1^{(n-2)} & \cdots & (-1)^{n-1}y_1^{(0)} \\ y_2^{(n-1)} & -y_2^{(n-2)} & & (-1)^{n-1}y_2^{(0)} \\ \vdots & & & \\ y_n^{(n-1)} & -y_n^{(n-2)} & & (-1)^{n-1}y_n^{(0)} \end{bmatrix} \cdot \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} y_1^{(n)} \\ y_2^{(n)} \\ \vdots \\ y_n^{(n)} \end{bmatrix} \tag{3.2.3}$$

gde smo stavili $\boldsymbol{y}^{(k)} = A^k\boldsymbol{y}^{(0)} = [y_1^{(k)}\ y_2^{(k)}\ \cdots\ y_n^{(k)}]^T\ \ (k = 1,2,\ldots,n)$. Primetimo da stepene matrice A^k ne treba izračunavati, već treba koristiti rekurzivni postupak

$$\boldsymbol{y}^{(k)} = A\boldsymbol{y}^{(k-1)} \quad (k = 1,\ldots,n). \tag{3.2.4}$$

Pod uslovom da je matrica dobijenog sistema linearnih jednačina (3.2.3) regularna, rešavanjem ovog sistema dobijamo koeficijente p_1, p_2, …, p_n. Ako je, međutim, matrica ovog sistema singularna treba promeniti početni vektor $\boldsymbol{y}^{(0)}$.

Izloženi metod je poznat kao metod KRILOVa.[144] Linearni potprostor generisan pomoću kvadratne matrice A reda n i n-dimenzionalnog vektora $\boldsymbol{b}$, kao linearna kombinacija $\mathcal{K}_m(A,\boldsymbol{b}) = \text{span}\{\boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{m-1}\boldsymbol{b}\}$ naziva se *potprostor* KRILOV*a reda m*.

Ilustrovaćemo primenu ovog metoda na nalaženje karakterističnog polinoma jedne matrice četvrtog reda.

Primer 3.2.1. Neka je

$$A = \begin{bmatrix} 3 & 2 & -2 & -1 \\ -1 & 3 & -1 & 0 \\ 1 & -2 & 4 & 1 \\ 3 & 0 & 1 & 3 \end{bmatrix}.$$

Ako uzmemo $\boldsymbol{y}^{(0)} = [1 \quad 0 \quad 0 \quad 0]^T$, pomoću (3.2.4) nalazimo redom

$$\boldsymbol{y}^{(1)} = A\boldsymbol{y}^{(0)} = \begin{bmatrix} 3 \\ -1 \\ 1 \\ 3 \end{bmatrix}, \quad \boldsymbol{y}^{(2)} = A\boldsymbol{y}^{(1)} = \begin{bmatrix} 2 \\ -7 \\ 12 \\ 19 \end{bmatrix},$$

$$\boldsymbol{y}^{(3)} = A\boldsymbol{y}^{(2)} = \begin{bmatrix} -51 \\ -35 \\ 83 \\ 75 \end{bmatrix}, \quad \boldsymbol{y}^{(4)} = A\boldsymbol{y}^{(3)} = \begin{bmatrix} -464 \\ -137 \\ 426 \\ 155 \end{bmatrix}.$$

Sistem (3.2.3) postaje

$$\begin{bmatrix} -51 & -2 & 3 & -1 \\ -35 & 7 & -1 & 0 \\ 83 & -12 & 1 & 0 \\ 75 & -19 & 3 & 0 \end{bmatrix} \cdot \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = \begin{bmatrix} -464 \\ -137 \\ 426 \\ 155 \end{bmatrix},$$

odakle dobijamo

$$p_1 = 13, \;\; p_2 = 67, \;\; p_3 = 151, \;\; p_4 = 120.$$

Prema tome, imamo

$$H(\lambda) = p(\lambda) = \lambda^4 - 13\lambda^3 + 67\lambda^2 - 151\lambda + 120. \quad \triangle$$

[144] ALEKSEY NIKOLAEVICH KRYLOV (1863 – 1945), ruski mornarički inženjer i matematičar.

2. LEVERRIERov metod i modifikacija FADDEEVa. LEVERRIERov[145] metod se bazira na poznatim NEWTONovim formulama za sume stepena svih nula polinoma. Neka su $\lambda_1,\dots,\lambda_n$ sopstvene vrednosti matrice A, tj. nule polinoma (3.2.2), pri čemu se svaka nula uzima onoliko puta kolika je njena višestrukost. Tada za sume

$$s_m = \lambda_1^m + \lambda_2^m + \cdots + \lambda_n^m \quad (m = 0,1,\dots,n)$$

važe NEWTONove formule (videti, na primer, [56, str. 241–242])

$$s_m - p_1 s_{m-1} + p_2 s_{m-2} - \cdots + (-1)^{m-1} p_{m-1} s_1 + (-1)^m m p_m = 0, \quad m = 1,\dots,n,$$

odakle imamo

$$\begin{aligned}
p_1 &= s_1,\\
p_2 &= -\frac{1}{2}(s_2 - p_1 s_1),\\
p_3 &= \frac{1}{3}(s_3 - p_1 s_2 + p_2 s_1),\\
&\vdots\\
p_n &= \frac{(-1)^{n-1}}{n}(s_n - p_1 s_{n-1} + \cdots + (-1)^{n-1} p_{n-1} s_1)
\end{aligned}$$

Prema tome, ako su sume s_m poznate možemo naći koeficijente karakterističnog polinoma. Primetimo da je

$$s_1 = \lambda_1 + \lambda_2 + \cdots + \lambda_n = \operatorname{tr} A = \sum_{i=1}^{n} a_{ii}.$$

Kako su $\lambda_1^m, \lambda_2^m, \dots, \lambda_n^m$ sopstvene vrednosti matrice A^m (videti teoremu 3.3.2) zaključujemo da je

$$s_m = \operatorname{tr} A^m.$$

Dakle, ako je $A^m = [a_{ij}^{(m)}]_{n\times n}$ imamo $s_m = \sum\limits_{i=1}^{n} a_{ii}^{(m)}$, pri čemu stepene matrice određujemo redom pomoću

$$A^m = A \cdot A^{m-1}.$$

[145] URBAIN JEAN JOSEPH LE VERRIER (1811 – 1877), francuski matematičar poznat po rezultatima u nebeskoj mehanici, a posebno za doprinos u otkriću planete Neptun.

Primer 3.2.2. Za matricu iz prethodnog primera imamo redom

$$A^2 = A \cdot A = \begin{bmatrix} 2 & 16 & -17 & -8 \\ -7 & 9 & -5 & 0 \\ 12 & -12 & 17 & 6 \\ 19 & 4 & 1 & 7 \end{bmatrix}, \quad A^3 = A \cdot A^2 = \begin{bmatrix} -51 & 86 & -96 & -43 \\ -35 & 23 & -15 & 2 \\ 83 & -46 & 62 & 23 \\ 75 & 48 & -31 & 3 \end{bmatrix},$$

$$A^4 = A \cdot A^3 = \begin{bmatrix} -464 & 348 & -411 & -174 \\ -137 & 29 & -11 & 26 \\ 426 & -96 & 151 & 48 \\ 155 & 356 & -319 & -97 \end{bmatrix}.$$

Kako je

$$\begin{aligned} s_1 &= \operatorname{tr} A = 3+3+4+3 = 13, \\ s_2 &= \operatorname{tr} A^2 = 2+9+17+7 = 35, \\ s_3 &= \operatorname{tr} A^3 = -51+23+62+3 = 37, \\ s_4 &= \operatorname{tr} A^4 = -464+29+151-97 = -381, \end{aligned}$$

imamo redom

$$\begin{aligned} p_1 &= 13, \\ p_2 &= -(35-13\cdot 3)/2 = 67, \\ p_3 &= (37-13\cdot 5+67\cdot 3)/3 = 151, \\ p_4 &= -(-381-13\cdot 7+67\cdot 5-151\cdot 3)/4 = 120. \quad \triangle \end{aligned}$$

Jednu modifikaciju LEVERRIERovog metoda dao je FADDEEV.[146] Ta modifikacija zahteva manji broj numeričkih operacija i sastoji se u sledećem (videti, takođe, [47]):

Umesto stepena A^m izračunava se niz matrica A_m $(m = 1, \ldots, n)$ pomoću formula

$$A_m = AB_{m-1}, \quad q_m = \frac{1}{m}\operatorname{tr} A_m, \quad B_m = A_m - q_m I,$$

pri čemu se za B_0 uzima jedinična matrica.

[146] DMITRII KONSTANTINOVICH FADDEEV (1907 – 1989), poznati ruski matematičar, čija je supruga VERA NIKOLAEVNA FADDEEVA (1906 – 1983), takođe, bila matematičar. Veći broj radova iz oblasti numeričke linearne algebre objavili su zajedno (videti, na primer, [24], [25], [26]).

Matematičkom indukcijom se može dokazati da je $q_m = (-1)^{m-1} p_m \quad (m = 1,\dots,n)$. Metod, takođe, omogućava dobijanje inverzne matrice. Naime, iz $A_n = q_n I = A(A_{n-1} - q_{n-1} I)$ imamo

$$A^{-1} = \frac{1}{q_n}\,(A_{n-1} - q_{n-1} I)\,. \tag{3.2.5}$$

Primer 3.2.3. Primenom metoda FADDEEVa za matricu iz prethodnih primera dobijamo redom

$$A_1 = A, \quad q_1 = \operatorname{tr} A_1 = 13, \quad B_1 = A_1 - 13I = \begin{bmatrix} -10 & 2 & -2 & -1 \\ -1 & -10 & -1 & 0 \\ 1 & -2 & -9 & 1 \\ 3 & 0 & 1 & -10 \end{bmatrix},$$

$$A_2 = AB_1 = \begin{bmatrix} -37 & -10 & 9 & 5 \\ 6 & -30 & 8 & 0 \\ -1 & 14 & -35 & -7 \\ -20 & 4 & -12 & 32 \end{bmatrix}, \quad q_2 = \frac{1}{2}\operatorname{tr} A_2 = -67,$$

$$B_2 = A_2 + 67I = \begin{bmatrix} 30 & -10 & 9 & 5 \\ 6 & 37 & 8 & 0 \\ -1 & 14 & 32 & -7 \\ -20 & 4 & -12 & 35 \end{bmatrix}, \; A_3 = AB_3 = \begin{bmatrix} 124 & 12 & -9 & -6 \\ -11 & 107 & -17 & 2 \\ -6 & -24 & 109 & 12 \\ 29 & -4 & 23 & 113 \end{bmatrix},$$

$$q_3 = \frac{1}{3}\operatorname{tr} A_3 = 151, \quad B_3 = A_3 - 151I = \begin{bmatrix} -27 & 12 & -9 & -6 \\ -11 & -44 & -17 & 2 \\ -6 & -24 & -42 & 12 \\ 29 & -4 & 23 & -38 \end{bmatrix},$$

$$A_4 = AB_3 = \begin{bmatrix} -120 & 0 & 0 & 0 \\ 0 & -120 & 0 & 0 \\ 0 & 0 & -120 & 0 \\ 0 & 0 & 0 & -120 \end{bmatrix}, \quad q_4 = \frac{1}{4}\operatorname{tr} A_4 = -120.$$

Dakle, $p_1 = q_1 = 13$, $p_2 = -q_2 = 67$, $p_3 = q_3 = 151$, $p_4 = -q_4 = 120$. $\triangle$

U daljem tekstu navodimo MATHEMATICA kôd za prethodnu modifikaciju FADDEEVa.

Kao što možemo videti, u modulu `Faddeev` se izračunavaju i štampaju:

1° niz matrica $\{A_m\}_{m=1}^{n}$, startujući sa $A_1 \equiv A$;

```
In[1]:= Faddeev[A_] :=
  Module[{m, n = Length[A], Ainv},
   A1 = A;
   q1 = Tr[A1];
   Print["A1 = ", MatrixForm[A1]];
   Print["q1 = Tr[A1] = ", q1];
   For[m = 2, m ≤ n, m++,
    Am = A.(Am-1 - qm-1 IdentityMatrix[n]); qm = 1/m Tr[Am];
    Print["A"m, " = ", MatrixForm[Am] ];
    Print["q"m, " = ", 1/m, "Tr[", "A"m, "] = ", qm];];
   Ainv = 1/qn (An-1 - qn-1 IdentityMatrix[n]);
   P[λ_] = λ^n - Sum[qm λ^(n-m), {m, 1, n}]; Print["P[λ] = ", P[λ]];
   Print["Ainv", " = ", MatrixForm[Ainv] ];
   Return[{P[λ], Ainv}];];
```

2° niz koeficijenata $\{q_m\}_{m=1}^n$;

3° inverzna matrica A^{-1}. (primenom formule (3.2.5)).

Iilustracije radi navodimo dobijene rezultate kada prethodni modul primenimo na sledeću matricu petog reda:

```
A={{-51, -2, 3, -1, 3}, {-35, 7, -1, 0, 1}, {83, -12, 1, 0, 4},
     {75, -19, 3, 0, -2}, {0, 1, 2, 3, 4}};
```

3. Metod DANILEVSKOG. Ovaj metod [17] se zasniva na transformaciji matrice A na tzv. FROBENIUSov oblik

$$F = \begin{bmatrix} f_1 & f_2 & \cdots & f_{n-1} & f_n \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & & 1 & 0 \end{bmatrix}, \tag{3.2.6}$$

pri čemu su matrice A i F slične, tj. postoji regularna matrica C takva da je $F = C^{-1}AC$ (videti definiciju 3.3.5, odeljak 2.3.3). S obzirom na to da slične matrice imaju identične karakteristične polinome, jednostavno se, na osnovu (3.2.6),

In[3]:= **Faddeev[A]**

$$A_1 = \begin{pmatrix} -51 & -2 & 3 & -1 & 3 \\ -35 & 7 & -1 & 0 & 1 \\ 83 & -12 & 1 & 0 & 4 \\ 75 & -19 & 3 & 0 & -2 \\ 0 & 1 & 2 & 3 & 4 \end{pmatrix}$$

$$_1 = \mathrm{Tr}[A_1] = -39$$

$$A_2 = \begin{pmatrix} 856 & -4 & -28 & 21 & -12 \\ 92 & 405 & -150 & 38 & -59 \\ -493 & -726 & 309 & -71 & 413 \\ 14 & -1062 & 360 & -81 & 132 \\ 356 & -31 & 96 & 129 & 175 \end{pmatrix}$$

$$_2 = \frac{1}{2}\mathrm{Tr}[A_2] = 832$$

$$A_3 = \begin{pmatrix} -1833 & -151 & 87 & -60 & -134 \\ 653 & -2154 & 549 & -269 & -1063 \\ 1819 & 3942 & -663 & 1732 & -2503 \\ -2139 & 5697 & -1011 & 382 & 2774 \\ 572 & -5189 & 268 & -2327 & -1465 \end{pmatrix}$$

$$_3 = \frac{1}{3}\mathrm{Tr}[A_3] = -1911$$

$$A_4 = \begin{pmatrix} 4028 & -1251 & 24 & -480 & 15 \\ 594 & -5547 & -182 & -3842 & 198 \\ 2745 & -26\,431 & 2953 & -9328 & 915 \\ -2244 & 15\,496 & -698 & 10\,461 & 1746 \\ 162 & 3976 & 1084 & 766 & 4037 \end{pmatrix}$$

$$_4 = \frac{1}{4}\mathrm{Tr}[A_4] = 3983$$

$$A_5 = \begin{pmatrix} 7482 & 0 & 0 & 0 & 0 \\ 0 & 7482 & 0 & 0 & 0 \\ 0 & 0 & 7482 & 0 & 0 \\ 0 & 0 & 0 & 7482 & 0 \\ 0 & 0 & 0 & 0 & 7482 \end{pmatrix}$$

$$_5 = \frac{1}{5}\mathrm{Tr}[A_5] = 7482$$

$$P[\lambda] = -7482 - 3983\,\lambda + 1911\,\lambda^2 - 832\,\lambda^3 + 39\,\lambda^4 + \lambda^5$$

$$\texttt{Ainv} = \begin{pmatrix} \frac{15}{2494} & -\frac{417}{2494} & \frac{4}{1247} & -\frac{80}{1247} & \frac{5}{2494} \\ \frac{99}{1247} & -\frac{4765}{3741} & -\frac{91}{3741} & -\frac{1921}{3741} & \frac{33}{1247} \\ \frac{915}{2494} & -\frac{26431}{7482} & -\frac{515}{3741} & -\frac{4664}{3741} & \frac{305}{2494} \\ -\frac{374}{1247} & \frac{7748}{3741} & -\frac{349}{3741} & \frac{3239}{3741} & \frac{291}{1247} \\ \frac{27}{1247} & \frac{1988}{3741} & \frac{542}{3741} & \frac{383}{3741} & \frac{9}{1247} \end{pmatrix}$$

dobija karakteristični polinom matrice A. Naime, ako det $(F-\lambda I)$ razvijemo po elementima prve kolone dobijamo

$$P(\lambda) = (f_1-\lambda)(-\lambda)^{n-1} - f_2(-\lambda)^{n-2} + \cdots + (-1)^{n-1} f_n,$$

tj.

$$P(\lambda) = (-1)^n(\lambda^n - f_1\lambda^{n-1} - f_2\lambda^{n-2} - \cdots - f_n).$$

Dakle, $p_m = (-1)^{m-1} f_m$ $(m = 1,\ldots,n)$.

Za vektor-vrstu $z = [z_1\ z_2\ \cdots\ z_n]$ definišimo kvadratnu matricu n-tog reda

$$G_r^{-1}(z) = \left[\begin{array}{c|c} I_{n-r} & O_{n-r,r} \\ \hline P_{r,n-r}(z) & R_r(z) \end{array}\right],$$

gde je I_{n-r} jedinična matrica reda $n-r$, $O_{n-r,r}$ nula matrica tipa $r\times(n-r)$, $P_{r,n-r}(z)$ pravougaona matrica tipa $r\times(n-r)$, koja u prvoj vrsti ima redom elemente $z_1, z_2, \ldots, z_{n-r}$, dok su svi ostali njeni elementi jednaki nuli, i najzad, $R_r(z)$ je kvadratna matrica reda r koja se razlikuje od jedinične matrice samo u prvoj vrsti, gde sadrži preostale elemente vektora z, tj. ima redom elemente $z_{n-r+1}, \ldots, z_n$.

Pretpostavimo da je $z_{n-r+1} \neq 0$. Na osnovu teoreme 3.1.3 (odeljak 2.3.1) lako se nalazi inverzna matrica za $G_r(z)$,

$$G_r^{-1}(z) = \left[\begin{array}{c|c} I_{n-r} & O_{n-r,r} \\ \hline -\dfrac{P_{r,n-r}(z)}{z_{n-r+1}} & R_r^{-1}(z) \end{array}\right],$$

gde je

$$R_r^{-1}(z) = \begin{bmatrix} \dfrac{1}{z_{n-r+1}} & -\dfrac{z_{n-r+2}}{z_{n-r+1}} & \cdots & -\dfrac{z_n}{z_{n-r+1}} \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}.$$

Neka je $A^{(k)}$ matrica n-tog reda čije su vrste vektori $\boldsymbol{a}_1^{(k)}, \ldots, \boldsymbol{a}_n^{(k)}$, tj.

$$A^{(k)} = \begin{bmatrix} \boldsymbol{a}_1^{(k)} \\ \vdots \\ \boldsymbol{a}_1^{(k)} \end{bmatrix}.$$

Stavimo sada da je $A^{(1)} = A$, gde je A data matrica čiju FROBENIUSovu formu tražimo.

Direktnim množenjem može se pokazati da se pomoću niza transformacija

$$\begin{aligned} A^{(2)} &= G_2(\boldsymbol{a}_n^{(1)})A^{(1)}G_2^{-1}(\boldsymbol{a}_n^{(1)}) && (a_{n,n-1}^{(1)} \neq 0), \\ A^{(3)} &= G_3(\boldsymbol{a}_n^{(2)})A^{(2)}G_3^{-1}(\boldsymbol{a}_n^{(2)}) && (a_{n-1,n-2}^{(2)} \neq 0), \\ &\vdots \\ A^{(n)} &= G_n(\boldsymbol{a}_2^{(n-1)})A^{(n-1)}G_n^{-1}(\boldsymbol{a}_2^{(n-1)}) && (a_{2,1}^{(n-1)} \neq 0), \end{aligned}$$

dolazi do FROBENIUSovog oblika $F = A^{(n)}$. Pri ovome je

$$C = G_2^{-1}(\boldsymbol{a}_n^{(1)})G_3^{-1}(\boldsymbol{a}_{n-1}^{(2)}) \cdots G_n^{-1}(\boldsymbol{a}_2^{(n-1)})$$

i

$$C^{-1} = G_n(\boldsymbol{a}_2^{(n-1)})G_{n-1}(\boldsymbol{a}_3^{(n-2)}) \cdots G_2(\boldsymbol{a}_n^{(1)}).$$

Prethodne transformacije egzistiraju pod navedenim uslovima.

Ovi uslovi mogu biti obezbeđeni dodatnom permutacijom vrsta i kolona. U slučaju kada je vektor-vrsta $\boldsymbol{a}_{n-k+1}^{(k)}$ nula vektor, problem se jednostavno redukuje na problem niže dimenzije.

Primer 3.2.4. Za matricu iz prethodnog primera imamo redom

$$G_2 = G_2(3,0,1,3) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 3 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad G_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -3 & 0 & 1 & -3 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$A^{(2)} = G_2 A G_2^{-1} = \begin{bmatrix} 9 & 2 & -2 & 5 \\ 2 & 3 & -1 & 3 \\ 16 & 4 & 1 & 4 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$G_3 = G_3(16,4,1,4) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 16 & 4 & 1 & 4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad G_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -4 & 1/4 & -1/4 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$A^{(3)} = G_3 A^{(2)} G_3^{-1} = \begin{bmatrix} 1 & 1/2 & -5/2 & 3 \\ -24 & 12 & -43 & 48 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$G_4 = G_4(-24,12,-43,48) = \begin{bmatrix} -24 & 12 & -43 & 48 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad G_4^{-1} = \begin{bmatrix} -\frac{1}{24} & \frac{1}{2} & -\frac{43}{24} & 2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$F = A^{(4)} = G_4 A^{(3)} G_4^{-1} = \begin{bmatrix} 13 & -67 & 151 & -120 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad \triangle$$

4.3.3 Metodi za dominantne sopstvene vrednosti

Vrlo često se u nekim primenama zahteva nalaženje samo maksimalne po modulu sopstvene vrednost i njoj odgovarajućeg sopstvenog vektora.

Neka su $\lambda_1, \ldots, \lambda_n$ sopstvene vrednosti i $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ odgovarajući sopstveni vektori matrice $A = [a_{ij}]_{n\times n}$. Ako je

$$|\lambda_1| = \cdots = |\lambda_r| > |\lambda_{r+1}| \geq \cdots \geq |\lambda_n|$$

kažemo da su $\lambda_1, \ldots, \lambda_r$ *dominantne sopstvene vrednosti* matrice A. U ovom odeljku razmatraćemo jedan metod za određivanje dominantne sopstvene vrednosti i odgovarajućeg sopstvenog vektora, kao i neke modifikacije ovog metoda. Pretpostavićemo, pritom, da su sopstveni vektori $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ linearno nezavisni, pa kao takvi oni čine jednu bazu u $\mathbb{R}^n$. Dakle, proizvoljan ne-nula vektor $\boldsymbol{v}_0$ se može izraziti pomoću

$$\boldsymbol{v}_0 = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i, \tag{3.3.1}$$

gde su α_i neki skalari. Definišimo sada iterativni proces

$$\boldsymbol{v}_k = A\boldsymbol{v}_{k-1} \quad (k = 1, 2, \ldots).$$

Tada je

$$\boldsymbol{v}_k = A\boldsymbol{v}_{k-1} = A^2\boldsymbol{v}_{k-2} = \cdots = A^k\boldsymbol{v}_0 = \sum_{i=1}^{n} \alpha_i A^k \boldsymbol{x}_i,$$

ili, s obzirom na (3.3.1) i tvrđenje teoreme 3.3.2 (odeljak 2.3.3),

$$\boldsymbol{v}_k = \sum_{i=1}^{n} \alpha_i \lambda_i^k \boldsymbol{x}_i. \tag{3.3.2}$$

Posebno je ovde interesantan slučaj kada imamo jednu dominantnu sopstvenu vrednost λ_1 $(r = 1)$. Pod pretpostavkom da je $\alpha_1 \neq 0$, na osnovu (3.3.2) imamo

$$\boldsymbol{v}_k = \alpha_i \lambda_i^k \left(\boldsymbol{x}_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \Big(\frac{\lambda_i}{\lambda_1} \Big)^k \boldsymbol{x}_i \right) = \alpha_1 \lambda_1^k (\boldsymbol{x}_1 + \varepsilon_k),$$

gde vektor $\varepsilon_k \to 0$, kada $k \to +\infty$.

Uvedimo sada oznaku $(\boldsymbol{y})_i$ za i-tu koordinatu nekog vektora $\boldsymbol{y}$. Tada je i-ta koordinata vektora $\boldsymbol{v}_k$

$$(\boldsymbol{v}_k)_i = \alpha_i \lambda_i^k \big((\boldsymbol{x}_1)_i + (\varepsilon_k)_i \big).$$

Kako je

$$\boldsymbol{v}_{k+1} = \alpha_1 \lambda_1^{k+1} (\boldsymbol{x}_1 + \varepsilon_{k+1}),$$

na osnovu prethodnog, za svako i $(1 \leq i \leq n)$ imamo

$$\frac{(\boldsymbol{v}_{k+1})_i}{(\boldsymbol{v}_k)_i} = \lambda_1 \frac{(\boldsymbol{x}_1)_i + (\varepsilon_{k+1})_i}{(\boldsymbol{x}_1)_i + (\varepsilon_k)_i} \to \lambda_1 \quad (k \to +\infty).$$

Na osnovu ove činjenice može se formulisati metod za određivanje dominantne sopstvene vrednosti λ_1, koji je poznat kao *metod stepenovanja*[147]. Vektor $\boldsymbol{v}_k$ je pri ovome jedna aproksimacija za nenormirani sopstveni vektor[148] koji odgovara dominantnoj sopstvenoj vrednosti. Pri praktičnoj realizaciji ovog metoda ide se na normiranje sopstvenog vektora, tj. na normiranje vektora $\boldsymbol{v}_k$ posle svakog iterativnog koraka. Normiranje se sprovodi deljenjem vektora $\boldsymbol{v}_k$ sa svojom koordinatom maksimalnog modula. Tako se metod stepenovanja može izraziti pomoću formula

$$\boldsymbol{z}_k = A\boldsymbol{v}_{k-1}, \quad \boldsymbol{v}_k = \frac{\boldsymbol{z}_k}{\gamma_k},$$

gde je γ_k koordinata vektora $\boldsymbol{z}_k$ sa najvećim modulom, tj. $\gamma = (z_k)_i$ i $|(z_k)_i| = \|\boldsymbol{z}_k\|$. Primetimo da $\gamma_k \to \lambda_1$ i $\boldsymbol{v}_k \to \boldsymbol{x}_1/\|\boldsymbol{x}_1\|_\infty$, kada $k \to +\infty$.

Brzina konvergencije ovog metoda zavisi od količnika $|\lambda_2/\lambda_1|$. Naime, važi

$$|\lambda_2 - \gamma_k| = O\Big(\Big|\frac{\lambda_2}{\lambda_1}\Big|^k\Big). \tag{3.3.3}$$

Primetimo da smo u izvodjenju ovog metoda pretpostavili da je $\alpha_1 \neq 0$, što znači da metod konvergira ako je λ_1 dominantna sopstvena vrednost i ako početni vektor $\boldsymbol{v}_0$ ima komponentu u pravcu koji odgovara sopstvenom vektoru $\boldsymbol{x}_1$. O ponašanju metoda bez ovih pretpostavki može se naći u radu [58], kao i u monografiji WILKINSONa [73, str. 570]. Praktično, zbog pojave grešaka zaokrugljivanja u iterativnom procesu, uslov $\alpha_1 \neq 0$ biće zadovoljen posle nekoliko koraka, iako polazna pretpostavka za vektor $\boldsymbol{v}_0$ nije ispunjena. Pri primeni ovog metoda često se unapred poznaje neka ocena za sopstveni vektor $\boldsymbol{x}_1$, pa se onda ona uzima kao početni vektor $\boldsymbol{v}_0$. Primenom perturbacione teorije mogu se dati izvesne ocene za (3.3.3) (videti, na primer, monografiju [35, str. 210–211]).

Prethodno razmatranje se odnosi na slučaj jedne dominantne sopstvene vrednosti. Slučajevi kada ima više dominantnih sopstvenih vrednosti razmatraju se slično kao i kod BERNOULLIevog metoda za odredjivanje dominantnih korena algebarskih jednačina (videti odeljak 5.3.3).

Primer 3.3.1. Neka je

$$A = \begin{bmatrix} -261 & 209 & -49 \\ -530 & 422 & -98 \\ -800 & 631 & -144 \end{bmatrix}$$

[147] Na engleskom jeziku: *power method*.

[148] Ako je $\boldsymbol{x}$ sopstveni vektor, tada je i $c\boldsymbol{x}$ $(c \neq 0)$, takodje, sopstveni vektor koji odgovara istoj sopstvenoj vrednosti.

čije su sopstvene vrednosti $\lambda_1 = 10$, $\lambda_2 = 4$, $\lambda_3 = 3$ (videti [35, str. 210]).

Uzimajući za početni vektor $\boldsymbol{v}_0 = [0 \quad 0 \quad -1]^T$, primenom metoda stepenovanja dobijamo rezultate koji su prikazani u tabeli 3.3.1. $\triangle$

Tabela 3.3.1.

k	γ_k	$(\boldsymbol{v}_k)_1$	$(\boldsymbol{v}_k)_2$	$(\boldsymbol{v}_k)_3$
1	144.	0.340278	0.680556	1.
2	13.2083	0.334911	0.669821	1.
3	10.7287	0.333774	0.667549	1.
4	10.2038	0.333463	0.666926	1.
5	10.0599	0.333372	0.666744	1.
6	10.0179	0.333345	0.666690	1.
7	10.0054	0.333337	0.666674	1.
8	10.0016	0.333334	0.666669	1.
9	10.0005	0.333334	0.666667	1.
10	10.0001	0.333333	0.666667	1.
11	10.0000	0.333333	0.666667	1.

S obzirom da je konvergencija metoda stepenovanja linearna za ubrzavanje konvergencije moze da se koristi AITKENov Δ^2-metod.

Jedan prost način za ubrzavanje konvergencije metoda stepenovanja, posebno kod matrica sa realnim sopstvenim vrednostima

$$\lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n, \tag{3.3.4}$$

je primena istog metoda na određivanje dominantne sopstvene vrednosti matrice $A - pI$, gde je parametar p na pogodan način izabran[149]. Naime, sopstvene vrednosti ovakve matrice su $\lambda_i - p$, dok sopstveni vektori ostaju nepromenjeni. Pod uslovom (3.3.4), dominantna sopstvena vrednost će biti $\lambda_1 - p$ ili $\lambda_n - p$. Neka je to prva od njih. Kako brzina konvergencije metoda stepenovanja zavisi od količnika $|\lambda_2/\lambda_1|$ to će brzina procesa primenjenog na $A - pI$ zavisiti od $|(\lambda_2 - p)/(\lambda_1 - p)|$ ili $|(\lambda_n - p)/(\lambda_n - p)|$ u zavisnosti od toga koji je od ovih količnika veći. Vrednosti parametra p za koje su ovi količnici manji od $|\lambda_2/\lambda_1|$

[149] Ovakva modifikacija se u anglo-saksonskoj literaturi sreće kao *shift of origin.*

daje bržu konvergenciju metoda, nego u standardnom slučaju kada je $p = 0$. Optimalna vrednost za p se jednostavno dobija, ako se poznaju vrednosti λ_2 i λ_n, kao

$$p = \frac{1}{2}(\lambda_n + \lambda_2).$$

Naravno, mi najčešće ove vrednosti ne poznajemo, ali je moguće, na primer, da znamo izvesne granice za ove sopstvene vrednosti, na osnovu kojih možemo približno odrediti parametar p.

Primer 3.3.2. U primeru 3.3.1 brzina konvergencije je, s obzirom na (3.3.3), dirigovana sa $(2/5)^k$. Ako za p uzmemo vrednost $(\lambda_1 + \lambda_2)/2 = 3.5$, brzina konvergencije će biti dirigovana sa $(1/13)^k$, što predstavlja značajno ubrzanje. Startujući od istog početnog vektora $\boldsymbol{v}_0$, sada se dobijaju rezultati koji su dati u tabeli 3.3.2.

Tabela 3.3.2.

k	γ_k	$(\boldsymbol{v}_k)_1$	$(\boldsymbol{v}_k)_2$	$(\boldsymbol{v}_k)_3$
1	147.5000	0.332203	0.664407	1.
2	5.9780	0.333428	0.666856	1.
3	6.5437	0.333326	0.666652	1.
4	6.4967	0.333334	0.666668	1.
5	6.5003	0.333333	0.666667	1.
6	6.5000	0.333333	0.666667	1.

Kao što možemo videti, ista tačnost je postignuta sa 6 iteracija kao u primeru 3.3.1. Primetimo da $\gamma_k \to \alpha_1 - 3.5 = 6.5$. $\triangle$

Na kraju razmotrimo slučaj kada je realna matrica $A = [a_{ij}]_{n\times n}$ simetrična. Tada su njene sopstvene vrednosti realni brojevi za koje se može pretpostaviti distribucija data sa (3.3.4).

Definicija 3.3.1. Ako je $\boldsymbol{x}$ ne nula vektor, veličina

$$r(\boldsymbol{x}) = \frac{(A\boldsymbol{x},\boldsymbol{x})}{(\boldsymbol{x},\boldsymbol{x})} = \frac{\boldsymbol{x}^T A\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}$$

se naziva RAYLEIGHov[150] količnik.

[150] JOHN WILLIAM STRUTT (LORD RAYLEIGH) (1842 – 1919), poznati engleski fizičar, dobitnik NOBELove nagrade za fikiku 1904. godine.

Na osnovu teoreme 3.4.6 (odeljak 2.3.4) za RAYLEIGHov količnik važi

$$\lambda_n \leq r(\boldsymbol{x}) \leq \lambda_1$$

Nalaženje najveće sopstvene vrednosti λ_1 može se interpretirati kao optimizacioni problem

$$\lambda_1 = \max_{\boldsymbol{x}\neq\boldsymbol{0}} r(\boldsymbol{x}), \tag{3.3.5}$$

s obzirom da se maksimum u (3.3.5) postiže kada je x sopstveni vektor koji odgovara sopstvenoj vrednosti λ_1. Rešavanje optimizacionog problema (3.3.5) se može sprovesti tako da se za $\boldsymbol{x}$ uzima vektor $\boldsymbol{x}_k$ koji se generiše pomoću metoda stepenovanja.

Neka su $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ ortonormirani sopstveni vektori, tj. $(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_j^T \boldsymbol{x}_i = \delta_{ij}$, gde je δ_{ij} KRONECKERova delta.

Kako je, na osnovu prethodnog,

$$\boldsymbol{v}_k = \sum_{i=1}^{n} \alpha_i \lambda_i^k \boldsymbol{x}_i \quad \text{i} \quad A\boldsymbol{v}_k = \sum_{i=1}^{n} \alpha_i \lambda_i^{k+1} \boldsymbol{x}_i,$$

imamo

$$(\boldsymbol{v}_k, \boldsymbol{v}_k) = \sum_{i=1}^{n} \alpha_i^2 \lambda_i^{2k} \quad \text{i} \quad (A\boldsymbol{v}_k, \boldsymbol{v}_k) = \sum_{i=1}^{n} \alpha_i^2 \lambda_i^{2k+1},$$

pa je

$$r(\boldsymbol{v}_k) = \frac{\sum\limits_{i=1}^{n} \alpha_i^2 \lambda_i^{2k+1}}{\sum\limits_{i=1}^{n} \alpha_i^2 \lambda_i^{2k}} = \lambda_1 + O\Big(\Big(\frac{\lambda_2}{\lambda_1}\Big)^{2k}\Big).$$

Prema tome, $r(\boldsymbol{v}_k) \to \lambda_1$ brže, nego γ_k kod osnovnog metoda stepenovanja.

4.3.4 Metodi za subdominantne sopstvene vrednosti

Pretpostavimo da su sopstvene vrednosti matrice A uređene, tj. da važi

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|.$$

U ovom odeljku biće razmatrani metodi za određivanje subdominantnih sopstvenih vrednosti, tj. $\lambda_2, \ldots, \lambda_m$ $(m < n)$. Izložićemo tri metoda.

1. Metod ortogonalizacije. Pretpostavimo, najpre, da je matrica A simetrična, i neka je, na primer, metodom stepenovanja odredjen sopstveni vektor $\boldsymbol{x}_1$ koji odgovara dominantnoj sopstvenoj vrednosti λ_1 ($|\lambda_1| > |\lambda_i|$, $i = 2,\dots,n$). Polazeći od proizvoljnog vektora $\boldsymbol{z}$ formirajmo vektor $\boldsymbol{v}_0$ koji je ortogonalan na $\boldsymbol{x}_1$. Tako imamo (videti GRAM–SCHMIDTov postupak ortogonalizacije, odeljak 2.1.7)

$$\boldsymbol{v}_0 = \boldsymbol{z} - \frac{(\boldsymbol{z},\boldsymbol{x}_1)}{(\boldsymbol{x}_1,\boldsymbol{x}_1)}\boldsymbol{x}_1. \tag{3.4.1}$$

Kako je $(\boldsymbol{v}_0,\boldsymbol{x}_1) = 0$, teorijski gledano niz $\boldsymbol{v}_k = A\boldsymbol{v}_{k-1}$ ($k = 1,2,\dots$) bi kod metoda stepenovanja mogao biti iskorišćen za odredjivanje λ_2 i odgovarajućeg sopstvenog vektora $\boldsymbol{x}_2$. Međutim, bez obzira što $\boldsymbol{v}_0$ nema komponentu u pravcu sopstvenog vektora $\boldsymbol{x}_1$, metod stepenovanja bi zbog prisustva grešaka zaokrugljivanja počeo, posle izvesnog broja koraka, da konvergira ka sopstvenom vektoru $\boldsymbol{x}_1$. O ovoj činjenici je bilo reči i u prethodnom odeljku.

Ovaj uticaj grešaka zaokrugljivanja moguće je odstraniti tzv. periodičnim „čišćenjem“ vektora $\boldsymbol{v}_0$ od komponente u pravcu $\boldsymbol{x}_1$. Ovo znači da posle, recimo r koraka, ponovo izračunamo $\boldsymbol{v}_0$ korišćenjem $\boldsymbol{v}_r$ umesto $\boldsymbol{z}$ u (3.4.1), tj. pomoću

$$\boldsymbol{v}_0 = \boldsymbol{v}_r - \frac{(\boldsymbol{v}_r,\boldsymbol{x}_1)}{(\boldsymbol{x}_1,\boldsymbol{x}_1)}\boldsymbol{x}_1.$$

Na ovaj način, ako je perioda „čišćenja“ dovoljno mala da ne dodje do značajne akumulacije grešaka zaokrugljivanja, metodom stepenovanja možemo odrediti sopstvenu vrednost λ_2 i sopstveni vektor $\boldsymbol{x}_2$.

Nastavljajući ovakav pristup možemo dalje odrediti λ_3 i $\boldsymbol{x}_3$.

U opštem slučaju, ako smo odredili $\lambda_1,\dots,\lambda_\nu$ i odgovarajuće vektore $\boldsymbol{x}_1,\dots,\boldsymbol{x}_\nu$ ($\nu < m$) moguće je odrediti $\lambda_{\nu+1}$ i $\boldsymbol{x}_{\nu+1}$ metodom stepenovanja, formirajući vektor $\boldsymbol{v}_0$ ortogona1an na $\boldsymbol{x}_1,\dots,\boldsymbol{x}_\nu$. Dakle, polazeći od proizvoljnog vektora $\boldsymbol{z}$ imamo

$$\boldsymbol{v}_0 = \boldsymbol{z}_0 - \sum_{i=1}^{\nu} \frac{(\boldsymbol{z},\boldsymbol{x}_i)}{(\boldsymbol{x}_i,\boldsymbol{x}_i)}\boldsymbol{x}_i, \tag{3.4.2}$$

što znači da vektor $\boldsymbol{v}_0$ ima komponente samo u pravcu preostalih sopstvenih vektora, tj. da je

$$\boldsymbol{v}_0 = \alpha_{\nu+1}\boldsymbol{x}_{\nu+1} + \cdots + \alpha_n\boldsymbol{x}_n.$$

Metod stepenovanja primenjen na $\boldsymbol{v}_0$ daje $\boldsymbol{x}_{\nu+1}$ i $\lambda_{\nu+1}$, ukoliko greške zaokrugljivanja nisu prisutne. S obzirom da ovo nije tačno, potrebno je često „čišćenje“ vektora $\boldsymbol{v}_k$ od komponenata u pravcu $\boldsymbol{x}_1,\dots,\boldsymbol{x}_\nu$. Drugim rečima, posle r koraka treba ponovo određivati $\boldsymbol{v}_0$ pomoću (3.4.2), uz korišćenje $\boldsymbol{v}_r$ umesto $\boldsymbol{z}$.

I u slučaju kada matrica A nije simetrična, ali ima potpun sistem sopstvenih vektora, prethodni orogonalizacioni postupak se može primeniti.

2. Metod inverzne iteracije. Ovaj metod se primenjuje na opštu matricu A i bazira se na rešavanju sistema jednačina

$$(A - pI)\boldsymbol{v}_k = \boldsymbol{v}_{k-1}, \tag{3.4.3}$$

gde je p konstanta, a $\boldsymbol{v}_0$ proizvoljan vektor. Sistem jednačina (3.4.3) se obično rešava GAUSSovim metodom eliminacije ili CHOLESKYevim metodom uz LR faktorizaciju matrice $B = A - pI$. Primetimo da je metod inverzne iteracije ekvivalentan metodu stepenovanja primenjenog na B. Prema tome, primenom metoda inverzne iteracije dobija se dominantna sopstvena vrednost matrice B, tj. $\mu_\nu = 1/(\lambda_\nu - p)$, za koju važi

$$\min_j |\lambda_j - p| = |\lambda_\nu - p|.$$

Sopstvena vrednost λ_ν je najbliža sopstvena vrednost matrice A broju p. Sopstveni vektor koji se pritom dobija isti je za matricu B i matricu A.

Pogodnim izborom parametra p mogu se, u principu, odrediti sve sopstvene vrednosti matrice A.

Kao i kod metoda stepenovanja, i ovde je pogodno izvršiti normiranje vektora $\boldsymbol{v}_k$ tako da imamo

$$B\boldsymbol{z}_k = \boldsymbol{v}_{k-1}, \quad \boldsymbol{v}_k = \frac{\boldsymbol{z}_k}{\gamma_k}, \tag{3.4.4}$$

gde je γ_k koordinata vektora $\boldsymbol{z}_k$ sa najvećim modulom.

Primer 3.4.1. Metodom inverzne iteracije za matricu

$$A = \begin{bmatrix} 4 & 1 & 4 \\ 1 & 10 & 1 \\ 4 & 1 & 10 \end{bmatrix}$$

odredićemo sopstvenu vrednost koja je najbliža broju $p = 9$, kao i odgovarajući sopstveni vektor.

Korišćenjem faktorizacije pomoću GAUSSovog metoda sa izborom glavnog elementa za matricu $B = A - 9I$ dobijamo

$$LR = PB,$$

gde su

$$L = \begin{bmatrix} 1 & 0 & 0 \\ -4/5 & 1 & 0 \\ -1/5 & 2/3 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} -5 & 1 & 4 \\ 0 & 9/5 & 21/5 \\ 0 & 0 & -1 \end{bmatrix}$$

i permutaciona matrica P određena indeksnim nizom $I = (1,3)$.

Sada se metod inverzne iteracije (3.4.4) može izraziti u obliku

$$Ly_k = Pv_{k-1}, \quad Rz_k = Ly_k, \quad v_k = \frac{z_k}{\gamma_k},$$

čijom primenom se dobijaju rezultati dati u tabeli 3.4.1. Za startni vektor smo

Tabela 3.4.1.

k	$(v_k)_1$	$(v_k)_2$	$(v_k)_3$	β_k
1	0.	1.	−1.	6.
2	−0.2	1.	−0.5	9.3
3	−0.17241	1.	−0.48276	9.34483
4	−0.17200	1.	−0.48000	9.34800
5	−0.17185	1.	−0.47980	9.34835
6	−0.17184	1.	−0.47977	9.34838

uzeli $v_0 = [1 \quad 0 \quad 0]^T$. U poslednjoj koloni tabele data je veličina $\beta_k = p + 1/\gamma_k$, koja daje aproksimaciju za odgovarajuću sopstvenu vrednost λ. Vidimo da je ta sopstvena vrednost približno jednaka 9.34838. $\triangle$

3. Metodi deflacije. Metodi iz ove klase se sastoje u konstrukciji niza matrica A_n $(= A), A_{n-1}, \ldots, A_1$, čiji je red jednak indeksu i pritom

$$\mathrm{Sp}\,(A_n) \supset \mathrm{Sp}\,(A_{n-1}) \supset \cdots \supset \mathrm{Sp}\,(A_1),$$

gde je $\mathrm{Sp}\,(A_k)$ spektar matrice A_k.

Opisaćemo sada jedan specijalan, ali važan slučaj metoda deflacije kada je matrica A hermitska.

Neka je $x = [x_1\ x_2\ \cdots\ x_n]^T$ sopstveni vektor matrice A koji odgovara sopstvenoj vrednosti λ i takav da je normiran

$$(x,x) = x^*x = \|x\|_E^2 = 1$$

i da mu je prva koordinata x_1 nenegativna.

Posmatrajmo matricu

$$P = I - 2\boldsymbol{w}\boldsymbol{w}^*, \tag{3.4.5}$$

gde je vektor $\boldsymbol{w} = [w_1\ w_2\ \cdots\ w_n]^T$ definisan pomoću prvog vektora $\boldsymbol{e}_1 = [1\ 0\ \cdots\ 0]^T$ iz prirodnog bazisa prostora $\mathbb{R}^n$ na sledeći način

$$\boldsymbol{w}^*\boldsymbol{w} = \|\boldsymbol{x}\|_E^2 = 1, \quad w_1 \geq 0, \tag{3.4.6}$$

$$P\boldsymbol{e}_1 = \boldsymbol{x}. \tag{3.4.7}$$

Matrica P ima oblik

$$P = \begin{bmatrix} 1-2w_1\overline{w}_1 & -2w_1\overline{w}_2 & \dots & -2w_1\overline{w}_n \\ -2w_2\overline{w}_1 & 1-2w_2\overline{w}_2 & & -2w_2\overline{w}_n \\ \vdots & & & \\ -2w_n\overline{w}_1 & -2w_n\overline{w}_2 & & 1-2w_n\overline{w}_n \end{bmatrix}. \tag{3.4.8}$$

Primetimo da je $P^* = P$ što znači da je i matrica P hermitska. Štaviše, s obzirom na (3.4.6), direktnim množenjem vidimo da je $P^*P = P^2 = I$ pa zaključujemo da je matrica P unitarna (videti definiciju 3.4.4, odeljak 2.3.4).

Na osnovu (3.4.7) nalazimo koordinate vektora w. Tako iz $1-2w_1\overline{w}_1 = x_1$ i $-2w_k\overline{w}_1 = x_k$ $(k = 2,\dots,n)$ sleduje

$$w_1 = \sqrt{\frac{1-x_1}{2}} \quad \text{i} \quad w_k = -\frac{x_k}{2w_1} \quad (k = 2,\dots,n).$$

Primetimo da je $\overline{w}_1 = w_1 > 0$.

Sada, na osnovu (3.4.7) i $A\boldsymbol{x} = \lambda\boldsymbol{x}$ nalazimo da je

$$AP\boldsymbol{e}_1 = P\boldsymbol{e}_1,$$

odakle zaključujemo da je

$$P^*AP\boldsymbol{e}_1 = \lambda\boldsymbol{e}_1,$$

tj. da je $\boldsymbol{e}_1$ sopstveni vektor matrice $B = P^*AP = PAP$. Primetimo, takodje, da je prva kolona u matrici B upravo vektor $\lambda\boldsymbol{e}_1$, tj.

$$B = \begin{bmatrix} \lambda & b_{12} & b_{13} & \dots & b_{1n} \\ 0 & b_{22} & b_{23} & & b_{2n} \\ 0 & b_{32} & b_{33} & & b_{3n} \\ \vdots & & & & \\ 0 & b_{n2} & b_{n3} & & b_{nn} \end{bmatrix} = \left[\begin{array}{c|c} \lambda & \boldsymbol{b}_{n-1}^T \\ \hline \mathbf{0}_{n-1} & A_{n-1} \end{array}\right],$$

gde smo sa A_{n-1} označili matricu reda $n-1$ koja se poklapa sa ograđenim blokom, $\mathbf{0}_{n-1}$ je nula vektor reda $n-1$, i najzad, $\boldsymbol{b}_{n-1}^T = [b_{12}\ b_{13}\ \cdots\ b_{1n}]^T$.

S obzirom da je matrica B slična (kažemo još i unitarno slična) sa matricom A, zaključujemo da je

$$\mathrm{Sp}\,(A_{n-1}) = \mathrm{Sp}\,(A_n) \setminus \{\lambda\} \quad (A_n = A).$$

Da bismo dobili matricu A_{n-2} postupićemo na sličan način. Umesto postojeće matrice P koristimo matricu

$$P_1 = \left[\begin{array}{c|c} 1 & \mathbf{0}_{n-1}^T \\ \hline \mathbf{0}_{n-1} & Q \end{array}\right],$$

gde je Q matrica reda $n-1$ oblika (3.4.5) i zadovoljava uslove (3.4.6) i (3.4.7) u odnosu na sopstveni vektor $\boldsymbol{y}$ i sopstvenu vrednost μ matrice A_{n-1}. Kako je $P_1^{-1} = P_1^* = P_1$ zaključujemo da je i matrica P_1 unitarna.

Sada matrica $C = P_1 B P_1 = P_1 P A P P_1$ ima oblik

$$C = \begin{bmatrix} \lambda & c_{12} & c_{13} & \dots & c_{1n} \\ 0 & \mu & c_{23} & & c_{2n} \\ 0 & 0 & c_{33} & & c_{3n} \\ \vdots & & & & \\ 0 & 0 & c_{n3} & & c_{nn} \end{bmatrix} = \left[\begin{array}{cc|ccc} \lambda & c_{12} & c_{13} & \cdots & c_{1n} \\ 0 & \mu & c_{23} & & c_{2n} \\ \hline 0 & 0 & & & \\ \vdots & & & A_{n-2} & \\ 0 & 0 & & & \end{array}\right],$$

gde je A_{n-2} matrica reda $n-2$. Nastavljajući ovakav postupak dolazimo do gornje trougaone matrice koja je unitarno slična sa polaznom matricom A. Imajući u vidu da je matrica A hermitska zaključujemo da je ona unitarno slična sa dijagonalnom matricom.

Izloženi postupak zahteva pre svakog koraka određivanje jedne sopstvene vrednosti i njoj odgovarajućeg sopstvenog vektora, što se može učiniti nekim od prethodno izloženih metoda. Tako, pre prvog koraka treba odrediti sopstvenu

vrednost λ i sopstveni vektor $\boldsymbol{x}$ matrice A, pre drugog koraka sopstvenu vrednost μ i sopstveni vektor $\boldsymbol{y}$ matrice $(n-1)$-og reda A_{n-1}, itd.

Jasno je da su sopstvene vrednosti matrice A dijagonalni elementi dobijene trougaone matrice, tj. $\lambda_1 = \lambda$, $\lambda_2 = \mu$, itd. Ostaje pitanje šta je sa sopstvenim vektorima matrice A? Jasno je za sada da je $\boldsymbol{x}_1 = \boldsymbol{x}$. Pokazaćemo kako se na osnovu dobijenih rezultata može naći drugi sopstveni vektor matrice A.

Neka su koordinate sopstvenog vektora $\boldsymbol{y}$ redom $y_2, \ldots, y_n$. U cilju nalaženja, najpre, sopstvenog vektora $\boldsymbol{y}$ matrice B stavimo $\boldsymbol{y}' = [y_i \; y_2 \; \cdots \; y_n]^T$ i pokušajmo da odredimo y_1.

Kako je

$$B\boldsymbol{y}' = \left[\begin{array}{c|c} \lambda & \boldsymbol{b}_{n-1}^T \\ - \; - & - \\ \mathbf{0}_{n-1} & A_{n-1} \end{array}\right] \cdot \left[\begin{array}{c} y_1 \\ -- \\ y_2 \end{array}\right] = \left[\begin{array}{c} \lambda y_1 + \boldsymbol{b}_{n-1}^T \boldsymbol{y} \\ - - - - - - - \\ A_{n-1}\boldsymbol{y} \end{array}\right],$$

tj.

$$B\boldsymbol{y}' = \left[\begin{array}{c} \lambda y_1 + \boldsymbol{b}_{n-1}^T \boldsymbol{y} \\ - - - - - - - \\ \mu \boldsymbol{y} \end{array}\right]$$

izlazi da mora biti

$$y_1 + \boldsymbol{b}_{n-1}^T \boldsymbol{y} = y_1.$$

Ako je $\lambda \neq \mu$, na osnovu prethodne jednakosti, dobijamo

$$y_1 = \frac{1}{\lambda - \mu} \boldsymbol{b}_{n-1}^T \boldsymbol{y} = \frac{1}{\lambda - \mu}(b_{12}y_2 + \cdots + b_{1n}y_n).$$

Sada jednostavno nalazimo sopstveni vektor $\boldsymbol{x}_2$ matrice A koji odgovara sopstvenoj vrednosti $\lambda_2 \neq \mu$. Zaista, kako je $PAP\boldsymbol{y}' = \mu \boldsymbol{y}$, tj. $A(P\boldsymbol{y}') = \mu(P\boldsymbol{y}')$ zaključujemo da je $\boldsymbol{x}_2 = P\boldsymbol{y}'$.

Slično se može postupiti i kod određivanja ostalih sopstvenih vektora.

4.3.5 JACOBIev metod

Ovim odeljkom počinjemo izlaganje metoda za rešavanje kompletnog problema sopstvenih vrednosti, tj. za određivanje svih sopstvenih vrednosti i odgovarajućih sopstvenih vektora. Metod koji izlažemo u ovom odeljku potiče od JACOBIa (1846), a primenljiv je za hermitske matrice. Metod se zasniva na transformaciji hermitske matrice A na dijagonalnu matricu D, čije su sopstvene vrednosti

dijagonalni elementi. Na osnovu teoreme 3.5.2 (odeljak 2.3.5) postoji unitarna matrica H takva da je

$$H^*AH = D. \tag{3.5.1}$$

Kako je $H^* = H^{-1}$, na osnovu (3.5.1), matrice A i D su slične pa imaju iste sopstvene vrednosti.

Razmotrimo najpre slučaj transformacije

$$B = R^*AR, \tag{3.5.2}$$

gde je R proizvoljna unitarna matrica.

Teorema 3.5.1. *Matrice B i A iz* (3.5.2) *imaju jednake* SCHMIDT*ove norme, tj.* $\varepsilon(B) = \varepsilon(A)$.

Dokaz. Matrice A i B su slične. Neka su $\lambda_i \quad (i = 1,2,\ldots,n)$ njihove sopstvene vrednosti. Kako je

$$\varepsilon(A)^2 = \sum_{i,j\neq 1}^{n} |a_{ij}|^2 = \operatorname{tr}(A^*A) = \operatorname{tr}(A^2) = \sum_{i=1}^{n} \lambda_i^2,$$

zaključujemo da je $\varepsilon(B) = \varepsilon(A)$. $\square$

Osnovna ideja u JACOBIevom metodu je u konstrukciji niza sličnih matrica $\{A_k\}_{k\in\mathbb{N}}$, startujući od $A_1 = A$, sa strategijom minimizacije veličine $\operatorname{vd}(A_k)$ na svakom koraku, gde je sa $\operatorname{vd}(A_k)$ označena „norma“ vandijagonalnih elemenata matrice A, tj.

$$\operatorname{vd}(A) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|^2.$$

Primetimo da je $\operatorname{vd}(A) + \operatorname{d}(A) = \varepsilon(A)$, gde je $\operatorname{d}(A) = \sum\limits_{i=1}^{n} |a_{ij}|^2$.

Strategija minimizacije treba da obezbedi

$$\operatorname{vd}(A_{k+1}) \leq \operatorname{vd}(A_k) \quad \text{i} \quad \lim_{k\to+\infty} \operatorname{vd}(A_k) = 0.$$

Ostaje pitanje kako generisati niz $\{A_k\}_{k\in\mathbb{N}}$, tj. kakve unitarne matrice koristiti za transformaciju A_k u A_{k+1}.

Posmatrajmo tzv. matricu rotacije $R = R(p,q)$ čiji su elementi

$$r_{pp} = \mathrm{e}^{\mathrm{i}\alpha}\cos\theta, \qquad r_{pq} = \mathrm{e}^{\mathrm{i}\beta}\sin\theta,$$

$$r_{qp} = -\mathrm{e}^{-\mathrm{i}\beta}\sin\theta, \quad r_{qq} = \mathrm{e}^{\mathrm{i}\alpha}\cos\theta,$$

$$r_{ij} = \delta_{ij} \quad \text{(u ostalim slučajevima)},$$

gde su θ, α, β realni brojevi. Direktnim množenjem pokazujemo da je $R^*R = I$, tj. da je matrica rotacije unitarna. Obično ovu matricu nazivamo elementarnom matricom rotacije po uglu θ u ravni (p,q).

Mada je JACOBIev metod primenljiv za hermitske matrice, jednostavnosti radi, izložićemo ga za slučaj kada je A realna simetrična matrica. U prilog ovome ide i činjenica da se problem sopstvenih vrednosti za hermitsku matricu A reda n može svesti na problem sopstvenih vrednosti za jednu simetričnu matricu čiji je red $2n$. Naime, tada se A može razložiti na dve realne matrice S i K u obliku $A = S + \mathrm{i}K$, gde je S simetrična, a K kososimetrična matrica. Sopstvena vrednost λ i sopstveni vektor $\boldsymbol{v} = \boldsymbol{x} + \mathrm{i}\boldsymbol{y}$ matrice A zadovoljavaju jednakost

$$\begin{bmatrix} S & -K \\ K & S \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}, \tag{3.5.3}$$

pri čemu ako je λ prosta sopstvena vrednost matrice A, onda je ona dvostruka za matricu reda $2n$ koja se pojavljuje u (3.5.3).

Iz prethodno navedenih razloga na dalje pretpostavljamo da je A realna simetrična matrica. U matrici rotacije možemo tada uzeti $\alpha = \beta = 0$, tako da je ova matrica sada ortogonalna. Prema tome, imamo

$$r_{pp} = \cos\theta = c, \qquad r_{pq} = \sin\theta = s,$$

$$r_{qp} = -\sin\theta = -s, \quad r_{qq} = \cos\theta = c,$$

$$r_{ij} = \delta_{ij} \quad \text{(u ostalim slučajevima)}.$$

Primetimo da će se elementi matrice B u transformaciji

$$B = R^T A R \tag{3.5.4}$$

poklapati sa odgovarajućim elementima matrica A, sem onih koji se nalaze u p-toj i q-toj vrsti i koloni. Štaviše, imamo

$$\begin{bmatrix} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{bmatrix} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{bmatrix} \begin{bmatrix} c & s \\ -s & c \end{bmatrix}, \tag{3.5.5}$$

odakle, s obzirom na tvrđenje teoreme 3.5.1, nalazimo

$$b_{pp}^2 + 2b_{pq}^2 + b_{qq}^2 = a_{pp}^2 + 2a_{pq}^2 + a_{qq}^2.$$

Kako je, opet na osnovu teoreme 3.5.1, $\varepsilon(A) = \varepsilon(B)$ imamo

$$\mathrm{vd}(B) = \varepsilon(B) - \mathrm{d}(B) = \varepsilon(A) - \mathrm{d}(B),$$

tj.

$$\begin{aligned}\mathrm{vd}(B) &= \mathrm{vd}(A) + \mathrm{d}(A) - \mathrm{d}(B)\\ &= \mathrm{vd}(A) + a_{pp}^2 + a_{qq}^2 - (b_{pp}^2 + b_{qq}^2)\\ &= \mathrm{vd}(A) + 2b_{pq}^2 - 2a_{pq}^2,\end{aligned}$$

odakle vidimo da će $\mathrm{vd}(B)$ biti minimizirano ako je $b_{pq} = 0$. Iz ovog uslova odredićemo $c = \cos\theta$ i $s = \sin\theta$ u matrici rotacije.

Kako je na osnovu (3.5.5)

$$b_{pq} = (c^2 - s^2)a_{pq} + cs(a_{pp} - a_{qq}),$$

nalazimo

$$(\cos^2\theta - \sin^2\theta)a_{pq} + \cos\theta\sin\theta(a_{pp} - a_{qq}) = 0,$$

tj.

$$r = \cot(2\theta) = \frac{a_{qq} - a_{pp}}{2a_{pq}} \quad (a_{pq} \neq 0).$$

Ako je $a_{pq} = 0$, imamo $c = 1$ i $s = 0$, što znači da je R jedinična matrica. Stavimo $\tan\theta = t$. Kako je $\cot(2\theta) = (1-t^2)/2t = \tau$, za određivanje t treba rešiti kvadratnu jednačinu

$$t^2 + 2\tau t - 1 = 0.$$

Obično se uzima rešenje ove jeanaćine sa manjim modulom, tj.

$$t = \frac{\operatorname{sgn}\tau}{|\tau| + \sqrt{\tau^2 + 1}},$$

što obezbeđuje uslov da je ugao rotacije $|\theta| \leq \pi/4$. Tada su elementi matrice rotacije

$$c = \frac{1}{\sqrt{1+t^2}} \quad \text{i} \quad s = tc.$$

Sa ovako određenom matricom rotacije R norma vandijagonalnih elemenata matrice B u (3.5.4) se minimizira, pri čemu se element na mestu (p,q) anulira. Ovakav postupak se ponavlja, tj. generiše se niz sličnih matrica

$$A_{k+1} = R_k^T A_k R_k \quad (k = 1,2,\dots), \tag{3.5.6}$$

startujući sa $A_1 = A$, pri čemu još ostaje otvoreno pitanje kako birati matricu rotacije $R_k = R_k(p,q)$, tj. parove (p,q), na svakom koraku u (3.5.6). Kod klasičnog JACOBIevog metoda za (p,q) se uzima pozicija dominantnog elementa, tj. vandijagonalnog elementa matrice $A_k = [a_{ij}^{(k)}]$ sa najvećim modulom. Dakle, (p,q) određujemo iz uslova

$$|a_{pq}^{(k)}| = \max_{i<j} |a_{ij}^{(k)}|.$$

Elementi matrice A_k koji se menjaju pri transformaciji (3.5.6) su samo oni koji se nalaze u p-toj i q-toj vrsti i koloni:

$$\left.\begin{aligned} a_{ip}^{(k+1)} &= a_{pi}^{(k+1)} = ca_{ip}^{(k)} - sa_{iq}^{(k)} \\ a_{iq}^{(k+1)} &= a_{qi}^{(k+1)} = sa_{ip}^{(k)} + ca_{iq}^{(k)} \end{aligned}\right\} \quad (i \neq p,q),$$

$$a_{pp}^{(k+1)} = c^2 a_{pp}^{(k)} - 2csa_{pq}^{(k)} + s^2 a_{qq}^{(k)},$$

$$a_{qq}^{(k+1)} = s^2 a_{pp}^{(k)} + 2csa_{pq}^{(k)} + c^2 a_{qq}^{(k)},$$

$$a_{pq}^{(k+1)} = a_{qp}^{(k+1)} = 0.$$

Za ovako konstruisani metod imamo $A_k \to D = [\lambda_i \delta_{ij}]$, kada $k \to +\infty$. Zaista, s obzirom na nejednakost

$$2(a_{pq}^{(k)})^2 \geq \frac{1}{N}\mathrm{vd}(A_k) \qquad \left(N = \frac{1}{2}n(n-1)\right),$$

imamo

$$\mathrm{vd}(A_{k+1}) = \mathrm{vd}(A_k) - 2(a_{pq}^{(k)})^2 \leq \left(1 - \frac{1}{N}\right)\mathrm{vd}(A_k),$$

tj.

$$\mathrm{vd}(A_k) \leq \left(1 - \frac{1}{N}\right)^{k-1} \mathrm{vd}(A),$$

odakle sledi $\lim\limits_{k\to+\infty} \mathrm{vd}(A_k) = 0$, odnosno $\lim\limits_{k\to+\infty} A_k = D$. Konvergencija JACOBIevog metoda je kvadratna u smislu da postoji konstanta M (> 0) takva da je, za dovoljno veliko k,

$$\mathrm{vd}(A_{k+N}) \leq M(\mathrm{vd}(A_k))^2.$$

Pomoću (3.5.6) mozemo pisati

$$A_{k+1} = H_k^T A H_k, \tag{3.5.7}$$

gde je $H_k = R_1 R_2 \cdots R_k$. Za dovoljno veliko k, matrica H_k se može tretirati kao dovoljno dobra aproksimacija za unitarnu matricu H koja se javlja u (3.5.1). JACOBIev iterativni proces (3.5.6) se obično prekida kada je $\mathrm{vd}(A_{k+1}) \leq \delta^2$, gde je δ unapred data tačnost.

U klasičnom JACOBIevom metodu dosta se vremena može utrošiti na određivanje pozicije dominantnog elementa. Jedna modifikacija klasičnog JACOBIevog metoda, poznata kao ciklični JACOBIev metod, uzima za (p,q) redom parove:

$$(1,2),(1,3),\ldots,(1,n);(2,3),\ldots,(2,n);\ldots;(n-1,n);(1,2),\ldots\ .$$

Može se pokazati da ovaj metod ima, takođe, kvadratnu konvergenciju.

Ako je k dovoljno veliko u (3.5.7), tako da je zadovoljen kriterijum za zaustavljanje procesa, tada možemo smatrati da je

$$H_k^T A H_k = A_{k+1} = D = [\lambda_i \delta_{ij}]_{n\times n},$$

odakle sleduje

$$AH_k = H_k D,$$

što znači da su kolone matrice H_k sopstveni vektori matrice A.

Štaviše, ovi sopstveni vektori čine ortogonalni skup. Matrica $H_k = [h_{ij}^{(k)}]_{n\times n}$ se može generisati rekurzivno pomoću matrice rotacije

$$H_k = H_{k-1} R_k(p,q) \quad (H_0 = I).$$

U skalarnom obliku imamo

$$\left.\begin{aligned} h_{ip}^{(k+1)} &= c h_{ip}^{(k)} - s h_{iq}^{(k)} \\ h_{iq}^{(k+1)} &= s h_{ip}^{(k)} - c h_{iq}^{(k)} \end{aligned}\right\} \quad (i = 1,2,\ldots,n),$$

$$h_{ij}^{(k+1)} = h_{ij}^{(k)} \quad \text{(u ostalim slučajevima)}.$$

Primer 3.5.1. Na matricu A iz primera 3.4.1 primenićemo klasičan JACOBIev metod. Stavimo

$$A_1 = A = \begin{bmatrix} 4 & 1 & 4 \\ 1 & 10 & 1 \\ 4 & 1 & 10 \end{bmatrix}.$$

Dominantni element je na poziciji $(p,q) = (1,3)$. Tada je

$$\tau = \frac{a_{33} - a_{11}}{2a_{13}} = \frac{3}{4}, \quad t = \frac{1}{2}, \quad c = c_1 = \frac{2}{\sqrt{5}}, \quad s = s_1 = \frac{1}{\sqrt{5}},$$

pa je

$$R_1 = R_1(1,3) = \begin{bmatrix} 2/\sqrt{5} & 0 & 1/\sqrt{5} \\ 0 & 1 & 0 \\ -1/\sqrt{5} & 0 & 2/\sqrt{5} \end{bmatrix},$$

$$H_1 = R_1, \quad A_2 = \begin{bmatrix} 2 & 1/\sqrt{5} & 0 \\ 1/\sqrt{5} & 10 & 3/\sqrt{5} \\ 0 & 3/\sqrt{5} & 12 \end{bmatrix}.$$

Nastavljajući ovaj postupak dobijamo sledeće rezultate.

a) Za elementarne matrice rotacije:

k	(p,q)	c_k	s_k
2	(2,3)	0.89376	0.44855
3	(1,2)	0.99852	0.05431
4	(1,3)	0.99982	0.01872

b) Za nizove $\{H_k\}$ i $\{A_k\}$ (matrice A_k su simetrične; elementi donjeg trougla nisu navedeni):

$$H_2 = \begin{bmatrix} 0.89443 & -0.20060 & 0.39970 \\ 0. & 0.89376 & 0.44855 \\ -0.44721 & -0.40119 & 0.79940 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 2. & 0.39970 & 0.20060 \\ & 9.32668 & 0. \\ & & 12.67332 \end{bmatrix},$$

$$H_3 = \begin{bmatrix} 0.90400 & -0.15172 & 0.39970 \\ -0.04854 & 0.89244 & 0.44855 \\ -0.42476 & -0.42489 & 0.79940 \end{bmatrix},$$

$$A_4 = \begin{bmatrix} 1.97826 & 0. & 0.20030 \\ & 9.34842 & 0.01089 \\ & & 12.67332 \end{bmatrix},$$

$$H_4 = \begin{bmatrix} 0.89636 & -0.15172 & 0.41655 \\ -0.05693 & 0.89244 & 0.44756 \\ -0.43965 & -0.42489 & 0.79131 \end{bmatrix},$$

$$A_5 = \begin{bmatrix} 1.97451 & -0.00020 & 0. \\ & 9.34842 & 0.01089 \\ & & 12.67707 \end{bmatrix}.$$

Svi rezultati su zaokrugljeni na 5 decimala.

Na osnovu A_5 i H_4 imamo da su sopstvene vrednosti matrice A

$$\lambda_1 \cong 1.97451, \quad \lambda_2 \cong 9.34842, \quad \lambda_3 \cong 12.67707,$$

i njima odgovarajući sopstveni vektori

$$\boldsymbol{x}_1 \cong \begin{bmatrix} 0.89636 \\ -0.05693 \\ -0.43965 \end{bmatrix}, \quad \boldsymbol{x}_2 \cong \begin{bmatrix} -0.15172 \\ 0.89244 \\ -0.42489 \end{bmatrix}, \quad \boldsymbol{x}_3 \cong \begin{bmatrix} 0.41655 \\ 0.44756 \\ 0.79131 \end{bmatrix}.$$

Primetimo da smo u primeru 3.4.1 odredili, metodom inverzne matrice, sopstvenu vrednost λ_2 i odgovarajući sopstveni vektor $\boldsymbol{x}_2$ (normiran u odnosu na koordinatu sa najvećim modulom). $\triangle$

Napomena 3.5.1. Ukoliko se ne zahteva rešavanje kompletnog problema sopstvenih vrednosti, matricu H_k nije potrebno generisati. Na primer, ako je potreban samo jedan sopstveni vektor, recimo $\boldsymbol{x}_m$, tada se on može jednostavno dobiti primenom rotacija na vektor $\boldsymbol{e}_m$, čija je m-ta koordinata jednaka jedinici, a sve ostale su jednake nuli.

4.3.6 GIVENSov i HOUSEHOLDERov metod

Kao što smo videli u prethodnom odeljku, JACOBIv metod transformiše simetričnu (ili, uopšte hermitsku) matricu na dijagonalnu posle beskonačno koraka. Vandijagonalni elementi (a_{pq} i a_{qp}) koji se anuliraju na određenom koraku primene JACOBIevog metoda, mogu u kasnijim koracima da postanu takvi da znatno odstupaju od nule, što je posebno izraženo kada je red matrice visok. Ova

činjenica usporava algoritam. U ovom odeljku izložicemo dva metoda kod kojih je navedeni nedostatak uklonjen. Prvi od ovih metoda razvio je W. J. GIVENS[151], a drugi A. S. HOUSEHOLDER[152]. Metodi su takvi da kroz konačan broj koraka transformišu polaznu realnu simetričnu matricu na simetričnu trodijagonalnu matricu. U narednom odeljku razmatraćemo problem sopstvenih vrednosti za simetrične trodijagonalne matrice. I GIVENSov i HOUSEHOLDERov metod se mogu jednostavno preneti na hermitske matrice, štaviše, njihova primena na opšte matrice dovodi do redukcije matrice na tzv. HESSENBERGovu formu ($a_{ij} = 0$ za $i \geq j+2$).

Kao i u prethodnom odeljku, i ovde ćemo izložiti pomenute metode za slučaj kada je data matrica A realna i simetrična.

GIVENSov metod. Ovaj metod redukcije [31], [32] (videti, takođe, [33], [34]) se zasniva na sukcesivnoj primeni tzv. GIVENSove transformacije, pri čemu se posle konačnog broja rotacija matrica A transformiše na trodijagonalnu matricu.

Za elemente matrice $A = [a_{ij}]_{n\times n}$ za čije indekse važi nejednakost $|i-j| > 1$ koristićemo termin *vantridijagonalni elementi*. Za njih ćemo uvesti takvo uredjenje da ćemo reći da je a_{ij} s-ti vantridijagonalni element ako je par (i,j) s-ti član cikličnog indeksnog niza

$$(1,3),(1,4),\ldots,(1,n),(2,4),(2,5),\ldots,(2,n),\ldots,(n-2,n).$$

Primetimo da ovaj niz sadrži ukupno

$$M = (n-2)+(n-1)+\cdots+1 = \frac{1}{2}(n-2)(n-1)$$

parova.

Neka je $A_1 = A$ i

$$A_{k+1} = G_k^T A_k G_k \quad (k = 1,2,\ldots,M), \tag{3.6.1}$$

gde su matrice G_k, tzv. dvodimenzionalne rotacije, izabrane tako da prvih k vantridijagonalnih elemenata matrice A_{k+1} bude jednako nuli.

Teorema 3.6.1. *Neka je A realna simetrična matrica, niz $A_k = [a_{ij}^{(k)}]_{n\times n}$ definisan pomoću* (3.6.1) *i neka je $(p-1,q)$ k-ti par cikličnog indeksnog niza vantridijagonalnih elemenata. Ako se matrica rotacije $G_k = [g_{ij}^{(k)}]_{n\times n}$ definiše pomoću*

[151] JAMES WALLACE GIVENS, JR. (1910 – 1993), američki matematičar i jedan od pionira u razvoju kompjuterskih nauka.

[152] ALSTON SCOTT HOUSEHOLDER (1904 – 1993), američki matematičar sa doprinosima u numeričkoj analizi i matematičkoj biologiji.

$$g_{pp}^{(k)} = g_{qq}^{(k)} = c, \quad g_{pq}^{(k)} = -g_{qp}^{(k)} = -s,$$

$$g_{ij}^{(k)} = \delta_{ij} \quad (\textit{u ostalim slučajevima}),$$

gde su

$$c = \cos\theta = \frac{1}{S} a_{p-1,p}^{(k)}, \quad s = \sin\theta = \frac{1}{S} a_{p-1,q}^{(k)},$$

$$S = \sqrt{(a_{p-1,p}^{(k)})^2 + (a_{p-1,q}^{(k)})^2},$$

ili sa $G_k = I$, *ako je* $a_{p-1,q}^{(k)} = 0$, *imamo:*

(a) *matrice* A_{k+1} *su realne i simetrične;*

(b) *prvih k vantridijagonalnih elemenata matrice* A_{k+1} *su nule* $(k = 1, \ldots, M)$*;*

(c) *matrica* A_{M+1} *je trodijagonalna.*

Dokaz prethodne GIVENSove teoreme se može dati matematičkom indukcijom. Primetimo jednu suštinsku razliku između GIVENSovog i JACOBIevog metoda. Naime, kod JACOBIevog metoda sa matricom rotacije $R_k = R_k(p,q)$ se anulira samo element na poziciji (p,q). U GIVENSovom metodu imamo da je u matrici A_k prvih $k-1$ vantridijagonalnih elemenata jednako nuli. Sa GIVENSovom rotacijom $G_k = G_k(p,q)$, odgovarajući elementi u matrici A_{k+1} ostaju nepromenjeni, tj. jednaki nuli. Kako je, međutim,

$$a_{p-1,q}^{(k+1)} = -a_{p-1,p}^{(k)} \sin\theta + a_{p-1,q}^{(k)} \cos\theta,$$

izborom

$$\tan\theta = \frac{a_{p-1,q}^{(k)}}{a_{p-1,p}^{(k)}}$$

imamo da je i k-ti element jednak nuli, tj. $a_{p-1,q}^{(k+1)} = 0$. Upravo ovakav izbor $\tan\theta$ je uzet za određivanje elemenata c i s u matrici rotacije G_k.

Na osnovu (3.6.1) imamo

$$\begin{aligned}
a_{pp}^{(k+1)} &= c^2 a_{pp}^{(k)} + 2cs\, a_{pq}^{(k)} + s^2 a_{qq}^{(k)}, \\
a_{pq}^{(k+1)} &= a_{qp}^{(k+1)} = (c^2 - s^2) a_{pq}^{(k)} + cs(a_{qq}^{(k)} + a_{pp}^{(k)}), \\
a_{qq}^{(k+1)} &= s^2 a_{pp}^{(k)} - 2cs\, a_{pq}^{(k)} + c^2 a_{qq}^{(k)},
\end{aligned}$$

$$\left.\begin{aligned} a_{ip}^{(k+1)} &= a_{pi}^{(k+1)} = ca_{ip}^{(k)} + sa_{iq}^{(k)} \\ a_{iq}^{(k+1)} &= a_{qi}^{(k+1)} = -sa_{ip}^{(k)} + ca_{iq}^{(k)} \end{aligned}\right\} \quad (i \neq p,q),$$
$$a_{ij}^{(k+1)} = a_{ji}^{(k+1)} = a_{ij}^{(k)} \quad \text{(u ostalim slučajevima).}$$

Primetimo da GIVENSov algoritam zahteva ukupno M korenovanja i aproksimativno $4n^3/3$ množenja.

Primer 3.6.1. Primenićemo GIVENSov metod na redukciju matrice

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 3 & 6 & 10 \\ 1 & 4 & 10 & 20 \end{bmatrix}$$

na trodijagonalni oblik. Na osnovu prethodnog, za to je potrebno $M = 3$ koraka. GIVENSove rotacije su, pritom, određene pomoću:

k	(p,q)	c_k	q_k
1	(2,3)	0.707107	0.707107
2	(2,4)	0.816496	0.577350
3	(3,4)	0.397360	0.917663

dok je odgovarajući niz simetričnih matrica A_k $(k = 2,3,4)$ (elementi donjeg trougla nisu navedeni):

$$A_2 = \begin{bmatrix} 1. & 1.414214 & 0. & 1. \\ & 7. & 2. & 9.899495 \\ & & 1. & 4.242641 \\ & & & 20. \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 1. & 1.732051 & 0. & 0. \\ & 20.666667 & 4.082493 & 9.428090 \\ & & 1. & 2.309401 \\ & & & 6.333333 \end{bmatrix},$$

$$A_4 = \begin{bmatrix} 1. & 1.732051 & 0. & 0. \\ & 20.666667 & 10.274023 & 0. \\ & & 7.175439 & 0.364642 \\ & & & 0.1578957895 \end{bmatrix}.$$

Svi rezultati su zaokrugljeni na šest decimala. $\triangle$

HOUSEHOLDERov metod. Ovaj metod [39] (videti, takođe, [40], [41]) se zasniva na korišćenju niza ortogonalnih transformacija oblika

$$H = I - 2\boldsymbol{w}\boldsymbol{w}^T, \quad \boldsymbol{w}^T\boldsymbol{w} = 1$$

sa pogodno izabranim vektorima $\boldsymbol{w}$. Nije teško pokazati da su ove matrice ortogonalne (videti odeljak 4.3.4, gde je razmatran opštiji slučaj).

Neka je A realna simetrična matrica reda n. Stavimo $A_1 = A$ i definišimo niz

$$A_{k+1} = H_k^T A_k H_k \quad (k = 1,2,\ldots,n-2), \tag{3.6.2}$$

gde je

$$H_k = I - 2\boldsymbol{w}\boldsymbol{w}^T. \tag{3.6.3}$$

Specijalni izbor vektora $\boldsymbol{w}$ na svakom koraku u HOUSEHOLDERovom metodu obezbeđuje da je matrica A_{n-1} trodijagonalna. Inače, sve matrice u nizu $\{A_k\}$ su realne i simetrične.

Strategija HOUSEHOLDERovog metoda je da se posle prvog koraka anuliraju vantridijagonalni elementi u prvoj vrsti (i koloni), posle drugog koraka vantridijagonalni elementi u drugoj vrsti (i koloni), itd. Pri ovome, prethodno anulrani elementi se ne menjaju. Dakle, za matricu $A_k = [a_{ij}^{(k)}]$ $(k = 2,\ldots,n-1)$ imamo

$$a_{ij}^{(k)} = 0 \qquad (1 \le i \le k-1 \wedge |i-j| > 1).$$

Da bi se ovakva transformacija obezbedila, za vektor $\boldsymbol{w}$ u (3.6.3) treba uzeti:

$$\boldsymbol{w} = \boldsymbol{0} \quad \text{ako je} \quad m_k = \sum_{j=k+2}^{n} (a_{kj}^{(k)})^2 = 0,$$

i

$$\boldsymbol{w} = \beta\boldsymbol{v} \quad \text{ako je} \quad m_k \neq 0.$$

Pri ovome, za koordinate vektora $\boldsymbol{v} = [v_1 \ \cdots \ v_n]^T$ treba uzeti

$$v_i = 0 \quad (i \le k), \quad v_{k+1} = 2Sy^2, \quad v_i = a_{ki}^{(k)} \quad (i \ge k+2),$$

gde su

$$S^2 = \sum_{j=k+1}^{n} (a_{kj}^{(k)})^2 \qquad \left(S = \operatorname{sgn}(a_{k,k+1}^{(k)})\sqrt{S^2} \ \text{ ako je } \ a_{k,k+1}^{(k)} \neq 0\right),$$

$$y = \frac{1}{2K}(S + a_{k,k+1}^{(k)}), \qquad 2K^2 = S^2 + a_{k,k+1}^{(k)} S, \qquad \beta = \frac{1}{2Sy}.$$

Određivanje matrice A_{k+1} u (3.6.2) može da se uprosti, s obzirom na činjenicu da je

$$\begin{aligned} A_{k+1} &= (I - 2\boldsymbol{w}\boldsymbol{w}^T)A_k(I - 2\boldsymbol{w}\boldsymbol{w}^T) \\ &= A_k - 2\boldsymbol{w}\boldsymbol{w}^T A_k - 2A_k\boldsymbol{w}\boldsymbol{w}^T + 4\boldsymbol{w}\boldsymbol{w}^T A_k \boldsymbol{w}\boldsymbol{w}^T, \end{aligned}$$

tj.

$$A_{k+1} = A_k - 2\beta^2(\boldsymbol{v}\boldsymbol{u}^T + \boldsymbol{u}\boldsymbol{v}^T),$$

gde su

$$\boldsymbol{u} = \boldsymbol{\xi} - a\boldsymbol{v}, \quad \boldsymbol{\xi} = A_k\boldsymbol{v}, \quad a = \beta^2 \boldsymbol{v}^T\boldsymbol{\xi}.$$

Primetimo da HOUSEHOLDERov metod zahteva $n-2$ korenovanja i aproksimativno $2n^3/3$ množenja, što je dva puta manje od broja množenja u GIVENSovom algoritmu.

Primer 3.6.2. Primenimo HOUSEHOLDERov metod na transformaciju matrice A iz prethodnog primera. Sada je potrebno samo dva koraka $(n-2=2)$.

Prvi korak $(k=1)$: Imamo $m_1 = 2 \neq 0$, $S^2 = 3$, $S = 1.732051$, $K = 1.538189$, $y = 0.888074$, $\beta = 0.325058$, $a = 10.479274$,

$$\boldsymbol{v} = \begin{bmatrix} 0. \\ 2.732051 \\ 1. \\ 1. \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} 4.732051 \\ 12.464102 \\ 24.196152 \\ 40.928203 \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} 4.732051 \\ -16.165808 \\ 13.716878 \\ 30.448929 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 1. & -1.732051 & 0. & 0. \\ & 20.666667 & -1.503206 & -10.163460 \\ & & 0.202565 & 0.666667 \\ & & & 7.130768 \end{bmatrix};$$

Drugi korak $(k=2)$: Sada imamo $m_2 = 103.295919 \neq 0$, $S^2 = 105.555556$, $S = -10.274023$, $K = 7.778160$, $y = -0.757070$, $\beta = 0.064283$, $a = 3.818322$,

$$\boldsymbol{v} = \begin{bmatrix} 0. \\ 0. \\ -11.777230 \\ -10.163460 \end{bmatrix}, \quad \boldsymbol{\xi} = \begin{bmatrix} 0. \\ 120.999532 \\ -9.161295 \\ -80.324767 \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} 0. \\ 120.999532 \\ 35.819733 \\ -41.507243 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 1. & -1.732051 & 0. & 0. \\ & 20.666667 & 10.274023 & 0. \\ & & 7.175439 & -0.364642 \\ & & & 0.157895 \end{bmatrix}.$$

Elementi donjeg trougla u simetričnim matricama A_2 i A_3 nisu navedeni. Primetimo da se dobijena trodijagonalna matrica razlikuje od one koja je dobijena GIVENSovim algoritmom. Naravno, ove dve matrice su slične, s obzirom na činjenicu da egzistira transformacija sličnosti sa dijagonalnom matricom $D = \text{diag}\,(1,-1,-1,1)$. $\triangle$

Napomenimo da se GIVENSovim i HOUSEHOLDERovim metodom hermitske matrice transformišu na hermitske trodijagonalne matrice. Štaviše, opšte kompleksne matrice se transformišu na HESSENBERGov oblik.

4.3.7 Problem sopstvenih vrednosti za simetrične trodijagonalne matrice

Neka je A realna simetrična trodijagonalna matrica reda n čije ćemo ne-nula elemente označiti sa

$$\begin{aligned} &a_{ii} = b_i \quad (i = 1,\ldots,n), \\ &a_{i,i-1} = a_{i-1,i} = c_i \quad (i = 2,\ldots,n). \end{aligned}$$

Sa $p_k(\lambda)$ označimo glavni minor reda k matrice $A - \lambda I$, tj.

$$p_k(\lambda) = \begin{vmatrix} b_1-\lambda & c_2 & & & \mathrm{O} \\ c_2 & b_2-\lambda & c_3 & & \\ & c_3 & b_3-\lambda & \ddots & \\ & & \ddots & \ddots & c_k \\ \mathrm{O} & & & c_k & b_k-\lambda \end{vmatrix}$$

i definišimo $p_0(\lambda) = 1$. Primetimo da je $p_1(\lambda) = b_1 - \lambda$.

Razvijanjem determinante $p_k(\lambda)$ po elementima poslednje vrste dobijamo

$$p_k(\lambda) = (b_k - \lambda)p_{k-1}(\lambda) - c_k^2\, p_{k-2}(\lambda).$$

Vrednost karakterističnog polinoma matrice A za vrednost λ, tj. $p_n(\lambda) = \det\,(A - \lambda I)$, možemo jednostavno odrediti, na osnovu prethodnog, korišćenjem tročlane rekurentne relacije

$$
\begin{aligned}
&p_k(\lambda) = (b_k - \lambda)p_{k-1}(\lambda) - c_k^2\, p_{k-2}(\lambda) \quad (k = 2, \ldots, n),\\
&p_0(\lambda) = 1, \quad p_1(\lambda) = b_1 - \lambda.
\end{aligned} \tag{3.7.1}
$$

Jedan jednostavan metod za određivanje sopstvenih vrednosti simetričnih trodijagonalnih matrica zasniva se na korišćenju rekurentne relacije (3.7.1), metoda polovljenja intervala (videti odeljak 5.1.5) i tvrđenju sledeće teoreme, koja se jednostavno dokazuje:

Teorema 3.7.1. *Neka su kod simetrične trodijagonalne matrice A reda n svi elementi $c_k \neq 0$. Tada važi:*

1° *nule svakog polinoma p_k $(k = 2, \ldots, n)$ su realne, različite i razdvojene nulama polinoma p_{k-1};*

2° *ako je $p_n(\lambda) \neq 0$, broj sopstvenih vrednosti matrice A manjih od λ je jednak broju promene znaka $s(\lambda)$ u nizu*

$$
p_0(\lambda), p_1(\lambda), \ldots, p_n(\lambda). \tag{3.7.2}
$$

Ako je neko $p_k(\lambda) = 0$, onda se na tom mestu u nizu (3.7.2) *može uzeti bilo koji znak, s obzirom da je $p_{k-1}(\lambda)p_{k+1}(\lambda) < 0$.*

Primetimo da u teoremi egzistira uslov $c_k \neq 0$ za svako $k = 2, \ldots, n$. Ako je, na primer, za neko $k = m$, $c_m = 0$, tada se problem pojednostavljuje jer se raspada na dva problema nižeg reda (m i $n - m$). Naime, matrica A postaje

$$
A = \begin{bmatrix} A' & O \\ O & A'' \end{bmatrix},
$$

gde su A' i A'' trodijagonalne simetrične matrice reda m i $n - m$ respektivno, i u tom slučaju je

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

Korišćenjem više vrednosti za λ moguće je sistematskom primenom teoreme 3.7.1, odrediti disjunktne intervale u kojima leže sopstvene vrednosti matrice λ. Dakle, ako nadjemo da je

$$
s(\lambda_1) = s_1 \quad \text{i} \quad s(\lambda_2) = s_2 = s_1 + 1 \quad (\lambda_1 < \lambda_2),
$$

na osnovu teoreme 3.7.1, imamo da u intervalu (λ_1, λ_2) leži samo jedna sopstvena vrednost matrice A. Tada se za njeno određivanje može iskoristiti jednostavni

metod polovljenja intervala (metod bisekcije), sužavajući ovaj polazni interval do zahtevane tačnosti (videti odeljak 5.1.5).

Za određivanje intervala u kojima leže sopstvene vrednosti može se koristiti i GERSHGORINova teorema (videti odeljak 4.3.1), na osnovu koje su ti intervali

$$\begin{aligned}
&[b_1 - |c_2|, b_1 + |c_2|],\\
&[b_i - |c_i| - |c_{i+1}|, b_i + |c_i| + |c_{i+1}|] \quad (i = 2, \ldots, n-1),\\
&[b_n - |c_n|, b_n + |c_n|].
\end{aligned}$$

Nažalost, ovi intervali nisu disjunktni, i u opštem slučaju ne sadrže samo po jednu sopstvenu vrednost matrice A.

Primer 3.7.1. Za datu matricu

$$A = \begin{bmatrix} 1 & 1 & & \\ 1 & 3 & 2 & \\ & 2 & 5 & 3 \\ & & 3 & 7 \end{bmatrix}$$

imamo

$$\begin{aligned}
&p_0(\lambda) = 1, \quad p_1(\lambda) = 1 - \lambda, \quad p_2(\lambda) = (3-\lambda)p_1(\lambda) - p_0(\lambda),\\
&p_3(\lambda) = (5-\lambda)p_2(\lambda) - 4p_1(\lambda), \quad p_4(\lambda) = (7-\lambda)p_3(\lambda) - 9p_2(\lambda).
\end{aligned}$$

Neka je $\lambda = 0$. Tada imamo $p_0(0) = 1$, $p_1(0) = 1$, $p_2(0) = 2$, $p_3(0) = 6$, $p_4(0) = 24$. Znaci u nizu (3.7.2) su redom $+++++$, što znači da nema promene znaka, tj. da je $s(0) = 0$. Prema teoremi 3.7.1, matrica A nema negativnih sopstvenih vrednosti, tj. ona je pozitivno definitna.

Uzimajući za λ redom vrednosti $1, 2, 4, 5, 7, 9, 10$ dobijamo sledeće rezultate: Na osnovu vrednosti za $s(\lambda)$ zaključujemo da se u intervalima $(0,1)$, $(1,2)$, $(4,5)$, $(9,10)$ nalazi po jedna sopstvena vrednost matrice A. Te sopstvene vrednosti na šest decimala su

$$\lambda_1 \cong 0.322548, \quad \lambda_2 \cong 1.745761, \quad \lambda_3 \cong 4.536620, \quad \lambda_4 \cong 9.395071.$$

Napomenimo da su ovo nule LAGUERREovog[153] polinoma L_4. $\triangle$

[153] EDMOND NICOLAS LAGUERRE (1834 – 1886), poznati francuski matematičar.

Tabela 3.7.1.

λ	$p_0(\lambda)$	$p_1(\lambda)$	$p_2(\lambda)$	$p_3(\lambda)$	$p_4(\lambda)$	$s(\lambda)$
1	1	0	−1	−4	−15	1
2	1	−1	−2	−2	8	2
4	1	−3	2	14	24	2
5	1	−4	7	16	−31	3
7	1	−6	23	−22	−207	3
9	1	−8	47	−156	−111	3
10	1	−9	62	−274	264	4

Određivanje $p_k(\lambda)$ pomoću (3.7.1), u aritmetici sa pokretnom tačkom, je stabilno. Nažalost, i kod matrica ne tako visokog reda, vrednosti $p_k(\lambda)$, za k koje je blisko n, mogu da izađu izvan opsega brojeva, tj. da ove vrednosti postanu veće od maksimalnog broja koji može biti zapisan u memoriji računara, ili pak, manje od minimalnog broja, što se onda tretira kao nula. Ova činjenica ne dozvoljava normiranje vrednosti za $p_k(\lambda)$ jer se praktično javljaju obe vrste prekoračenja. Navedena teškoća dolazi posebno do izražaja kada matrica A ima bliske sopstvene vrednosti.

Jedan bolji pristup ovoj problematici ([10], [74]) je konstrukcija niza $q_k(\lambda)$, definisanog pomoću

$$q_k(\lambda) = p_k(\lambda)/p_{k-1}(\lambda) \quad (k = 1,\ldots,n).$$

Broj $s(\lambda)$ je sada broj negativnih članova niza $\{q_k(\lambda)\}$. Na osnovu (3.7.1) dobijamo rekurentnu relaciju za $q_k(\lambda)$,

$$q_k(\lambda) = b_k - \lambda - c_k^2/q_{k-1}(\lambda) \quad (k = 2,\ldots,n),$$

gde je $q_1(\lambda) = b_1 - \lambda$.

Jedan efikasan metod, tzv. QR algoritam, može se uspešno primeniti, u specijalnom slučaju, i na trodijagonalne matrice, o čemu će biti reči u narednom odeljku.

4.3.8 LR i QR algoritmi

Ovaj odeljak posvećujemo tzv. faktorizacionim metodima. Prvi takav algoritam za rešavanje problema sopstvenih vrednosti za proizvoljnu matricu A opisao je H. RUTISHAUSER[154] ([62]) 1958. godine, nazivajući ga LR transformacijom.

[154] HEINZ RUTISHAUSER (1918 – 1970), poznati švajcarski matematičar, jedan od pionira moderne numeričke matematike i kompjuterskih nauka.

Metod se sastoji u konstrukciji niza matrica $\{A_k\}_{k\in\mathbb{N}}$ startujući od $A_1 = A$, na sledeći način: Matrica A_k se faktorizuje na donju trougaonu matricu L_k sa jediničnom dijagonalom i gornju trougaonu matricu R_k, tj.

$$A_k = L_k R_k, \tag{3.8.1}$$

a zatim se naredni član niza određuje množenjem dobijenih faktora u obrnutom redosledu, tj.

$$A_{k+1} = R_k L_k.$$

Primetimo da su matrice A_{k+1} i A_k slične jer su povezane transformacijom sličnosti

$$A_{k+1} = L_k^{-1} A_k L_k. \tag{3.8.2}$$

Faktorizacija (3.8.1) se može izvesti GAUSSovim metodom eliminacije.

Ako stavimo

$$L^{(k)} = L_1 \cdots L_k \quad \text{i} \quad R^{(k)} = R_k \cdots R_1,$$

na osnovu (3.8.2), imamo

$$L^{(k)} A_{k+1} = A L^{(k)},$$

odakle je

$$L^{(k)} R^{(k)} = L^{(k-1)} A_k R^{(k-1)} = A L^{(k-1)} R^{(k-1)}.$$

Iterirajući poslednju jednakost dobijamo

$$L^{(k)} R^{(k)} = A^2 L^{(k-2)} R^{(k-2)} = \cdots = A^k,$$

što znači da je $L^{(k)} R^{(k)}$ faktorizacija matrice A^k. Koristeći ove činjenice, RUTISHAUSER [62] (videti, takođe, [73]) je pokazao da pod određenim uslovima niz matrica $\{A_k\}$ konvergira ka nekoj gornje trougaonoj matrici, čiji elementi na glavnoj dijagonali daju sopstvene vrednosti matrice A. Obično se LR metod primenjuje na matrice prethodno svedene na gornju HESSENBERGovu formu ($a_{ij} = 0$ za $i \geq j+2$). Ako je nekim od metoda opšta matrica svedena na donju HESSENBERGovu formu, LR metod primenjujemo na transponovanu matricu koja ima iste sopstvene vrednosti. Sve matrice u nizu $\{A_k\}$ imaju HESSENBERGov oblik. Ubrzavanje konvergencije niza $\{A_k\}$ može biti učinjeno uvođenjem pogodnog pomeranja (šifta) p_k, tako da umesto A_k, faktorizujemo $B_k = A_k - p_k I = L_k R_k$, pri čemu je, dalje, $A_{k+1} = p_k I + R_k L_k$.

Nažalost, LR algoritam ima više nedostataka (videti monografiju WILKINSONa [73]). Na primer, faktorizacija ne egzistira za svaku matricu. Mnogo bolji

faktorizacioni metod razvili su nezavisno FRANCIS[155] [30] i KUBLANOVSKAJA[156] [50], u kome je matrica L zamenjena sa unitarnom matricom Q. Tako se dobija QR algoritam[157] definisan pomoću

$$A_k = Q_k R_k, \quad A_{k+1} = R_k Q_k \quad (k = 1, 2, \dots), \tag{3.8.3}$$

startujući od $A_1 = A$. Primetimo da je $A_{k+1} = Q_k^* A_k Q_k$.

Ako stavimo

$$Q^{(k)} = Q_1 \cdots Q_k \quad \text{i} \quad R^{(k)} = R_k \cdots R_1, \tag{3.8.4}$$

slično kao kod LR metoda, nalazimo

$$Q^{(k)} A_{k+1} = A Q^{(k)} \quad \text{i} \quad Q^{(k)} R^{(k)} = A^k. \tag{3.8.5}$$

Teorema 3.8.1. *Ako je matrica A regularna, tada egzistira dekompozicija $A = QR$, gde je Q unitarna, a R gornje trougaona matrica. Štaviše, ako su dijagonalni elementi matrice R pozitivni, dekompozicija je jedinstvena.*

QR faktorizacija (3.8.3) se može izvesti korišćenjem unitarnih matrica oblika $I - 2\boldsymbol{w}\boldsymbol{w}^*$. Tako, u cilju transformacije A_k na R_k, tj. redukcije kolona u A_k, imamo

$$(I - 2\boldsymbol{w}_{n-1}\boldsymbol{w}_{n-1}^*) \cdots (I - 2\boldsymbol{w}_1 \boldsymbol{w}_1^*) A_k = R_k. \tag{3.8.6}$$

Matrica Q_k je tada

$$Q_k = (I - 2\boldsymbol{w}_1 \boldsymbol{w}_1^*) \cdots (I - 2\boldsymbol{w}_{n-1}\boldsymbol{w}_{n-1}^*). \tag{3.8.7}$$

QR algoritam je efikasan ako polazna matrica ima (gornju) HESSENBERGovu formu. Tada se prethodno pomenute unitarne matrice svode na dvodimenzionalne rotacije. Sve matrice A_k imaju HESSENBERGovu formu. Dakle, problem sopstvenih vrednosti za opštu matricu je najpogodnije rešavati kroz dva koraka. Najpre, svesti matricu na HESSENBERGov oblik, a zatim primeniti QR algoritam.

U specijalnom slučaju kada je polazna matrica trodijagonalna, matrice A_k u QR algoritmu su takođe trodijagonalne. U tom slučaju, uz korišćenje pogodno odabranog pomeraja p_k, QR algoritam postaje veoma efikasan za rešavanje problema sopstvenih vrednosti za trodijagonalne matrice.

[155] JOHN G.F. FRANCIS (1934 –), engleski naučnik u oblasti kompjuterskih nauka.

[156] VERA NIKOLAEVNA KUBLANOVSKAJA (1920 – 2012), poznata ruska matematičarka u oblasti linearne algebre.

[157] Smatra se da je QR algoritam jedan od deset najznačajnih algoritama u dvadesetom veku.

Uvodjenjem pomeraja p_k, formule (3.8.3) postaju

$$A_k - p_k I = Q_k R_k, \quad A_{k+1} = p_k I + R_k Q_k \quad (k = 1, 2, \dots). \tag{3.8.8}$$

Pretpostavimo da je $A(= A_1)$ simetrična trodijagonalna realna matrica. Sve ostale matrice $A_k = [a_{ij}^{(k)}]_{n\times n}$ su isto takve. Uprošćenja radi, uvedimo notaciju iz odeljka 4.3.7, tj.

$$a_{ii}^{(k)} = b_i^{(k)} \quad (i = 1, \dots n),$$
$$a_{i,i-1}^{(k)} = a_{i-1,1}^{(k)} = c_i^{(k)} \quad (i = 2, \dots, n).$$

i pretpostavimo da su $c_i^{(1)} \neq 0 \ \ (i = 2, \dots, n)$. Tada matrica A_1 ima sve različite sopstvene vrednosti.

Postoje dva načina za izbor pomeraja p_k. Prvi način je da se za p_k uzme vrednost elementa iz donjeg desnog ugla matrice A_k, tj. $p_k = b_n^{(k)}$. Drugi način izbora p_k je takav da se za p_k, uzima ona sopstvena vrednost matrice tipa 2×2

$$\begin{bmatrix} b_{n-1}^{(k)} & c_n^{(k)} \\ c_n^{(k)} & b_n^{(k)} \end{bmatrix} \tag{3.8.9}$$

koja je bliža vrednosti $b_n^{(k)}$. Ovakav izbor

$$p_k = b_n^{(k)} + d - \operatorname{sgn}(d)\sqrt{d^2 + \left(c_n^{(k)}\right)^2}, \quad d = \frac{1}{2}\left(b_{n-1}^{(k)} - b_n^{(k)}\right) \tag{3.8.10}$$

potiče od WILKINSONa i daje bržu konvergenciju QR algoritma u odnosu na izbor $p_k = b_n^{(k)}$. U oba slučaja, inače imamo konvergenciju takvu da je

$$c_n^{(k)} c_{n-1}^{(k)} \to 0 \quad (k \to +\infty).$$

Ako $c_n^{(k)}$ postane zanemarljivo malo (na primer, mašinska nula), možemo uzeti da je $b_n^{(k)}$ jedna sopstvena vrednost matrice A i da pritom poslednju vrstu i kolonu u A_k izostavimo, tako da na dalje, rešavamo problem dimenzije $n-1$. Međutim, ako je $c_{n-1}^{(k)}$ zanemarljivo malo, dok je $c_n^{(k)}$ značajno, možemo odmah odrediti dve sopstvene vrednosti matrice A. To su, zapravo, sopstvene vrednosti matrice (3.8.9). Proces dalje nastavljamo tako što u matrici A_k izostavljamo poslednje dve vrste i dve kolone, a zatim primenjujemo algoritam (3.8.8) na problem dimenzije $n-2$. Na ovaj način, QR algoritam postaje veoma efikasan jer proizvodi deflaciju reda matrice.

U praktičnoj primeni QR algoritma, dekompoziciju (3.8.3), tj. (3.8.8), nije potrebno eksplicitno izračunavati. Naime, moguće je direktno odrediti (3.8.5), tj.

$$A_{k+1} = Q^{(k)T} A Q^{(k)},$$

gde je sada $Q^{(k)}$ ortogonalna matrica odredjena sa (3.8.4). Primetimo da matrica $Q^{(k)}$ konvergira ka ortogonalnoj matrici sopstvenih vektora matrice A (uporediti sa slučajem kod JACOBIevog metoda, odeljak 4.3.5).

Kao što je već rečeno, unitarne (ortogonalne) matrice koje se pojavljuju u (3.8.6) i (3.8.7), svode se na dvodimenzionalne rotacije.

Neka je $Z_p = Z_p^{(k)} = I - 2\boldsymbol{w}_p \boldsymbol{w}_p^T$, $\boldsymbol{w}_p = [w_{p1} \ \cdots \ w_{pn}]^T$, gde smo za koordinate vektora $\boldsymbol{w}_p$ uzeli

$$w_{pp} = \sin\frac{\theta}{2}, \quad w_{p,p+1} = \cos\frac{\theta}{2}, \quad w_{pi} = 0 \quad (i \neq p, p+1).$$

Na osnovu (4.4.8) imamo

$$Z_p = \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & O & & \\ & & 1 & & & & & \\ & & & c & s & & & \leftarrow p \\ & & & s & -c & & & \leftarrow p+1 \\ & & & & & 1 & & \\ & O & & & & & \ddots & \\ & & & & & & & 1 \end{bmatrix}$$

$$\uparrow \quad \uparrow$$
$$p \ \ p+1$$

gde su $c = \cos\theta$ i $s = \sin\theta$.

Jednostavnosti radi, u matrici A_k izostavimo indeks iteracije, tj. uzmimo da je $b_i^{(k)} = b_i$ i $c_i^{(k)} = c_i$. Izaberimo sada ugao $\theta = \theta_1^{(k)} = \theta_1$, tako da su elementi prve kolone ispod glavne dijagonale u matrici $Z_1 A_k$ jednaki nuli. Tada imamo $\cot\theta_1 = b_1/c_2$, tj.

$$c = \frac{b_1}{\sqrt{b_1^2 + c_2^2}} \qquad \text{i} \qquad s = \frac{c_2}{\sqrt{b_1^2 + c_2^2}}.$$

U matrici

$$Z_1 A_k Z_1 = \begin{bmatrix} b_1' & c_2' & d_1 & 0 & \dots & 0 & 0 \\ c_2' & b_2' & c_3' & 0 & & 0 & 0 \\ d_1 & c_3' & b_3 & c_4 & & 0 & 0 \\ 0 & 0 & c_4 & b_4 & & 0 & 0 \\ \vdots & & & & & & \\ 0 & 0 & 0 & 0 & & b_{n-1} & c_n \\ 0 & 0 & 0 & 0 & & c_n & b_n \end{bmatrix}$$

elementi označeni sa primom su oni elementi matrice A_k koji se menjaju pri navedenoj transformaciji. Matrica A_{k+1} se dobija pomoću

$$A_{k+1} = Z_{n-1} \cdots Z_2 Z_1 A_k Z_1 Z_2 \cdots Z_{n-1},$$

gde se $Z_2, \dots, Z_{n-1}$ konstruišu na sličan način kao i Z_1, tako da matrica A_{k+1} postane trodijagonalna. Proizvod svih ortogonalnih (dvodimenzionalnih) rotacija

$$Z = \prod_{k=1}^{\infty} (Z_1^{(k)} Z_2^{(k)} \cdots Z_{n-1}^{(k)})$$

daje matricu sopstvenih vektora. Naime, ovde je

$$\prod_{j=1}^{k} (Z_1^{(j)} Z_2^{(j)} \cdots Z_{n-1}^{(j)}) = Q^{(k)}.$$

Sada možemo rekurzivno formulisati QR algoritam za određivanje jedne sopstvene vrednosti λ i odgovarajućeg ortonormiranog vektora $\boldsymbol{x} = Z\boldsymbol{e}_1$ realne simetrične trodijagonalne matrice. Ovde je $\boldsymbol{x}^T\boldsymbol{x} = 1$ i $\boldsymbol{e}_1 = [1\ 0\ \cdots\ 0]^T$.

Neka su $\lambda^{(k)}$ i $\boldsymbol{y}^{(k)} = [y_1^{(k)}\ y_2^{(k)}\ \cdots\ y_n^{(k)}]^T$ redom aproksimacije za sopstvenu vrednost λ i sopstveni vektor $\boldsymbol{x}$ u k-tom iterativnom koraku $(k = 1, 2, \dots)$.

Startujući od $\boldsymbol{y}^{(1)} = \boldsymbol{e}_1$, tj. $y_1^{(1)} = 1$, $y_i^{(1)} = 0$ $(i = 2, \dots, n)$, k-ti iterativni korak se može iskazati na sledeći način.

Za $p = 1, 2, \dots, n-1$ određujemo

$$\alpha := [(\overline{c}_p^{(k)})^2 + (d_p^{(k)})^2]^{1/2}, \quad c := \overline{c}_p^{(k)}/\alpha, \quad s := d_p^{(k)}/\alpha,$$
$$b_p^{(k+1)} := s^2\overline{b}_p^{(k)} + 2cs\widetilde{c}_{p+1}^{(k)} + s^2 b_{p+1}^{(k)},$$
$$\overline{b}_{p+1}^{(k)} := s^2\overline{b}_p^{(k)} - 2cs\widetilde{c}_{p+1}^{(k)} + c^2 b_{p+1}^{(k)},$$
$$c_p^{(k+1)} := c\overline{c}_p^{(k)} + sd_p^{(k)} = \alpha,$$
$$\overline{c}_p^{(k)} := (\overline{b}_p^{(k)} - b_{p+1}^{(k)})cs + \widetilde{c}_{p+1}^{(k)}(s^2 - c^2),$$
$$\widetilde{c}_{p+2}^{(k)} := -cc_{p+2}^{(k)}, \quad d_{p+1}^{(k)} := sc_{p+2}^{(k)},$$
$$y_p^{(k+1)} := c\overline{y}_p^{(k)} + sy_{p+1}^{(k)}, \quad \overline{y}_{p+1}^{(k)} := s\overline{y}_p^{(k)} + cy_{p+1}^{(k)},$$

pomoću

$$d_1^{(k)} := c_2^{(k)}, \quad \overline{c}_1^{(k)} := b_1^{(k)} - \lambda^{(k)},$$
$$\overline{b}_1^{(k)} := b_1^{(k)}, \quad \widetilde{c}_2^{(k)} := c_2^{(k)}, \quad \overline{y}_1^{(k)} := y_1^{(k)}$$

i $\lambda_k := p_k$, gde je p_k sopstvena vrednost matrice (3.8.9) određena pomoću (3.8.10).

Iterativni proces se prekida kada, na primer, $c_n^{(k)}$ postane dovoljno malo. Kao što smo ranije rekli, u tom slučaju uzimamo da je $\lambda := b_n^{(k)}$ i $\boldsymbol{x} := \boldsymbol{y}^{(k)}$, odbacujemo poslednju vrstu i poslednju kolonu u matrici A_{k+1}, a zatim kompletan iterativni postupak ponavljamo nad ovom matricom reda $n-1$. Tada određujemo drugu sopstvenu vrednost i odgovarajući sopstveni vektor (videti metod deflacije, odeljak 4.3.4), itd. Slično postupamo ukoliko je $c_{n-1}^{(k)}$ zanemarljivo, a $c_n^{(k)}$ značajno. U tom slučaju određujemo dve sopstvene vrednosti istovremeno.

Primetimo da nizovi $\overline{b}_p^{(k)}$, $\overline{c}_p^{(k)}$, $\widetilde{c}_p^{(k)}$, $\overline{y}_p^{(k)}$ ne zahtevaju dodatni memorijski prostor. Naime, oni se mogu memorisati na istim mestima gde se memorišu nizovi $b_p^{(k)}$, $c_p^{(k)}$, $y_p^{(k)}$.

Navedena modifikacija QR algoritma za trodijagonalne matrice igra značajnu ulogu kod dobro poznatog GOLUB[158]–WELSCHovog[159] algoritma [37] koji je

[158] GENE HOWARD GOLUB (1932 – 2007), poznati američki matematičar sa značajnim doprinosima u numeričkoj analizi, a posebno u numeričkoj linearnoj algebri. Bio je profesor na Stanford univerzitetu i član više nacionalnih akademija nauka. Osnivač je danas prestižnih naučnih časopisa *SIAM Journal on Scientific Computing* i *SIAM Journal on Matrix Analysis and Applications*.

[159] Pretraživanjem po internetu, kao i u komunikacijama sa mnogim relevantnim matematičarima u svetu, nismo uspeli da dođemo do podataka o koautoru GOLUBovog rada [37], JOHN H. WELSCHu.

doprineo značajnom progresu u numeričkoj konstrukciji parametara (čvorova i težinskih koeficijenata) GAUSS-CHRISTOFFELovih[160] kvadraturnih formula.

Slično QR algoritmu razvijen je i QL algoritam [14], gde je L donja trougaona matrica, a Q unitarna matrica. Takođe, razvijen je i tzv. implicitni QL algoritam [20]. U oba od pomenutih radova dati su, u to vreme, i odgovarajući programi na ALGOL jeziku.

[160] ELWIN BRUNO CHRISTOFFEL (1829 – 1900), poznati nemački matematičar i fizičar.

Literatura

1. G. ALEFELD, *Zur Konvergenz eines Verfahrens von D. J. Evans zur iterativen Verbesserung einer Nëherung für die inverse Matrix,* Numer. Math. **39** (1982), 163–173.
2. J. ALBRECHT, *Fehlerabschätzungen bei Relaxationsverfahren zur numerischen Auflödsung linearer Gleichungssysteme,* Numer. Math. **3** (1961), 188–201.
3. J. ALBRECHT, *Monotone Iteration und ihre Verwendung zur Lösung linearer Gleichungssysteme,* Numer. Math. **3** (1961), 345–358.
4. J. ALBRECHT, *Bemerkungen zum Iterationsverfahren von Schulz zur Matrixinversion,* Z. Angew. Math. Mech. **41** (1961), 262–263.
5. M. ALTMAN, *An optimum cubically convergent iterative method of inverting a linear bounded operator in Hilbert space,* J. Math. **10** (1960), 1107–1113.
6. E. ANDERSON, Z. BAI, C. BISCHOF, S. BLACKFORD, J. DEMMEL, J. DONGARRA, J. DU CROZ, A. GREENBAUM, S. HAMMARLING, A. MCKENNEY, D. SORENSEN, *LAPACK Users' Guide* (Third ed.), Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.
7. R. ANSORGE, *Über ein Iterationsverfahren von G. Schulz zur Ermittlung der Reziproken einer Matrix,* Z. Angew. Math. Mech. **39** (1959), 164–165.
8. N. APOSTOLATOS, U. KULISCH, *Über die Konvergenz das Relaxationsverfahrens bei nichtnegativen und diagonaldominanten Matricen,* Computing **2** (1967), 17–24.
9. N. S. BAHVALOV, *Čislennye metody,* Nauka, Moskva, 1973.
10. W. BARTH, R. S. MARTIN, J. H. WILKINSON, *Calculation of the eigenvalues of a symmetric tridiagonal matrix by the method bisection,* Numer. Math. **9** (1967), 386–393.
11. A. BJORCK, G. DAHLQUIST, *Numerische Methoden,* R. Oldenbourg Verlag, München–Wien, 1972.
12. Z. BOHTE, *Numeričke metode*, Državna založba Slovenije, Ljubljana, 1978.
13. Z. BOHTE, M. PETKOVŠEK, *Gaussian elimination for diagonally dominant matrices*. In: *Numerical Methods and Approximation Theory* (ed. by G.V. Milovanović), Faculty of Electronic Engineering, Niš, 1984, pp. 1–6.
14. H. BOWDLER, R. S. MARTIN, C. REINSCH, J. H. WILKINSON, *The QR and QL algorithms for symmetric matrices,* Numer. Math. **11** (1968), 293–306.
15. L. COLLATZ, *Über die Konvergenzkriterion bei Iterationsverfahren für lineare Gleichungssysteme,* Math. Z. **53** (1950), 149–161.
16. L. COLLATZ, *Funktionanalysis und numerische Mathematik,* Springer-Verlag, Berlin-Heidelberg-New York, 1964.
17. A. M. DANILEVSKIĬ, *The numerical solution of the secular equation* Matem. Sbornik **44** (2) (1937), 169–171 (na ruskom).
18. J. W. DEMMEL, *Applied Numerical Linear Algebra,* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.
19. B. P. DEMIDOVIČ, I. A. MARON, *Foundations of Numerical Analysis,* Nauka, Moscow, 1966 (na ruskom).
20. A. DUBRULLE, R. S. MARTIN, J. H. WILKINSON, *The implicit QL algorithm,* Numer. Math. **12** (1968), 377–383.
21. W. DÜCK, *Fehlerabschätzungen für das Iterationsverfahren von Schulz zur Bestimmung der Inversen einer Matrix,* Z. Angew. Math. Mech. **40** (1960), 192–194.
22. D. J. EVANS, *An implicit iterative process for matrix inversion,* Internat. J. Comput. Math. **9** (1981), 335–341.
23. D. K. FADDEEV, *On the conditionality of matrices,* Trudy Mat. Inst. Steklov. **53** (1959), 387–391 (na ruskom).
24. D. K. FADDEEV, V. N. FADDEEVA, *Ill-conditioned systems of linear equations,* Ž. Vyčisl. Mat. i Mat. Fiz. **1** (1961), 412–417 (na ruskom).

25. D. K. FADDEEV, V. N. FADDEEVA, *Computational Methods of Linear Algebra,* W. H. Freeman and Co., San Francisco – London, 1963.
26. D. K. FADDEEV, V. N. FADDEEVA, *Natural norms in algebraic processes,* SIAM J. Numer. Anal. **7** (1970), 520–531.
27. D. A. FLANDERS, G. SHORTLEY, *Numerical determination of fundamental modes,* J. Appl. Phys. **21** (1950), 1326–1332.
28. G. E. FORSYTHE, *Solving linear algebraic equations can be interesting,* Bull. Amer. Math. Soc. **59** (1953), 299–329.
29. G. FORSYTHE, C. B. MOLER, *Computer solution of linear algebraic systems,* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1967.
30. J. G. F. FRANCIS, *The QR transformation – a unitary analogue to the LR transformation,* Comput. J. **4** (1961/62), 265–271; 332–345.
31. W. GIVENS, *A method of computing eigenvalues and eigenvectors suggested by classical results on symmetric matrices,* Appl. Math. Ser. Nat. Bur. Stand. **29** (1953), 117–122.
32. W. GIVENS, *Numerical computation of the characteristic values of a real symmetric matrix,* Rep. ORNL 1574. Oak Ridge National Laboratory, Oak Ridge, Tenn. (1954), vi+107 pp.
33. W. GIVENS, *The characteristic value-vector problem,* J. Assoc. Comput. Mach. **4** (1957), 298–307.
34. W. GIVENS, *Computation of plane unitary rotations transforming a general matrix to triangular form,* J. Soc. Indust. Appl. Math. **6** (1958), 26–50.
35. G. H. GOLUB, C. F. VAN LOAN, *Matrix Computations,* The Johns Hopkins University Press, Baltimore, Maryland, 1984.
36. G. H. GOLUB, R. S. VARGA, *Chebyshev semi-iterative methods, succesive overrelaxation iterative methods and second order Richardson iterative methods,* Numer. Math. **3** (1961), 147–168.
37. G. H. GOLUB, J. H. WELSCH, *Calculation of Gauss quadrature rules,* Math. Comp. **23** (1969), 221–230.
38. A. R. GOURLAY, G. A. WATSON, *Computational Methods for Matrix Eigenproblems,* John Wiley and Sons, Chichester-New York, 1973.
39. A. S. HOUSEHOLDER, *Unitary triangularization of a nonsymmetric matrix*, J. Assoc. Comput. Mach. **5** (1958), 339–342.
40. A. S. HOUSEHOLDER, *The approximate solution of matrix problems,* J. Assoc. Comp. Mach. **5** (1958), 205–243.
41. A. S. HOUSEHOLDER, *Generated error in rotational tridiagonalization,* J. Assoc. Comput. Mach. **5** (1958), 335–338.
42. A. HOUSEHOLDER, *The Theory of Matrices in Numerical Analysis,* Blaisdell Publ. Co., New York, 1964.
43. M. R. HESTENES, E. STIEFEL, *Methods of conjugate gradients for solving linear systems,* J. Res. Nat. Bur. Stand. **49** (1952), 409–436.
44. N. J. HIGHAM, *Accuracy and Stability of Numerical Algorithms,* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.
45. F. B. HILDEBRAND, *Introduction to Numerical Analysis,* McGraw-Hill, Inc., New York, 1974.
46. H. HOTELLING, *Some new methods in matrix calculation,* Amer. Math. Statist. **14** (1943), 1–34.
47. S.-H. HOU, *A simple proof of the Leverrier–Faddeev characteristic polynomial algorithm,* SIAM Rev. **40** (1998), 706–709.
48. E. ISAACSON, H. B. KELLER, *Analysis of Numerical Methods,* John Wiley and Sons, Inc., New York, 1966.
49. V. I. KRYLOV, V. V. BOBKOV, P. I. MONASTYRNYIĬ, *Numerical Methods of Higher Mathematics. Vol. 1,* Izdat. "Vyšéĭšaja Škola", Minsk, 1972 (na ruskom).

50. V. N. KUBLANOVSKAJA, *Some algorithms for the solution of the complete problem of eigenvalues,* Ž. Vyčisl. Mat. i Mat. Fiz. **1** (1961), 555–570 (na ruskom).
51. C. LANSZOS, *Solution of systems of linear equations by minimized iterations,* J. Res. Nat. Bur. Stand. **49** (1952), 33.
52. P. B. MADIĆ, *Lose rešljivi sistemi linearnih algebarskih jednačina i njihovo rešavanje* (Doktorska disertacija), Beograd, 1965.
53. G. V. MILOVANOVIĆ, *Numerička analiza, I deo,* Naučna knjiga, Beograd, 1985.
54. G. V. MILOVANOVIĆ, Đ. R. ĐORĐEVIĆ, Programiranje numeričkih metoda na FORTRAN jeziku. Institut za dokumentaciju zaštite na radu "Edvard Kardelj", Niš, 1981.
55. G. V. MILOVANOVIĆ, R. Ž. ĐORĐEVIĆ, *Linearna algebra,* Elektronski fakultet u Nišu, Niš, 2004.
56. D. S. MITRINOVIĆ, D. Ž. DJOKOVIĆ, *Polinomi i matrice,* ICS, Beograd, 1975.
57. B. N. PARLETT, *The Symmetric Eigenvalue Problem,* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980.
58. B. N. PARLETT, W. G. POOLE, *A geometric theory for the QR, LU, and power iterations,* SIAM J. Numer. Anal. **10** (1973), 389–412.
59. M. PETKOV, *Čisleni metodi na algebrata,* Nauka i izkustvo, Sofija, 1974.
60. W. V. PETRYSHYN, *On the inversion of matrices and linear operators,* Proc. Amer. Math. Soc. **16** (1965), 893–901.
61. E. REICH, *On the convergence of the classical iterative method of solving linear simultaneous equations,* Ann. Math. Statist. **20** (1949), 448–451.
62. H. RUTISHAUSER, *Solution of eigenvalue problems with the LR-transformation,* Appl. Math. Ser. Nat. Bur. Stand. **49** (1958), 47–81.
63. S. SCHECHTER, *Relaxation methods for linear equations,* Comm. Pure and Appl. Math. **12** (1959), 313–335.
64. G. SCHULZ, *Iterative Berechnung der reziproken Matrix,* Z. Angew. Math. Mech. **13** (1933), 57–59.
65. R. V. SOUTWELL, *Relaxation Methods in Theoretical Physics. 2 vols,* Oxford University Press. Fair Lawn., New Jersey, 1956.
66. G. W. STEWART, *Conjugate direction methods for solving systems of linear equations,* Numer. Math. **21** (1973), 285–297.
67. E. STIEFEL, *Relaxationsmethoden bester Strategie zur Lösung linear Gleihungssysteme,* Comm. Math. Helv. **29** (1955), 157–179.
68. J. STOER, *Einführung in die Numerische Mathematik I,* Springer-Verlag, Berlin – Heidelberg – New York, 1973.
69. J. STOER, R. BULIRSCH, *Einführung in die Numerische Mathematik II,* Springer-Verlag, Berlin – Heidelberg – New York, 1973.
70. V. STRASSEN, *Gaussian elimination is not optimal,* Numer. Math. **13** (1969), 354–356.
71. L. N. TREFETHEN, D. BAU, III, *Numerical Linear Algebra,* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.
72. R. S. VARGA, *Matrix Iterative Analysis,* Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1962.
73. J. H. WILKINSON, *The Algebraic Eigenvalue Problem,* Oxford University Press, London, 1965.
74. J. H. WILKINSON, C. REINSCH, *Handbook for Automatic Computation. Vol. II Linear Algebra,* Springer-Verlag, Berlin –Heidelberg – New York, 1971.
75. D. YOUNG, *Iterative methods for solving partial difference equations of elliptic type,* Trans. Amer. Math. Soc. **76** (1954), 92–111.
76. D. M. YOUNG, *Iterative Solution of Large Systems,* Academic Press, New York, 1971.
77. D. M. YOUNG, R. T. GREGORY, *A Survey of Numerical Mathematics. Vol. II,* Addison-Wesley Publ. Co., Reading, Massachusetts, 1973.

5. NELINEARNE JEDNAČINE I SISTEMI

5.1 NELINEARNE JEDNAČINE

5.1.1 Osnovne napomene

Ovo poglavlje je posvećeno konstrukciji metoda za rešavanje nelinearnih jednačina oblika $f(x) = 0$, bez obzira na to da li su one algebarske ili transcendentne. Zbog specifičnosti koje poseduju algebarske jednačine i zbog važnosti koje one u primenama imaju, u literaturi je razvijen ogroman broj metoda za njihovo rešavanje. U našem izlaganju ovom problemu posvećujemo posebno poglavlje.

U glavi 3, a posebno u odeljku 3.1.2, data je opšta teorija o egzistenciji i jedinstvenosti rešenja nelinearnih jednačina oblika

$$f(x) = 0, \tag{1.1.1}$$

kao i o iterativnim procesima za nalaženje ovih rešenja. U ovom poglavlju razmatraćemo konstrukciju konkretnih iterativnih metoda za rešavanje jednačine (1.1.1). Svi metodi koji će biti izloženi primenjuju se na određivanje izolovanih korena jednačina (videti, na primer, [56], [103], [104]).

5.1.2 NEWTONov metod

NEWTONov ili kako se često naziva NEWTON-RAPHSONov[161] metod predstavlja osnovni metod za nalaženje izolovanih korena nelinearnih jednačina.

Neka je na segmentu $[\alpha, \beta]$ izolovan jedinstven prost koren $x = a$ jednačine (1.1.1) i neka $f \in C^1[\alpha, \beta]$ i $f'(x) \neq 0$ za svako $x \in [\alpha, \beta]$. Izaberimo tačku $x_0 \in (\alpha, \beta)$. Tada, na osnovu TAYLORove formule, imamo

[161] JOSEPH RAPHSON (?1648 – ?1715), engleski matematičar o čijem životu se malo zna uključujući i tačne godine rođenja i smrti.

$$f(a) = f(x_0) + f'(x_0)(a-x_0) + O((a-x_0)^2), \tag{1.2.1}$$

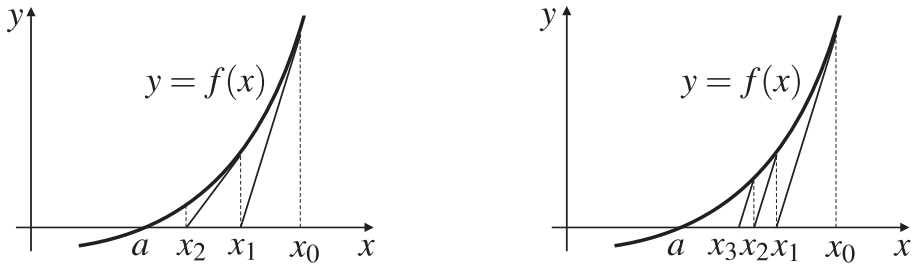
gde je $\xi = x_0 + \theta(a-x_0)$ $(0 < \theta < 1)$. Imajući u vidu da je $f(a) = 0$, zanemarivanjem poslednjeg člana na desnoj strani u jednakosti (1.2.1), dobijamo

$$a \approx x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Sa x_1 označimo desnu stranu u poslednjoj približnoj jednakosti, tj.

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{1.2.2}$$

Geometrijski x_1 predstavlja apscisu tačke preseka tangente na krivu $y = f(x)$, u tački $(x_0, f(x_0))$, sa x-osom (videti sl. 1.2.1, levo).



Slika 1.2.1. Geometrijska interpretacija NEWTONovog (levo) i modifikovanog NEWTONovog metoda (desno)

Jednakost (1.2.2) sugeriše konstrukciju iterativnog procesa

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \ldots, \tag{1.2.3}$$

koji je poznat kao NEWTONOV metod ili metod tangente.

Pređimo sada na ispitivanje konvergencije iterativnog procesa (1.2.3), uvodeći dopunsku pretpostavku za funkciju f. Naime, pretpostavimo da $f \in C^2[\alpha, \beta]$.

Kako je iterativna funkcija ϕ, kod NEWTONovog metoda, određena sa

$$\phi(x) = x - \frac{f(x)}{f'(x)},$$

diferenciranjem dobijamo

$$\phi'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}. \tag{1.2.4}$$

Primetimo da je $\phi(a) = a$ i $\phi'(a) = 0$. Kako je, na osnovu učinjenih pretpostavki za f, funkcija ϕ' neprekidna na $[\alpha, \beta]$ i kako je $\phi'(a) = 0$, to postoji okolina tačke $x = a$, u oznaci $U(a)$, u kojoj je

$$|\phi'(x)| = \left| \frac{f(x)f''(x)}{f'(x)^2} \right| \leq q < 1. \tag{1.2.5}$$

Teorema 1.2.1. *Ako $x_0 \in U(a)$, niz $\{x_k\}_{k\in\mathbb{N}_0}$, generisan pomoću* (1.2.3)*, konvergira ka tački $x = a$, pri čemu je*

$$\lim_{k\to+\infty} \frac{x_{k+1} - a}{(x_k - a)^2} = \frac{f''(a)}{2f'(a)}. \tag{1.2.6}$$

Dokaz. Iz (1.2.3) sleduje

$$x_{k+1} - a = x_k - a - \frac{f(x_k) - f(a)}{f'(x_k)},$$

tj.

$$f(x) - f(a) = f'(x_k)(x_k - a) - f'(x_k)(x_{k+1} - a).$$

S druge strane, na osnovu TAYLORove formule, imamo

$$f(a) - f(x_k) = f'(x_k)(a - x_k) + \frac{1}{2}f''(\xi_k)(a - x_k)^2,$$

gde je ξ_k neka tačka između x_k i a.

Ako poslednje dve jednakosti saberemo dobijamo

$$0 = -f'(x_k)(x_{k+1} - a) + \frac{1}{2}f''(\xi_k)(a - x_k)^2.$$

Kako je, na osnovu učinjene pretpostavke, $f'(x_k) \neq 0$, iz poslednje jednakosti sleduje

$$\frac{x_{k+1} - a}{(x_k - a)^2} = \frac{f''(\xi_k)}{2f'(x_k)}. \tag{1.2.7}$$

Da bismo dokazali konvergenciju niza $\{x\}_{k\in N_0}$ dovoljno je primetiti da ϕ preslikava $U(a)$ u $U(a)$. Tada, imajući u vidu (1.2.5) vidimo da ϕ zadovoljava uslove teoreme 1.2.2 iz odeljka 3.1.2, odakle sleduje konvergencija iterativnog procesa (1.2.3), pri proizvoljnom $x_0 \in U(a)$.

Kako $x_k \to a$, kada $k \to +\infty$, i kako je f'' neprekidna funkcija, iz (1.2.7) sleduje (1.2.6). □

Primer 1.2.1. Odredićemo rešenje jednačine

$$f(x) = x - \cos x = 0$$

na segmentu $[0, \pi/2]$ primenom NEWTONovog metoda

$$x_{k+1} = x_k - \frac{x_k - \cos x_k}{1 + \sin x_k} = \frac{x_k \sin x_k + \cos x_k}{1 + \sin x_k}, \quad k = 0, 1, \dots .$$

Primetimo da je $f'(x) = 1 + \sin x > 0$ za svako $x \in (0, \pi/2)$. Startujući sa $x_0 = 1$, kao i u primeru 2.2.1 iz odeljka 3.2.2, dobijamo niz iteracija:

k	x_k
0	1.
1	0.7503638678402439
2	0.7391128909113617
3	0.7390851333852838
4	0.7390851332151607
5	0.7390851332151606

Na osnovu poslednje dve iteracije zaključujemo da je rešenje posmatrane jednačine određeno sa šesnaest tačnih decimala. $\triangle$

Primer 1.2.2. Primenićemo NEWTONov metod na rešavanje jednačine $f(x) = x^n - a = 0$ $(a > 0,\ n > 1)$ u cilju da dobijamo iterativnu formulu za određivanje n-tog korena iz pozitivnog broja a,

$$x_{k+1} = x_k - \frac{x_k^n - a}{n x_k^{n-1}} = \frac{1}{n}\left\{(n-1)x_k + \frac{a}{x_k^{n-1}}\right\}, \quad k = 0, 1, \dots .$$

Specijalan slučaj ove formule, za $n = 2$, naveden je na samom početku ove knjige u odeljku 1.1.1. $\triangle$

Kod primene NEWTONog metoda često se nameće pitanje kako izabrati početnu vrednost x_0 tako da niz $\{x_k\}_{k\in\mathbb{N}}$ bude monoton. Jedan odgovor na ovo pitanje potiče još od FOURIERa. Naime, ako f'' ne menja znak na $[\alpha, \beta]$ i ako se x_0 izabere tako da je $f(x_0)f''(x_0) > 0$, niz $\{x_k\}_{k\in\mathbb{N}}$ biće monoton. Ovo tvrđenje sleduje iz (1.2.4).

Na osnovu teoreme 1.2.1 zaključujemo da NEWTONov metod primenjen na određivanje prostog korena $x = a$ ima kvadratnu konvergenciju ako je $f''(a) \neq 0$. U tom slučaju, faktor konvergencije (asimptotska konstanta greške) je

$$C_2 = \left| \frac{f''(a)}{2f'(a)} \right| .$$

Slučaj $f''(a) = 0$ treba posebno analizirati. Naime, ako pretpostavimo da $f \in C^3[\alpha, \beta]$, može se dokazati da je

$$\lim_{k \to +\infty} \frac{x_{k+1} - a}{(x_k - a)^3} = \frac{f(a)}{3f'(a)}.$$

Primer 1.2.3. Posmatrajmo jednačinu

$$f(x) = x^3 - 3x^2 + 4x - 2 = 0.$$

Kako je $f(0) = -2$ i $f(1.5) = 0.625$ zaključujemo da na segmentu $[0, 1.5]$ data jednačina ima koren. S druge strane, $f'(x) = 3x^2 - 6x + 4 = 3(x-1)^2 + 1 > 0$, što znači da je koren prost pa možemo primeniti NEWTONov metod.

Startujući sa $x_0 = 1.5$ i korišćenjem aritmetike dovoljno velike dužineu, na primer sa 60 dekadnih cifara u programskom paketu MATHEMATICA, dobijamo sledeće iteracije:

k	x_k	e_k
0	1.5	$5.00(-1)$
1	1.14285714285714285714285714285714285714285714285714285714286	$1.41(-1)$
2	1.00549450549450549450549450549450549450549450549450549450549	$5.49(-3)$
3	1.00000033172369936942641986865731847166930331720382132432726	$3.32(-7)$
4	1.00000000000000000007300615826690374163861987594028874395515	$7.30(-20)$
5	1.00000000000000000000000000000000000000000000000000000000078	$7.78(-58)$

Tačna vrednost korena je $a = 1$, s obzirom da je $f(x) = (x-1)^3 + (x-1)$. Primetimo da je u ovom slučaju $f''(1) = 0$. Ako stavimo $e_k = x_k - a$, $k = 0, 1, \ldots$, imamo

$$e_{k+1} = \frac{2e_k^3}{1 + 3e_k^2}, \quad k = 0, 1, \ldots,$$

tj. $e_{k+1} \simeq 2e_k^3$, kada $k \to +\infty$. Zaista, u ovom slučaju imamo $f'''(1)/(3f'(1)) = 2$. Numeričke vrednosti za e_k su takođe navedene u prethodnoj tabeli. $\triangle$

U cilju smanjivanja broja računskih operacija, često se koristi sledeća modifikacija NEWTONovog metoda

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)}, \quad k = 0, 1, \dots \ . \tag{1.2.8}$$

Geometrijski x_{k+1} predstavlja apscisu tačke preseka x-ose sa pravom koja prolazi kroz tačku $(x_k, f(x_k))$ i koja je paralelna sa tangentom krive $y = f(x)$ postavljene u tački $(x_0, f(x_0))$ (videti sl. 1.2.1, desno).

Iterativna funkcija modifikovanog NEWTONovog metoda je

$$\phi_1(x) = x - \frac{f(x)}{f'(x_0)}.$$

Kako je $\phi_1(a) = a$ i $\phi_1'(a) = 1 - f'(a)/f'(x_0)$, zaključujemo da metod ima red konvergencije jedan, tj. važi

$$x_{k+1} - a \cong \Big(1 - \frac{f'(a)}{f'(x_0)}\Big)(x_k - a) \quad (k \to +\infty),$$

pri čemu je uslov

$$\Big|1 - \frac{f'(x)}{f'(x_0)}\Big| \leq q < 1,$$

analogan uslovu (1.2.5).

NEWTONov metod može se razmatrati i kao specijalan slučaj tzv. uopštenog NEWTONovog metoda

$$x_{k+1} = x_k - \frac{\psi(x_k) f(x_k)}{\psi'(x_k) f(x_k) + \psi(x_k) f'(x_k)}, \quad k = 0, 1, \dots, \tag{1.2.9}$$

gde je ψ data diferencijabilna funkcija.

Za $\psi(x) = 1$, (1.2.9) se svodi na standardni NEWTONov metod (1.2.3).

Za $\psi(x) = x^p$, gde je p parametar, iz (1.2.9) sleduje formula

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k) + \dfrac{p}{x_k} f(x_k)}, \quad k = 0, 1, \dots,$$

tj.

$$x_{k+1} = x_k \left\{1 - \frac{f(x_k)}{x_k f'(x_k) + p f(x_k)}\right\}, \quad k = 0, 1, \dots \ . \tag{1.2.10}$$

Metod definisan formulom (1.2.10) razmatran je u radu [19], dok je specijalni slučaj za $p = 1 - n$, u slučaju kada je f algebarski polinom stepena n, bio tretiran ranije u radu [125].

Na kraju navedimo još jednu modifikaciju NEWTONovog metoda, koja se sastoji u sukcesivnoj primeni formula

$$y_k = x_k - \frac{f(x_k)}{f'(x_k)}, \quad x_{k+1} = y_k - \frac{f(y_k)}{f'(x_k)}, \quad k = 0, 1, \ldots. \tag{1.2.11}$$

Slično dokazu teoreme 1.2.1, za ovaj dvo–koračni metod mogu se dokazati granične vrednosti

$$\lim_{k\to+\infty} \frac{x_{k+1} - a}{(y_k - a)(x_k - a)} = \frac{f''(a)}{f'(a)}$$

i

$$\lim_{k\to+\infty} \frac{x_{k+1} - a}{(x_k - a)^3} = \frac{1}{2}\left(\frac{f''(a)}{f'(a)}\right)^2,$$

gde je $\lim\limits_{k\to+\infty} x_k = a$. Dakle, dvo–koračni iterativni proces definisan formulama (1.2.11) ima kubnu konvergenciju.

O nekim više–stepenim metodima biće reči u odeljku 5.1.8.

5.1.3 NEWTONov metod za višestruke nule

Posmatramo jednačinu $f(x) = 0$, koja na segmentu $[\alpha, \beta]$ ima koren $x = a$ višestrukosti m (≥ 2). Ako pretpostavimo da $f \in C^{m+1}[\alpha, \beta]$, tada je

$$f(a) = f'(a) = \cdots = f^{(m-1)}(a) = 0, \quad f^{(m)}(a) \neq 0.$$

Naime, u ovom slučaju, f se može predstaviti u obliku

$$f(x) = (x - a)^m g(x), \tag{1.3.1}$$

gde $g \in C^{m+1}[\alpha, \beta]$ i $g(a) \neq 0$.

Iz (1.3.1) sleduje

$$f'(x) = m(x - a)^{m-1} g(x) + (x - a)^m g'(x)$$

i

$$\Delta(x) = \frac{f(x)}{f'(x)} = \frac{(x - a)g(x)}{mg(x) + (x - a)g'(x)} \quad (x \neq a).$$

Ako stavimo $\Delta(a) := \lim_{x\to a} \Delta(x)$, tada je $\Delta(a) = 0$.

Iterativna funkcija NEWTONovog metoda, primenjenog na određivanje višestrukog korena, na osnovu prethodnog postaje

$$\phi(x) = x - \frac{(x-a)g(x)}{mg(x) + (x-a)g'(x)}.$$

Kako je

$$\phi(a) = a, \quad \phi'(a) = 1 - \frac{1}{m}, \quad \frac{1}{2} \le \phi'(a) < 1 \quad (m \ge 2)$$

i ϕ' neprekidna funkcija, zaključujemo da postoji okolina korena $x = a$ u kojoj je $|\phi'(x)| \le q < 1$, odakle sleduje da je NEWTONov metod i u ovom slučaju konvergentan, ali sa redom konvergencije jedan.

Ukoliko nam je unapred poznat red višestrukosti korena, tada se NEWTONov metod može modifikovati tako da ima red konvergencije dva. Naime, treba samo uzeti

$$x_{k+1} = x_k - m\frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \dots \ . \tag{1.3.2}$$

Napomena 1.3.1. Formalno, formula (1.3.2) predstavlja NEWTONov metod primenjen na rešavanje jednačine

$$F(x) = \sqrt[m]{f(x)} = 0.$$

Teorema 1.3.1. *Ako je x_0 izabrano dovoljno blisko korenu $x = a$, čiji je red višestrukosti m, tada niz $\{x_k\}_{k\in\mathbb{N}_0}$, definisan pomoću* (1.3.2)*, konvergira ka a, pri čemu je*

$$\frac{x_{k+1} - a}{(x_k - a)^2} \simeq \frac{1}{m(m+1)} \frac{f^{(m+1)}(a)}{f^{(m)}(a)} \quad (k \to +\infty).$$

Dokaz ove teoreme može se naći, na primer, u [93].

Ukoliko red višestrukosti m nije poznat, tada se umesto jednačine $f(x) = 0$ može rešavati jednačina $f(x)/f'(x) = 0$, čiji su svi koreni prosti. NEWTONov metod primenjen na ovu jednačinu daje formulu

$$x_{k+1} = x_k - \left[\frac{\dfrac{f(x)}{f'(x)}}{\left(\dfrac{f(x)}{f'(x)}\right)'} \right]_{x=x_k}, \qquad k = 0, 1, \dots,$$

tj. formulu

$$x_{k+1} = x_k - \frac{f(x_k)f'(x_k)}{f'(x_k)^2 - f(x_k)f''(x_k)}, \quad k = 0,1,\ldots,$$

čiji je red konvergencije dva. Primetimo da se ova funkcija dobija iz (1.2.9) uzimajući $\psi(x) = 1/f'(x)$.

5.1.4 Metod sečice

Ako se u NEWTONovom metodu vrednost izvoda $f'(x_k)$ aproksimira pomoću podeljene razlike $\dfrac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$ dobijamo metod sečice

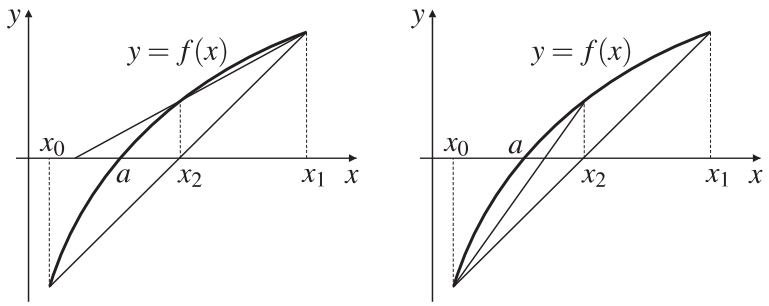
$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k), \quad k = 1,2,\ldots, \tag{1.4.1}$$

koji pripada klasi metoda oblika (2.1.2) (odeljak 3.2.1). Za startovanje iterativnog procesa (1.4.1) potrebne su dve početne vrednosti x_0 i x_1. Geometrijska interpretacija metoda sečice data je na sl. 1.4.1 (levo).

Neka na segmentu $[\alpha,\beta]$ postoji jedinstven koren $x = a$ jednačine $f(x) = 0$. Za ispitivanje konvergencije iterativnog procesa (1.4.1) pretpostavimo da funkcija $f \in C^2[\alpha,\beta]$ i $f'(x) = 0$ za svako $x \in [\alpha,\beta]$.

Ako stavimo $e_k = x_k - a$ $(k = 0,1,\ldots)$, iz (1.4.1) sleduje

$$e_{k+1} = e_k - \frac{e_k - e_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k). \tag{1.4.2}$$



Slika 1.4.1. Geometrijska interpretacija metoda sečice (levo) i metoda *regula falsi* (desno)

Kako je

$$f(x_k) = f'(a)e_k + \frac{1}{2}f''(a)e_k^2 + O(e_k^3)$$

i

$$\frac{f(x_k) - f(x_{k-1})}{e_k - e_{k-1}} = f'(a) + \frac{1}{2}(e_k + e_{k-1})f''(a) + O(e_{k-1}^2)$$

zamenom u (1.4.2) dobijamo

$$e_{k+1} = e_k\left(1 - \frac{f'(a) + \frac{1}{2}e_k f''(a) + O(e_k^2)}{f'(a) + \frac{1}{2}(e_k + e_{k-1})f''(a) + O(e_{k-1}^2)}\right),$$

odakle je

$$e_{k+1} = e_k\left[1 - \left(1 + \frac{1}{2}e_k\frac{f''(a)}{f'(a)} + O(e_k^2)\right)\left(1 - \frac{1}{2}(e_k + e_{k-1})\frac{f''(a)}{f'(a)} + O(e_{k-1}^2)\right)\right],$$

tj.

$$e_{k+1} = e_k e_{k-1}\frac{f''(a)}{2f'(a)}\big(1 + O(e_{k-1})\big). \tag{1.4.3}$$

Da bismo odredili red konvergencije i faktor konvergencije pođimo od

$$|e_{k+1}| = C_r|e_k|^r|1 + O(e_k)|. \tag{1.4.4}$$

Tada, na osnovu (1.4.3) i (1.4.4), dobijamo

$$|e_{k+1}| = C_r|e_k|^r|1 + O(e_k)| = C_r\big(C_r|e_{k-1}|^r\big)^r|1 + O(e_{k-1})|$$

$$= C_r|e_{k-1}|^r|e_{k-1}|\left|\frac{f''(a)}{2f'(a)}\right||1 + O(e_{k-1})|,$$

odakle sleduje

$$r^2 - r - 1 = 0 \quad \text{i} \quad C_r = \left|\frac{f''(a)}{2f'(a)}\right|^{1/r}.$$

Red konvergencije r dobijamo kao pozitivno rešenje dobijene kvadratne jednačine, tj. $r = (1 + \sqrt{5})/2 \cong 1.62$. Faktor konvergencije je

$$C_r = \left|\frac{f''(a)}{2f'(a)}\right|^{(\sqrt{5}-1)/2}.$$

Napomena 1.4.1. U literaturi se za rešavanje jednačine

$$x = g(x) \tag{1.4.5}$$

sreće tzv. metod WEIGSTEINa (videti [137]), kod koga se polazeći od x_0 generiše niz $\{x_k\}_{k\in\mathbb{N}}$ pomoću

$$\begin{aligned} x_1 &= g(x_0), \\ x_{k+1} &= g(x_k) - \frac{(g(x_k)-g(x_{k-1}))(g(x_k)-x_k)}{(g(x_k)-g(x_{k-1}))-(x_k-x_{k-1})}, \quad k = 1,2,\dots . \end{aligned} \tag{1.4.6}$$

U radu [134] je pokazano da je ovaj metod, ustvari, metod sečice sa početnim vrednostima x_0 i $x_1 = g(x_0)$. Naime, ako jednačinu (1.4.5) predstavimo u obliku

$$f(x) = g(x) - x = 0, \tag{1.4.7}$$

smenom (1.4.7) u (1.4.6), dobijamo (1.4.1).

Metod sečice može se modifikovati tako da je

$$x_{k+1} = x_k - \frac{x_k - x_0}{f(x_k) - f(x_0)} f(x_k), \quad k = 1,2,\dots . \tag{1.4.8}$$

Ovaj metod se često naziva metod *regula falsi*. Za razliku od metoda sečice, gde je dovoljno uzeti $x_1 \neq x_0$, kod ovog metoda x_1 i x_0 treba uzeti sa različitih strana u odnosu na odnosu na koren $x = a$. Geometrijska interpretacija metoda regula falsi data je na sl. 1.4.1 (desno).

Iterativna funkcija kod modifikovanog metoda sečice je

$$\phi(x) = x - \frac{x - x_0}{f(x) - f(x_0)} f(x) = \frac{x_0 f(x) - x f(x_0)}{f(x) - f(x_0)}.$$

Ako pretpostavimo da $f \in C^1[\alpha,\beta]$, tada je

$$\phi'(x) = \frac{f(x_0)}{f(x) - f(x_0)} \left[\frac{x - x_0}{f(x) - f(x_0)} f'(x) - 1 \right].$$

Kako je $\phi(a) = a$ i $\phi'(x) \neq 0$, zaključujemo da iterativni proces (1.4.8), ukoliko je konvergentan, ima red konvergencije jedan. Uslov konvergencije, u ovom slučaju, je dat pomoću

$$|\phi'(x)| \leq q \leq 1 \quad (f(x) \neq f(x_0)),$$

za svako $x \in [\alpha,\beta] \setminus \{x_0\}$.

Primer 1.4.1. Primenom teoreme 2.4.1 (odeljak 3.2.4) na iterativni proces (1.4.8) dobijemo iterativni proces drugog reda

$$x_{k+1} = \frac{x_0 g(x_k) - x_k h(x_k)}{g(x_k) - h(x_k)}, \quad k = 1, 2, \ldots,$$

gde su

$$g(x) = \frac{f(x) - f(x_0)}{x - x_0} \quad \text{i} \quad h(x) = \frac{f'(x) f(x_0)}{f(x)}.$$

$\triangle$

Ako se izvod $f'(x_k)$ u NEWTONovom metodu zameni konačnom razlikom u tački x_k, sa korakom $h = f(x_k)$, tj.

$$f'(x_k) \cong \frac{f(x_k + f(x_k)) - f(x_k)}{f(x_k)},$$

dobija se metod STEFFENSENa

$$x_{k+1} = x_k - \frac{f(x_k)^2}{f(x_k + f(x_k)) - f(x_k)}, \quad k = 0, 1, \ldots . \tag{1.4.9}$$

Neka je $x = a$ jedinstven prost koren jednačine $f(x) = 0$ na segmentu $[\alpha, \beta]$ i neka $f \in C^2[\alpha, \beta]$. Metod STEFFENSENa je interesantan jer ima red konvergencije dva, a da pri tome iterativna funkcija

$$\phi(x) = x - \frac{f(x)^2}{f(x + f(x)) - f(x)}$$

ne sadrži izvod f'. Da bismo odredili asimptotsku konstantu greške metoda (1.4.9), pođimo od TAYLORove formule

$$f(x + f(x)) = f(x) + f'(x) f(x) + \frac{1}{2} f''(\xi) f(x)^2,$$

gde je $\xi = x + \theta f(x)$ $(0 < \theta < 1)$. Tada je

$$\phi(x) = x - \frac{f(x)/f'(x)}{1 + \frac{1}{2} f''(\xi) f(x)/f'(x)}.$$

S obzirom na to da postoji okolina tačke $x = a$, u oznaci $U(a)$, u kojoj je

$$\left| \frac{1}{2} \cdot \frac{f''(\xi) f(x)}{f'(x)} \right| < 1,$$

imamo

$$\phi(x) = x - \frac{f(x)}{f'(x)} \left(1 - \frac{1}{2} \frac{f''(\xi)}{f'(x)} f(x) + O(f(x)^2) \right) \quad (x \in U(a)).$$

Kako je

$$\frac{f(x)}{f'(x)} = x - a - \frac{f''(a)}{2f'(a)}(x-a)^2 + O((x-a)^3) \quad (x \in U(a)),$$

iz poslednje jednakosti sleduje

$$\Phi(x) - a \sim \frac{f''(a)}{2f'(a)}(f'(a)+1)(x-a)^2 \quad (x \to a),$$

što znači da je asimptotska konstanta greške

$$C_2 = \left| \frac{f''(a)}{2f'(a)}(f'(a)+1) \right|.$$

5.1.5 Metod polovljenja intervala

Neka je na segmentu $[\alpha, \beta]$ izolovan prost koren $x = a$ jednačine

$$f(x) = 0, \tag{1.5.1}$$

gde $f \in C[\alpha, \beta]$. Metod polovljenja intervala za rešavanje jednačine (1.5.1) sastoji se u konstrukciji niza intervala $\{(x_k, y_k)\}_{k \in \mathbb{N}}$ takvog da je

$$y_{k+1} - x_{k+1} = \frac{1}{2}(y_k - x_k), \quad k = 1, 2, \ldots,$$

i pri tome da je

$$\lim_{k \to +\infty} x_k = \lim_{k \to +\infty} y_k = a.$$

Navedeni proces konstrukcije intervala se prekida, na primer, kada dužina intervala postane manja od unapred zadatog malog pozitivnog broja ε.

Metod polovljenja intervala je veoma jednostavan i može se iskazati kroz sledeća četiri koraka:

1° $k := 0,\ x_1 := \alpha,\ y_1 := \beta;$

2° $k := k+1,\ z_k := \dfrac{1}{2}(x_k + y_k);$

3° ako je

$$f(z_k)f(x_k)\begin{cases} <0 & \text{uzeti} \quad x_{k+1} := x_k,\ y_{k+1} := z_k, \\ >0 & \qquad\quad x_{k+1} := z_k,\ y_{k+1} := y_k, \\ =0 & \qquad\quad \text{kraj izračunavanja } a := z_k; \end{cases}$$

4° ako je

$$|y_{k+1} - x_{k+1}|\begin{cases} \geq \varepsilon & \text{preći na } 2^\circ, \\ < \varepsilon & z_{k+1} := \frac{1}{2}(x_{k+1} + y_{k+1}) \\ & \text{kraj izračunavanja } a := z_{k+1}. \end{cases}$$

Primetimo da za grešku u aproksimaciji z_{k+1} važi ocena

$$|z_{k+1} - a| \leq \frac{1}{2^{k+1}}(\beta - \alpha).$$

5.1.6 SCHRÖDERov razvoj

Neka je funkcija $f : [\alpha, \beta] \to \mathbb{R}$ diferencijabilna i takva da je $f'(x) \neq 0$ za svako $x \in [\alpha, \beta]$. S obzirom na to da je tada f striktno monotona na $[\alpha, \beta]$ to postoji njena inverzna funkcija F koja je, takođe, diferencijabilna. Naime,

$$F'(y) = \frac{dx}{dy} = \frac{1}{f'(x)} \quad (y = f(x)).$$

Štaviše, ako je f dva puta diferencijabilna funkcija na $[\alpha, \beta]$, tada je

$$F''(y) = -\frac{f''(x)}{f'(x)^3}.$$

Problem nalaženja viših izvoda funkcije F, pod pretpostavkom da je ona dovoljan broj puta diferencijabilna, može biti vrlo komplikovan. Kako SCHRÖDERov[162] razvoj [118], o kome će biti reči u ovom odeljku, zahteva poznavanje viših izvoda

[162] ERNST SCHRÖDER (1841 – 1902), nemački matematičar.

funkcije F to ćemo najpre izložiti jedan rekurzivni postupak za rešavanje pomenutog problema.

Pretpostavimo da je funkcija f diferencijabilna $(n+1)$ puta na $[\alpha,\beta]$, kao i to da je

$$F^{(k)}(y)=\frac{X_k}{(f')^{2k-1}} \quad (k=1,\dots,n+1), \tag{1.6.1}$$

gde je X_k polinom po $f',f'',\dots,f^{(k)}$ i $f^{(i)}\equiv f^{(i)}(x)$ za $i=1,\dots,n+1$.

Primetimo da je formula (1.6.1) tačna za $k=1$ i $k=2$, pri čemu su $X_1=1$ i $X_2=-f''$.

Pretpostavimo sada da je formula (1.6.1) tačna za neko $k\in\{1,\dots,n\}$. Kako je X_k polinom po $f',\dots,f^{(k)}$ i

$$X_k'=\frac{dX_k}{dx}=\frac{d}{dx}X_k(f',\dots,f^{(k)})=\sum_{i=1}^{k}\frac{\partial X_k}{\partial f^{(i)}}f^{(i+1)}$$

polinom po $f',\dots,f^{(k+1)}$, to iz (1.6.1) sleduje

$$\begin{aligned}F^{(k+1)}(y)&=\frac{d}{dx}\left(\frac{X_k}{(f')^{2k-1}}\right)\frac{dx}{dy}\\&=\frac{f'X_k'-(2k-1)X_kf''}{(f')^{2k+1}}=\frac{X_{k+1}}{(f')^{2k+1}},\end{aligned}$$

gde je

$$X_{k+1}=f'X_k'-(2k-1)X_kf'' \tag{1.6.2}$$

polinom po $f',\dots,f^{(k+1)}$.

Izvodi funkcije F su dati sa (1.6.1), pri čemu se niz $\{X_k\}_{k\in\mathbb{N}}$ određuje pomoću rekurentne relacije (1.6.2) polazeći od $X_1=1$.

Prvih pet članova niza $\{X_k\}$ su

$$\begin{aligned}&X_1=1,\\&X_2=-f'',\\&X_3=-f'f'''+3f''^2,\\&X_4==-f'^2f^{(4)}+10f'f''f'''-15f''^3,\\&X_5=-f'^3f^{(5)}+15f'^2f''f^{(4)}+10f'^2f'''^2-105f'f''^2f'''+105f''^4.\end{aligned}$$

Pretpostavimo sada da funkcija f na segmentu $[\alpha,\beta]$ ima jednu prostu nulu $x=a$, čiju okolinu označimo sa $U(a)$. Ako stavimo $h=-f(x)/f'(x)$ $(x\in U(a))$, tada je $0=f(x)+hf'(x)$, odakle je

$$a=F(0)=F(h+hf').$$

Neka je $f\in C^{n+1}[\alpha,\beta]$. Tada, na osnovu TAYLORove formule, imamo

$$a=\sum_{k=0}^{n}\frac{1}{k!}F^{(k)}(y)(hf')^k+\frac{F^{(n+1)}(\overline{y})}{(n+1)!}(hf')^{n+1},$$

gde je $\overline{y}=f+thf'=(1-t)f=\theta f$ $(t,\theta\in(0,1))$. Najzad, korišćenjem formule (1.6.1) dobijamo SCHRÖDERov razvoj

$$a-x=\sum_{k=1}^{n}\frac{1}{k!}X_k\Big(\frac{f'}{f'},\frac{f''}{f'},\dots,\frac{f^{(k)}}{f'}\Big)h^k+O(f(x)^{n+1}),$$

tj.

$$a-x=h-\frac{f''}{2f'}h^2+\frac{3f''^2-f'f'''}{6f'^2}h^3 \tag{1.6.3}$$
$$+\frac{10f'f''f'''-f'^2f^{(4)}-15f''^3}{24f'^3}h^4+\cdots.$$

Napomena 1.6.1. Ako je funkcija f analitička, u prethodnom razvoju može se uzeti da $n\to+\infty$.

Ako za analitičku funkciju f (sa $f'(a)\neq 0$) definišemo koeficijente

$$c_k=\frac{1}{k!}\frac{f^{(k)}(a)}{f'(a)},\quad k=2,3,\dots, \tag{1.6.4}$$

i stavimo $e:=x-a$ $(x\in U(a))$, tada je (videti, na primer, [75])

$$\begin{aligned}\frac{f(x)}{f'(x)}&=e-c_2e^2+2(c_2^2-c_3)e^3-(4c_2^3-7c_3c_2+3c_4)e^4\\&\quad+(8c_2^4-20c_3c_2^2+10c_4c_2+6c_3^2-4c_5)e^5\\&\quad-[16c_2^5-52c_3c_2^3+28c_4c_2^2+(33c_3^2-13c_5)c_2-17c_3c_4+5c_6]e^6\\&\quad+O(e^7),\end{aligned} \tag{1.6.5}$$

što je, u stvari, inverzna formula od (1.6.3). Za dobijanje razvoja (1.6.5) možemo jednostavno koristititi pogodnost simboličkog izračunavanja što je obezbeđeno u softverskom paketu MATHEMATICA. Dakle, za

$$\frac{f(x)}{f'(x)} = \frac{f(a)+f'(a)e+\dfrac{f''(a)}{2!}e^2+\dfrac{f'''(a)}{3!}e^3+\cdots}{f'(a)+f''(a)e+\dfrac{f'''(a)}{2!}e^2+\cdots} = \frac{e+c_2e^2+c_3e^3+\cdots}{1+2c_2e+3c_3e^2+\cdots},$$

jednostavan kôd daje razvoj:

```
In[1]:= c[1] = 1; fkrozf1[e_, n_, m_] :=
         Series[Sum[c[k] e^k, {k, 1, n}] / Sum[k c[k] e^(k - 1), {k, 1, n}],
          {e, 0, m}]

In[2]:= fkrozf1[e, 6, 6]

Out[2]= e - c[2] e^2 + (2 c[2]^2 - 2 c[3]) e^3 + (-4 c[2]^3 + 7 c[2] c[3] - 3 c[4]) e^4 +
         (8 c[2]^4 - 20 c[2]^2 c[3] + 6 c[3]^2 + 10 c[2] c[4] - 4 c[5]) e^5 +
         (-16 c[2]^5 + 52 c[2]^3 c[3] - 33 c[2] c[3]^2 - 28 c[2]^2 c[4] + 17 c[3] c[4] +
          13 c[2] c[5] - 5 c[6]) e^6 + O[e]^7
```

Na primer, u slučaju NEWTONovog metoda (1.2.3), za koji je

$$\phi(x) = \phi_N(x) = x - \frac{f(x)}{f'(x)}, \tag{1.6.6}$$

imamo

$$\begin{aligned}\phi_N(x) - a = {} & c_2e^2 - 2(c_2^2 - c_3)e^3 + (4c_2^3 - 7c_3c_2 + 3c_4)e^4 \\ & -(8c_2^4 - 20c_3c_2^2 + 10c_4c_2 + 6c_3^2 - 4c_5)e^5 \\ & +[16c_2^5 - 52c_3c_2^3 + 28c_4c_2^2 + (33c_3^2 - 13c_5)c_2 \\ & \qquad -17c_3c_4 + 5c_6]e^6 + O(e^7).\end{aligned} \tag{1.6.7}$$

Formula (1.6.5) je korisna u asimptotskoj analizi iterativnih procesa.

5.1.7 Metodi višeg reda i računska efikasnost iterativnih procesa

U ovom odeljku ukazaćemo na neke načine za dobijanje iterativnih procesa, čiji je red konvergencije veći od dva, pri čemu pretpostavljamo da jednačina

$$f(x) = 0 \tag{1.7.1}$$

na segmentu $[\alpha, \beta]$ ima jedinstven prost koren $x = a$, kao i da je funkcija f dovoljan broj puta neprekidno diferencijabilna na $[\alpha, \beta]$. Kao što smo videli u odeljku 5.1.2, NEWTONov iterativni metod (1.2.3) ima red konvergencije $r = 2$ i pri tome zahteva dva funkcionalna izračunavanja po iteraciji, vrednost funkcije i vrednost njenog izvoda u tački x_k. STEFFENSENov metod (1.4.9) je drugog reda i zahteva takođe dva funkcionalna izračunavanja po iteraciji, $f(x_k)$ i $f(x_k + f(x_k))$. S druge strane, modifikovani NEWTONov metod (1.2.8) zahteva samo jedno izračunavanje funkcije po iteraciji, ali mu je zato konvergencija linearna. Kod metoda sečice (1.4.1), koji pripada klasi metoda sa memorijom (2.1.2), red konvergencije je $r = (1 + \sqrt{5})/2 \approx 1.62$, uz samo jedno funkcionalno izračunavanje $f(x_k)$, dok se druga neophodna vrednost $f(x_{k-1})$ prenosi iz prethodne iteracije. Nije teško zaključiti da se viši red konvergencije može postići većim brojem funkcionalnih izračunavanja (vrednosti funkcije f i/ili njenih izvoda) po iteraciji.

Dakle, kod određivanja korena jednačine (1.7.1), sa zahtevanom tačnošću, dobro je, s jedne strane, postići to sa manjim brojem iteracija, što se obezbeđuje bržom konvergencijom iterativnog procesa, ali, sa druge strane, i sa manjim brojem operacija po iteraciji, što je suprotno prethodnom zahtevu. Stoga je neophodno načiniti kompromis između ova dva zahteva uvođenjem tzv. *računske efikasnosti* (videti: TRAUB[163] [129, str. 260–264], OSTROWSKI [93, str. 20])

$$E = E(\phi, f) = r^{1/d}, \tag{1.7.2}$$

gde je r red konvergencije procesa definisanog iterativnom funkcijom ϕ i d ukupan broj novih funkcionalnih izračunavanja (vrednosti funkcije f i/ili njenih izvoda) koji se pojavljuju u iterativnoj funkciji ϕ.

U skladu sa definicijom (1.7.2), računske efikasnosti NEWTONovog i STEFFENSENovog metoda su jednake $2^{1/2} \approx 1.414$, dok je računska efikasnost metoda sečice veća i iznosi $(1 + \sqrt{5})/2 \approx 1.618$. Modifikovani NEWTONov metod (1.2.8) ima računsku efikasnost jednaku jedinici.

U daljem tekstu izložićemo neke standardne načine za dobijanje iterativnih procesa višeg reda.

1. Posmatramo SCHRÖDERov razvoj (1.6.3). Uzimajući konačan broj prvih članova na desnoj strani ovog razvoja možemo dobiti niz iterativnih formula:

[163] JOSEPH FREDERICK TRAUB (1932 –), američki naučnik u oblasti kompjuterskih nauka, rođen u Nemačkoj.

$$\phi_2(x) = x+h = x - \frac{f(x)}{f'(x)},$$

$$\phi_3(x) = \phi_2(x) - \frac{f''}{2f'}h^2 = x - \frac{f(x)}{f'(x)} - \frac{f''(x)f(x)^2}{2f'(x)^3},$$

$$\begin{aligned}\phi_4(x) &= \phi_3(x) + \frac{3f''^2 - f'f'''}{6f'^2}h^3 \\ &= x - \frac{f(x)}{f'(x)} - \frac{f''(x)f(x)^2}{2f'(x)^3} - \frac{f(x)^3}{6f'(x)^4}\left(3\frac{f''(x)^2}{f'(x)} - f'''(x)\right),\end{aligned}$$

itd.

Primetimo da je ϕ_2 iterativna funkcija NEWTONovog metoda.

Kako je, u prvoj aproksimaciji, h jednako $a-x$ $(x \to a)$, na osnovu (1.6.3) imamo

$$\phi_m(x) - a = O(h^m) = O((x-a)^m), \quad m = 2,3,\ldots,$$

kada $x \to a$, što znači da iterativni proces

$$x_{k+1} = \phi_m(x_k), \quad k = 0,1,\ldots, \tag{1.7.3}$$

primenjen na određivanje korena jednačine (1.7.1), ima red konvergencije najmanje m. Takođe, na osnovu SCHRÖDERovog razvoja možemo dobiti i vrednosti limesa

$$L_m = \lim_{k\to+\infty} \frac{x_{k+1}-a}{(x_k-a)^m} \quad (m = 2,3,\ldots).$$

Naime, imamo

$$\begin{aligned}L_2 &= \frac{f''(a)}{2f'(a)} = c_2, \\ L_3 &= \frac{1}{2}\left(\frac{f''(a)}{f'(a)}\right)^2 - \frac{1}{6}\frac{f'''(a)}{f'(a)} = 2c_2^2 - c_3, \\ L_4 &= \frac{5}{8}\left(\frac{f''(a)}{f'(a)}\right)^3 + \frac{1}{24}\frac{f^{iv}(a)}{f'(a)} - \frac{5}{12}\frac{f''(a)f'''(a)}{f'(a)^2} = 5c_2(c_2^2 - c_3) + c_4,\end{aligned}$$

gde su c_2, c_3, c_4 dati pomoću (1.6.4).

Formule (1.7.3) se često nazivaju ČEBIŠEVljeve iterativne formule. Njihove računske efikasnosti za $m = 3$ i $m = 4$ su $E(\phi_3, f) = 3^{1/3} \approx 1.442$ i $E(\phi_4, f) = 4^{1/4} \approx 1.414$, respektivno.

Korišćenjem HERMITEove interpolacione formule[164] za funkciju f u tačkama $x = x_{k-1}$ i $x = x_k$ može se dobiti iterativna formula (videti [114])

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f(x_k)^2}{2f'(x_k)}\overline{f''}(x_k), \quad k = 1,2,\ldots, \tag{1.7.4}$$

gde je

$$\overline{f''}(x_k) = -\frac{6}{\varepsilon_k^2}\big[f(x_k) - f(x_{k-1})\big] + \frac{2}{\varepsilon_k}\big[2f'(x_k) + f'(x_{k-1})\big]$$

i $\varepsilon_k = x_k - x_{k-1}$. Red konvergencije ovog procesa je $r = 1+\sqrt{3} \simeq 2.73$. Primetimo da iterativna funkcija ovog procesa predstavlja jednu modifikaciju ČEBIŠEVljeve funkcije ϕ_3. Računska efikasnost ovako dobijene formule sa memorijom je veća nego kod originalnog ČEBIŠEVljevog metoda i iznosi $(1+\sqrt{3})^{1/2} \approx 1.653$.

U radu [80] MILOVANOVIĆ i PETKOVIĆ[165] su razmatrali modifikaciju funkcije ϕ_3 korišćenjem aproksimacije

$$f''(x_k) \cong \frac{f'(x_k+\varepsilon_k) - f'(x_k)}{\varepsilon_k}$$

pri čemu $\varepsilon_k \to 0$, kada $k \to +\infty$. Odgovarajući iterativni proces je tada

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f(x_k)^2}{2f'(x_k)^3} \cdot \frac{f'(x_k+\varepsilon_k) - f'(x_k)}{\varepsilon_k}. \tag{1.7.5}$$

Teorema 1.7.1. *Neka $f \in C^3[\alpha,\beta]$ i neka je $\varepsilon_k = x_{k-1} - x_k$. Tada iterativni proces* (1.7.5) *ima red konvergencije $r = 1+\sqrt{2}$, tj. važi*

$$|x_{k+1} - a| \sim C_r |x_k - a|^r \quad (k \to +\infty),$$

gde je $C_r = |f'''(a)/(4f'(a))|^{1/\sqrt{2}}$.

Dokaz. Za $\varepsilon_k = x_{k-1} - x_k$, formula (1.7.5) postaje

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f(x_k)^2}{2f'(x_k)^3} \cdot \frac{f'(x_k) - f'(x_{k-1})}{x_k - x_{k-1}}. \tag{1.7.6}$$

[164] Interpolacioni procesi biće razmatrani u trećoj knjizi iz ove serije *Numerička analiza i teorija aproksimacija*. Inače, za niz detalja o interpolacionim procesima i mnogim primenama videti monografiju [65].

[165] MIODRAG S. PETKOVIĆ (1948 –), srpski matematičar.

Na osnovu SCHRÖDERovog razvoja (1.6.3), za x_k koje je dovoljno blisko korenu $x = a$, imamo

$$a - x_k = h_k - \frac{h_k^3 s_k}{2} + \frac{h_k^3}{6}(3s_k^2 - r_k) + O(h_k^4),$$

gde smo uveli notaciju

$$h_k = -\frac{f(x_k)}{f'(x_k)}, \quad s_k = \frac{f''(x_k)}{f'(x_k)}, \quad r_k = \frac{f'''(x_k)}{f'(x_k)}.$$

S druge strane, primenom TAYLORove formule, imamo

$$f'(x_{k-1}) = f'(x_k) + \varepsilon_k f''(x_k) + \frac{1}{2}\varepsilon_k^2 f'''(\zeta_k),$$

gde je $\zeta_k = x_k + \theta\varepsilon_k \ \ (0 < \theta < 1)$.

Kako je

$$\varepsilon_k = x_{k-1} - x_k = \frac{f(x_{k-1})}{f'(x_{k-1})} + O(h_{k-1}^2) = -h_{k-1} + O(h_{k-1}^2),$$

na osnovu prethodnog, imamo

$$x_{k-1} - a = \frac{h_k^2}{2f'(x_k)}\left[f''(x_k) + \frac{f'(x_k) - f'(x_{k-1})}{x_{k-1} - x_k} + O(h_k)\right],$$

tj.

$$\begin{aligned} x_{k+1} - a &= \frac{h_k^2}{4}\left[\frac{f'''(\zeta_k)}{f'(x_k)}\varepsilon_k + O(h_k)\right] \\ &= \frac{h_k^2}{4}\left[\frac{f'''(\zeta_k)}{f'(x_k)}\big(-h_{k-1} + O(h_{k-1}^2)\big) + O(h_k)\right]. \end{aligned}$$

Neka je dalje $e_k = x_k - a$. Kako je $e_k = -h_k + O(h_k^2)$, zaključujemo da je

$$h_k = -e_k + O(e_k^2) \quad \text{i} \quad h_{k-1} = -e_{k-1} + O(e_{k-1}^2).$$

Sada imamo

$$\begin{aligned} e_{k+1} &= \frac{f'''(\zeta_k)}{4f'(x_k)}\big(e_k + O(e_k^2)\big)^2\big(e_{k-1} + O(e_{k-1}^2) + O(e_k)\big) \\ &= \frac{f'''(\zeta_k)}{4f'(x_k)}\big(e_k^2 e_{k-1} + O(e_k^2 e_{k-1}^2)\big) \sim \frac{f'''(\zeta_k)}{4f'(x_k)} e_k^2 e_{k-1}. \end{aligned}$$

Kada $k \to +\infty$, $x_k \to a$, pri čemu je

$$|e_{k+1}| \sim \Big| \frac{f'''(a)}{4f'(a)} \Big| \, |e_k|^2 |e_{k-1}|.$$

Iz poslednje asimptotske relacije, slično kao kod metoda sečice (odeljak 5.1.4), nalazimo red konvergencije iz jednačine $r^2 - 2r - 1 = 0$. Dakle, $r = 1 + \sqrt{2}$. Asimptotska konstanta greške je

$$C_r = \Big| \frac{f'''(a)}{4f'(a)} \Big|^{1/\sqrt{2}}. \quad \square$$

Napomena 1.7.1. Iterativni proces (1.7.6) ima manji red konvergencije nego proces (1.7.4), a samim tim i nešto manju računsku efikasnost, tj. $(1+\sqrt{2})^{1/2} \approx 1.554$. S druge strane, iterativna formula procesa (1.7.6) je ipak nešto jednostavnija i zahteva ukupno manji broj operacija nego formula (1.7.4).

Slično prethodnoj teoremi, u pomenutom radu [80], dokazan je i sledeći rezultat.

Teorema 1.7.2. *Neka $f \in C^3[\alpha,\beta]$ i neka je $\varepsilon_k = f(x_k)$. Tada iterativni proces* (1.7.5) *ima red konvergencije tri, tj. važi*

$$|x_{k+1} - a| \sim C_3 |x_k - a|^3 \quad (k \to +\infty),$$

gde je

$$C_3 = \frac{1}{12f'(a)^2} \, \Big| \, 3f'(a)^2 f'''(a) + 2f'(a)f'''(a) - 6f''(a)^2 \, \Big| \, .$$

U slučajevima kada je izračunavanje izvoda jako komplikovano, uvođenjem aproksimacija za $f'(x)$ i $f''(x)$ pomoću

$$\text{(1.7.7)} \qquad \begin{cases} f'(x) \simeq \overline{f'}(x) = \dfrac{f(x+f(x)) - f(x-f(x))}{2f(x)} \\ f''(x) \simeq \overline{f''}(x) = \dfrac{f(x+f(x)) - 2f(x) + f(x-f(x))}{f(x)^2} \end{cases}$$

u radu [79] razmatrana je iterativna funkcija

$$\phi_3^*(x) = x - \frac{f(x)}{\overline{f'}(x)} - \frac{f(x)^2 \overline{f''}(x)}{2\overline{f'}(x)^3}$$

i dokazan sledeći rezultat.

Teorema 1.7.3. *Neka $f \in C^3[\alpha,\beta]$. Iterativni proces $x_{k+1} = \phi_3^*(x_k)$, $k = 0,1,\ldots$, ima red konvergencije $r = 3$, tj. važi*

$$L_3^* = \lim_{k\to+\infty} \frac{x_{k+1}-a}{(x_k-a)^3} = L_3 + \frac{1}{6} f'(a) f'''(a),$$

gde je L_3 dato ranije.

Računska efikasnost je ista kao i kod originalnog ČEBIŠEVljevog metoda za $m = 3$ i iznosi $E(\phi_3^*, f) = 3^{1/3} \approx 1.442$.

2. Primenimo sada metod, koji je definisan teoremom 2.4.1 iz odeljka 3.2.4, na ubrzavanje NEWTONovog procesa.

Kako je iterativna funkcija kod NEWTONovog procesa data sa $\phi(x) = x - f(x)/f'(x)$, imamo

$$x_{k+1} = x_k - \frac{x_k - \phi(x_k)}{1 - \frac{1}{2}\phi'(x_k)}, \quad k = 0,1,\ldots,$$

tj.

$$x_{k+1} = x_k - \frac{2f(x_k)f'(x_k)}{2f'(x_k)^2 - f(x_k)f''(x_k)}, \quad k = 0,1,\ldots\ . \tag{1.7.8}$$

Metod (1.7.8), na osnovu pomenute teoreme, ima red konvergencije najmanje tri. Ovaj metod je u literaturi poznat kao metod *tangentnih hiperbola* [115] ili kao HALLEYev[166] metod.

Za metod (1.7.8) lako se može dobiti asimptotska formula

$$\frac{x_{k+1}-a}{(x_k-a)^3} \sim \left(\frac{f''(a)}{2f'(a)}\right)^2 - \frac{f'''(a)}{6f'(a)} = c_2^2 - c_3 \quad (k \to +\infty),$$

gde su koeficijenti c_2 i c_3 definisani u (1.6.4).

Napomena 1.7.2. U radu [43] data je modifikacija metoda (1.7.8) korišćenjem aproksimacije $f''(x_k) \simeq (f'(x_k) - f'(x_{k-1}))/(x_k - x_{k-1})$. Može se dokazati da je

$$x_{k+1} - a \simeq \frac{f'''(a)}{4f'(a)}(x_k-a)^2(x_{k-1}-a) \quad (k \to +\infty),$$

odakle izlazi da je red konvergencije ovako modifikovanog metoda sa memorijom $r = 1+\sqrt{2} \simeq 2.414$. Njegova računska efikasnost iznosi $(1+\sqrt{2})^{1/2} \approx 1.554$, što je veće od računske efikasnosti HALLEYevog metoda, tj. od $3^{1/3} \approx 1.442$.

[166] EDMOND HALLEY (1656 – 1742), engleski astronom, geofizičar i matematičar, poznat po izračunavanju putanje komete koja danas nosi njegovo ime.

Korišćenjem aproksimacija (1.7.7), MILOVANOVIĆ i KOVAČEVIĆ[167] su u radu [79] razmatrali iterativnu funkciju

$$\psi_3^*(x) = x - \frac{2f(x)\overline{f'}(x)}{2\overline{f'}(x)^2 - f(x)\overline{f''}(x)},$$

koja predstavlja jednu modifikaciju HALLEYevog metoda (1.7.8). Konvergencija procesa je kubna, dok je računska efikasnost kao i kod HALLEYevog metoda.

3. Sukcesivna primena metoda, definisanog teoremom 2.4.4 (odeljak 3.2.4), na ubrzavanje Newtonovog metoda daje niz iterativnih formula, koje su po obliku slične ČEBIŠEVljevim formulama (1.7.3).

4. Izaberemo niz funkcija $\{\psi_i\}_{i=0,1,\ldots,n}$ dovoljan broj puta diferencijabilnih na $[\alpha,\beta]$. U radu [131] je predložena konstrukcija iterativnog procesa za rešavanje jednačine (1.7.1) u obliku

$$x_{k+1} = \phi_n(x_k), \quad k = 0,1,\ldots, \tag{1.7.9}$$

gde je

$$\phi_n(x) = x - \frac{D_n(x)}{\Delta_n(x)},$$

dok su funkcije Δ_n i D_n date pomoću

$$\Delta_n(x) = \begin{vmatrix} (f\psi_0)' & (f\psi_1)' & \ldots & (f\psi_n)' \\ (f\psi_0)'' & (f\psi_1)'' & & (f\psi_n)'' \\ \vdots & & & \\ (f\psi_0)^{(n+1)} & (f\psi_1)^{(n+1)} & & (f\psi_n)^{(n+1)} \end{vmatrix},$$

$$D_n(x) = \begin{vmatrix} f\psi_0 & f\psi_1 & \ldots & f\psi_n \\ (f\psi_0)'' & (f\psi_1)'' & & (f\psi_n)'' \\ \vdots & & & \\ (f\psi_0)^{(n+1)} & (f\psi_1)^{(n+1)} & & (f\psi_n)^{(n+1)} \end{vmatrix}.$$

Može se pokazati da je

$$\phi_n(a) = a,\ \phi_n'(a) = \phi_n''(a) = \cdots = \phi_n^{(n+1)}(a) = 0,\ \phi_n^{(n+2)}(a) \neq 0,$$

[167] MILAN A. KOVAČEVIĆ (1953 –), srpski matematičar.

što znači da iterativni proces (1.7.9) ima red konvergencije $n+2$.

Navešćemo nekoliko specijalnih slučajeva formule (1.7.9).

1° Ako je $n=0$ i $\psi_0(x)\equiv\psi(x)$, (1.7.9) se svodi na uopšten NEWTONov metod (1.2.9).

2° Ako je $n=1$, $\psi_0(x)\equiv 1$, $\psi_1(x)\equiv x$, (1.7.9) se svodi na (1.7.8).

3° Ako je $n=2$, $\psi_0(x)\equiv 1$, $\psi_1(x)\equiv x$, $\psi_2(x)\equiv x^2$, iz (1.7.9)) sleduje formula dobijena u radu [57] (videti, takođe, [58] i [59]), čija je iterativna funkcija data sa

$$\phi_2(x)=x-f(x)\begin{vmatrix} f'(x) & f(x) \\ \frac{1}{2}f''(x) & f'(x)\end{vmatrix}\cdot\begin{vmatrix} f'(x) & f(x) & 0 \\ \frac{1}{2}f''(x) & f'(x) & f(x) \\ \frac{1}{6}f'''(x) & \frac{1}{2}f''(x) & f'(x)\end{vmatrix}^{-1}. \tag{1.7.10}$$

5. Posmatrajmo formulu (1.2.10), gde je p parametar. Sa $U(a)$ označimo okolinu korena $x=a$ jednačine (1.7.1).

Teorema 1.7.4. *Neka je funkcija f četiri puta neprekidno–diferencijabilna u $U(a)$ i neka niz $\{x_k\}_{k\in\mathbb{N}_0}$ $(x_0\in U(a))$, definisan pomoću* (1.2.10)*, konvergira ka a. Tada, pri $k\to+\infty$, važi asimptotska formula*

$$x_{k+1}-a\sim\left[\frac{p}{x_k}+\frac{s_k}{2}\right]h_k^2+\left[\left(\frac{p}{x_k}\right)-\frac{1}{6}\left(3s_k^2-r_k\right)\right]h_k^3 + \left[\left(\frac{p}{x_k}\right)^3-\frac{1}{24}\left(10r_ks_k-t_k-15s_k^3\right)\right]h_k^4,$$

gde su

$$h_k=-\frac{f(x_k)}{f'(x_k)},\ \ s_k=\frac{f''(x_k)}{f'(x_k)},\ \ r_k=\frac{f'''(x_k)}{f'(x)},\ \ t_k=\frac{f^{iv}(x_k)}{f'(x)}.$$

Dokaz. Kako je, na osnovu (1.2.10),

$$x_{k+1}-x_k=\frac{x_kf(x_k)}{x_kf'(x_k)+pf(x_k)}=\frac{h_k}{1-\dfrac{p}{x_k}h_k}$$

i kako je $|ph_k/x_k|<1$ $(x_k\in U(a))$, nalazimo

$$x_{k+1}-x_k=h_k+\frac{p}{x_k}h_k^2+\left(\frac{p}{x_k}\right)^2h_k^3+\left(\frac{p}{x_k}\right)^3h_k^4+O(h_k^5).$$

S druge strane, na osnovu (1.6.3), imamo

$$a - x_k = h_k - \frac{1}{2}s_k h_k^2 + \frac{1}{6}(3s_k^2 - r_k)h_k^3 + \frac{1}{24}(10r_k s_k - t_k - 15s_k^3)h_k^4 + O(h_k^5).$$

Na osnovu poslednjih jednakosti dobijamo tvrđenje teoreme. □

Posmatrajmo sada formulu za dve vrednosti parametra p ($p = p_1$ i $p = p_2$), tj.

$$x_{k+1}^{(\nu)} = x_k\left(1 - \frac{f(x_k)}{x_k f'(x_k) + p_\nu f(x_k)}\right) \qquad (\nu = 1,2) \tag{1.7.11}$$

i stavimo

$$x_{k+1} = \frac{1}{2}\big(x_{k+1}^{(1)} + x_{k+1}^{(2)}\big). \tag{1.7.12}$$

Tada, na osnovu teoreme 1.7.4, imamo

$$\begin{aligned} x_{k+1} - a \simeq{}& \frac{1}{2}\left[\frac{p_1 + p_2}{x_k} + s_k\right]h_k^2 + \frac{1}{2}\left[\frac{p_1^2 + p_2^2}{x_k^2} - \frac{1}{3}\big(s_k^2 - r_k\big)\right]h_k^3 \\ &+ \frac{1}{2}\left[\frac{p_1^3 + p_2^3}{x_k^3} - \frac{1}{12}\big(10r_k s_k - t_k - 15s_k^3\big)\right]h_k^4. \end{aligned} \tag{1.7.13}$$

Ako u formulama (1.7.11) izaberemo parametre p_1 i p_2, takve da je

$$p_1 + p_2 = -s_k x_k\,, \quad p_1^2 + p_2^2 = \frac{1}{3}\big(3s_k^2 - r_k\big)x_k^2\,, \tag{1.7.14}$$

formula (1.7.13) postaje

$$x_{k+1} - a \sim C_k h_k^4 \quad (k \to +\infty),$$

gde smo sa C_k označili koeficijent u formuli (1.7.13) uz h_k^4, sa vrednostima za p_1 i p_2 koje se dobijaju iz (1.7.14). Naime, p_1 i p_2 su rešenja kvadratne jednačine

$$p^2 + s_k x_k p + \frac{1}{6}x_k^2 r_k = 0,$$

tj.

$$p^2 + x_k\frac{f''(x_k)}{f'(x_k)}\,p + \frac{1}{6}x_k^2\,\frac{f'''(x_k)}{f'(x_k)} = 0. \tag{1.7.15}$$

Na osnovu prethodnog može se formulisati i dokazati sledeća teorema (videti [20]).

Teorema 1.7.5. *Neka je funkcija f četiri puta neprekidno-diferencijabilna u $U(a)$ i neka su p_1 i p_2 rešenja jednačine* (1.7.15). *Tada iterativni proces definisan formulama* (1.7.11) *i* (1.7.12) *ima red konvergencije četiri, tj. važi*

$$x_{k+1} - a \sim C(x_k - a)^4 \quad (k \to +\infty),$$

gde je

$$C = \frac{3f''(a)^3 - 4f'(a)f''(a)f'''(a) + f'(a)^2 f^{iv}(a)}{24f'(a)^3}.$$

Napomena 1.7.3. Iterativni proces definisan formulama (1.7.11) i (1.7.12) može se eksplicitno izraziti u obliku

$$x_{k+1} = \phi(x_k), \quad k = 0, 1, \ldots,$$

gde je

$$\phi(x) = x - 3f(x)\frac{2f'(x)^2 - f(x)f''(x)}{6f'(x)\big(f'(x)^2 - f(x)f''(x)\big) + f(x)^2 f'''(x)}.$$

Primetimo da je ova formula ekvivalentna sa formulom (1.7.10).

Iterativni proces definisan formulama (1.7.11) i (1.7.12) pripada tzv. klasi metoda dvostrukog približavanja. Naime, kod ovakvih metoda jedna od aproksimacija $x_{k+1}^{(\nu)}$ $(\nu = 1, 2)$ je manja, a druga veća od korena $x = a$. U opštem slučaju, ako je

$$x_{k+1}^{(\nu)} = \phi_\nu(x_k) \qquad (\nu = 1, 2) \tag{1.7.16}$$

i

$$x_{k+1} = \frac{1}{2}\big(x_{k+1}^{(1)} + x_{k+1}^{(2)}\big), \quad k = 0, 1, \ldots, \tag{1.7.17}$$

i pri tome

$$\big(\phi_1(x_k) - a\big)\big(\phi_2(x_k) - a\big) < 0, \quad k = 0, 1, \ldots,$$

imamo metod dvostranog približavanja.

Jedan bolji izbor funkcija ϕ_ν $(\nu = 1, 2)$, nego što je (1.7.11), daje takođe metod četvrtog reda, ali sa manjim brojem numeričkih operacija [100]. Naime, ako uzmemo

$$\phi_1(x_k) = x_k + \frac{h_k}{1 + h_k s_k} \tag{1.7.18}$$

i

$$\phi_2(x_k) = x_k + h_k\Big(1 - \frac{1}{3}h_k^2 r_k\Big), \tag{1.7.19}$$

gde su veličine h_k, s_k, r_k definisane u teoremi 1.7.4, dobijamo metod dvostranog približavanja za koji važi sledeći rezultat.

Teorema 1.7.6. *Neka je funkcija f četiri puta neprekidno–diferencijabilna u $U(a)$. Tada iterativni proces, definisan pomoću* (1.7.16)–(1.7.19)*, ima red konvergencije četiri, tj. važi*

$$x_{k+1} - a \sim C(x_k - a)^4 \quad (k \to +\infty),$$

gde je

$$C = \frac{10f'(a)f''(a)f'''(a) - f'(a)^2 f^{iv}(a) - 27f''(a)^3}{24f'(a)^3}.$$

6. Iterativne formule trećeg reda za određivanje višestrukih korena jednačina razmatrane su u radovima [16], [31], [76].

5.1.8 Više-koračni iterativni metodi

U ovom odeljku izložićemo ukratko tzv. klasu više–koračnih metoda bez memorije za nalaženje prostog korena $x = a$ jednačine (1.7.1) na segmentu $[\alpha, \beta]$. Sledeći jednostavan primer ukazuje da ima smisla razmatrati više–koračne iterativne metode. Naime, na kraju odeljka 5.1.2 pomenuli smo dvo–koračni iterativni proces definisan formulama (1.2.11) koji ima kubnu konvergenciju, ali je, kako možemo izračunati, njegova računska efikasnost veća nego kod NEWTONovog metoda, s obzirom na to da je $3^{1/3} > 2^{1/2}$. Zaista, ako se dve iteracije NEWTONovog metoda posmatraju kao jedna iteracija, onda je red konvergencije $2^2 = 4$ i pri tome se zahtevaju četiri funkcionalna izračunavanja, što znači da je računska efikasnost procesa $4^{1/4} = 2^{1/2}$.

Još bolji primer je dvo–koračni iterativni proces OSTROWSKOG [93]

$$\begin{cases} y_k = x_k - \dfrac{f(x_k)}{f'(x_k)}, \\ x_{k+1} = y_k - \dfrac{f(x_k)}{f(x_k) - 2f(y_k)} \dfrac{f(y_k)}{f'(x_k)}, \quad k = 0, 1, \ldots, \end{cases} \tag{1.8.1}$$

koji takođe zahteva tri funkcionalna izračunavanja $f(x_k)$, $f'(x_k)$ i $f(y_k)$, a kako ćemo kasnije videti ima red konvergencije $r=4$, tako da mu je računska efikasnost $4^{1/3}\approx 1.587$.

Dakle, neka je pomoću n iterativnih funkcija $\Phi_\nu\colon [\alpha,\beta]^\nu \to [\alpha,\beta]$, $\nu = 1,2,\dots,n$, definisana k-ta iteracija n-koračnog iterativnnog metoda

$$\text{(1.8.2)}\qquad \begin{cases} x_{k,1} = \Phi_1(x_{k,0}),\\ x_{k,2} = \Phi_2(x_{k,0},x_{k,1}),\\ \quad\vdots\\ x_{k,n} = \Phi_n(x_{k,0},x_{k,1},\dots,x_{k,n-1}), \end{cases}$$

gde su $x_{k,0}=x_k$ i $x_{k,n}=x_{k+1}$. KUNG[168] i TRAUB [56] su 1974. godine postavili hipotezu da n-koračni metod (1.8.2) koji zahteva $n+1$ funkcijskih izračunavanja ima red konvergencije najviše 2^n. Za takav metod kažemo da je *optimalni n-koračni iterativni proces* i njegova računska efikasnost, na osnovu (1.7.2), iznosi $E=2^{n/(n+1)}$.

U daljem tekstu razmatraćemo samo neke dvo–koračne i tro–koračne iterative procese, pri čemu ćemo koristiti jednostavniju notaciju kao i kod procesa (1.8.1). U analizi konvergencije ovih procesa za nalaženje prostog korena $x=a$ jednačine $f(x)=0$, pored razvoja (1.6.5), tj.

$$\text{(1.8.3)}\qquad \frac{f(x_k)}{f'(x_k)} = e_k + a_2e_k^2 + a_3e_k^3 + a_4e_k^4 + a_5e_k^5 + \cdots, \quad e_k = x_k - a,$$

gde su

$$a_2 = -c_2, \quad a_3 = 2(c_2^2 - c_3), \quad a_4 = -4c_2^3 + 7c_3c_2 - 3c_4,$$
$$a_5 = 8c_2^4 - 20c_3c_2^2 + 10c_4c_2 + 6c_3^2 - 4c_5, \ \dots,$$

a koeficijenti c_k definisani pomoću (1.6.4), koristićemo i razvoj za $f(y_k)/f(x_k)$, gde se $y_k = \phi_N(x_k) = a + \widehat{e}_k$ određuje NEWTONovim metodom, kao i još neke razvoje istog tipa. Zahvaljujući simboličkom izračunavanju koje je obezbeđeno u paketu MATHEMATICA, pomenuti razvoji se jednostavno implementiraju. Pri ovome, pretpostavljamo da je funkcija f dovoljan broj puta diferencijabilna u okolini $U(a)$.

[168] HSIANG-TSUNG KUNG (1945 –), američki naučnik u oblasti kompjuterskih nauka, kineskog porekla.

Kako je $\widehat{e}_k = -(a_2e_k^2 + a_3e_k^3 + a_4e_k^4 + a_5e_k^5 + \cdots)$ za

$$u_k = \frac{f(y_k)}{f(x_k)} = \frac{f(a) + f'(a)\widehat{e}_k + \dfrac{f''(a)}{2!}\,\widehat{e}_k^2 + \cdots}{f(a) + f'(a)e_k + \dfrac{f''(a)}{2!}\,e_k^2 + \cdots} = \frac{\widehat{e}_k + c_2\widehat{e}_k^2 + \cdots}{e_k + c_2e_k^2 + \cdots}$$

važi razvoj

$$u_k = c_2e_k + (-3c_2^2 + 2c_3)e_k^2 + (8c_2^3 - 10c_2c_3 + 3c_4)e_k^3 + \cdots . \tag{1.8.4}$$

1. Dvo–koračni iterativni metodi. Metod OSTROWSKOG (1.8.1) je prvi dvokoračni optimalni iterativni proces. Pomenimo ovde još jedan takav metod koji je, na osnovu TRAUBove teorije [129, str. 197–204], dobijen u radu [44] (videti, takođe, [45])

$$x_{k+1} = x_k - \frac{1}{2}\frac{f(x_k)}{f'(x_k)} + \frac{f(x_k)}{f'(x_k) - 3f'\left(x_k - \frac{2}{3}f(x_k)/f'(x_k)\right)}, \quad k = 0, 1, \ldots ,$$

koji se može predstaviti i u obliku

$$\begin{cases} y_k = x_k - \dfrac{f(x_k)}{f'(x_k)}, \\ x_{k+1} = \dfrac{x_k + y_k}{2} + \dfrac{f(x_k)}{f'(x_k) - 3f'((x_k + 2y_k)/3)}, \quad k = 0, 1, \ldots . \end{cases} \tag{1.8.5}$$

Metod zahteva tri funkcionalna izračunavanja – izračunavanje vrednosti funkcije $f(x_k)$ i njenog izvoda u tačkama x_k i $(x_k + 2y_k)/3$. Za red konvergencije važi sledeći rezultat.

Teorema 1.8.1. *Neka je funkcija f dovoljan broj puta diferencijabilna u $U(a)$ i x_0 dovoljno blisko nuli $x = a$. Iterativni proces definisan formulama* (1.8.5) *ima red konvergencije četiri, za koji važi*

$$e_{k+1} = \left(c_2^3 - c_2c_3 + \frac{1}{9}c_4\right)e_k^4 + O(e_k^5), \tag{1.8.6}$$

gde je $e_k = x_k - a$ i koeficijenti c_k definisani pomoću (1.6.4).

Dokaz. Neka je $\widehat{e}_k = y_k - a$. Druga formula u (1.8.5) se može predstaviti u obliku

$$e_{k+1} = \frac{1}{2}(e_k + \widehat{e}_k) + \frac{f(x_k)}{f'(x_k)} \cdot \frac{1}{1 - 3\dfrac{f'(w_k)}{f'(x_k)}}, \tag{1.8.7}$$

gde je $w_k = (x_k + 2y_k)/3 = a + \theta_k$, $\theta_k = (e_k + 2\widehat{e}_k)/3$ i

$$\frac{f'(w_k)}{f'(x_k)} = \frac{f'(a) + f''(a)\theta_k + \dfrac{f'''(a)}{2!}\,\theta_k^2 + \cdots}{f'(a) + f''(a)e_k + \dfrac{f'''(a)}{2!}\,e_k^2 + \cdots} = \frac{1 + 2c_2\theta_k + 3c_3\theta_k^2 + \cdots}{1 + 2c_2 e_k + 3c_3 e_k^2 + \cdots},$$

tj.

$$\frac{f'(w_k)}{f'(x_k)} = 1 - \frac{4}{3}\,c_2 e_k + \Big(4c_2^2 - \frac{8}{3}\,c_3\Big)e_k^2 - \frac{8}{27}\left(36c_2^3 - 45c_2c_3 + 13c_4\right)e_k^3 + O(e_k^4).$$

Tada, korišćenjem (1.8.3) i razvoja (1.8.7) po e_k, dobijamo (1.8.6). □

U radu [53] dobijena je direktna (jednoparametarska) generalizacija metoda OSTROWSKOG (1.8.1).

Teorema 1.8.2. *Neka je funkcija f dovoljan broj puta diferencijabilna u $U(a)$ i x_0 dovoljno blisko nuli $x = a$. Dvo–koračni iterativni proces*

$$\begin{cases} \quad y_k = x_k - \dfrac{f(x_k)}{f'(x_k)}, \\[2ex] x_{k+1} = y_k - \dfrac{f(x_k) + bf(y_k)}{f(x_k) + (b-2)f(y_k)}\,\dfrac{f(y_k)}{f'(x_k)}, \quad k = 0, 1, \dots\ , \end{cases} \tag{1.8.8}$$

ima red konvergencije četiri, tj. važi

$$e_{k+1} = \left((1+2b)c_2^3 - c_2c_3\right)e_k^4 + O(e_k^5), \tag{1.8.9}$$

gde je $e_k = x_k - a$ i koeficijenti c_k definisani pomoću (1.6.4).

Dokaz. Kao i u dokazu prethodne teoreme neka je $\widehat{e}_k = y_k - a$. Druga formula u (1.8.8) se može predstaviti u obliku

$$e_{k+1} = \widehat{e}_k - \frac{f(x_k)}{f'(x_k)} \cdot \frac{(1 + bu_k)u_k}{1 + (b-2)u_k} = \widehat{e}_k - (e_k - \widehat{e}_k) \cdot \frac{(1 + b\,u_k)u_k}{1 + (b-2)u_k}, \tag{1.8.10}$$

gde je u_k određeno sa (1.8.4).

Razvijanjem desne strane u (1.8.10) po e_k dobijamo (1.8.9). □

Napomena 1.8.1. Za $b = 0$ metod (1.8.8) se svodi na metod OSTROWSKOG, za koji važi

$$e_{k+1} = \left(c_2^3 - c_2 c_3\right) e_k^4 + O(e_k^5).$$

Nedavno je u radu [101] generalisan metod (1.8.8), zamenom druge formule sa

$$x_{k+1} = y_k - g(u_k)\,\frac{f(y_k)}{f'(x_k)}, \quad u_k = \frac{f(y_k)}{f(x_k)},$$

gde je g dva puta neprekidno–diferencijabilna funkcija u okolini nule i zadovoljava uslove $g(0) = 1$ i $g'(0) = 2$. Tada je odgovarajući dvo–koračni metod optimalan, za koji važi

$$e_{k+1} = \left((5 - g''(0)/2)c_2^3 - c_2 c_3\right) e_k^4 + O(e_k^5).$$

U radu se, pored racionalne funkcije iz (1.8.8), navode i još neki primeri za funkciju g sa zahtevanim uslovima. M.S. PETKOVIĆ je sa saradnicima razmatrao i druge familije dvo–koračnih metoda u kojima je prva formula (NEWTONov metod) zamenjena STEFFENSENovim metodom (1.4.9) (videti, na primer, [102], [103]).

2. Tro–koračni iterativni metodi. Prva dva optimalna tro-koračna iterativna procesa proizilaze iz teorije koju su izveli KUNG i TRAUB u njihovom radu [56] iz 1974. Metodi su reda osam i zahtevaju četiri funkcionalna izračunavanja. Prvi od njih je

$$\begin{cases} y_k = x_k - \dfrac{f(x_k)}{f'(x_k)}, \\ z_k = y_k - \dfrac{f(x_k)^2 f(y_k)}{f'(x_k)(f(x_k) - f(y_k))^2}, \\ x_{k+1} = z_k - \dfrac{f(x_k)^2 f(y_k)}{f(y_k) - f(z_k)} \left[\dfrac{1}{f(x_k) - f(z_k)} \left(\dfrac{x_k - z_k}{f(x_k) - f(z_k)} - \dfrac{1}{f'(x_k)} \right) \right. \\ \qquad\qquad \left. - \dfrac{f(y_k)}{(f(x_k) - f(y_k))^2 f'(x_k)} \right], \quad k = 0, 1, \dots , \end{cases}$$

koji zahteva izračunavanje $f(x_k)$, $f'(x_k)$, $f(y_k)$ i $f(z_k)$, a drugi je bez upotrebe izvoda, zasnovan na primeni STEFFENSENovog metoda,

$$\begin{cases} y_k = x_k - \dfrac{f(x_k)^2}{f(x_k+f(x_k))-f(x_k)}, \\ z_k = y_k - \dfrac{f(x_k+f(x_k))f(y_k)}{(f(x_k+f(x_k))-f(x_k))f[x_k,y_k]}, \\ x_{k+1} = z_k - \dfrac{f(x_k+f(x_k))f(y_k)\,(y_k-x_k+f(x_k)/f[x_k,z_k])}{(f(y_k)-f(z_k))(f(x_k+f(x_k))-f(z_k))} + \dfrac{f(y_k)}{f[y_k,z_k]}, \end{cases}$$

$k = 0, 1, \ldots$, gde je $f[x,y] = (f(x)-f(y))/(x-y)$ oznaka za tzv. podeljenu razliku.

I pored toga što se u analizi konvergencije više–koračnih metoda od matematičkog aparata koristi samo TAYLORov razvoj, zbog glamaznosti izraza koji se pojavljuju u takvoj analizi, ova oblast je bila dugo zapostavljena. Buran razvoj informatičke tehnologije, uz pojavu mogućnosti simboličkih izračunavanja i korišćenja aritmetike proizvoljne preciznosti, omogućio je progres i u ovoj oblasti. Samo nakon 2000. godine pojavilo se nekoliko stotina radova koji tretiraju iterativne procese, uključujući i više–koračne, od kojih mnogi nisu optimalni u prethodno definisanom smislu. Na primer, u radu [85] autori su razmatrali više tro–koračnih metoda, koji u svim varijantama sadrži formulu

$$x_{k+1} = x_k - \frac{f(x_k)f'(x_k)}{f'(x_k)^2 - \lambda f(x_k)f''(x_k)},$$

koja se očigledno za $\lambda = 0$ svodi na NEWTONov, a za $\lambda = 1/2$ na HALLEYev method (1.7.8). Kao varijantu sa maksimalnim redom konvergencije oni daju trokoračni metod

$$\begin{cases} y_k = x_k - \dfrac{f(x_k)}{f'(x_k)}, \\ z_k = y_k - \dfrac{f(y_k)f'(y_k)}{f'(y_k)^2 - \frac{1}{2}f(y_k)f''(y_k)}, \\ x_{k+1} = z_k - \dfrac{(y_k-z_k)f(z_k)}{f(y_k)-2f(z_k)}, \end{cases} \tag{1.8.11}$$

dokazujući da važi $e_{k+1} = (-c_3c_2^5 + c_2^7)e_k^8 + O(e_k^9)$. Međutim, (1.8.11) zahteva šest funkcionalnih izračunavanja $f(x_k)$, $f(y_k)$, $f(z_k)$, $f'(x_k)$, $f'(y_k)$ i $f''(y_k)$. Bez dodatnih funkcionalnih izračunavanja, MILOVANOVIĆ i CVETKOVIĆ[169] [75] su dokazali da je mnogo bolji izbor iterativne formule

[169] ALEKSANDAR S. CVETKOVIĆ (1972 –), srpski matematičar.

$$x_{k+1} = S(y_k, z_k) = z_k - \frac{f(z_k)}{f'(y_k) + (z_k - y_k) f''(y_k)}$$

umesto treće formule u (1.8.11). U tom slučaju odgovarajući tro-koračni metod ima red konvergencije 10, sa asimptotikom

$$e_{k+1} = 3c_2^5 c_3 (c_3 - c_2^2) e_k^{10} + O(e_k^{11}).$$

U istom radu [75] razmatran je i tro–koračni metod

$$\text{(1.8.12)} \qquad \begin{cases} y_k &= x_k - \dfrac{f(x_k)}{f'(x_k)}, \\ z_k &= y_k - \dfrac{(x_k - y_k) f(y_k)}{f(x_k) - 2f(y_k)}, \\ x_{k+1} &= z_k - \dfrac{f(z_k) f'(z_k)}{f'(z_k)^2 - \dfrac{1}{2} f(z_k) \dfrac{f'(z_k) - f'(x_k)}{z_k - x_k}}, \end{cases}$$

kao i optimalna varijanta sa smanjenim brojem funkcionalnih izračunavanja na četiri, uvođenjem aproksimacije za $f'(z_k)$:

$$\text{(1.8.13)} \qquad f'(z_k) \approx \widetilde{f}'(z_k) = p_k f(x_k) + q_k f(y_k) + r_k f(z_k) + w_k f'(x_k),$$

gde su koeficijenti p_k, q_k, r_k i w_k dobijeni HERMITEovom interpolacijom[170] (videti [72, str. 51–58]) u obliku

$$p_k = \frac{(y_k - z_k)(z_k + 2y_k - 3x_k)}{(x_k - y_k)^2 (x_k - z_k)}, \quad q_k = \frac{(x_k - z_k)^2}{(x_k - y_k)^2 (y_k - z_k)},$$

$$r_k = \frac{3z_k - 2y_k - x_k}{(x_k - z_k)(y_k - z_k)}, \qquad w_k = \frac{y_k - z_k}{x_k - y_k}.$$

Teorema 1.8.3. *Neka je funkcija f dovoljan broj puta diferencijabilna u $U(a)$ i x_0 dovoljno blisko nuli $x = a$. Tro–korački iterativni proces* (1.8.12) *ima red konvergencije devet, pri čemu je*

$$e_{k+1} = -\frac{3}{2} c_3 c_2^2 (c_2^2 - c_3)^2 e_k^9 + O(e_k^{10}).$$

Ako je u trećoj formuli u (1.8.12) $f'(z_k)$ *zamenjeno sa* $\widetilde{f}'(z_k)$*, prema formuli* (1.8.13)*, odgovarajući tro–korački metod je optimalan sa redom konvergencije osam, pri čemu je*

[170] Interpolacioni procesi biće razmatrani u posebnoj knjizi iz ove serije.

$$e_{k+1} = (c_2^2 - c_3)c_2^2 c_4 e_k^8 + O(e_k^9),$$

gde je $e_k = x_k - a$ *i koeficijenti* c_k *definisani pomoću* (1.6.4).

Nekoliko godina nakon ovog optimalnog tro–koračnog metoda [75] pojavio se čitav niz tro–koračnih i više–koračnih metoda (videti, na primer, [6], [98], [104], [89], [101], [130], [136], [21]).

5.2 SISTEMI NELINEARNIH JEDNAČINA

5.2.1 Uvodne napomene

Ovo poglavlje je posvećeno problemu rešavanja sistema nelinearnih jednačina. Opštosti radi, najpre je tretiran slučaj rešavanja operatorskih jednačina u BANACHovom prostoru, a zatim je rezultat primenjen na sistem nelinearnih jednačina.

Kao što je rečeno u odeljku 3.1.1, sistem nelinearnih jednačina

$$f_i(x_1, \ldots, x_n) = 0, \quad i = 1, \ldots, n, \tag{2.1.1}$$

gde su $f_i : \mathbb{R}^n \to \mathbb{R}$ $(i = 1, \ldots, n)$ date funkcije, može se tretirati kao specijalan slučaj operatorske jednačine

$$Fu = \theta, \tag{2.1.2}$$

gde je F operator koji preslikava BANACHov prostor X u BANACHov prostor Y i θ nula–vektor prostora Y. Naime, treba samo uzeti da je $X = Y = \mathbb{R}^n$, $u = \boldsymbol{x} = [x_1 \ldots x_n]^T$, $\theta = [0 \ldots 0]^T$,

$$Fu = \boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_n(x_1, \ldots, x_n) \end{bmatrix}. \tag{2.1.3}$$

Osnovni metod za rešavanje operatorske jednačine (2.1.2), ali i za rešavanje sistema jednačina (2.1.1), je metod NEWTON–KANTOROVIČa, koji predstavlja uopštenje Newtonovog metoda (1.2.3). U našem izlaganju ovaj metod NEWTON–KANTOROVIČa biće razmatran za opšti slučaj operatorske jednačine (2.1.2).

S obzirom na to da metod NEWTON-KANTOROVIČa zahteva nalaženje inverznog operatora $[F'_{(u)}]^{-1}$, koje ponekad može biti komplikovano, ili čak nemoguće, u literaturi je u poslednje vreme razrađena čitava klasa tzv. kvazi–njutnovskih

metoda, kod kojih se koriste izvesne aproksimacije pomenutog operatora. U našem izlaganju, od ovih metoda obradićemo samo gradijentni metod, i njega ćemo prezentovati kao metod za minimizaciju izvesne funkcionele.

O pomenutim metodima postoji opširna literatura (videti, na primer, [10], [49], [51], [92], [113]).

5.2.2 Metod NEWTON-KANTOROVIČa

Osnovni iterativni metod za rešavanje jednačine (2.1.2) je metod NEWTON–KANTOROVIČa, koji predstavlja generalizaciju klasičnog NEWTONovog metoda (1.2.3). Fundamentalne rezultate vezane za egzistenciju i jedinstvenost rešenja jednačine (2.1.2) i konvergenciju metoda je dao KANTOROVIČ[171] (videti [49]). Takođe, ovaj metod je razmatran i od strane drugih autora (videti, na primer, [54], [87], [28], [135]).

Pretpostavimo da jednačina (2.1.2) ima rešenje $u = a$ i da je operator $F : X \to Y$ FRÉCHET–diferencijabilan u konveksnoj okolini D tačke a. Metod NEWTON–KANTOROVIČa se zasniva na linearizaciji jednačine (2.1.2). Naime, neka je nađeno približno rešenje u_k. Tada za nalaženje sledeće aproksimacije u_{k+1}, jednačinu (2.1.2) zamenimo jednačinom

$$Fu_k + F'_{(u_k)}(u - u_k) = \theta\,. \tag{2.2.1}$$

Ako za operator $F'_{(u_k)}$ postoji inverzni operator $\Gamma(u_k) = [F'_{(u_k)}]^{-1}$, iz (2.2.1) dolazimo do iterativnog metoda

$$u_{k+1} = u_k - \Gamma(u_k)Fu_k, \quad k = 0, 1, \ldots, \tag{2.2.2}$$

koji je poznat kao metod NEWTON-KANTOROVIČa. Početna vrednost u_0 za generisanje niza $\{u_k\}_{k\in\mathbb{N}}$ uzima se iz D, a njen izbor predstavlja dosta težak problem.

Metod (2.2.2) se može predstaviti u obliku

$$u_{k+1} = Tu_k, \quad k = 0, 1, \ldots,$$

gde je

$$Tu = u - \Gamma(u)Fu.$$

[171] LEONID VITALIYEVICH KANTOROVICH (na ruskom: Леонид Витальевич Канторович) (1912 – 1986), poznati sovjetski matematičar i ekonomista. Dobitnik je NOBELove nagrade u oblasti ekonomije za razvoj metoda za optimalnu alokaciju resursa.

Teorema 2.2.1. *Neka je operator F dva puta* FRÉCHET*–diferencijabilan na D, pri čemu za svako $u \in D$ postoji operator $\Gamma(u)$. Ako su operatori $\Gamma(u)$ i $F''_{(u)}$ ograničeni i $u_0 \in D$ dovoljno blisko tački a, iterativni proces* (2.2.2) *ima red konvergencije najmanje dva.*

Dokaz. Neka su $\|\Gamma(u)\| \le m$ i $\|F''_{(u)}\| \le M_2$ za svako $u \in D$. Na osnovu TAYLORove formule imamo

$$Fa = \theta = F_u + F'_{(u)}(a-u) + W(u, a-u), \tag{2.2.3}$$

gde je

$$\|W(u,a-u)\| \le \frac{1}{2} \sup_{t\in[0,1]} \|F''_{(u+t(a-u))}\| \cdot \|a-u\|^2 \le \frac{M_2}{2}\|u-a\|^2.$$

Kako je

$$Tu - a = u - \Gamma(u)Fu - a = \Gamma(u)\left[F'_{(u)}(u-a) - Fu\right],$$

na osnovu (2.2.3), dobijamo

$$Tu - a = \Gamma(u)W(u, a-u),$$

odakle sleduje

$$\|Tu-a\| \le \|\Gamma(u)\| \cdot \|W(u,a-u)\| \le \frac{1}{2} m M_2 \|u-a\|^2,$$

tj.

$$\|Tu-a\| = O\big(\|u-a\|^2\big). \quad \square$$

U daljem razmatranju smatraćemo da je D kugla $K[u_0, R]$, gde je u_0 početna vrednost niza $\{u_k\}_{k\in\mathbb{N}_0}$.

Ako je ispunjen LIPSCHTIZov uslov,

$$\big(u, v \in K[u_0, R]\big) \qquad \|F'_{(u)} - F'_{(v)}\| \le L\|u-v\|, \tag{2.2.4}$$

iz

$$Fu - Fv - F'_{(v)}(u-v) = \int_0^1 \big[F'_{(v+t(u-v))} - F'_{(v)}\big](u-v)\,\mathrm{d}t$$

sleduje nejednakost

$$\|Fu - Fv - F'_{(v)}(u-v)\| \le \frac{L}{2}\|u-v\|^2. \tag{2.2.5}$$

Teorema 2.2.2. *Neka je operator F* FRÉCHET*–diferencijabilan u kugli $K[u_0,R]$ i zadovoljava uslov* (2.2.4) *i neka su tačne nejednakosti*

$$\|\Gamma_0\| \le b_0\,, \quad \|\Gamma_0 F u_0\| \le \eta_0\,, \quad h_0 = b_0 L \eta_0 \le \frac{1}{2}\,, \tag{2.2.6}$$

gde je $\Gamma_0 = \Gamma(u_0)$.

Ako je

$$R \ge r_0 = \frac{1-\sqrt{1-2h_0}}{h_0}\,\eta_0\,, \tag{2.2.7}$$

niz $\{u_k\}_{k\in\mathbb{N}_0}$, koji se generiše pomoću (2.2.2)*, konvergira rešenju $a \in K[u_0,r_0]$ jednačine* (2.1.2).

Dokaz. Neka su nizovi $\{b_k\}$, $\{\eta_k\}$, $\{h_k\}$, $\{r_k\}$ definisani sa

$$b_{k+1} = \frac{b_k}{1-h_k}\,, \qquad \eta_{k+1} = \frac{h_k}{2(1-h_k)}\,\eta_k\,,$$

$$h_{k+1} = b_{k+1} L \eta_{k+1}\,, \quad r_k = \frac{1-\sqrt{1-2h_{k+1}}}{h_{k+1}}\,\eta_k\,.$$

Dokazaćemo da niz $\{u_k\}_{k\in\mathbb{N}_0}$ postoji i da je

$$\|\Gamma(u_k)\| \le b_k\,, \quad \|\Gamma(u_k)Fu_k\| \le \eta_k\,, \quad h_k \le \frac{1}{2}\,, \tag{2.2.8}$$

i

$$K[u_k,r_k] \subset K[u_{k-1},\, r_{k-1}]\,. \tag{2.2.9}$$

Dokaz izvodimo indukcijom.

Za $k=0$ nejednakosti (2.2.8) su tačne jer se svode na (2.2.6).

Pretpostavimo da su tačne i za $k=m$. Kako je, na osnovu (2.2.2) i induktivne pretpostavke

$$\|u_{m+1}-u_m\| = \|\Gamma(u_m)Fu_m\| \le \eta_m$$

i kako je $r_m > \eta_m$, sledi da $u_{m+1} \in K[u_m,r_m]$, a tim pre $u_m \in K[u_0,R]$, odakle zaključujemo da $F'_{(u_{m+1})}$ postoji.

Operator $\Gamma(u_{m+1})$ takođe postoji, s obzirom na to da se može predstaviti u obliku

$$\Gamma(u_{m+1}) = \left[I + \Gamma(u_m)\left(F'_{(u_{m+1})} - F'_{(u_m)}\right)\right]^{-1}\Gamma(u_m),$$

tj.

$$\Gamma(u_{m+1}) = \sum_{i=0}^{+\infty}(-1)^i\left[\Gamma(u_m)\left(F'_{(u_{m+1})} - F'_{(u_m)}\right)\right]^i\Gamma(u_m), \tag{2.2.10}$$

jer je, na osnovu (2.2.4),

$$\lambda = \left\| \Gamma(u_m)\left(F'_{(u_{m+1})} - F'_{(u_m)}\right) \right\| \le b_m L\|u_{m+1} - u_m\| \le h_m \le \frac{1}{2} < 1.$$

Iz (2.2.10) sleduje

$$\|\Gamma(u_{m+1})\| \le \sum_{i=0}^{\infty}\lambda^i b_m = \frac{b_m}{1-b_m} = b_{m+1}\,, \tag{2.2.11}$$

čime je dokazana prva nejednakost u (2.2.8).

Kako je $Fu_{m+1} = Fu_{m+1} - Fu_m - F'_{(u_m)}(u_{m+1} - u_m)$, na osnovu (2.2.5) sleduje nejednakost

$$\|Fu_{m+1}\| \le \frac{L}{2}\|u_{m+1} - u_m\|^2 \le \frac{L}{2}\,\eta_m^2\,, \tag{2.2.12}$$

odakle, na osnovu (2.2.11), dobijamo

$$\|\Gamma(u_{m+1})Fu_{m+1}\| \le \frac{b_m L\eta_m^2}{2(1-h_m)}\,\eta_m = \eta_{m+1}\,,$$

što predstavlja drugu nejednakost u (2.2.8) za $k = m+1$.

Kako je

$$h_{m+1} = B_{m+1}L\eta_{m+1} = \frac{b_m}{1-h_m}\,L\,\frac{h_m}{2(1-h_m)}\,\eta_m = \frac{1}{2}\Big(\frac{h_m}{1-h_m}\Big)^2 \le \frac{1}{2}\,,$$

dokazana je i treća nejednakost u (2.2.8).

Za dokaz inkluzije (2.2.9) pretpostavimo da tačka $u \in K[u_{m+1}, r_{m+1}]$. Tada iz $\|u - u_{m+1}\| \le r_{m+1}$ sleduje nejednakost

$$\|u - u_m\| \le \|u - u_{m+1}\| + \|u_{m+1} - u_m\| \le r_{m+1} + \eta_m\,. \tag{2.2.13}$$

S obzirom na to da je $r_{m+1} + \eta_m = r_m$ dokaz je završen.

Kako je $h_k \leq 1/2$ iz definicije niza $\{\eta_k\}$ sleduje nejednakost

$$\eta_{k+1} \leq \frac{1}{2}\eta_k\,,$$

odakle zaključujemo da je $\{\eta_k\}$ nula–niz. Tada je

$$\lim_{k\to+\infty} r_k = \lim_{k\to+\infty} \frac{1-\sqrt{1-2h_k}}{h_k}\,\eta_k = \lim_{k\to+\infty} \frac{2\eta_k}{1+\sqrt{1-2h_k}} = 0\,,$$

što znači da niz $\{u_k\}$ konvergira ka nekoj tački $a \in K[u_0, r_0]$. Tačka a je rešenje jednačine (2.1.2), s obzirom da je, na osnovu (2.2.12),

$$\lim_{m\to+\infty} \|Fu_{m+1}\| \leq \frac{L}{2} \lim_{m\to+\infty} \eta_m^2 = 0,$$

tj.

$$\lim_{m\to+\infty} Fu_m = \theta. \quad \square$$

Teorema 2.2.3. *Ako su ispunjeni uslovi prethodne teoreme, tada je*

$$\|u_k - a\| \leq \frac{1}{2^{k-1}}\, 2(h_0)^{2^k-1}\eta_0 \qquad (k \in \mathbb{N}). \tag{2.2.14}$$

Dokaz. Primetimo najpre da je

$$\frac{h_k}{1-h_k} \leq 2h_k \qquad \left(0 \leq h_k \leq \frac{1}{2}\right).$$

Tada je, na osnovu definicije nizova $\{\eta_k\}$ i $\{h_k\}$,

$$\eta_{k+1} \leq h_k\eta_k \quad \text{i} \quad h_{k+1} \leq 2h_k^2\,,$$

odakle sleduje

$$h_k \leq \frac{1}{2}(2h_0)^{2^k}$$

i

$$\begin{aligned}\eta_k &\leq h_{k-1}h_{k-2}\eta_{k-2} \leq \cdots \leq h_{k-1}h_{k-2}\cdots h_0\eta_0 \\ &\leq \frac{1}{2^k}\,(2h_0)^{1+2+\cdots+2^{k-1}}\eta_0 = \frac{1}{2^k}(2h_0)^{2^k-1}\eta_0\,.\end{aligned}$$

Najzad, na osnovu (2.2.13), za $u = a$ i $m = k$, imamo

$$\|u_k - a\| \leq r_k = \frac{2}{1+\sqrt{1-2h_k}}\eta_k \leq 2\eta_k\,,$$

tj. (2.2.14). $\square$

Da bi se izbeglo određivanje inverznog operatora $\Gamma(u) = [F'_{(u)}]^{-1}$ pri svakom koraku, metod NEWTON–KANTOROVIČa može se modifikovati na sledeći način

$$u_{k+1} = u_k - \Gamma_0 F u_k, \quad k = 0,1,2,\dots \,, \tag{2.2.15}$$

gde je $\Gamma_0 = \Gamma(u_0)$. Uvođenjem operatora T pomoću

$$Tu = u - \Gamma_0 F u, \tag{2.2.16}$$

modifikovani metod (2.2.15) se može predstaviti u obliku

$$u_{k+1} = Tu_k, \quad k = 0,1,2,\dots \,.$$

Pretpostavimo da su sada ispunjeni sledeći uslovi:

1° operator F je FRÉCHET–diferencijabilan u kugli $K[u_0,R]$,

2° $F'_{(u)}$ zadovoljava uslov (2.2.4),

3° operator Γ_0 postoji i

$$\|\Gamma_0\| \le b_0\,, \quad \|\Gamma_0 F u_0\| \le \eta_0\,.$$

Tada važi sledeća teorema.

Teorema 2.2.4. *Ako su ispunjeni uslovi*

$$h_0 = b_0 L \eta_0 \le \frac{1}{2}$$

i

$$r_0 = \frac{1-\sqrt{1-2h_0}}{h_0}\,\eta_0 \le R$$

niz koji se generiše pomoću (2.2.15) *konvergira ka rešenju* $a \in K[u_0,r_0]$ *jednačine* (2.1.1).

Dokaz. Na osnovu teoreme 2.2.2, jednačina (2.1.2) ima rešenje $a \in K[u_0,r_0]$. Dokažimo sada da je operator T, uveden pomoću (2.2.16), kontrakcija na $K[u_0,r_0]$.

Neka $u,v \in K[u_0,r_0]$. Tada iz

$$Tu - Tv = u - v - \Gamma_0(Fu - Fv) = \Gamma_0 \int_0^1 \left[F'_{(u_0)} - F'_{\left(v+t(u-v)\right)}\right](u-v)\,\mathrm{d}t \tag{2.2.17}$$

i uslova (2.2.4) sleduje

$$\|Tu - Tv\| \leq b_0 L r_0 \|u - v\|, \tag{2.2.18}$$

gde je

$$q = b_0 L r_0 = 1 - \sqrt{1 - 2h_0} < 1,$$

odakle zaključujemo da je T kontrakcija na kugli $K[u_0, r_0]$. Dokažimo još da je $TK[u_0, r_0] \subset K[u_0, r_0]$.

Neka je $u \in K[u_0, r_0]$. Tada, na osnovu (2.2.17) i (2.2.4), imamo

$$\|Tu - u_0\| \leq \|Tu - Tu_0\| + \|Tu_0 - u_0\| \leq \frac{b_0 L}{2}\|u - u_0\|^2 + \eta_0\,,$$

tj.

$$\|Tu - u_0\| \leq \frac{b_0 L r_0^2}{2} + \eta_0 = r_0\,.$$

S obzirom da $T : K[u_0, r_0] \to K[u_0, r_0]$, zaključujemo da jednačina (2.1.2) ima jedinstveno rešenje $a \in K[u_0, r_0]$. Tačka a predstavlja graničnu vrednost niza, koji se generiše pomoću modifikovanog metoda NEWTON–KANTOROVIČa. □

Nejednakost (2.2.18) za $v = a$, postaje

$$\|Tu - a\| \leq q\|u - a\|\,,$$

što znači da je iterativni proces (2.2.15) prvog reda.

5.2.3 Metod NEWTON–KANTOROVIČa za sistem nelinearnih jednačina

U ovom odeljku daćemo primenu metoda NEWTON–KANTOROVIČa na rešavanje sistema nelinearnih jednačina

$$f_i(x_1, \ldots, x_n) = 0, \quad i = 1, \ldots, n. \tag{2.3.1}$$

Ovde je $X = Y = \mathbb{R}^n$, $u = \boldsymbol{x} = [x_1 \ldots x_n]^T$ i F definisano pomoću (2.1.3). Ako je F FRÉCHET–diferencijabilan operator (videti primer 2.4.2 iz odeljka 2.2.4), tada je

$$F'_{(u)} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & & \\ \dfrac{\partial f_n}{\partial x_1} & & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix} = W(\boldsymbol{x})\,,$$

tj. $W(\boldsymbol{x})$ je JACOBIeva matrica za $\boldsymbol{f} = [f_1 \ \dots \ f_n]^T$. Ako je $\det\big(W(\boldsymbol{x})\big) \neq 0$, tj. ako je matrica $W(\boldsymbol{x})$ regularna, metod NEWTON–KANTOROVIČa za rešavanje sistema jednačina (2.3.1) je dat sa

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - W^{-1}\big(\boldsymbol{x}^{(k)}\big)\boldsymbol{f}\big(\boldsymbol{x}^{(k)}\big), \quad k = 0, 1, \dots, \tag{2.3.2}$$

gde je $\boldsymbol{x}^{(k)} = \big[x_1^{(k)} \ \dots \ x_n^{(k)}\big]^T$. Vrlo često se u literaturi ovaj metod sreće kao NEWTON–RAPHSONov metod.

Do metoda (2.3.2) možemo doći i znatno elementarnije, linearizacijom sistema jednačina (2.3.1) u okolini aproksimacije $\boldsymbol{x}^{(k)}$. Neka je $\boldsymbol{a} = [a_1 \ \dots \ a_n]^T$ tačno rešenje datog sistema jednačina. Korišćenjem TAYLORovog razvoja za funkcije koje se pojavljuju u (2.3.1) dobijamo

$$\begin{aligned} f_i(a_1, \dots, a_n) = f_i\big(x_1^{(k)}, \dots, x_n^{(k)}\big) &+ \frac{\partial f_i}{\partial x_1}\big(a_1 - x_1^{(k)}\big) + \cdots \\ &+ \frac{\partial f_i}{\partial x_n}\big(a_n - x_n^{(k)}\big) + r_i^{(k)}, \qquad i = 1, \dots, n, \end{aligned}$$

gde se parcijalni izvodi na desnoj strani ovih jednakosti izračunavaju u tački $\boldsymbol{x}^{(k)}$. Veličina $r_i^{(k)}$ je odgovarajući ostatak u TAYLORovoj formuli.

Kako je $f_i(a_1, \dots, a_n) = 0$, $i = 1, \dots, n$, prethodni sistem jednakosti se može predstaviti u matričnom obliku

$$\boldsymbol{0} = \boldsymbol{f}\big(\boldsymbol{x}^{(k)}\big) + W\big(\boldsymbol{x}^{(k)}\big)\big(\boldsymbol{a} - \boldsymbol{x}^{(k)}\big) + \boldsymbol{r}^{(k)},$$

gde je $\boldsymbol{r}^{(k)} = \big[r_1^{(k)} \ \dots \ r_n^{(k)}\big]^T$. Ako je JACOBIeva matrica za $\boldsymbol{f}$ regularna, tada imamo

$$\boldsymbol{a} = \boldsymbol{x}^{(k)} - W^{-1}\big(\boldsymbol{x}^{(k)}\big)\boldsymbol{f}\big(\boldsymbol{x}^{(k)}\big) - W^{-1}\big(\boldsymbol{x}^{(k)}\big)\boldsymbol{r}^{(k)}.$$

Zanemarivanjem poslednjeg člana na desnoj strani ove jednakosti, umesto vektora $\boldsymbol{a}$ dobićemo njegovu novu aproksimaciju, koju ćemo označiti sa $\boldsymbol{x}^{(k+1)}$. Tako dobijamo (2.3.2).

Kao što smo u prethodnom izlaganju videli, metod (2.3.2) se može modifikovati u smislu da se inverzna matrica od $W(\boldsymbol{x})$ ne određuje u svakoj iteraciji, već samo u prvoj. Dakle,

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - W^{-1}\big(\boldsymbol{x}^{(0)}\big)\boldsymbol{f}\big(\boldsymbol{x}^{(k)}\big), \quad k = 0, 1, \dots, \tag{2.3.3}$$

Napomena 2.3.1. Modifikovani metod (2.3.3) se može shvatiti kao metod proste iteracije

$$\boldsymbol{x}^{(k+1)} = T\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k)} + \Lambda \boldsymbol{f}\left(\boldsymbol{x}^{(k)}\right), \quad k = 0, 1, \ldots,$$

sa matricom Λ određenom iz uslova da je izvod od operatora T nula–operator, tj. da je $I + \Lambda W\left(\boldsymbol{x}^{(0)}\right)$ nula matrica. Ako je $W\left(\boldsymbol{x}^{(0)}\right)$ regularna matrica, tada imamo $\Lambda = W^{-1}\left(\boldsymbol{x}^{(0)}\right)$.

Teoreme dokazane u prethodnom odeljku se mogu prilagoditi za slučaj sistema nelinearnih jednačina, pri čemu se uslovi za konvergenciju procesa (2.3.2) i (2.3.3) mogu iskazati na različite načine, što zavisi od uvedene norme u X. Tako na primer, uzimajući za normu u $\mathbb{R}^n$

$$\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_\infty = \max_i |x_i|$$

i pretpostavljajući da $f \in C^2(D)$, gde je D kugla $K[x^{(0)}, R]$, iz teoreme 2.2.2 sleduje rezultat.

Posledica 2.3.1. *Neka su u D ispunjeni sledeći uslovi:*

$$s_{ij} = \sum_{k=1}^{n} \left\| \frac{\partial^2 f_i}{\partial x_j \partial x_k} \right\| \leq N \quad (i, j = 1, \ldots, n); \tag{2.3.4}$$

$$\left\| \boldsymbol{f}\left(\boldsymbol{x}^{(0)}\right) \right\| \leq Q, \quad \left\| W^{-1}\left(\boldsymbol{x}^{(0)}\right) \right\| \leq b; \tag{2.3.5}$$

$$\Delta_0 = \det W\left(\boldsymbol{x}^{(0)}\right) \neq 0, \quad h = nNQb^2 \leq \frac{1}{2}. \tag{2.3.6}$$

Tada, ako je

$$R \geq r = \frac{1 - \sqrt{1 - 2h}}{h} Qb,$$

metod NEWTON–KANTOROVIČ*a* (2.3.2) *konvergira ka rešenju* $a \in K[\boldsymbol{x}^{(0)}, r]$.

Dokaz. Neka je $a_{ij} = \left.\dfrac{\partial f_i}{\partial x_j}\right|_{\boldsymbol{x}=\boldsymbol{x}^{(0)}}$ i neka su sa A_{ij} označeni kofaktori elemenata a_{ij} JACOBIeve matrice $W\left(\boldsymbol{x}^{(0)}\right) = [a_{ij}]$. Tada je, na osnovu (2.3.5),

$$\|\Gamma_0\| = \left\| W^{-1}\left(\boldsymbol{x}^{(0)}\right) \right\| = \max_i \frac{1}{|\Delta_0|} \sum_{j=1}^{n} |A_{ji}| \leq b.$$

Dalje, na osnovu saglasnosti norme matrice sa normom vektora, imamo

$$\| W^{-1}(\boldsymbol{x}^{(0)})\boldsymbol{f}(\boldsymbol{x}^{(0)}) \| \le \| W^{-1}(\boldsymbol{x}^{(0)}) \| \cdot \| \boldsymbol{f}(\boldsymbol{x}^{(0)}) \| \le Qb\,.$$

Primetimo da pod uslovom (2.3.4), $W(\boldsymbol{x})$ zadovoljava LIPSCHTIZov uslov (2.2.4) sa konstantom $L = nN$. Zaista, za dva proizvoljna vektora $\boldsymbol{x} = [x_1 \ldots x_n]^T$ i $\boldsymbol{y} = [y_1 \ldots y_n]^T$ imamo

$$\begin{aligned}
\|W(\boldsymbol{x}) - W(\boldsymbol{y})\| &= \max_i \sum_{j=1}^{n} \left| \frac{\partial f_i(\boldsymbol{x})}{\partial x_j} - \frac{\partial f_i(\boldsymbol{y})}{\partial x_j} \right| \\
&= \max_i \sum_{j=1}^{n} \left| \sum_{j=1}^{n} \frac{\partial^2 f_i(\xi)}{\partial x_j \partial x_k} (x_k - y_k) \right| \\
&= \max_i \sum_{j=1}^{n} \big(N\|\boldsymbol{x} - \boldsymbol{y}\| \big) = nN\|\boldsymbol{x} - \boldsymbol{y}\|\,,
\end{aligned}$$

gde je $\boldsymbol{\xi} = \boldsymbol{y} + \theta(\boldsymbol{x} - \boldsymbol{y}),\ 0 < \theta < 1$.

Na osnovu prethodnog vidimo da su, pod uslovima (2.3.6), ispunjeni uslovi (2.2.6) u teoremi 2.2.2, pa je ovaj rezultat posledica teoreme 2.2.2. □

Napomena 2.3.2. Kako za $0 < h \le 1/2$ važi $(1 - \sqrt{1-2h})/h \le 2$, za r u posledici 2.3.1 možemo uzeti $r = 2Qb$.

Napomena 2.3.3. Modifikovani metod NEWTON–KANTOROVIČa (2.3.3) konvergira, takođe, pod uslovima datim u posledici 2.3.1.

Primer 2.3.1. Posmatrajmo sistem nelinearnih jednačina

$$\begin{aligned}
f_1(x_1, x_2) &= 9x_1^2 x_2 + 4x_2^2 - 36 = 0, \\
f_2(x_1, x_2) &= 16x_2^2 - x_1^4 + x_2 + 1 = 0,
\end{aligned}$$

koji ima rešenje u prvom kvadrantu $(x_1, x_2 > 0)$. Da bismo se u to uverili dati sistem jednačina predstavimo u obliku

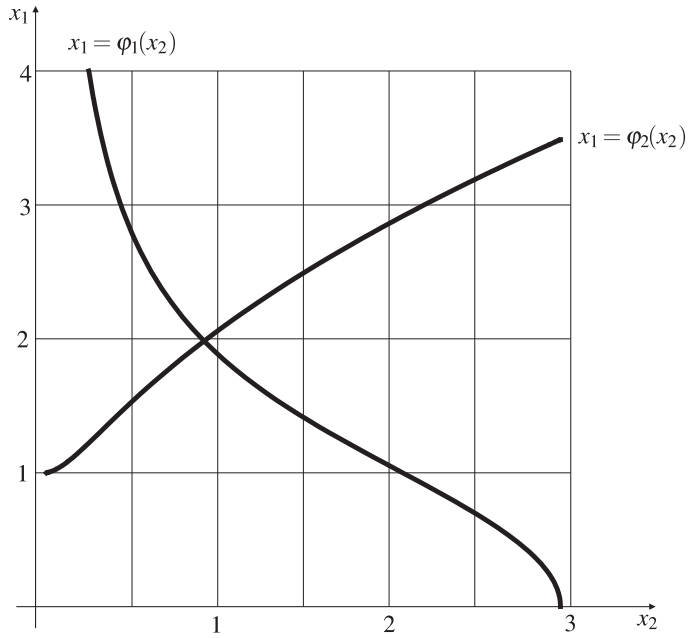
$$x_1^2 = \frac{4}{9x_2}\big(9 - x_2^2\big)\,, \quad x_1^4 = 1 + x_2 + 16x_2^2\,,$$

odakle nalazimo grane datih implicitnih funkcija u prvom kvadrantu

$$x_1 = \varphi_1(x_2) = \frac{2}{3}\sqrt{\frac{9-x_2^2}{x_2}} \qquad (0 < x_2 \leq 3),$$

$$x_1 = \varphi_2(x_2) = \sqrt{1+x_2+16x_2^2} \qquad (x_2 \geq 0),$$

koje su prikazane na slici 2.3.1.



Slika 2.3.1. Grafici funkcija $x_1 = \varphi_\nu(x_2)$, $\nu = 1,2$, u prvom kvadrantu

Kako se rešenje datog sistema jednačina $\boldsymbol{a} = (a_1, a_2)$ nalazi u okolini tačke $(2,1)$ (videti sliku 2.3.1), za početni vektor ćemo uzeti $\boldsymbol{x}^{(0)} = [2\ 1]^T$, tj. $x_1^{(0)} = 2$ i $x_2^{(0)} = 1$.

Kako su

$$\frac{\partial f_1}{\partial x_1} = 18x_1x_2, \quad \frac{\partial f_1}{\partial x_2} = 9x_1^2 + 8x_2, \quad \frac{\partial f_2}{\partial x_1} = -4x_1^3, \quad \frac{\partial f_2}{\partial x_2} = 32x_2 + 1;$$

$$\frac{\partial^2 f_1}{\partial x_1^2} = 18x_2, \quad \frac{\partial f_1^2}{\partial x_1 \partial x_2} = 18x_1, \quad \frac{\partial^2 f_1}{\partial x_2^2} = 8,$$

$$\frac{\partial^2 f_2}{\partial x_1^2} = -12x_1^2, \quad \frac{\partial^2 f_2}{\partial x_1 \partial x_2} = 0, \quad \frac{\partial^2 f_2}{\partial x_2^2} = 32,$$

imamo

$$W(\boldsymbol{x}) = \begin{bmatrix} 18x_1x_2 & 9x_1^2 + 8x_2 \\ -4x_1^3 & 32x_2 + 1 \end{bmatrix},$$

$$W^{-1}(\boldsymbol{x}) = \frac{1}{\Delta(\boldsymbol{x})} \begin{bmatrix} 32x_2 + 1 & -(9x_1^2 + 8x_2) \\ 4x_1^3 & 18x_1x_2 \end{bmatrix},$$

gde je

$$\Delta(\boldsymbol{x}) = 18x_1x_2(32x_2 + 1) + 4x_1^3(9x_1^2 + 8x_2) > 0.$$

Veličine s_{ij}, koje se pojavljuju u (2.3.4), su u ovom primeru

$$s_{11} = 18(x_1 + x_2), \quad s_{12} = 2(9x_1 + 4), \quad s_{21} = 12x_1^2, \quad s_{22} = 32.$$

Na dalje, stavimo $f_i^{(k)} \equiv f_i\big(x_1^{(k)}, x_2^{(k)}\big)$, $i = 1,2$, i $\Delta_k = \Delta\big(x^{(k)}\big)$, $k = 0,1,\dots$. Skalarni oblik metoda NEWTON–KANTOROVIČa (2.3.2) je

$$x_1^{(k+1)} = x_1^{(k)} - \frac{1}{\Delta_k}\Big\{\big(32x_2^{(k)} + 1\big)f_1^{(k)} - \big(9(x_1^{(k)})^2 + 8x_2^{(k)}\big)f_2^{(k)}\Big\},$$

$$x_2^{(k+1)} = x_2^{(k)} - \frac{1}{\Delta_k}\Big\{4(x_1^{(k)})^3 f_1^{(k)} + 18x_1^{(k)}x_2^{(k)} f_2^{(k)}\Big\}.$$

Kako su

$$\|\boldsymbol{f}\big(\boldsymbol{x}^{(0)}\big)\| = \max\Big\{\big|\, f_1^{(0)}\,\big|, \big|\, f_2^{(0)}\,\big|\Big\} = \max\{4,\ 2\} = 4\,,$$

$$W^{-1}\big(\boldsymbol{x}^{(0)}\big) = \frac{1}{2596}\begin{bmatrix} 33 & -44 \\ 32 & 36 \end{bmatrix},$$

$$\|W^{-1}\big(\boldsymbol{x}^{(0)}\big)\| = \frac{1}{2596}\max\{33 + 44,\ 32 + 36\} = \frac{77}{2596} < 0.03,$$

tj. $Q = 4,\ b = 0.03$, saglasno napomeni 2.3.2, za r se može uzeti $r = 2Qb = 0.24$. Tada uslov da se $\boldsymbol{x}$ nađe u kugli $K[\boldsymbol{x}^{(0)}, r]$, tj. uslov

$$\|\boldsymbol{x} - \boldsymbol{x}^{(0)}\| = \max_i \left| x_i - x_i^{(0)} \right| \leq r = 0.24,$$

daje $1.76 \leq x_1 \leq 2.24$ i $0.76 \leq x_2 \leq 1.24$, pa imamo

$$s_{11} \leq 18(1.24 + 2.24) = 62.64, \quad s_{12} = 2(9 \cdot 2.24 + 4) = 48.32,$$
$$s_{21} \leq 12 \cdot 2.24^2 = 60.2112, \quad s_{22} = 32.$$

Dakle, možemo uzeti $N = 63$.

Kako je $h = nNQb^2 = 2 \cdot 63 \cdot 4 \cdot 0.03^2 = 0.4536 < 1/2$, zaključujemo da će prethodno dati iterativni proces, za izabrani početni vektor, konvergirati ka rešenju. Rezultati dobijeni primenom ovog procesa dati su u tabeli 2.3.1. Kao i uvek, brojevi u zagradama označavaju decimalni eksponent. Primetimo da se poslednje dve iteracije poklapaju na 12 značajnih cifara. Inače, poslednja iteracija daje tačno rešenje datog sistema jednačina, zaokrugljeno na 16 cifara. △

Tabela 2.3.1.

k	$x_1^{(k)}$	$x_2^{(k)}$	$f_1^{(k)}$	$f_2^{(k)}$
0	2.	1.	4.00	2.00
1	1.983050847457627	0.9229583975346687	7.31(−2)	8.81(−2)
2	1.983707108973573	0.9207432150674075	−2.87(−5)	6.83(−5)
3	1.983708733954053	0.9207426370180257	−1.03(−11)	−5.70(−11)
4	1.983708733953144	0.9207426370189653		

Poseban problem kod primene metoda NEWTON–KANTOROVIČa je izbor startnih vrednostu, tj. početnog vektora $\boldsymbol{x}^{(0)}$, koji bi obezbedio konvergenciju procesa. Jedno rešenje ovog problema dato je 1974. godine u radu [55]. Naime, za njegovo nalaženje predlažen je iterativni proces

$$\boldsymbol{x}_{m+1} = \boldsymbol{x}_m - W^{-1}(\boldsymbol{x}_m)\big(\boldsymbol{f}(\boldsymbol{x}_m) - \alpha_m \boldsymbol{f}(\boldsymbol{x}_0)\big), \quad m = 0, 1, \ldots, \tag{2.3.7}$$

gde je

$$\alpha_m = \max\left\{0,\ 1 - \frac{1}{2nN\|\boldsymbol{f}(\boldsymbol{x}_0)\|}\left(\frac{1}{\|W^{-1}(\boldsymbol{x}_m)\|^2} + \frac{3}{4}\sum_{i<m}\frac{1}{\|W^{-1}(\boldsymbol{x}_i)\|^2}\right)\right\},$$

startujući sa nekom grubom aproksimacijom $\boldsymbol{x}_0$. Proces (2.3.7) se može interpretirati kao iteracija metoda NEWTON–KANTOROVIČa iz tačke $\boldsymbol{x}_m$ na sistem

$$\boldsymbol{f}(\boldsymbol{x}) = \alpha_m \boldsymbol{f}(\boldsymbol{x}_0).$$

Primetimo da $\alpha_m \in [0,1]$. Ako je $\det W(\boldsymbol{x}_0) \neq 0$, tada se može pokazati da je, takođe, $\det W(\boldsymbol{x}_m) \neq 0$ i da je

$$2nN\|W^{-1}(\boldsymbol{x}_m)\|^2 \|\boldsymbol{f}(\boldsymbol{x}_m) - \alpha_m \boldsymbol{f}(\boldsymbol{x}_m)\| \leq 1,$$

gde je $\{\alpha_m\}$ nerastući niz, tj. $\alpha_{m+1} \leq \alpha_m$.

Neka je $G(\subset \mathbb{R}^n)$ konveksna oblast koja sadrži rešenje $\boldsymbol{a}$. Pod pretpostavkama

$$\boldsymbol{f}(\boldsymbol{x}) \in C^2(G), \quad \det W(\boldsymbol{x}) \neq 0 \ \ (\boldsymbol{x} \in G), \quad \boldsymbol{f}(\boldsymbol{x}) \neq \boldsymbol{f}(\boldsymbol{y}) \Leftrightarrow \boldsymbol{x} \neq \boldsymbol{y} \ \ (\boldsymbol{x}, \boldsymbol{y} \in G),$$

u pomentom radu [55] dokazana je sledeća teorema.

Teorema 2.3.1. *Za svako $\boldsymbol{x}_0 \in G$, posle konačnog broja iteracija s u* (2.3.7)*, dolazi se do $\boldsymbol{x}_s$ za koji su ispunjeni uslovi za konvergenciju metoda* NEWTON–KANTOROVIČ*a, pri čemu je $\alpha_m = 0$ za svako $m \geq s$.*

Dakle, kao početni vektor $\boldsymbol{x}^{(0)}$ za metod NEWTON–KANTOROVIČa može se uzeti vektor $\boldsymbol{x}_s$, tj. $\boldsymbol{x}^{(0)} = \boldsymbol{x}_s$.

U odeljku 5.2.5 ramatraćemo jedan opšti princip kojim se razrešava problem startnih vrednosti kod metoda NEWTON–KANTOROVIČa i njemu sličnih metoda i obezbeđuje konvergencija takvih iterativnih procesa.

5.2.4 Gradijentni metod

Neka je dat sistem nelinearnih jednačina (2.3.1), čiji je matrični oblik (videti (2.1.3))

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}. \tag{2.4.1}$$

Gradijentni metod za rešavanje datog sistema jednačina zasniva se na minimizaciji funkcionele

$$U(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(x_1, \ldots, x_n)^2 = \big(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x})\big).$$

Lako je videti da važi ekvivalencija $U(\boldsymbol{x}) = 0 \Leftrightarrow \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$.

Pretpostavimo da jednačina (2.4.1) ima jedinstveno rešenje $\boldsymbol{x} = \boldsymbol{a}$, za koje funkcionela U dostiže minimum. Neka je $\boldsymbol{x}^{(0)}$ početna aproksimacija ovog rešenja. Konstruišemo niz $\{\boldsymbol{x}^{(k)}\}$ takav da je

$$U\left(\boldsymbol{x}^{(0)}\right) > U\left(\boldsymbol{x}^{(1)}\right) > U\left(\boldsymbol{x}^{(2)}\right) > \cdots,$$

koristeći strategiju minimizacije idući po pravcu najvećeg pada, tj. prateći pravac gradijenta funkcije U. Dakle, uzećemo sledeću iteraciju u obliku

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \lambda_k \nabla U\left(\boldsymbol{x}^{(k)}\right), \quad k = 0, 1, \ldots, \tag{2.4.2}$$

gde je

$$\nabla U(\boldsymbol{x}) = \operatorname{grad} U(\boldsymbol{x}) = \left[\frac{\partial U}{\partial x_1} \quad \cdots \quad \frac{\partial U}{\partial x_n}\right]^T.$$

Parametar λ_k određujemo iz uslova da skalarna funkcija S, definisana pomoću $S(t) = U\left(\boldsymbol{x}^{(k)} - t\,\nabla U\left(\boldsymbol{x}^{(k)}\right)\right)$, ima minimum u tački $t = \lambda_k$. S obzirom na to da je jednačina $S'(t) = 0$ nelinearna, izvršićemo njenu linearizaciju u okolini $t = 0$. U tom slučaju imamo

$$L_i^{(k)} = f_i\left(\boldsymbol{x}^{(k)} - t\,\nabla U\left(\boldsymbol{x}^{(k)}\right)\right) = f_i\left(\boldsymbol{x}^{(k)}\right) - t\left(\nabla f_i\left(\boldsymbol{x}^{(k)}\right), \nabla U\left(\boldsymbol{x}^{(k)}\right)\right)$$

pa je linearizovana jednačina

$$\sum_{i=1}^{n} L_i^{(k)} \frac{\mathrm{d}}{\mathrm{d}t} L_i^{(k)} = -\sum_{i=1}^{n} L_i^{(k)} \left(\nabla f_i\left(\boldsymbol{x}^{(k)}\right), \nabla U\left(x^{(k)}\right)\right) = 0,$$

odakle dobijamo

$$\lambda_k = t = \frac{\sum\limits_{i=1}^{n} H_i f_i\left(\boldsymbol{x}^{(k)}\right)}{\sum\limits_{i=1}^{n} H_i^2}, \tag{2.4.3}$$

gde smo stavili $H_i = \left(\nabla f_i\left(\boldsymbol{x}^{(k)}\right),\ \nabla U\left(\boldsymbol{x}^{(k)}\right)\right)$, $i = 1, \ldots, n$.

Kako je

$$\frac{\partial U}{\partial x_j} = \frac{\partial}{\partial x_j}\left\{\sum_{i=1}^{n} f_i(\boldsymbol{x})^2\right\} = 2\sum_{i=1}^{n} f_i(\boldsymbol{x}) \frac{\partial f_i(\boldsymbol{x})}{\partial x_j},$$

imamo $\nabla U(\boldsymbol{x}) = 2W^T(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})$, gde je $W(\boldsymbol{x})$ JACOBIeva matrica.

Shodno prethodnom, (2.4.3) se svodi na

$$\lambda_k = \frac{1}{2} \cdot \frac{\left(\boldsymbol{f}^{(k)}, W_k W_k^T \boldsymbol{f}^{(k)}\right)}{\left(W_k W_k^T \boldsymbol{f}^{(k)}, W_k W_k^T \boldsymbol{f}^{(k)}\right)},$$

gde su $\boldsymbol{f}^{(k)} = \boldsymbol{f}\left(\boldsymbol{x}^{(k)}\right)$ i $W_k = W\left(\boldsymbol{x}^{(k)}\right)$. Tada se gradijentni metod (2.4.2) može predstaviti u obliku

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - 2\lambda_k W_k^T \boldsymbol{f}\left(\boldsymbol{x}^{(k)}\right), \quad k = 0, 1, \ldots .$$

Kao što vidimo, ulogu inverzne matrice $W^{-1}(\boldsymbol{x}^{(k)})$, koja se javlja kod metoda NEWTON–KANTOROVIČa, preuzela je matrica $2\lambda_k W_k^T$.

Primer 2.4.1. Na rešavanje sistema jednačina iz primera 2.3.1 primenićemo gradijentni metod startujući opet sa istim početnim vektorom $\boldsymbol{x}^{(0)} = [2 \quad 1]^T$. Odgovarajući rezultati su dati u tabeli 2.4.1. △

Tabela 2.4.1.

k	$x_1^{(k)}$	$x_2^{(k)}$	$2\lambda_k$
0	2.	1.	3.057873950(−4)
1	1.975537008	0.9259994504	5.387476894(−4)
2	1.983210179	0.9201871306	5.395536228(−4)
3	1.983643559	0.9207840032	5.355965389(−4)
4	1.983705230	0.9207387845	3.393286035(−4)
5	1.983708270	0.9207429317	5.355733542(−4)
6	1.983708709	0.9207426096	3.393325162(−4)
7	1.983708731	0.9207426391	5.359906237(−4)
8	1.983708734	0.9207426368	3.377933010(−4)
9	1.983708734	0.9207426370	

Primetimo da je konvergencija znatno sporija, nego kod metoda NEWTON–KANTOROVIČa, s obzirom na to da je gradijentni metod prvog reda.

Metod gradijenata se uspešno koristi u mnogim optimizacionim problemima nelinearnog programiranja. U literaturi je poslednjih godina razrađen veliki broj metoda, posebno gradijentnog tipa, na osnovu kojih je formiran veći broj paket programa za rešavanje zadataka nelinearnog programiranja (videti, na primer, [83]).

5.2.5 Homotopija i metod nastavljanja

Kao što smo videli u prethodnom izlaganju, metod NEWTON–KANOROVIČa i ostali metodi pokazuju brzu konvergenciju ako su početne vrednosti izabrane dovoljno blizu rešenja. Taj uslov se obezbeđuje jednostavno pomoću tzv. *metoda nastavljanja*[172]. Naime, kod ovih metoda polazimo od problema čije rešenje znamo, a onda postepeno menjamo (deformišemo) problem, težeći pri tome da dođemo do problema koji želimo da rešimo. U ovom odeljku ukratko izlažemo jednu ideju kako realizovati ovakvu strategiju (za istorijske detalje videti [3]).

Neka je zadat problem rešavanja operatorske jednačine (2.1.2), tj. $Fu = \theta$, za koju nemamo dovoljno dobre početne uslove. Ovde je, da ponovimo, F operator koji preslikava BANACHov prostor X u BANACHov prostor Y i θ nula–vektor prostora Y. Međutim, uvek možemo izabrati neki jednostavan problem $Gu = \theta$ ($G\colon X \to Y$) za koji znamo rešenje; na primer, jednačina $Gu = Fu - Fu_0 = \theta$ (ili još jednostavnije, $Gu = u - u_0 = \theta$) ima rešenje $u = u_0$. Tada možemo formulisati novi problem, sa tzv. *funkcijom homotopije* $\Phi\colon X \times [0,1] \to Y$,

$$\Phi(u,t) = tFu + (1-t)Gu = \theta, \tag{2.5.1}$$

gde je t skalar koji pripada intervalu $[0,1]$.[173] Dakle, glavna ideja je da se dati problem umetne u jedno–parametarsku familiju problema, gde $t \in [0,1]$. Jasno je, da je za $t = 0$ rešenje problema (2.5.1) poznato jer se problem svodi na $Gu = \theta$ (u konkretnom primeru, to rešenje je $u = u_0$), dok se za $t = 1$ dobija originalni problem,

$$\Phi(u,1) = Fu = \theta.$$

Strategija rešavanja problema se sastoji u povećanju parametra, na primer sa vrednosti t na vrednost $\hat{t} > t$ i rešavanju problema $\Phi(u,\hat{t}) = \theta$, uz dobre početne uslove koji su obezbeđeni kao rešenje dobijeno u prethodnom koraku za problem $\Phi(u,t) = \theta$, ukoliko je $\Delta t = \hat{t} - t$ dovoljno malo. Dakle, povećavanjem parametra t rešavamo problem (2.5.1), na primer u tačkama $t_0, t_1, \ldots,$

$$0 = t_0 < t_1 < t_2 < \cdots < t_m = 1,$$

tako da posle m koraka dolazimo do rešenja originalnog problema kada parametar postane $t = t_m = 1$.

[172] Na enleskom: *continuation methods*.

[173] U opštem slučaju, *homotopija* između dve funkcije $f, g\colon X \to Y$ je neprekidno preslikavanje $h\colon X \times [0,1] \to Y$ takvo da su $h(x,0) = g(x)$ i $h(x,1) = f(x)$. Ako takvo preslikavanje postoji, tada kažemo da je f *homotopno* na g. Ovo je jedna relacija ekvivalencije među preslikavanjima iz X u Y, gde X i Y mogu biti bilo koja dva topološka prostora.

Koristeći vektorsko označavanje kao i u odeljku 5.2.3, u daljem tekstu razmatraćemo problem rešavanja sistema nelinearnih jednačina (2.3.1), tj.

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}, \tag{2.5.2}$$

primenom homotopije $h\colon \mathbb{R}^n \times [0,1] \to \mathbb{R}^n$, date sa

$$\boldsymbol{h}(\boldsymbol{x},t) = t\boldsymbol{f}(\boldsymbol{x}) + (1-t)\boldsymbol{g}(\boldsymbol{x}), \tag{2.5.3}$$

pri čemu znamo rešenje problema $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0}$ $(\boldsymbol{g}\colon \mathbb{R}^n \to \mathbb{R}^n)$.

U primeni metoda nastavljanja često se koristi jednostavna homotopija, sa $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_0)$, gde je $\boldsymbol{x}_0$ proizvoljna tačka iz $\mathbb{R}^n$. U tom slučaju, (2.5.3) se svodi na

$$\boldsymbol{h}(\boldsymbol{x},t) = \boldsymbol{f}(\boldsymbol{x}) + (t-1)\boldsymbol{f}(\boldsymbol{x}_0), \tag{2.5.4}$$

gde $\boldsymbol{x}_0 \in \mathbb{R}^n$ rešenje problema za $t = t_0 = 0$.

Ako jednačina $\boldsymbol{h}(\boldsymbol{x},t) = \boldsymbol{0}$, gde je $\boldsymbol{h}$ dato sa (2.5.3), ima jedinstveno rešenje za svaku vrednost parametra $t \in [0,1]$, tada je ono funkcija od t pa je $\boldsymbol{x}(t)$ jedinstvena tačka u $\mathbb{R}^n$ za koju je $\boldsymbol{h}(\boldsymbol{x}(t),t) = \boldsymbol{0}$. Skup

$$\big\{\boldsymbol{x}(t) \;\big|\; 0 \le t \le 1\big\} \tag{2.5.5}$$

se tada može interpretirati kao parametrizirana kriva u $\mathbb{R}^n$, sa početkom u poznatoj tački $\boldsymbol{x}(0)$ i krajem u rešenju datog problema (2.5.2), $\boldsymbol{x}(1)$. Dakle, metodom nastavljanja određujemo putanju, tj. krivu (2.5.5), izračunavajući, shodno prethodnom, tačake $\boldsymbol{x}(t_0)$, $\boldsymbol{x}(t_1)$, …, $\boldsymbol{x}(t_m)$. Pomenimo da u nekim slučajevima mogu nastati problemi kod nalaženja krive (2.5.5). Na primer, mogu se pojaviti *povratne tačke*, može nastupiti tzv. *bifurkacija*, može doći do prekida kada za neko $t < 1$ ne postoji rešenje, ili pak da kriva „odlazi“ u beskonačnost. Ukoliko se pojavi neki od ovih slučajeva, preporučuje se promena homotopske funkcije.

Pretpostavimo sada da su funkcije $t \mapsto \boldsymbol{x}(t)$ i $(\boldsymbol{x},t) \mapsto \boldsymbol{h}(\boldsymbol{x},t)$ diferencijabilne. Tada, na osnovu poznate teoreme o implicitnoj funkciji, možemo odrediti $\boldsymbol{x}'(t)$ i opisati krivu (2.5.5) diferencijalnom jednačinom. Neka je operator $T\colon \mathbb{R}^n \to \mathbb{R}^n$ definisan pomoću $T\boldsymbol{x} = \boldsymbol{h}(\boldsymbol{x},t) = [h_1 \;\ldots\; h_n]^T$, gde su

$$h_k = h_k(x_1,\ldots,x_n,t), \quad k = 1,\ldots,n,$$

i t parametar. Na osnovu primera 2.4.2 iz odeljka 2.2.4, zaključujemo da je FRÉCHETov izvod, u ovom slučaju, kvadratna JACOBIeva matrica reda n,

$$T'_{(\boldsymbol{x})} = W(\boldsymbol{h}) = \begin{bmatrix} \dfrac{\partial h_1}{\partial x_1} & \cdots & \dfrac{\partial h_1}{\partial x_n} \\ \vdots & & \\ \dfrac{\partial h_n}{\partial x_1} & & \dfrac{\partial h_n}{\partial x_n} \end{bmatrix}.$$

Kako $\boldsymbol{x}$ zavisi od t, diferenciranjem $\boldsymbol{h}(\boldsymbol{x}(t),t) = \mathbf{0}$ po parametru t, dobijamo

$$W(\boldsymbol{h})\boldsymbol{x}'(t) + \frac{\partial}{\partial t}\boldsymbol{h}(\boldsymbol{x},t) = 0,$$

odakle sleduje

$$\boldsymbol{x}'(t) = -W^{-1}(\boldsymbol{h})\frac{\partial}{\partial t}\boldsymbol{h}(\boldsymbol{x},t).$$

Na primer, u slučaju homotopije (2.5.4), koju ćemo u daljem tekstu tretirati, imamo

$$\boldsymbol{x}'(t) = -W^{-1}(\boldsymbol{f})\boldsymbol{f}(\boldsymbol{x}_0), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0. \tag{2.5.6}$$

Jednačina (2.5.6) predstavlja diferencijalnu jednačinu za krivu (2.5.5), sa poznatim početnim uslovima $\boldsymbol{x}(0)$. Takvi problemi su poznati kao CAUCHYevi problemi za diferencijalne jednačine (videti [73]). Numeričkom integracijom CAUCHYevog problema (2.5.6) na intervalu $0 \le t \le 1$, dolazimo do rešenja $\boldsymbol{x}(1)$, koje u našem slučaju predstavlja rešenje originalnog problema (2.5.2).

Sledeći rezultat daje uslove pod kojima se metod nastavljanja uvek može primeniti na rešavanje sistema nelinearnih jednačina (2.5.2) (videti [92, str. 231]).

Teorema 2.5.1. *Ako je preslikavanje $\boldsymbol{f}\colon \mathbb{R}^n \to \mathbb{R}^n$ neprekidno–diferencijabilno i ako je $\|W^{-1}(\boldsymbol{f})\| \le M$ na $\mathbb{R}^n$, tada za svako $\boldsymbol{x}_0 \in \mathbb{R}^n$ postoji jedinstvena kriva* (2.5.5) *u $\mathbb{R}^n$ takva da je $\boldsymbol{f}(\boldsymbol{x}(t)) + (t-1)\boldsymbol{f}(\boldsymbol{x}_0) = \mathbf{0}$, $0 \le t \le 1$, pri čemu je funkcija $t \mapsto \boldsymbol{x}(t)$ neprekidno–diferencijabilno rešenje* CAUCHY*evog problema* (2.5.6).

Primer 2.5.1. Na rešavanje sistema nelinearnih jednačina[174]

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} \sin x + \mathrm{e}^y - 3 \\ y^2 + 2y - x - 1 \end{bmatrix}, \quad \boldsymbol{x} = (x,y) \in \mathbb{R}^2, \tag{2.5.7}$$

primenićemo homotopiju (2.5.4). Neka je $\boldsymbol{x}_0 = (a,b) \in \mathbb{R}^2$ proizvoljna tačka i $\boldsymbol{f}(\boldsymbol{x}_0) = (A,B) \in \mathbb{R}^2$, gde su $A = \sin a + \mathrm{e}^b - 3$ i $B = b^2 + 2b - a - 1$.

[174] Sličan primer je razmatran u [11, str. 239–240].

Tabela 2.5.1.

t	$\widehat{x}$	$\widehat{y}$
0	0.	2.
1/4	0.4740635917747482	1.78734393196246
	0.4282666498422128	1.77101664199978
1/2	0.8132263964992279	1.52466178200962
	0.7691681845319902	1.50391664106551
3/4	1.0630075387581043	1.21305412137179
	1.0020191136602829	1.18016079331117
1	1.1651928359432943	0.81598851445457
	1.0691468191394507	0.75302901232354
	1.0711605959382172	0.75247279731052
	1.0711621056097559	0.75247313976845
	1.0711621056106288	0.75247313976866
	1.0711621056106288	0.75247313976866

Kako su

$$W(\boldsymbol{h}) = W(\boldsymbol{f}) = \begin{bmatrix} \cos x & \mathrm{e}^y \\ -1 & 2(y+1) \end{bmatrix} \quad \text{i} \quad W^{-1}(\boldsymbol{f}) = \frac{1}{\Delta} \begin{bmatrix} 2(y+1) & -\mathrm{e}^y \\ 1 & \cos x \end{bmatrix},$$

gde je $\Delta = \det W(\boldsymbol{f}) = 2(y+1)\cos x + \mathrm{e}^y$, metod NEWTON-KANTOROVIČa primenjen na (2.5.4) ima oblik

$$\begin{bmatrix} \widehat{x} \\ \widehat{y} \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} - \frac{1}{\Delta} \begin{bmatrix} 2(y+1) & -\mathrm{e}^y \\ 1 & \cos x \end{bmatrix} \left(\begin{bmatrix} \sin x + \mathrm{e}^y - 3 \\ y^2 + 2y - x - 1 \end{bmatrix} + (t-1) \begin{bmatrix} A \\ B \end{bmatrix} \right).$$

Kao početnu tačku uzećemo $(a,b) = (0,2) \in \mathbb{R}^2$ i niz parametara $t_0 = 0$, $t_1 = 1/4$, $t_2 = 1/2$, $t_3 = 3/4$ i $t_4 = 1$. Rešenje za $t_0 = 0$ je poznato i to ćemo uzeti kao početno rešenje za primenu metoda NEWTON–KANTOROVIČa za $t = t_1 = 1/4$. Dobijene vrednosti uzimamo kao početne za sledeći nivo $t = t_2 = 1/2$, itd. Vrednosti iteracija su date u tabeli 2.5.1, pri čemu je dovoljno uzeti samo po dve iteracije na nivoima za $t < 1$. Primetimo da se na završnom nivou $t = 1$ poslednje dve iteracije poklapaju, tako da se te vrednosti mogu uzeti kao rešenja datog sistema jednačina.

Pokazaćemo sada još jedan način rešavanja zasnovan na rešavanju CAUCHY-evog problema (2.5.6), koji se u ovom slučaju svodi na

$$\boldsymbol{x}'(t) = -\frac{1}{\Delta} \begin{bmatrix} 2(y+1) & -\mathrm{e}^y \\ 1 & \cos x \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \frac{-1}{2(y+1)\cos x + \mathrm{e}^y} \begin{bmatrix} 2A(y+1) - B\mathrm{e}^y \\ A + B\cos x \end{bmatrix}.$$

U programskom paketu MATHEMATICA postoji funkcija `NDSolve` za numeričku integraciju sistema diferencijalnih jednačina, sa zadatim početnim uslovima. U našem primeru, odgovarajući programski kôd je

```
In[1]:= a = 0; b = 2;
        ff = {Sin[x] + Exp[y] - 3, y^2 + 2 y - x - 1};
        {A, B} = ff /. {x → a, y → b};
        det = 2 (y[t] + 1) Cos[x[t]] + Exp[y[t]];
        ds = -{2 A (y[t] + 1) - B Exp[y[t]], A + B Cos[x[t]]} / det;
        resenje = NDSolve[{x'[t] == ds[[1]], y'[t] == ds[[2]],
            x[0] == a, y[0] == b}, {x, y}, {t, 0, 1}]
```

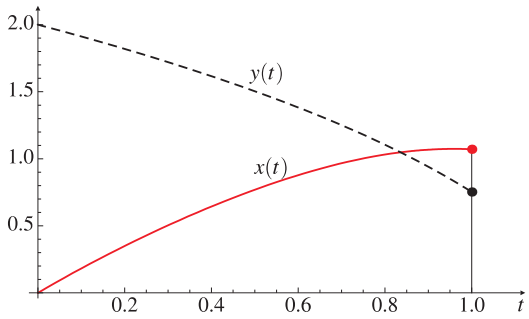
gde smo, kao i ranije, uzeli $(a,b) = (0,2) \in \mathbb{R}^2$. Funkcija `NDSolve` daje rešenje u obliku interpolacionog polinoma,[175] koje je prezentovano kao

```
{{x->InterpolatingFunction[{{0.,1.}},<>],
  y->InterpolatingFunction[{{0.,1.}},<>]}}
```

Dobijeno rešenje omogućava da se naredbom

```
In[7]:= Plot[Evaluate[{x[t], y[t]} /. resenje], {t, 0, 1}]
```

mogu skicirati grafici za $x(t)$ i $y(t)$, kada $0 \le t \le 1$ (vdeti sliku 2.5.1).

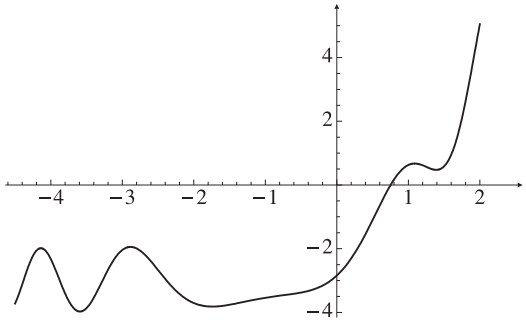


Slika 2.5.1. Krive $x(t)$ i $y(t)$ za $0 \le t \le 1$

Odgovarajuće vrednosti za $t = 1$ se mogu izračunati naredbom:

[175] Problem interpolacije biće razmatran u posebnoj knjizi iz ove serije.

```
In[8]:= Evaluate[{x[1], y[1]} /. resenje[[1]]]

Out[8]= {1.07116, 0.752473}
```



Slika 2.5.2. Grafik funkcije $y \mapsto \mathrm{e}^y + \sin(y^2 + 2y - 1) - 3$

Najzad, napomenimo da se eliminacijom nepoznate x, sistem jednačina (2.5.7) transformiše na jednačinu $\mathrm{e}^y + \sin(y^2 + 2y - 1) - 3 = 0$, koja ima jedinstveno rešenje (videti grafik na slici 2.5.2) i ono se može naći naredbom `FindRoot`

```
In[1]:= FindRoot[Exp[y] + Sin[y^2 + 2 y - 1] - 3 == 0, {y, 1}]

Out[1]= {y → 0.752473}
```

Za startnu vrednost smo uzeli $y_0 = 1$. $\triangle$

Primer 2.5.2. Razmotrićemo sada jedan složeniji sistem od tri nelinearne jednačine

$$\text{(2.5.8)} \qquad \begin{cases} f_1(x,y,z) = \quad 2\mathrm{e}^{-x} - \cos 2y + 2z^2 \quad = 0, \\ f_2(x,y,z) = 400x^3 - 15x^2y^2 - 8z - 31 = 0, \\ f_3(x,y,z) = 10x^2 - 4y + (1+z)y^3 + z^2 = 0. \end{cases}$$

Kao i u prethodnom primeru, primenićemo homotopiju (2.5.4), gde početnu tačku i odgovarajuću vrednost funkcije $\boldsymbol{f}(\boldsymbol{x}) = [f_1 \;\; f_2 \;\; f_3]^T \in \mathbb{R}^3$ označavamo sa $\boldsymbol{x}_0 = (a,b,c) \in \mathbb{R}^3$ i $\boldsymbol{f}(\boldsymbol{x}_0) = (A,B,C) \in \mathbb{R}^3$, respektivno. U ovom slučaju za $W(\boldsymbol{f})$ dobijamo

$$W(\boldsymbol{f}) = \begin{bmatrix} -2\mathrm{e}^{-x} & 2\sin 2y & 4z \\ 30x(40x - y^2) & -30x^2y & -8 \\ 20x & 3y^2(z+1) - 4 & y^3 + 2z \end{bmatrix},$$

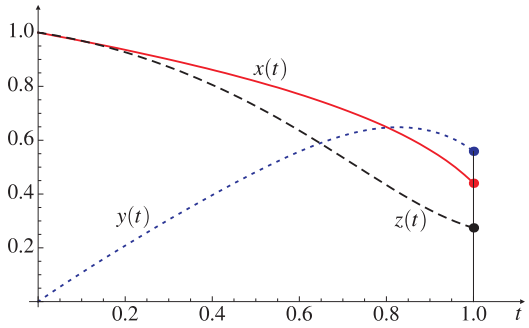
sa determinantom

$$\Delta = 2400x^3yz - 8\left[2e^{-x} - 15x(40x-y^2)z\right]\left[-4+3y^2(1+z)\right]$$
$$-320x\sin 2y + 60x(y^3+2z)\left[xye^{-x} - (40x-y^2)\sin 2y\right].$$

MATHEMATICA kôd koji sleduje obezbeđuje konstrukciju i rešavanje konturnog problema (2.5.6) za $0 \le t \le 1$:

```
In[1]:= f1[x_, y_, z_] := 2 Exp[-x] - Cos[2 y] + 2 z^2 - 1;
        f2[x_, y_, z_] := 400 x^3 - 15 x^2 y^2 - 8 z - 31;
        f3[x_, y_, z_] := 10 x^2 - 4 y + (1 + z) y^3 + z^2;
        ff = {f1[x, y, z], f2[x, y, z], f3[x, y, z]};
        W = Transpose[{D[ff, x], D[ff, y], D[ff, z]}];
        Winv = Inverse[W] // Simplify;
        ds = -Winv.{aA, aB, aC} /. {x → x[t], y → y[t], z → z[t]};
        a = 1; b = 0; c = 1; {aA, aB, aC} = ff /. {x → a, y → b, z → c};
        resenje = NDSolve[{x'[t] == ds[[1]], y'[t] == ds[[2]],
            z'[t] == ds[[3]], x[0] == a, y[0] == b, z[0] == c}, {x, y, z}, {t, 0, 1}]
```

gde je kao startna tačka uzeta $(a,b,c) = (1,0,1) \in \mathbb{R}^3$.

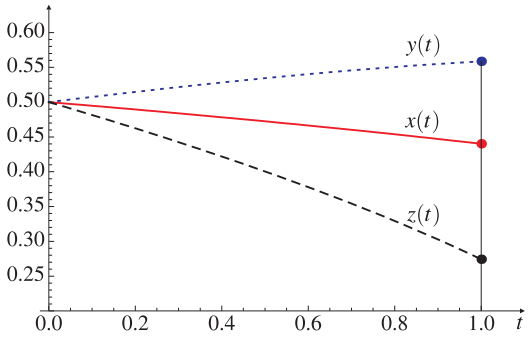


Slika 2.5.3. Krive $x(t)$, $y(t)$ i $z(t)$, $0 \le t \le 1$, za startnu tačku $(a,b,c) = (1,0,1)$

Grafici za krive $x(t)$, $y(t)$ i $z(t)$ su prikazani na slici 2.5.3. Ako, međutim, stavimo $a = b = c = 1/2$, odgovarajući grafici su dati na slici 2.5.4.

Za vrednost $\boldsymbol{x}(1)$, što je inače rešenje datog sistema jednačina (2.5.8), dobijamo:

```
{0.440113, 0.558554, 0.274187}.
```



Slika 2.5.4. Krive $x(t)$, $y(t)$ i $z(t)$, $0 \le t \le 1$, za startnu tačku $(a,b,c) = (1/2, 1/2, 1/2)$

Ako želimo rešenje sa većom preciznošću, na primer sa 60 cifara, potrebno je napraviti nekoliko iteracija pomoću metoda NEWTON–KANTOROVIČa, sa ovim početnim vrednostima, uz korišćenje aritmetike dovoljne preciznosti.

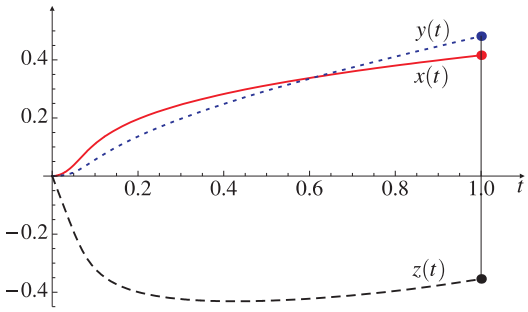
```
In[10]:= x0 = {440 113, 558 554, 274 187} / 10^6;
         Do[x0 = N[({x, y, z} - Winv.ff) /. {x → x0[[1]], y → x0[[2]], z → x0[[3]]}, 60];
          Print[x0], {k, 5}]

{0.440113440680433332221915362054428278489414810061331651385026,
 0.558553896465869236815724913151199590955591319010822754737179,
 0.274187131995499953525765641004647947308145406487311046900080}

{0.440113440679987975823044266121985318430866462851031370832 32,
 0.558553896465257969460425690948914798416103879911623762444 42,
 0.274187131995824525159918964712121391205037688066833719864 31}

{0.440113440679987975823043816450677919414944630619141662447 83,
 0.558553896465257969460425170482472071015263782910033320529 0,
 0.274187131995824525159918682549521004389169993587345020435 4}

{0.440113440679987975823043816450677919414944630618682171964 75,
 0.558553896465257969460425170482472071015263782909541281291 0,
 0.274187131995824525159918682549521004389169993587131519085}

{0.440113440679987975823043816450677919414944630618682171964 47,
 0.558553896465257969460425170482472071015263782909541281291,
 0.274187131995824525159918682549521004389169993587131519085}
```

Najzad, napomenimo i to da ukoliko izaberemo početnu tačku $(a,b,c) = (0,0,0)$, odgovarajući CAUCHYev problem daje drugo rešenje sistema jednačina (2.5.8). U tom slučaju, grafici za krive $x(t)$, $y(t)$ i $z(t)$ su prikazani na slici 2.5.5.

Odgovarajuće rešenje sistema jednačina (2.5.8) je:

```
{0.415862, 0.481786, -0.354293},
```

što je, u stvari, vrednost $\boldsymbol{x}(1)$. $\triangle$



Slika 2.5.5. Krive $x(t)$, $y(t)$ i $z(t)$, $0 \le t \le 1$, za startnu tačku $(a,b,c) = (0,0,0)$

Napomena 2.5.1. Sličan način za opis putanje (2.5.5) je dat u radovima [24, 25, 26].

5.3 ALGEBARSKE JEDNAČINE

5.3.1 Uvodne napomene

Zbog važnosti koje zauzimaju algebarske jednačine u nauci i tehnici, ovo poglavlje posvećujemo numeričkim metodama za rešavanje numeričkih algebarskih jednačina. Poseban odeljak posvećujemo lokalizaciji korena jednačina, imajući pri tome u vidu izbor početnih rešenja u iterativnim procesima. Pored standardnih metoda kakvi su BERNOULLIjev i NEWTON–HORNERov metod, izložićemo i dva metoda trećeg reda, kao i metode za simultano nalaženje korena.

Jedan od najstarijih problema u matematici je rešavanje algebarskih jednačina, koji je i danas aktuelan, s obzirom da ogroman broj tehničkih problema to zahteva.

Opšti oblik algebarske jednačine n-tog stepena je

$$P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0 \quad (a_0 \neq 0), \tag{3.1.1}$$

gde je n prirodan broj, a koeficijenti a_ν, $\nu = 0,1,\ldots,n$, realni ili kompleksni brojevi.

Koreni ove jednačine $x_1,\ldots,x_n$ povezani su sa koeficijentima a_ν, $\nu = 0,1,\ldots,n$, pomoću VIÈTEovih[176] formula

[176] FRANÇOIS VIÈTE (1540 – 1603), francuski matematičar.

$$
\begin{aligned}
\sigma_1 &= \sum_i x_i &&= -\frac{a_1}{a_0},\\
\sigma_2 &= \sum_{i<j} x_i x_j &&= \frac{a_2}{a_0},\\
\sigma_3 &= \sum_{i<j<k} x_i x_j x_k &&= -\frac{a_3}{a_0},\\
&\vdots\\
\sigma_n &= x_1 x_2 \cdots x_n &&= (-1)^n \frac{a_n}{a_0}.
\end{aligned}
\tag{3.1.2}
$$

Za jednačinu (3.1.1) se kaže da je *numerička* ako svi njeni koeficijenti a_ν, $\nu = 0, 1, \ldots, n$, imaju numeričke vrednosti. Ako bar jedan od njih ima oblik opšteg broja, onda se radi o opštoj algebarskoj jednačini. U vezi s tim se i metodi za rešavanje algebarskih jednačina mogu podeliti na metode za rešavanje opštih algebarskih jednačina i na metode za rešavanje numeričkih jednačina.

Metodi prve grupe daju rešenje u obliku formule. Dokazano je da su opšte algebarske jednačine, u opštem slučaju, rešive samo za $n \leq 4$ (za detalje videti, na primer, [69]). Međutim, numeričke algebarske jednačine je uvek moguće rešiti sa određenom tačnošću. Dobijeni koreni su tada u obliku približnog broja.

Ogroman broj metoda za numeričko rešavanje algebarskih jednačina prilagođen je digitalnim računskim mašinama i oni se uglavnom mogu podeliti u dve grupe. U prvu grupu spadaju metodi za izračunavanje korena sa zadatom tačnošu bez korišćenja približnih vrednosti korena. Druga grupa sadrži metode za izračunavanje korena na osnovu, već poznatih približnih vrednosti korena.

Metodi jedne i druge grupe koji se najviše koriste u primenama biće izloženi u ovom poglavlju. Među ovim metodima biće onih koji se koriste za određivanje samo jednog korena (na primer, najvećeg po modulu), kao i onih za simultano određivanje više korena (na primer, dva ili svih n).

Ukoliko je nekim od metoda određeno $k(< n)$ korena $x_1, x_2, \ldots, x_k$, tada se za određivanje ostalih korena rešava jednačina nižeg stepena

$$Q(x) = \frac{P(x)}{(x - x_1) \cdots (x - x_2)} = 0.$$

Ovakav pristup u sukcesivnom određivanju korena jednačina poznat je kao *metod deflacije*.

S obzirom na to da se koreni $x_1, \ldots, x_k$ određuju sa ograničenom tačnošću, jasno je da koeficijenti jednačine $Q(x) = 0$ neće biti apsolutno tačni, što znači da se

tačnost korena $x_{k+1},\dots,x_n$ narušava, bez obzira na metod kojim se oni određuju. Naime, koreni $x_{k+1},\dots,x_n$ neće biti, u opštem sučaju, koreni jednačine $Q(x)=0$.

Proučimo sada uticaj promene koeficijenata algebarske jednačine (3.1.1) na promenu njenog prostog korena $x=x_k$. Ne umanjujući opštost razmatranja, uzmimo da je $a_0=1$. Pod pretpostavkom da koeficijenti a_ν, $\nu=0,1,\dots,n$, postanu redom $a_\nu+\varepsilon_\nu$, gde je $|\varepsilon_\nu|\ll|a_\nu|$, $\nu=1,\dots,n$, oredićemo grešku korena x_k. Naime, u tom slučaju, koren x_k postaje $x_k+\xi_k$.

Kako je

$$(x_k+\xi_k)^n+(a_1+\varepsilon_1)(x_k+\xi_k)^{n-1}+\cdots \\ +(a_{n-1}+\varepsilon_{n-1})(x_k+\xi_k)+(a_n+\varepsilon_n)=0,$$

korišćenjem aproksimacije

$$(x_k+\xi_k)^m\cong x_k^m+mx_k^{m-1}\xi_k\quad(m=2,\dots,n)$$

i zanemarivanjem članova oblika $\xi_k\varepsilon_i$, dobijamo

$$P(x_k)+P'(x_k)\xi_k+(\varepsilon_1x_k^{n-1}+\varepsilon_2x_k^{n-2}+\cdots+\varepsilon_{n-1}x_k+\varepsilon_n)\cong 0,$$

odakle, s obzirom na $P(x_k)=0$ i $P'(x_k)\neq 0$, sleduje

$$\xi_k\cong\frac{1}{P'(x_k)}\left(\varepsilon_1x_k^{n-1}+\varepsilon_2x_k^{n-2}+\cdots+\varepsilon_{n-1}x_k+\varepsilon_n\right).$$

Ako pretpostavimo da su granice apsolutnih grešaka koeficijenata jednake, tj. da je $|\varepsilon_i|\le\varepsilon$, na osnovu poslednje približne jednakosti dobijamo granicu apsolutne greške korena, u oznaci Δ,

$$\Delta\cong\frac{1}{|P'(x_k)|}\left(|x_k|^{n-1}+|x_k|^{n-2}+\cdots+|x_k|+1\right)\varepsilon,$$

tj.

$$\Delta\cong\begin{cases}\dfrac{\left(|x_k|^n-1\right)\varepsilon}{|P'(x_k)|(|x_k|-1)}, & |x_k|\neq 1,\\[2ex] \dfrac{n\varepsilon}{|P'(x_k)|}, & |x_k|=1,\end{cases}$$

U radu [140], WILKINSON navodi primer jednačine

$$P(x)=(x+1)(x+2)\cdots(x+20)=0,$$

sa svim realnim korenima $x_k = -k$, $k = 1, 2, \ldots, 20$, tj.

$$\begin{aligned} P(x) = x^{20} &+ 210x^{19} + 20615x^{18} + 1256850x^{17} + 53327946x^{16} + 1672280820x^{15} \\ &+ 40171771630x^{14} + 756111184500x^{13} + 11310276995381x^{12} \\ &+ 135585182899530x^{11} + 1307535010540395x^{10} \\ &+ 10142299865511450x^9 + 63030812099294896x^8 \\ &+ 311333643161390640x^7 + 1206647803780373360x^6 \\ &+ 3599979517947607200x^5 + 8037811822645051776x^4 \\ &+ 12870931245150988800x^3 + 13803759753640704000x^2 \\ &+ 8752948036761600000x + 2432902008176640000 = 0, \end{aligned}$$

kod koje male promene u koeficijentima izazivaju velike promene u korenima. Naime, ako koeficijent $a_1 = 210$ promenimo za samo $\varepsilon_1 = 2^{-23} \cong 1.2 \times 10^{-7}$, promene u malim korenima su neznatne, ali koren $x_{20} = -20$ postaje -20.8. Štaviše, pojavljuju se i konjugovano–kompleksni koreni i to čak pet parova. Ovo je primer tzv. slabo uslovljenog polinoma.

Pri sukcesivnom određivanju korena jednačine (3.1.1) i njenoj redukciji na niži stepen (metod deflacije), preporučuje se određivanje korena polazeći od najmanjeg po modulu.

Poseban problem kod rešavanja algebarskih jednačina je određivanje višestrukih korena ili korena koji su dovoljno bliski (kvazi–višestrukost). Ovaj problem nećemo posebno razmatrati. Napomenimo samo jedan dobro poznati postupak za eliminaciju višestrukih korena koji se sastoji u sledećem. Za polinom P i njegov izvodni polinom P' odredimo najveći zajednički delilac Q i formiramo jednačinu

$$R(x) = \frac{P(x)}{Q(x)} = 0.$$

Iz teorije polinoma je poznato (videti, na primer, [86] ili [77, IV glava]) da svaka višestruka nula reda k polinoma P je istovremeno i nula reda $k-1$ polinoma P', pa i polinoma Q, odakle zaključujemo da polinom R, dobijen deobom polinoma P polinomom Q, sadrži sve nule polinoma P. Naravno, sve nule polinoma R su proste, što znači da problem rešavanja jednačine $R(x) = 0$ nije više tako složen.

Za određivanje najvećeg zajedničkog delioca dva polinoma koristi se Euklidov algoritam (videti [77, IV glava])). Napomenimo da je najveći zajednički delilac jedinstven sa tačnošću do na multiplikativnu konstantu.

Primer 3.1.1. Neka je dat polinom

$$P(x) = x^5 + 10x^4 + 42x^3 + 92x^2 + 105x + 50.$$

S obzirom na to da je najveći zajednički delilac za P i P' polinom Q, definisan sa $Q(x) = x^2 + 4x + 5$, imamo

$$R(x) = x^3 + 6x^2 + 13x + 10. \quad \triangle$$

5.3.2 Granice korena algebarskih jednačina

Posmatrajmo jednačinu (3.1.1) u kompleksnoj ravni, tj. neka je $x = z$. Stavimo

$$a = \max\{|a_1|, |a_2|, \ldots, |a_n|\} \quad \text{i} \quad a' = \max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}.$$

Teorema 3.2.1. *Svi koreni jednačine* (3.1.1) *leže u prstenu*

$$\frac{|a_n|}{a' + |a_n|} < |z| < 1 + \frac{a}{|a_0|}. \tag{3.2.1}$$

Dokaz. Pretpostavimo da je $|z| > 1$. Kako je

$$|P(z)| \geq \big| |a_0 z^n| - |a_1 z^{n-1} + \cdots + a_{n-1} z + a_n| \big|,$$

i

$$\begin{aligned} |a_1 z^{n-1} + \cdots + a_{n-1} z + a_n| &\leq a(|z|^{n-1} + |z|^{n-2} + \cdots + |z| + 1) \\ &= a\frac{|z|^n - 1}{|z| - 1} < \frac{a|z|^n}{|z| - 1}, \end{aligned}$$

zaključujemo da je $|P(z)| > 0$, ako je

$$|a_0 z^n| - a\frac{|z|^n}{|z| - 1} \geq 0,$$

tj.

$$|z| \geq 1 + \frac{a}{|a_0|}.$$

Kako vrednosti za z koje zadovoljavaju prethodnu nejednakost ne mogu biti koreni jednačine (3.1.1), zaključujemo da svi koreni ove jednačine zadovoljavaju suprotnu nejednakost, tj. leže u unutrašnjosti kruga

$$|z| = 1 + \frac{a}{|a_0|}.$$

Uvođenjem smene $z = 1/\zeta$, jednačina (3.1.1) postoje

$$a_0 + a_1\zeta + \cdots + a_{n-1}\zeta^{n-1} + a_n\zeta^n = 0. \tag{3.2.2}$$

Na osnovu prethodnog, koreni jednačine (3.2.2) zadovoljavaju nejednakost

$$|\zeta| = \frac{1}{|z|} < 1 + \frac{a'}{|a_n|},$$

odakle dobijamo donju granicu

$$|z| > \frac{|a_n|}{a' + |a_n|}. \quad \square$$

Primer 3.2.1. Na osnovu teoreme 3.2.1, ispitajmo u kojoj oblasti leže koreni jednačine

$$P(z) = z^6 - \left(4 + \frac{3\mathrm{i}}{2}\right) z^5 + (3 + 5\mathrm{i})z^4 + \left(\frac{7}{2} - 2\mathrm{i}\right) z^3 - \left(\frac{5}{2} + \frac{9\mathrm{i}}{2}\right) z^2 + 2iz - (2 + 2\mathrm{i}) = 0.$$

Kako su

$$a = \max\{|a_1|, |a_2|, \ldots, |a_6|\} = |a_2| = |3 + 5\mathrm{i}| = \sqrt{34},$$

$$a' = \max\{|a_0|, |a_1|, \ldots, |a_5|\} = |a_2| = \sqrt{34},$$

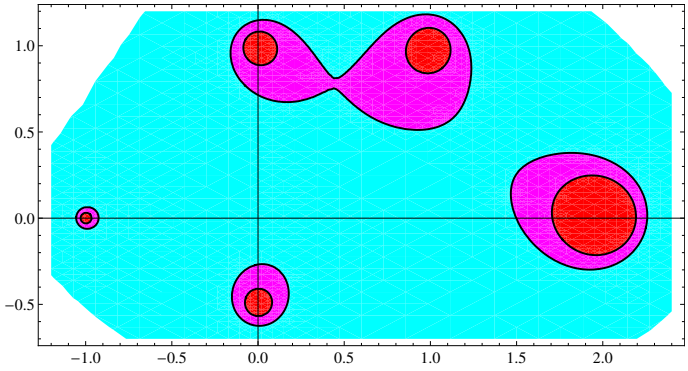
imamo

$$\frac{2}{2 + \sqrt{17}} < |z| < 1 + \sqrt{34},$$

tj. $0.33 < |z| < 6.83$. Očigledno, ovaj rezultat je dosta grub, što se vidi iz sledeće geometrijske interpretacije.

Na slici 3.2.1 su prikazane nivo linije $|P(x + \mathrm{i}y)| = c$ za $c = 1$ i $c = 2$ u pravougaoniku $\{(x, y) \mid -1.2 \le x \le 2.4, -0.7 \le y \le 1.2\}$. Sa slike možemo identifikovati pet tamnih (crvenih) oblasti u kojima je $|P(x + \mathrm{i}y)| < 1$. Jasno je da su nule polinoma lokalizovane u ovim oblastima i one su redom:

$$z_1 = -1, \quad z_2 = -\frac{\mathrm{i}}{2}, \quad z_3 = \mathrm{i}, \quad z_4 = 1 + \mathrm{i}, \quad z_5 = z_6 = 2. \quad \triangle$$



Slika 3.2.1. Nivo linije $|P(x+\mathrm{i}y)| = c$ za $c = 1$ i $c = 2$

Posmatrajmo sada jednačinu

$$P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0, \tag{3.2.3}$$

sa realnim koeficijentima i $a_0 > 0$.

Sledeća teorema daje granicu realnih korena ove jednačine.

Teorema 3.2.2. *Neka je A maksimum apsolutnih vrednosti negativnih koeficijenata jednačine* (3.2.3) *i neka je a_m prvi negativni koeficijent u nizu $a_0, a_1, \ldots, a_n$. Tada su svi pozitivni koreni ove jednačine, ukoliko postoje, manji od*

$$R = 1 + \sqrt[m]{\frac{A}{a_0}}.$$

Posledica 3.2.1. *Ako su svi koeficijenti jednačine* (3.2.3) *nenegativni, jednačina nema pozitivnih korena.*

Napomena 3.2.1. Donja granica pozitivnih korena se može odrediti na sličan način, uz prethodno uvođenje smene $x = 1/y$ u jednačini (3.2.3).

Primer 3.2.2. Odredićemo granice realnih korena jednačine

$$2x^9 + x^7 - x^4 + 19x^3 - 24x^2 + 11 = 0. \tag{3.2.4}$$

Kako su $a_0 = 2$, $A = 24$, $m = 5$, imamo

$$R = 1 + \sqrt[5]{\frac{24}{2}} < 2.65.$$

Za određivanje donje granice pozitivnih korena posmatrajmo jednačinu

$$11y^9 - 24y^7 + 19y^6 - y^5 + y^2 + 2 = 0.$$

Ovde je $\overline{a}_0 = 11$, $\overline{A} = 24$, $\overline{m} = 2$, pa je

$$\overline{R} = 1 + \sqrt{\frac{24}{11}} < 2.48,$$

tj. donja granica pozitivnih korena je $r = 1/\overline{R} > 0.40$. Dakle, ukoliko postoje, pozitivni koreni jednačine (3.2.4) leže u intervalu $(0.40,\ 2.65)$.

Smenom $x = -t$, jednačina (3.2.4) postaje

$$2t^9 + t^7 + t^4 + 19t^3 + 24t^2 - 11 = 0,$$

Primenom sličnog postupka nalazimo da pozitivni koreni ove jednačine mogu ležati samo u intervalu $(0.40,\ 2.21)$.Dakle, ukoliko postoje, negativni koreni jednačine (3.2.4) leže u intervalu $(-2.21,\ -0.40)$.

Interesantno je primetiti da jednačina (3.2.4) ima samo jedan realan koren, čija je približna vrednost -0.560228, dok su ostali koreni konjugovano kompleksni brojevi:

```
{{x -> -1.358 - 0.795441 I}, {x -> -1.358 + 0.795441 I},
{x -> -0.560228},
{x -> -0.157371 - 1.58822 I}, {x -> -0.157371 + 1.58822 I},
{x -> 0.885953 - 0.932499 I}, {x -> 0.885953 + 0.932499 I},
{x -> 0.90953 - 0.336561 I}, {x -> 0.90953 + 0.336561 I}}
```

Ove vrednosti su dobijene u paketu MATHEMATICA. $\triangle$

Za nalaženje gornje granice pozitivnih korena jednačine (3.2.3) može se koristiti i NEWTONov metod koji se sastoji u sledećem.

Ako za $x = a > 0$ važe nejednakosti

$$P^{(k)}(a) \geq 0 \quad (k = 0, 1, \ldots, n),$$

za gornju granicu korena se može uzeti vrednost $R = a$, što sleduje iz TAYLORove formule

$$P(x) = \sum_{k=0}^{n} \frac{1}{k!} P^{(k)}(a)(x-a)^k.$$

Naime, u tom slučaju, za $x > a$ imamo $P(x) > 0$.

Sledećim postupkom se može se dobiti bolja granica. Predstavimo $P(x)$ u obliku

$$P(x) = Q_1(x) - Q_2(x) + Q_3(x) - \cdots + Q_{2m-1}(x) - Q_{2m}(x),$$

gde su $Q_1(x)$ suma prvih članova $P(x)$ sa pozitivnim koeficijentima, počev sa $a_0x^n, -Q_2(x)$ suma narednih članova $P(x)$ sa negativnim koeficijentima, itd., pri čemu se poslednji član $-Q_{2m}(x)$ ili sastoji iz članova sa negativnim koeficijentima ili je jednak nuli. Ako su c_j, $j = 1,\ldots,m$, pozitivni brojevi takvi da je

$$Q_{2j-1}(c_j) - Q_{2j}(c_j) > 0 \quad (j = 1,\ldots,m),$$

za gornju granicu pozitivnih korena može se uzeti broj

$$R = \max(c_1,\ldots,c_m).$$

Navedimo sada jedan rezultat, koji često može biti korisniji od teoreme 3.2.1. Neka je

$$P(z) = z^n + a_1z^{n-1} + a_2z^{n-2} + \cdots + a_{n-1}z + a_n.$$

Uvođenjem smene $z = w - a_1/n$, polinom P se transformiše u polinom Q, kod koga je koeficijent uz w^{n-1} jednak nuli. Dakle,

$$Q(w) = P\Big(w - \frac{a_1}{n}\Big) = w^n + c_2w^{n-2} + \cdots + c_{n-1}w + c_n.$$

Definišimo sada polinom S pomoću

$$S(w) = w^n - |c_2|w^{n-2} - \cdots - |c_{n-1}|w - |c_n|.$$

Teorema 3.2.3. *Ako je bar jedan od koeficijenata $c_k \neq 0$, $k = 2,\ldots,n$, sve nule polinoma P leže u oblasti*

$$\Big|z + \frac{a_1}{n}\Big| \leq r,$$

gde je r pozitivna nula polinoma S.

Dokaz. Neka je $|w| > r$. Pod pretpostavkom da je bar jedan od koeficijenata polinoma $c_k \neq 0$, $k = 2,\ldots,n$, i s obzirom na to da je $S(r) = 0$, imamo

$$\Big|\frac{c_2}{w^2} + \cdots + \frac{c_n}{w^n}\Big| < \frac{|c_2|}{r^2} + \cdots + \frac{|c_n|}{r^n} = 1.$$

Kako je

$$Q(w) = w^n \left(1 + \frac{c_2}{w^2} + \cdots + \frac{c_n}{w^n}\right),$$

na osnovu prethodnog imamo

$$|Q(w)| \geq |w|^n \left(1 - \left|\sum_{\nu=2}^{n} \frac{c_\nu}{w^\nu}\right|\right) > 0,$$

odakle zaključujemo da je $Q(w) \neq 0$ za svako w za koje je $|w| > r$. Drugim rečima, nule polinoma Q nalaze se u krugu $|w| \leq r$, što je s obzirom na smenu $z = w - a_1/n$, ekvivalentno sa tvrđenjem teoreme. □

Primer 3.2.3. Neka je

$$P(z) = z^5 - 10z^4 + 43z^3 - 104z^2 + 150z - 100. \tag{3.2.5}$$

Smenom $z = w - (-10)/5 = w + 2$ dobijamo

$$Q(w) = P(w+2) = w^5 + 3w^3 - 6w^2 + 10w$$

i

$$S(w) = w^5 - 3w^3 - 6w^2 - 10w.$$

Na osnovu teoreme 3.2.2, zaključujemo da je jedinstveni pozitivni koren $w = r$ jednačine $S(w) = 0$ manji od $R = 1 + \sqrt{10} \cong 4.16$. Ova granica se može poboljšati. Naime, nije teško ustanoviti da je $r < 3$. Tada, na osnovu teoreme 3.2.3, zaključujemo da se nule polinoma (3.2.5) nalaze u krugu $|z-2| \leq r < 3$. Napomenimo da su tačne nule polinoma P redom $z_1 = 2$, $z_{2,3} = 1 \pm 2\mathrm{i}$, $z_{4,5} = 3 \pm \mathrm{i}$.

Teorema 3.2.1 daje grube granice $0.4 < |z| < 151$. △

5.3.3 BERRNOULLIjev metod

U ovom odeljku izložićemo jedan vrlo jednostavan metod za određivanje korena algebarskih jednačina bez korišćenja približnih vrednosti ovih korena. Prve ideje o ovom metodu dao je D. BERNOULLI[177] 1728. godine. Sa pojavom digitalnih računskih mašina ovaj metod je znatno usavršen.

Neka je data jednačina

$$a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n = 0 \quad (a_0 \neq 0) \tag{3.3.1}$$

[177] DANIEL BERNOULLI (1700 – 1782), poznati holandsko–švajcarski matematičar.

sa realnim koeficijentima. Ne umanjujući opštost možemo uzeti $a_0 = 1$. Jednačina (3.3.1) se može shvatiti kao karakteristična jednačina linearne diferencne jednačine n-tog reda

$$y_{n+k} + a_1 y_{n+k-1} + \cdots + a_{n-1} y_{k+1} + a_n y_k = 0. \tag{3.3.2}$$

Rešenje ove jednačine određeno je korenima karakteristične jednačine (3.3.1). Naime, ako su x_ν, $\nu = 1, \ldots, n$, koreni jednačine (3.3.1), tada je

$$y_k = C_1 \phi_1(k) + \cdots + C_n \phi_n(k)$$

rešenje diferencne jednačine (3.3.2), gde su funkcije ϕ_i određene korenima x_ν, $\nu = 1, \ldots, n$, respektivno (videti odeljak 1.3.1), a konstante C_ν, $\nu = 1, \ldots, n$, zavise od početnih uslova $y_0, y_1, \ldots, y_{n-1}$.

Ne umanjujući opštost uzećemo da je

$$|x_1| \geq |x_2| \geq \cdots \geq |x_n|.$$

Koren sa najvećim modulom zvaćemo *dominantnim korenom*. Očigledno je da jednačina (3.3.1) može imati:

(a) jedan dominantan koren;

(b) više dominantnih korena.

Ukoliko postoji samo jedan dominantan koren x_1, tada je on realan.

Više dominantnih korena imamo kada je

$$|x_1| = |x_2| = \cdots = |x_r| > |x_{r+1}| \geq \cdots \geq |x_n|,$$

pri čemu treba razlikovati sledeće slučajeve.

(b.1) Dominantni koreni realni i jednaki (višestruki koren reda r), tj.

$$x_1 = \cdots = x_r.$$

(b.2) Dominantni koreni realni i suprotni po znaku (jedan višestruk reda p, a drugi reda $r - p$), tj.

$$x_1 = \cdots = x_p = -x_{p+1} = \cdots = -x_r > 0,$$

(b.3) Dominantni koreni konjugovano kompleksni (u parovima), tj.

$$x_1 = \overline{x}_2,\ x_3 = \overline{x}_4,\ \ldots,\ x_{2m-1} = \overline{x}_{2m} \quad (r = 2m),$$

Primetimo da, u ovom slučaju, među dominantnim korenima može biti i višestrukih.

(b.4) Među dominantnim korenima ima i realnih i konjugovano kompleksnih.

BERNOULLIjev metod omogućuje nalaženje dominantnih korena jednačine (3.3.1), pri čemu se koristi niz $\{y_k\}_{k\in\mathbb{N}_0}$, koji se generiše pomoću (3.3.2) sa početnim vrednostima $y_0, y_1, \ldots, y_{n-1}$.

Razmotrimo najpre slučaj **(1)**, tj. slučaj kada imamo jedan dominantan koren x_1. Tada je

$$y_k = C_1 x_1^k + C_2\phi_2(k) + \cdots + C_n\phi_n(k). \tag{3.3.3}$$

Nije teško pokazati da je

$$\lim_{k\to+\infty} \frac{\phi_i(k)}{x_i^k} = 0 \quad (i = 2, \ldots, n). \tag{3.3.4}$$

Korišćenjem niza $\{y_k\}$ konstruisaćemo niz $\{u_k\}$ pomoću $u_k = y_{k+1}/y_k$. Tada je, s obzirom na (3.3.3),

$$u_k = x_1 \frac{1 + \dfrac{C_2}{C_1}\dfrac{\phi_2(k+1)}{x_1^{k+1}} + \cdots + \dfrac{C_n}{C_1}\dfrac{\phi_n(k+1)}{x_1^{k+1}}}{1 + \dfrac{C_2}{C_1}\dfrac{\phi_2(k)}{x_1^k} + \cdots + \dfrac{C_n}{C_1}\dfrac{\phi_n(k)}{x_1^k}}$$

pri čemu smo pretpostavili da je $C_1 \neq 0$.

Na osnovu (3.3.4) imamo

$$\lim_{k\to+\infty} u_k = x_1, \tag{3.3.5}$$

tj. niz $\{u_k\}$ konvergira ka dominantnom korenu x_1. Za primenu ovog metoda treba obezbediti uslov $C_1 \neq 0$, što zavisi od početnih uslova. Na primer, ovaj uslov je ispunjen ako se uzme

$$y_0 = y_1 = \cdots = y_{n-2} = 0, \quad y_{n-1} = 1.$$

Ispitajmo sada konvergenciju niza $\{u_k\}$ pod uslovom da je koren x_2 realan i takav da je $|x_2| > |x_\nu|$, $\nu = 3, \ldots, n$. Sa e_k označimo grešku $u_k - x_1$. Kako je, pod navedenim uslovom, $\phi_2(k) = x_2^k$, za dovoljno veliko k imamo

$$e_k = \frac{y_{k+1}}{y_k} - x_1 \cong \frac{C_2}{C_1}(x_2 - x_1)\left(\frac{x_2}{x_1}\right)^k,$$

tj. $e_{k+1} \cong (x_2/x_1)e_k$, odakle zaključujemo da je konvergencija linearna.

U slučaju **(b.1)**, umesto rešenja (3.3.3) imamo rešenje

$$y_k = (C_1 + C_2 k + \cdots + C_r k^{r-1})x_1^r + C_{r+1}\phi_{r+1}(k) + \cdots + C_n\phi_n(k).$$

Može se pokazati da i u ovom slučaju važi (3.3.5).

Primer 3.3.1. Neka je data jednačina

$$x^4 - x^3 - 3x^2 - 7x - 6 = 0. \tag{3.3.6}$$

Diferencna jednačina (3.3.2), u ovom slučaju, postaje

$$y_{k+4} - y_{k+3} - 3y_{k+2} - 7y_{k+1} - 6y_k = 0.$$

Startujući sa $y_0 = y_1 = y_2 = 0, y_3 = 1$, na osnovu prethodnog, dobijamo niz $\{y_k\}$, a zatim niz $\{u_k\}$ (tabela 3.3.1).

Tabela 3.3.1.

k	y_k	u_k
3	1	1.
4	1	4.
5	4	3.5
6	14	2.7857143
7	39	2.9487179
8	115	3.0782609
9	354	2.9830508
10	1056	2.9895833
11	3157	3.0069686
12	9493	

Primetimo da je $x_1 = 3$ dominantan koren jednačine (3.3.6). $\triangle$

U slučaju **(b.2)**, imamo

$$\begin{aligned} y_k = &\big[C_1 + C_2 k + \cdots + C_p k^{p-1} + (-1)^k\big(C_{p+1} + \cdots + C_r k^{r-p-1}\big)\big]x_1^k \\ &+ C_{r+1}\phi_{r+1}(k) + \cdots + C_n\phi_n(k), \end{aligned}$$

odakle lako možemo zaključiti da niz $\{u_k\}$, gde je $u_k = y_{k+1}/y_k$, divergira. Mogućno je, međutim, i u ovom slučaju odrediti x_1, ali pomoću niza $\{v_k\}$, gde je $v_k = y_{2k+2}/y_{2k}$. Naime, ovde je

$$\lim_{k\to+\infty} v_k = x_1^2.$$

Primer 3.3.2. Posmatrajmo jednačinu $x^3 + 0.5x^2 - 4x - 2 = 0$, čiji su koreni $x_1 = -x_2 = 2$ i $x_3 = -0.5$. Startujući sa $y_0 = y_1 = 0$, $y_2 = 1$, određujemo nizove $\{y_k\}, \{u_k\}, \{v_k\}$ (tabela 3.3.2).

Tabela 3.3.2.

k	y_k	u_k	$v_{k/2}$
2	1.	−0.5	4.25
3	−0.5	−0.85	
4	4.25	−0.9705882	4.25
5	−4.125	−4.3787879	
6	18.0625	−0.9429066	4.0147059
7	−17.03125	−4.2577982	
8	72.51562	−0.9412842	4.0009158
9	−68.25781	4.2504866	
10	290.12891		

Primetimo da niz $\{u_k\}$ divergira, a da niz $\{v_k\}$ konvergira, što je kriterijum za egzistenciju slučaja **(b.2)**. Niz $\{v_k\}$ konvergira ka $x_1^2 = 4$. $\triangle$

Razmotrimo sada slučaj **(b.3)**. Jednostavnosti radi neka je $r = 2$, tj. neka postoji samo jedan par konjugovano kompleksnih korena, koji su dominantni. Neka su to $x_1 = \rho e^{i\theta}$ i $x_2 = \rho e^{-i\theta}$. Tada je

$$C_1\phi_1(k) + C_2\phi_2(k) = \rho^k\left[A_1\cos k\theta + A_2\sin k\theta\right] = A\rho^k\cos(k\theta + \psi),$$

pri čemu postoji određena veza između konstanti koje se pojavljuju u ovoj jednakosti. Štaviše, za dovoljno veliko k, važi

$$y_k = A\rho^k\cos(k\theta + \psi) + O\left(|x_3|^k\right). \tag{3.3.7}$$

Lako je videti da ranije definisani nizovi $\{u_k\}$ i $\{v_k\}$, u ovom slučaju, divergiraju. Zato definišimo sada nove nizove $\{s_k\}$ i $\{t_k\}$ pomoću

$$s_k = \begin{vmatrix} y_k & y_{k+1} \\ y_{k-1} & y_k \end{vmatrix} = y_k^2 - y_{k-1}y_{k+1},$$

$$t_k = \begin{vmatrix} y_{k+1} & y_{k+2} \\ y_{k-1} & y_k \end{vmatrix} = y_{k+1}y_k - y_{k-1}y_{k+2}.$$

Tada, na osnovu (3.3.7), za dovoljno veliko k, imamo

$$s_k \cong A^2\rho^{2k}\left[\cos^2(k\theta+\psi) - \cos((k-1)\theta+\psi)\cos((k+1)\theta+\psi)\right]$$

i

$$t_k \cong A^2\rho^{2k+1}\left[\cos((k+1)\theta+\psi)\cos(k\theta+\psi) - \cos((k-1)\theta+\psi)\cos((k+2)\theta+\psi)\right],$$

tj.

$$s_k \cong A^2\rho^{2k}\sin^2\theta \quad \text{i} \quad t_k \cong 2A^2\rho^{2k+1}\sin^2\theta\cos\theta,$$

odakle zaključujemo da je

$$\lim_{k\to+\infty}\frac{s_{k+1}}{s_k} = \rho^2 \quad \text{i} \quad \lim_{k\to+\infty}\frac{t_k}{2s_k} = \rho\cos\theta.$$

Dakle, u posmatranom slučaju, niz $\{s_{k+1}/s_k\}$ konvergira ka ρ^2, a niz $\{t_k/(2s_k)\}$ ka realnom delu dominantnog korena.

Primer 3.3.3. Neka je data jednačina $x^3 + x^2 + x - 1 = 0$, kod koje je par konjugovano kompleksnih korena dominantan. Polazeći od $y_0 = y_1 = 0$ i $y_2 = 1$, na osnovu

$$y_{k+3} = -y_{k+2} - y_{k+1} + y_k, \quad k = 0, 1, \ldots,$$

konstruišemo niz $\{y_k\}$:

$$\{y_k\} = \{0,0,1,-1,0,2,-3,1,4,-2,-1,7,-8,0,15,-23,8,30,-61,39,52,\ldots\}.$$

Kako je $s_{18} = (-61)^2 - 39\cdot 30 = 2551$, $s_{19} = 39^2 - (-61)\cdot 52 = 4693$, $t_{18} = 39\cdot(-61) - 52\cdot 30 = -3939$, dobijamo

$$\frac{s_{19}}{s_{18}} \cong 1.8397 \quad \text{i} \quad \frac{t_{18}}{2s_{18}} \cong -0.7720,$$

odakle je

$$x_1 \cong -0,7720 + \text{i}\left(1,8397 - (-0,7720)^2\right)^{1/2} \cong -0,7720 + \text{i}\,1,1152$$

i $x_2 = \overline{x}_1$. $\triangle$

Slučaj tri dominantna korena, od kojih je jedan realan, a dva konjugovano kompleksna, razmatran je u radu [67].

Napomena 3.3.1. Ako se u jednačini (3.3.1) izvrši smena $x = 1/z$, i na rešavanje dobijene jednačine primeni BERNOLLIjev metod mogu se odrediti najmanji po modulu koreni jednačine (3.3.1).

5.3.4 Dva metoda trećeg reda

Pretpostavimo da su sve nule ξ_i, $i = 1,\dots,n$, polinoma

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

realne, različite i uređene po veličini, $\xi_1 < \xi_2 < \cdots < \xi_n$. Tada, na osnovu ROLLEove teoreme, izvodni polinom P' ima $n-1$ nula koje su, takođe, realne i međusobno različite. Označimo ih redom sa $\eta_1,\dots,\eta_{n-1}$, tako da je

$$\xi_1 < \eta_1 < \xi_2 < \cdots < \eta_{n-1} < \xi_n.$$

Sada na skupu $X = \mathbb{R} \setminus \{\xi_1, \eta_1, \xi_2, \dots, \eta_{n-1}, \xi_n\}$ definišimo funkciju $a : X \to \{\xi_1, \xi_2, \dots \xi_n\}$ pomoću

$$a(x) = \begin{cases} \xi_1, & x < \xi_1, \\ \xi_i, & \xi_i < x < \eta_i,\ i = 1,\dots,n-1, \\ \xi_{i+1}, & \eta_i < x < \xi_{i+1},\ i = 1,\dots,n-1, \\ \xi_n, & x > \xi_n. \end{cases}$$

Primetimo da u svakom intervalu između $a(x)$ i x, izraz $P(x)P'(x)$ ne menja znak. Naime, važi

$$\operatorname{sgn}(x - a(x)) = \operatorname{sgn}(P(x)P'(x)). \tag{3.4.1}$$

Kako je $P(x) = \prod_{i=1}^{n}(x - \xi_i)$, imamo

$$\frac{P'(x)}{P(x)} = \sum_{i=1}^{n} \frac{1}{x - \xi_i}, \quad H(x) = \frac{P'(x)^2 - P(x)P''(x)}{P(x)^2} = \sum_{i=1}^{n} \frac{1}{(x-\xi_i)^2}. \tag{3.4.2}$$

S obzirom da je $H(x) > 0$, definišimo funkciju K pomoću

$$K(x) = \frac{\operatorname{sgn}(P(x)P'(x))}{\sqrt{H(x)}}.$$

Nije teško videti da je

$$K(x) = \frac{P(x)/P'(x)}{\sqrt{1 - P(x)P''(x)/P'(x)^2}}. \tag{3.4.3}$$

Uzimajući $x_0 = x$ $(x \in X)$ razmotrimo iterativni proces

$$x_{k+1} = x_k - K(x_k), \quad k = 0, 1, \ldots, \tag{3.4.4}$$

koji je u literaturi poznat kao *metod kvadratnog korena* [93].

Teorema 3.4.1. *Niz* $\{x_k\}_{k\in\mathbb{N}_0}$ *koji se generiše pomoću* (3.4.4) *konvergira monotono ka nuli* $a = a(x)$*, pri čemu je konvergencija kubna, tj. važi*

$$\lim_{k\to+\infty} \frac{x_{k+1} - a}{(x_k - a)^3} = \frac{3P''(a)^2 - 4P'(a)P'''(a)}{24P'(a)^2}. \tag{3.4.5}$$

Dokaz. Neka je početna vrednost $x_0 = x < a(x)$. Tada je, s obzirom na (3.4.1), $\operatorname{sgn}(P(x)P'(x)) = -1$, odakle zaključujemo da je $K(x) < 0$, tj. da je $x_1 = x_0 - K(x_0) > x_0$. S druge strane, na osnovu druge jednakosti u (3.4.2), imamo

$$H(x) > \frac{1}{(x - a(x))^2},$$

odakle sleduje $a(x) - x > H(x)^{-1/2} = -K(x)$, tj. $x_1 < a(x)$. Dakle, važi $x_0 < x_1 < a(x)$ i $a(x_0) = a(x_1)$. Nastavljajući sa ovakvim rezonovanjem dokazujemo da je

$$x_0 < x_1 < x_2 < \cdots < a(x),$$

što znači da niz $\{x_k\}_{k\in\mathbb{N}_0}$ monotono konvergira ka nekoj tački $a \le a(x)$. Kako je, u tom slučaju, $\lim\limits_{k\to+\infty}(x_{k+1} - x_k) = 0$, na osnovu (3.4.4) zaključujemo da je $K(a) = 0$ $(\Rightarrow P(a) = 0)$, tj. da je $a \le a(x)$.

Slučaj $x_0 = x > a(x)$ dokazuje se potpuno simetrično.

Da bismo dokazali jednakost (3.4.5), uvedimo oznaku $h = P(x)/P'(x)$ i pustimo da $x \to a$. Na osnovu (3.4.3) i (3.4.4) imamo

$$x_1 - x = -K(x) = h\left(1 + \frac{P''(x)}{P'(x)}\right)^{-1/2},$$

tj.

$$x_1 - x = h - \frac{1}{2}h^2\frac{P''(x)}{P'(x)} + \frac{3}{8}h^3\left(\frac{P''(x)}{P'(x)}\right)^2 + O(h^4)$$

S druge strane, na osnovu SCHRODERovog razvoja (1.6.3) imamo

$$a - x = h - h^2\frac{P''(x)}{2P'(x)} + h^3\frac{3P''(x)^2 - P'(x)P'''(x)}{6P'(x)^2} + O(h^4).$$

Ako poslednju jednakost oduzmemo od prethodne jednakosti, dobijamo

$$x_1 - x = -h^3\frac{3P''(x)^2 - 4P'(x)P'''(x)}{24P'(x)^2} + O(h^4),$$

odakle, s obzirom da je $h \sim a - x$ $(x \to a)$ sleduje (3.4.5), uz prethodnu zamenu x sa x_k $(k \to +\infty)$. □

Dakle, metod kvadratnog korena ima kubnu konvergenciju. Međutim, može se razmatrati slučaj kada je a višestruka nula reda p; tada je kovergencija prvog reda i važi

$$\lim_{k\to+\infty}\frac{x_{k+1} - a}{x_k - a} = 1 - \frac{1}{\sqrt{p}}.$$

U tesnoj vezi sa metodom kvadratnog korena je LAGUERREov metod

$$x_{k+1} = x_k - \frac{nP(x_k)}{P'(x_k) + \operatorname{sgn}(P'(x_k))\sqrt{G(x_k)}}, \quad k = 0, 1, \ldots, \tag{3.4.6}$$

gde je $x_0 = x$ $(x \in X)$ i

$$G(t) = (n-1)\big[(n-1)P'(t)^2 - nP(t)P''(t)\big].$$

Metod (3.4.6) se može predstaviti i u obliku

$$x_{k+1} = x_k - \frac{nP(x_k)/P'(x_k)}{1 + (n-1)\sqrt{1 - \dfrac{n}{n-1}P(x_k)P''(x_k)/P'(x_k)^2}}, \quad k = 0, 1, \ldots .$$

Red konvergencije LAGUERREovog metoda je tri za proste nule, a jedan za višestruke. Slično dokazu teoreme 3.4.1, može se dokazati odgovarajuća teorema za ovaj metod. Formula (3.4.5) ostaje u važnosti, s tim što konstantu 3 (uz $P''(a)^2$) na desnoj strani ove formule, treba zameniti konstantom $3(n-2)/(n-1)$. LAGUERREov metod je efikasniji od metoda kvadratnog korena, na šta ukazuje sledeći primer.

Primer 3.4.1. Neka je

$$P(x) = (x-1)(x-2)(x-3) = x^3 - 6x^2 + 11x - 6.$$

Tabela 3.4.1.

	pomoću (3.4.4)	pomoću (3.4.6)
x_1	0.85714286	0.98840803
x_2	0.99858769	0.99999991

Startujući sa $x_0 = 0$, dobijamo vrednosti koje su navedene u tabeli 3.4.1. $\triangle$

U radu [31] razmatrana je jedna opštija klasa metoda, koja kao partikularne slučajeve sadrži prethodno navedena dva metoda.

5.3.5 NEWTON-HORNERov metod

NEWTONov metod izložen u odeljku 5.1.2 može se uspešno primeniti za određivanje nula polinoma

$$P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n \quad (a_i \in \mathbb{R}), \tag{3.5.1}$$

pri čemu se vrednosti za $P(x_k)$ i $P'(x_k)$ izračunavaju po HORNERovoj šemi (videti odeljak 1.3.4).

Ako obeležimo $p_j(x) = \dfrac{1}{j!}P^{(j)}(x)$ $(j = 0,1,\ldots,n)$, tada je, na osnovu TAYLORove formule,

$$P(x) = p_0(x_k) + p_1(x_k)(x-x_k) + \cdots + p_n(x_k)(x-x_k)^n.$$

Pretpostavimo da polinom P ima realnu nulu $x = \xi$. NEWTONov metod (1.2.3) za određivanje ove nule postaje

$$x_{k+1} = x_k - \frac{p_0(x_k)}{p_1(x_k)}, \quad k = 0,1,\ldots . \tag{3.5.2}$$

Kao što je poznato, vrednost za $p_0(x_k)$ može se dobiti iz n koraka HORNERovom šemom, kao $p_0(x_k) = b_n$, gde su

$$b_0 = a_0, \quad b_j = b_{j-1}x_k + a_j, \quad j = 1,\dots,n. \tag{3.5.3}$$

U stvari b_j, $j = 0,1,\dots,n-1$, su koeficijenti polinoma dobijenog deljenjem $P(x)$ sa $x - x_k$, tj.

$$P(x) = p_1(x)(x - x_k) + p_0(x_k)$$

i

$$p_1(x) = b_0x^{n-1} + b_1x^{n-2} + \cdots + b_{n-1}.$$

Slično, deljenjem $p_1(x)$ sa $x - x_k$ dobijamo

$$p_2(x) = c_0x^{n-2} + \cdots + c_{n-2},$$

tj. važi

$$p_1(x) = p_2(x)(x - x_k) + p_1(x_k).$$

Ovde je $c_0 = b_0$, $c_j = c_{j-1}x_k + b_j$, $j = 1,\dots,n-1$, i $p_1(x_k) = c_{n-1}$.

Navedena deljenja se šematski mogu prikazati u obliku

	a_0	a_1	$\cdots$	a_j	$\cdots$	a_{n-1}	a_n
x_k		b_0x_k		$b_{j-1}x_k$		$b_{n-2}x_k$	$b_{n-1}x_k$
	b_0	b_1		b_j		b_{n-1}	$b_n = p_0(x_k)$
x_k		c_0x_k		$c_{j-1}x_k$		$c_{n-2}x_k$	
	c_0	c_1		c_j		$c_{n-1} = p_1(x_0)$	

Dakle, korišćenjem HORNERove šeme i formule (3.5.2) može se odrediti realna nula $x = \xi$. Napomenimo da proces (3.5.2) ima kvadratnu konvergenciju ukoliko je nula prosta.

S obzirom da polinom (3.5.1) može imati i konjugovano-kompleksne nule, navedeni metod se može prilagoditi i za ovaj slučaj.

Posmatrajmo polinom (3.5.1) u kome je x zamenjeno sa $z = x + \mathrm{i}y$. Tada, razdvajanjem realnog i imaginarnog dela u $P(z)$, imamo

$$P(z) = u(x,y) + \mathrm{i}v(x,y),$$

gde su u i v harmonijske funkcije. Izračunavanje vrednosti ovih funkcija u tački (x_k, y_k) mogućno je po istom postupku kao što je (3.5.3), s tim što b_j treba zameniti sa $\alpha_j + \mathrm{i}\beta_j$, $j = 0,1,\dots,n$. Tada je

$$\begin{aligned}
&\alpha_0 = a_0, \quad \beta_0 = 0,\\
&\alpha_j = \alpha_{j-1}x_k - \beta_{j-1}y_k + a_j, \quad \beta_j = \alpha_{j-1}y_k + \beta_{j-1}x_k, \quad j = 1,\dots,n,\\
&u_k = u(x_k,y_k) = \alpha_n, \quad v_k = v(x_k,y_k) = \beta_n.
\end{aligned}$$

Neka je $z = \xi + \mathrm{i}\eta$ koren jednačine $P(z) = 0$. S obzirom da je jednačina ekvivalentna sistemu jednačina

$$u(x,y) = 0, \quad v(x,y) = 0, \tag{3.5.4}$$

to se za određivanje korena $z = \xi + \mathrm{i}\eta$ može konstruisati niz $z_k = x_k + \mathrm{i}y_k$, $k = 0,1,\dots$, na primer, pomoću metoda NEWTON-KANTOROVIČa. Imajući u vidu da funkcije u i v zadovoljavaju CAUCHY-RIEMANNove uslove,

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial y},$$

jednostavno se pokazuje da je metod NEWTON-KANTOROVIČa, u ovom slučaju, ekvivalentan sa kompleksnom verzijom NEWTONovog metoda

$$z_{k+1} = z_k - \frac{P(z_k)}{P'(z_k)}, \quad k = 0,1,\dots\ . \tag{3.5.5}$$

Vrednost

$$P'(z_k) = u'_k + \mathrm{i}v'_k = \Big(\frac{\partial u}{\partial x} + \mathrm{i}\frac{\partial v}{\partial x}\Big)\Big|_{x=x_k,y=y_k}$$

može se odrediti rekurzivno pomoću

$$\begin{aligned}
&\gamma_0 = \alpha_0 = a_0, \quad \delta_0 = 0,\\
&\gamma_j = \gamma_{j-1}x_k - \delta_{j-1}y_k + \alpha_j, \quad \delta_j = \gamma_{j-1}y_k + \delta_{j-1}x_k + \beta_j, \quad j = 1,\dots,n-1.
\end{aligned}$$

Naime, $u'_k = \gamma_{n-1}$ i $v'_k = \delta_{n-1}$.

Primer 3.5.1. Za $z = 1+\mathrm{i}$ odredićemo vrednost polinoma $P(z) = z^3 + 2z^2 + z - 1$ i njegovog izvoda $P'(z)$. Na osnovu prethodno datih jednakosti imamo:

$$\begin{aligned}
&\alpha_0 = 1, && \beta_0 = 0,\\
&\alpha_1 = 1\cdot 1 - 0\cdot 1 + 3 = 4, && \beta_1 = 1\cdot 1 + 0\cdot 1 = 1,\\
&\alpha_2 = 4\cdot 1 - 1\cdot 1 + 1 = 4, && \beta_2 = 4\cdot 1 + 1\cdot 1 = 5,\\
&\alpha_3 = 4\cdot 1 - 5\cdot 1 - 1 = -2, && \beta_3 = 4\cdot 1 + 5\cdot 1 = 9
\end{aligned}$$

i

$$\begin{aligned}
&\gamma_0 = 1, \quad \delta_0 = 0,\\
&\gamma_1 = 1\cdot 1 - 0\cdot 1 + 4 = 5, \qquad \delta_1 = 1\cdot 1 + 0\cdot 1 + 1 = 2,\\
&\gamma_2 = 5\cdot 1 - 2\cdot 1 + 4 = 7, \qquad \delta_2 = 5\cdot 1 + 2\cdot 1 + 5 = 12,
\end{aligned}$$

pa je $P(1+\mathrm{i}) = -2+9\mathrm{i}$, a $P'(1+\mathrm{i}) = 7+12\mathrm{i}$. $\triangle$

Korišćenjem prethodno datih rekurentnih relacija, formula (3.5.5) se svodi na

$$\text{(3.5.6)} \qquad \begin{cases} x_{k+1} = x_k - \dfrac{u_k u_k' + v_k v_k'}{(u_k')^2 + (v_k')^2} \\ y_{k+1} = y_k - \dfrac{v_k u_k' - u_k v_k'}{(u_k')^2 + (v_k')^2} \end{cases} \qquad (k = 0, 1, \ldots).$$

Kod primene NEWTON-HORNERovog metoda obično se startuje sa dovoljno malim vrednostima (po modulu) x_0 i y_0 i primenom formula (3.5.6) određuje se najmanja nula po modulu $z_1 = \xi_1 + \mathrm{i}\eta_1$. Ako je $|\eta_1| < \varepsilon$ (ε dovoljno mali pozitivan broj, na primer $\varepsilon = 10^{-7}$) nula z_1 se tretira kao realna, u protivnom nula se uzima kao kompleksna. S obzirom da su koeficijenti polinoma (3.5.1) realni, nule se javljaju u parovima, pa je na ovaj način određena i nula $z_2 = \xi_1 - \mathrm{i}\eta_1$. Ponavljajući izloženi postupak na polinomu $P(z)/(z-\xi_1)$, odnosno $P(z)/((z-\xi_1)^2 + \eta_1^2)$ određuju se ostale nule polinoma P.

Na kraju napomenimo da se formule (3.5.6) vrlo često sreću u programima za numeričko rešavanje algebarskih jednačina. Na primer, program POLRT (iz nekad popularne FORTRAN kolekcije 1130 Scientific Subroutine Package (1130-CM-02X) firme IBM) sačinjen je po formulama (3.5.6).

5.3.6 JENKINS-TRAUBov algoritam

Neka su a_ν, $\nu = 1, \ldots, n$, u opštom slučaju, kompleksni koeficijenti polinoma

$$\text{(3.6.1)} \qquad P(z) = z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n,$$

tj.

$$\text{(3.6.2)} \qquad P(z) = \prod_{\nu=1}^{m} (z - r_\nu)^{m_\nu},$$

gde su $r_1, \ldots, r_m$ različite nule polinoma sa odgovarajućim višestrukostima m_ν, $\nu = 1, \ldots, m$, čiji je zbir $\sum_{\nu=1}^{m} m_\nu = n$.

JENKINS[178]-TRAUBov algoritam je tro-koračni iterativni proces [46], [47] (videti, takođe, knjigu RALSTONa[179] i RABINOWITZa[180] [114, str. 383–392]) sa globalnom konveregencijom za određivanje svih nula polinoma i odgovarajućih višestrukosti, pri čemu se sprovodi metod deflacije pomenut u uvodnom odeljku 5.3.1, polazeći od najmanjih nula po modulu u cilju izbegavanja numeričke nestabilnosti.

U daljem tekstu sa $p(z)$ biće označen sâm polazni polinom $P(z)$ ili polinom dobijen metodom deflacije na nekom koraku.

Na osnovu (3.6.2) imamo

$$p'(z) = \sum_{\nu=1}^{m} m_\nu Q_\nu(z),$$

gde su

$$Q_\nu(z) = \frac{p(z)}{z - r_\nu}, \quad \nu = 1, \ldots, m. \tag{3.6.3}$$

Glavna ideja metoda je generisanje niza polinoma $H^{(k)}(z)$ u obliku

$$H^{(k)}(z) = \sum_{\nu=1}^{m} c_\nu^{(k)} Q_\nu(z), \tag{3.6.4}$$

startujući sa $H^{(0)}(z) = p'(z)$, tj. $c_\nu^{(0)} = m_\nu$, $\nu = 1, \ldots, m$. Ako bismo obezbedili takav izbor ovog niza da $H^{(k)}(z) \to c_1^{(k)} Q_1(z)$, tj. da

$$d_\nu^{(k)} = \frac{c_\nu^{(k)}}{c_1^{(k)}} \to 0, \quad \nu = 2, \ldots, m, \tag{3.6.5}$$

tada bi niz $\{t_k\}$, definisan sa

[178] MICHAEL A. JENKINS, američki matematičar.

[179] ANTHONY RALSTON (1930 –), američki matematičar. Sada je profesor emeritus za kompjuterske nauke i matematiku (State University of New York at Buffalo).

[180] PHILIP RABINOWITZ (1926 – 2006), izraelski matematičar, rođen u SAD, radio u Americi i Izraelu (Weizmann Institute of Science). Posebno je poznat po doprinosima u oblasti numeričke integracije.

$$t_{k+1} = s_k - \frac{p(s_k)}{\widehat{H}^{(k+1)}(s_k)}, \tag{3.6.6}$$

gde je

$$\widehat{H}^{(k)}(z) = \frac{H^{(k)}(z)}{\sum_{\nu=1}^{m} c_\nu^{(k)}} = \frac{\sum_{\nu=1}^{m} c_\nu^{(k)} Q_\nu(z)}{\sum_{\nu=1}^{m} c_\nu^{(k)}}$$

monični polinom i $\{s_k\}$ proizvoljni niz kompleksnih brojeva, bio dobra aproksimacija za nulu r_1. Napomenimo da proizvoljnost niza $\{s_k\}$ treba shvatiti u smislu da je $p(s_k) \neq 0$. U protivnom, s_k je nula polinoma $p(z)$ pa se zato polinom redukuje na niži stepen (deflacija), tj. sprovodi se deljenje $p(z)$ faktorom $z - s_k$.

Teorema 3.6.1. *Pod uslovima* (3.6.5)*, niz* $\{t_{k+1}\}$ *konvergira ka nuli* r_1 *polinoma* (3.6.1).

Dokaz. Kako je za $\nu = 1$ i $z = s_k$, na osnovu (3.6.3), $p(s_k) = (s_k - r_1)Q_1(s_k)$, jednakost (3.6.6) se može izraziti u obliku

$$\begin{aligned} t_{k+1} &= s_k - \frac{p(s_k) \sum_{\nu=1}^{m} c_\nu^{(k+1)}}{\sum_{\nu=1}^{m} c_\nu^{(k+1)} Q_\nu(s_k)} \\ &= s_k - \frac{(s_k - r_1)Q_1(s_k)c_1^{(k+1)} \left(1 + \sum_{\nu=1}^{m} \frac{c_\nu^{(k+1)}}{c_1^{(k+1)}}\right)}{c_1^{(k+1)} Q_1(s_k) \left(1 + \sum_{\nu=1}^{m} \frac{c_\nu^{(k+1)}}{c_1^{(k+1)}} \frac{Q_\nu(s_k)}{Q_1(s_k)}\right)} \\ &= s_k - (s_k - r_1) \frac{1 + \sum_{\nu=1}^{m} d_\nu^{(k+1)}}{1 + \sum_{\nu=1}^{m} d_\nu^{(k+1)} \frac{Q_\nu(s_k)}{Q_1(s_k)}}. \end{aligned}$$

Dakle, ako su ispunjeni uslovi $d_\nu^{(k)} \to 0$, $\nu = 2, \ldots, m$, očigledno $t_{k+1} \to r_1$, kada $k \to +\infty$. □

Postavlja se pitanje kako generisati niz polinoma (3.6.4). U pomenutim radovima [46] i [47], JENKINS i TRAUB predložli su sledeći izbor

$$H^{(k+1)}(z) = \frac{1}{z-s_k}\left[H^{(k)}(z) - \frac{H^{(k)}(s_k)}{p(s_k)}\,p(z)\right],$$

što je uvek definisano kad god je $p(s_k) \neq 0$.

Kako, na osnovu (3.6.4) i (3.6.3), važe jednakosti

$$H^{(k)}(z) = p(z)\sum_{\nu=1}^{m}\frac{c_\nu^{(k)}}{z-r_\nu} \quad \text{i} \quad \frac{H^{(k)}(s_k)}{p(s_k)} = \frac{1}{p(s_k)}\sum_{\nu=1}^{m} c_\nu^{(k)} Q_\nu(s_k) = \sum_{\nu=1}^{m}\frac{c_\nu^{(k)}}{s_k-r_\nu},$$

zaključujemo da je

$$\begin{aligned} H^{(k)}(z) &= \frac{p(z)}{z-s_k}\left\{\sum_{\nu=1}^{m}\frac{c_\nu^{(k)}}{z-r_\nu} - \sum_{\nu=1}^{m}\frac{c_\nu^{(k)}}{s_k-r_\nu}\right\} \\ &= \sum_{\nu=1}^{m}\frac{c_\nu^{(k)}p(z)}{z-s_k}\left(\frac{1}{z-r_\nu}-\frac{1}{s_k-r_\nu}\right) = \sum_{\nu=1}^{m} c_\nu^{(k+1)} Q_\nu(z), \end{aligned}$$

gde je, za svako $\nu = 1,\ldots,m$,

$$c_\nu^{(k+1)} = \frac{c_\nu^{(k)}}{r_\nu - s_k} = \frac{c_\nu^{(k-1)}}{(r_\nu-s_k)(r_\nu-s_{k-1})} = \cdots = \frac{c_\nu^{(0)}}{(r_\nu-s_k)(r_\nu-s_{k-1})\cdots(r_\nu-s_0)},$$

tj.

$$c_\nu^{(k+1)} = \frac{m_\nu}{\prod_{j=0}^{k}(r_\nu - s_j)}, \quad \nu = 1,\ldots,m. \tag{3.6.7}$$

Dakle, iz uslova $p(s_j) \neq 0$ sleduje da je $c_\nu^{(k+1)} \neq 0$ za svako $\nu = 1,\ldots,m$.

Pozabavimo se sada izborom vrednosti za s_k. Kao što je na početku rečeno, u ovom tro-koračnom postupku sprovodi se metod deflacije, polazeći od najmanjih nula po pomdulu. Zato u prvom koraku biramo $s_k = 0$ za $k = 0,1,\ldots,M-1$, tako da koeficijenti velikih nula po modulu imaju zanemarljivo male vrednosti, a da se oni koji odgovaraju malim nulama istaknu, što izbor $s_k = 0$ obezbeđuje. Zaista, u tom slučaju, na osnovu (3.6.7), imamo

$$d_\nu^{(k)} = \frac{m_\nu}{m_1}\left(\frac{r_1}{r_\nu}\right)^k, \quad \nu = 2,\ldots,m.$$

U tom slučaju, ako je postojala nula r_1 takva da je

$$|r_1| < |r_2| \leq |r_3| \leq \cdots \leq |r_m|,$$

nastavak procesa (3.6.6) sa $s_k = 0$ bi generisao niz $\{t_k\}$ koji konvergira ka r_1 brzinom geometrijske progresije sa količnikom $|r_1/r_2|$. Ovaj uslov, međutim, ne mora uvek važiti, a može se desiti i slučaj da uslov važi, ali da je pomenuti količnik blizak jedinici. Ovo znači da tada imamo tzv. *klaster nula*[181].

Dakle, u ovom prvom koraku, iz pomenutog razloga bira se $s_k = 0$ za fiksiran broj iteracija M i naknadno analizira eventualno postojanje klastera nula. U programskoj implementaciji (kod nalaženja nula polinoma stepena $n \leq 50$) se najčešće uzima $M = 5$ (videti, na primer, [114, str. 385]).

Drugi korak se bavi razdvanjem nula koje su (skoro) jednake po apsolutnoj vrednosti, pomoću iteracije sa $s_k = s$, gde je kompleksan broj s izabran tako da je bliži jednoj nuli, u oznaci r_1, u odnosu na sve ostale nule, i sa takvim izborom za s_k imamo

$$d_\nu^{(k+1)} = \frac{m_\nu}{m_1}\left(\frac{r_1}{r_\nu}\right)^M \left(\frac{r_1 - s}{r_\nu - s}\right)^{k-M+1}, \quad \nu = 2,\ldots,m, \tag{3.6.8}$$

gde $k = M, M+1, \ldots, L-1$. Dakle, ako je $|r_1 - s| < |r_\nu - s|$, $\nu = 2,\ldots,m$, zaključujemo da $d_\nu^{(L)} \to 0$, kada $L \to +\infty$, tj. da niz $\{t_k\}$ konvergira ka r_1.

Kao početna ocena za s se bira kompleksan broj $s = re^{i\theta}$, gde je r jedinstvena pozitivana nula jednačine

$$z^n + |a_1|z^{n-1} + \cdots + |a_{n-1}|z - |a_n| = 0,$$

koja predstavlja donju granicu modula svih nula datog polinoma (3.6.1), saglasno klasičnom CAUCHYevom rezultatu (videti [84, str. 244]), a ugao θ se teorijski može uzeti kao slučajni broj[182]. Kao i u prethodnom koraku, niz $\{t_k\}$ ima linearnu konvergenciju ka r_1, sa količnikom $|(r_1 - s|)/(r_2 - s)|$, gde je r_2 nula najbliža tački s, izuzimajući r_1. Broj iteracija u ovom drugom koraku je $L - M$, i to je teorijski zavisno od distribucije nula polinoma, ali se praktično L određuje na osnovu zadovoljenja uslova

$$|t_k - t_{k-1}| \leq \frac{1}{2}|t_{k-1}| \qquad \text{i} \qquad |t_{k-1} - t_{k-2}| \leq \frac{1}{2}|t_{k-2}|.$$

[181] Na engleskom: *a cluster of zeros*.

[182] U izboru θ često se primenjuju neki heuristički pristupi kako je to, na primer, navedeno u knjizi [114, str. 386].

Međutim, ako se prethodni uslovi ne ostvare posle nekog razumnog broja iteracija (npr. desetak iteracija), postupak se ponavlja sa nekim drugim (slučajnim) uglom θ, startujući opet sa $k = M$.

Pretpostavljajući da t_k daje dobru aproksimaciju za r_1, u trećem koraku se koristi promenljiva vrednost $s_k = t_k$, $k = L, L+1, \ldots$, koja se generiše, saglasno sa (3.6.6), pomoću

$$s_L = t_L = s - \frac{p(s)}{\widehat{H}^{(L)}(s)}, \quad s_{k+1} = s_k - \frac{p(s_k)}{\widehat{H}^{(k+1)}(s_k)}, \quad k = L, L+1, \ldots ,$$

i na taj način se obezbeđuje kvadratna konvergencija ka nuli r_1.

Za prethodno opisani JENKINS-TRAUBov trokoračni algoritam može se dokazati teorema o globalnoj konvergenciji [47] (videti, takođe, [114, str. 387–390]). Pre formulisanja ove teoreme uvedimo sledeća označavanja

$$R = \min_{2 \le \nu \le m} |r_1 - r_\nu|, \qquad D_L = \sum_{\nu=2}^{m} \frac{|c_\nu^{(L)}|}{|c_1^{(L)}|} = \sum_{\nu=2}^{m} |d_\nu^{(L)}| \quad (c_1^{(L)} \neq 0),$$

$$\tau_L = \frac{2D_L}{1 - D_L}, \qquad C_j = \frac{|s_{L+j+1} - r_1|}{|s_{L+j} - r_1|^2}.$$

Teorema 3.6.2. *Pod uslovima*

$$|s_L - r_1| < \frac{R}{2} \quad i \quad D_L < \frac{1}{3}, \tag{3.6.9}$$

JENKINS-TRAUB*ov algoritam konvergira, tj.* $s_k \to r_1$, *kada* $k \to +\infty$.

Štaviše, konvergencija je kvadratna, sa faktorom konvergencije C_j *koji teži nuli kada* $k = L + j \to +\infty$, *tj.*

$$C_j \le \frac{2}{R} \tau_L^{j(j-1)/2}.$$

Napomenimo da u trećem koraku algoritma $s_k = t_k$, kao i da uslov $D_L < 1/3$ obezbeđuje da je $\tau_L < 1$. Inače, uslovi (3.6.9) se uvek mogu obezbediti ako se s izabere takvo da je $|s - r_1| < |s - r_\nu|$ za svako $\nu = 2, \ldots, m$. Zaista, kako je na osnovu (3.6.8) za $k = L - 1$,

$$D_L = \sum_{\nu=2}^{m} |d_\nu^{(L)}| = \sum_{\nu=2}^{m} \frac{m_\nu}{m_1} \left| \frac{r_1}{r_\nu} \right|^M \left| \frac{r_1 - s}{r_\nu - s} \right|^{L-M},$$

tada za fiksirano M i neko dovoljno veliko L može se obezbediti da je $D_L < 1/3$ i $2D_L/(1 - D_L)|s - r_1| < R/2$. Poslednja nejednakost obezbeđuje prvi uslov u (3.6.9), tj. $|s_L - r_1| < R/2$.

JENKINS-TRAUBov algoritam je primenjen u mnogim softverskim paketima za nalaženje nula polinoma.

5.3.7 Numerička faktorizacija polinoma

Faktorizacija algebarskih polinoma je veoma važan problem koji se pojavljuje ne samo u matematičkim oblastima, već i mnogim primenama u drugim oblastima nauke i tehnike. Faktorizacija je posebno bitna kod simboličkih izračunavanja ili u tzv. *kompjuterskoj algebri* kod algoritama i odgovarajućih softvera za rad sa matematičkim izrazima i jednačinama u simboličkom obliku. U ovom odeljku bavićemo se samo klasičnim problemom numeričke faktorizacije polinoma nad poljem $\mathbb{R}$ ili $\mathbb{C}$. U literaturi postoji više razvijenih metoda za rešavanje ovakvih problema faktorizacije, počev od BAIRSTOWljevog[183] metoda [4] i metoda LINa [60, 61]. Mnogi od tih metoda imaju kvadratnu konvergenciju, ali obično zahtevaju poznavanje dovoljno bliske startne vrednosti za faktorizaciju. U preglednom radu [37], HOUSEHOLDER i STEWART[184] pominju, takođe, DANDELIN[185]–GRAEFFEov[186] metod[187], kao i *qd* algoritam, mada oni nisu primarno za tu namenu. Jedan broj ovih metoda se može povezati i sa algoritmom koji je 1941. godine predložio SEBASTIÃO E SILVA[188] u [119], a tridesetak godina kasnije generalisao HOUSEHOLDER [34] (videti takođe članke [35], [123], [12]).

Pomenimo ovde još i SAMELSONov[189] metod [116] iz 1959. godine, koji predstavlja generalizaciju iterativnog postupka koji su nešto ranije dobili BAUER i SAMELSON [5]. Sâm SAMELSON [116] pominje vezu njegovog metoda sa BAIRSTOWljevim metodom. Naime, uzimajući monični algebarski polinom na polju kompleksnih brojeva, sa nulama $z_1, z_2, \ldots, z_n$, tj.

$$P(z) = z^n + p_1 z^{n-1} + \cdots + p_{n-1} z + p_n = \prod_{k=1}^{n} (z - z_k), \tag{3.7.1}$$

[183] LEONARD BAIRSTOW (1880 – 1963), britanski aerodinamičar.

[184] G. W. (PETE) STEWART (1940 –), poznati američki naučnik u oblasti numeričke linearne algebre i kompjuterskih nauka i urednik u mnogim naučnim časopisima. Sada je profesor emeritus na Maryland univerzitetu (SAD).

[185] GERMINAL PIERRE DANDELIN (1794 – 1847), francuski matematičar, vojnik i profesor inženjerstva.

[186] KARL HEINRICH GRÄFFE (1799 – 1873), nemački matematičar.

[187] Metod su nezavisno otkrili i razvili DANDELIN 1826. i GRÄFFE 1837. godine. Glavne ideje ovog metoda je otkrio 1834. godine i poznati ruski matematičar u oblasti geometrije NIKOLAI IVANOVICH LOBACHEVSKY (1792 – 1856), sa čijim se imenom često povezuje ovaj metod (videti [33]).

[188] JOSÉ SEBASTIÃO E SILVA (1914 – 1972), portugalski matematičar.

[189] KLAUS SAMELSON (1918 – 1980), nemački matematičar i fizičar.

on traži faktorizaciju pomoću dva faktora

$$u(z) = (z-z_1)(z-z_2)\cdots(z-z_m) \quad \text{i} \quad v(z) = (z-z_{m+1})(z-z_{m+2})\cdots(z-z_n).$$

Neka su p i q monični polinomi stepena m i $n-m$ koji aproksimiraju u i v, respektivno. Tada njegova kvadratno konvergentna iterativna procedura određuje poboljšane aproksimacije p^* i q^* pomoću formule

$$p^*q + q^*p = P + pq. \tag{3.7.2}$$

Ako su p i q relativno prosti, tada su polinomi p^* i q^* jedinstveno određeni pomoću (3.7.2). SAMELSONova iteracija je nezavisno bila otkrivena 1969. godine i od strane STEWARTa [121], koji je okarakterisao pq^* kao linearnu kombinaciju od P, q, zq, …, $z^{m-1}q$ koja je deljiva sa p. O nekim metodima faktorizacije videti [36], [37], [122], [117].

Mi ćemo se ovde zadržati najpre na PREŠIĆevom[190] metodu [110] iz 1966. godine (videti, takođe, i opširniju verziju rada [111]), a zatim ćemo se osvrnuti na varijantu faktorizacionog metoda, poznatog u literaturi kao GRAUov[191] metod, i na kraju ćemo analizirati BAIRSTOWljevev metod. GRAU [29] je u svom pristupu koristio NEWTONov tip aproksimacije za simultano numeričko određivanje kompletnog skupa faktora datog polinoma. Inače, PREŠIĆev pristup faktorizaciji je mnogo elegantniji od GRAUovog, a uz to se pojavio i pet godina ranije[192].

1. PREŠIĆev metod faktorizacije. Inspirisan samo rezultatima svog profesora MARKOVIĆa,[193] Prešić [110, 111] je razvio iterativni metod za numeričku faktorizaciju algebarskih polinoma sa m $(2 \le m \le n)$ faktora.

Neka je P monični algebarski polinom nad poljem kompleksnih brojeva dat sa (3.7.1) i neka je predstavljen u faktorizovanom obliku

[190] SLAVIŠA B. PREŠIĆ (1933 – 2008), srpski matematičar sa doprinosima u više matematičkih oblasti.

[191] Za ovog autora ne posedujemo podatke, sem da je A.A. GRAU.

[192] Nažalost, prva skraćena verzija PREŠIĆevog rada [110] se pojavila na samo dve stranice na francuskom jeziku u opštem časopisu Francuske akademije nauka (C. R. Acad. Sci. Paris), dok je kompletna verzija rada štampana na srpskom jeziku dve godine kasnije (1968) u časopisu *Mat. Vesnik*, tako da je rad ostao nezapažen. Takođe, postojao je još jedan hendikep u prezentaciji PREŠIĆevih radova, što je prenaglašen deo o $1-1-\cdots-1$ faktorizaciji, a što će se kasnije pokazati da je to već poznat WEIERSTRASSov rezultat iz 1903 (videti odeljak 5.3.8). Inače, PREŠIĆevi radovi sadrže dokaz opšte faktorizacije. S druge strane, GRAU je svoj rad publikovao 1971. godine u prestižnom specijalizovanom časopisu *SIAM J. Numer. Anal.* [29]. U međuvremenu (1969) pojavio se i rad J. DVORČUKa [18] o faktorizaciji polinoma na kvadratne faktore primenom NEWTONovog metoda.

[193] DRAGOLJUB MARKOVIĆ (1903 – 1965), srpski matematičar i utemeljivač moderne algebre u Srbiji.

$$P(z) = A_1(z)A_2(z)\cdots A_m(z) \qquad (2 \le m \le n), \tag{3.7.3}$$

gde su $A_\nu(z)$ monični polinomi stepena n_ν, tj.

$$A_\nu(z) = \sum_{i=0}^{n_\nu} a_{\nu i} z^{n_\nu - i}, \qquad a_{\nu 0} = 1 \quad (\nu = 1,2,\ldots,m), \tag{3.7.4}$$

gde je $\sum_{\nu=1}^{m} n_\nu = n$. Slučaj $m = 2$ je pomenut u uvodnom delu ovog odeljka.

Pretpostavljajući da su (kompleksne) nule polinoma (3.7.1) proste, PREŠIĆ je formulisao tzv. $n_1 - n_2 - \cdots - n_m$ faktorizaciju, u kojoj su sukcesivno iterirani monični faktori

$$A_\nu^{(k)}(z) = \sum_{i=0}^{n_\nu} a_{\nu i}^{(k)} z^{n_\nu - i}, \quad a_{\nu 0}^{(k)} = 1 \qquad (\nu = 1,2,\ldots,m), \tag{3.7.5}$$

određeni pomoću

$$A_1^{(k+1)}A_2^{(k)}\cdots A_m^{(k)} + A_1^{(k)}A_2^{(k+1)}\cdots A_m^{(k)} + \cdots + A_1^{(k)}A_2^{(k)}\cdots A_m^{(k+1)}$$
$$-(m-1)A_1^{(k)}A_2^{(k)}\cdots A_m^{(k)} \equiv P,$$

tj.

$$A_1^{(k)}(z)A_2^{(k)}(z)\cdots A_m^{(k)}(z)\left(\sum_{\nu=1}^{m}\frac{A_\nu^{(k+1)}(z)}{A_\nu^{(k)}(z)} - m + 1\right) \equiv P(z). \tag{3.7.6}$$

Uzimajući koeficijente $a_{\nu i}$ polinoma (3.7.4) kao koordinate n-dimenzionalnog vektora

$$\boldsymbol{a} = [a_{11}\ a_{12}\ \cdots\ a_{1n_1}\ a_{21}\ a_{22}\ \cdots\ a_{2n_2}\ \cdots\ a_{m1}\ a_{m2}\ \cdots\ a_{mn_m}]^T \tag{3.7.7}$$

i $a_{\nu i}^{(k)}$ (koeficijenti iteriranih faktora (3.7.5)) kao koordinate odgovarajućeg n-dimenzionalnog vektora $\boldsymbol{a}^{(k)}$, PREŠIĆ je uočio da (3.7.6) implicira sledeći sistem linearnih jednačina

$$A_n(\boldsymbol{a}^{(k)})\boldsymbol{a}^{(k+1)} = \boldsymbol{b}_n(\boldsymbol{a}^{(k)}, \boldsymbol{p}), \tag{3.7.8}$$

sa matricom A_n tipa $n \times n$, koja zavisi samo od $\boldsymbol{a}^{(k)}$, a slobodni član je n-dimenzionalni vektor $\boldsymbol{b}_n$, koji zavisi, takođe, od $\boldsymbol{a}^{(k)}$ i koeficijenata polinoma (3.7.1), tj. od $\boldsymbol{p} = [p_1\ p_2\ \cdots\ p_n]^T$.

PREŠIĆ dalje zaključuje da postoji okolina V vektora $\boldsymbol{a} \in \mathbb{C}^n$, takva da se (3.7.8) može izraziti u obliku

$$\boldsymbol{a}^{(k+1)} = F(\boldsymbol{a}^{(k)}) \qquad (k = 0,1,\ldots;\ \boldsymbol{a}^{(k)} \in V), \tag{3.7.9}$$

gde je $F : V \to V$ dovoljan broj puta diferencijabilni operator (u FRÉCHETovom smislu). Praktično, on je dokazao da je $F(\boldsymbol{a}) = \boldsymbol{a}$ i da je $F'_{(\boldsymbol{a})}$ nula operator, tako da važi

$$\|\boldsymbol{a}^{(k+1)} - \boldsymbol{a}\| = O\big(\|\boldsymbol{a}^{(k)} - \boldsymbol{a}\|^2\big) \qquad \Big(\boldsymbol{a} = \lim_{k\to+\infty} \boldsymbol{a}^{(k)}\Big).$$

Dakle, PREŠIĆev rezultat možemo sumirati u sledećoj teoremi, kao što je dato u našem preglednom radu [74]:

Teorema 3.7.1. *Postoji okolina V vektora $\boldsymbol{a} \in \mathbb{C}^n$ tako da za proizvoljno $\boldsymbol{a}^{(0)} \in V$, iterativni proces* (3.7.9) *ima kvadratnu konvergenciju ka $\boldsymbol{a}$.*

Prema tome,

$$\lim_{k\to+\infty} A_\nu^{(k)}(z) = A_\nu(z) \qquad (\nu = 1,2,\ldots,m)$$

daje faktorizaciju (3.7.3).

U specijalnom slučaju, PREŠIĆ [111] je izveo formule za $2-2-2$ faktorizaciju polinoma šestog stepena. Ovde kao ilustraciju navodimo prostiji slučaj faktorizacije $2-2$ za polinom četvrtog stepena $P(z) = z^4 + p_1 z^3 + p_2 z^2 + p_3 z + p_4$, sa

$$A_1(z) = z^2 + a_{11}z + a_{12}, \quad A_2(z) = z^2 + a_{21}z + a_{22}.$$

U tom slučaju, (3.7.8) postaje

$$\begin{aligned}
a_{11}^{(k+1)} \qquad\qquad &+ \quad a_{21}^{(k+1)} \qquad\qquad &= b_1^{(k)},\\
a_{21}^{(k)}a_{11}^{(k+1)} + \quad a_{12}^{(k+1)} &+ a_{11}^{(k)}a_{21}^{(k+1)} + \quad a_{22}^{(k+1)} &= b_2^{(k)},\\
a_{22}^{(k)}a_{11}^{(k+1)} + a_{21}^{(k)}a_{12}^{(k+1)} &+ a_{12}^{(k)}a_{21}^{(k+1)} + a_{11}^{(k)}a_{22}^{(k+1)} &= b_3^{(k)},\\
a_{22}^{(k)}a_{12}^{(k+1)} \qquad\qquad &+ a_{12}^{(k)}a_{22}^{(k+1)} &= b_4^{(k)},
\end{aligned}$$

gde su

$$\begin{aligned}
b_1^{(k)} &= p_1, & b_2^{(k)} &= p_2 + a_{11}^{(k)}a_{21}^{(k)},\\
b_3^{(k)} &= p_3 + a_{11}^{(k)}a_{22}^{(k)} + a_{12}^{(k)}a_{21}^{(k)}, & b_4^{(k)} &= p_4 + a_{12}^{(k)}a_{22}^{(k)}.
\end{aligned}$$

Rešavanjem ovog sistema jednačina dobijamo iterativnu proceduru oblika (3.7.9).

Napomena 3.7.1. Ovaj slučaj $(m=2)$ daje SAMELSONovu iteraciju.

Primer 3.7.1. Neka je dat polinom $P(z)=z^4+4z^3+z^2+2z-8$, tj. $p_1=4$, $p_2=1$, $p_3=2$ i $p_4=-8$. U cilju nalaženja $2-2$ faktorizacije, kao startne polinome uzećemo

$$A_1^{(0)}(z)=z^2+\frac{3}{2}z+3 \quad \text{i} \quad A_2^{(0)}(z)=z^2+\frac{5}{2}z-3,$$

tj. početni vector $\boldsymbol{a}^{(0)}=[3/2\ \ 3\ \ 5/2\ \ -3]^T$. Na osnovu prethodno datog sistema jednačina[194] za slučaj $2-2$ faktorizacije, primenom jednostavne procedure realizovane u paketu MATHEMATICA,

```
Presic22[p_,a_] := Module[{b,aa}, b=p+{0, a[[1]]a[[3]],
     a[[1]]a[[4]]+a[[2]]a[[3]], a[[2]]a[[4]]};
         aa = {{1,0,1,0},{a[[3]],1,a[[1]],1},
           {a[[4]],a[[3]],a[[2]],a[[1]]},{0,a[[4]],0,a[[2]]}};
         Return[LinearSolve[aa,b]]];
```

dobijamo sledeći niz vektora

$$\begin{aligned}
\boldsymbol{a}^{(1)} &= \begin{bmatrix}0.91666666667 & 1.75000000000 & 3.08333333333 & -3.91666666667\end{bmatrix}^T,\\
\boldsymbol{a}^{(2)} &= \begin{bmatrix}1.00271694416 & 1.99971390746 & 2.99728305584 & -4.01254506425\end{bmatrix}^T,\\
\boldsymbol{a}^{(3)} &= \begin{bmatrix}1.00000212280 & 1.99999553954 & 2.99999787720 & -4.00000715540\end{bmatrix}^T,\\
\boldsymbol{a}^{(4)} &= \begin{bmatrix}0.99999999999 & 1.99999999999 & 3.00000000000 & -4.00000000000\end{bmatrix}^T.
\end{aligned}$$

Jednostavno je proveriti da su kvadratni faktori datog polinoma, zaista,

$$A_1(z)=z^2+z+2 \quad \text{i} \quad A_2(z)=z^2+3z-4. \quad \triangle$$

Korišćenjem prethodnih ideja o faktorizaciji polinoma, PETRIĆ[195] i PREŠIĆ [105] su razmatrali problem simultanog određivanja svih rešenja sistema algebarskih jednačina

$$\begin{aligned}
J_1(x,y) &\equiv A_1x^2+2B_1xy+C_1y^2+2D_1x+2E_1y+F_1=0,\\
J_2(x,y) &\equiv A_2x^2+2B_2xy+C_2y^2+2D_2x+2E_2y+F_2=0.
\end{aligned}$$

[194] Koordinate vektora (3.7.7) označili smo standardno sa a_k, $k=1,2,\ldots,n$, gde je $k=(\nu-1)n_\nu+i$, $i=1,\ldots,n_\nu$; $\nu=1,\ldots,s$.

[195] JOVAN J. PETRIĆ (1930 – 1997), srpski matematičar, posebno poznat u oblasti operacionih istraživanja i matematičkih optimizacija. Pod njegovim rukovodstvom, autor ove knjige je radio 1971. godine diplomski rad pod naslovom *„Metoda za simultano nalaženje nula algebarskih jednačina i njena primena na ispitivanje stabilnosti sistema automatske regulacije i rešavanje nekih diferencijalnih, diferencnih i transcendentnih jednačina.“*

2. GRAUov pristup faktorizaciji. Veoma slično PREŠIĆevom metodu, GRAU [29] je razmatrao monični polinom $P(z)$ stepena n, definisan sa (3.7.1), i m moničnih polinoma $\varphi_\nu(z)$, $\nu = 1,\dots,m$, redom stepena n_ν, tj.

$$\varphi_\nu(z) = z^{n_\nu} + \sum_{j=0}^{n_\nu-1} z^j s_{ij}, \quad \nu = 1,\dots,m,$$

sa $\sum_{\nu=1}^{m} n_\nu = n$, tako da monični polinom stepena n,

$$g(z) = \prod_{\nu=1}^{m} \varphi_\nu(z),$$

predstavlja aproksimaciju za $P(z)$.

Pretpostavljajući da se, za svako ν, odgovarajući tačni faktor polinoma $P(z)$ razlikuje od $\varphi_\nu(z)$ za polinom stepena najviše $n_\nu - 1$, tj.

$$\Delta\varphi_\nu(z) = \sum_{j=0}^{n_\nu-1} z^j \Delta s_{\nu j},$$

tada je

$$g(z) + \Delta g(z) = \prod_{\nu=1}^{m} [\varphi_\nu(z) + \Delta\varphi_\nu(z)] \equiv P(z), \tag{3.7.10}$$

odakle se izjednačavanjem koeficijenata uz stepene od z može dobiti sistem od n nelinearnih jednačina za određivanje n nepoznatih veličina $\Delta s_{\nu j}$ ($\nu = 1,\dots,m;\ j = 0,1,\dots,n_\nu - 1$).

Glavna ideja GRAUovog pristupa je u linearizaciji prethodnog identiteta, tj. u aproksimaciji $\Delta g(z)$ sa totalnim diferencijalom od $g(z)$ u odnosu na $s_{\nu j}$. Dakle,

$$\Delta g(z) \approx \mathrm{d}g(z) = \sum_{\nu=1}^{m} h_\nu(z)\Delta\varphi_\nu(z) = \sum_{\nu=1}^{m} h_\nu(z) \sum_{j=0}^{n_\nu-1} z^j \Delta s_{\nu j}, \tag{3.7.11}$$

gde su $h_\nu(z)$ polinomi $g(z)/\varphi_\nu(z)$. Sada, korišćenjem (3.7.10) i (3.7.11), dobijamo (približni) identitet

$$P(x) - g(x) \equiv \sum_{\nu=1}^{m} h_\nu(z)\Delta\varphi_\nu(z) = \sum_{\nu=1}^{m} h_\nu(z) \sum_{j=0}^{n_\nu-1} z^j \Delta s_{\nu j}, \tag{3.7.12}$$

odakle se, izjednačavanjem koeficijenata uz odgovarajuće stepene od z, dolazi do sistema od n linearnih jednačina za određivanje n nepoznatih veličina $\Delta s_{\nu j}$, što se dalje može rešiti standardnim metodama.

Neke generalizacije GRAUovog metoda dali su CARSTENSEN[196] [14] i CARSTENSEN i SAKURAI[197] [15].

3. BAIRSTOWljev metod. Metod se koristi za određivanje kvadratnih faktora (delilaca) polinoma, na osnovu kojih se mogu odrediti kompleksne nule polinoma.

Neka je

$$P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n \quad (a_0 \neq 0)$$

polinom sa realnim koeficijentima. S obzirom da se kompleksne nule ovog polinoma javljaju kao konjugovano kompleksni parovi, to polinom P ima realne kvadratne faktore.[198]

U cilju određivanja jednog takvog faktora $\widehat{m}(x) = x^2 + \widehat{p}x + \widehat{q}$, pretpostavimo da nam je poznata izvesna aproksimacija $m_0(x) = x^2 + p_0x + q_0$ $(p_0, q_0 \in \mathbb{R})$.

Pre nego što damo osnovne jednakosti koje definišu BAIRSTOWljev metod izvešćemo jednu rekurentnu relaciju koja se odnosi na deljenje $P(x)$ kvadratnim faktorom $m(x) = x^2 + px + q$.

Neka su polinomi Q i R u jednakosti

$$P(x) = m(x)Q(x) + R(x) \tag{3.7.13}$$

dati sa

$$Q(x) = b_2x^{n-2} + \cdots + b_{n-1}x + b_n \quad \text{i} \quad R(x) = ux + v.$$

Upoređivanjem koeficijenata uz odgovarajuće stepene na levoj i desnoj strani jednakosti (3.7.13) dobijamo

$$\begin{aligned} a_0 &= b_2, \quad a_1 = b_3 + pb_2, \\ a_{i-2} &= b_i + pb_{i-1} + qb_{i-2}, \quad i = 4, \ldots, n, \\ a_{n-1} &= pb_n + qb_{n-1} + u, \quad a_n = qb_n + v. \end{aligned}$$

Za $b_0 = b_1 = 0$, iz prethodnih jednakosti sleduje

[196] CARSTEN CARSTENSEN (1962 –), nemački matematičar, bavi se numeričkom analizom, a posebno metodima za parcijalne diferencijalne jednačine, kao i odgovarajućim primenama.

[197] TETSUYA SAKURAI (1961 –), japanski matematičar, bavi se numeričkom analizom, posebno iterativnim procesima i paralelnim algoritmima.

[198] Kvadratni faktori sa realnim koeficijentima.

$$b_i = a_{i-2} - pb_{i-1} - qb_{i-2}, \quad i = 2, \ldots, n, \tag{3.7.14}$$

i

$$u = a_{n-1} - pb_n - qb_{n-1} = b_{n+1}, \quad v = a_n - qb_n, \tag{3.7.15}$$

gde smo b_{n+1} definisali prirodnim produženjem rekurentne relacije (3.7.14) za $i = n+1$.

Na osnovu (3.7.14) i (3.7.15) zaključujemo da koeficijenti polinoma Q i R na jedinstven način zavise od p i q.

Dakle, za određivanje kvadratnog faktora polinoma P dovoljno je naći rešenje sistema jednačina

$$u(p,q) = 0, \quad v(p,q) = 0.$$

BAIRSTOWljev metod se zasniva na rešavanju ovog sistema jednačina primenom metoda NEWTON-KANTOROVIČa. Naime, polazeći od (p_0, q_0) generiše se niz parova $\{(p_k, q_k)\}_{k=1,2,\ldots}$ pomoću

$$\begin{bmatrix} p_{k+1} \\ q_{k+1} \end{bmatrix} = \begin{bmatrix} p_k \\ q_k \end{bmatrix} - W_k^{-1} \begin{bmatrix} u(p_k, q_k) \\ v(p_k, q_k) \end{bmatrix}, \quad k = 0, 1, \ldots, \tag{3.7.16}$$

gde je JACOBIeva matrica data sa

$$W_k = \begin{bmatrix} \dfrac{\partial u}{\partial p} & \dfrac{\partial u}{\partial q} \\ \dfrac{\partial v}{\partial p} & \dfrac{\partial v}{\partial q} \end{bmatrix}_{p=p_k, q=q_k}$$

Da bismo odredili elemente ove matrice uvedimo oznake

$$r_i = \frac{\partial b_i}{\partial p}, \quad t_i = \frac{\partial b_i}{\partial q}, \quad i = 0, 1, \ldots, n+1,$$

i

$$s_i = r_i - t_{i+1}, \quad i = 0, 1, \ldots, n.$$

Diferenciranjem po p i q prethodno dokazanih rekurentnih relacija za koeficijente b_i, dobijamo

$$r_i = -b_{i-1} - p\frac{\partial b_{i-1}}{\partial p} - q\frac{\partial b_{i-2}}{\partial p} = -b_{i-1} - pr_{i-1} - qr_{i-2} \tag{3.7.17}$$

i

$$t_i = -p\frac{\partial b_{i-1}}{\partial q} - b_{i-2} - q\frac{\partial b_{i-2}}{\partial q} = -pt_{i-1} - b_{i-2} - qt_{i-2}, \tag{3.7.18}$$

gde je $i = 2,\ldots,n+1$. Ako od jednakosti (3.7.17) oduzmemo jednakost (3.7.18) u kojoj je indeks i zamenjen sa $i+1$, dobijamo

$$s_i = -ps_{i-1} - qs_{i-2}, \quad i = 2,\ldots,n. \tag{3.7.19}$$

Kako je $b_0 = b_1 = 0$ i $b_2 = a_0$, imamo $r_0 = r_1 = r_2 = 0$, $t_0 = t_1 = t_2 = 0$, $s_0 = s_1 = 0$, što zajedno sa (3.7.19) daje

$$s_i = r_i - t_{i+1} = 0, \quad \text{tj.} \quad \frac{\partial b_{i+1}}{\partial q} = \frac{\partial b_i}{\partial p}, \quad i = 0,1,\ldots,n.$$

Kako je na osnovu prethodnog

$$W_k = \begin{bmatrix} r_{n+1} & r_n \\ -qr_n & -(b_n + qr_{n-1}) \end{bmatrix}_{p=p_k, q=q_k},$$

iterativni proces (3.7.16) postaje

$$\begin{bmatrix} p_{k+1} \\ q_{k+1} \end{bmatrix} = \begin{bmatrix} p_k \\ q_k \end{bmatrix} - \frac{1}{D}\begin{bmatrix} -(b_n + qr_{n-1})b_{n+1} - r_n(a_n - qb_n) \\ qr_n b_{n+1} + r_{n+1}(a_n - qb_n) \end{bmatrix}_{p=p_k, q=q_k}$$

gde je $D = qr_n^2 - (b_n + qr_{n-1})r_{n+1}$.

Naravno, iterativni proces dobijen na ovaj način egzistira ako je $D \neq 0$. Red konvergencije je dva, s obzirom da se koristi metod NEWTON-KANTOROVIČa.

Jedna opštija klasa metoda BAIRSTOWljevog tipa razmatrana je u radu [9]. Naime, pokazano je da je BAIRSTOWljev metod samo jedan od metoda iz čitave familije algoritama za određivanje kvadratnih faktora polinoma. U radu se sugeriše izbor pogodnog metoda iz ove familije za rešavanje konkretnog problema.

5.3.8 Metodi za simultano određivanje korena

U novije vreme sve više se proučavaju metodi za simultano (istovremeno) određivanje svih korena algebarske jednačine

$$P(z) = z^n + a_1 z^{n-1} + \cdots + a_{n-1}z + a_n = 0, \tag{3.8.1}$$

gde su a_ν, $\nu = 1,\dots,n$, u opštom slučaju, kompleksni koeficijenti.

Pretpostavimo da su koreni $r_1,\dots,r_n$ algebarske jednačine (3.8.1) međusobno različiti i neka su njihove približne vrednosti u k-tom iterativnom koraku $z_i^{(k)}$, $i = 1,\dots,n$. S obzirom da je $r_i = z_i^{(k)} + \Delta z_i^{(k)}$, $i = 1,\dots,n$, gde su $\Delta z_i^{(k)}$ odgovarajuće greške pojedinih korena, važi identičnost

$$\prod_{i=1}^{n}\bigl(z - \bigl(z_i^{(k)} + \Delta z_i^{(k)}\bigr)\bigr) = P(z),$$

odakle, razvijanjem u red po stepenima od $\Delta z_i^{(k)}$, sleduje

$$\prod_{i=1}^{n}\bigl(z - z_i^{(k)}\bigr) - \sum_{i=1}^{n}\Delta z_i^{(k)}\left(\prod_{\substack{m=1\\ m\neq i}}^{n}\bigl(z - z_m^{(k)}\bigr)\right) + \sum_{i,j}\Delta z_i^{(k)}\Delta z_j^{(k)}\left(\prod_{\substack{m=1\\ m\neq i,j}}^{n}\bigl(z - z_m^{(k)}\bigr)\right) - \cdots = P(z).$$

Pretpostavljajući da su greške $\Delta z_i^{(k)}$ dovoljno male po modulu, na levoj strani u poslednjoj jednakosti možemo uzeti samo prva dva člana. Ako u tako dobijenoj jednakosti stavimo redom $z := z_i^{(k)}$, $i = 1,\dots,n$, dobijamo

$$\Delta z_i^{(k)} = -\frac{P\bigl(z_i^{(k)}\bigr)}{\prod\limits_{\substack{m=1\\ m\neq i}}^{n}\bigl(z_i^{(k)} - z_m^{(k)}\bigr)} \quad (i = 1,\dots,n). \tag{3.8.2}$$

Na ovaj način dolazimo do simultanog iterativnog procesa

$$z_i^{(k+1)} = z_i^{(k)} + \Delta z_i^{(k)} \quad (i = 1,\dots,n;\ k = 0,1,\dots). \tag{3.8.3}$$

Ako definišemo polinom n-tog stepena pomoću $Q_k(z) = \prod\limits_{m=1}^{n}\bigl(z - z_m^{(k)}\bigr)$, tada se (3.8.2) može predstaviti u obliku koji podseća na NEWTONovu korekciju, tj.

$$\Delta z_i^{(k)} = -\frac{P\bigl(z_i^{(k)}\bigr)}{Q_k'\bigl(z_i^{(k)}\bigr)} \quad (i = 1,\dots,n).$$

Primetimo da polinom

$$Q_k(z) = z^n - \sigma_1 z^{n-1} + \sigma_2 z^{n-2} - \cdots + (-1)^n\sigma_n, \tag{3.8.4}$$

gde su $\sigma_1, \sigma_2, \ldots, \sigma_n$ elementarne simetrične funkcije nula $z_1^{(k)}, z_2^{(k)}, \ldots, z_n^{(k)}$ (videti (3.1.2)), teži ka $P(z)$, kada $z_i^{(k)} \to r_i$ $(i = 1, \ldots, n)$. U novo uvedenoj notaciji, iterativni proces (3.8.3) postaje

$$z_i^{(k+1)} = z_i^{(k)} - \frac{P\left(z_i^{(k)}\right)}{Q_k'\left(z_i^{(k)}\right)} \quad (i = 1, \ldots, n;\ k = 0, 1, \ldots). \tag{3.8.5}$$

Jedno kraće izvođenje formula (3.8.5) može se dati polazeći od faktorizacije

$$P(z) = \prod_{m=1}^{n} (z - r_m) = (z - r_i) \prod_{\substack{m=1 \\ m \neq i}}^{n} (z - r_m),$$

tj. od identiteta koji iz njega sleduje

$$r_i = z - \frac{P(z)}{\prod\limits_{\substack{m=1 \\ m \neq i}}^{n} (z - r_m)}. \tag{3.8.6}$$

Uzimajući za r_m $(m \neq i)$ približne vrednosti $z_m^{(k)}$, tj. $r_m := z_m^{(k)}$ i $z := z_i^{(k)}$, desna strana u (3.8.6), naravno, neće biti jednaka tačnoj nuli r_i, već će to biti neka nova aproksimacija za tu nulu, koju ćemo ozanačiti sa $z_i^{(k+1)}$. Tako dobijamo iterativni proces (3.8.5).

Formule (3.8.5) su otkrivane više puta od strane većeg broja autora. Zato se mogu i sresti više načina izvođenja ovih formula. Primetimo da formule (3.8.5) podsećaju na klasičan NEWTONov metod, ali se od njega suštinski razlikuju. Naime, ovde imamo n iterativnih formula koje generišu n nizova koji su međusobno zavisni; $(k+1)$-vi član jednog niza zavisi od k-tih članova svih nizova. Napomenimo da se ove formule mogu naći u sabranim delima WEIERSTRASSa. Naime, WEIERSTRASS [138] ih je koristio u vezi sa dokazom osnovne teoreme algebre. Međutim, za određivanje nula polinoma, ove formule se počinju koristiti tek u drugoj polovini dvadesetog veka.

Iterativni proces (3.8.5) ima kvadratnu konvergenciju. On je, u stvari, ekvivalentan metodu NEWTON-KANTOROVIČa, primenjenog na rešavanje sistema nelinearnih jednačina, dobijenog iz VIÈTEovih formula (3.1.2) za polinom (3.8.1).

Da bismo ovo dokazali, definisaćemo polinome $R^{(\nu)}(z)$ stepena $n-1$, izostavljanjem po jedne nule u polinomu $Q_k(z)$, i pri tome ćemo, u cilju uprošćenja, izostavljati indeks k u oznakama za polinom i odgovarajuće nule, tj. jednostavno ćemo pisati $Q(z)$ umesto $Q_k(z)$ i z_ν umesto $z_\nu^{(k)}$. Dakle,

$$R^{(\nu)}(z)=\frac{Q_k(z)}{z-z_\nu}=\prod_{\substack{j=1\\ j\neq\nu}}^{n}(z-z_j),\quad \nu=1,\dots,n,$$

čiji je razvijeni oblik, u skladu sa (3.8.4),

$$R^{(\nu)}(z)=z^{n-1}-\sigma_1^{(\nu)}z^{n-2}+\sigma_2^{(\nu)}z^{n-3}-\cdots+(-1)^{n-1}\sigma_{n-1}^{(\nu)},$$

gde su $\sigma_1^{(\nu)},\sigma_2^{(\nu)},\dots,\sigma_{n-1}^{(\nu)}$ odgovarajuće simetrične funkcije iz kojih je isključena nula z_ν. Nije teško zaključiti da je $Q_k'(z_\nu)=R^{(\nu)}(z_\nu)$, $\nu=1,\dots,n$.

Lema 3.8.1. *Za bilo koji polinom $Q_k(z)$ sa prostim nulama z_ν, $\nu=1,\dots,n$, matrica*

$$W=\begin{bmatrix}1 & 1 & \cdots & 1\\ \sigma_1^{(1)} & \sigma_1^{(2)} & & \sigma_1^{(n)}\\ \vdots & & & \\ \sigma_{n-1}^{(1)} & \sigma_{n-1}^{(2)} & & \sigma_{n-1}^{(n)}\end{bmatrix}$$

je regularna i njena inverzna matrica je data sa

$$W^{-1}=\begin{bmatrix}D_1z_1^{n-1} & -D_1z_1^{n-2} & \cdots & (-1)^{n-1}D_1\\ D_2z_2^{n-1} & -D_2z_2^{n-2} & \cdots & (-1)^{n-1}D_2\\ \vdots & & & \\ D_nz_n^{n-1} & -D_nz_n^{n-2} & \cdots & (-1)^{n-1}D_n\end{bmatrix},$$

gde su $D_\nu=1/Q_k'(z_\nu)$, $\nu=1,\dots,n$.

Predstavimo sada iterativne formule (3.8.5) u vektorskom obliku

$$\boldsymbol{z}^{(k+1)}=T(\boldsymbol{z}^{(k)}),\quad k=0,1,\dots, \tag{3.8.7}$$

gde su $T(\boldsymbol{z})=\boldsymbol{z}-\boldsymbol{e}(\boldsymbol{z})$ i

$$\boldsymbol{z}=\begin{bmatrix}z_1\\ z_2\\ \vdots\\ z_n\end{bmatrix},\quad \boldsymbol{e}(\boldsymbol{z})=\begin{bmatrix}e_1\\ e_2\\ \vdots\\ e_n\end{bmatrix},\quad e_\nu=\frac{P(z_\nu)}{Q'(z_\nu)},\quad Q(z)=\prod_{j=1}^{n}(z-z_j).$$

Primenićemo sada metod NEWTON-KANTOROVIČa na sistem nelinearnih jednačina, dobijenog iz VIÈTEovih formula (3.1.2) za polinom (3.8.1), tj. na

$$\boldsymbol{f}(z) = \begin{bmatrix} \sigma_1 + a_1 \\ \sigma_2 - a_2 \\ \vdots \\ \sigma_n - (-1)^n a_n \end{bmatrix} = \mathbf{0}.$$

U radu TOŠIĆa[199] i MILOVANOVIĆa [128] je dokazano da se, u ovom slučaju, metod NEWTON-KANTOROVIČa

$$z^{(k+1)} = z^{(k)} - W^{-1}(z^{(k)})\boldsymbol{f}(z(k)), \quad k = 0, 1, \ldots,$$

svodi na (3.8.7). Zaista, JACOBIeva matrica $W(z)$ je upravo ona koja se pojavljuje u lemi 3.8.1, tako da je korekcija

$$W^{-1}(z)\boldsymbol{f}(z) = \begin{bmatrix} D_1 z_1^{n-1} & -D_1 z_1^{n-2} & \cdots & (-1)^{n-1} D_1 \\ D_2 z_2^{n-1} & -D_2 z_2^{n-2} & \cdots & (-1)^{n-1} D_2 \\ \vdots & & & \\ D_n z_n^{n-1} & -D_n z_n^{n-2} & \cdots & (-1)^{n-1} D_n \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 + a_1 \\ \sigma_2 - a_2 \\ \vdots \\ \sigma_n - (-1)^n a_n \end{bmatrix} = \boldsymbol{e}(z).$$

Da bi simultani metod (3.8.5) bio definisan neophodno je da sve startne vrednosti budu međusobno različite, tj. $z_i^{(0)} \neq z_j^{(0)}$ $(i \neq j)$. Može se pokazati (videti, na primer, [17]) da metod ima jedno interesantno svojstvo. Naime, bez obzira na izbor startnih vrednosti (sem da budu međusobno različite), posle prve iteracije imamo da je

$$\sum_{i=1}^{n} z_i^{(1)} = -a_1, \tag{3.8.8}$$

gde je a_1 koeficijent u jednačini (3.8.1). Naravno, nije teško primetiti da je posle svake iteracije zbir aproksimacija $z_i^{(k)}$ jednak $-a_1$. Daćemo sada jedan interesantan dokaz jednakosti (3.8.8).

[199] DOBRILO Đ. TOŠIĆ (1932–), dugogodišnji profesor matematike na Elektrotehničkom fakultetu u Beogradu. Na početku karijere radio je u Institutu za nuklearne nauke u Vinči i na Elektronskom fakultetu u Nišu. Poznat je po radovima iz oblasti Fizike jonizovanog gasa i Fizike plazme, a u matematici iz oblasti Numeričke kompleksne analize i Specijalnih funkcija.

Kako su startne vrednosti međusobno različite, zaključujemo da racionalna funkcija $z \mapsto P(z)/Q_0(z)$ ima samo proste polove. Neka je $r > \max_i |z_i^{(0)}|$. Tada je

$$I = \frac{1}{2\pi \mathrm{i}} \oint\limits_{|z|=r} \frac{P(z)}{Q_0(z)}\,\mathrm{d}z = \sum \operatorname{Res} \frac{P(z)}{Q_0(z)} = \sum_{i=1}^{n} \frac{P(z_i^{(0)})}{Q_0'(z_i^{(0)})}.$$

S druge strane, zamenom $w = 1/z$, poslednji integral postaje

$$I = \frac{1}{2\pi \mathrm{i}} \oint\limits_{|w|=1/r} \frac{1}{w^2} \frac{P(1/w)}{Q_0(1/w)}\,\mathrm{d}w = \lim_{w\to 0} \frac{\mathrm{d}}{\mathrm{d}w} \left(\frac{w^n P(1/w)}{w^n Q_0(1/w)} \right),$$

tj.

$$I = a_i + \sum_{i=1}^{n} z_i^{(0)}.$$

Sumiranjem (3.8.5) za $i = 1, \ldots, n$, pri $k = 0$, na osnovu prethodnog, dobijamo

$$\sum_{i=1}^{n} z_i^{(1)} = \sum_{i=1}^{n} z_i^{(0)} - \left(a_1 + \sum_{i=1}^{n} z_i^{(0)} \right) = -a_1.$$

Navešćemo sada dve teoreme koje se odnose na konvergenciju metoda (3.8.5). Dokazi ovih rezultata mogu se dati matematičkom indukcijom (videti [17, 109]).

Teorema 3.8.1. *Ako je*

$$d = \min_{i\neq j} |r_i - r_j| \quad i \quad |z_i^{(0)} - r_i| < \rho \quad (i = 1, \ldots, n),$$

gde je

$$\rho = d \frac{(1+q)^{1/(n-1)} - 1}{2(1+q)^{(n-1)} - 1} \quad (0 < q < 1),$$

metod (3.8.5) *konvergira, pri čemu je, za svako* $k \in \mathbb{N}$,

$$|z_i^{(k)} - r_i| \leq \rho q^{2^k - 1} \quad (i = 1, \ldots, n).$$

Teorema 3.8.2. *Neka su početne vrednosti* $z_i^{(0)}$ *međusobno različiti kompleksni brojevi za koje važe nejednakosti*

$$\frac{|P(z_i^{(0)})|}{|Q_0'(z_i^{(0)})|} \leq \frac{|Q_0'(z_i^{(0)})|}{6nM(0)^{n-2}} \quad (i = 1, \ldots, n),$$

gde je $M(0)=\max\limits_{i\neq j}\left|z_i^{(0)}-z_j^{(0)}\right|$, *tada simultani iterativni proces* (3.8.5) *konvergira, tj.* $\lim\limits_{k\to+\infty} z_i^{(k)}=r_i$, $i=1,\ldots,n$. *Štaviše, za svako* $i=1,\ldots,n$, *svi članovi niza* $\{z_i^{(k)}\}_{k\in N}$ *se nalaze u krugu*

$$K_i=K\left(z_i^{(0)}\right)=\left\{z \;\middle|\; \left|z-z_i^{(1)}\right|\leq\left|z_i^{(1)}-z_i^{(0)}\right|\right\}.$$

U radovima [110] i [111] S. B. PREŠIĆ je razmatrao postupak za faktorizaciju polinoma, koji se u slučaju linearnih faktora svodi na metod (3.8.5). U radu [107] (videti, takođe, [108]) razmatran je metod za određivanje $m(<n)$ nula polinoma (3.8.1). Jedan drugačiji način izvođenja ovog metoda i generalizacije za slučaj višestrukih nula polinoma dat je u radu [62].

Polazeći od iterativne formule (3.8.5) i disjunktnih realnih intervala $X_i^{(0)}=\left[x_i^{(0)}-\varepsilon_i^{(0)},x_i^{(0)}+\varepsilon_i^{(0)}\right]$ koji sadrže nule r_i, $i=1,\ldots,n$, HERZBERGER[200] [41, 42] je formulisao intervalni metod za simultano određivanje nula polinoma P:

$$X_i^{(k+1)}=\left\{x_i^{(k)}-\frac{P\left(x_i^{(k)}\right)}{\prod\limits_{\substack{m=1\\m\neq i}}^{n}\left(x_i^{(k)}-X_m^{(k)}\right)}\right\}\bigcap X_i^{(k)}\quad(i=1,\ldots,n;\ k=0,1,\ldots)$$

dokazujući da za svako $i=1,\ldots,n$ i svako $m\in\mathbb{N}_0$ važi:

(a) $X_i^{(k+1)}\subset X_i^{(k)}$;

(b) $r_i\in X_i^{(k)}$;

(c) postoji konstanta $q(>0)$ takva da je $\varepsilon^{(k+1)}\leq q\left(\varepsilon^{(k)}\right)^2$, gde je $\varepsilon^{(k)}=\max\limits_i\varepsilon_i^{(k)}$.

Opšti slučaj kompleksnih nula razmatrao je PETKOVIĆ [95] korišćenjem kompleksne kružne aritmetike.[201] Naime, neka su nule polinoma P izolovane u disjunktnim kružnim intervalima $Z_i^{(0)}=\left\{z_i^{(0)};\varepsilon_i^{(0)}\right\}$, $i=1,\ldots,n$, i neka je u (3.8.6)

[200] JÜRGEN HERZBERGER (1940 – 2009), nemački matematičar.

[201] Kompleksni kružni interval $Z=\{c;r\}$ je disk u kompleksnoj ravni s centrom u tački c i poluprečnikom r, tj. $Z=\{z\mid|z-c|\leq r\}$. Tačka a se predstavlja pomoću $\{a;0\}$. Kompleksna kružna aritmetika je definisana nad kružnim intervajima $Z_k=\{c_k,r_k\}$ $(k=1,2)$ pomoću

$$Z_1\pm Z_2=\{c_1\pm c_2;r_1\pm r_2\},\quad Z_1Z_2=\{c_1c_2;|c_2|r_1+|c_1|r_2+r_1r_2\},$$
$$Z_1\div Z_2=Z_1Z_2^{-1}\ (0\notin Z_2),\quad Z^{-1}=\{\alpha\overline{c};\alpha r\}\ (\alpha=|c|^2-r^2;\ 0\notin Z_2).$$

Za detalje videti [39].

uzeto $z = z_i^{(0)}$. Imajući u vidu da $r_m \in Z_m^{(0)}$ $(m = 1, \ldots, n)$, na osnovu osobine inkluzivne izotonosti sleduje

$$r_i \in Z_i^{(1)} = \left\{ z_i^{(0)} - \frac{P\left(z_i^{(0)}\right)}{\prod\limits_{\substack{m=1 \\ m \neq i}}^{n} \left(z_i^{(0)} - Z_m^{(0)}\right)} \right\} \bigcap Z_i^{(0)}.$$

Poslednja formula sugeriše iterativni metod

$$Z_i^{(k+1)} = \left\{ z_i^{(k)} - \frac{P\left(z_i^{(k)}\right)}{\prod\limits_{\substack{m=1 \\ m \neq i}}^{n} \left(z_i^{(k)} - Z_m^{(k)}\right)} \right\} \bigcap Z_i^{(k)} \quad (i = 1, \ldots, n;\ k = 0, 1, \ldots). \tag{3.8.9}$$

Korišćenjem kompleksne kružne aritmetike, u pomenutom radu [95], iterativni metod (3.8.9) je sveden na oblik

$$Z_i^{(k+1)} = \left\{ z_i^{(k)} - - \frac{P\left(z_i^{(k)}\right)}{\prod\limits_{\substack{m=1 \\ m \neq i}}^{n} \left(z_i^{(k)} - z_m^{(k)}\right)}; \frac{a^{(k)}(n-1)\left|P\left(z_i^{(k)}\right)\right|\varepsilon^{(k)}}{\left(\rho^{(k)}\right)^n} \right\} \bigcap Z_i^{(k)},$$

gde su

$$\varepsilon^{(k)} = \max_i \varepsilon_i^{(k)}, \quad \rho^{(k)} = \min_{i \neq j} \left\{ \left|z_i^{(k)} - z_j^{(k)}\right| - \varepsilon_j^{(k)} \right\}$$

i $a^{(k)}$ pozitivan broj za koji važi $a^{(k)} \geq \left(1 + \varepsilon^{(k)}/\rho^{(k)}\right)^{n-1}$.

Pri praktičnoj realizaciji simultanih metoda najčešće se koristi GAUSS-SEIDELov pristup, tj. korišćenje aproksimacija dobijenih u istoj iteraciji. Ovakav postupak ne samo da ubrzava konvergenciju osnovnog metoda bez dodatnih izračunavanja, već je i pogodan sa stanovišta zauzeća memorijskog prostora. Analiza konvergencije ovako modifikovanih metoda je omogućena uvođenjem koncepta R-reda konvergencije (videti odeljak 3.2.5).

Osnovna modifikacija procesa (3.8.5) dobijena pomoću GAUSS-SEIDELovog pristupa je

$$z_i^{(k+1)} = z_i^{(k)} - \frac{P\left(z_i^{(k)}\right)}{\prod\limits_{m=1}^{i-1} \left(z_i^{(k)} - z_m^{(k+1)}\right) \prod\limits_{m=i+1}^{n} \left(z_i^{(k)} - z_m^{(k)}\right)} \quad (i = 1, \ldots, n). \tag{3.8.10}$$

Programska realizacija ove varijante je jednostavnija u odnosu na osnovni oblik (3.8.5). Novo izračunata aproksimacija $z_i^{(k)}$ se odmah smešta u istu poziciju

gde je bila vrednost $z_i^{(k)}$, tako da po obavljenoj iteraciji (po svim nulama) nema zamene $z_i^{(k)} := z_i^{(k+1)}$, $i = 1, \ldots, n$.

Izložićemo sada još neke simultane metode koji se javljaju u primenama. Najpre, uvedimo izvesna označavanja koja omogućavaju bolju preglednost kod definisanja dve susedne iteracije:

1° aproksimacije nula u k-toj iteraciji $z_1^{(k)}, \ldots, z_n^{(k)}$ označavaćemo prosto sa $z_1, \ldots, z_n$, a nove aproksimacije koje dobijamo primenom simultanog metoda redom sa $\hat{z}_1, \ldots, \hat{z}_n$;

2° $Q(z) = \prod_{m=1}^{n} (z - z_m)$;

3° $W_i = P(z_i)/Q'(z_i)$ (WEIERSTRASSova korekcija, tj. korekcija kod metoda (3.8.5));

4° $N_i = P(z_i)/P'(z_i)$ (NEWTONova korekcija).

U novo uvedenoj notaciji, osnovni metod (3.8.5) postaje

$$\hat{z}_i = z_i - W_i \quad (i = 1, \ldots, n), \tag{3.8.11}$$

dok je metod (3.8.10)

$$\hat{z}_i = z_i - \frac{P(z_i)}{\prod_{m=1}^{i-1} (z - \hat{z}_m) \prod_{m=i+1}^{n} (z - z_m)} \quad (i = 1, \ldots, n). \tag{3.8.12}$$

Ranije smo dokazali da osnovni metod (3.8.11) ima kvadratnu konvergenciju. U radu [2] dokazano je da je R-red konvergencije iterativnog procesa (3.8.12), u zavisnosti od stepena polinoma, jednak $r(n) = 1 + \sigma_n$, gde je σ_n jedinstven pozitivan koren jednačine $\sigma^n - \sigma - 1 = 0$.

Jedna druga modifikacija procesa (3.8.11), sa kubnom konvergencijom, koja koristi popravku W_i (videti, na primer, [90]) data je pomoću

$$\hat{z}_i = z_i - \frac{P(z_i)}{\prod_{\substack{m=1 \\ m \neq i}}^{n} (z_i - z_m + W_m)} \quad (i = 1, \ldots, n). \tag{3.8.13}$$

Dalje ubrzavanje konvergencije može se postići kombinacijom formula (3.8.12) i (3.8.13). Na primer, PETKOVIĆ i MILOVANOVIĆ [99]) su razmatrali metod

$$\hat{z}_i = z_i - \frac{P(z_i)}{\prod_{m=1}^{i-1} (z_i - \hat{z}_m) \prod_{m=i+1}^{n} (z_i - z_m + W_m)} \quad (i = 1, \ldots, n), \tag{3.8.14}$$

čiji je R-red konvergencije $r(n) = 1 + \sigma_n$, gde je σ_n jedinstven pozitivan koren jednačine $\sigma^n - \sigma - \sum_{k=0}^{n-1} \sigma^k = 0$.

Jedan drugačiji pristup modifikaciji osnovne formule (3.8.11) dat je u radu [8]:

$$\hat{z}_i = z_i - \frac{W_i}{1 + \sum\limits_{\substack{m=1 \\ m\neq i}}^{n} \frac{W_m}{z_i - z_m}} \quad (i = 1, \ldots, n). \tag{3.8.15}$$

Konvergencija ovog procesa je kubna. Do formule (3.8.15) se može jednostavno doći konstrukcijom LAGRANGEovog interpolacionog polinoma[202] za funkciju (polinom) $z \mapsto P(z) - Q(z)$ u čvorovima $z_1, \ldots, z_n$, tj.

$$P(z) - Q(z) = \sum_{m=1}^{n} P(z_m) \frac{Q(z)}{(z - z_m) Q'(z_m)}.$$

Ako za z uzmemo bilo koju nulu polinoma r_i, iz poslednje jednakosti dobijamo

$$r_i = z_i - \frac{W_i}{1 - \sum\limits_{\substack{m=1 \\ m\neq i}}^{n} \frac{W_m}{z_m - r_i}} \quad (i = 1, \ldots, n). \tag{3.8.16}$$

Najzad, stavljajući $r_i := z_i$, desna strana u (3.8.16) daje aproksimaciju za nulu r_i. Označavajući ovu aproksimaciju sa $\hat{z}_i$ dobijamo formulu (3.8.15).

Slično konstrukciji formule (3.8.13), na osnovu (3.8.15) se može dobiti formula četvrtog reda (videti [91]):

$$\hat{z}_i = z_i - \frac{W_i}{1 + \sum\limits_{\substack{m=1 \\ m\neq i}}^{n} \frac{W_m}{z_i - W_i - z_m}} \quad (i = 1, \ldots, n). \tag{3.8.17}$$

Koristeći logaritamski izvod polinoma P moguće je izvesti formulu [64] (videti, takođe, [1], [7], [22])

$$\hat{z}_i = z_i - \frac{1}{\frac{1}{N_i} - \sum\limits_{\substack{m=1 \\ m\neq i}}^{n} \frac{1}{z_i - z_m}} \quad (i = 1, \ldots, n), \tag{3.8.18}$$

[202] Interpolacioni procesi biće razmatrani u posebnoj knjizi iz ove serije.

koja ima kubnu konvergenciju. Naime, na osnovu

$$\frac{P'(z)}{P(z)} = \sum_{m=1}^{n} \frac{1}{z - r_m},$$

važi

$$r_i = z - \frac{1}{\dfrac{P'(z)}{P(z)} - \displaystyle\sum_{\substack{m=1 \\ m\neq i}}^{n} \frac{1}{z - r_m}},$$

odakle zamenom $z := z_i$, $r_m := z_m$ $(m \neq i)$, na desnoj strani u prethodnoj jednakosti dobijamo novu aproksimaciju r_i. Ako ovu aproksimaciju označimo sa $\hat{z}_i$ dobijamo formulu (3.8.18). Korišćenjem GAUSS-SEIDELovog pristupa može se dobiti poboljšani metod

$$\hat{z}_i = z_i - \frac{1}{\dfrac{1}{N_i} - \displaystyle\sum_{m=1}^{i-1} \frac{1}{z_i - \hat{z}_m} - \sum_{m=i+1}^{n} \frac{1}{z_i - z_m}} \quad (i = 1, \ldots, n), \tag{3.8.19}$$

koji ima R-red konvergencije $r(n) = 2 + \sigma_n (> 3)$, gde je σ_n jedinstven pozitivni koren jednačine $\sigma^n - \sigma - 2 = 0$ (videti, ALEFELD[203] i HERZBERGER [2]).

Sa NEWTONovom korekcijom, u radu [90] je dobijena sledeća modifikacija formule (3.8.18):

$$\hat{z}_i = z_i - \frac{1}{\dfrac{1}{N_i} - \displaystyle\sum_{\substack{m=1 \\ m\neq i}}^{n} \frac{1}{z_i - z_m + N_m}} \quad (i = 1, \ldots, n), \tag{3.8.20}$$

čiji je red konvergencije četiri.

Analogno formuli (3.8.14) može se dobiti ubrzani iterativni proces bez dodatnih izračunavanja (videti rad MILOVANOVIĆa i PETKOVIĆa [81]):

$$\hat{z}_i = z_i - \frac{1}{\dfrac{1}{N_i} - \displaystyle\sum_{m=1}^{i-1} \frac{1}{z - \hat{z}_m} - \sum_{m=i+1}^{n} \frac{1}{z_i - z_m + N_m}} \quad (i = 1, \ldots, n) \tag{3.8.21}$$

čiji je R-red konvergencije $r(n) = 2(1 + \sigma_n)$, gde je $\sigma_n \in (1, 2)$ jedinstveni pozitivni koren jednačine $\sigma^n - \sigma - 1 = 0$.

[203] GÖTZ ALEFELD (1941 –), nemački matematičar.

Ilustrovaćemo određivanje R-reda konvergencije na primeru metoda (3.8.21). Neka su

$$d = \min_{i \neq j} |r_i - r_j|, \quad v_j^{(k)} = z_j^{(k)} - r_j, \quad \hat{g}_i^{(k)} = \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{z_i^{(k)} - r_i}{z_i^{(k)} - r_j}$$

$$g_i^{(k)} = \sum_{j=1}^{i-1} \frac{\left(z_i^{(k)} - r_i\right)\left(r_j - z_j^{(k+1)}\right)}{\left(z_i^{(k)} - r_j\right)\left(z_i^{(k)} - z_j^{(k+1)}\right)} + \sum_{j=i+1}^{n} \frac{\left(z_i^{(k)} - r_i\right)\left(r_j - w_j^{(k)}\right)}{\left(z_i^{(k)} - r_j\right)\left(z_i^{(k)} - w_j^{(k)}\right)},$$

gde je $w_j^{(k)}$ aproksimacija dobijena po NEWTONovom metodu, tj.

$$w_j^{(k)} = z_j^{(k)} - P\left(z_j^{(k)}\right) / P'\left(z_j^{(k)}\right).$$

Nije teško pokazati da se NEWTONov metod može predstaviti u obliku

$$w_j^{(k)} - r_j = \frac{\hat{g}_j^{(k)}}{1 + \hat{g}_j}\left(z_j^{(k)} - r_j\right) = \frac{\hat{g}_j^{(k)}}{1 + \hat{g}_j^{(k)}} v_j^{(k)}. \tag{3.8.22}$$

Slično, za iterativni proces (3.8.21) imamo

$$z_i^{(k+1)} = r_i + \frac{g_i^{(k)}}{1 + g_i^{(k)}}\left(z_i^{(k)} - r_i\right) \qquad (i = 1, \ldots, n;\ k = 0, 1, \ldots),$$

odakle je

$$v_i^{(k+1)} = \frac{g_i^{(k)}}{1 + g_i^{(k)}} v_i^{(k)} \qquad (i = 1, \ldots, n;\ k = 0, 1, \ldots). \tag{3.8.23}$$

Pretpostavimo da su početne aproksimacije odabrane tako da su zadovoljeni uslovi

$$\left|v_i^{(0)}\right| < \frac{1}{q} = \frac{d}{2n-1} \quad (i = 1, \ldots, n). \tag{3.8.24}$$

Tada za $i \neq j$ imamo

$$\left|z_i^{(0)} - r_j\right| \geq |r_i - r_j| - \left|z_i^{(0)} - r_i\right| > d - \frac{d}{2n-1},$$

$$\left|z_i^{(0)} - z_j^{(0)}\right| \geq \left|z_i^{(0)} - r_j\right| - \left|z_j^{(0)} - r_j\right| > \left(d - \frac{d}{2n-1}\right) - \frac{1}{2n-1}$$

odakle je

$$\left|z_i^{(0)} - r_j\right| > \frac{2(n-1)}{q} \quad \text{i} \quad \left|z_i^{(0)} - z_j^{(0)}\right| > \frac{2n-3}{q} \geq \frac{1}{q}.$$

Odredimo sada ocenu za $\left|g_i^{(0)}\right|$. Na osnovu (3.8.24) i prethodnih nejednakosti imamo

$$\left|\hat{g}_i^{(0)}\right| \leq \left|v_i^{(0)}\right| \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{1}{\left|z_i^{(0)} - r_j\right|} < \frac{q}{2}\left|v_i^{(0)}\right| < \frac{1}{2}. \tag{3.8.25}$$

Tada iz (3.8.22), za $k = 0$, sleduje

$$\left|r_j - w_j^{(0)}\right| \leq \frac{\left|\hat{g}_j^{(0)}\right|}{1 - \left|\hat{g}_j^{(0)}\right|}\left|v_j^{(0)}\right| < q\left|v_i^{(0)}\right|^2 < \frac{1}{q}. \tag{3.8.26}$$

Kako je, za $i \neq j$

$$\left|z_i^{(0)} - w_j^{(0)}\right| \geq \left|z_i^{(0)} - r_j\right| - \left|r_j - w_j^{(0)}\right| > \frac{2n-3}{q} \geq \frac{1}{q},$$

korišćenjem prethodnih ocena imamo

$$\frac{\left|r_j - w_j^{(0)}\right|}{\left|z_i^{(0)} - r_j\right|\left|z_i^{(0)} - w_j^{(0)}\right|} < \frac{q^3}{2(n-1)}\left|v_j^{(0)}\right|^2. \tag{3.8.27}$$

Nađimo sada ocene za $\left|v_i^{(1)}\right|$ i $\left|g_i^{(0)}\right|$. Na osnovu (3.8.23) dobijamo

$$\left|v_i^{(1)}\right| \leq \frac{\left|g_i^{(0)}\right|}{1 - \left|g_i^{(0)}\right|}\left|v_i^{(0)}\right| \qquad (i = 1, \ldots, n). \tag{3.8.28}$$

Kako je

$$g_1^{(0)} = \left| \sum_{j=2}^{n} \frac{\left(z_1^{(0)} - r_1\right)\left(r_j - w_j^{(0)}\right)}{\left(z_1^{(0)} - r_j\right)\left(z_1^{(0)} - w_j^{(0)}\right)} \right|,$$

korišćenjem (3.8.24) i (3.8.27), nalazimo da je

$$\left|g_1^{(0)}\right| < \frac{q^3}{2(n-1)} \cdot \frac{n-1}{q^2}\left|v_1^{(0)}\right| < \frac{1}{2},$$

odakle, s obzirom na (3.8.28), zaključujemo da je $\left|v_1^{(1)}\right| < \left|v_1^{(0)}\right| < \dfrac{1}{q}$.

Na osnovu prethodnog razmatranja i nejednakosti

$$\left|z_2^{(0)} - z_1^{(1)}\right| \geq \left|z_2^{(0)} - r_1\right| - \left|v_1^{(1)}\right| > \frac{1}{q},$$

sukcesivno za $i = 2, \ldots, n$, dobijamo sledeće ocene

$$\left|g_i^{(0)}\right| < \left|v_i^{(0)}\right| \left(\frac{q^2}{2(n-1)} \sum_{j=1}^{i-1} \left|v_j^{(1)}\right| + \frac{q^3}{2(n-1)} \sum_{j=i+1}^{n} \left|v_j^{(0)}\right|^2 \right) < \frac{1}{2},$$

$$\left|v_i^{(1)}\right| < \frac{q^2}{n-1} \left|v_i^{(0)}\right|^2 \left(\sum_{j=1}^{i-1} \left|v_j^{(1)}\right| + q \sum_{j=i+1}^{n} \left|v_j^{(0)}\right|^2 \right) < \frac{1}{q}.$$

Matematičkom indukcijom se sada može dokazati da za svako $k = 0, 1, \ldots$ važe nejednakosti

$$\left|v_i^{(k+1)}\right| < \frac{q^2}{n-1} \left|v_i^{(k)}\right|^2 \left(\sum_{j=1}^{i-1} \left|v_j^{(k+1)}\right| + q \sum_{j=i+1}^{n} \left|v_j^{(k)}\right|^2 \right) < \frac{1}{q}.$$

Uvođenjem smene $q\left|v_i^{(k)}\right| = h_i^{(k)}$, poslednja nejednakost postaje

$$h_i^{(k+1)} < \frac{1}{n-1} \left(h_i^{(k)}\right)^2 \left(\sum_{j=1}^{i-1} h_i^{(k+1)} + \sum_{j=i+1}^{n} \left(h_i^{(k)}\right)^2 \right), \tag{3.8.29}$$

gde je $i = 1, \ldots, n$ i $k = 0, 1, \ldots$. Na osnovu uslova (3.8.24) imamo da je $h_i^{(0)} < 1$ $(i = 1, \ldots, n)$. Ako stavimo $h = \max\limits_i h_i^{(0)}$, tada je

$$h_i^{(0)} \leq h < 1 \qquad (i = 1, \ldots, n).$$

Na osnovu (3.8.29) zakjučujemo da je iterativni proces (3.8.21) konvergentan. Stavimo dalje da je

$$h_i^{(k+1)} \leq h^{u_i^{(k+1)}} \quad (i = 1, \ldots, n;\ k = 0, 1, \ldots).$$

Ako definišemo matricu B pomoću $B = 2A$, gde je

$$A = \begin{bmatrix} 1 & 1 & & & \\ & 1 & 1 & & \mathbf{0} \\ & & \ddots & \ddots & \\ & \mathbf{0} & & & 1 & 1 \\ 1 & 1 & 0 & \cdots & 0 & 1 \end{bmatrix},$$

vektori $\boldsymbol{u}^{(k)} = \left[u_1^{(k)} \ \cdots \ u_n^{(k)}\right]^T$ mogu biti sukcesivno izračunati pomoću

$$\boldsymbol{u}^{(k+1)} = B\boldsymbol{u}^{(k)} \quad (k = 0, 1, \ldots),$$

startujući sa $\boldsymbol{u}^{(0)} = [1 \ \cdots \ 1]^T$. Ovo se može dokazati matematičkom indukcijom. Sličnim razmatranjem, u radu [2], pokazuje se da za R-red konvergencije iterativnog procesa (3.8.21) važi

$$O_R((3.8.21), \boldsymbol{r}) \geq \rho(B) = 2\rho(A) \quad \left(\boldsymbol{r} = [r_1 \ \cdots \ r_n]^T\right), \tag{3.8.30}$$

gde su $\rho(A)$ i $\rho(B)$ spektralni radijusi matrica A i B.

Karakteristični polinom matrice A je $f_n(\lambda) = (\lambda - 1)^n - (\lambda - 1) - 1$. Ako stavimo $\lambda = \sigma - 1$ dobijamo da je $\overline{f}_n(\sigma) = f_n(1 + \lambda) = \sigma^n - \sigma - 1$. Kako je $\overline{f}_n(1) = -1 < 0$ i $\overline{f}_n(2) = 2^n - 3 > 0$, zaključujemo da postoji nula σ_n u intervalu $(1, 2)$. Nije teško pokazati da je ovo jedina pozitivna nula polinoma $\overline{f}_n$, i da je $\rho(A) = 1 + \sigma_n$. Najzad, na osnovu (3.8.30) dobijamo

$$O_R((3.8.21), \boldsymbol{r}) \geq 2(1 + \sigma_n),$$

što je trebalo i dokazati.

Tabela 3.8.1.

n	(3.8.12)	(3.8.14)	(3.8.19)	(3.8.21)
3	2.325	3.148	3.521	4.649
4	2.221	3.066	3.353	4.441
5	2.167	3.032	3.267	4.335
6	2.135	3.016	3.215	4.269
7	2.113	3.008	3.180	4.226
8	2.097	3.004	3.154	4.194
9	2.085	3.002	3.135	4.170
10	2.076	3.001	3.125	4.152

Za iterativne procese (3.8.12), (3.8.14), (3.8.19), (3.8.21) kod kojih je R-red konvergencije zavisan od n, u tabeli 3.8.1 dajemo vrednosti za $r(n)$, kada je $n = 3, 4, \ldots, 10$.

Primetimo da $r(n)$ teži redu konvergencije osnovnog (ne ubrzanog) metoda, kada $n \to +\infty$.

U radu [97] izvedena je familija metoda za simultano određivanje nula polinoma, korišćenjem razvoja racionalne funkcije u parcijalne razlomke. Svi metodi ove familije imaju kubnu konvergenciju za proste nule.

U više radova (videti, na primer, [27, 82, 96, 120]) razmatrani su i simultani metodi za određivanje višestrukih nula polinoma.

Literatura

1. O. ABERTH, *Iteration methods dor finding all zeros of a polynomial simultaneously*, Math. Comp. **27** (1973), 339–344.
2. G. ALEFELD, J. HERZBERGER, *On the convergence speed of some algorithms for the simultaneous approximation of polynomial roots*, SIAM J. Numer. Anal. **11** (1974), 237–243.
3. J. C. ALEXANDER, J. A. YORKE, *The homotopy continuation method: Numerically implementable topological procedures,* Trans. Amer. Math. Soc. **242** (1978). 271–284.
4. L. BAIRSTOW, *Investigations relating to the stability of the aeroplane*, Rep. & Memo. **154**, Advisory Committee for Aeronautics, 1914.
5. F. L. BAUER, K. SAMELSON, *Polynomkerne und Iterationsverfahren*, Math. Z. **67** (1957), 93–98.
6. W. BI, Q. WU, H. REN, *A new family of eight-order iterative methods for solving nonlinear equations*, Appl. Math. Comput. **214** (2009), 236–245.
7. W. BÖRSCH-SUPAN, *A posteriori error bounds for the zeros of polynomials*, Numer. Math. **5** (1963), 380–398.
8. W. BÖRSCH-SUPAN, *Residuenabschätzung für Polynom-Nullstellen mittels Lagrange-Interpolation*, Numer. Math. **14** (1970), 287–296.
9. K. W. BRODLIE, *On Bairstow's method for the solution of polynomial equations*, Math. Comp. **29** (1975), 816–826.
10. C. G. BROYDEN, *Quasi-Newton methods and their application to function minimisation*, Math. Comp. **21** (1967), 368–381.
11. W. CHENEY, *Analysis for Applied Mathematics*, Graduate Texts in Mathematics, Springer Verlag, New York, 2001.
12. S. P. CHUNG, *Generaliztion and acceleration of an algorithm of Sebastião e Silva and its duals*, Numer. Math. **25** (1976), 365–377.
13. L. COLLATZ, *Functionalanalysis und Numerische Mathematik*, Springer Verlag, Berlin – Heidelberg – New York, 168.
14. C. CARSTENSEN, *On Grau's method for simultaneous factorization of polynomials,* SIAM J. Numer. Anal. **29** (1992), 601–613.
15. C. CARSTENSEN, T. SAKURAI, *Simultaneous factorization of a polynomial by rational approximation,* J. Comput. Appl. Math. **61** (1995), 165–178.
16. J. C. DAUBISSE, *Sur une méthode de résolution numérique d'équations algebriques enzparticulier dans le cas de racines multiples*, Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fiz. N-498 – N-541 (1975), 163–166.
17. K. DOČEV, *A modified Newton method for the simultaneous approximation of all roots of the given algebraic equation*, Fiz.- Mat. Spis. Bulgar. Akad. Nauk **5** (1962), 136–139 (na bugarskom).
18. J. DVORČUK, *Factorization of a polynomial into quadratic factors by Newton method*, Apl. Mat. **14** (1969), 54–80.
19. L. N. ĐORĐEVIĆ, *An iterative solution of algebraic equations with a parameter to accelerate convergence,* Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fiz. No 412 – No 460 (1973), 179–182.
20. L. N. ĐORĐEVIĆ, G. V. MILOVANOVIĆ, *A combined iterative formula for solving equations*, Informatica **78**, Bled 1978, 3(207).
21. J. DŽUNIĆ, M. S. PETKOVIĆ, LJ. D. PETKOVIĆ, *A family of optimal three-point methods for solving nonlinear equations using two parametric functions*, Appl. Math. Comput. **217** (2011), 7612–7619.
22. L. W. EHRLICH, *A modified Newton method for polynomials*, Comm. ACM **10** (1967), 107–108.

23. C. B. GARCIA, F. J. GOULD, *Relations between several path-following algorithms and local and global Newton methods*, SIAM Rev. **22** (1980), 263–274.
24. C. B. GARCIA, W. I. ZANGWILL, *An approach to homotopy and degree theory,* Math. Oper. Res. **4** (1979), 390–405.
25. C. B. GARCIA, W. I. ZANGWILL, *Finding all solutions to polynomial systems and other systems of equations*, Math. Programming **16** (1979), 159–176.
26. C. B. GARCIA, W. I. ZANGWILL, *Pathways to Solutions, Fixed Points and Equilibria*, Prentice Hall, Inc., Englewood Cliffs, New Jersy, 1981.
27. I. GARGANTINI, *Parallel square-root iterations for multiple roots*, Comput. Math. Appl. **6** (1980), 279–288.
28. M. K. GAVURIN, *Nonlinear functional equations and continuous analogues of iteration methods,* Izv. Vyssh. Uchebn. Zaved. Mat. **5** (1958), 18–30 (na ruskom).
29. A. A. GRAU, *The simultaneous Newton improvement of a complete set of approximate factors of a polynomial*, SIAM J. Numer. Anal. **8** (1971), 425–438.
30. E. HALLEY, *A new, exact, and easy method of finding the roots of any equations generally, and that without any previous reduction,* Philos. Trans. Roy. Soc. London **18** (1694), 136–148 (na latinskom).
31. E. HANSEN, M. PATRICK, *A family of root finding methods,* Numer. Math. **27** (1977), 257–269.
32. A. S. HAUSEHOLDER, *Principles of Numerical Analysis,* Dover, New York, 1953.
33. A S. HOUSEHOLDER, *Dandelin, Lobačevskiĭ, or Graeffe?*, Amer. Math. Monthly **66** (1959), 464–466.
34. A S. HOUSEHOLDER, *Generalizations of an algorithm of Sebastião e Silva*, Numer. Math. **16** (1970/71), 375–382.
35. A S. HOUSEHOLDER, *Postscript to: "Generalizations of an algorithm of Sebastião e Silva"*, Numer. Math. **20** (1972/73), 205–207.
36. A.S. HOUSEHOLDER, G.W. STEWART, *Comments on "Some iterations for factoring polynomials"*, Numer. Math. **13** (1969), 470–471.
37. A.S. HOUSEHOLDER, G.W. STEWART, *The numerical factorization of a polynomial*, SIAM Rev. **13** (1971), 38–46.
38. P. HENRICI, *Uniformly convergent algorithms for the simultaneous determination of all zeros of a polynomial*, Studies in Numer. Anal. 1969, pp. 1–8.
39. P. HENRICI, *Circular arithmetic and the determination of polynomial zeros*, Conference on Application of Numerical Analysis, In: Lect. Notes Math. 228, Springer-Verlag, Berlin-Heidelberg-New York, 1971, pp. 86–92.
40. P. HENRICI, *Applied and Computational Complex Analysis, Vol.I,* Wiley-Interscience, New York, 1974.
41. J. HERZBERGER, *Über ein Verfahren zur Bestimmung reealer Nullstellen mit Anwendung auf Parallelrechnung*, Elektron. Rechenanal. **14** (1972), 250–254.
42. J. HERZBERGER, *Bemerkungen zu einem Verfahren von R.E. Moore*, Z. Angew. Math. Mech. **53** (1973), 356–358.
43. L. N. HROMOVA, *Ob odnoĭ modifikaciĭ metoda kasatel'nyh giperbol*, Taškent. Ord. Trud. Kras. Znam. Gos. Univ. im V.I. Lenina **476** (1975), 35–41.
44. P. JARRATT, *Some fourth order multipoint iterative methods for solving equations*, Math. Comp. **20** (1966), 434–437. 398–400.
45. P. JARRATT, *Some efficient fourth order multipoint methods for solving equations*, BIT **9** (1969), 119–124.
46. M. A. JENKINS, J. F. TRAUB, *A three-stage algorithm for real polynomials using quadratic iteration*, SIAM J. Numer. Anal. **7** (1970), 545–566.
47. M. A. JENKINS, J. F. TRAUB, *A three-stage variable-shift iteration for polynomial zeros in relation to generalized Rayleigh iteration*, Numer. Math. **14** (1970), 252–263.

48. B. JOVANOVIĆ, *A method for obtaining iterative formulae of higher order*, Mat. Vesnik **9** (24) (1972), 365–369.
49. L. V. KANTOROVIČ, *Functional Analysis and Applied Mathematics*, Uspehi Matem. Nauk (N.S.) **3** (1948), 89–185 (na ruskom).
50. L. V. KANTOROVICH, G. P. AKILOV, *Functional Analysis*, Pergamon Press, Oxford–Elmsford, N.Y., 1982 (originalno izdanje na ruskom: Nauka, Moskva, 1977).
51. C. T. KELLEY, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, 1995.
52. D. KINCAID, W. CHENEY, *Numerical Analysis*, Brooks/Cole, Pacific Grove, CA, 1991.
53. R. R. KING, *A family of fourth-order methods for nonlinear equations*, SIAM J. Numer. Anal. **10** (1973), 876–879.
54. M. A. KRASNOSEL'SKIĬ, G. M. VAĬNIKKO, P. P. ZABREĬKO, JA. B. RUTICKIĬ, V. JA. STECENKO, *Približennoe rešenie operatornyh uravneniĭ*, Nauka, Moskva, 1969.
55. O. JU. KUL'ČICKIĬ, L. I. ŠIMELEVIČ, *The problem of finding the initial approximation for Newton's method*, Zh. Vychisl. Mat. i Mat. Fiz. **14** (1974), 1016–1018 (na ruskom).
56. H. T. KUNG, J. F. TRAUB, *Optimal order of one-point and multipoint iteration*, J. Assoc. Comput. Math. **21** (1974), 643–651.
57. D. K. LIKA, *Ob iteracionnyh metodah vysšego porjadka*, Tezisi Dokl. 2-ĭ Nauč.- tehn. respubl. konf. Moldaviĭ. Kišinev, 1965, 13–16.
58. D. K. LIKA, *Certain modifications of iteration processes for solution of nonlinear functional equations,* 1965 Problems Theory Control Systems., pp. 38–49, Akad. Nauk Moldav. SSR, Kishinev (na ruskom).
59. D. K. LIKA, *An iteration process for nonlinear functional equations*, 1965 Studies in Algebra and Math. Anal., pp. 134–139 Izdat. "Karta Moldovenjaske", Kishinev (na ruskom).
60. SHIH-NGE LIN, *A method of successive approximations of evaluating the real and complex roots of cubic and higher-order equations*, J. Math. Phys. Mass. Inst. Tech. **20** (1941), 231–242.
61. SHIH-NGE LIN, *A method for finding roots of algebraic equations*, J. Math. Phys. Mass. Inst. Tech. **22** (1943), 60–77.
62. G. LOIZOU, *Une Note sur le procédé itératif de M^{me} Marica D. Prešić*, C. R. Acad. Sci. Paris, **295** (1982), 707–710.
63. G. LOIZOU, *Higher-order iteration functions for simultaneously approximating polynomial zeros*, Internat. J. Comput. Math. **14** (1983), 46–58.
64. H. J. MAEHLY, *Zur iterativen Auflösung algebraischer Gleichungen*, Z. Angew. Math. Phys. **5** (1954), 260–263.
65. G. MASTROIANNI, G. V. MILOVANOVIĆ, *Interpolation Processes – Basic Theory and Applications*. Springer Verlag, Berlin – Heidelberg – New York, 2008.
66. J. M. MCNAMEE, V. Y. PAN, *Efficient polynomial root-refiners: A survey and new record efficiency estimates* Appl. Math. Comp. **63** (2012), 239–354.
67. M. MIGNOTTE, *Note sur la méthode Bernoulli*, Numer. Math. 325–326.
68. G. V. MILOVANOVIĆ, *Contribution and influence of S.B. Prešic to numerical factorization of polynomials*, In: A Tribute to S. B. Prešic (Papers Celebrating his 65th Birthday), pp. 47–56, Mathematics Institut SANU, Belgrade, 2001.
69. G. V. MILOVANOVIĆ, *Ekstremalni problemi i nejednakosti sa polinomima,* Zavod za udžbenike, Beograd, 2012.
70. G. V. MILOVANOVIĆ, *A method to accelerate iterative processes in Banach space,* Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fiz. No 461–No 497 (1974), 67–71.
71. G. V. MILOVANOVIĆ, *Numerička analiza, I deo,* Naučna knjiga, Beograd, 1985.
72. G. V. MILOVANOVIĆ, *Numerička analiza, II deo,* Naučna knjiga, Beograd, 1991.
73. G. V. MILOVANOVIĆ, *Numerička analiza, III deo,* Naučna knjiga, Beograd, 1991.

74. G. V. MILOVANOVIĆ, *Contribution and influence of S.B. Prešic to numerical factorization of polynomials*, In: A Tribute to S. B. Prešic (Papers Celebrating his 65th Birthday), pp. 47 – 56, Mathematical Institut SANU, Belgrade, 2001.
75. G. V. MILOVANOVIĆ, A. S. CVETKOVIĆ, *A note on three-step iterative methods for nonlinear equations,* Stud. Univ. Babeş-Bolyai Math. **52** (2007), 137–146.
76. G. V.MILOVANOVIĆ, L. N.ĐORĐEVIĆ, *On some iterative formulas of the third order,* Informatica 78, Bled 1978, 3(202).
77. G. V. MILOVANOVIĆ, R. Ž. ĐORĐEVIĆ, *Linearna algebra,* Elektronski fakultet u Nišu, Niš, 2004.
78. G. V. MILOVANOVIĆ, R. Ž. ĐORĐEVIĆ, *Matematička analiza I,* Elektronski fakultet u Nišu, Niš, 2005.
79. G. V. MILOVANOVIĆ, M. A. KOVAČEVIĆ, *Dva iterativna procesa trećeg reda u kojima se ne pojavljuju izvodi,* Zbornik radova sa IV znanstvenog skupa PPPR, Stubičke Toplice 1982, 473–478.
80. G. V. MILOVANOVIĆ, M. S. PETKOVIĆ, *On some modifications of a third order method for solving equations,* Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fiz. No 678–No 715 (1980), 63–67.
81. G. V. MILOVANOVIĆ, M. S. PETKOVIĆ, *On the convergence order of a modified method for simultaneous finding polynomial zeros,* Computing **30** (1983), 171–178.
82. G. V. MILOVANOVIĆ, M. S. PETKOVIĆ, *Metodi visokog reda za simultano određivanje višestrukih nula polinoma,* Zbornik radova sa V znanstvenog skupa PPPR, Stubičke Toplice 1983, 95–100.
83. G. V. MILOVANOVIĆ, P. S. STANIMIROVIĆ, *Simbolička implementacija nelinearne optimizacije*, Elektronski fakultet u Nišu, Niš, 2002.
84. G.V. MILOVANOVIĆ, D.S. MITRINOVIĆ, TH.M. RASSIAS, *Topics in Polynomials: Extremal Problems, Inequalities, Zeros,* World Scientific Publ. Co., Singapore – New Jersey – London – Hong Kong, 1994.
85. N. A. MIR, T. ZAMAN, *Some quadrature based three-step iterative methods for non-linear equations,* Appl. Math. Comput. **193** (2007), 366–373.
86. D. S. MITROVIĆ, D. Ž. ĐOKOVIĆ, *Polinomi i matrice,* ICS, Beograd, 1975.
87. I. P. MYSOVSKIH, *On the convergence of L. V. Kantorovič's method of solution of functional equations and its applications,* Dokl. Akad. Nauk SSSR (N.S.) **70** (1950), 565–568 (na ruskom).
88. A. W. NOUREIN, *An iteration formula for simultaneous determination of the zeroes of a polynominal,* J. Comput. Appl. Math. **1** (1975), 251–254.
89. B. NETA, M. S. PETKOVIĆ, *Construction of optimal order nonlinear solvers using inverse interpolation*, Appl. Math. Comput. **217** (2010), 2448–2455.
90. A. W. NOUREIN, *An improvement on iteration formula for simultaneous determination of the zeroes of a polynomial,* Internat. J. Comput. Math. **3** (1977), 109–112.
91. A. W. NOUREIN, *An improvement of Nourein's method for the simultaneous determination of the zeroes of a polynomial,* J. Comput. Appl. Math. **2** (1977), 241–252.
92. J. M. ORTEGA, W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, SIAM, Philadelphia, 2000.
93. A. OSTROWSKI, *Solution of Equations and Systems of Equations*, Academic Press, New York, 1966.
94. V. Y. PAN, *Solving a polynomial equation: Some history and recent progress*, SIAM Rev. **39** (1997), 187–220.
95. M. S. PETKOVIĆ, *Some interval methods of the second order for the simultaneous approximation of polynomial roots,* Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fiz. N-634 – N-677 (1979), 74–81.

96. M. S. PETKOVIĆ, *Generalized root iterations for the simultaneous determination of multiple complex zeros*, Z. Angew. Math. Mech. **62** (1982), 37–55.
97. M. S. PETKOVIĆ, *A family of simultaneous method for the determination of polynomial complex zeros*, Internat. J. Comput. Math. **11** (1982), 285–296.
98. M. S. PETKOVIĆ, *On a general class of multipoint root-finding methods of high computational efficiency*, SIAM. J. Numer. Anal. **47** (2010), 4402–4414.
99. M. S. PETKOVIĆ, G. V. MILOVANOVIĆ, *A note on some improvements of the simultaneous method for determination of polynomial zeros*, J. Comput. Appl. Math. **9** (1983), 65–69.
100. M. S. PETKOVIĆ, LJ. D. PETKOVIĆ, *On a method of two-sided approaching for solving equations*, Freiburger Intervall-Berichte **10** (1980), 1–10.
101. M. S. PETKOVIĆ, LJ. D. PETKOVIĆ, *Families of optimal multipoint methods for solving nonlinear equations: a survey*, Appl. Anal. Discrete Math. **4** (2010), 1–22.
102. M. S. PETKOVIĆ, J. DŽUNIĆ, LJ. D. PETKOVIĆ, *A family of two-point methods with memory for solving nonlinear equations*, Appl. Anal. Discrete Math. **5** (2011), 298–317.
103. M. S. PETKOVIĆ, S. ILIĆ, J. DŽUNIĆ, *Derivative free two-point methods with and without memory for solving nonlinear equations*, Appl. Math. Comp. **217** (2010), 1887–1895.
104. M. S. PETKOVIĆ, LJ.D. PETKOVIĆ, J. DŽUNIĆ, A class of three-point root-solvers of optimal order of convergence, Appl. Math. Comput. **216** (2010), 671–676.
105. J. J. PETRIĆ, S. B. PREŠIĆ, *An algorithm for the solution of* 2×2 *system of nonlinear algebraic equations*, Publ. Inst. Math. (Beograd) (N.S.) **12** (26) (1971), 85–94.
106. J. PETRIĆ, M. JOVANOVIĆ, S. STAMATOVIĆ, *Algoritam for simultaneous determination of all roots of algebraic polynomial equations*, Mat. Vesnik **9(24)** (1972), 325–332.
107. M. D. PREŠIĆ, *Un procédé itératif pour déterminer k zéros d'un polynome*, C. R. Acad. Sci. Paris. **273** (1971), 446–449.
108. M. D. PREŠIĆ, *Jedan iterativni postupak za jednovremeno određivanje k realnih rešenja jednačine na polju realnih brojeva*, Mat. Vesnik **10 (25)** (1973), 299–308.
109. M. D. PREŠIĆ, *A convergence theorem for a method for simultaneous determination of all zeros of a polynomial*, Publ. Inst. Math. **28(42)** (1980), 158–168.
110. S. B. PREŠIĆ, *Procédé itératif pour la factorisation des polynomes*, C. R. Acad. Sci. Paris. **262** (1966), 862–863.
111. S. B. PREŠIĆ, *Jedan iterativni postupak za faktorizaciju polinoma*, Mat. Vesnik **5 (20)** (1968), 205–216.
112. A. V. PROKOPČENKO, *Iteration processes of higher orders*, Zh. Vychisl. Mat. i Mat. Fiz. **14** (1974), 230–233 (na ruskom).
113. L. RALL, *Computational solution of nonlinear operator equations,* New York, 1969.
114. A. RALSTON, PH. RABINOWITZ, *A First Course in Numerical Analysis*, Second edition, Dover Publications, Inc., Mineola, New York, 2001.
115. G. S. SALEHOV, *On the convergence of the process of tangent hyperbolas,* Dokl. Akad. Nauk SSSR (N.S.) **82** (1952), 525–528 (na ruskom).
116. K. SAMELSON, *Factorisierung von Polynomes durch funktionale Iteration*, Bayer. Akad. Wiss. Math. Natur. Kl. Abh. **95** (1959), 1–26.
117. J. SCHRÖDER, *Factorization of polynomials by generalized Newton procedures*, In: Constructive Aspects of the Fundamental Theorem of Algebra (B. Dejon and P. Henrici, eds.), John Wiley, New York, 1969, 295–320.
118. E. SCHRÖDER, *On infinitely many algorithms for solving equations,* Math. Ann. **2** (1870), 317–365 (na nemačkom).
119. J. SEBASTIÃO E SILVA, *Sur une méthode d'approximation semblable à celle de Gräffe* Portugal. Math. **2** (1941), 271–279.
120. H. SEMERDZIEV, *A method for simultaneous finding all roots of algebraic equations with given multiplicity*, C. D. Acad. Bulgare Sci. **35** (1982), 1057–1060 (na bugarskom).

121. G. W. STEWART *Some iterations for factoring a polynomial*, Numer. Math. **13** (1969), 458–470.
122. G. W. STEWART *On Samelson's iteration for factoring polynomials*, Numer. Math. **15** (1970), 306–314.
123. G. W. STEWART *On the convergence of Sebastião e Silva's method for finding a zero of a polynomial*, SIAM Rev. **12** (1970), 458–460.
124. J. STOER, *Einführung in die Numerische Mathematik I,* Springer Verlag, Berlin – Heidelberg – New York, 1972.
125. O. N. TIHONOV, *O bystrom vyčislenij naibol'ših kornej mnogočlena,* Zap. Leningrad. gorn. in-ta 48, 3 (1968), 36–41.
126. O. N. TIHONOV, *A generalization of Newton's method for computing the roots of algebraic polynomials,* Izv. Vyssh. Uchebn. Zaved. Mat. **1976**, no. 6 (169), 122–124 (na ruskom).
127. D. TOŠIĆ, *Uvod u numeričku analizu,* Naučna knjiga, Beograd.
128. D. Đ. TOŠIĆ, G. V. MILOVANOVIĆ, *An application of Newton's method to simultaneous determination of zeros of a polynomial,* Univ. Beograd. Publ. Elektrotehn. Fak. Ser. Mat. Fiz. No 412 – No 460 (1973), 175–177.
129. J. F. TRAUB, *Iterative Methods for the Solution of Equations,* Prentice-Hall, New Jersey, 1964.
130. R. THUKRAL, M.S. PETKOVIĆ, *Family of three-point methods of optimal order for solving nonlinear equations*, J. Comput. Appl. Math. **233** (2010), 2278–2284.
131. V. A. VARJUHIN, S. A. KAS'JANJUK, *Iteration methods for sharpening the roots of equations,* Zh. Vychisl. Mat. i Mat. Fiz. **9** (1969), 684–687.
132. V. A. VARJUHIN, S. A. KAS'JANJUK, *The iteration methods of the solution of nonlinear systems,* Zh. Vychisl. Mat. i Mat. Fiz. **10** (1970), 1533–1536.
133. V. A. VARJUHIN, S. A. KAS'JANJUK, *IA class of iterative procedures for the solution of systems of nonlinear equations,* Zh. Vychisl. Mat. i Mat. Fiz. **17** (1977), 1123–1131.
134. V. M. VERBUK, D. I. MIL'MAN, *The Wegstein method as a modification of the method of secants,* Zh. Vychisl. Mat. i Mat. Fiz. **17** (1977), 507–508 (na ruskom).
135. B. A. VERTGEIM, *On certain methods of approximate solutions of non-linear functional equations in Banach spaces,* Uspekhi Mat. Nauk. **12** (1957), 166–169 (na ruskom).
136. X. WANG, L. LIU, *New eighth-order iterative methods for solving nonlinear equations*, J. Comput. Appl. Math. **234** (2010), 1611–1620.
137. J. H. WEGSTEIN, *Accelerating convergence of iterative processes,* Comm. ACM **1** (1958), 9–13.
138. K. WEIERSTRASS, *Neuer Beweis des Satzes, dass jede Ganze Rationale Function Einer Veranderlichen dargestellt werden kann als ein Product aus Linearen Functionen darstelben Veranderlichen,* Ges. Werke **3** (1903), 251–296.
139. W. WERNER, *Iterative solution of nonlinear systems of equations,* In: Lect. Notes Math. 953, Springer-Verlag, Berlin – Heidelberg – New York, pp. 188–202.
140. J. H. WILKINSON, *The evaluation of zeros of ill-conditioned polynomials,* Numer. Math. **1** (1959), 150–166.
141. D. M. YOUNG, R. T. GREGORY, *A Survey of Numerical Mathematics,* Addison–Wesly Publ. Co., Reading, Massachusetts, 1972.
142. V. L. ZAGUSKIN, *Handbook of Numerical Methods for the Solution of Algebraic and Transcendental Equations,* Pergamon Press, New York – London –Paris, 1961.

Indeks

GRADIMIR V. MILOVANOVIĆ

NUMERIČKA ANALIZA I TEORIJA APROKSIMACIJA

UVOD U NUMERIČKE PROCESE I REŠAVANJE JEDNAČINA

Biblioteka KARAMATA

Prvo izdanje 2014. godine.

Izdavač
ZAVOD ZA UDŽBENIKE
Obilićev venac 5, Beograd
www.zavod.co.rs

Likovni urednik
mr TIJANA RANČIĆ

Grafički urednik
ALEKSANDAR RADOVANOVIĆ

Prelom teksta i korektura
GRADIMIR V. MILOVANOVIĆ

Tiraž 700

Obim: 29,5 štamparskih tabaka
Format: 165×235 mm

Rukopis predat u štampu avgusta 2014. godine.
Štampanje završeno avgusta 2014. godine.
Štampa: Službeni glasnik, Beograd