UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO
ODDELEK ZA MATEMATIKO

Andrej Perne

KONSTRUKCIJA SPEKTRALNIH METOD Z NEPERIODIČNIMI TRIGONOMETRIČNIMI VRSTAMI

Doktorska disertacija

MENTOR: prof. dr. Bojan Orel

Ljubljana, 2012

Kratek povzetek vsebine

Ortogonalni polinomi so poleg Fourierove vrste eno izmed orodij, ki se v teoriji aproksimacije najpogosteje uporablja. Posebne lastnosti trigonometričnih funkcij in polinomov zagotavljajo učinkovito računanje ter stabilne in konvergentne numerične rešitve. Spektralne metode so poleg metod končnih razlik in končnih elementov pomembno orodje za reševanje robnih problemov tako pri navadnih kot pri parcialnih diferencialnih enačbah. V prvem delu doktorske disertacije so opisane osnovne lastnosti Fourierove vrste in ortogonalnih polinomov ter nekateri osnovni pristopi za konstrukcijo spektralnih metod z osnovnimi orodji za analizo konvergence in napake.

V nadaljevanju sta predstavljeni dve neklasični družini ortogonalnih polinomov, tj. poldomenski polinomi Čebiševa prve in druge vrste ter pripadajoča poldomenska Čebišev-Fourierova vrsta. Obe družini sta konstruirani z uporabo modificiranega algoritma Čebiševa za izračun rekurzivnih koeficientov v tričlenski rekurzivni formuli. Aproksimacija s kvadratom integrabilnih funkcij s poldomensko Čebišev-Fourierovo vrsto vrne primerljive rezultate kot aproksimacija s Fourierovo vrsto ali z vrsto Čebiševa.

V osrednjem delu doktorske disertacije je konstruiran nov razred Čebišev-Fourierovih kolokacijskih spektralnih metod za reševanje linearnih dvotočkovnih robnih problemov z Dirichletovimi robnimi pogoji, kjer numerično rešitev problema iščemo v obliki odrezane poldomenske Čebišev-Fourierove vrste, spektralne koeficiente pa izračunamo z metodo kolokacije. Analiza konvergence in napake pokaže, da so te metode primerljive s standardnimi, kjer iščemo rešitev v obliki Fourierove vrste za periodične ali v obliki vrste Čebiševa za neperiodične probleme. Nov razred metod konstruiramo tudi za nekatere evolucijske robne probleme, tj. za posplošene toplotne in valovne enačbe.

Numerični zgledi potrjujejo teoretične rezultate in prikazujejo primerljivost napake numerične rešitve dobljene z novimi ali s standardnimi metodami. Kljub temu pa je računska zahtevnost neprimerljiva, saj v primeru poldomenske Čebišev-Fourierove vrste ni na voljo orodja za izračun koeficientov, ki bi bilo primerljivo s hitro Fourierovo transformacijo.

Math. Subj. Class. (2010): 65L10, 65L60, 65L20, 65L70, 65T40, 65N35

Ključne besede: *spektralne metode*, *ortogonalni polinomi*, *dvotočkovni robni problemi*, *kolokacija*, *aproksimacija*, *poldomenski polinomi Čebiševa prve in druge vrste*, *poldomenska Čebišev-Fourierova vrsta*, *posplošena toplotna enačba*, *posplošena valovna enačba*

Abstract

Orthogonal polynomials are, along with Fourier series, one of the most widely used tools in the theory of approximation. Specific properties of trigonometric functions and polynomials assure efficient computation as well as stable and convergent numerical solutions. Spectral methods are, besides finite difference and finite element methods, an important tool for solving boundary value problems for ordinary as well as partial differential equations. In the first part of the doctoral thesis, some basic properties of the Fourier series and orthogonal polynomials as well as some basic approaches for the construction of spectral methods with fundamental tools for convergence and error analysis are described.

In the sequel, two non-classical families of orthogonal polynomials are presented, i.e., the half-range Chebyshev polynomials of the first and second kind as well as the corresponding half-range Chebyshev-Fourier series. Both families are constructed via the modified Chebyshev algorithm used for the computation of the recursive coefficients for the three-term recurrence relation. The approximation of square integrable functions with half-range Chebyshev-Fourier series yields comparable results to the approximation with Fourier or Chebyshev series.

In the central part of the doctoral thesis, a new class of Chebyshev-Fourier collocation spectral methods for solving linear two-point boundary value problems with Dirichlet boundary conditions is constructed. We seek for the numerical solution in the form of the truncated half-range Chebyshev-Fourier series, where spectral coefficients are computed using the collocation method. Convergence and error analysis shows that these methods are comparable with standard ones, where the solution is approximated with the Fourier series for periodic or with the Chebyshev series for non-periodic problems. We construct a new class of methods also for some evolutive boundary value problems, i.e., for generalized heat and wave equations.

Numerical examples confirm theoretical results and show the comparability of the error of the numerical solution obtained with the new or the standard methods. Yet, computational costs are not comparable, because in the case of half-range Chebyshev-Fourier series there does not exist a tool for the computation of coefficients being comparable with fast Fourier transform.

Math. Subj. Class. (2010): 65L10, 65L60, 65L20, 65L70, 65T40, 65N35

Key words: *spectral methods*, *orthogonal polynomials*, *two-point boundary value problems*, *collocation*, *approximation*, *half-range Chebyshev polynomials of the first and second kind*, *half-range Chebyshev-Fourier series*, *generalized heat equation*, *generalized wave equation*

Kazalo

Slike

Poglavje 1

Uvod

Eden izmed standardnih problemov v numerični matematiki je aproksimacija dane funkcije v nekem končnorazsežnem podprostoru izbranega funkcijskega prostora. V nadaljevanju naj bo to prostor s kvadratom integrabilnih funkcij na intervalu $[-1, 1]$ glede na utež w, ki ga označimo z $L^2_w(-1, 1)$.

Problem 1.1 *Naj bo* $\{\phi_0, \phi_1, \ldots, \phi_N\}$*, kjer je* $N \in \mathbb{N}_0$*, množica baznih funkcij podprostora* $X_N \subset L^2_w(-1, 1)$ *in naj bo* $f \in L^2_w(-1, 1)$*. Tedaj iščemo tako funkcijo* $f_N \in X_N$ *oblike*

$$f_N(x) = \sum_{k=0}^{N} f_k \phi_k(x), \quad x \in [-1, 1], \tag{1.1}$$

da bo L^2_w *norma razlike* $\|f - f_N\|_{L^2_w}$ *čim manjša.*

Standardna izbora baznih funkcij sta množici trigonometričnih funkcij za aproksimacijo periodičnih ter ortogonalnih polinomov za aproksimacijo neperiodičnih funkcij. V prvem primeru je X_N prostor trigonometričnih polinomov stopnje kvečjemu N, funkcija f_N pa običajno odrezana Fourierova vrsta, v drugem primeru pa je X_N prostor polinomov stopnje kvečjemu N, funkcija f_N pa polinom stopnje kvečjemu N, ki je v mnogih primerih razvit po polinomih Čebiševa prve vrste v odrezano vrsto Čebiševa. Oba izbora privedeta do rešitve problema 1.1, kar je podrobno opisano v različnih knjigah s tega področja, npr. v K. Atkinson in W. Han [6], J. P. Boyd [7], C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang [11], D. Gottlieb in S. A. Orszag [28] ter L. N. Trefethen [60].

Za rešitev problema 1.1 je potrebno izračunati koeficiente f_k v razvoju (1.1) dane funkcije po baznih funkcijah. V primeru koeficientov Fourierove vrste, ki jim krajše pravimo Fourierovi koeficienti, lahko to učinkovito izvedemo z uporabo hitre Fourierove transformacije (FFT), ki je opisana v članku J. W. Cooley in J. W. Tukey [15]. To je stabilna in dobro poznana metoda, ki za gladke oz. analitične periodične funkcije izredno hitro konvergira, napaka pa pada eksponentno glede na naraščajočo vrednost N. Pri

funkcijah, ki niso dovolj gladke ali niso periodične, pa z uporabo Fourierove vrste nastopijo težave, ki se kažejo kot oscilacije v okolici točk nezveznosti. Vzrok temu je *Gibbsov fenomen* (ang. *Gibbs phenomenon*), ki je opisan npr. v J. W. Gibbs [26] ali C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang [11]. Poleg tega Fourierovi koeficienti za neperiodične funkcije padajo bistveno počasneje kot za periodične funkcije.

Obstaja več možnosti za odpravo teh težav, ki so opisane npr. v D. Gottlieb in C. W. Shu [29] ter E. Tadmor [55]. Ena izmed možnosti je, da uporabimo neko transformacijo, ki dano funkcijo prevede na periodično, in izračunamo Fourierove koeficiente za tako transformirano funkcijo. Najpogosteje se uporablja transformacija, ki vodi do polinomov Čebiševa prve vrste in je opisana npr. v J. P. Boyd [7], B. Fornberg in D. M. Sloan [19] ter L. N. Trefethen [57]. Drug pristop je nedavno predstavil D. Huybrechs v članku [35], kjer je analiziral problem, podan v J. P. Boyd [8] ter O. P. Bruno, Y. Han in M. M. Pohlman [9].

Problem 1.2 *Za $T > 1$, naj bo $\mathcal{G}_N$ prostor $2T$-periodičnih funkcij $g \in \mathcal{G}_N$ oblike*

$$g(x) = \frac{a_0}{2} + \sum_{k=1}^{N} \left(a_k \cos \frac{\pi k x}{T} + b_k \sin \frac{\pi k x}{T} \right). \tag{1.2}$$

Fourierova razširitev funkcije $f \in L^2(-1,1)$, definirane na intervalu $[-1,1]$, na interval $[-T,T]$, je rešitev optimizacijskega problema

$$g_N := \arg \min_{g \in \mathcal{G}_N} \|f - g\|_{L^2}. \tag{1.3}$$

Huybrechs je v [35] obravnaval problem 1.2 za funkcijo $f \in L^2(-1,1)$, ki ni nujno gladka oz. analitična ali periodična. Glavna ideja za zagotovitev eksponentne natančnosti Fourierove vrste je, da dano funkcijo f razširimo do funkcije g, ki je periodična na večjem intervalu $[-T,T]$, kjer je $T > 1$. Fourierova vrsta tako razširjene funkcije je očitno konvergentna po točkah k funkciji f na osnovnem intervalu $[-1,1]$. Za izbiro $T = 2$ je Huybrechs predlagal več numeričnih metod za reševanje problema 1.2. Poleg dokaza obstoja in enoličnosti je karakteriziral rešitev z dvema neklasičnima družinama ortogonalnih polinomov, ki se imenujeta *poldomenski polinomi Čebiševa prve in druge vrste* (ang. *half-range Chebyshev polynomials of the first and second kind*). Ti polinomi so v nekem smislu sorodni klasičnim polinomom Čebiševa prve in druge vrste, saj imajo enaki uteži kot slednji, le da so ortogonalni na krajšem intervalu. V večini primerov je konvergenca eksponentna.

Ortogonalni polinomi so zelo pomembno orodje, ki se v numerični analizi pa tudi na drugih področjih matematike izredno veliko uporablja. Primeri uporabe so npr. v teoriji aproksimacije, pri konstrukciji numeričnih integracijskih metod Gaussovega tipa ter pri reševanju robnih problemov s spektralnimi metodami. Klasične družine ortogonalnih polinomov (Hermiteovi, Laguerreovi in Jacobijevi polinomi) so podrobno obravnavane v

različnih knjigah s tega področja, npr. v M. Abramowitz in I. A. Stegun [1], T. S. Chihara [13] ter G. Szegö [54]. Kratek pregled lastnosti nekaterih standardnih polinomov Jacobijevega tipa (Legendreovi polinomi ter polinomi Čebiševa prve in druge vrste) je v nadaljevanju.

Naš cilj je konstrukcija spektralnih metod za reševanje robnih problemov, ki bo temeljila na uporabi poldomenskih polinomov Čebiševa prve in druge vrste, ki jih opišemo z uporabo tričlenske rekurzivne formule. V ta namen potrebujemo učinkovit algoritem za izračun rekurzivnih koeficientov. Primerna izbira je uporaba modificiranega algoritma Čebiševa, ki je predstavljen v člankih R. A. Sack in A. F. Donovan [50] ter J. C. Wheeler [61]. Podrobnosti konstrukcije rekurzivnih koeficientov preko modificiranih momentov za preslikane monične Legendreove polinome so opisane v B. Orel in A. Perne [46].

Podobno kot iz trigonometričnih funkcij sestavimo Fourierovo vrsto ali iz polinomov Čebiševa prve vrste vrsto Čebiševa, iz poldomenskih polinomov Čebiševa prve in druge vrste sestavimo *poldomensko Čebišev-Fourierovo (HCF) vrsto* (ang. *half-range Chebyshev-Fourier series*), ki je prav tako primerna za reševanje aproksimacijskega problema 1.1. Obravnava nekaterih potrebnih orodij za konstrukcijo spektralnih metod, npr. odvajanje in množenje HCF vrst, je v članku B. Orel in A. Perne [46].

Robni problemi so, podobno kot problem aproksimacije, zelo pogosti v numerični analizi. Iščemo lahko numerično rešitev robnega problema v eni ali več dimenzijah pri navadnih ali parcialnih diferencialnih enačbah. V tem delu se bomo omejili na nekatere posebne tipe, npr. na linearne dvotočkovne robne probleme v eni dimenziji.

Problem 1.3 *Iščemo rešitev y linearnega dvotočkovnega robnega problema oblike*

$$\mathcal{L}y(x) = f(x), \quad x \in [-1, 1], \tag{1.4}$$

z robnimi pogoji

$$\mathcal{B}y(x) = 0, \quad x \in \{-1, 1\}, \tag{1.5}$$

kjer je $\mathcal{L}$ linearni diferencialni operator

$$\mathcal{L} = \alpha(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \beta(x)\frac{\mathrm{d}}{\mathrm{d}x} + \gamma(x)I, \tag{1.6}$$

I identiteta in $\mathcal{B}$ množica linearnih robnih diferencialnih operatorjev.

Za reševanje robnih problemov drugega ali višjega reda pri navadnih diferencialnih enačbah (ODE) imamo na voljo precej različnih numeričnih metod. Poleg dobro poznanih metod, kot sta metoda končnih razlik (FDM) in metoda končnih elementov (FEM), imamo na voljo spektralne metode (SM), ki jih bomo v tej doktorski disertaciji obširno obravnavali.

Dobro znano dejstvo je, da spektralne metode aproksimirajo točno rešitev v nekem končnorazsežnem podprostoru izbranega Hilbertovega prostora. V nasprotju z metodo končnih elementov oz. končnih razlik, kjer so bazne funkcije definirane lokalno (tj. samo na majhnem intervalu), so bazne funkcije, ki jih uporabljamo pri spektralnih metodah, definirane globalno (tj. na celem intervalu, kjer je definiran obravnavani problem).

Glede na to, ali je dan problem periodičen ali neperiodičen, poznamo različne pristope pri konstrukciji spektralnih metod. V primeru periodičnih problemov je naravno, da za množico baznih funkcij vzamemo trigonometrične funkcije. Z drugimi besedami, numerično rešitev iščemo v obliki odrezane Fourierove vrste. Pri tem za diskretizacijo intervala, na katerem je problem definiran, uporabimo ekvidistantne delilne točke. V primeru neperiodičnih problemov pa za bazne funkcije vzamemo ortogonalne polinome, najpogosteje polinome Čebiševa prve vrste, lahko pa tudi Legendreove polinome. Numerično rešitev iščemo v obliki odrezane vrste Čebiševa.

Poznamo dva tipa delilnih točk Čebiševa. Pri spektralnih metodah za neperiodične probleme običajno za diskretizacijo intervala, na katerem je problem definiran, uporabljamo točke Čebiševa druge vrste. To so točke, kjer polinomi Čebiševa prve vrste dosežejo ekstremne vrednosti. Po drugi strani so točke Čebiševa prve vrste ničle polinomov Čebiševa prve vrste. Običajno z uporabo točk Čebiševa za reševanje neperiodičnih problemov dosežemo boljše rezultate kot z uporabo ekvidistantnih točk, ker so prve gostejše blizu roba območja kot blizu sredine intervala. Takšna porazdelitev točk pomaga pri premagovanju težav, ki jih povzročata tako Gibbsov (glej J. W. Gibbs [26]) kot Rungejev fenomen (glej C. Runge [49]), tj. z uporabo neekvidistantnih točk Čebiševa se izognemo oscilacijam v okolici robov danega intervala. Glede na metodo izračuna spektralnih koeficientov ločimo različne tipe spektralnih metod. Najpogostejše so: Galerkinova metoda, Tau metoda ter metoda kolokacije. Spektralne kolokacijske metode običajno imenujemo psevdospektralne metode.

V literaturi je bilo nekaj poskusov, kako rešiti neperiodične probleme s trigonometričnimi baznimi funkcijami. B. Adcock je v svojih člankih [2] in [3] rešil problem z uporabo modificirane Fourierove vrste, kjer je za izračun spektralnih koeficientov vrste uporabil Galerkinovo metodo. D. Huybrechs pa je v članku [35] predlagal uporabo množice trigonometričnih baznih funkcij, ki vsebuje tako sinusne in kosinusne funkcije kot tudi sinusne in kosinusne funkcije polovičnih kotov. Le-te so združene v poldomenske polinome Čebiševa prve in druge vrste ter organizirane v obliko HCF vrste.

V tem delu bomo numerično rešitev problema 1.3 namesto s Fourierovo vrsto ali vrsto Čebiševa aproksimirali s poldomensko Čebišev-Fourierovo vrsto, kar je podrobno obravnavano v članku B. Orel in A. Perne [45]. Konstrukcijo novega razreda spektralnih metod izvedemo z uporabo orodij, ki so opisana v člankih D. Huybrechs [35] ter B. Orel in A. Perne [46]. Spektralne koeficiente vrste izračunamo z uporabo metode kolokacije. Opisan pristop

vodi h konstrukciji psevdospektralnih metod za reševanje neperiodičnih robnih problemov z uporabo orodij za reševanje periodičnih robnih problemov. Kljub temu, da obravnavamo zgolj Dirichletove robne pogoje, metode ni težko posplošiti na Neumannove ali mešane (Robinove) robne pogoje. Prav tako lahko napravimo posplošitev na linearne robne probleme višjega reda. Omejitev za interval je le stvar poenostavitve, saj je dobro znano, kako poljuben interval $[a, b]$ preslikamo na interval $[-1, 1]$ in obratno.

Poleg linearnih dvotočkovnih robnih problemov v eni ali več (dveh) dimenzijah, nas zanimajo tudi nekateri tipi linearnih evolucijskih robnih problemov pri parcialnih diferencialnih enačbah (PDE). Obravnavali bomo posplošene toplotne enačbe paraboličnega tipa ter posplošene valovne enačbe hiperboličnega tipa.

Problem 1.4 *Iščemo rešitev u posplošene toplotne enačbe oblike*

$$u_t = \alpha(x,t)u_{xx} + \beta(x,t)u_x + \gamma(x,t)u + \delta(x,t), \quad x \in [-1,1], \quad t \geq 0, \tag{1.7}$$

z začetnim pogojem

$$u(x,0) = f(x), \quad x \in [-1,1] \tag{1.8}$$

ter s konsistentnima robnima pogojema

$$u(-1,t) = g(t), \quad u(1,t) = h(t), \quad t \geq 0. \tag{1.9}$$

Problem 1.5 *Iščemo rešitev u posplošene valovne enačbe oblike*

$$u_{tt} = \alpha(x,t)u_{xx} + \beta(x,t)u_x + \gamma(x,t)u + \delta(x,t), \quad x \in [-1,1], \quad t \geq 0, \tag{1.10}$$

z začetnima pogojema

$$u(x,0) = f_1(x), \quad u_t(x,0) = f_2(x), \quad x \in [-1,1] \tag{1.11}$$

ter s konsistentnima robnima pogojema

$$u(-1,t) = g(t), \quad u(1,t) = h(t), \quad t \geq 0. \tag{1.12}$$

Pri obeh evolucijskih problemih 1.4 in 1.5 uporabimo za diskretizacijo po prostorski spremenljivki eno izmed spektralnih metod, za diskretizacijo po časovni spremenljivki pa eno izmed metod za reševanje začetnih problemov za navadne diferencialne enačbe (ODE) oblike

$$u' = f(t,u), \quad u(t_0) = u_0. \tag{1.13}$$

Uporabimo lahko Magnusove metode, metode Runge-Kutta ter številne druge, ki so podrobno opisane npr. v M. Abramowitz in I. A. Stegun [1], W. Gautschi [22], E. Hairer, S. P. Nørsett in G. Wanner [33], E. Hairer,

C. Lubich in G. Wanner [32], E. Isaacson in H. B. Keller [36], A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett in A. Zanna [38] ter A. Iserles [37].

Doktorska disertacija je organizirana na sledeč način. V poglavju 2 obravnavamo ortogonalne sisteme. Definiciji ortogonalnosti v razdelku 2.1 sledi opis in obravnava konvergenčnih lastnosti Fourierove vrste v razdelku 2.2 ter opis klasičnih družin ortogonalnih polinomov s poudarkom na konvergenčnih lastnostih vrste Čebiševa v razdelku 2.3.

Poglavje 3 je posvečeno kratkemu uvodu v teorijo spektralnih metod, kjer v razdelku 3.1 opišemo metodo uteženega residuala, v razdelku 3.2 izbor primernih baznih funkcij, v razdelku 3.3 tri metode za izračun spektralnih koeficientov (Galerkinova metoda, Tau metoda in metoda kolokacije) ter v razdelku 3.4 konstrukcijo dveh razredov spektralnih metod (Fourierove metode in metode Čebiševa) ter Clenshaw-Curtisovo kvadraturno formulo. Razdelek 3.5 predstavi spektralne metode na dveh modelnih linearnih dvotočkovnih robnih problemih, razdelek 3.6 pa je posvečen predstavitvi osnovnih orodij za analizo konvergence in napake pri spektralnih metodah. V poglavju 4 je prikazana konstrukcija spektralnih metod za posplošene toplotne in valovne enačbe v razdelkih 4.1 in 4.2.

Poglavje 5 je posvečeno poldomenskim polinomom Čebiševa. V razdelku 5.1 je predstavljena tričlenska rekurzivna formula, v razdelku 5.2 pa je opisan modificiran algoritem Čebiševa za izračun rekurzivnih koeficientov, kar omogoča konstrukcijo poldomenskih polinomov Čebiševa prve in druge vrste v razdelku 5.3 ter obravnavo njihovih lastnosti v razdelku 5.4.

Problema 1.1 in 1.2 obravnavamo v poglavju 6, kjer v razdelku 6.1 najprej definiramo novo ortonormalno bazo v končnorazsežnem podprostoru prostora $L^2(-1,1)$, nato pa definiramo poldomensko Čebišev-Fourierovo vrsto ter analiziramo njeno konvergenco v razdelku 6.2. Primerjava kvalitete aproksimacij, ki jih dobimo z uporabo različnih vrst (Fourierova vrsta, vrsta Čebiševa, poldomenska Čebišev-Fourierova vrsta), je opisana v razdelku 6.3.

Linearne robne probleme v eni dimenziji obravnavamo v poglavju 7. Numerično rešitev iščemo v obliki poldomenske Čebišev-Fourierove vrste. Razdelek 7.1 je posvečen konstrukciji dveh operatorskih matrik za odvajanje in množenje HCF vrst. V razdelku 7.2 konstruiramo kolokacijsko spektralno metodo za reševanje problema 1.3, v razdelku 7.3 pa analiziramo napako in konvergenco te metode. Poglavje zaključuje razdelek 7.4, kjer je prikazanih nekaj numeričnih zgledov skupaj s primerjavo Čebišev-Fourierove kolokacijske metode (CFC) s kolokacijsko metodo Čebiševa (CC).

V poglavju 8 konstruiramo nov razred spektralnih metod z uporabo poldomenske Čebišev-Fourierove vrste tako za reševanje posplošene toplotne enačbe paraboličnega tipa v razdelku 8.1 kot tudi posplošene valovne enačbe hiperboličnega tipa v razdelku 8.2, tj. problemov 1.4 in 1.5. Oba razdelka vsebujeta nekaj numeričnih zgledov, vendar brez analize napake in konvergence. Poglavje 9 zaključuje doktorsko disertacijo.

Poglavje 2

Ortogonalni sistemi

2.1 Ortogonalnost

Linearne probleme običajno obravnavamo v prostorih s skalarnim produktom. To nam omogoča definicijo norme preko pojma skalarnega produkta ter vpeljavo pojma ortogonalnosti dveh elementov (funkcij). Le-ta je namreč izredno pomemben na različnih področjih numerične analize, npr. pri aproksimaciji in numeričnemu reševanju diferencialnih enačb. Podrobnejši pregled teorije prostorov s skalarnim produktom najdemo npr. v monografiji K. Atkinson in W. Han [6]. Posebnega pomena so polni prostori s skalarnim produktom, ki jim pravimo *Hilbertovi prostori*. Zanimajo nas predvsem prostori funkcij, npr. prostor uteženih s kvadratom integrabilnih funkcij $L^2_w(a,b)$, ki je Hilbertov prostor s skalarnim produktom

$$(f,g) = \int_a^b f(x)\, g(x)\, w(x)\, \mathrm{d}x, \qquad f,g \in L^2_w(a,b). \tag{2.1}$$

Pripadajoča norma je definirana s predpisom

$$\|f\|_{L^2_w} = \sqrt{(f,f)}, \qquad f \in L^2_w(a,b). \tag{2.2}$$

Funkcija w je nenegativna in se imenuje *utež*. Pomembni prostori v analizi numeričnih metod, predvsem pri analizi konvergence in napake Fourierovih spektralnih metod, so *prostori Soboljeva*, npr. $H^m(a,b)$ ali $H^m_p(a,b)$.

Definicija 2.1 Prostor Soboljeva $H^m(a,b)$, $m \in \mathbb{N}_0$, *je prostor funkcij* $f \in L^2(a,b)$, *za katere velja, da vsi šibki odvodi do reda* m *pripadajo prostoru* $L^2(a,b)$*:*

$$H^m(a,b) = \left\{ f \in L^2(a,b) : f^{(k)} \in L^2(a,b), 0 \le k \le m \right\}. \tag{2.3}$$

Prostor Soboljeva $H^m(a,b)$ je Hilbertov prostor glede na skalarni produkt

$$(f,g)_m = \sum_{k=0}^{m} \int_a^b f^{(k)}(x)\, g^{(k)}(x)\, \mathrm{d}x, \tag{2.4}$$

ki je opremljen z normo

$$\|f\|_m = \left(\sum_{k=0}^{m} \|f^{(k)}\|_{L^2}^2 \right)^{1/2}. \tag{2.5}$$

Podrobnosti o prostorih Soboljeva, vključno z dokazom spodnjega izreka, so opisane npr. v monografijah G. Leoni [42] ter W. P. Ziemer [63].

Izrek 2.2 *Prostor $C^\infty(a,b)$ je gost podprostor v $H^m(a,b)$ za vsak $m \in \mathbb{N}_0$.*

V analizi Fourierovih spektralnih metod potrebujemo definicijo prostora Soboljeva za periodične funkcije. Naj bodo le-te 2π-periodične na intervalu $(-\pi, \pi)$. Tedaj za vsak $m \in \mathbb{N}_0$ definiramo

$$H_p^m(-\pi,\pi) = \left\{ f \in H^m(-\pi,\pi) : f^{(k)}(-\pi) = f^{(k)}(\pi), 0 \le k \le m-1 \right\}. \tag{2.6}$$

Izmed mnogo različnih neenakosti, ki veljajo v Hilbertovih prostorih $\mathcal{H}$, omenimo eno pomembnejših, tj. *Cauchy-Schwarzovo neenakost*

$$|(f,g)| \le \|f\| \, \|g\|, \qquad f,g \in \mathcal{H}. \tag{2.7}$$

Enakost velja natanko tedaj, ko sta funkciji f in g linearno odvisni. Pravimo, da sta funkciji $f, g \in \mathcal{H}$ *ortogonalni*, če je $(f,g) = 0$. Množica funkcij $\{f_k\}_{k\ge 0} \subset \mathcal{H}$ tvori *ortogonalni sistem*, če velja $(f_k, f_j) = 0$ za $j \neq k$. Če funkcije $\{f_k\}_{k\ge 0}$ tvorijo ortogonalni sistem in so baza prostora $\mathcal{H}$, potem pravimo, da tvorijo *ortogonalno bazo*. Primeri ortogonalnih baz:

1. *Trigonometrične funkcije* $\{1, \cos(k\,\cdot), \sin(k\,\cdot)\}_{k=0}^\infty$ tvorijo ortogonalno bazo prostora $L^2(-\pi,\pi)$.

2. *Legendreovi polinomi* $\{L_k\}_{k=0}^\infty$ tvorijo ortogonalno bazo prostora $L^2(-1,1)$.

3. *Polinomi Čebiševa prve vrste* $\{T_k\}_{k=0}^\infty$ tvorijo ortogonalno bazo prostora $L_w^2(-1,1)$, kjer je $w(x) = \frac{1}{\sqrt{1-x^2}}$.

Legendreovi polinomi in polinomi Čebiševa (prve vrste) so posebni primeri iz družine *Jacobijevih ortogonalnih polinomov* (glej npr. M. Abramowitz in I. A. Stegun [1]).

Naj bo množica funkcij $\{\phi_k\}_{k=0}^\infty \subset L_w^2(a,b)$ ortogonalni sistem v prostoru $L_w^2(a,b)$, ki ni nujno baza, pač pa vsaj ogrodje. Tedaj lahko vsako s kvadratom integrabilno funkcijo $f \in L_w^2(a,b)$ razvijemo v (Fourierovo) vrsto po (baznih) funkcijah tega ortogonalnega sistema

$$f(x) = \sum_{k=0}^{\infty} a_k \phi_k(x), \tag{2.8}$$

kjer koeficiente a_k v razvoju vrste izračunamo po formuli

$$a_k = \frac{(f, \phi_k)}{\|\phi_k\|^2}. \tag{2.9}$$

2.2 Fourierova vrsta

Naravno vprašanje v mnogih problemih numerične analize, npr. pri reševanju toplotne ali valovne enačbe s Fourierovo metodo ločitve spremenljivk, je, ali se da dano funkcijo zapisati kot trigonometrično vrsto, ki ji pravimo *Fourierova vrsta*. Podrobno obravnavo Fourierove analize najdemo npr. v M. Abramowitz in I. A. Stegun [1], K. Atkinson in W. Han [6], J. P. Boyd [7] ter C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang [11].

Naj bo $f \in L^1(-\pi, \pi)$ integrabilna funkcija. Tedaj je njena Fourierova vrsta definirana z

$$F(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \big(a_k \cos(kx) + b_k \sin(kx)\big), \tag{2.10}$$

kjer Fourierove koeficiente a_k in b_k izračunamo s formulama

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(kx)\, \mathrm{d}x, \quad k \geq 0, \tag{2.11}$$

$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(kx)\, \mathrm{d}x, \quad k \geq 1. \tag{2.12}$$

Fourierova vrsta (2.10) se v primeru, ko je f liha funkcija, poenostavi v *sinusno Fourierovo vrsto*, v primeru, ko je f soda, pa v *kosinusno Fourierovo vrsto*. Pri obravnavanju trigonometričnih vrst se uporabljajo standardne formule iz trigonometrije (npr. adicijski izreki, formule za dvojne in polovične kote).

Za izračun Fourierovih koeficientov a_k in b_k (2.11 – 2.12) uporabimo *hitro Fourierovo transformacijo* (FFT), oz. algoritem z njeno implementacijo, ki sta ga leta 1965 predstavila J. W. Cooley in J. W. Tukey [15]. Algoritem, oz. njegova diskretna verzija, tj. *diskretna Fourierova transformacija* (DFT), je podrobno predstavljen tudi v P. Henrici [34]. Pomembna prednost tega algoritma je v njegovi računski učinkovitosti, saj Fourierove koeficiente do reda N izračuna v $\mathcal{O}(N \log N)$ operacijah.

Konvergenca Fourierove vrste ni samoumevna. Fourierova vrsta F definirana z enačbami (2.10 – 2.12), ne konvergira nujno, če pa konvergira po točkah, ne konvergira nujno k funkciji f. Velja pa, da Fourierova vrsta konvergira v L^2 normi. Za $f \in L^2(-\pi, \pi)$ velja

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \big(a_k \cos(kx) + b_k \sin(kx)\big). \tag{2.13}$$

Dokaz spodnjega izreka ter splošnejšo teorijo konvergence Fourierove vrste najdemo npr. v K. Atkinson in W. Han [6], pa tudi v drugih monografijah, ki obravnavajo harmonično analizo.

Izrek 2.3 *Naj bo $f \in L^2(-\pi, \pi)$ in*

$$P_N f(x) = \frac{a_0}{2} + \sum_{k=1}^{N} \big(a_k \cos(kx) + b_k \sin(kx) \big), \tag{2.14}$$

kjer so koeficienti a_k in b_k definirani v (2.11 – 2.12). Tedaj

$$\|P_N f - f\|_{L^2} \to 0, \qquad N \to \infty \tag{2.15}$$

za vsak $f \in L^2(-\pi, \pi)$ natanko tedaj, ko obstaja konstanta $C > 0$, da je

$$\|P_N f\|_{L^2} \leq C \, \|f\|_{L^2} \tag{2.16}$$

za vsak $N \geq 1$ in vsak $f \in L^2(-\pi, \pi)$.

Posledica konvergence v L^2 normi je *Parsevalova enakost*

$$\|f\|_{L^2}^2 = \pi \left(\frac{|a_0|^2}{2} + \sum_{k=1}^{\infty} \big(|a_k|^2 + |b_k|^2 \big) \right). \tag{2.17}$$

Konvergenca Fourierove vrste je odvisna tudi od stopnje gladkosti dane funkcije. V moderni harmonični analizi se pogosto uporabljajo prostori Soboljeva $H_p^m(-\pi, \pi)$ (2.6), ki sovpadajo s prostori $(m-1)$-krat zvezno odvedljivih, 2π-periodičnih funkcij. Za $f \in H_p^m(-\pi, \pi)$, $m \geq 0$ in $0 \leq \ell \leq m$, veljata oceni (glej C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang [11], podrazdelek 5.1.2)

$$\|f - P_N f\|_{L^2} \leq C_1 \, N^{-m} \, \|f^{(m)}\|_{L^2}, \tag{2.18}$$

$$\|f - P_N f\|_{H_p^\ell} \leq C_2 \, N^{\ell - m} \, \|f^{(m)}\|_{L^2}, \tag{2.19}$$

kjer sta $C_1 > 0$ in $C_2 > 0$ konstanti, $P_N f$ pa je ortogonalna projekcija funkcije f na prostor trigonometričnih polinomov stopnje kvečjemu N, ki je definirana z enačbo (2.14). Oceni sledita iz Parsevalove enakosti (2.17). Velja, da operatorja odvajanja in projeciranja komutirata

$$(P_N f)' = P_N f'.$$

V primeru analitičnih, 2π-periodičnih funkcij, je konvergenca izredno hitra (glej npr. A. Iserles in S. P. Nørsett [39]), saj obstajata taki konstanti $C > 0$ in $\alpha > 0$, da za vse Fourierove koeficiente velja $|a_k|, |b_k| \leq C \mathrm{e}^{-\alpha k}$ za $k \geq 0$. Če izpustimo pogoj 2π-periodičnosti, koeficienti padajo bistveno počasneje: $|a_k|, |b_k| = \mathcal{O}(1/k)$, $k \gg 1$.

2.3 Ortogonalni polinomi

Ortogonalni polinomi predstavljajo pomembno orodje v mnogih problemih numerične analize. Podrobno obravnavo najdemo npr. v M. Abramowitz in I. A. Stegun [1], W. Han in K. Atkinson [6], T. S. Chihara [13], W. Gautschi [22], J. Shen, T. Tang in L. Wang [51] ter G. Szegö [54]. Zaporedje polinomov $\{p_n\}_{n=0}^{\infty}$ je ortogonalno v prostoru $L_w^2(a,b)$ glede na utež w, če je stopnja polinoma p_n enaka n in velja

$$(p_n, p_m) = \int_a^b p_n(x)\, p_m(x)\, w(x)\, \mathrm{d}x = \gamma_n \delta_{mn}, \tag{2.20}$$

kjer je $\delta_{mn} = \begin{cases} 1, & m = n, \\ 0, & m \neq n, \end{cases}$ Kroneckerjev delta in $\gamma_n = \|p_n\|^2$. Sistem ortogonalnih polinomov $\{p_n\}_{n=0}^{\infty}$ lahko dobimo tako, da na monomski bazi $\{1, x, x^2, x^3, \ldots\}$ uporabimo Gram-Schmidtov postopek. Različne izbire intervala (a,b) in uteži w vodijo do različnih družin ortogonalnih polinomov. Tri klasične družine so:

1. $(a,b) = (-1,1)$, $w(x) = (1-x)^{\alpha}(1+x)^{\beta}$, $-1 < \alpha, \beta < 1$: *Jacobijevi polinomi* $P_n^{(\alpha,\beta)}$.
2. $(a,b) = (0,\infty)$, $w(x) = x^{\alpha}\mathrm{e}^{-x}$: *Laguerreovi polinomi* L_n^{α}.
3. $(a,b) = (-\infty,\infty)$, $w(x) = \mathrm{e}^{-x^2}$: *Hermiteovi polinomi* H_n.

V nadaljevanju nas bodo zanimali predvsem posebni primeri Jacobijevih polinomov: Legendreovi polinomi $\{L_n\}_{n=0}^{\infty}$ ($\alpha = \beta = 0$, podrazdelek 2.3.2) ter polinomi Čebiševa prve $\{T_n\}_{n=0}^{\infty}$ in druge vrste $\{U_n\}_{n=0}^{\infty}$ ($\alpha = \beta = \pm\frac{1}{2}$, podrazdelek 2.3.3). Tričlensko rekurzivno formulo podrobneje obravnavamo v razdelku 5.2. Naj bo $\{p_n\}_{n=0}^{\infty}$ sistem ortogonalnih polinomov. Polinom p_n je stopnje n za vsak $n \geq 0$ in ima natanko n ničel na intervalu $(-1,1)$. Za $k \geq 0$ so polinomi p_{2k} sode, polinomi p_{2k+1} pa lihe funkcije.

2.3.1 Momentni funkcional, momentno zaporedje

V večini literature, ki obravnava ortogonalne polinome, npr. v G. Szegö [54], najdemo klasičen pristop k obravnavi le-teh. Drugačen pristop ponuja T. S. Chihara v [13]. Ortogonalne polinome uvede z uporabo momentnega zaporedja in momentnih funkcionalov. Za naše potrebe zadostuje, da zapišemo nekaj osnovnih definicij in lastnosti.

Naj bo $\{\mu_k\}_{k=0}^{\infty}$ zaporedje kompleksnih števil, ki ga imenujemo *momentno zaporedje*, in naj bo V vektorski prostor vseh polinomov. Linearni funkcional $\mathcal{L} : V \to \mathbb{C}$, ki zadošča pogoju $\mathcal{L}(x^k) = \mu_k$ za vsak k, se imenuje *momentni funkcional*. Iz definicije direktno sledi, da za polinom $p(x) = \sum_{k=0}^{n} c_k x^k$ velja $\mathcal{L}(p) = \sum_{k=0}^{n} c_k \mu_k$. Zaporedje $\{p_k\}_{k=0}^{\infty}$ se imenuje

ortogonalno polinomsko zaporedje glede na momentni funkcional $\mathcal{L}$, če je za vsak $k, \ell \geq 0$ p_k polinom stopnje k, $\mathcal{L}(p_\ell\, p_k) = 0$ za $\ell \neq k$ in $\mathcal{L}(p_k^2) \neq 0$. Standarden primer momentnega funkcionala je integral

$$\mathcal{L}(f) := \int_a^b f(x)\, w(x)\, \mathrm{d}x, \tag{2.21}$$

kjer je f integrabilna funkcija na intervalu (a, b) in w integrabilna nenegativna funkcija na intervalu (a, b), ki se imenuje *utež*.

Pristop z momenti je zanimiv in uporaben predvsem zato, ker omogoča izračun vrednosti momentnega funkcionala, ki je običajno podan z integralom (2.21), brez računanja integralov. Pri tem mora biti množica momentov $\{\mu_k\}$ znana.

2.3.2 Legendreovi polinomi

Legendreovi polinomi L_n, $n \geq 0$, so klasični ortogonalni polinomi (glej npr. M. Abramowitz, I. A. Stegun [1], T. S. Chihara [13] ali G. Szegö [54]), ki spadajo v družino Jacobijevih polinomov ($\alpha = \beta = 0$). V prostoru $L^2(-1, 1)$ so ortogonalni glede na utež $w(x) = 1$, in normalizirani tako, da je $L_n(1) = 1$. Običajno so definirani z

$$L_0(x) = 1, \quad L_n(x) = \frac{1}{2^n n!} \frac{\mathrm{d}^n}{\mathrm{d}x^n}[(x^2 - 1)^n], \quad n \geq 1. \tag{2.22}$$

Legendreovi polinomi so ortogonalni

$$\int_{-1}^{1} L_m(x) L_n(x)\, \mathrm{d}x = \begin{cases} 0, & m \neq n, \\ \frac{2}{2n+1}, & m = n. \end{cases} \tag{2.23}$$

Dobimo jih kot rešitve Legendreove diferencialne enačbe

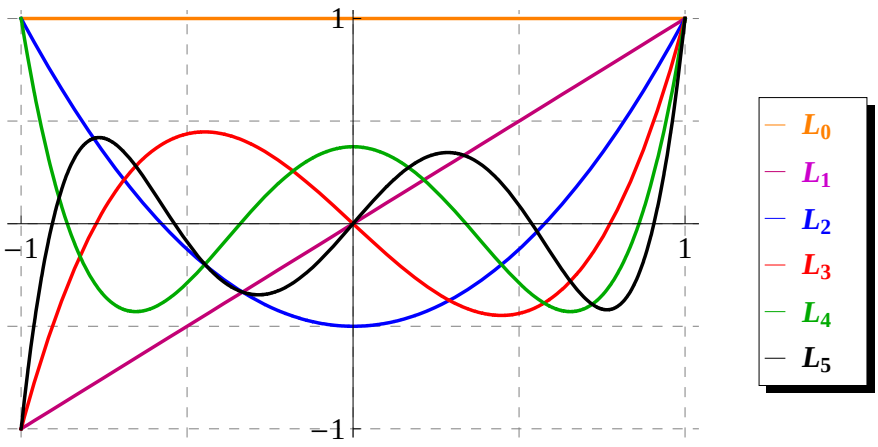
$$-\left[(1 - x^2) L_n'(x)\right]' = n(n+1) L_n(x), \quad n \geq 0. \tag{2.24}$$

Poleg tega zadoščajo tričlenski rekurzivni formuli

$$n L_n(x) = (2n - 1) x L_{n-1}(x) - (n - 1) L_{n-2}(x), \quad n \geq 2, \tag{2.25}$$

kjer je $L_0(x) = 1$ in $L_1(x) = x$. Velja, da je $L_n(-1) = (-1)^n$ in $|L_n(x)| \leq 1$ za vsak $n \geq 0$ in $x \in [-1, 1]$. Prvih šest Legendreovih polinomov je zapisanih spodaj in prikazanih na sliki 2.1:

$$\begin{aligned} L_0(x) &= 1, \\ L_1(x) &= x, \\ L_2(x) &= \tfrac{3}{2}x^2 - \tfrac{1}{2}, \\ L_3(x) &= \tfrac{5}{2}x^3 - \tfrac{3}{2}x, \\ L_4(x) &= \tfrac{35}{8}x^4 - \tfrac{15}{4}x^2 + \tfrac{3}{8}, \\ L_5(x) &= \tfrac{63}{8}x^5 - \tfrac{35}{4}x^3 + \tfrac{15}{8}x. \end{aligned}$$



Slika 2.1: Grafi prvih šestih Legendreovih polinomov L_n: L_0 oranžna črta, L_1 vijolična, L_2 modra, L_3 rdeča, L_4 zelena in L_5 črna črta.

Pri konstrukciji spektralnih metod je pomembno poznati zvezo med Legendreovimi polinomi in njihovimi odvodi. Velja rekurzivna formula

$$L'_{n+1}(x) = (2n+1)L_n(x) + L'_{n-1}(x), \quad n \geq 1, \tag{2.26}$$

od koder za $n \geq 0$ sledi

$$L'_{n+1}(x) = (2n+1)L_n(x) + (2(n-2)+1)L_{n-2}(x) + (2(n-4)+1)L_{n-4}(x) + \cdots . \tag{2.27}$$

Odvodi prvih šestih Legendreovih polinomov so:

$$\begin{aligned}
L'_0(x) &= 0, \\
L'_1(x) &= 1 = L_0(x), \\
L'_2(x) &= 3x = 3L_1(x), \\
L'_3(x) &= \tfrac{15}{2}x^2 - \tfrac{3}{2} = 5L_2(x) + L_0(x), \\
L'_4(x) &= \tfrac{35}{2}x^3 - \tfrac{15}{2}x = 7L_3(x) + 3L_1(x), \\
L'_5(x) &= \tfrac{315}{8}x^4 - \tfrac{105}{4}x^2 + \tfrac{15}{8} = 9L_4(x) + 5L_2(3) + L_0(x).
\end{aligned}$$

Za Legendreove polinome je momentni funkcional (podrazdelek 2.3.1) definiran z

$$\mathcal{L}(f) := \int_{-1}^{1} f(x)\,\mathrm{d}x, \tag{2.28}$$

momentno zaporedje pa je podano z ($k \geq 0$)

$$\mu_k = \begin{cases} 0, & k \text{ lih}, \\ \frac{2}{k+1}, & k \text{ sod}. \end{cases} \tag{2.29}$$

Opazimo, da momenti s k padajo proti 0.

Poljubno integrabilno funkcijo $f \in L^1(-1,1)$ lahko zapišemo z *Legendreovo vrsto*, ki je opisana npr. v C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang [11] ali P. Grandclément [30], in je sestavljena iz Legendreovih polinomov L_k, $k \geq 0$,

$$f(x) = \sum_{k=0}^{\infty} a_k L_k(x). \tag{2.30}$$

Legendreove koeficiente a_k, $k \geq 0$, izračunamo s formulo

$$a_k = \frac{2k+1}{2} \int_{-1}^{1} f(x) L_k(x)\, \mathrm{d}x. \tag{2.31}$$

S $P_N f(x) = \sum_{k=0}^{N} a_k L_k(x)$ označimo odrezano Legendreovo vrsto pri nekem odreznem številu N. Rezanje vrste je ekvivalentno ortogonalni projekciji funkcije f na prostor polinomov stopnje kvečjemu N. Za dano m-krat zvezno odvedljivo funkcijo $f \in \mathcal{C}^m(-1,1)$, $m \geq 0$, veljata spodnji oceni o kvaliteti aproksimacije (glej C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang [11], podrazdelek 5.4.2)

$$\|f - P_N f\|_{L^2} \leq C_1\, N^{-m}\, \|f^{(m)}\|_{L^2}, \tag{2.32}$$

$$\|f - P_N f\|_{\infty} \leq C_2\, N^{1/2-m}\, V(f^{(m)}), \tag{2.33}$$

kjer sta $C_1 > 0$ in $C_2 > 0$ konstanti. Funkcija $f^{(m)}$ naj ima omejeno totalno variacijo $V(f^{(m)})$. *Totalna variacija* funkcije f na intervalu $[a,b]$ je definirana z

$$V(f) = \sup_{a=x_0<x_1<\cdots<x_n=b} \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|, \tag{2.34}$$

kjer vzamemo supremum po vseh delitvah intervala $[a,b]$ s končnim številom točk, tj. po vseh množicah $n+1$ točk, kjer je $a = x_0 < x_1 < \cdots < x_n = b$ in n poljuben (glej [11], dodatek A.8). Totalna variacija je *omejena* na intervalu $[a,b]$, če je število $V(f)$ končno. Vsaka funkcija z omejeno totalno variacijo je omejena.

V nasprotju s Fourierovo vrsto, operatorja odvajanja in projeciranja v splošnem ne komutirata

$$(P_N f)' \neq P_{N-1} f'.$$

2.3.3 Polinomi Čebiševa prve in druge vrste

Polinomi Čebiševa T_n, $n \geq 0$, prve vrste so družina klasičnih ortogonalnih polinomov v Hilbertovem prostoru $L^2_w(-1,1)$, ki so ortogonalni glede na utež $w(x) = 1/\sqrt{1-x^2}$, in normalizirani tako, da je $T_n(1) = 1$ (glej npr. M. Abramowitz, I. A. Stegun [1], T. S. Chihara [13] ali G. Szegö [54]).

Podobno kot Legendreovi polinomi spadajo v družino Jacobijevih polinomov ($\alpha = \beta = -\frac{1}{2}$). Skalarni produkt v prostoru $L^2_w(-1,1)$ je definiran z enačbama (2.1) in (2.2), kjer je $(a,b) = (-1,1)$.

Polinomi Čebiševa prve vrste so povsem karakterizirani z lastnostjo

$$T_n(\cos\theta) = \cos n\theta, \quad n \geq 0, \tag{2.35}$$

ki opisuje dejstvo, da je $\cos n\theta$ polinom v $\cos\theta$. So ortogonalni

$$\int_{-1}^{1} \frac{T_m(x)T_n(x)}{\sqrt{1-x^2}}\,\mathrm{d}x = \begin{cases} 0, & m \neq n, \\ \pi, & m = n = 0, \\ \frac{\pi}{2}, & m = n > 0. \end{cases} \tag{2.36}$$

Nadalje zadoščajo diferencialni enačbi

$$-\left[\sqrt{1-x^2}T_n'(x)\right]' = n^2 \frac{T_n(x)}{\sqrt{1-x^2}}, \quad n \geq 0, \tag{2.37}$$

poleg tega pa zadoščajo tudi tričlenski rekurzivni formuli

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), \quad n \geq 2, \tag{2.38}$$

kjer je $T_0(x) = 1$ in $T_1(x) = x$. Velja, da je $T_n(-1) = (-1)^n$ in $|T_n(x)| \leq 1$ za vsak $n \geq 0$ in $x \in [-1,1]$. Opazimo tudi, da ti polinomi oscilirajo med -1 in 1. Prvih šest polinomov Čebiševa prve vrste je zapisanih spodaj in prikazanih na sliki 2.2:

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_2(x) &= 2x^2 - 1, \\ T_3(x) &= 4x^3 - 3x, \\ T_4(x) &= 8x^4 - 8x^2 + 1, \\ T_5(x) &= 16x^5 - 20x^3 + 5x. \end{aligned}$$

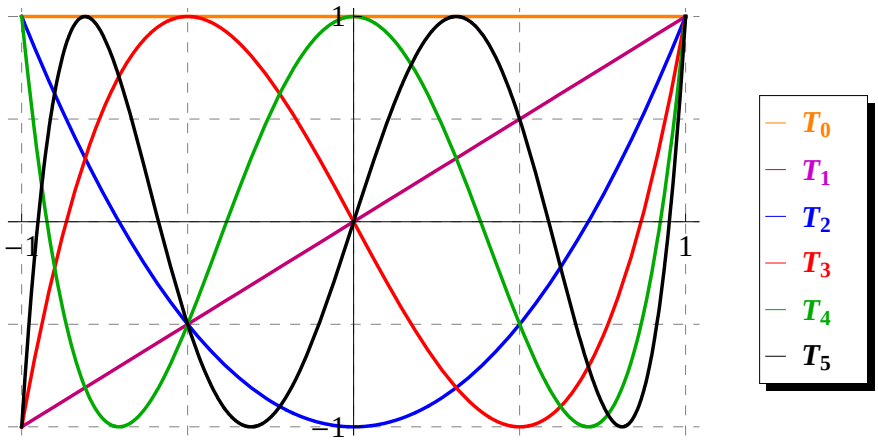
Podobno kot pri Legendreovih polinomih nas zanima zveza med polinomi Čebiševa prve vrste in njihovimi odvodi. Velja rekurzivna formula

$$\tfrac{1}{n+1}T_{n+1}'(x) = 2T_n(x) + \tfrac{1}{n-1}T_{n-1}'(x), \tag{2.39}$$

od koder za $n \geq 0$ sledi

$$T_{2n}'(x) = 4n\left(T_{2n-1}(x) + T_{2n-3}(x) + \cdots + T_1(x)\right), \tag{2.40}$$

$$T_{2n+1}'(x) = (2n+1)\left(2T_{2n}(x) + 2T_{2n-2}(x) + \cdots + 2T_2(x) + T_0(x)\right). \tag{2.41}$$



Slika 2.2: Grafi prvih šestih polinomov Čebiševa prve vrste T_n: T_0 oranžna črta, T_1 vijolična, T_2 modra, T_3 rdeča, T_4 zelena in T_5 črna črta.

Odvodi prvih šestih polinomov Čebiševa prve vrste so:

$$\begin{aligned}
T_0'(x) &= 0,\\
T_1'(x) &= 1 = T_0(x),\\
T_2'(x) &= 4x = 4T_1(x),\\
T_3'(x) &= 12x^2 - 3 = 6T_2(x) + 3T_0(x),\\
T_4'(x) &= 32x^3 - 16x = 8T_3(x) + 8T_1(x),\\
T_5'(x) &= 80x^4 - 60x^2 + 5 = 10T_4(x) + 10T_2(3) + 5T_0(x).
\end{aligned}$$

Poleg odvajanja je pomembna tudi povezava med polinomi Čebiševa prve vrste in njihovimi medsebojnimi produkti. Veljata formuli, ki sledita iz lastnosti (2.35) in adicijskih izrekov za trigonometrične funkcije

$$T_n^2(x) = \tfrac{1}{2}\left(T_0(x) + T_{2n}(x)\right), \tag{2.42}$$

$$T_n(x)\cdot T_m(x) = \tfrac{1}{2}\left(T_{n-m}(x) + T_{n+m}(x)\right), \quad m \le n. \tag{2.43}$$

Momentni funkcional za polinome Čebiševa prve vrste je

$$\mathcal{L}(f) := \int_{-1}^{1} \tfrac{f(x)}{\sqrt{1-x^2}}\,\mathrm{d}x, \tag{2.44}$$

momentno zaporedje pa je podano z ($k \ge 0$, $\ell \ge 0$)

$$\mu_k = \begin{cases} 0, & k = 2\ell + 1,\\ \frac{\binom{2\ell}{\ell}\pi}{4^\ell}, & k = 2\ell. \end{cases} \tag{2.45}$$

Opazimo, da momenti s k padajo proti 0.

Polinomi Čebiševa U_n, $n \geq 0$, druge vrste so družina klasičnih ortogonalnih polinomov v Hilbertovem prostoru $L^2_w(-1,1)$, ki so ortogonalni glede na utež $w(x) = \sqrt{1-x^2}$, in normalizirani tako, da je $U_n(1) = n+1$. Tudi ti polinomi spadajo v družino Jacobijevih polinomov ($\alpha = \beta = \frac{1}{2}$).

Polinomi Čebiševa druge vrste so karakterizirani z lastnostjo

$$U_n(\cos\theta) = \frac{\sin(n+1)\theta}{\sin\theta}, \quad n \geq 0. \tag{2.46}$$

So ortogonalni

$$\int_{-1}^{1} U_m(x)U_n(x)\sqrt{1-x^2}\,\mathrm{d}x = \begin{cases} 0, & m \neq n, \\ \frac{\pi}{2}, & m = n. \end{cases} \tag{2.47}$$

Nadalje zadoščajo diferencialni enačbi

$$-\left[\sqrt{(1-x^2)^3}U_n'(x)\right]' = n(n+2)\sqrt{1-x^2}U_n(x), \quad n \geq 0, \tag{2.48}$$

poleg tega pa zadoščajo tudi tričlenski rekurzivni formuli

$$U_n(x) = 2xU_{n-1}(x) - U_{n-2}(x), \tag{2.49}$$

kjer je $U_0(x) = 1$ in $U_1(x) = 2x$. Velja, da je $U_n(-1) = (n+1)(-1)^n$ in $|U_n(x)| \leq n+1$ za vsak $n \geq 0$ in $x \in [-1,1]$. Prvih šest polinomov Čebiševa druge vrste je zapisanih spodaj in prikazanih na sliki 2.3:

$$\begin{aligned} U_0(x) &= 1, \\ U_1(x) &= 2x, \\ U_2(x) &= 4x^2 - 1, \\ U_3(x) &= 8x^3 - 4x, \\ U_4(x) &= 16x^4 - 12x^2 + 1, \\ U_5(x) &= 32x^5 - 32x^3 + 6x. \end{aligned}$$

Velja rekurzivna zveza med polinomi Čebiševa druge vrste in njihovimi odvodi

$$U_n'(x) = 2nU_{n-1}(x) + U_{n-2}'(x), \tag{2.50}$$

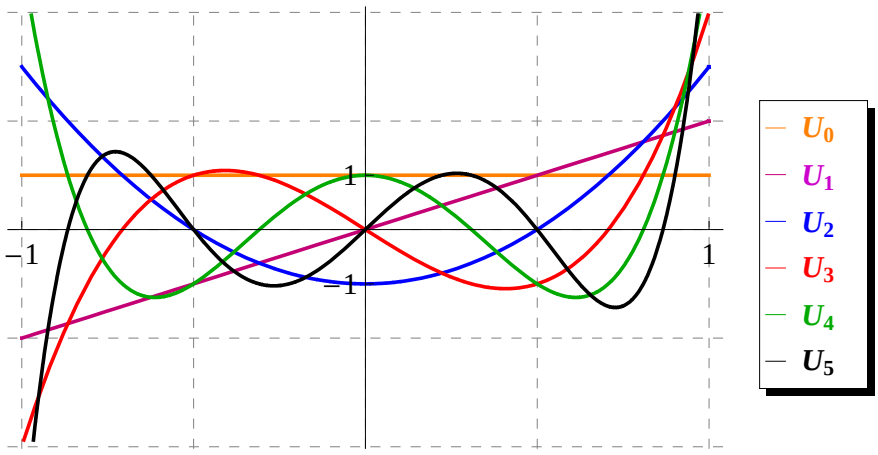
od koder za $n \geq 0$ sledi

$$U_{2n}'(x) = 4\left(nU_{2n-1}(x) + (n-1)U_{2n-3}(x) + \cdots + U_1(x)\right), \tag{2.51}$$

$$U_{2n+1}'(x) = 2\left((2n+1)U_{2n}(x) + (2n-1)U_{2n-2}(x) + \cdots + U_0(x)\right). \tag{2.52}$$

Odvodi prvih šestih polinomov Čebiševa druge vrste so:

$$\begin{aligned} U_0'(x) &= 0, \\ U_1'(x) &= 2 = 2U_0(x), \\ U_2'(x) &= 8x = 4U_1(x), \\ U_3'(x) &= 24x^2 - 4 = 6U_2(x) + 2U_0(x), \\ U_4'(x) &= 64x^3 - 24x = 8U_3(x) + 4U_1(x), \\ U_5'(x) &= 160x^4 - 96x^2 + 6 = 10U_4(x) + 6U_2(3) + 2U_0(x). \end{aligned}$$



Slika 2.3: Grafi prvih šestih polinomov Čebiševa druge vrste U_n: U_0 oranžna črta, U_1 vijolična, U_2 modra, U_3 rdeča, U_4 zelena in U_5 črna črta.

Momentni funkcional za polinome Čebiševa druge vrste je

$$\mathcal{L}(f) := \int_{-1}^{1} f(x)\sqrt{1-x^2}\,\mathrm{d}x, \tag{2.53}$$

momentno zaporedje pa je podano z ($k \geq 0$, $\ell \geq 0$)

$$\mu_k = \begin{cases} 0, & k = 2\ell+1, \\ \frac{\binom{2\ell}{\ell}\pi}{2^{2\ell+1}(\ell+1)}, & k = 2\ell. \end{cases} \tag{2.54}$$

Opazimo, da momenti s k padajo proti 0.

2.3.4 Vrsta Čebiševa

Podobno kot v primeru trigonometričnih funkcij ali Legendreovih polinomov, kjer dano funkcijo razvijemo v Fourierovo oz. Legendreovo vrsto, lahko vsako integrabilno funkcijo $f \in L^1_w(-1,1)$ razvijemo v *vrsto Čebiševa*, ki je opisana npr. v C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang [11], P. Grandclément [30] ter J. Shen, T. Tang in L. Wang [51], in je sestavljena iz polinomov Čebiševa prve vrste T_k, $k \geq 0$,

$$f(x) = \sum_{k=0}^{\infty} a_k T_k(x). \tag{2.55}$$

Koeficiente Čebiševa a_k, $k \geq 0$, izračunamo s formulama

$$a_0 = \frac{1}{\pi}\int_{-1}^{1} \frac{f(x)T_0(x)}{\sqrt{1-x^2}}\,\mathrm{d}x = \frac{1}{\pi}\int_0^{\pi} f(\cos\theta)\,\mathrm{d}\theta, \tag{2.56}$$

$$a_k = \frac{2}{\pi}\int_{-1}^{1} \frac{f(x)T_k(x)}{\sqrt{1-x^2}}\,\mathrm{d}x = \frac{2}{\pi}\int_0^{\pi} f(\cos\theta)\cos(k\theta)\,\mathrm{d}\theta. \tag{2.57}$$

Vrsta Čebiševa je ekvivalentna kosinusni Fourierovi vrsti ($x = \cos\theta$). V praksi pogosto uporabljamo aproksimacijo z odrezano vrsto Čebiševa pri nekem odreznem številu N

$$f(x) \approx P_N f(x) := \sum_{k=0}^{N} a_k T_k(x). \tag{2.58}$$

Rezanje vrste je ekvivalentno ortogonalni projekciji funkcije f na prostor polinomov stopnje kvečjemu N. Koeficiente te vrste učinkovito izračunamo z uporabo *diskretne kosinusne transformacije (DCT)*. Podobno lahko vsako integrabilno funkcijo $f \in L^1_w(-1,1)$ aproksimiramo z odrezano vrsto Čebiševa, ki je sestavljena iz polinomov Čebiševa druge vrste U_k.

Za dano m-krat zvezno odvedljivo funkcijo $f \in \mathcal{C}^m(-1,1)$, $m \geq 0$ veljata spodnji oceni o kvaliteti aproksimacije (glej C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang [11], podrazdelek 5.5.2)

$$\|f - P_N f\|_{L^2_w} \leq C_1\, N^{-m}\, \|f^{(m)}\|_{L^2_w}, \tag{2.59}$$

$$\|f - P_N f\|_{\infty} \leq \frac{C_2\,(1+\ln N)}{N^m} \sum_{k=0}^{m} \|f^{(k)}\|_{\infty}, \tag{2.60}$$

kjer sta $C_1 > 0$ in $C_2 > 0$ konstanti. Podobno kot v primeru Legendreovih polinomov, operatorja odvajanja in projeciranja v splošnem ne komutirata

$$(P_N f)' \neq P_{N-1} f'.$$

Iz povedanega sledi, da je tako kot pri Fourierovi oz. Legendreovi vrsti konvergenca vrste Čebiševa odvisna od stopnje gladkosti dane funkcije. Veljajo spodnji izreki, katerih dokaze lahko najdemo npr. v J. P. Boyd [7] ali L. N. Trefethen [60].

Izrek 2.4 *Naj bo $P_N f(x) = \sum_{n=0}^{N} a_n T_n(x)$ odrezana vrsta Čebiševa za dano funkcijo $f \in L^1_w(-1,1)$. Tedaj je*

$$|f(x) - P_N f(x)| \leq \sum_{n=N+1}^{\infty} |a_n| \tag{2.61}$$

za vse f, N in $x \in [-1,1]$.

Koeficienti Čebiševa a_n padajo proti 0, ko gre $n \to \infty$. Za funkcije $f \in \mathcal{C}^m(-1,1)$ dobimo sledeče rezultate.

Izrek 2.5 *Naj ima dana funkcija f za nek $m \geq 0$ absolutno zvezen $(m-1)$-vi odvod $f^{(m-1)}$ na intervalu $[-1,1]$ (če je $m > 0$) in m-ti odvod $f^{(m)}$ z omejeno totalno variacijo $V(f^{(m)})$ za $m \geq 1$. Tedaj za $n \geq m+1$ za koeficiente Čebiševa velja ocena*

$$|a_n| \leq \frac{2V(f^{(m)})}{\pi n(n-1)\cdots(n-m)} \leq \frac{2V(f^{(m)})}{\pi(n-m)^{m+1}}. \tag{2.62}$$

Izrek 2.6 *Naj ima dana funkcija f za nek $m \geq 0$ absolutno zvezen $(m-1)$-vi odvod $f^{(m-1)}$ na intervalu $[-1,1]$ (če je $m > 0$) in m-ti odvod $f^{(m)}$ z omejeno totalno variacijo $V(f^{(m)})$ za $m \geq 1$. Naj bo dalje $P_N f(x) = \sum_{n=0}^{N} a_n T_n(x)$ njena odrezana vrsta Čebiševa. Tedaj za vsak $N > m$ velja ocena*

$$\|f - P_N f\| \leq \frac{2V(f^{(m)})}{\pi m(N-m)^m}. \tag{2.63}$$

Iz ocene napake v zadnjem izreku sledi, da za gladke funkcije $f \in \mathcal{C}^\infty$, še posebej pa za analitične funkcije, napaka pada hitreje od vsake potence od $1/N$. Govorimo o *eksponentnem padanju* oz. *spektralni konvergenci*. To je pomembna lastnost spektralnih metod, ki jih bomo obravnavali v nadaljevanju. Veljata spodnja rezultata.

Izrek 2.7 *Naj bo f analitična funkcija na intervalu $[-1,1]$, ki ima analitično nadaljevanje na odprto elipso E_ρ, kjer je $\rho > 1$ vsota glavnih polosi, ter naj zadošča pogoju $|f(x)| \leq M$ za nek $M > 0$. Tedaj za koeficiente Čebiševa velja*

$$|a_n| \leq 2M\rho^{-n}, \tag{2.64}$$

kjer je $|a_0| \leq M$.

Izrek 2.8 *Naj bo f analitična funkcija na intervalu $[-1,1]$, ki ima analitično nadaljevanje na odprto elipso E_ρ, kjer je $\rho > 1$ vsota glavnih polosi, ter naj zadošča pogoju $|f(x)| \leq M$ za nek $M > 0$. Naj bo dalje $P_N f(x) = \sum_{n=0}^{N} a_n T_n(x)$ njena odrezana vrsta Čebiševa. Tedaj za vsak $N \geq 0$ velja*

$$\|f - P_N f\| \leq \frac{2M\rho^{-N}}{\rho - 1}. \tag{2.65}$$

Poglavje 3

Uvod v spektralne metode

Za reševanje robnih problemov tako pri navadnih (npr. problem 1.3), kot tudi pri parcialnih diferencialnih enačbah (npr. problem 1.4 ali problem 1.5), imamo na voljo različne numerične metode. Le-te lahko klasificiramo po več kriterijih. Osnovna ideja je, da iskano rešitev u danega problema aproksimiramo s funkcijo $\tilde{u}$, ki je podana v obliki končne vrste

$$\tilde{u}(x) = \sum_{n=0}^{N} \tilde{u}_n \phi_n(x),$$

kjer so ϕ_n *bazne funkcije*, npr. polinomi ali trigonometrični polinomi stopnje n, in iščemo *spektralne koeficiente* $\tilde{u}_n$. Glede na obliko baznih funkcij konstruiramo različne razrede numeričnih metod. Eden izmed kriterijev je *globalnost* oz. *lokalnost* baznih funkcij, tj., ali so bazne funkcije definirane in netrivialne na celotnem območju, kjer je definiran robni problem, ali zgolj na neki manjši podmnožici tega območja. Matrika koeficientov sistema enačb za izračun koeficientov $\tilde{u}_n$, ki jo dobimo pri konstrukciji numerične metode, je posledično velika in razpršena, oz. majhna in polna. Glede na ta kriterij ločimo tri standardne razrede metod.

1. *Metode končnih razlik* (FDM) so bile razvite v petdesetih letih 20. stoletja. Za bazne funkcije ϕ_n vzamemo lokalne polinome nizkega reda. Prednost teh metod je sorazmerno preprosta konstrukcija, ceno pa plačamo z reševanjem velikega sistema linearnih enačb za izračun koeficientov $\tilde{u}_n$. Matrika koeficientov tega sistema je sicer velika, toda razpršena, oz. ima pasovno strukturo.

2. *Metode končnih elementov* (FEM) so bile razvite v šestdesetih letih 20. stoletja. Za bazne funkcije ϕ_n vzamemo lokalne gladke funkcije. Podobno kot pri metodah končnih razlik dobimo zaradi lokalnosti baznih funkcij velik sistem linearnih enačb, katerega matrika koeficientov je velika, toda razpršena, oz. ima pasovno strukturo.

3. *Spektralne metode* (SM) so bile razvite v sedemdesetih letih 20. stoletja. Za bazne funkcije ϕ_n v nasprotju s FDM ali FEM vzamemo globalne gladke funkcije, tipično trigonometrične funkcije ali ortogonalne polinome (npr. Legendreove polinome ali polinome Čebiševa). Za primerljivo kvaliteto aproksimacije potrebujemo bistveno manj baznih funkcij, zato je matrika sistema linearnih enačb sedaj majhna, toda polna. Zahtevnejša konstrukcija spektralnih metod v primerjavi z metodami končnih razlik je poplačana z dejstvom, da lahko spektralne metode s primernimi adaptacijami uporabimo pri reševanju raznovrstnih problemov. Tipično lahko z njimi dosežemo visoko stopnjo natančnosti z omejenim številom računskih operacij. Računska učinkovitost (časovna in prostorska) se še posebej izkaže pri večdimenzionalnih problemih. Poleg tega je uporaba spektralnih metod nepričakovano učinkovita v mnogih primerih, kjer imamo opravka s funkcijami, ki niso gladke, morda celo nezvezne. Za analitične funkcije pada napaka aproksimacije glede na N eksponentno, in ne zgolj polinomsko.

V nadaljevanju tega dela obravnavamo zgolj spektralne metode.

3.1 Metoda uteženega residuala

Spektralne metode uporabljamo za reševanje robnih problemov tako pri navadnih, kot tudi pri parcialnih diferencialnih enačbah. Podroben pregled spektralnih metod najdemo v različnih člankih, npr. v B. Fornberg in D. M. Sloan [19], P. Grandclément [30], P. Grandclément in J. Novak [31] ter monografijah, npr. v J. P. Boyd [7], C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang [10] in [11], B. Fornberg [18], D. Gottlieb in S. A. Orszag [28], B. Mercier [44], J. Shen, T. Tang in L. Wang [51] ter L. N. Trefethen [56], [57] in [60]. Namen tega dela ni v podrobni predstavitvi spektralnih metod, pač pa bralca vpeljati v svet teh metod. Tako se bomo v tem poglavju omejili na reševanje linearnih dvotočkovnih robnih problemov v eni dimenziji, ki so dani s problemom 1.3 in so oblike

$$\mathcal{L}u(x) = f(x), \qquad x \in [-1, 1], \tag{3.1}$$

z robnimi pogoji

$$\mathcal{B}u(x) = 0, \qquad x \in \{-1, 1\}, \tag{3.2}$$

kjer je $\mathcal{L}$ linearni diferencialni operator (1.6) in $\mathcal{B}$ par linearnih robnih diferencialnih operatorjev, ki ustrezajo Dirichletovim, Neumannovim ali mešanim (Robinovim) robnim pogojem. V nadaljevanju se omejimo na Dirichletove robne pogoje.

Numerično rešitev robnega problema (3.1 – 3.2) iščemo v končnorazsežnem podprostoru $\mathcal{P}_N$ izbranega Hilbertovega prostora $\mathcal{H}$ nad intervalom

$[-1, 1]$, ki je opremljen s skalarnim produktom

$$(p, q) = \int_{-1}^{1} p(x)\, q(x)\, w(x)\, \mathrm{d}x, \tag{3.3}$$

kjer je w nenegativna *utež*. *Spektralno bazo* prostora $\mathcal{P}_N$, ki je običajno ortogonalna, označimo z $\{\phi_0, \phi_1, \ldots, \phi_N\}$, približno rešitev problema (3.1 – 3.2) pa zapišemo kot linearno kombinacijo *baznih funkcij*

$$\tilde{u}(x) = \sum_{k=0}^{N} \tilde{u}_k \phi_k(x). \tag{3.4}$$

V praksi je torej potrebno poiskati spektralne koeficiente $\tilde{u}_k$, ki nastopajo v razvoju (3.4). Funkcija $\tilde{u}$ je tedaj dopustna numerična rešitev, če zadošča robnim pogojem (tj. zadošča enačbi (3.2) na strojno natančnost) in zanjo velja, da je residual

$$r = \mathcal{L}\tilde{u} - f \tag{3.5}$$

dovolj majhen. Za zapis kriterija, ki določa, kaj je to majhen residual, se opremo na primerno število (npr. $N + 1$) *testnih funkcij* ψ_i, $i = 0, 1, \ldots, N$, za katere zahtevamo, da je njihov skalarni produkt (3.3) z residualom r (3.5) enak 0, torej

$$(\psi_i, r) = \int_{-1}^{1} r(x)\, \psi_i(x)\, w(x)\, \mathrm{d}x = 0, \qquad i = 0, 1, \ldots, N. \tag{3.6}$$

Gornji pogoj lahko aproksimiramo z diskretnim skalarnim produktom

$$(\psi_i, r)_N = \sum_{j=0}^{N} w_j\, r(x_j)\, \psi_i(x_j) = 0, \qquad i = 0, 1, \ldots, N, \tag{3.7}$$

kjer je $\{x_j\}_{j=0}^{N}$ množica predhodno določenih *kolokacijskih točk*, $\{w_j\}_{j=0}^{N}$ pa so *uteži*, ki pripadajo izbrani kvadraturni formuli. Za naše potrebe naj bo število kolokacijskih točk enako številu baznih oz. testnih funkcij. Če je metoda konvergentna, se natančnost numerične rešitve z naraščajočim N povečuje, kar pomeni, da je le-ta vedno bližja točni rešitvi. Z različnimi izbirami spektralnih baznih ϕ_k in testnih funkcij ψ_i, lahko konstruiramo različne tipe spektralnih metod. Nekaj najbolj razširjenih z uporabo na modelnem problemu bomo predstavili v nadaljevanju. Pri konstrukciji spektralnih metod se tako pojavita dve glavni vprašanji:

1. Kako primerno izbrati bazne funkcije ϕ_k?
2. Kako določiti spektralne koeficiente $\tilde{u}_k$ v razvoju vrste (3.4)?

Na zastavljeni vprašanji bomo odgovorili v naslednjih dveh razdelkih.

3.2 Izbor baznih funkcij

Izbor baznih funkcij določajo naslednje tri zahteve.

1. Aproksimacija, podana s končno vrsto $\sum_{k=0}^{N} \tilde{u}_k \phi_k(x)$, mora glede na N hitro konvergirati k funkciji u vsaj za dovolj gladke funkcije.

2. Za dane koeficiente $\tilde{u}_k$ mora biti določitev koeficientov $\tilde{v}_k$, za katere velja
$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\sum_{k=0}^{N} \tilde{u}_k \phi_k(x)\right) = \sum_{k=0}^{N} \tilde{v}_k \phi_k(x), \tag{3.8}$$
učinkovita.

3. Pretvorba med spektralnimi koeficienti $\tilde{u}_k$, $k = 0, 1, \ldots, N$, in funkcijskimi vrednostmi $\tilde{u}(x_j)$, kjer je $\{x_j\}_{j=0}^{N}$ množica danih vozlov, mora biti hitro in učinkovito izvedljiva.

Danim zahtevam zadoščajo *trigonometrične funkcije* in *ortogonalni polinomi*, predvsem *polinomi Čebiševa*. Ločimo dva tipa problemov, in sicer *periodične* in *neperiodične probleme*.

Pri periodičnih problemih za bazne funkcije izberemo trigonometrične funkcije $\{1, \cos(k\,\pi\,\cdot), \sin(k\,\pi\,\cdot)\}_{k=0}^{\infty}$, rešitev pa aproksimiramo s *Fourierovo vrsto* (2.10) (glej razdelek 2.2). Prvi dve zahtevi sta izpolnjeni direktno, tretja zahteva pa je bila zadovoljivo izpolnjena leta 1965, ko je bila v članku J. W. Cooley in J. W. Tukey [15] opisana *hitra Fourierova transformacija* (FFT).

Poleg izbora baznih funkcij je pomemben tudi izbor vozlov na danem intervalu, na katerem iščemo rešitev. V primeru periodičnih problemov interval $[-1, 1]$ razdelimo enakomerno z *ekvidistantnimi točkami*

$$x_j = -1 + jh, \quad h = \frac{2}{N}, \quad j = 0, 1, \ldots, N. \tag{3.9}$$

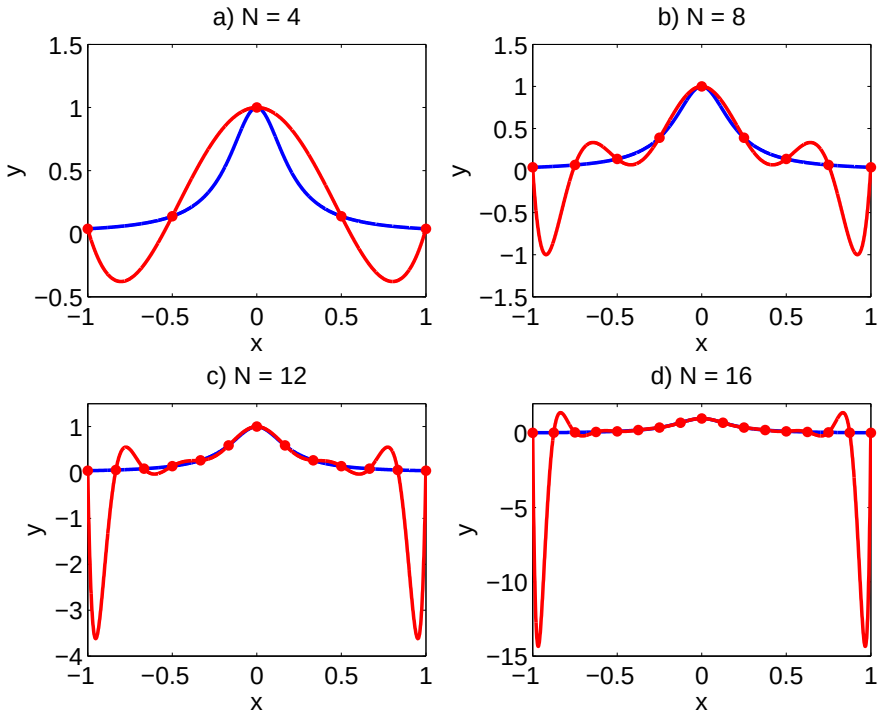
Na sliki 3.2 spodaj so prikazane ekvidistantne točke za primer $N = 8$.

Pri neperiodičnih problemih izberemo za bazne funkcije ortogonalne polinome Jacobijevega tipa, predvsem Legendreove polinome ali polinome Čebiševa, pri čemer so najpogostejša izbira polinomi Čebiševa prve vrste $\{T_k\}_{k=0}^{\infty}$, rešitev pa aproksimiramo z vrsto Čebiševa (2.55) (glej podrazdelek 2.3.4). Izbira trigonometričnih funkcij v tem primeru ni dobra, saj zaradi nenaravne uporabe periodičnih baznih funkcij v neperiodičnem problemu prva zahteva ni izpolnjena. V primeru nezveznosti pa se lahko pojavi tudi *Gibbsov fenomen*. Konvergenca k točni rešitvi je zato počasna, saj koeficienti v Fourierovi vrsti padajo z N samo kot $\mathcal{O}(1/N)$. Druga težava zaradi uporabe ekvidistantnih točk je pojav *Rungejevega fenomena*, saj se v okolici robnih točk pojavijo oscilacije.

Na sliki 3.1 je prikazana interpolacija *Rungejeve funkcije* $f(x) = \frac{1}{1+25x^2}$ na intervalu $[-1, 1]$ z interpolacijskim polinomom Čebiševa

$$I_N^{CH} f(x) := \sum_{k=0}^{N} \hat{a}_k T_k(x), \tag{3.10}$$

ki se ujema z dano funkcijo f v $N+1$ ekvidistantnih točkah za različne vrednosti N ($N = 4, 8, 12$ in 16). Opazimo, da interpolacijski polinom Čebiševa dobro aproksimira dano funkcijo f na sredini intervala, blizu robov pa se pojavijo oscilacije, ki se z naraščajočim številom interpolacijskih delilnih točk N večajo. To kaže na to, da zaporedje interpolacijskih polinomov Čebiševa $I_N^{CH} f$ ne konvergira k dani funkciji f, ko gre N proti neskončno.



Slika 3.1: Rungejev fenomen: interpolacija Rungejeve funkcije $f(x) = \frac{1}{1+25x^2}$ (modra črta) na intervalu $[-1, 1]$ z interpolacijskim polinomom Čebiševa $I_N^{CH} f$ za $N + 1$ ekvidistantnih točk (rdeča črta) za a) $N = 4$, b) $N = 8$, c) $N = 12$ in d) $N = 16$. Vozli so označeni z rdečimi pikami.

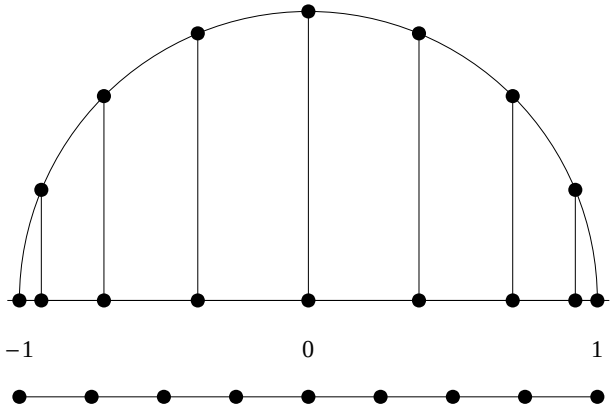
Pri konstrukciji spektralnih metod imamo dve možnosti za določitev numerične rešitve v obliki končne vrste. Prva je projekcija, kjer neskončno vrsto odrežemo pri nekem N, druga pa je interpolacija, kjer rešitev interpoliramo s končno vrsto sestavljeno iz baznih funkcij do nekega N, ki se z

neskončno vrsto ujema v $N+1$ delilnih točkah. V tem delu bomo večinoma obravnavali projekcijske metode, vendar je moč Rungejev fenomen bolj nazorno prikazati z rešitvijo interpolacijskega problema.

Opisanim težavam se običajno izognemo, če rešitev danega problema aproksimiramo z vrsto Čebiševa, interval $[-1,1]$ pa razdelimo s *točkami Čebiševa (druge vrste)*

$$x_j = -\cos\left(\frac{\pi j}{N}\right), \quad j = 0, 1, \ldots, N. \tag{3.11}$$

Slika 3.2 zgoraj prikazuje točke Čebiševa za primer $N = 8$. Pripadajoča polkrožnica nad intervalom $[-1,1]$ je razdeljena ekvidistantno.



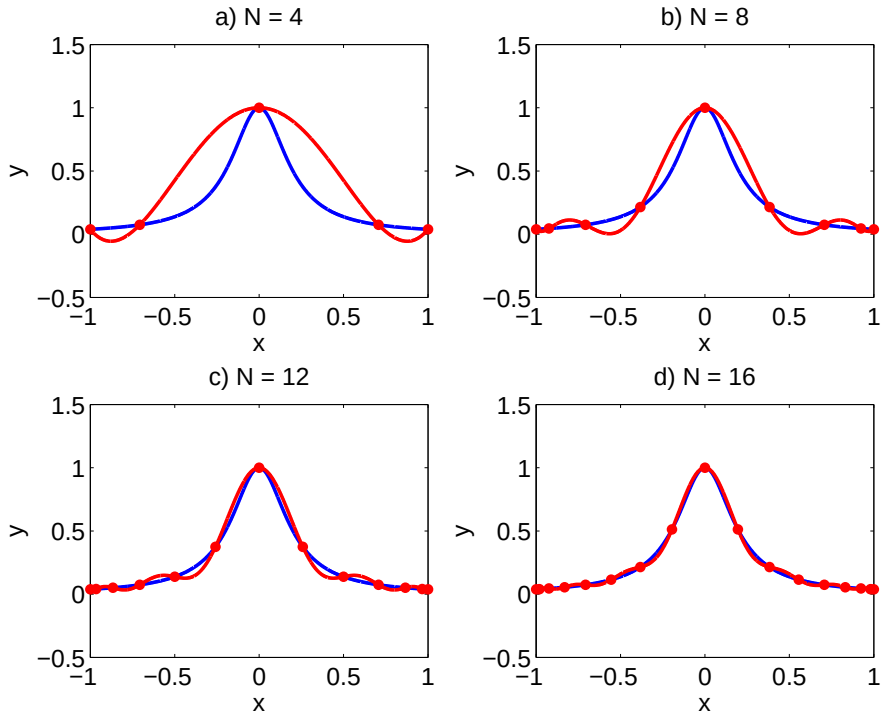
Slika 3.2: Ekvidistantne točke (spodaj) in točke Čebiševa (zgoraj) za $N = 8$.

Vozli, ki so podani v formuli (3.11) se imenujejo točke Čebiševa druge vrste. To so tiste točke, kjer polinomi Čebiševa prve vrste dosežejo ekstremne vrednosti ± 1. Glavna lastnost in hkrati razlog za njihovo uporabo je dejstvo, da se te točke gostijo proti robu intervala, kar prepreči oscilacije blizu roba, ki smo jim priča pri uporabi ekvidistantnih točk.

Ničle polinomov Čebiševa (tj. točke Čebiševa prve vrste) ležijo med točkami Čebiševa druge vrste. Ko govorimo o točkah Čebiševa, imamo običajno v mislih točke druge vrste. Te namreč vsebujejo tudi robova danega intervala, kar omogoča obravnavo robnih pogojev. V odvisnosti od zahtev danega problema pa lahko uporabimo tudi točke Čebiševa prve vrste.

Na sliki 3.3 pa je prikazana interpolacija Rungejeve funkcije na intervalu $[-1,1]$ z interpolacijskim polinomom Čebiševa $I_N^{CH} f$ (3.10), ki se ujema z dano funkcijo f v $N+1$ točkah Čebiševa za za različne vrednosti N ($N = 4, 8, 12$ in 16). Opazimo, da interpolacijski polinom Čebiševa sedaj dobro aproksimira dano funkcijo f na celotnem intervalu, saj oscilacij blizu

robov ni. Rungejevega fenomena v tem primeru ni. To kaže na to, da zaporedje interpolacijskih polinomov Čebiševa $I_N^{CH} f$ konvergira k dani funkciji f, ko gre N proti neskončno.



Slika 3.3: Rungejev fenomen: interpolacija Rungejeve funkcije $f(x) = \frac{1}{1+25x^2}$ (modra črta) na intervalu $[-1, 1]$ z interpolacijskim polinomom Čebiševa $I_N^{CH} f$ za $N+1$ točk Čebiševa (rdeča črta) za a) $N = 4$, b) $N = 8$, c) $N = 12$ in d) $N = 16$. Vozli so označeni z rdečimi pikami.

Tri zahteve iz začetka tega razdelka so v primeru aproksimacije z vrsto Čebiševa izpolnjene. Znano dejstvo je, da dosežejo Gaussove integracijske formule, kjer kot vozle uporabimo ničle ortogonalnih polinomov, visoko stopnjo natančnosti. Poleg tega je interpolacijski polinom Čebiševa $I_N^{CH} f$ (3.10), ki je definiran kot linearna kombinacija polinomov Čebiševa prve vrste stopnje kvečjemu N, za vsako funkcijo f v točkah Čebiševa zelo blizu optimalnega Lagrangeovega interpolacijskega polinoma $I_N^{OPT} f$ v maksimum normi, saj velja

$$\|f - I_N^{CH} f\| \leq \left(1 + \Lambda_N^{CH}\right) \|f - I_N^{OPT} f\|, \tag{3.12}$$

kjer je $\Lambda_N^{CH} = \mathcal{O}(\log N)$ *Lebesguova konstanta*. Prva zahteva je tako izpolnjena. Pogoj (3.8) v drugi zahtevi je izpolnjen z relacijo

$$\begin{bmatrix} 1 & 0 & -\frac{1}{2} & & & & \\ & \frac{1}{4} & 0 & -\frac{1}{4} & & & \\ & & \frac{1}{6} & 0 & -\frac{1}{6} & & \\ & & & \frac{1}{8} & 0 & -\frac{1}{8} & \\ & & & & \ddots & \ddots & \ddots \\ & & & & & \frac{1}{2N-4} & 0 \\ & & & & & & \frac{1}{2N-2} \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{N-2} \\ b_{N-1} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ \vdots \\ a_{N-1} \\ a_N \end{bmatrix}, \quad (3.13)$$

kjer so a_k koeficienti odrezane vrste Čebiševa, b_k pa koeficienti odvoda te odrezane vrste. Relacija (3.13) sledi iz formul (2.40 – 2.41). Tretja zahteva pa je izpolnjena z uporabo diskretne kosinusne transformacije.

3.3 Metode za izračun koeficientov

Spektralne koeficiente $\tilde{u}_k$ iskane numerične rešitve (3.4) problema (3.1 – 3.2) lahko določimo na različne načine, najpogosteje pa se uporabljajo naslednje tri metode:

1. *Galerkinova metoda*,
2. *Tau metoda*,
3. *kolokacijska* ali *psevdospektralna metoda*.

V vseh primerih pa nas zanima residual r, ki je določen z enačbo (3.5). Izpolnjena morata biti dva pogoja: residual naj bo karseda majhen in robni pogoji naj bodo izpolnjeni.

3.3.1 Galerkinova metoda

Osnovna ideja Galerkinove metode je v tem, da iz $N+1$ originalnih baznih funkcij $\{\phi_k\}_{k=0}^{N}$ tvorimo novo množico $N-1$ funkcij $\{G_n\}_{n=0}^{N-2}$, ki vse zadoščajo robnim pogojem. V tem primeru govorimo o *Galerkinovi bazi*, rešitev danega problema pa nato razvijemo po tej bazi. Izbira baze je seveda pomembna, konstrukcija pa ni enolična, saj več različnih izbir zadošča danim robnim pogojem. Običajno pa je, da Galerkinovo bazo izberemo tako, da se na enostaven način izraža v osnovni bazi. Za primer vzemimo homogene Dirichletove robne pogoje $u(-1) = u(1) = 0$ in osnovno bazo sestavljeno iz polinomov Čebiševa prve vrste T_k, $k = 0, 1, \ldots, N$. Z upoštevanjem lastnosti $T_k(1) = 1$ in $T_k(-1) = (-1)^k$ za vsak k zapišemo elemente Galerkinove baze G_n, $n = 0, 1, \ldots, N-2$, kot

$$G_{2\ell}(x) = T_{2\ell+2}(x) - T_0(x), \quad \ell \geq 0, \quad (3.14)$$

$$G_{2\ell+1}(x) = T_{2\ell+3}(x) - T_1(x), \quad \ell \geq 0. \quad (3.15)$$

Opazimo, da je namesto $N+1$ originalnih baznih funkcij za isti red aproksimacije potrebno upoštevati le $N-1$ Galerkinovih baznih funkcij. Razlog je v tem, da nova baza že zadošča robnim pogojem in imamo tako dva pogoja manj. V primeru, ko so osnovne bazne funkcije ortogonalni polinomi, je potrebno poudariti, da Galerkinove baze v splošnem ne sestavljajo ortogonalni polinomi.

Naj bo $M \in \mathbb{R}^{(N-1)\times(N+1)}$ transformacijska matrika, ki povezuje Galerkinovo in osnovno bazo. V primeru, ko je Galerkinova baza podana z enačbama (3.14 – 3.15) in je $\phi_k = T_k$ za vsak k, je transformacija med bazama podana z enačbo

$$G_n(x) = \sum_{k=0}^{N} M_{nk} T_k(x), \quad n = 0, 1, \ldots, N-2, \tag{3.16}$$

kjer je M_{nk} element matrike M, ki leži v n-ti vrstici in k-tem stolpcu. V primeru $N = 4$ je matrika M, ki pripada transformaciji polinomov Čebiševa prve vrste v Galerkinovo bazo za homogene robne pogoje, enaka

$$M = \begin{bmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3.17}$$

Numerično rešitev $\tilde{u}$ iščemo v razvoju po Galerkinovi bazi:

$$\tilde{u}(x) = \sum_{n=0}^{N-2} \tilde{u}_n^G G_n(x), \tag{3.18}$$

kjer so $\tilde{u}_n^G$ *Galerkinovi koeficienti*. Definiramo še matriko $L \in \mathbb{R}^{(N+1)\times(N+1)}$, ki pripada linearnemu operatorju $\mathcal{L}$, podanemu z enačbo (3.1). Če je numerična rešitev podana z odrezano vrsto (3.4), velja

$$\mathcal{L}\tilde{u}(x) = \sum_{k=0}^{N} \tilde{u}_k \mathcal{L}\phi_k(x) = \sum_{k=0}^{N} \tilde{u}_k \left(\sum_{j=0}^{N} L_{jk} \phi_j(x) \right), \tag{3.19}$$

kjer je L_{jk} element matrike L, ki leži v j-ti vrstici in k-tem stolpcu. V primeru linearnega operatorja

$$\mathcal{L}u = \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} - 4\frac{\mathrm{d}u}{\mathrm{d}x} + 4u, \tag{3.20}$$

dobimo z uporabo baze $\phi_k = T_k$ za $N = 4$ matriko

$$L = \begin{bmatrix} 4 & -4 & 4 & -12 & 32 \\ 0 & 4 & -16 & 24 & -32 \\ 0 & 0 & 4 & -24 & 48 \\ 0 & 0 & 0 & 4 & -32 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix}. \tag{3.21}$$

Enačbo (3.19) z uporabo Galerkinove baze in enačbe (3.18) zapišemo kot

$$\mathcal{L}\tilde{u}(x) = \sum_{n=0}^{N-2} \tilde{u}_n^G \sum_{j=0}^{N} \sum_{k=0}^{N} M_{jk} L_{jn} \phi_k(x). \tag{3.22}$$

Pri Galerkinovi metodi za testne funkcije vzamemo kar elemente Galerkinove baze, $\psi_i = G_i$, $i = 0, 1, \ldots, N-2$. Zahtevamo, da je residual r, definiran z enačbo (3.5), ortogonalen na kar se da veliko testnih funkcij

$$(G_i, r) = (G_i, \mathcal{L}\tilde{u} - f) = 0, \quad 0 \leq i \leq N-2. \tag{3.23}$$

Izraz $(G_i, \mathcal{L}\tilde{u})$ lahko izračunamo z uporabo enačbe (3.22), kjer elemente Galerkinove baze s pomočjo transformacijske matrike M zapišemo po osnovnih baznih funkcijah ϕ_k. Desno stran f diferencialne enačbe (3.1) prav tako zapišemo po osnovnih baznih funkcijah ϕ_k

$$f(x) = \sum_{k=0}^{N} \tilde{f}_k \phi_k(x). \tag{3.24}$$

Za uporabo Galerkinove baze v razvoju desne strani diferencialne enačbe namreč ni nobenega pravega razloga, saj funkcija f v splošnem ne zadošča robnim pogojem (3.2). Ko združimo vse skupaj ter uporabimo transformacijsko matriko M za prehod iz Galerkinove v osnovno bazo, dobimo Galerkinov sistem linearnih enačb za izračun Galerkinovih koeficientov $\tilde{u}_n^G$

$$\sum_{n=0}^{N-2} \tilde{u}_n^G (G_i, \mathcal{L}G_n) - (G_i, f) = 0, \quad 0 \leq i \leq N-2, \tag{3.25}$$

ki ga za $0 \leq i \leq N-2$ zapišemo v obliki

$$\sum_{n=0}^{N-2} \tilde{u}_n^G \sum_{j=0}^{N} \sum_{k=0}^{N} \sum_{\ell=0}^{N} M_{i\ell} M_{jk} L_{jn} (\phi_\ell(x), \phi_k(x)) = \sum_{k=0}^{N} \sum_{\ell=0}^{N} M_{i\ell} \tilde{f}_k (\phi_\ell(x), \phi_k(x)). \tag{3.26}$$

Sistem je dobro definiran, matrika koeficientov pa obrnljiva. Rešitev Galerkinovega sistema so koeficienti $\tilde{u}_n^G$ v razvoju numerične rešitve po Galerkinovi bazi. S ponovno uporabo transformacijske matrike M dobimo numerično rešitev, razvito po osnovni spektralni bazi

$$\tilde{u}(x) = \sum_{k=0}^{N} \left(\sum_{n=0}^{N-2} M_{nk} \tilde{u}_n^G \right) \phi_k(x), \tag{3.27}$$

kjer so $\tilde{u}_k$ koeficienti tega razvoja

$$\tilde{u}_k = \sum_{n=0}^{N-2} M_{nk} \tilde{u}_n^G. \tag{3.28}$$

3.3.2 Tau metoda

Pri Tau metodi se za razliko od Galerkinove metode zadovoljimo z osnovno spektralno bazo, za testne funkcije pa vzamemo prav tako elemente ϕ_i spektralne baze, npr. polinome Čebiševa prve vrste T_i. Podobno kot prej zahtevamo, da je residual r, definiran z enačbo (3.5), ortogonalen na karseda veliko testnih funkcij

$$(\phi_i, r) = (\phi_i, \mathcal{L}\tilde{u} - f) = 0, \quad 0 \leq i \leq N-2. \tag{3.29}$$

Z uporabo operatorske matrike L (3.19), enačbo (3.29) zapišemo kot sistem linearnih enačb za izračun spektralnih koeficientov $\tilde{u}_k$

$$\begin{aligned}
\sum_{k=0}^{N} \tilde{u}_k \left(\phi_i, \mathcal{L}\phi_k\right) &= \left(\phi_i, f\right), \quad i = 0, 1, \ldots, N-2, \\
\sum_{k=0}^{N} \tilde{u}_k \left(\phi_i(x), \sum_{j=0}^{N} L_{jk}\phi_j(x)\right) &= \left(\phi_i(x), \sum_{j=0}^{N} \tilde{f}_j \phi_j(x)\right), \\
\sum_{k=0}^{N}\sum_{j=0}^{N} L_{jk}\tilde{u}_k \left(\phi_i(x), \phi_j(x)\right) &= \sum_{j=0}^{N} \tilde{f}_j \left(\phi_i(x), \phi_j(x)\right), \\
\sum_{j=0}^{N} L_{jk}\tilde{u}_k &= \tilde{f}_k.
\end{aligned} \tag{3.30}$$

Pri tem so koeficienti $\tilde{f}_k$ spektralni koeficienti v razvoju desne strani f enačbe (3.1) in so podani z enačbo (3.24).

Sistem enačb (3.30) ne upošteva robnih pogojev, zato jih moramo dodati, preden ga rešimo. Pri Tau metodi so robni pogoji določeni z dodatnimi enačbami. V primeru Dirichletovih robnih pogojev $u(-1) = A$ in $u(1) = B$ imamo dodatni enačbi, ki se za splošne bazne funkcije ϕ_k glasita

$$u(-1) = \sum_{k=0}^{N} \tilde{u}_k \phi_k(-1) = A, \tag{3.31}$$

$$u(1) = \sum_{k=0}^{N} \tilde{u}_k \phi_k(1) = B. \tag{3.32}$$

Še preprostejši enačbi dobimo za izbor $\phi_k = T_k$

$$u(-1) = \sum_{k=0}^{N} (-1)^k \tilde{u}_k = A, \tag{3.33}$$

$$u(1) = \sum_{k=0}^{N} \tilde{u}_k = B. \tag{3.34}$$

V splošnem naj bo $\{g_1, g_2\}$ ortonormirana baza na množici $\{-1, 1\}$. Operator $\mathcal{B}\phi_k(x)$ razvijemo po teh funkcijah

$$\mathcal{B}\phi_k(x) = b_{1k}g_1(x) + b_{2k}g_2(x), \quad k = 0, 1, \ldots, N. \tag{3.35}$$

Robni pogoji

$$\mathcal{B}u(x) = \sum_{k=0}^{N} \tilde{u}_k \left(b_{1k}g_1(x) + b_{2k}g_2(x)\right) = 0 \tag{3.36}$$

tedaj implicirajo enačbi

$$\sum_{k=0}^{N} b_{1k}\tilde{u}_k = 0, \quad \sum_{k=0}^{N} b_{2k}\tilde{u}_k = 0. \tag{3.37}$$

Enačbi (3.31) in (3.32) za robna pogoja dodamo sistemu linearnih enačb (3.30). Tako dobljeni (relaksirani) sistem je dobro definiran, matrika koeficientov sistema pa obrnljiva. Spektralne koeficiente $\tilde{u}_k$ v razvoju numerične rešitve $\tilde{u}$ dobimo kot rešitev tega sistema.

3.3.3 Kolokacijska metoda

Kolokacijska metoda se nekoliko razlikuje od prejšnjih dveh. Podobno kot pri Tau metodi, numerično rešitev razvijemo po osnovni spektralni bazi, testne funkcije v tem primeru pa v nasprotju z obema gornjima metodama niso več elementi spektralne baze, pač pa so izbrane tako, da so enake 0 v eni kolokacijski točki. Primerna izbira je npr. Diracova delta funkcija $\delta(\,\cdot\, - x_n)$, ki je definirana z $\delta(x) = \begin{cases} 1, & x = 0, \\ 0, & \text{sicer.} \end{cases}$ Residualni pogoj (3.6), ki zahteva, da je residual r, definiran z enačbo (3.5), ortogonalen na karseda veliko testnih funkcij, je ekvivalenten pogoju, da je residual enak 0 v karseda veliko točkah x_n

$$\mathcal{L}\tilde{u}(x_n) = f(x_n), \quad 0 \leq n \leq N. \tag{3.38}$$

Z uporabo operatorske matrike L (3.19), enačbo (3.38) zapišemo kot sistem linearnih enačb za izračun spektralnih koeficientov $\tilde{u}_k$

$$\sum_{k=0}^{N}\sum_{j=0}^{N} L_{jk}\tilde{u}_k\phi_j(x_n) = f(x_n), \quad 0 \leq n \leq N. \tag{3.39}$$

Pri tem so x_n *kolokacijske točke* na intervalu $[-1, 1]$

$$-1 = x_0 < x_1 < x_2 < \cdots < x_{N-1} < x_N = 1. \tag{3.40}$$

Podobno kot pri Tau metodi, je potrebno dodati robne pogoje, ki so definirani z enačbami (3.31) in (3.32). Sistem (3.39) relaksiramo tako, da prvo in zadnjo enačbo nadomestimo z enačbama za robna pogoja, kar je ekvivalentno temu, da kolokacijskih enačb ne zapišemo v robnih točkah. Rešitev kolokacijskega sistema, ki je dobro definiran in obrnljiv, so spektralni koeficienti $\tilde{u}_k$ v razvoju numerične rešitve $\tilde{u}$.

3.4 Konstrukcija spektralnih metod

Spektralne metode lahko konstruiramo z uporabo različnih baznih funkcij. Na tem mestu se omejimo na trigonometrične funkcije, oz. Fourierovo vrsto za periodične probleme, iz katerih dobimo *razred Fourierovih spektralnih metod* ter na polinome Čebiševa prve vrste, oz. vrsto Čebiševa za neperiodične probleme, iz katerih dobimo *razred spektralnih metod Čebiševa*.

Spektralne metode za reševanje diferencialnih enačb lahko konstruiramo tako, da funkcije odvajamo v *fizičnem* ali *faznem prostoru*. V obeh primerih konstruiramo operatorske *matrike odvodov* (ang. *differentiation matrix*), ki jih množimo bodisi z vektorjem funkcijskih vrednosti (fizični prostor), bodisi z vektorjem spektralnih koeficientov funkcije (fazni prostor). V tem delu se bomo omejili na konstrukcijo spektralnih metod v faznem prostoru. Višje odvode opišemo s potenciranjem matrike odvodov.

Poleg matrike odvodov potrebujemo tudi *multiplikacijske matrike* (ang. *multiplication matrix*), ki opisujejo transformacijo spektralnih koeficientov iskane funkcije v koeficiente produkta iskane funkcije z dano funkcijo. Za razliko od matrike odvodov, ki je določena samo z izbiro baznih funkcij, je multiplikacijska matrika določena tako z izbiro baznih funkcij kot s spektralnimi koeficienti v razvoju dane funkcije po izbrani bazi.

3.4.1 Fourierove spektralne metode

Vsako analitično funkcijo f lahko razvijemo v Fourierovo vrsto na intervalu $[-1, 1]$, kjer je Fourierova vrsta (2.10) po pretvorbi iz intervala $[-\pi, \pi]$ na interval $[-1, 1]$

$$f(x) = a_0 + \sum_{n=0}^{\infty} \big(a_n \cos(n\pi x) + b_n \sin(n\pi x)\big). \tag{3.41}$$

Fourierove koeficiente izračunamo s transformiranimi formulami (2.11) in (2.12)

$$a_0 = \frac{1}{2}\int_{-1}^{1} f(x)\,\mathrm{d}x, \tag{3.42}$$

$$a_n = \int_{-1}^{1} f(x)\cos(n\pi x)\,\mathrm{d}x, \tag{3.43}$$

$$b_n = \int_{-1}^{1} f(x)\sin(n\pi x)\,\mathrm{d}x. \tag{3.44}$$

To naredimo učinkovito s hitro Fourierovo transformacijo (FFT) (glej članek J. W. Cooley in J. W. Tukey [15]), ki nam omogoča izračun prvih $2N + 1$ Fourierovih koeficientov z $\mathcal{O}(N \log N)$ operacijami.

Fourierove spektralne metode konstruiramo za reševanje periodičnih robnih problemov. Pri tem rešitev modelnega linearnega dvotočkovnega robnega problema (3.1 – 3.2) aproksimiramo z odrezano Fourierovo vrsto

$$u(x) \approx u_N(x) = a_0 + \sum_{n=1}^{N} \big(a_n \cos(n\pi x) + b_n \sin(n\pi x)\big), \tag{3.45}$$

kjer je N odrezno število vrste. Pripadajoče odvode aproksimiramo z odvodi odrezane Fourierove vrste (3.45)

$$\begin{aligned} u'(x) \approx u'_N(x) &= \sum_{n=1}^{N} \big(- n\pi a_n \sin(n\pi x) + n\pi b_n \cos(n\pi x)\big) \\ &= \sum_{n=1}^{N} \Big(\tilde{a}_n \cos(n\pi x) + \tilde{b}_n \sin(n\pi x)\Big), \qquad (3.46) \\ u''(x) \approx u''_N(x) &= \sum_{n=1}^{N} \big(- n^2\pi^2 a_n \cos(n\pi x) - n^2\pi^2 b_n \sin(n\pi x)\big) \\ &= \sum_{n=1}^{N} \Big(\tilde{\tilde{a}}_n \cos(n\pi x) + \tilde{\tilde{b}}_n \sin(n\pi x)\Big). \qquad (3.47) \end{aligned}$$

Transformacijo koeficientov a_n in b_n (3.42 – 3.44) v koeficiente $\tilde{a}_n$ in $\tilde{b}_n$ (3.46) opišemo z matriko odvodov $D \in \mathbb{R}^{(2N+1)\times(2N+1)}$, saj so zveze med koeficienti linearne. Elementi d_{ij} matrike odvodov so podani s predpisom

$$d_{ij} = \begin{cases} i\pi, & 2 \le i \le N+1,\ j = N+i, \\ -i\pi, & N+2 \le i \le 2N+1,\ j = i-N, \\ 0, & \text{sicer.} \end{cases} \tag{3.48}$$

Transformacijo koeficientov a_n in b_n v koeficiente $\tilde{\tilde{a}}_n$ in $\tilde{\tilde{b}}_n$ (3.47) pa opišemo z matriko D^2. Velja

$$\tilde{\mathbf{v}} = D\,\mathbf{v} \quad \text{in} \quad \tilde{\tilde{\mathbf{v}}} = D^2\,\mathbf{v}, \tag{3.49}$$

kjer je $\mathbf{v} = (a_0, a_1, \ldots, a_N, b_1, \ldots, b_N)^T$, $\tilde{\mathbf{v}} = (\tilde{a}_0, \tilde{a}_1, \ldots, \tilde{a}_N, \tilde{b}_1, \ldots, \tilde{b}_N)^T$ in $\tilde{\tilde{\mathbf{v}}} = (\tilde{\tilde{a}}_0, \tilde{\tilde{a}}_1, \ldots, \tilde{\tilde{a}}_N, \tilde{\tilde{b}}_1, \ldots, \tilde{\tilde{b}}_N)^T$. V primeru $N = 2$ sta transformacijski matriki D in D^2 enaki

$$D = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \pi & 0 \\ 0 & 0 & 0 & 0 & 2\pi \\ 0 & -\pi & 0 & 0 & 0 \\ 0 & 0 & -2\pi & 0 & 0 \end{bmatrix}, \; D^2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -\pi^2 & 0 & 0 & 0 \\ 0 & 0 & -4\pi^2 & 0 & 0 \\ 0 & 0 & 0 & -\pi^2 & 0 \\ 0 & 0 & 0 & 0 & -4\pi^2 \end{bmatrix}. \tag{3.50}$$

Naj bo g neka znana funkcija, ki jo lahko razvijemo v Fourierovo vrsto, in naj bo g_N njena odrezana Fourierova vrsta

$$g(x) \approx g_N(x) = g_0 + \sum_{n=1}^{N} \big(g_n \cos(n\pi x) + h_n \sin(n\pi x)\big). \tag{3.51}$$

Produkt odrezanih Fourierovih vrst za funkcijo g (3.51) in iskano rešitev u (3.45) aproksimiramo z odrezano Fourierovo vrsto

$$g_N(x)\, u_N(x) \approx \hat{a}_0 + \sum_{n=1}^{N} \Big(\hat{a}_n \cos(n\pi x) + \hat{b}_n \sin(n\pi x)\Big). \tag{3.52}$$

Za transformacijo koeficientov a_n in b_n (3.42 – 3.44) v koeficiente $\hat{a}_n$ in $\hat{b}_n$ (3.52) potrebujemo operatorsko multiplikacijsko matriko, saj je zveza med koeficienti ponovno linearna, kar pomeni, da matrika $F \in \mathbb{R}^{(2N+1)\times(2N+1)}$ za množenje odrezanih Fourierovih vrst g_N in u_N obstaja in je odvisna od koeficientov g_n in h_n (3.51). Opazimo, da je produkt odrezan pri istem odreznem številu N kot oba faktorja. Velja

$$\hat{\mathbf{v}} = F\, \mathbf{v}, \tag{3.53}$$

kjer je $\hat{\mathbf{v}} = (\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_N, \hat{b}_1, \ldots, \hat{b}_N)^T$.

Elementi matrike F so določeni z uporabo adicijskih izrekov za trigonometrične funkcije. Matrika F je bločna

$$F = \begin{bmatrix} g_0 & \frac{1}{2}\mathbf{g} & \frac{1}{2}\mathbf{h} \\ \mathbf{g}^T & F_1 & F_2 \\ \mathbf{h}^T & F_3 & F_4 \end{bmatrix}, \tag{3.54}$$

kjer sta $\mathbf{g} = (g_1, \ldots, g_N)$ in $\mathbf{h} = (h_1, \ldots, h_N)$ vektorja Fourierovih koeficientov dane funkcije g_N (3.51). Matrika F brez prve vrstice in prvega stolpca pa je simetrična, saj velja $F_1^T = F_1$, $F_3 = F_2^T$ in $F_4^T = F_4$. Bloki $F_1, F_2, F_3, F_4 \in \mathbb{R}^{N\times N}$ v primeru $N = 4$ so

$$\begin{aligned}
F_1 &= \frac{1}{2}\begin{bmatrix} 2g_0+g_2 & g_1+g_3 & g_2+g_4 & g_3 \\ g_1+g_3 & 2g_0+g_4 & g_1 & g_2 \\ g_2+g_4 & g_1 & 2g_0 & g_1 \\ g_3 & g_2 & g_1 & 2g_0 \end{bmatrix}, \\
F_2 &= \frac{1}{2}\begin{bmatrix} h_2 & h_1+h_3 & h_2+h_4 & h_3 \\ -h_1+h_3 & h_4 & h_1 & h_2 \\ -h_2+h_4 & -h_1 & 0 & h_1 \\ -h_3 & -h_2 & -h_1 & 0 \end{bmatrix}, \\
F_3 &= F_2^T, \\
F_4 &= \frac{1}{2}\begin{bmatrix} 2g_0-g_2 & g_1-g_3 & g_2-g_4 & g_3 \\ g_1-g_3 & 2g_0-g_4 & g_1 & g_2 \\ g_2-g_4 & g_1 & 2g_0 & g_1 \\ g_3 & g_2 & g_1 & 2g_0 \end{bmatrix}.
\end{aligned}$$

Bloka F_2 in F_3 sta v primeru, ko je funkcija g soda, ničelni matriki, v primeru, ko je funkcija g liha, pa sta bloka F_1 in F_4 ničelni matriki. Kot zgled zapišimo matriko F za množenje z znano (liho) funkcijo $g(x) = x$ za $N = 3$

$$F = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.3183 & -0.1592 & 0.1061 \\ 0 & 0 & 0 & 0 & -0.1592 & 0.4244 & -0.1592 \\ 0 & 0 & 0 & 0 & -0.2122 & 0 & 0.3183 \\ 0 & 0 & 0 & 0 & 0.1592 & -0.3183 & 0 \\ 0.6366 & -0.1592 & -0.2122 & 0.1592 & 0 & 0 & 0 \\ -0.3183 & 0.4244 & 0 & -0.3183 & 0 & 0 & 0 \\ 0.2122 & -0.1592 & 0.3183 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

3.4.2 Spektralne metode Čebiševa

Vsako analitično funkcijo f lahko razvijemo v vrsto Čebiševa (2.55) na intervalu $[-1, 1]$, kjer koeficiente Čebiševa izračunamo s formulami (2.56) in (2.57). To naredimo učinkovito z diskretno kosinusno transformacijo (DCT), ki nam omogoča izračun prvih $N+1$ koeficientov z $\mathcal{O}(N \log N)$ operacijami.

Spektralne metode Čebiševa konstruiramo za reševanje neperiodičnih robnih problemov. Pri tem rešitev modelnega linearnega dvotočkovnega robnega problema (3.1 – 3.2) aproksimiramo z odrezano vrsto Čebiševa (2.58)

$$u(x) \approx u_N(x) = \sum_{n=0}^{N} a_n T_n(x), \tag{3.55}$$

kjer je N odrezno število vrste. Pripadajoče odvode aproksimiramo z odvodi odrezane vrste Čebiševa (3.55)

$$u'(x) \approx u'_N(x) = \sum_{n=0}^{N-1} b_n T_n(x), \tag{3.56}$$

$$u''(x) \approx u''_N(x) = \sum_{n=0}^{N-2} c_n T_n(x). \tag{3.57}$$

Iščemo zvezo med koeficienti a_n, b_n in c_n. Zveza med koeficienti a_n (3.55) in b_n (3.56) je podana z relacijo (3.13). Velja

$$\mathbf{b} = D\,\mathbf{a}, \tag{3.58}$$

kjer je $\mathbf{a} = (a_0, a_1, \ldots, a_N)^T$, $\mathbf{b} = (b_0, b_1, \ldots, b_N)^T$ in $b_N = 0$. Pri tem je $D \in \mathbb{R}^{(2N+1)\times(2N+1)}$ matrika odvodov, katere elementi d_{ij} so podani s predpisom

$$d_{ij} = \begin{cases} j-1, & i = 1 \text{ in } j \text{ sod}, \\ 2j-2, & 2 \le i \le N \text{ in } j = i+1, i+3, \ldots, N+1, \\ 0, & \text{sicer}. \end{cases} \tag{3.59}$$

Zvezo med koeficienti a_n in c_n (3.57) pa opišemo z matriko D^2. Velja

$$\mathbf{c} = D^2\,\mathbf{a}, \tag{3.60}$$

kjer je $\mathbf{c} = (c_0, c_1, \ldots, c_N)^T$ in $c_{N-1} = c_N = 0$. V primeru $N = 4$ sta transformacijski matriki D in D^2 enaki

$$D = \begin{bmatrix} 0 & 1 & 0 & 3 & 0 \\ 0 & 0 & 4 & 0 & 8 \\ 0 & 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & 8 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad D^2 = \begin{bmatrix} 0 & 0 & 4 & 0 & 32 \\ 0 & 0 & 0 & 24 & 0 \\ 0 & 0 & 0 & 0 & 48 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3.61}$$

Naj bo g neka znana funkcija, ki jo lahko razvijemo v vrsto Čebiševa, in naj bo g_N njena odrezana vrsta Čebiševa

$$g(x) \approx g_N(x) = \sum_{n=0}^{N} g_n T_n(x). \tag{3.62}$$

Produkt odrezanih vrst Čebiševa za funkcijo g (3.62) in iskano rešitev u (3.55) aproksimiramo z odrezano vrsto Čebiševa

$$g_N(x)\,u_N(x) \approx \sum_{n=0}^{N} \hat{a}_n T_n(x). \tag{3.63}$$

Za transformacijo koeficientov a_n (3.55) v koeficiente $\hat{a}_n$ (3.63) potrebujemo operatorsko multiplikacijsko matriko, saj je zveza med koeficienti ponovno linearna, kar pomeni, da matrika $F \in \mathbb{R}^{(2N+1)\times(2N+1)}$ za množenje odrezanih vrst Čebiševa g_N in u_N obstaja in je odvisna od koeficientov g_n (3.62). Opazimo, da je produkt odrezan pri istem odreznem številu N kot oba faktorja. Velja

$$\hat{\mathbf{a}} = F\,\mathbf{a}, \tag{3.64}$$

kjer je $\hat{\mathbf{a}} = (\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_N)^T$.

Elementi matrike F so določeni z zvezami (2.42) in (2.43). Matrika F brez prve vrstice in prvega stolpca je simetrična

$$F = \begin{bmatrix} g_0 & \frac{1}{2}g_1 & \frac{1}{2}g_2 & \cdots & \frac{1}{2}g_{N-1} & \frac{1}{2}g_N \\ g_1 & g_0 + \frac{1}{2}g_2 & \frac{1}{2}g_1 + \frac{1}{2}g_3 & \cdots & \frac{1}{2}g_{N-2} + \frac{1}{2}g_N & \frac{1}{2}g_{N-1} \\ g_2 & \frac{1}{2}g_1 + \frac{1}{2}g_3 & \frac{1}{2}g_0 + \frac{1}{2}g_4 & \cdots & \frac{1}{2}g_{N-3} & \frac{1}{2}g_{N-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ g_{N-1} & \frac{1}{2}g_{N-2} + \frac{1}{2}g_N & \frac{1}{2}g_{N-3} & \cdots & g_0 & \frac{1}{2}g_1 \\ g_N & \frac{1}{2}g_{N-1} & \frac{1}{2}g_{N-2} & \cdots & \frac{1}{2}g_1 & g_0 \end{bmatrix}. \tag{3.65}$$

Kot zgled zapišimo matriko F za množenje z znano (sodo) funkcijo $g(x) = \cos x$ za $N = 4$

$$F = \begin{bmatrix} 0.7652 & 0 & -0.1149 & 0 & 0.0025 \\ 0 & 0.6503 & 0 & -0.1124 & 0 \\ -0.2298 & 0 & 0.7652 & 0 & -0.1149 \\ 0 & -0.1124 & 0 & 0.7652 & 0 \\ 0.0050 & 0 & -0.1149 & 0 & 0.7652 \end{bmatrix}.$$

3.4.3 Clenshaw-Curtisova kvadraturna formula

Pri računanju koeficientov Fourierove vrste (2.11 – 2.12) in vrste Čebiševa (2.56 – 2.57), je potrebno, največkrat numerično, izračunati integrale oblike

$$I = \int_{-1}^{1} f(x)\ w(x)\,\mathrm{d}x,$$

kjer je $w(x)$ nenegativna utež. To lahko storimo z uporabo kvadraturnih formul Newton-Cotesovega (periodični problemi z ekvidistantnimi delilnimi vozli) ali Gaussovega tipa (neperiodični problemi z delilnimi vozli Čebiševa). Oba razreda integracijskih metod sta v literaturi široko obdelana, npr. v M. Abramowitz in I. A. Stegun [1], W. Gautschi [22] ter E. Isaacson in H. B. Keller [36]. Učinkovitejše so metode Gaussovega tipa, kjer so prosti parametri poleg *integracijskih uteži* w_k tudi *integracijski vozli* x_k. Gaussovo integracijsko pravilo na intervalu $[-1, 1]$ zapišemo kot

$$\int_{-1}^{1} f(x)\ w(x)\,\mathrm{d}x \approx \sum_{k=0}^{N} w_k f(x_k). \tag{3.66}$$

Integracijska formula je točna za polinome določenega reda, ki je odvisen od izbire kolokacijskih točk (vozlov). Za Gaussove točke (vsi vozli ležijo v notranjosti intervala) je formula točna za polinome reda $2N + 1$, za Gauss-Radaujeve točke (en vozel je fiksno določen kot levo oz. desno krajišče intervala) za polinome reda $2N$, za Gauss-Lobattove točke (dva vozla sta fiksno določena kot obe krajišči) pa za polinome reda $2N-1$. Gaussove kvadraturne formule so najboljša možna izbira, če gledamo red aproksimacije, vendar so tudi druge formule koristne, predvsem z vidika upoštevanja robnih pogojev. V nadaljevanju bomo tako uporabljali Gauss-Lobattove integracijske formule. Vozli kvadraturnih formul Gaussovega tipa so običajno ničle ortogonalnih polinomov. Integracijska formula je tako poimenovana po tipu vozlov ter po ortogonalnih polinomih, katerih ničle so vozli, npr. Legendre-Gauss, Čebišev-Gauss-Lobatto. V primeru Čebišev-Gauss-Lobattovih kvadraturnih formul velja

$$x_k = -\cos\frac{\pi k}{N}, \quad k = 0, 1, \ldots, N, \tag{3.67}$$

$$w_k = \frac{\pi}{N}, \quad w_0 = w_N = \frac{\pi}{2N}, \quad k = 1, 2, \ldots, N-1. \tag{3.68}$$

Vozle in uteži Gaussovih kvadraturnih formul izračunamo v $\mathcal{O}(N^2)$ operacijah, kjer rešimo tridiagonalni problem lastnih vrednosti, kar je opisano npr. v G. H. Golub in J. H. Welsch [27].

Poleg standardnih integracijskih metod, lahko za izračun spektralnih koeficientov Fourierove vrste ali vrste Čebiševa uporabimo tudi razred spektralnih Clenshaw-Curtisovih kvadraturnih formul, kjer integracijske vozle in uteži z uporabo FFT izračunamo z $\mathcal{O}(N \log N)$ operacijami. Ideja, ki sta jo uporabila avtorja C. W. Clenshaw in A. R. Curtis v članku [14], je, da optimalne vozle nadomestimo s točkami Čebiševa (3.11), tj. z ekstremnimi vrednostmi polinomov Čebiševa prve vrste.

Funkcijo f v tako izbranih točkah interpoliramo s polinomom p stopnje kvečjemu N. Iščemo vrednost integrala

$$I_N = \int_{-1}^{1} p(x)\,\mathrm{d}x = \int_{0}^{\pi} p(\cos\theta)\sin\theta\,\mathrm{d}\theta, \tag{3.69}$$

kjer uvedemo novo spremenljivko $x = \cos\theta$. Funkcijo $F(\theta) = p(\cos\theta)$ razvijemo v kosinusno Fourierovo vrsto

$$F(\theta) = p(\cos\theta) = \sum_{k=0}^{N} a_k \cos k\theta,$$

kar je ekvivalentno temu, da polinom p razvijemo v vrsto Čebiševa. Koeficiente a_k izračunamo z uporabo FFT, natančneje z uporabo DCT. Clenshaw-Curtisova kvadraturna formula se glasi

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx \int_{-1}^{1} F(\theta)\sin\theta\,\mathrm{d}\theta = \sum_{\substack{k=0 \\ k \text{ sod}}}^{N} \frac{2a_k}{1-k^2}. \tag{3.70}$$

Koeficienta pri $k = 0$ in $k = N$ sta polovična.

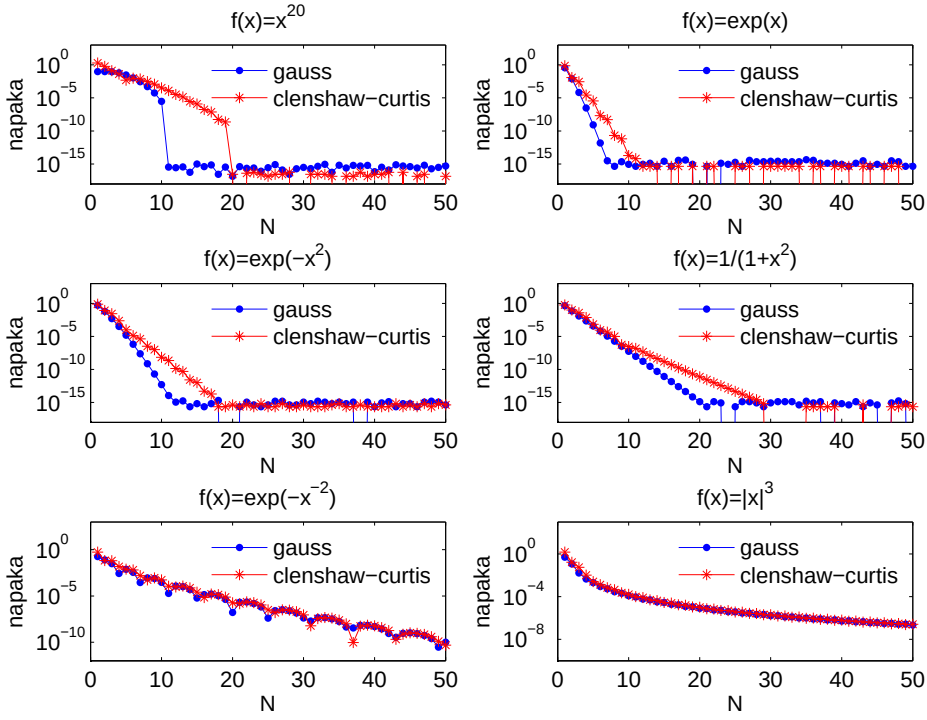
Ocena napake kvadraturne formule za m-krat zvezno odvedljive funkcije $f \in \mathcal{C}^\infty$ je primerljiva z oceno napake za Gaussove kvadraturne formule

$$|I - I_N| \leq \frac{64\,\|f^{(m)}\|_\infty}{15\pi m(2N+1-m)^m}. \tag{3.71}$$

Primerjava med obema metodama je obširno opisana v L. N. Trefethen [57] in [58], od koder je vzet spodnji zgled.

Na sliki 3.4 je prikazano padanje maksimalne absolutne vrednosti napake v odvisnosti od odreznega števila N za izračun integrala $\int_{-1}^{1} f(x)\,\mathrm{d}x$ za funkcije $f(x) \in \left\{x^{20}, \mathrm{e}^x, \mathrm{e}^{-x^2}, \frac{1}{1+x^2}, \mathrm{e}^{-x^{-2}}, |x|^3\right\}$, ki so različnih stopenj gladkosti. Integrali so izračunani z dvema numeričnima kvadraturnima formulama: Clenshaw-Curtisovo spektralno metodo, in Čebišev-Gaussovo metodo. Predvsem v primeru funkcij, ki niso gladke, opazimo, da sta metodi

povsem primerljivi. Za preostale funkcije je iz slike razvidno, da maksimalna absolutna napaka s Čebišev-Gaussovo kvadraturno formulo doseže strojno natančnost za manjše vrednosti N kot napaka s Clenshaw-Curtisovo formulo. Na prvi sliki levo zgoraj za primer polinoma stopnje 20 vidimo razliko v redu aproksimacije, saj je Čebišev-Gaussova kvadraturna formula reda $2N+1$ in zato dobimo točno vrednost integrala že za $N = 11$, Clenshaw-Curtisova formula pa je reda $N+1$ in zato dobimo točno vrednost integrala šele za $N = 20$. Za manjše vrednosti N sta metodi dokaj primerljivi.



Slika 3.4: Primerjava maksimalnih absolutnih vrednosti napake pri izračunu integrala $\int_{-1}^{1} f(x)\,\mathrm{d}x$ v odvisnosti od N za dve integracijski metodi: Clenshaw-Curtisovo spektralno metodo (rdeče zvezde in črta), in Čebišev-Gaussovo metodo (modre pike in črta) ter 6 funkcij padajoče gladkosti.

3.5 Numerični primeri

Konstrukcijo spektralnih metod bomo sklenili z dvema zgledoma uporabe za modelna dvotočkovna robna problema (3.1) s homogenimi Dirichletovimi robnimi pogoji (3.2), ki imata konstantne oz. nekonstantne koeficiente. Oba rešimo s spektralnimi metodami Čebiševa, kjer spektralne koeficiente izračunamo z uporabo kolokacije, Galerkinove ter Tau metode. Rezultate,

ki so dobljeni s temi tremi metodami, primerjamo z rezultati, ki so dobljeni z metodami končnih razlik drugega in četrtega reda. Za delilne (kolokacijske) točke intervala $[-1,1]$ vzamemo točke Čebiševa (3.11).

Primer 3.1 Kot prvi modelni problem vzemimo primer, ki ga predlaga P. Grandclément v članku [30]. Rešujemo enačbo

$$y'' - 4y' + 4y = \mathrm{e}^x - \frac{4\mathrm{e}}{1+\mathrm{e}^2}, \quad -1 \le x \le 1,$$

skupaj s homogenimi Dirichletovimi robnimi pogoji

$$y(-1) = y(1) = 0,$$

ki ima enolično analitično rešitev

$$y(x) = \mathrm{e}^x - \frac{\sinh(1)}{\sinh(2)}\mathrm{e}^{2x} - \frac{\mathrm{e}}{1+\mathrm{e}^2}.$$

Problem, ki ga rešujemo, je linearna DE drugega reda s konstantnimi koeficienti. Opazimo, da rešitev ni polinomska. Na levi strani imamo linearni operator $\mathcal{L}$, ki je definiran v (3.20). Matriko L, ki mu pripada, dobimo kot linearno kombinacijo matrik D in D^2 (3.61) ter identitete. V primeru, ko je $N = 4$, dobimo matriko, ki je podana z enačbo (3.21)

$$\begin{aligned} L &= D^2 - 4D + 4I \\ &= \begin{bmatrix} 0 & 0 & 4 & 0 & 32 \\ 0 & 0 & 0 & 24 & 0 \\ 0 & 0 & 0 & 0 & 48 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} - 4 \begin{bmatrix} 0 & 1 & 0 & 3 & 0 \\ 0 & 0 & 4 & 0 & 8 \\ 0 & 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & 8 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} + 4 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 4 & -4 & 4 & -12 & 32 \\ 0 & 4 & -16 & 24 & -32 \\ 0 & 0 & 4 & -24 & 48 \\ 0 & 0 & 0 & 4 & -32 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix}. \end{aligned}$$

Numerično rešitev y_{num} aproksimiramo z odrezano vrsto Čebiševa

$$y_{num}(x) = \sum_{n=0}^{N} \tilde{u}_n T_n(x).$$

Galerkinova baza je podana z enačbami

$$\begin{aligned} G_0(x) &= T_2(x) - T_0(x) = 2x^2 - 2, \\ G_1(x) &= T_3(x) - T_1(x) = 4x^3 - 4x, \\ G_2(x) &= T_4(x) - T_0(x) = 8x^4 - 8x^2, \end{aligned}$$

kjer vse funkcije G_n, $n = 0, 1, 2$, zadoščajo homogenim robnim pogojem $G_n(\pm 1) = 0$. Transformacijska matrika M med spektralno in Galerkinovo bazo je za primer $N = 4$ podana z (3.17).

Galerkinov sistem za izračun Galerkinovih koeficientov je podan z

$$M\ S\ L\ M^T\ \tilde{\mathbf{u}}^G = M\ S\ \mathbf{v},$$

kjer je $\mathbf{v}$ vektor spektralnih koeficientov v razvoju funkcije desne strani f v vrsto Čebiševa

$$\mathbf{v} = \left(\tilde{f}_0, \tilde{f}_1, \ldots, \tilde{f}_N\right)^T,$$

$\tilde{\mathbf{u}}^G$ vektor Galerkinovih koeficientov

$$\tilde{\mathbf{u}}^G = \left(\tilde{u}_0^G, \tilde{u}_1^G, \ldots, \tilde{u}_{N-2}^G\right)^T,$$

matrika S pa diagonalna matrika kvadratov norm polinomov Čebiševa prve vrste z elementi

$$S_{ij} = \begin{cases} \pi, & i = j = 0, \\ \frac{\pi}{2}, & i = j > 0, \\ 0, & i \neq j. \end{cases}$$

Galerkinov sistem je v primeru $N = 4$

$$\begin{bmatrix} 2\pi & -4\pi & -4\pi \\ 8\pi & -8\pi & 0 \\ 0 & 8\pi & -26\pi \end{bmatrix} \cdot \begin{bmatrix} \tilde{u}_0^G \\ \tilde{u}_1^G \\ \tilde{u}_2^G \end{bmatrix} = \begin{bmatrix} 0.521 \\ -1.705 \\ 0.103 \end{bmatrix}.$$

Koeficiente $\tilde{\mathbf{u}} = (\tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_N)^T$ v razvoju po spektralni bazi dobimo z množenjem vektorja $\tilde{\mathbf{u}}^G$ s transformacijsko matriko M

$$\tilde{\mathbf{u}} = M\ \tilde{\mathbf{u}}^G.$$

V konkretnem primeru za $N = 4$ dobimo Galerkinove koeficiente

$$\tilde{u}_0^G \simeq -0.160,\ \tilde{u}_1^G \simeq -0.092 \text{ in } \tilde{u}_2^G \simeq -0.029$$

ter spektralne koeficiente

$$\tilde{u}_0 \simeq 0.189,\ \tilde{u}_1 \simeq 0.092,\ \tilde{u}_2 \simeq -0.160,\ \tilde{u}_3 \simeq -0.092 \text{ in } \tilde{u}_4 \simeq -0.029.$$

Sistem za izračun spektralnih koeficientov $\tilde{u}_n$ po Tau metodi, je podan z enačbo

$$\tilde{L}\ \tilde{\mathbf{u}} = \tilde{\mathbf{v}},$$

kjer je vektor $\tilde{\mathbf{u}}$ podan zgoraj, matrika $\tilde{L}$ in vektor $\tilde{\mathbf{v}}$ pa sta določena tako, da upoštevamo robne pogoje. Matriko $\tilde{L}$ tako dobimo iz L, ko zadnji dve vrstici nadomestimo z vrsticama vrednosti polinomov Čebiševa prve vrste v robnih točkah ± 1 ($T_n(1) = 1$ in $T_n(-1) = (-1)^n$ za vsak $n \geq 0$), vektor

$\tilde{\mathbf{v}}$ pa iz $\mathbf{v}$, ko zadnja dva elementa nadomestimo z Dirichletovima robnima pogojema. Ker imamo homogene robne pogoje, sta zadnja elementa enaka 0. V primeru $N = 4$ je sistem po Tau metodi

$$\begin{bmatrix} 4 & -4 & 4 & -12 & 32 \\ 0 & 4 & -16 & 24 & -32 \\ 0 & 0 & 4 & -24 & 48 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \tilde{u}_0 \\ \tilde{u}_1 \\ \tilde{u}_2 \\ \tilde{u}_3 \\ \tilde{u}_4 \end{bmatrix} = \begin{bmatrix} -0.03 \\ 1.13 \\ 0.27 \\ 0 \\ 0 \end{bmatrix}.$$

V konkretnem primeru za $N = 4$ dobimo spektralne koeficiente

$$\tilde{u}_0 \simeq 0.146,\ \tilde{u}_1 \simeq 0.079,\ \tilde{u}_2 \simeq -0.122,\ \tilde{u}_3 \simeq -0.079 \text{ in } \tilde{u}_4 \simeq -0.024.$$

Za kolokacijski sistem potrebujemo kolokacijsko matriko, tj. matriko vrednosti polinomov Čebiševa prve vrste v kolokacijskih točkah x_i (3.11). Elementi kolokacijske matrike $C = [c_{ij}]_{i,j}$ so podani kot

$$c_{ij} = T_j(x_i). \tag{3.72}$$

V primeru $N = 4$ so kolokacijski vozli: $x_0 = -1$, $x_1 = -\frac{\sqrt{2}}{2}$, $x_2 = 0$, $x_3 = \frac{\sqrt{2}}{2}$ in $x_4 = 1$, kolokacijska matrika pa

$$C = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 \\ 1 & -\frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} & -1 \\ 1 & 0 & -1 & 0 & 1 \\ 1 & \frac{\sqrt{2}}{2} & 0 & -\frac{\sqrt{2}}{2} & -1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Sistem za izračun spektralnih koeficientov $\tilde{u}_n$ je podan z enačbo

$$\hat{L}\ \tilde{\mathbf{u}} = \hat{\mathbf{v}},$$

kjer je vektor $\tilde{\mathbf{u}}$ podan zgoraj, matrika $\hat{L}$ in vektor $\hat{\mathbf{v}}$ pa sta določena tako, da upoštevamo robne pogoje. Matriko $\hat{L}$ dobimo tako, da v produktu $C\ L$ prvo in zadnjo vrstico nadomestimo s prvo in zadnjo vrstico matrike C, vektor $\hat{\mathbf{v}}$ pa tako, da v produktu $C\ \mathbf{v}$ prvi in zadnji element nadomestimo z Dirichletovima robnima pogojema. Ker imamo homogene robne pogoje, sta prvi in zadnji element enaka 0. V primeru $N = 4$ je kolokacijski sistem

$$\begin{bmatrix} 1 & -1 & 1 & -1 & 1 \\ 4 & -6.83 & 15.3 & -26.1 & 28 \\ 4 & -4 & 0 & 12 & -12 \\ 4 & -1.17 & -7.31 & 2.14 & 28 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \tilde{u}_0 \\ \tilde{u}_1 \\ \tilde{u}_2 \\ \tilde{u}_3 \\ \tilde{u}_4 \end{bmatrix} = \begin{bmatrix} 0 \\ -0.80 \\ -0.30 \\ 0.73 \\ 0 \end{bmatrix}.$$

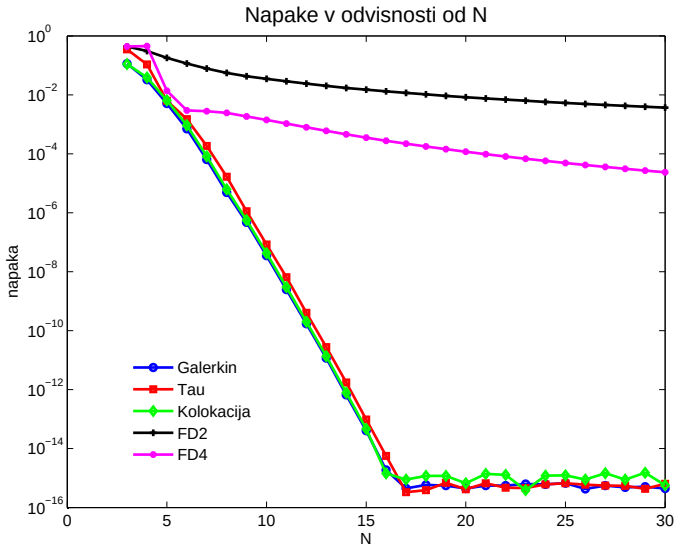
V konkretnem primeru za $N = 4$ dobimo spektralne koeficiente

$$\tilde{u}_0 \simeq 0.188,\ \tilde{u}_1 \simeq 0.089,\ \tilde{u}_2 \simeq -0.157,\ \tilde{u}_3 \simeq -0.089 \text{ in } \tilde{u}_4 \simeq -0.031.$$

Vrednosti numeričnih rešitev v kolokacijskih točkah x_i (3.11) dobimo za vse tri obravnavane metode tako, da kolokacijsko matriko C pomnožimo z vektorjem spektralnih koeficientov $\tilde{\mathbf{u}}$

$$\mathbf{y}_{num} = C\ \tilde{\mathbf{u}}.$$

Tu je $\mathbf{y}_{num}$ vektor rešitev, ki je dobljen z eno izmed spektralnih metod. V našem primeru dobimo $\mathbf{y}_{gal}$ za Galerkinovo, $\mathbf{y}_{tau}$ za Tau metodo ter $\mathbf{y}_{col}$ za metodo kolokacije. Tako dobljene rešitve lahko med seboj primerjamo.



Slika 3.5: Primerjava maksimalne absolutne vrednosti napake v odvisnosti od N za Galerkinovo metodo (modra črta), Tau metodo (rdeča črta), metodo kolokacije (zelena črta) ter dve metodi končnih razlik drugega (črna črta) in četrtega reda (roza črta) za primer 3.1.

Na sliki 3.5 je prikazana primerjava med različnimi metodami. Numerične rešitve, ki so dobljene s tremi spektralnimi metodami: Galerkinovo (modra črta), Tau (rdeča črta) in kolokacijo (zelena črta), eksponentno hitro konvergirajo k točni rešitvi, saj maksimalna absolutna vrednost napake v odvisnosti od odreznega števila N eksponentno hitro pada. Ta rezultat bomo potrdili v naslednjem primeru. Slika potrjuje visoko stopnjo natančnosti že za sorazmerno majhne vrednosti N. Govorimo o spektralni natančnosti teh metod. Medtem ko sta Galerkinova metoda in kolokacija povsem primerljivi,

se Tau metoda obnaša za odtenek slabše. Konvergenca metod končnih razlik je bistveno slabša, saj sta metodi drugega (črna črta), oz. četrtega reda (roza črta).

Primer 3.2 Drugi modelni primer je linearni robni problem z nekonstantnimi koeficienti. Rešujemo enačbo

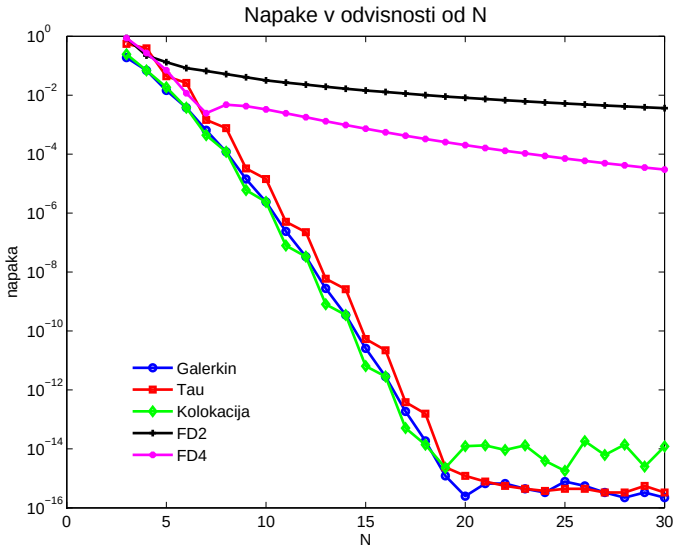
$$y'' + xy' - 2x^2 y = \pi x \cos(\pi x) - (\pi^2 + 2x^2)\sin(\pi x), \quad -1 \leq x \leq 1,$$

skupaj s homogenimi Dirichletovimi robnimi pogoji

$$y(-1) = y(1) = 0,$$

ki ima enolično nepolinomsko analitično rešitev

$$y(x) = \sin(\pi x).$$



Slika 3.6: Primerjava maksimalne absolutne vrednosti napake v odvisnosti od N za Galerkinovo metodo (modra črta), Tau metodo (rdeča črta), metodo kolokacije (zelena črta) ter dve metodi končnih razlik drugega (črna črta) in četrtega reda (roza črta) za primer 3.2.

Na sliki 3.6 je prikazana primerjava med različnimi metodami: Galerkinovo metodo (modra črta), Tau metodo (rdeča črta), kolokacijo (zelena črta), metodo končnih razlik drugega reda (črna črta) ter metodo končnih razlik četrtega reda (roza črta). Slika kaže na enako stopnjo konvergence kot v primeru 3.1, tj. eksponentno za vse tri spektralne metode ter polinomsko za obe metodi končnih razlik.

Eksponentna konvergenca metode pomeni, da maksimalna absolutna vrednost napake $e_N = |y - y_N|$ pada glede na N po formuli

$$e_N = C\,\mathrm{e}^{-\alpha N}, \tag{3.73}$$

kjer sta $C > 0$ in $\alpha > 0$ konstanti. Polinomska konvergenca metode pa pomeni, da maksimalna absolutna vrednost napake $e_N = |y - y_N|$ pada glede na N po formuli

$$e_N = C\,N^{-r}, \tag{3.74}$$

kjer sta $C > 0$ in $r > 0$ konstanti. Koeficient r pomeni red metode.

V tabeli 3.1 so v drugem stolpcu prikazane maksimalne absolutne vrednosti napake e_N^{col} za kolokacijsko spektralno metodo za pripadajoča števila členov vrste N iz prvega stolpca. V tretjem stolpcu pa so prikazane vrednosti za koeficient α_{col}, ki ga v vsaki vrstici izračunamo iz dveh zaporednih napak

$$\alpha = \log\left(e_{N_1}/e_{N_2}\right)/(N_2 - N_1). \tag{3.75}$$

Vrednosti za α_{col} se z naraščajočim N bližajo neki konstantni vrednosti, kar potrjuje eksponentno padanje napake za to metodo. Podoben rezultat dobimo tudi za Galerkinovo in Tau metodo, ki pa sta iz tabele izpuščeni.

N	e_N^{col}	α_{col}	e_N^{fd2}	r_{fd2}	e_N^{fd4}	r_{fd4}
4	2.4435e-01		8.6603e-01		8.6603e-01	
6	1.8443e-02	1.2920	1.3204e-01	4.6387	7.0577e-02	6.1835
8	4.3896e-04	1.8690	6.5936e-02	2.4137	2.4387e-03	11.6978
10	6.0129e-06	2.1453	4.0543e-02	2.1795	4.2485e-03	-2.4877
12	7.8660e-08	2.1683	2.6789e-02	2.2727	2.4330e-03	3.0575
14	8.0537e-10	2.2908	1.9389e-02	2.0970	1.3103e-03	4.0148
16	6.4300e-12	2.4148	1.4429e-02	2.2130	7.2692e-04	4.4121
18	5.0000e-14	2.4406	1.1325e-02	2.0563	4.2181e-04	4.6209

Tabela 3.1: Maksimalne absolutne vrednosti napake ter koeficienti α in r v odvisnosti od N za metodo kolokacije ter metodi končnih razlik drugega in četrtega reda.

Poleg tega so v tabeli 3.1 v četrtem in šestem stolpcu prikazane maksimalne absolutne vrednosti napake e_N^{fd2} in e_N^{fd4} za metodi končnih razlik drugega in četrtega reda za pripadajoča števila členov vrste N iz prvega stolpca. V petem in sedmem stolpcu pa so prikazane vrednosti za koeficienta r_{fd2} in r_{fd4}, ki ju v vsaki vrstici izračunamo iz dveh zaporednih napak

$$r = \log\left(e_{N_1}/e_{N_2}\right)/\log\left(N_1/N_2\right). \tag{3.76}$$

Vrednosti za r_{fd2} in r_{fd4} se z naraščajočim N bližajo konstantnim vrednostim, ki so blizu 2 in 4, kar potrjuje polinomsko padanje napake za ti dve metodi, ki sta torej drugega in četrtega reda.

3.6 Osnovna orodja za analizo napake

V prejšnjih razdelkih tega poglavja smo opisali osnovna načela spektralnih metod, predvsem izbiro baznih funkcij ter shem za prostorsko diskretizacijo robnih problemov glede na metodo uteženega residuala. V tem razdelku pa bomo obravnavali nekatera osnovna orodja za analizo stabilnosti in konvergence za numerične sheme iz razreda spektralnih metod. Podrobno teorijo najdemo npr. v monografijah C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang [11] ter J. Shen, T. Tang in L. Wang [51]. Kot vselej v tem poglavju nas zanima modelni linearni dvotočkovni robni problem (3.1 – 3.2)

$$\mathcal{L}u(x) = f(x), \quad x \in [-1,1], \quad \mathcal{B}u(x) = 0, \quad x \in \{-1,1\},$$

kjer sta $\mathcal{L}$ in $\mathcal{B}$ linearna diferencialna operatorja in f dana funkcija na intervalu $[-1,1]$. Problem (3.1 – 3.2) zapišemo v *šibki formulaciji*

$$\text{Poišči tak } u \in X, \text{ da velja } a(u,v) = F(v) \text{ za vsak } v \in Y, \tag{3.77}$$

kjer je X linearna ogrinjača množice baznih funkcij, Y množica testnih funkcij, F linearni funkcional na Y, $a(\cdot,\cdot)$ pa bilinearna forma na $X \times Y$. Običajno privzamemo, da sta X in Y Hilbertova prostora.

V nadaljevanju potrebujemo nekaj osnovnih definicij iz funkcionalne analize. Nekaj podrobnosti o ortogonalnosti smo opisali že v razdelku 2.1. *Hilbertov prostor* je Banachov prostor opremljen s skalarnim produktom in pripadajočo normo. V njem velja *Cauchy-Schwarzova neenakost* (2.7).

Definicija 3.3 *Linearen funkcional $F : X \to \mathbb{R}$ je* zvezen *oz.* omejen*, če obstaja taka konstanta $c > 0$, za katero za vsak $u \in X$ velja*

$$|F(u)| \leq c\,\|u\|. \tag{3.78}$$

Množica linearnih funkcionalov X' na Hilbertovem prostoru X je Hilbertov prostor, ki ga imenujemo *dualen prostor* prostora X.

Definicija 3.4 *Naj bo X Hilbertov prostor, ki je opremljen z normo $\|\cdot\|$. Funkcional $a(\cdot,\cdot) : X \times X \to \mathbb{R}$ je* bilinearna forma*, če za vsak $u, v, w \in X$ in $\alpha, \beta \in \mathbb{R}$ velja*

$$\begin{aligned} a(\alpha u + \beta v, w) &= \alpha a(u,w) + \beta a(v,w), \\ a(u, \alpha v + \beta w) &= \alpha a(u,w) + \beta a(u,w). \end{aligned}$$

Z drugimi besedami to pomeni, da sta za fiksen $u \in X$ funkcionala $a(u,\cdot) : X \to \mathbb{R}$ in $a(\cdot,u) : X \to \mathbb{R}$ linearna. Bilinearna forma je simetrična*, če je $a(u,v) = a(v,u)$ za vsak $u, v \in X$.*

Naslednji dve definiciji podajata potrebna pogoja za stabilnost in konvergenco, ki nastopata v izrekih in lemah v nadaljevanju.

Definicija 3.5 *Bilinearna forma $a(\cdot,\cdot)$ na Hilbertovem prostoru X je* zvezna *(ang.* continuous*), če obstaja taka konstanta $A > 0$, da za vsak $u, v \in X$ velja*

$$|a(u,v)| \leq A \, \|u\| \, \|v\|. \tag{3.79}$$

Definicija 3.6 *Bilinearna forma $a(\cdot,\cdot)$ na Hilbertovem prostoru X je* pozitivno definitna *(ang.* coercive*) na X, če obstaja taka konstanta $\alpha^* > 0$, da za vsak $u \in X$ velja*

$$a(u,u) \geq \alpha^* \, \|u\|^2. \tag{3.80}$$

Prvi osnovni rezultat za obstoj, enoličnost in stabilnost numerične rešitve je Lax-Milgramova lema (glej P. D. Lax in A. N. Milgram [41]), kjer v (3.77) vzamemo $X = Y$. Posplošitev tega izreka za $X \neq Y$ z upoštevanjem šibkejšega inf-sup pogoja namesto pogoja pozitivne definitnosti najdemo v referencah [11] in [51].

Izrek 3.7 (Lax-Milgramova lema) *Naj bo X Hilbertov prostor, naj bo nadalje $a(\cdot,\cdot) : X \times X \to \mathbb{R}$ zvezna in pozitivno definitna bilinearna forma ter naj bo $F : X \to \mathbb{R}$ linearni funkcional iz X', kjer je X' dual prostora X. Tedaj ima variacijski problem*

$$\textit{Poišči tak } u \in X, \textit{ da velja } a(u,v) = F(v) \textit{ za vsak } v \in X \tag{3.81}$$

enolično rešitev. Poleg tega velja neenakost

$$\|u\| \leq \frac{1}{\alpha^*} \|F\|_{X'}. \tag{3.82}$$

Naj bo kot prej $X = Y$. Privzamemo, da je $X_N \subseteq X$ in za vsak $v \in X$ velja $\inf_{v_N \in X_N} \|v - v_N\|_X \to 0$, ko gre $N \to \infty$. Galerkinova aproksimacija problema (3.81) je

$$\text{Poišči tak } u_N \in X_N, \text{ da velja } a(u_N, v_N) = F(v_N) \text{ za vsak } v_N \in X_N. \tag{3.83}$$

Stabilnost in konvergenca Galerkinove sheme sledita iz spodnjega izreka (glej J. Céa [12]).

Izrek 3.8 (Céa lema) *Pod pogoji Lax-Milgramove leme (izrek 3.7) ima Galerkinova metoda (3.83) enolično rešitev $u_N \in X_N$, za katero velja*

$$\|u_N\|_X \leq \frac{1}{\alpha^*} \|F\|_{X'}. \tag{3.84}$$

Če je u rešitev problema (3.81), velja neenakost

$$\|u - u_N\|_X \leq \frac{A}{\alpha^*} \inf_{v_N \in X_N} \|u - v_N\|_X, \tag{3.85}$$

kjer sta konstanti A in α^* podani z neenačbama (3.79) in (3.80).

Dokaz: Ker je X_N podprostor prostora X, z uporabo Lax-Milgramove leme za (3.83) sledi obstoj in enoličnost rešitve u_N ter ocena za stabilnost (3.84). Za dokaz konvergence vzamemo v enačbi problema (3.81) $v = v_N$ in odštejemo enačbo iz problema (3.83), da dobimo $a(u - u_N, v_N) = 0$ za vsak $v_N \in X_N$. To, skupaj s pogojema zveznosti (3.79) in pozitivne definitnosti (3.80), da za vsak $v_N \in X_N$

$$\begin{aligned} \alpha^* \|u - u_N\|_X^2 &\leq a(u - u_N, u - u_N) = a(u - u_N, u - v_N) \\ &\leq A\|u - u_N\|_X \|u - v_N\|_X, \end{aligned}$$

od koder sledi ocena (3.85). □

V analizi napake za spektralne metode običajno vzamemo, da je v_N ortogonalna projekcija u na prostor X_N, ki jo označimo s $P_N u$. Iz neenačbe (3.85) sledi ocena

$$\|u - u_N\|_X \leq \frac{A}{\alpha^*} \|u - P_N u\|_X. \tag{3.86}$$

Ocena napake numerične metode za reševanje robnega problema tako sledi iz ocene napake za aproksimacijo s projekcijo, ki je običajno oblike

$$\|u - P_N u\|_X \leq C\ N^{-\sigma(m)}\ \|u\|_{H^m}, \tag{3.87}$$

kjer je $C > 0$ pozitivna konstanta, ki je neodvisna od N, $\sigma(m) > 0$ pa funkcija, ki je odvisna od stopnje gladkosti m in jo imenujemo *red konvergence*. Prostor H^m je *prostor Soboljeva* (2.3) na intervalu $[-1, 1]$.

Pogosto je priporočljivo, da namesto zvezne bilinearne forme in zveznih linearnih funkcionalov vzamemo pripadajoče aproksimacije le-teh. Problem (3.83) v diskretni obliki zapišemo kot

$$\text{Poišči tak } u_N \in X_N, \text{ da velja } a_N(u_N, v_N) = F_N(v_N) \text{ za vsak } v_N \in X_N, \tag{3.88}$$

kjer sta $a_N(\cdot, \cdot)$ in $F_N(\cdot)$ primerni aproksimaciji za $a(\cdot, \cdot)$ in $F(\cdot)$.

Spodnji izrek (glej G. Strang [52]), ki je uporaben tudi za dokaz stabilnosti in konvergence kolokacijske sheme, je posplošitev izreka 3.8.

Izrek 3.9 (Strangova lema) *Naj veljajo pogoji Lax-Milgramove leme (izrek 3.7), kjer nadalje predpostavimo, da je $a_N(\cdot, \cdot) : X_N \times X_N \to \mathbb{R}$ zvezna in pozitivno definitna bilinearna forma ter $F_N : X_N \to \mathbb{R}$ linearni funkcional iz X_N', kjer je X_N' dual prostora X_N in $X_N \subset X$, in naj obstaja tak $\alpha^* > 0$, neodvisen od N, za katerega za vsak $v \in X_N$ velja*

$$a_N(v, v) \geq \alpha^*\ \|v\|_X^2. \tag{3.89}$$

Tedaj ima problem (3.88) enolično rešitev $u_N \in X_N$, za katero velja

$$\|u_N\|_X \leq \frac{1}{\alpha^*} \sup_{0 \neq v_N \in X_N} \frac{|F_N(v_N)|}{\|v_N\|_X}. \tag{3.90}$$

Če je u rešitev problema (3.81), velja neenakost

$$\begin{aligned}\|u-u_N\|_X \;\leq\; & \inf_{w_N\in X_N}\left\{\left(1+\frac{A}{\alpha^*}\right)\|u-w_N\|_X\right.\\ & \left.+\frac{1}{\alpha^*}\sup_{0\neq v_N\in X_N}\frac{|a(w_N,v_N)-a_N(w_N,v_N)|}{\|v_N\|_X}\right\}\\ & +\frac{1}{\alpha^*}\sup_{0\neq v_N\in X_N}\frac{|F(v_N)-F_N(v_N)|}{\|v_N\|_X},\end{aligned}\tag{3.91}$$

kjer je konstanta A podana z neenačbo (3.79).

Dokaz: Obstoj in enoličnost rešitve u_N ter ocena za stabilnost (3.90) sledijo, podobno kot pri Céa lemi, iz Lax-Milgramove leme. Dokaz neenakosti (3.91) pa se nekoliko razlikuje od dokaza neenakosti (3.85). Naj bo za vsak $w_N \in X_N$ definiran $e_N = u_N - w_N$. Z uporabo neenakosti (3.89) in enakosti v (3.81) in (3.88) dobimo

$$\begin{aligned}\alpha^*\|e_N\|_X^2 &\leq a_N(e_N,e_N)\\ &= a(u-w_N,e_N)+a(w_N,e_N)-a_N(w_N,e_N)+F_N(e_N)-F(e_N).\end{aligned}$$

Ker je rezultat za $e_N = 0$ trivialen, dobimo iz neenačbe (3.79) za $e_N \neq 0$

$$\begin{aligned}\alpha^*\|e_N\|_X \;\leq\; & A\|u-w_N\|_X+\frac{|a(w_N,e_N)-a_N(w_N,e_N)|}{\|e_N\|_X}\\ & +\frac{|F(e_N)-F_N(e_N)|}{\|e_N\|_X}\\ \leq\; & A\|u-w_N\|_X+\sup_{0\neq v_N\in X_N}\frac{|a(w_N,v_N)-a_N(w_N,v_N)|}{\|v_N\|_X}\\ & +\sup_{0\neq v_N\in X_N}\frac{|F(v_N)-F_N(v_N)|}{\|v_N\|_X}.\end{aligned}$$

Iz definicije e_N in trikotniške neenakosti sledi

$$\|u-u_N\|_X \leq \|u-w_N\|_X+\|e_N\|_X.$$

Neenakost (3.91) sledi z uporabo infimuma glede na vse $w_N \in X_N$ na gornji neenačbi. □

Izrek 3.9 bomo v nadaljevanju uporabili za dokaz konvergence in analizo napake novega razreda kolokacijskih Čebišev-Fourierovih spektralnih metod.

Poglavje 4

Spektralne metode za linearne evolucijske enačbe

V tem poglavju bomo opisali konstrukcijo kolokacijskih metod Čebiševa za reševanje (posplošenih) toplotnih enačb paraboličnega ter (posplošenih) valovnih enačb hiperboličnega tipa, tj. iščemo rešitev linearnih evolucijskih problemov 1.4 in 1.5 skupaj z robnimi ter začetnimi pogoji. Opravka imamo z eno ali več prostorskimi ter časovno spremenljivko katere diskretiziramo ločeno. Numerično rešitev tako iščemo v dveh korakih.

V tem delu se omejimo le na probleme v eni dimenziji, npr. porazdelitev toplote na palici, valovanje na vpeti struni. Za diskretizacijo po prostorski spremenljivki uporabimo kolokacijske (psevdo)spektralne metode, ki so podrobno opisane v poglavju 3.

Za diskretizacijo po časovni spremenljivki pa uporabimo eno izmed standardnih metod za reševanje začetnih problemov za navadne diferencialne enačbe (ODE): linearne eno- ali večstopenjske metode, metode tipa prediktor-korektor, najpogostejša izbira pa so standardne metode Runge-Kutta. Podrobnosti najdemo v vrsti monografij s tega področja, npr. v M. Abramowitz in I. A. Stegun [1], W. Gautschi [22], E. Hairer, S. P. Nørsett in G. Wanner [33], E. Isaacson in H. B. Keller [36] ter A. Iserles [37]. Drugi primeren razred so metode geometrijske integracije, oz. metode Liejevih grup, katerih pomemben predstavnik je Magnusova metoda. Le-te so obširno opisane npr. v monografijah E. Hairer, C. Lubich in G. Wanner [32], E. Hairer, S. P. Nørsett in G. Wanner [33] ter v preglednem članku A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett in A. Zanna [38]. Metode Liejevih grup so podrobno obravnavane tudi v magistrskem delu A. Perne [47], z naslovom *Metode Liejevih grup in parcialne diferencialne enačbe*.

Zanimajo nas samo linearni evolucijski problemi, kjer po diskretizaciji prostorske spremenljivke dobimo začetne probleme oblike

$$\dot{\mathbf{u}} = \mathbf{f}(t, \mathbf{u}), \quad \mathbf{u}(t_0) = \mathbf{u}_0, \tag{4.1}$$

kjer je t časovna spremenljivka, $\mathbf{u}$ pa bodisi skalarna (toplotna enačba) bodisi vektorska funkcija (valovna enačba) odvisna od t, ter $\dot{\mathbf{u}}$ pomeni odvod po t. Nelinearnih problemov v tem delu ne bomo obravnavali.

Standardna eksplicitna 4-stopenjska metoda Runge-Kutta četrtega reda za reševanje začetnih problemov oblike (4.1) je:

$$\begin{aligned} \mathbf{k}_1 &:= \mathbf{f}(t_n, \mathbf{u}_n), \\ \mathbf{k}_2 &:= \mathbf{f}\left(t_n + \tfrac{1}{2}h, \mathbf{u}_n + \tfrac{1}{2}h\mathbf{k}_1\right), \\ \mathbf{k}_3 &:= \mathbf{f}\left(t_n + \tfrac{1}{2}h, \mathbf{u}_n + \tfrac{1}{2}h\mathbf{k}_2\right), \\ \mathbf{k}_4 &:= \mathbf{f}(t_{n+1}, \mathbf{u}_n + h\mathbf{k}_3), \\ \mathbf{u}_{n+1} &:= \mathbf{u}_n + \tfrac{1}{6}h\left(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4\right), \end{aligned} \tag{4.2}$$

kjer je $t_n = t_0 + nh$, torej $t_{n+1} = t_n + h$, $\mathbf{u}_n \approx \mathbf{u}(t_n)$, $n \in \mathbb{N}$, in $h > 0$ časovni korak.

Začetni problemi (4.1) so linearni, za naše potrebe pa tudi homogeni, zato jih lahko za potrebe metod Liejevih grup zapišemo v obliki

$$\dot{\mathbf{u}} = A(t)\mathbf{u}, \quad \mathbf{u}(t_0) = \mathbf{u}_0, \tag{4.3}$$

kjer je A matrika koeficientov sistema navadnih diferencialnih enačb. Eksplicitna Magnusova metoda četrtega reda za reševanje začetnih problemov oblike (4.3) je:

$$\begin{aligned} A_1 &:= A\left(t_n + c_1 h\right), \\ A_2 &:= A\left(t_n + c_2 h\right), \\ \sigma &:= \tfrac{1}{2}h(A_1 + A_2) + \tfrac{\sqrt{3}}{12}h^2[A_2, A_1], \\ \mathbf{u}_{n+1} &:= \exp(\sigma)\mathbf{u}_n, \end{aligned} \tag{4.4}$$

kjer sta $c_1 = \frac{1}{2} - \frac{\sqrt{3}}{6}$ in $c_2 = \frac{1}{2} + \frac{\sqrt{3}}{6}$ Gauss-Legendreova vozla. Kot prej je $t_n = t_0 + nh$, torej $t_{n+1} = t_n + h$, $u_n \approx u(t_n)$, $n \in \mathbb{N}$, in $h > 0$ časovni korak. Pri tem je $[A_2, A_1] = A_2A_1 - A_1A_2$ matrični komutator in exp matrična eksponentna funkcija.

V nadaljevanju poglavja uporabljamo okrajšave za parcialne odvode: $u_x \equiv \frac{\partial u}{\partial x}$, $u_{xx} \equiv \frac{\partial^2 u}{\partial x^2}$, $u_t \equiv \frac{\partial u}{\partial t}$ in $u_{tt} \equiv \frac{\partial^2 u}{\partial t^2}$.

4.1 Kolokacijska metoda Čebiševa za posplošene toplotne enačbe

Linearni evolucijski problemi paraboličnega tipa (problem 1.4) so oblike

$$u_t = \alpha(x,t)u_{xx} + \beta(x,t)u_x + \gamma(x,t)u + \delta(x,t), \tag{4.5}$$

kjer je $x \in [-1, 1]$ in $t \geq 0$, koeficientne funkcije α, β, γ in δ pa so v splošnem odvisne tako od prostorske x, kot tudi od časovne spremenljivke t. Poleg enačbe je podan začetni pogoj

$$u(x, 0) = f(x), \quad x \in [-1, 1] \tag{4.6}$$

ter Dirichletovi robni pogoji

$$u(-1, t) = g(t), \quad u(1, t) = h(t), \quad t \geq 0, \tag{4.7}$$

ki naj bodo konsistentni: $g(0) = f(-1)$, $h(0) = f(1)$.

Reševanje problema (4.5 – 4.7) razdelimo na dva dela. V prvem koraku z uporabo (psevdo)spektralne metode diskretiziramo interval $[-1, 1]$, v drugem koraku pa uporabimo bodisi standardno Runge-Kutta (4.2), bodisi Magnusovo metodo četrtega reda (4.4) za diskretizacijo po časovni spremenljivki (reševanje linearne ODE). V primeru, ko so koeficientne funkcije α, β in γ neodvisne od t, je matrika σ, ki nastopa v Magnusovi metodi, konstantna, od koder sledi, da zadošča napraviti en sam korak po času, saj je ta metoda točna za konstantne matrike. S podobnimi posplošenimi problemi paraboličnega tipa se je ukvarjal A. Y. Suhov v članku [53].

Točno rešitev u diferencialne enačbe (4.5) aproksimiramo z odrezano vrsto Čebiševa P^N (2.55)

$$u(x, t) \approx P^N(x, t) = \sum_{k=0}^{N} u_k(t) T_k(x), \tag{4.8}$$

kjer so koeficienti Čebiševa u_k funkcije časovne spremenljivke t in kjer je N odrezno število vrste, ki je povezano s številom iskanih spektralnih koeficientov. Za odrezno število N imamo tako $N+1$ koeficientov. Parcialne odvode v enačbi (4.5) prav tako aproksimiramo z odrezanimi vrstami Čebiševa tako, da odvajamo vrsto (4.8)

$$u_t(x, t) \approx P_t^N(x, t) = \sum_{k=0}^{N} \dot{u}_k(t) T_k(x), \tag{4.9}$$

$$u_x(x, t) \approx P_x^N(x, t) = \sum_{k=0}^{N-1} u'_k(t) T_k(x), \tag{4.10}$$

$$u_{xx} \approx P_{xx}^N(x, t) = \sum_{k=0}^{N-2} u''_k(t) T_k(x). \tag{4.11}$$

Pri tem so koeficienti $\dot{u}_k$ odvodi osnovnih koeficientov u_k po spremenljivki t, koeficiente u_k, u'_k in u''_k pa povezujeta matrični enačbi

$$\mathbf{u}' = D\,\mathbf{u} \quad \text{in} \quad \mathbf{u}'' = D^2\,\mathbf{u}, \tag{4.12}$$

kjer je D operatorska matrika odvajanja (3.59), ki jo dobimo kot inverz matrike definirane z enačbo (3.13). Spektralne koeficiente odrezanih vrst Čebiševa P^N, P_x^N in P_{xx}^N po vrsti označimo z

$$\begin{aligned}\mathbf{u}(t) &= (u_0(t), u_1(t), \ldots, u_N(t))^T, \\ \mathbf{u}'(t) &= \left(u_0'(t), u_1'(t), \ldots, u_N'(t)\right)^T, \\ \mathbf{u}''(t) &= \left(u_0''(t), u_1''(t), \ldots, u_N''(t)\right)^T,\end{aligned}$$

kjer so koeficienti u_N', u_{N-1}'' in u_N'' konstantno enaki 0.

Nato z odrezanimi vrstami Čebiševa aproksimiramo tudi koeficientne funkcije α, β ter γ

$$\alpha(x,t) \approx \alpha_N(x,t) = \sum_{k=0}^{N} a_k(t) T_k(x), \tag{4.13}$$

$$\beta(x,t) \approx \beta_N(x,t) = \sum_{k=0}^{N} b_k(t) T_k(x), \tag{4.14}$$

$$\gamma(x,t) \approx \gamma_N(x,t) = \sum_{k=0}^{N} c_k(t) T_k(x), \tag{4.15}$$

kjer z

$$\begin{aligned}\mathbf{a}(t) &= (a_0(t), a_1(t), \ldots, a_N(t))^T, \\ \mathbf{b}(t) &= (b_0(t), b_1(t), \ldots, b_N(t))^T, \\ \mathbf{c}(t) &= (c_0(t), c_1(t), \ldots, c_N(t))^T\end{aligned}$$

označimo pripadajoče spektralne koeficiente. Poleg tega z

$$\begin{aligned}\tilde{\mathbf{a}}(t) &= (\tilde{a}_0(t), \tilde{a}_1(t), \ldots, \tilde{a}_N(t))^T, \\ \tilde{\mathbf{b}}(t) &= \left(\tilde{b}_0(t), \tilde{b}_1(t), \ldots, \tilde{b}_N(t)\right)^T, \\ \tilde{\mathbf{c}}(t) &= (\tilde{c}_0(t), \tilde{c}_1(t), \ldots, \tilde{c}_N(t))^T\end{aligned}$$

po vrsti označimo koeficiente produktov $\alpha_N P_{xx}^N$, $\beta_N P_x^N$ in $\gamma_N P^N$, ki nastopajo v aproksimaciji enačbe (4.5) z odrezanimi vrstami Čebiševa. Povezava med koeficienti $\tilde{\mathbf{a}}$, $\tilde{\mathbf{b}}$, $\tilde{\mathbf{c}}$ in $\mathbf{u}$, $\mathbf{u}'$, $\mathbf{u}''$ je podana z matričnimi enačbami

$$\tilde{\mathbf{a}} = F_\alpha\, \mathbf{u}'', \quad \tilde{\mathbf{b}} = F_\beta\, \mathbf{u}' \quad \text{in} \quad \tilde{\mathbf{c}} = F_\gamma\, \mathbf{u}, \tag{4.16}$$

kjer so F_α, F_β in F_γ operatorske multiplikacijske matrike (3.65), ki pripadajo posameznim koeficientnim funkcijam.

Nadalje razvijemo tudi nehomogeni del enačbe δ v odrezano vrsto Čebiševa

$$\delta(x,t) \approx \delta_N(x,t) = \sum_{k=0}^{N} d_k(t) T_k(x), \tag{4.17}$$

kjer z

$$\mathbf{d}(t) = (d_0(t), d_1(t), \ldots, d_N(t))^T$$

označimo vektor pripadajočih spektralnih koeficientov.

Za izračun spektralnih koeficientov u_k uporabimo metodo kolokacije, kjer za kolokacijske vozle izberemo točke Čebiševa (3.11) $x_i = -\cos\left(\frac{i\pi}{N}\right)$, $i = 0, 1, \ldots, N$, da dobimo sistem navadnih diferencialnih enačb prvega reda

$$\sum_{k=0}^{N} \dot{u}_k(t) T_k(x_i) = \sum_{k=0}^{N} \Big(\tilde{a}_k(t) + \tilde{b}_k(t) + \tilde{c}_k(t) + d_k(t)\Big) T_k(x_i) \tag{4.18}$$

za notranje kolokacijske točke x_i, $i = 1, 2, \ldots, N-1$. Naj bo nadalje $C \in \mathbb{R}^{(N+1)\times(N+1)}$ kolokacijska matrika vrednosti polinomov Čebiševa v kolokacijskih vozlih

$$C = \begin{bmatrix} T_0(x_0) & T_1(x_0) & \cdots & T_N(x_0) \\ T_0(x_1) & T_1(x_1) & \cdots & T_N(x_1) \\ \vdots & \vdots & & \vdots \\ T_0(x_N) & T_1(x_N) & \cdots & T_N(x_N) \end{bmatrix}. \tag{4.19}$$

in naj bo $L \in \mathbb{R}^{(N+1)\times(N+1)}$ diferencialna operatorska matrika

$$L = F_\alpha\, D^2 + F_\beta\, D + F_\gamma, \tag{4.20}$$

ki pripada diferencialni enačbi (4.5). Vse operatorske matrike D, F_α, F_β in F_γ so reda $(N+1)\times(N+1)$. Sistem (4.18) lahko zapišemo v matrični obliki

$$\tilde{C}\ \dot{\mathbf{u}} = \tilde{C}\ L\ \mathbf{u} + \tilde{C}\ \mathbf{d}, \tag{4.21}$$

kjer z

$$\dot{\mathbf{u}}(t) = (\dot{u}_0(t), \dot{u}_1(t), \ldots \dot{u}_N(t))^T$$

označimo vektor spektralnih koeficientov odrezane vrste Čebiševa P_t^N in je $\tilde{C} \in \mathbb{R}^{(N-1)\times(N+1)}$ matrika C brez prve in zadnje vrstice.

Nazadnje upoštevamo še Dirichletove robne pogoje tako, da sistemu (4.18) dodamo enačbi

$$P^N(-1, t) = \sum_{k=0}^{N} (-1)^k u_k(t) = g(t), \tag{4.22}$$

$$P^N(1, t) = \sum_{k=0}^{N} u_k(t) = h(t). \tag{4.23}$$

Na ta način iz sistema linearnih diferencialnih enačb dobimo sistem linearnih diferencialno-algebraičnih enačb (DAE), ki ga lahko rešimo z uporabo katere izmed metod za reševanje sistemov DAE. Druga možnost, ki pa ne deluje

vedno, zagotovo pa v primeru homogenih robnih pogojev, je, da enačbi (4.22) in (4.23) odvajamo po časovni spremenljivki:

$$\sum_{k=0}^{N}(-1)^k \dot{u}_k(t) = \dot{g}(t), \qquad \sum_{k=0}^{N} \dot{u}_k(t) = \dot{h}(t), \tag{4.24}$$

in sistemu (4.18) dodamo enačbi (4.24). Sistem (4.21) tako zapišemo v obliki

$$\dot{\mathbf{u}}(t) = U\ \mathbf{u}(t) + \tilde{\mathbf{d}}(t), \tag{4.25}$$

kjer je matrika U dobljena kot produkt inverzne kolokacijske matrike C^{-1} z matriko $\tilde{C}\ L$ iz sistema (4.21), ki ji dodamo zgoraj in spodaj po eno vrstico samih ničel, vektor $\tilde{\mathbf{d}}$ pa je dobljen tako, da matriko C^{-1} pomnožimo z nehomogenim delom $\tilde{C}\ \mathbf{d}$ sistema (4.21), ki mu zgoraj in spodaj dodamo odvoda robnih pogojev, ki sta podana z enačbo (4.24).

Koeficientne funkcije u_k izračunamo kot rešitev sistema linearnih diferencialnih enačb prvega reda (4.25), kjer prvi (homogeni) del sistema rešimo z Magnusovo metodo (4.4) ali z Runge-Kutta metodo (4.2) četrtega reda. Vektor začetnih vrednosti dobimo iz začetnega pogoja (4.6), ko funkcijo f razvijemo v odrezano vrsto Čebiševa

$$f(x) \approx \sum_{k=0}^{N} \lambda_k T_k(x), \tag{4.26}$$

kjer z $\mathbf{u}_0 = (\lambda_0, \lambda_1, \ldots \lambda_N)^T$ označimo vektor pripadajočih spektralnih koeficientov.

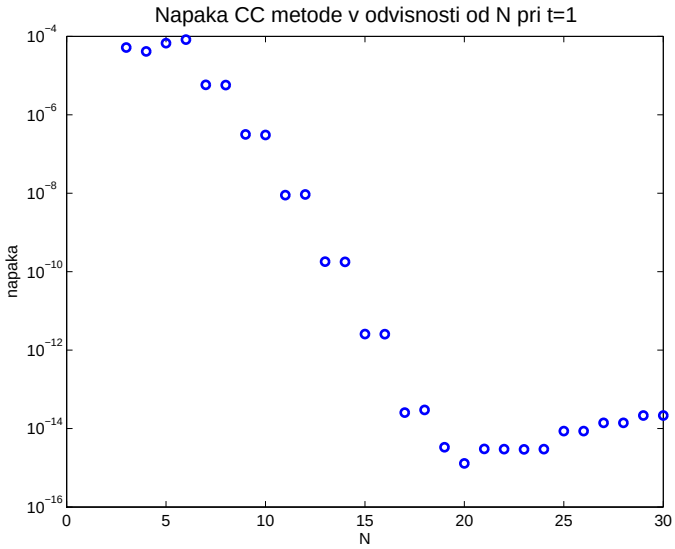
Primer 4.1 Kot zgled vzemimo toplotno enačbo na intervalu $[-1, 1]$

$$u_t = u_{xx}, \quad -1 \le x \le 1, \quad t \ge 0,$$

skupaj s homogenimi Dirichletovimi robnimi pogoji $g(t) = h(t) = 0$ in začetnim pogojem $f(x) = \sin(\pi x)$, ki ima enolično analitično rešitev

$$u(x,t) = \mathrm{e}^{-t\pi^2} \sin(\pi x).$$

Na sliki 4.1 je prikazano padanje maksimalne absolutne vrednosti napake v odvisnosti od odreznega števila N ob času $t = 1$. Napaka eksponentno hitro pada, kar kaže na spektralno natančnost kolokacijske spektralne metode Čebiševa (CC). To potjujejo tudi podatki v tabeli 4.1, kjer so v drugem stolpcu prikazane maksimalne absolutne vrednosti napake e_N^{col} za CC metodo za pripadajoče vrednosti za N, ki so podane v prvem stolpcu. V tretjem stolpcu pa so prikazane vrednosti za koeficient α_{col}, ki ga v vsaki vrstici izračunamo iz dveh zaporednih napak (3.75). Te vrednosti se z naraščajočim N bližajo neki konstantni vrednosti, kar potrjuje eksponentno padanje napake za CC metodo. Dejstvo, da sta dve zaporedni napaki skoraj povsem enaki, lahko pripišemo temu, da je rešitev liha funkcija glede na prostorsko spremenljivko.



Slika 4.1: Padanje maksimalne absolutne vrednosti napake v odvisnosti od odreznega števila N ob času $t = 1$ za kolokacijsko spektralno metodo Čebiševa za toplotno enačbo.

N	e_N^{col}	α_{col}
4	5.1723e-05	
5	6.6883e-05	-0.1285
8	5.8076e-06	1.2219
10	3.1660e-07	1.4546
12	8.9140e-09	1.7850
14	1.7994e-10	1.9514
16	2.5700e-12	2.1235
18	3.0000e-14	2.3054

Tabela 4.1: Maksimalne absolutne vrednosti napake ter koeficienti α v odvisnosti od N za kolokacijsko spektralno metodo Čebiševa (CC).

4.2 Kolokacijska metoda Čebiševa za posplošene valovne enačbe

Linearni evolucijski problemi hiperboličnega tipa (problem 1.5) so oblike

$$u_{tt} = \alpha(x,t)u_{xx} + \beta(x,t)u_x + \gamma(x,t)u + \delta(x,t), \tag{4.27}$$

kjer je $x \in [-1,1]$ in $t \geq 0$, koeficientne funkcije α, β, γ in δ pa so v splošnem odvisne tako od prostorske x, kot tudi od časovne spremenljivke t. Poleg

enačbe sta podana začetna pogoja

$$u(x,0) = f_1(x), \quad u_t(x,0) = f_2(x), \quad x \in [-1,1] \tag{4.28}$$

ter Dirichletovi robni pogoji

$$u(-1,t) = g(t), \quad u(1,t) = h(t), \quad t \geq 0. \tag{4.29}$$

ki naj bodo konsistentni: $g(0) = f_1(-1)$, $h(0) = f_1(1)$, $g'(0) = f_2(-1)$, $h'(0) = f_2(1)$.

Reševanje problema (4.27 – 4.29) razdelimo na dva dela. V prvem koraku z uporabo (psevdo)spektralne metode diskretiziramo interval $[-1,1]$ z enakim postopkom kot za probleme paraboličnega tipa. V drugem koraku pa za diskretizacijo po časovni spremenljivki najprej uvedemo novo spremenljivko, da (linearno) diferencialno enačbo drugega reda prevedemo na sistem dveh (linearnih) diferencialnih enačb prvega reda, ki ga rešimo bodisi s standardno Runge-Kutta (4.2), bodisi z Magnusovo metodo četrtega reda (4.4).

Konstrukcijo psevdospektralne metode Čebiševa (CC) za reševanje posplošenih valovnih enačb (4.27) hiperboličnega tipa izvedemo podobno kot za reševanje posplošenih toplotnih enačb (4.5) paraboličnega tipa v razdelku 4.1. Točno rešitev u diferencialne enačbe (4.27) aproksimiramo z odrezano vrsto Čebiševa P^N (4.8)

$$u(x,t) \approx P^N(x,t) = \sum_{k=0}^{N} u_k(t) T_k(x),$$

kjer je N odrezno število. Parcialne odvode v enačbi (4.27) prav tako aproksimiramo z odrezanimi vrstami Čebiševa (4.9 – 4.11), kjer poleg naštetih potrebujemo še drugi odvod po t

$$u_{tt}(x,t) \approx P^N_{tt}(x,t) = \sum_{k=0}^{N} \ddot{u}_k(t) T_k(x). \tag{4.30}$$

Pri tem so koeficienti $\ddot{u}_k$ drugi odvodi osnovnih koeficientov u_k po spremenljivki t. Koeficientne funkcije α, β in γ aproksimiramo z odrezanimi vrstami Čebiševa (4.13 – 4.15), nehomogeni del enačbe δ pa z vrsto (4.17).

Za izračun spektralnih koeficientov u_k uporabimo metodo kolokacije, kjer za kolokacijske vozle izberemo točke Čebiševa (3.11) $x_i = -\cos\left(\frac{i\pi}{N}\right)$, $i = 0, 1, \ldots, N$, da dobimo sistem diferencialnih enačb drugega reda

$$\sum_{k=0}^{N} \ddot{u}_k(t) T_k(x_i) = \sum_{k=0}^{N} \Big(\tilde{a}_k(t) + \tilde{b}_k(t) + \tilde{c}_k(t) + d_k(t) \Big) T_k(x_i) \tag{4.31}$$

za notranje kolokacijske točke x_i, $i = 1, 2, \ldots, N-1$. Naj bodo C kolokacijska matrika (4.19), $\tilde{C}$ njena podmatrika brez prve in zadnje vrstice ter

L diferencialna operatorska matrika (4.20), ki pripada diferencialni enačbi (4.27) in kjer so D, F_α, F_β in F_γ matrike dane z enačbami (4.12) in (4.16). Sistem (4.31) lahko zapišemo v matrični obliki

$$\tilde{C}\ \ddot{\mathbf{u}} = \tilde{C}\ L\ \mathbf{u} + \tilde{C}\ \mathbf{d}, \tag{4.32}$$

kjer z

$$\ddot{\mathbf{u}}(t) = (\ddot{u}_0(t), \ddot{u}_1(t), \dots \ddot{u}_N(t))^T$$

označimo vektor spektralnih koeficientov odrezane vrste Čebiševa P_{tt}^N.

Nazadnje upoštevamo še Dirichletove robne pogoje tako, da sistemu (4.31) dodamo enačbi (4.22) in (4.23) in dobimo sistem DAE, katerega rešitev obravnavamo enako kot v razdelku 4.1, da dobimo sistem (4.32) v obliki

$$\ddot{\mathbf{u}}(t) = U\ \mathbf{u}(t) + \tilde{\mathbf{d}}(t). \tag{4.33}$$

Rešitev tega sistema so iskane koeficientne funkcije u_k. Prvi (homogeni) del sistema (4.33) z uvedbo novih spremenljivk $\mathbf{y}_1 = \mathbf{u}$ in $\mathbf{y}_2 = \dot{\mathbf{u}}$ za katere velja

$$\begin{aligned} \dot{\mathbf{y}}_1 &= \dot{\mathbf{u}} = \mathbf{y}_2, \\ \dot{\mathbf{y}}_2 &= \ddot{\mathbf{u}} = U\,\mathbf{y}_1, \end{aligned}$$

prevedemo na sistem linearnih diferencialnih enačb prvega reda

$$\begin{bmatrix} \dot{\mathbf{y}}_1 \\ \dot{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} 0 & I \\ U & 0 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}. \tag{4.34}$$

Sistem (4.34) rešimo z Magnusovo metodo (4.4) ali z Runge-Kutta metodo (4.2) četrtega reda. Vektor začetnih vrednosti dobimo iz začetnih pogojev (4.28), ko funkciji f_1 in f_2 razvijemo v odrezani vrsti Čebiševa

$$f_1(x) \approx \sum_{k=0}^{N} \lambda_k T_k(x), \qquad f_2(x) \approx \sum_{k=0}^{N} \mu_k T_k(x), \tag{4.35}$$

kjer z $\mathbf{u}_0 = (\lambda_0, \lambda_1, \dots \lambda_N, \mu_0, \mu_1, \dots \mu_N)^T$ označimo združen vektor pripadajočih spektralnih koeficientov.

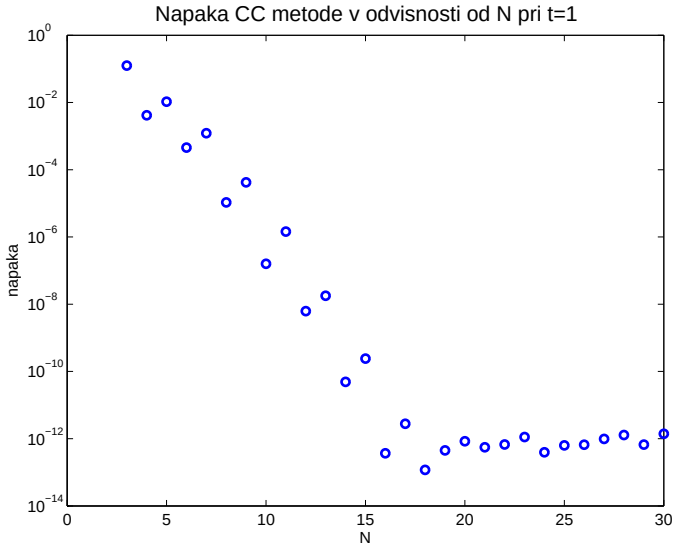
Primer 4.2 Kot zgled vzemimo valovno enačbo na intervalu $[-1, 1]$

$$u_{tt} = u_{xx}, \quad -1 \le x \le 1, \quad t \ge 0,$$

skupaj s homogenimi Dirichletovimi robnimi pogoji $g(t) = h(t) = 0$ in začetnima pogojema $f_1(x) = -\frac{1}{\pi^2}\sin(\pi x)$ in $f_2(x) = 0$, ki ima enolično analitično rešitev

$$u(x,t) = -\tfrac{1}{\pi^2}\cos(\pi t)\sin(\pi x).$$

Na sliki 4.2 je prikazano padanje maksimalne absolutne vrednosti napake v odvisnosti od odreznega števila N ob času $t = 1$. Napaka eksponentno hitro pada, kar kaže na spektralno natančnost kolokacijske spektralne metode Čebiševa (CC). Podobno kot v primeru 4.1 bi lahko preverili red konvergence s tabeliranjem vrednosti za maksimalne absolutne napake v odvisnosti od N in uporabo enačbe (3.75). Postopek je enak, zato ta izračun na tem mestu izpustimo. Dejstvo, da sta dve zaporedni napaki skoraj povsem enaki, lahko ponovno pripišemo temu, da je rešitev liha funkcija glede na prostorsko spremenljivko.



Slika 4.2: Padanje maksimalne absolutne vrednosti napake v odvisnosti od odreznega števila N ob času $t = 1$ za kolokacijsko spektralno metodo Čebiševa za valovno enačbo.

Poglavje 5

Poldomenski polinomi Čebiševa

5.1 Tričlenska rekurzivna formula

Ortogonalne polinome π_k, ki so ortogonalni na intervalu (a, b) glede na nenegativno utež w, običajno konstruiramo z uporabo *tričlenske rekurzivne formule*

$$\begin{aligned}
\pi_{-1}(x) &= 0, \\
\pi_0(x) &= \text{const.}, \\
\tilde{\pi}_{k+1}(x) &= (x - \alpha_k)\pi_k(x) - \beta_k \pi_{k-1}(x), \quad k \geq 0, \qquad (5.1) \\
\pi_{k+1}(x) &= \frac{\tilde{\pi}_{k+1}(x)}{\|\tilde{\pi}_{k+1}(x)\|}, \quad k \geq 0, \qquad (5.2)
\end{aligned}$$

kjer so za vsak $k \geq 0$ rekurzivni koeficienti α_k in β_k enaki

$$\begin{aligned}
\alpha_k &= \int_a^b x\ \pi_k^2(x)\ w(x)\,\mathrm{d}x, \qquad (5.3) \\
\beta_k &= \int_a^b x\ \pi_k(x)\ \pi_{k-1}(x)\ w(x)\,\mathrm{d}x. \qquad (5.4)
\end{aligned}$$

Na žalost je računanje rekurzivnih koeficientov α_k in β_k s formulami (5.3) in (5.4) numerično nestabilno. Z uporabo standardne aritmetike v plavajoči vejici (IEEE) izgubimo vso natančnost že za zelo majhne vrednosti k (npr. $k = 12$). Na voljo imamo vsaj dva pristopa, da se izognemo težavam z nestabilnostjo. Prva možnost je uporaba orodij za simbolno računanje, npr. programskih paketov `Mathematica` [62] ali `Maple`, kjer povečamo število decimalnih mest (števk) in posledično računamo z večjo natančnostjo. Ta pristop ni priporočljiv, saj že za sorazmerno majhne vrednosti k potrebujemo nesorazmerno veliko število dodatnih števk, ki zelo hitro raste s k. Druga možnost je implementacija stabilnega *modificiranega algoritma*

Čebiševa za izračun rekurzivnih koeficientov, ki deluje v standardni aritmetiki s plavajočo vejico.

5.2 Modificiran algoritem Čebiševa

Modificiran algoritem Čebiševa sta leta 1972 najprej opisala R. A. Sack in A. F. Donovan v članku [50]. Neodvisno pa je J. C. Wheeler leta 1974 v članku [61] predstavil alternativen, učinkovit algoritem s časovno zahtevnostjo $\mathcal{O}(N^2)$, ki je ekvivalenten algoritmu, vpeljanem v [50]. V nadaljevanju bomo opisali in uporabili Wheelerjev pristop. Podroben opis tega algoritma najdemo v različnih člankih in knjigah avtorja W. Gautschi [20], [21], [22], [23], [24] ter [25], v članku B. Fisher in G. H. Golub [17] ter v knjigi W. H. Press, B. P. Flannery, S. A. Teukolsky in W. T. Vetterling [48].

Cilj, ki ga zasledujemo z uporabo modificiranega algoritma Čebiševa, je izračun prvih N koeficientov α_j in β_j, ki določajo tričlensko rekurzivno formulo

$$\begin{aligned} \pi_{-1}(x) &\equiv 0, \\ \pi_0(x) &\equiv 1, \\ \pi_{j+1}(x) &= (x-\alpha_j)\pi_j(x) - \beta_j \pi_{j-1}(x), \qquad j = 0, 1, 2, \ldots \end{aligned}$$

za družino moničnih ortogonalnih polinomov π_j, ki so definirani na intervalu (a, b) in zadoščajo pogojem ortogonalnosti

$$\int_a^b \pi_k(x)\, \pi_\ell(x)\, w(x)\, \mathrm{d}x = 0, \quad k \neq \ell. \tag{5.5}$$

Tako dobimo prvih N ortogonalnih polinomov. Glavna ideja je, da namesto prvih $2N$ *potenčnih momentov*

$$\mu_j = \int_a^b x^j\, w(x)\, \mathrm{d}x, \qquad j = 0, 1, \ldots, 2N-1, \tag{5.6}$$

ki nastopajo v izračunu po formulah (5.3 – 5.4), izračunamo in uporabimo prvih $2N$ *modificiranih momentov*

$$\nu_j = \int_a^b p_j(x)\, w(x)\, \mathrm{d}x, \qquad j = 0, 1, \ldots, 2N-1, \tag{5.7}$$

kjer je w utež za ortogonalno družino, za katero iščemo rekurzivne koeficiente. Polinomi p_j so znani in ortogonalni na istem intervalu (a, b) kot družina iskanih polinomov π_j, ki jih želimo konstruirati, vendar z drugačno utežjo. Zadoščajo rekurzivni relaciji

$$\begin{aligned} p_{-1}(x) &\equiv 0, \\ p_0(x) &\equiv 1, \\ p_{j+1}(x) &= (x-a_j)p_j(x) - b_j p_{j-1}(x), \qquad j = 0, 1, 2, \ldots, \end{aligned}$$

kjer so koeficienti a_j in b_j znani eksplicitno.

Pristop k izvedbi modificiranega algoritma Čebiševa, ki ga je razvil Wheeler v [61], temelji na izračunu vmesnih vrednosti, oz. *mešanih momentov*

$$\sigma_{k,\ell} = \int_a^b \pi_k(x)\, p_\ell(x)\, w(x)\, \mathrm{d}x, \qquad k, \ell \geq -1. \tag{5.8}$$

Te vrednosti nato uporabimo za izračun rekurzivnih koeficientov. Iz pogojev ortogonalnosti sledi, da je $\sigma_{k,\ell} = 0$ za $\ell < k$. Vhodni parametri za algoritem so znani rekurzivni koeficienti a_j in b_j ter modificirani momenti ν_j, ki pa morajo biti izračunani natančno. Izhodni parametri so rekurzivni koeficienti α_j in β_j.

Algoritem: Inicializacija:

$$\begin{aligned}
&\alpha_0 = a_0 + \frac{\nu_1}{\nu_0}, \quad \beta_0 = \nu_0,\\
&\sigma_{-1,\ell} = 0, \qquad \ell = 1, 2, \ldots, 2N-2,\\
&\sigma_{0,\ell} = \nu_\ell, \qquad \ell = 0, 1, \ldots, 2N-1.
\end{aligned}$$

Za $k = 1, 2, \ldots, N-1$ izračunaj

$$\begin{aligned}
\sigma_{k,\ell} &= \sigma_{k-1,\ell+1} - (\alpha_{k-1} - a_\ell)\sigma_{k-1,\ell} - \beta_{k-1}\sigma_{k-2,\ell} + b_\ell \sigma_{k-1,\ell-1},\\
&\qquad\qquad \ell = k, k+1, \ldots, 2N-k-1,\\
\alpha_k &= a_k + \frac{\sigma_{k,k+1}}{\sigma_{k,k}} - \frac{\sigma_{k-1,k}}{\sigma_{k-1,k-1}},\\
\beta_k &= \frac{\sigma_{k,k}}{\sigma_{k-1,k-1}}.
\end{aligned}$$

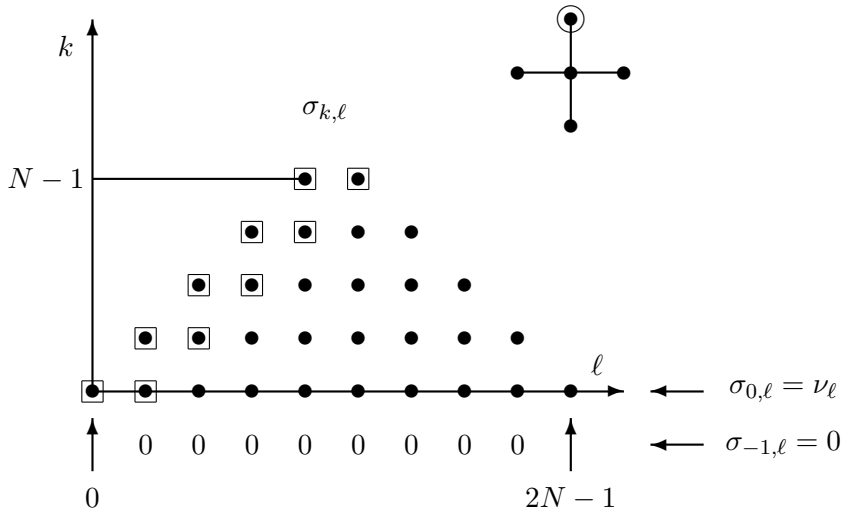
Shematično lahko modificiran algoritem Čebiševa predstavimo kot je prikazano na sliki 5.1. Vmesne vrednosti $\sigma_{k,\ell}$ izračunamo s pomočjo pettočkovne sheme (na sliki desno zgoraj), kjer računamo obkroženi element iz preostalih štirih elementov. Za določitev vseh potrebnih vmesnih vrednosti zadošča izračunati le tiste vmesne vrednosti $\sigma_{k,\ell}$, ki so označene s črno piko, rekurzivne koeficiente α_j in β_j pa izračunamo zgolj z uporabo tistih vmesnih vrednosti, ki so občrtane s kvadratkom.

Modificiran algoritem Čebiševa deluje za monične ortogonalne polinome. Običajno pa ti polinomi niso monični, zato potrebujemo normalizacijske faktorje, tj. norme polinomov

$$\|\pi_0\|^2 = \nu_0, \tag{5.9}$$

$$\|\pi_j\|^2 = \beta_j \|\pi_{j-1}\|^2, \qquad j = 1, 2, \ldots. \tag{5.10}$$

Ker nas v nadaljevanju poglavja zanimata dve družini nestandardnih ortogonalnih polinomov, ki nista monični, je potrebno iz rekurzivnih koeficientov



Slika 5.1: Shema za izračun vmesnih vrednosti $\sigma_{k,\ell}$ v modificiranem algoritmu Čebiševa.

α_j in β_j za monične polinome izračunati rekurzivne koeficiente α_j^* in β_j^* za pripadajoče normalizirane polinome

$$\alpha_j^* = \alpha_j, \qquad j = 0, 1, \ldots, \tag{5.11}$$

$$\beta_j^* = \beta_j \frac{\|\pi_{j-1}\|}{\|\pi_j\|}, \qquad j = 1, 2, \ldots. \tag{5.12}$$

Formuli (5.11) in (5.12) direktno sledita iz enostavne primerjave tričlenskih rekurzivnih formul za monične in normalizirane polinome.

5.3 Poldomenski polinomi Čebiševa

Glavni cilj doktorske disertacije je konstrukcija spektralnih metod, kjer numerično rešitev danega linearnega dvotočkovnega ali evolucijskega robnega problema aproksimiramo z neperiodičnimi trigonometričnimi vrstami na danem intervalu $[-1, 1]$. Z drugimi besedami to pomeni, da numerično rešitev zapišemo v obliki funkcijske vrste, ki je periodična na širšem intervalu, v konkretnem primeru na intervalu $[-2, 2]$. V tako zapisani vrsti nastopajo tako sinusne in kosinusne funkcije, kot tudi sinusne in kosinusne funkcije polovičnih kotov, ki so organizirane kot ortogonalni polinomi. Te

nestandardne polinome, ki so ortogonalni na intervalu $[0,1]$ in ki jih imenujemo *poldomenski polinomi Čebiševa prve in druge vrste* (ang. *half-range Chebyshev polynomials of the first and second kind*), je prvi vpeljal D. Huybrechs v članku [35]. Poleg tega je predlagal tudi uporabo funkcijske vrste, ki jo imenujemo *poldomenska Čebišev-Fourierova (HCF) vrsta* (ang. *half-range Chebyshev-Fourier series*).

Huybrechs je v svojem delu obravnaval problem aproksimacije 1.1 oz. 1.2, nas pa bodo v nadaljevanju tega dela zanimali robni problemi 1.3, 1.4 in 1.5, kjer numerično rešitev aproksimiramo v obliki HCF vrste. Konstrukcija poldomenskih polinomov Čebiševa prve in druge vrste z uporabo modificiranega algoritma Čebiševa ter njihove lastnosti, vključno z nekaterimi lastnostmi HCF vrste so podrobno obravnavane v članku B. Orel in A. Perne [46].

Definicija 5.1 *Naj bo T_k^h tisto normalizirano zaporedje ortogonalnih polinomov, ki zadošča pogojema*

$$\int_0^1 T_k^h(x) x^\ell \frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x = 0, \quad \ell = 0,\ldots,k-1, \tag{5.13}$$

$$\frac{4}{\pi}\int_0^1 \left(T_k^h(x)\right)^2 \frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x = 1. \tag{5.14}$$

Tedaj se elementi množice $\{T_k^h\}_{k=0}^\infty$ imenujejo poldomenski polinomi Čebiševa prve vrste.

Definicija 5.2 *Naj bo U_k^h tisto normalizirano zaporedje ortogonalnih polinomov, ki zadošča pogojema*

$$\int_0^1 U_k^h(x) x^\ell \sqrt{1-x^2}\,\mathrm{d}x = 0, \quad \ell = 0,\ldots,k-1, \tag{5.15}$$

$$\frac{4}{\pi}\int_0^1 \left(U_k^h(x)\right)^2 \sqrt{1-x^2}\,\mathrm{d}x = 1. \tag{5.16}$$

Tedaj se elementi množice $\{U_k^h\}_{k=0}^\infty$ imenujejo poldomenski polinomi Čebiševa druge vrste.

Zgornji družini ortogonalnih polinomov imata enaki uteži kot klasični družini polinomov Čebiševa prve in druge vrste, le da so ortogonalni na krajšem intervalu $[0,1]$ namesto na intervalu $[-1,1]$. Tako definirani polinomi obstajajo in so enolično določeni, saj sta obe uteži, ki nastopata v definicijah, pozitivni in integrabilni. Za podrobnosti glej K. Atkinson in W. Han [6] ali T. S. Chihara [13]. Norma je v obeh primerih enaka

$$\|T_k^h\|^2 = \|U_k^h\|^2 = \frac{\pi}{4}$$

za vsak $k \geq 0$.

5.3.1 Poldomenski polinomi Čebiševa prve vrste

Poldomenski polinomi Čebiševa prve vrste so ortogonalni na intervalu $[0,1]$ z utežjo $w(x) = 1/\sqrt{1-x^2}$. Pripadajoči momentni funkcional za to družino ortogonalnih polinomov je podan z integralom

$$\mathcal{L}_T(\pi) := \int_0^1 \pi(x) \frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x, \tag{5.17}$$

zaporedje potenčnih momentov $\mu_k := \mathcal{L}_T(x^k)$ pa je karakterizirano s spodnjo lemo.

Lema 5.3 *Potenčni momenti momentnega zaporedja za poldomenske polinome Čebiševa prve vrste so*

$$\mu_{2k} = \frac{\binom{2k}{k}\pi}{2^{2k+1}}, \quad k = 0, 1, \ldots, \tag{5.18}$$

$$\mu_{2k+1} = \frac{(2k)!!}{(2k+1)!!}, \quad k = 0, 1, \ldots. \tag{5.19}$$

Dokaz: Elemente momentnega zaporedja izračunamo iz formule (5.17) za $\pi(x) = x^n$, kjer uvedemo novo spremenljivko $x = \cos t$, $\mathrm{d}x = -\sin t\,\mathrm{d}t$, $\sqrt{1-x^2} = \sin t$

$$\mu_n = \int_0^1 \frac{x^n}{\sqrt{1-x^2}}\,\mathrm{d}x = \int_0^{\frac{\pi}{2}} \cos^n t\,\mathrm{d}t.$$

Pri izračunu gornjega integrala ločimo dva primera, in sicer za sode in lihe potence.

i) Za $n = 2k$, $k \geq 0$:

$$\mu_{2k} = \int_0^{\frac{\pi}{2}} \cos^{2k} t\,\mathrm{d}t = \frac{\sqrt{\pi}\,\Gamma(k+\frac{1}{2})}{2\Gamma(k+1)} = \frac{\binom{2k}{k}\pi}{2^{2k+1}}.$$

ii) Za $n = 2k+1$, $k \geq 0$:

$$\mu_{2k+1} = \int_0^{\frac{\pi}{2}} \cos^{2k+1} t\,\mathrm{d}t = \frac{\sqrt{\pi}\,\Gamma(k+1)}{2\Gamma(k+\frac{3}{2})} = \frac{(2k)!!}{(2k+1)!!}.$$

V obeh primerih upoštevamo lastnosti gama funkcije: $\Gamma(n+1) = n\,\Gamma(n)$ in $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. □

Za konstrukcijo rekurzivnih koeficientov (5.3) in (5.4) z modificiranim algoritmom Čebiševa uporabimo kot znano družino ortogonalnih polinomov p_k Legendreove polinome L_k preslikane na interval $[0,1]$. Ker morajo biti

polinomi, ki jih želimo uporabiti za modificiran algoritem Čebiševa, monični, je množica polinomov p_k podana s predpisom

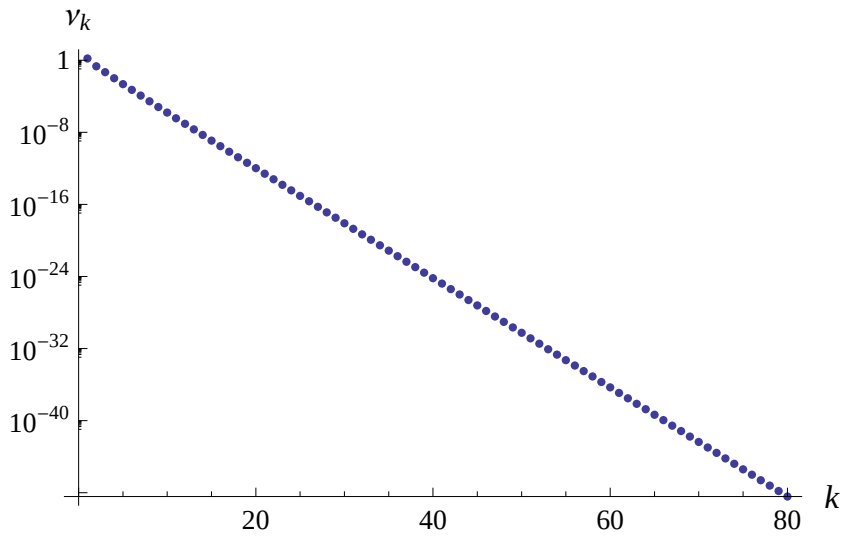
$$p_k(x) = \frac{(k!)^2}{(2k)!} L_k(2x-1). \tag{5.20}$$

Rekurzivni koeficienti za tričlensko rekurzivno formulo za preslikane monične Legendreove polinome so

$$a_k = \frac{1}{2}, \qquad k = 0, 1, \ldots, \tag{5.21}$$

$$b_k = \frac{1}{4(4-k^{-2})}, \qquad k = 1, 2, \ldots. \tag{5.22}$$

Modificirane momente ν_k, ki so definirani z enačbo (5.7), izračunamo z uporabo orodij za simbolno računanje (`Mathematica`) iz potenčnih momentov μ_k, ki so podani z enačbama (5.18 – 5.19) ter prikazani na sliki 5.2.



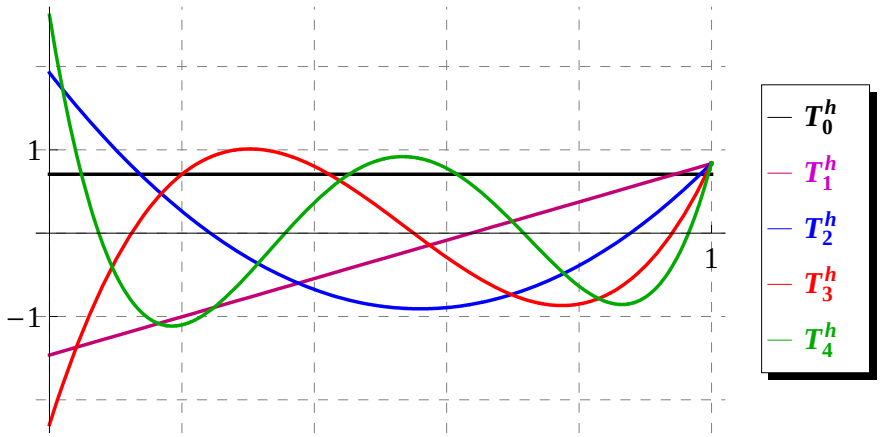
Slika 5.2: Vrednosti modificiranih momentov ν_k za poldomenske polinome Čebiševa prve vrste.

Modificiran algoritem Čebiševa vrne rekurzivne koeficiente $\widetilde{\alpha}_k$ in $\widetilde{\beta}_k$ za monične poldomenske polinome Čebiševa $\widetilde{\pi}_k$ prve vrste ter po potrebi norme teh polinomov. To nam omogoča izračun rekurzivnih koeficientov α_k in β_k za poldomenske polinome Čebiševa T_k^h prve vrste, ki so definirani s formulama (5.11) in (5.12).

Prvi štirje poldomenski polinomi Čebiševa prve vrste so:

$$\begin{aligned} T_0^h(x) &= \frac{\sqrt{2}}{2}, \\ T_1^h(x) &= \frac{\pi x-2}{\sqrt{\pi^2-8}}, \\ T_2^h(x) &= \frac{(6\pi^2-48)x^2-4\pi x+32-3\pi^2}{\sqrt{9\pi^4-160\pi^2+704}}, \\ T_3^h(x) &= \frac{(540\pi^3-5280\pi)x^3+(6144-648\pi^2)x^2+(4032\pi-405\pi^3)x+414\pi^2-4096}{\sqrt{18225\pi^6-477576\pi^4+4106880\pi^2-11534336}}, \end{aligned}$$

grafi prvih petih polinomov pa so prikazani na sliki 5.3.



Slika 5.3: Grafi prvih petih poldomenskih polinomov Čebiševa prve vrste T_n^h: T_0^h črna črta, T_1^h vijolična, T_2^h modra, T_3^h rdeča in T_4^h zelena črta.

5.3.2 Poldomenski polinomi Čebiševa druge vrste

Poldomenski polinomi Čebiševa druge vrste so prav tako ortogonalni na intervalu $[0,1]$ z utežjo $w(x)=\sqrt{1-x^2}$. Pripadajoči momentni funkcional za to družino ortogonalnih polinomov je podan z integralom

$$\mathcal{L}_U(\pi) := \int_0^1 \pi(x)\sqrt{1-x^2}\,\mathrm{d}x, \tag{5.23}$$

zaporedje potenčnih momentov $\mu_k := \mathcal{L}_U(x^k)$ pa je karakterizirano s spodnjo lemo.

Lema 5.4 *Potenčni momenti momentnega zaporedja za poldomenske polinome Čebiševa druge vrste so*

$$\mu_{2k} = \frac{\binom{2k}{k}\pi}{2^{2k+2}(k+1)}, \quad k=0,1,\ldots, \tag{5.24}$$

$$\mu_{2k+1} = \frac{(2k)!!}{(2k+3)!!}, \quad k=0,1,\ldots. \tag{5.25}$$

Dokaz: Elemente momentnega zaporedja izračunamo iz formule (5.23) za $\pi(x) = x^n$, kjer uvedemo novo spremenljivko $x = \cos t$, $\mathrm{d}x = -\sin t\,\mathrm{d}t$, $\sqrt{1-x^2} = \sin t$

$$\mu_n = \int_0^1 x^n \sqrt{1-x^2}\,\mathrm{d}x = \int_0^{\frac{\pi}{2}} \cos^n(t)(1-\cos^2 t)\,\mathrm{d}t.$$

Pri izračunu gornjega integrala ločimo dva primera, in sicer za sode in lihe potence.

i) Za $n = 2k$, $k \geq 0$:

$$\mu_{2k} = \int_0^{\frac{\pi}{2}} \cos^{2k}(t)(1-\cos^2 t)\,\mathrm{d}t = \frac{\sqrt{\pi}\,\Gamma(k+\frac{1}{2})}{4(k+1)\Gamma(k+1)} = \frac{\binom{2k}{k}\pi}{2^{2k+2}(k+1)}.$$

ii) Za $n = 2k+1$, $k \geq 0$:

$$\mu_{2k+1} = \int_0^{\frac{\pi}{2}} \cos^{2k+1}(t)(1-\cos^2 t)\,\mathrm{d}t = \frac{\sqrt{\pi}\,\Gamma(k+1)}{2(2k+3)\Gamma(k+\frac{3}{2})} = \frac{(2k)!!}{(2k+3)!!}.$$

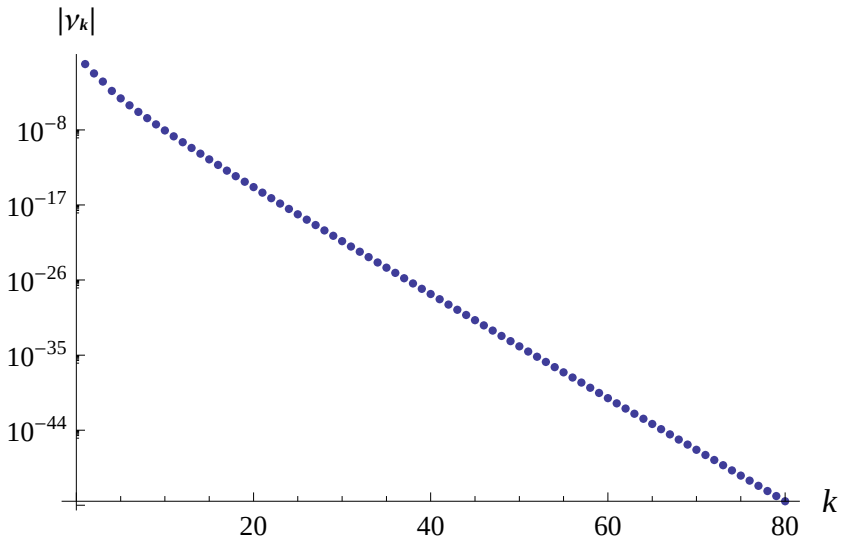
V obeh primerih upoštevamo lastnosti gama funkcije: $\Gamma(n+1) = n\,\Gamma(n)$ in $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. □

Za konstrukcijo rekurzivnih koeficientov (5.3) in (5.4) z modificiranim algoritmom Čebiševa ponovno uporabimo kot znano družino ortogonalnih polinomov p_k monične Legendreove polinome (5.20) preslikane na interval $[0,1]$ z rekurzivnimi koeficienti (5.21) in (5.22). Modificirane momente ν_k, ki so definirani z enačbo (5.7), izračunamo z uporabo orodij za simbolno računanje (`Mathematica`) iz potenčnih momentov μ_k, ki so podani z enačbama (5.24 – 5.25) ter prikazani na sliki 5.4.

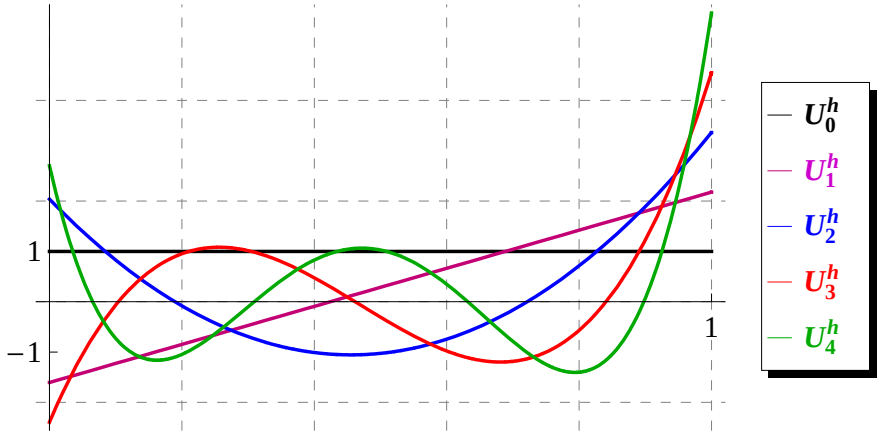
Modificiran algoritem Čebiševa vrne rekurzivne koeficiente $\widetilde{\alpha}_k$ in $\widetilde{\beta}_k$ za monične poldomenske polinome Čebiševa $\widetilde{\pi}_k$ druge vrste ter po potrebi norme teh polinomov. To nam omogoča izračun rekurzivnih koeficientov α_k in β_k za poldomenske polinome Čebiševa U_k^h druge vrste, ki so definirani s formulama (5.11) in (5.12).
Prvi štirje poldomenski polinomi Čebiševa druge vrste so:

$$\begin{aligned}
U_0^h(x) &= 1,\\
U_1^h(x) &= \tfrac{6\pi x-8}{\sqrt{9\pi^2-64}},\\
U_2^h(x) &= \tfrac{(180\pi^2-1280)x^2-144\pi x+512-45\pi^2}{\sqrt{2025\pi^4-33984\pi^2+139264}},\\
U_3^h(x) &= \tfrac{(63000\pi^3-609280\pi)x^3+(655360-72000\pi^2)x^2+(317440\pi-31500\pi^3)x+26400\pi^2-262144}{\sqrt{62015625\pi^6-1594080000\pi^4+13391052800\pi^2-36507222016}},
\end{aligned}$$

grafi prvih petih polinomov pa so prikazani na sliki 5.5.



Slika 5.4: Absolutne vrednosti modificiranih momentov $|\nu_k|$ za poldomenske polinome Čebiševa druge vrste.



Slika 5.5: Grafi prvih petih poldomenskih polinomov Čebiševa druge vrste U_n^h: U_0^h črna črta, U_1^h vijolična, U_2^h modra, U_3^h rdeča in U_4^h zelena črta.

5.4 Lastnosti poldomenskih polinomov Čebiševa

V tem razdelku bomo obravnavali nekatere osnovne lastnosti poldomenskih polinomov Čebiševa prve in druge vrste, ki smo jih definirali in konstruirali v razdelku 5.3. Najprej bomo zapisali pomožen rezultat, nato pa še dva pomembnejša izreka.

Lema 5.5 *Naj bo p_n dana družina ortogonalnih polinomov glede na utež $w(z) = h(z)(1-z)^\alpha(1-z)^\beta$ na intervalu $[-1,1]$, kjer je h strogo pozitivna realna analitična funkcija na intervalu $[-1,1]$ in naj bosta $\alpha, \beta > -1$ realni vrednosti. Tedaj velja*

$$p_n(1) \sim n^{\alpha+1/2}, \qquad p_n(-1) \sim n^{\beta+1/2}, \qquad n \to \infty.$$

Dokaz: Dokaz leme je podan v članku D. Huybrechs [35]. Omejimo se na krajišče intervala $z = 1$. Dokaz za krajišče $z = -1$ sledi zaradi simetrije. Iz izreka 1.13 v članku A. B. J. Kuijlaars, K. T. -R. McLaughlin, W. Van Assche in M. Vanlessen [40], dobimo asimptotsko obnašanje moničnih ortogonalnih polinomov π_n za $z \in (1-\delta, 1)$, kjer je δ dovolj majhen

$$\begin{aligned}\pi_n(z) \quad \sim \quad & \frac{n^{1/2}}{2^n} a_1(z) \Big(\cos\left(a_2(z)\right) J_\alpha(n \arccos z) \\ & + \sin\left(a_3(z)\right) J'_\alpha(n \arccos z) + \mathcal{O}\left(\tfrac{1}{n}\right) \Big).\end{aligned}$$

Tu so funkcije a_j, $j = 1, 2, 3$, znane eksplicitno in neodvisne od n, J_α pa je standardna Besselova funkcija reda α. Iz izreka 1.6 v [40] dobimo asimptotsko relacijo med moničnimi in ortogonalnimi polinomi

$$p_n(z) \sim 2^n \pi_n(z), \quad n \to \infty.$$

Iz opažanja, da je $J_\alpha(z) \sim z^\alpha$, ko gre $n \to \infty$ (glej enačbo (9.1.7) v [1]), sledi rezultat leme. □

Izrek 5.6 *Poldomenski polinomi Čebiševa prve in druge vrste zadoščajo relacijam*

$$T_k^h(0) \sim (-1)^k \sqrt{k}, \quad T_k^h(1) \sim 1, \quad U_k^h(0) \sim (-1)^k \sqrt{k}, \quad U_k^h(1) \sim k,$$

za $k \to \infty$.

Dokaz: Dokaz izreka je podan v članku D. Huybrechs [35]. Uteži za poldomenske polinome Čebiševa prve in druge vrste imata obliko

$$w(z) = h(z)(1-z)^\alpha(1-z)^\beta,$$

ki je določena do linearne preslikave $z = 2x - 1$, ki interval $[0, 1]$ preslika na interval $[-1, 1]$, natančno. Sedaj uporabimo izrek 5.5 z $\alpha = -\frac{1}{2}$ in $\beta = 0$ za poldomenske polinome Čebiševa prve vrste ter z $\alpha = \frac{1}{2}$ in $\beta = 0$ za poldomenske polinome Čebiševa druge vrste. □

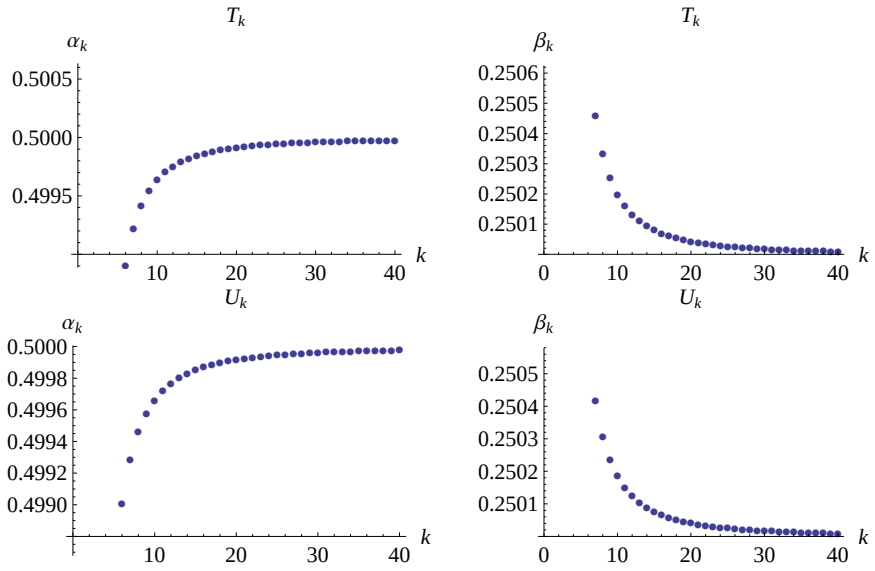
Izrek 5.7 *Rekurzivni koeficienti α_k in β_k v tričlenski rekurzivni formuli (5.1) za poldomenske polinome Čebiševa T_k^h prve in U_k^h druge vrste zadoščajo relacijam*

$$\alpha_k \to \frac{1}{2}, \quad \beta_k \to \frac{1}{4}, \quad k \to \infty.$$

Dokaz: Dokaz tega izreka sledi direktno iz izreka 12.7.1 in enačb (3.2.1) in (3.2.2) v knjigi G. Szegö [54] (str. 307–308), ko interval $[-1, 1]$ s primerno uvedbo nove spremenljivke $t = (x + 1)/2$ preslikamo na interval $[0, 1]$. □

Iz formul (5.21) in (5.22) opazimo, da za monične preslikane Legendreove polinome velja, da je $\alpha_k = \frac{1}{2}$ in $\beta_k \to \frac{1}{16}$, $k \to \infty$. Po normalizaciji dobimo podoben rezultat za preslikane Legendreove polinome kot v izreku 5.7, tj. $\alpha_k = \frac{1}{2}$ in $\beta_k \to \frac{1}{4}$, $k \to \infty$. Podobno za rekurzivne koeficiente preslikanih polinomov Čebiševa prve in druge vrste velja, da je $\alpha_k = \frac{1}{2}$ in $\beta_k = \frac{1}{4}$, $k \to \infty$ z izjemo β_1.

Na sliki 5.6 je prikazano asimptotsko obnašanje rekurzivnih koeficientov za obe neklasični družini ortogonalnih polinomov definiranih v podrazdelkih 5.3.1 in 5.3.2.



Slika 5.6: Rekurzivni koeficienti α_k (zgornja leva slika) in β_k (zgornja desna slika) za poldomenske polinome Čebiševa T_k^h prve vrste ter rekurzivni koeficienti α_k (spodnja leva slika) in β_k (spodnja desna slika) za poldomenske polinome Čebiševa U_k^h druge vrste.

Poglavje 6

Aproksimacija s poldomensko Čebišev-Fourierovo vrsto

6.1 Ortonormalna baza

V tem razdelku bomo konstruirali ortonormalno bazo prostora funkcij na intervalu $[-1,1]$, ki jo predlaga Huybrechs v [35] in jo sestavljajo trigonometrične funkcije, ki so organizirane kot poldomenski polinomi Čebiševa prve in druge vrste.

Za dano s kvadratom integrabilno funkcijo $f \in L^2(-1,1)$ iščemo Fourierovo vrsto g_N na intervalu $[-2,2]$ oblike

$$g_N(x) = \frac{a_0}{2} + \sum_{k=1}^{N} \left(a_k \cos \frac{\pi k x}{2} + b_k \sin \frac{\pi k x}{2} \right). \tag{6.1}$$

Pri tem želimo vse izračune izvesti na intervalu $[-1,1]$. Funkcijski prostor $\mathcal{G}_N \subset L^2(-1,1)$ naj bo linearna ogrinjača množice

$$D_N := C_N \cup S_N, \tag{6.2}$$

kjer je

$$C_N = \left\{ \tfrac{1}{\sqrt{2}} \right\} \cup \left\{ \cos \tfrac{\pi k \cdot}{2} \right\}_{k=1}^{N} \quad \text{in} \quad S_N = \left\{ \sin \tfrac{\pi k \cdot}{2} \right\}_{k=1}^{N}. \tag{6.3}$$

Opazimo, da množico C_N sestavljajo sode, množico S_N pa lihe funkcije. Množici imata pomembno vlogo v analizi konvergence. Dokaz spodnje leme dobimo v [35].

Lema 6.1 *Množica D_∞ je sestavljena iz vseh lastnih funkcij Laplaceovega operatorja na intervalu $[-1,1]$, kjer imamo bodisi homogene Dirichletove, bodisi homogene Neumannove robne pogoje.*

Dokaz: Z upoštevanjem razlike med sodimi in lihimi vrednostmi indeksa k v (6.3) lahko množico D_∞ razbijemo na dve podmnožici

$$L^N := \left\{\tfrac{1}{\sqrt{2}}\right\} \cup \{\cos(\pi k\,\cdot)\}_{k=1}^{\infty} \cup \left\{\sin\left(\pi\left(k+\tfrac{1}{2}\right)\,\cdot\right)\right\}_{k=1}^{\infty} \tag{6.4}$$

in

$$L^D := \left\{\cos\left(\pi\left(k+\tfrac{1}{2}\right)\,\cdot\right)\right\}_{k=1}^{\infty} \cup \{\sin(\pi k\,\cdot)\}_{k=1}^{\infty}. \tag{6.5}$$

Množica L^N je sestavljena iz vseh lastnih funkcij Laplaceovega operatorja na intervalu $[-1,1]$ glede na homogene Neumannove robne pogoje, množica L^D pa iz vseh lastnih funkcij Laplaceovega operatorja na intervalu $[-1,1]$ glede na homogene Dirichletove robne pogoje. □

Množici L^N in L^D sta ortonormalni bazi prostora $L^2(-1,1)$. Sledi, da je množica D_∞ ogrodje prostora $L^2(-1,1)$. Velja spodnja posledica.

Posledica 6.2 *Množica D_∞ je* prilegajoče se ogrodje *(ang.* tight frame*) prostora $L^2(-1,1)$ z* ogrodno mejo *(ang.* frame bound*)* 2. *To pomeni, da ogrodje zadošča posplošeni Parsevalovi enakosti*

$$\sum_k |(f,\phi_k)|^2 = 2\|f\|, \quad f \in L^2(-1,1), \tag{6.6}$$

kjer so ϕ_k elementi množice D_∞.

Čeprav množica D_∞ ni baza prostora $L^2(-1,1)$, pač pa le prilegajoče se ogrodje, je za vsak končen N množica D_N baza končnorazsežnega podprostora v $L^2(-1,1)$, kar pomeni, da so vse funkcije iz množice D_N linearno neodvisne. Zanima nas ortonormalna baza na intervalu $[-1,1]$. Ker so sode in lihe funkcije iz C_N in S_N medsebojno ortogonalne na simetričnem intervalu $[-1,1]$, iskanje ortonormalne baze naravno razpade na dva manjša dela. Linearni ogrinjači množic C_N in S_N označimo s

$$\mathcal{C}_N := \mathrm{Lin}(C_N) \quad \text{in} \quad \mathcal{S}_N := \mathrm{Lin}(S_N), \tag{6.7}$$

kjer je $\mathcal{C}_N$ $(N+1)$-razsežen prostor sodih in $\mathcal{S}_N$ N-razsežen prostor lihih funkcij. Znano dejstvo je, da je $\cos(kx) = T_k(\cos x)$ polinom v $\cos x$ stopnje k. Podobno opazimo, da velja

$$\cos\tfrac{\pi k x}{2} = T_k\left(\cos\tfrac{\pi x}{2}\right), \tag{6.8}$$

$$\sin\tfrac{\pi(k+1)x}{2} = U_k\left(\cos\tfrac{\pi x}{2}\right)\sin\tfrac{\pi x}{2}. \tag{6.9}$$

Dokaza spodnjih dveh izrekov lahko najdemo v [35].

Izrek 6.3 *Naj bo $\{T_k^h\}$, $k \geq 0$, množica poldomenskih polinomov Čebiševa prve vrste opisanih v definiciji 5.1. Tedaj je množica*

$$\left\{T_k^h(\cos\tfrac{\pi\,\cdot}{2})\right\}_{k=0}^{N}$$

ortonormalna baza prostora $\mathcal{C}_N$ na intervalu $[-1,1]$.

Dokaz: Naj bo $g \in \mathcal{C}_N \setminus \mathcal{C}_{N-1}$. Ker je g soda funkcija na $[-1,1]$, se lahko omejimo na interval $[0,1]$. Preslikava $y = \cos\frac{\pi x}{2}$ slika interval $[0,1]$ vase in je obrnljiva s predpisom $x = \frac{2}{\pi}\cos^{-1} y$. Iz lastnosti (6.8) sledi, da je $g(\frac{2}{\pi}\cos^{-1} y)$ polinom v y na $[0,1]$, ki ga označimo s P_g. Poleg tega vsak polinom p stopnje kvečjemu N ustreza funkciji $p(\cos\frac{\pi x}{2}) \in \mathcal{C}_N$, saj polinomi Čebiševa prve vrste stopnje manjše ali enake N tvorijo bazo prostora polinomov stopnje kvečjemu N.

Ugotovimo, da je funkcija g ortogonalna na poljubno funkcijo $\tilde{g} \in \mathcal{C}_{N-1}$, ker je

$$\begin{aligned}\int_{-1}^{1} g(x)\tilde{g}(x)\,\mathrm{d}x &= 2\int_{1}^{0} g\left(\tfrac{2}{\pi}\cos^{-1} y\right)\tilde{g}\left(\tfrac{2}{\pi}\cos^{-1} y\right)\left(-\tfrac{2}{\pi}\right)\tfrac{1}{\sqrt{1-y^2}}\,\mathrm{d}y \\ &= \frac{4}{\pi}\int_{0}^{1} P_g(y)P_{\tilde{g}}(y)\frac{1}{\sqrt{1-y^2}}\,\mathrm{d}x = 0\end{aligned}$$

Če je $g(x) = T_N^h(\cos\frac{\pi x}{2})$, je $P_g(y) = T_N^h(y)$ in g je ortogonalna na vse funkcije v prostoru $\mathcal{C}_{N-1}$. Normalizacija, podana z enačbama (5.13 – 5.14) v definiciji 5.1, natančno ustreza normalizaciji funkcije $T_k^h(\cos\frac{\pi\cdot}{2})$ iz prostora $L^2(-1,1)$. Izrek je s tem dokazan. □

Izrek 6.4 *Naj bo $\{U_k^h\}$, $k \geq 0$, množica poldomenskih polinomov Čebiševa druge vrste opisanih v definiciji 5.2. Tedaj je množica*

$$\left\{U_k^h(\cos\tfrac{\pi\cdot}{2})\sin\tfrac{\pi\cdot}{2}\right\}_{k=0}^{N-1}$$

ortonormalna baza prostora $\mathcal{S}_N$ na intervalu $[-1,1]$.

Dokaz: Naj bo $g \in \mathcal{S}_N \setminus \mathcal{S}_{N-1}$. Ker je g liha funkcija, je $g(x)/\sin\frac{\pi x}{2}$ dobro definirana in soda funkcija na $[-1,1]$, zato se lahko ponovno omejimo na interval $[0,1]$ in uporabimo substitucijo $y = \cos\frac{\pi x}{2}$. Iz lastnosti (6.9) sledi, da je

$$\frac{g(\frac{2}{\pi}\cos^{-1} y)}{\sin(\frac{\pi}{2}\frac{2}{\pi}\cos^{-1} y)} = \frac{g(\frac{2}{\pi}\cos^{-1} y)}{\sqrt{1-y^2}}$$

polinom v y na $[0,1]$, ki ga označimo s Q_g. Poleg tega vsak polinom q stopnje kvečjemu N ustreza funkciji $q(\cos\frac{\pi x}{2})\sin\frac{\pi x}{2} \in \mathcal{S}_N$, saj polinomi Čebiševa druge vrste stopnje manjše ali enake $N-1$ tvorijo bazo prostora polinomov stopnje kvečjemu $N-1$.

Ugotovimo, da je funkcija g ortogonalna na poljubno funkcijo $\tilde{g} \in \mathcal{S}_{N-1}$, ker je

$$\begin{aligned}\int_{-1}^{1} g(x)\tilde{g}(x)\,\mathrm{d}x &= 2\int_{1}^{0} \frac{g\left(\frac{2}{\pi}\cos^{-1} y\right)}{\sqrt{1-y^2}}\frac{\tilde{g}\left(\frac{2}{\pi}\cos^{-1} y\right)}{\sqrt{1-y^2}}(1-y^2)\left(-\tfrac{2}{\pi}\right)\frac{1}{\sqrt{1-y^2}}\,\mathrm{d}y \\ &= \frac{4}{\pi}\int_{0}^{1} Q_g(y)Q_{\tilde{g}}(y)\sqrt{1-y^2}\,\mathrm{d}x = 0\end{aligned}$$

Če je $g(x) = U^h_{N-1}(\cos\frac{\pi x}{2})\sin\frac{\pi x}{2}$, je $Q_g(y) = U^h_{N-1}(y)$ in g je ortogonalna na vse funkcije iz prostora $\mathcal{S}_{N-1}$. Normalizacija, podana z enačbama (5.15 – 5.16) v definiciji 5.2, natančno ustreza normalizaciji $U^h_k(\cos\frac{\pi\cdot}{2})\sin\frac{\pi\cdot}{2}$ v prostoru $L^2(-1,1)$. Izrek je s tem dokazan. □

6.2 Poldomenska Čebišev-Fourierova vrsta

D. Huybrechs je v članku [35] obravnaval optimizacijski problem 1.2. Za izbiro $T = 2$ je za vsako s kvadratom integrabilno funkcijo $f \in L^2(-1,1)$ dokazal obstoj in enoličnost rešitve, ki jo je karakteriziral z uporabo ortonormalne baze iz razdelka 6.1 in ortogonalne projekcije v obliki *poldomenske Čebišev-Fourierove (HCF) vrste*

$$f(x) = \sum_{k=0}^{\infty} a_k T^h_k(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{\infty} b_k U^h_k(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2}, \tag{6.10}$$

kjer koeficiente a_k in b_k izračunamo s formulama

$$a_k = \int_{-1}^{1} f(x) T^h_k(\cos\tfrac{\pi x}{2})\,\mathrm{d}x, \tag{6.11}$$

$$b_k = \int_{-1}^{1} f(x) U^h_k(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2}\,\mathrm{d}x. \tag{6.12}$$

Rešitev problema 1.2 je podana s spodnjim izrekom, ki je dokazan v [35].

Izrek 6.5 *Za dano funkcijo $f \in L^2(-1,1)$, je rešitev problema 1.2 odrezana poldomenska Čebišev-Fourierova vrsta*

$$g_N(x) = \sum_{k=0}^{N} a_k T^h_k(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k U^h_k(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2}. \tag{6.13}$$

Dokaz: Trditev izreka sledi direktno iz ortogonalne projekcije na ortonormalni bazi, ki sta opisani v izrekih 6.3 in 6.4. □

Poldomenska Čebišev-Fourierova vrsta (6.10) je sestavljena iz trigonometričnih funkcij $\cos\frac{\pi k\,\cdot}{2}$ in $\sin\frac{\pi(k+1)\,\cdot}{2}$, ki so z uporabo enakosti (6.8) in (6.9) razvite po poldomenskih polinomih Čebiševa prve in druge vrste definiranih in opisanih v razdelku 5.3. Ti polinomi so razviti po funkciji $\cos\frac{\pi\cdot}{2}$ in nekateri pomnoženi s funkcijo $\sin\frac{\pi\cdot}{2}$. To pomeni, da je vrsta (6.10) reorganizirana Fourierova vrsta, zato lahko pri analizi konvergence uporabimo mnoga orodja iz analize konvergence Fourierovih vrst za periodične probleme, čeprav v osnovi obravnavamo neperiodične probleme.

V nadaljevanju bomo obravnavali konvergenco poldomenske Čebišev-Fourierove vrste (6.10). Le-ta konvergira za vsako funkcijo, hitrost konvergence pa je odvisna od stopnje gladkosti oz. analitičnosti funkcije in

je v večini primerov eksponentna in ne zgolj polinomska. Ta lastnost je značilna tudi za konvergenco Fourierove vrste za periodične funkcije ter vrste Čebiševa za neperiodične funkcije. Za dokaz konvergence se omejimo le na analitične funkcije f. Spodnje leme in izrek so dokazane v [35].

Lema 6.6 *Naj bo funkcija f soda in analitična. Tedaj je f periodična na intervalu $[-2,2]$ natanko tedaj, ko je f soda glede na število 2, tj. graf funkcije f je zrcalen glede na premico $x = 2$.*

Lema 6.7 *Naj bo funkcija f liha in analitična. Tedaj je f periodična na intervalu $[-2,2]$ natanko tedaj, ko je f liha glede na število 2, tj. graf funkcije f je zrcalen glede na točko $(2,0)$.*

Vsako funkcijo f lahko zapišemo kot vsoto sode in lihe funkcije

$$f(x) = f_s(x) + f_\ell(x), \tag{6.14}$$

kjer je denimo

$$f_s(x) = \frac{f(x) + f(-x)}{2} \quad \text{in} \quad f_\ell(x) = \frac{f(x) - f(-x)}{2}.$$

Definirajmo funkciji

$$f_1(y) = f_s\left(\tfrac{2}{\pi}\cos^{-1} y\right) = f_s(x) \tag{6.15}$$

in

$$f_2(y) = \frac{f_\ell\left(\tfrac{2}{\pi}\cos^{-1} y\right)}{\sqrt{1-y^2}} = \frac{f_\ell(x)}{\sqrt{1-y^2}} \tag{6.16}$$

ter konstanto

$$E = 3 + 2\sqrt{2}. \tag{6.17}$$

Lema 6.8 *Naj bo funkcija f soda in analitična na okolici intervala $[-1,1]$ in naj bo*

$$\tilde{f}(y) = f\left(\tfrac{2}{\pi}\cos^{-1} y\right).$$

Če je $\tilde{f}$ analitična na območju omejenem z elipso, ki ima gorišči v točkah 0 in 1, vsota glavnih polosi elipse pa je enaka $\frac{\rho}{2}$, potem velja

$$\rho \leq E,$$

razen v primeru, ko je f analitična in periodična na intervalu $[-2,2]$.

Izrek 6.9 *Naj bo $\tilde{f}(y) = f(x)$, kjer je $x = \frac{2}{\pi}\cos^{-1} y$ analitična funkcija na območju omejenem z elipso z večjo polosjo dolžine R in goriščema v točkah 0 in 1. Pripadajočo domeno, na kateri je analitična funkcija f, označimo z*

$D(R)$. Če je f analitična na domeni $D(R)$, kjer je $R > \frac{1}{2}$, tedaj rešitev g_N problema 1.2 zadošča

$$\|f - g_N\|_{L^2} \sim \rho^{-N}, \tag{6.18}$$

kjer je

$$\rho = \min(E, 2R + \sqrt{4R^2 - 1}), \tag{6.19}$$

razen v primeru, ko je f analitična in periodična na intervalu $[-2, 2]$.

Dokaz: Najprej uporabimo lemo 6.8 za funkciji f_s in $f_\ell / \sin \frac{\pi \cdot}{2}$. Obe izbrani funkciji sta sodi. Funkcija f_s je na domeni $D(R)$ analitična po konstrukciji, medtem ko ima lahko funkcija $f_\ell / \sin \frac{\pi \cdot}{2}$ pole v točkah $x = \pm 2n$ za $n \geq 0$.

Če je f_ℓ periodična na intervalu $[-2, 2]$, potem sledi po lemi 6.6, da je enaka 0 v vseh možnih polih. Torej je funkcija $f_\ell / \sin \frac{\pi \cdot}{2}$ analitična na domeni $D(R)$. V nasprotnem primeru, ko funkcija f_ℓ ni periodična, je $f_\ell / \sin \frac{\pi \cdot}{2}$ analitična samo na območju $D = D(R) \cap D(\frac{3}{2})$.

Naj bosta f_1 in f_2 funkciji definirani z enačbama (6.15) in (6.16). Označimo s p_N polinom stopnje N, ki je najboljša aproksimacija f_1 po metodi najmanjših kvadratov glede na utež $\frac{4}{\pi} \frac{1}{\sqrt{1-y^2}}$, ter s q_N polinom stopnje $N-1$, ki je najboljša aproksimacija f_2 po metodi najmanjših kvadratov glede na utež $\frac{4}{\pi}\sqrt{1-y^2}$. Ker so sode in lihe funkcije na intervalu $[-1, 1]$ ortogonalne, velja zveza

$$\|f - g_N\|_{L^2}^2 = \|f_s - p_N(\cos \tfrac{\pi x}{2})\|_{L^2}^2 + \|f_\ell - q_N(\cos \tfrac{\pi x}{2}) \sin \tfrac{\pi x}{2}\|_{L^2}^2. \tag{6.20}$$

Naj bo ρ definiran z enačbo (6.19). Iz izreka 2.8 sledi

$$\frac{4}{\pi} \int_0^1 (f_1(y) - p_N(y))^2 \frac{1}{\sqrt{1-y^2}} \, \mathrm{d}y \sim \rho^{-2N} \tag{6.21}$$

in

$$\frac{4}{\pi} \int_0^1 (f_2(y) - q_N(y))^2 \sqrt{1-y^2} \, \mathrm{d}y \sim \rho^{-2N}. \tag{6.22}$$

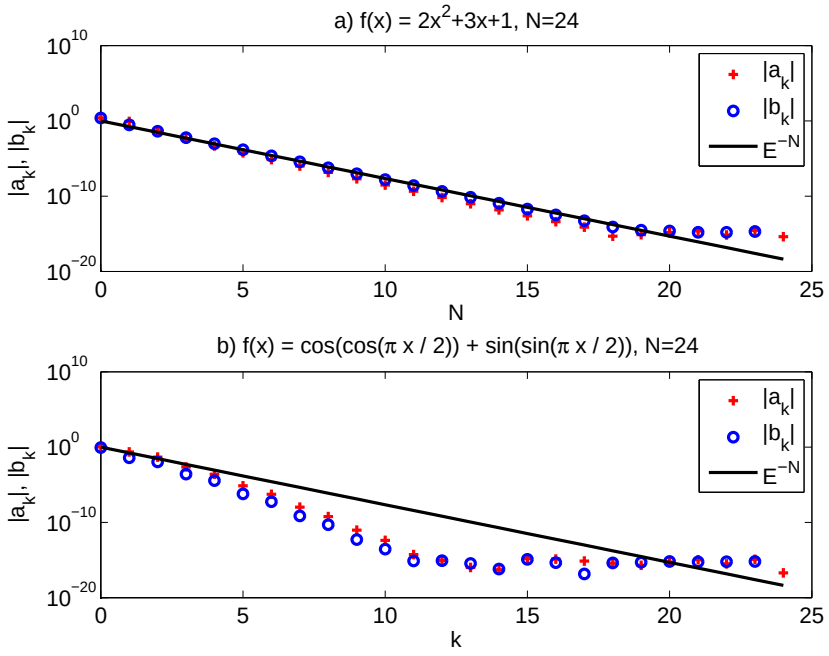
Opazimo, da sta integrala v relacijah (6.21) in (6.22) natančno normi desne strani enačbe (6.20), kjer upoštevamo $y = \cos \frac{\pi x}{2}$. Izrek je s tem dokazan. □

Posledica 6.10 *Pod pogoji izreka 6.9, koeficienti a_k in b_k, ki so definirani z enačbama (6.11) in (6.12), v razvoju poldomenske Čebišev-Fourierove vrste g_N (6.13), zadoščajo pogoju*

$$a_k, b_k \sim \rho^{-N}.$$

Slika 6.1 prikazuje padanje absolutnih vrednosti koeficientov a_k in b_k v razvoju dane funkcije v poldomensko Čebišev-Foureirovo vrsto za dve funkciji pri fiksni izbiri odreznega števila N. Funkcija $f(x) = 2x^2 + 3x + 1$

je cela, toda ni periodična, medtem ko je funkcija $f(x) = \cos(\cos\frac{\pi x}{2}) + \sin(\sin\frac{\pi x}{2})$ cela in periodična na intervalu $[-2, 2]$. V obeh izbranih primerih je $N = 24$. Na obeh slikah je vidna stopnja konvergence E^{-N}, ki je napovedana v posledici 6.10, kjer je E definiran z enačbo (6.17). Hitrost konvergence poldomenske Čebišev-Fourierove vrste je v obeh primerih eksponentna.

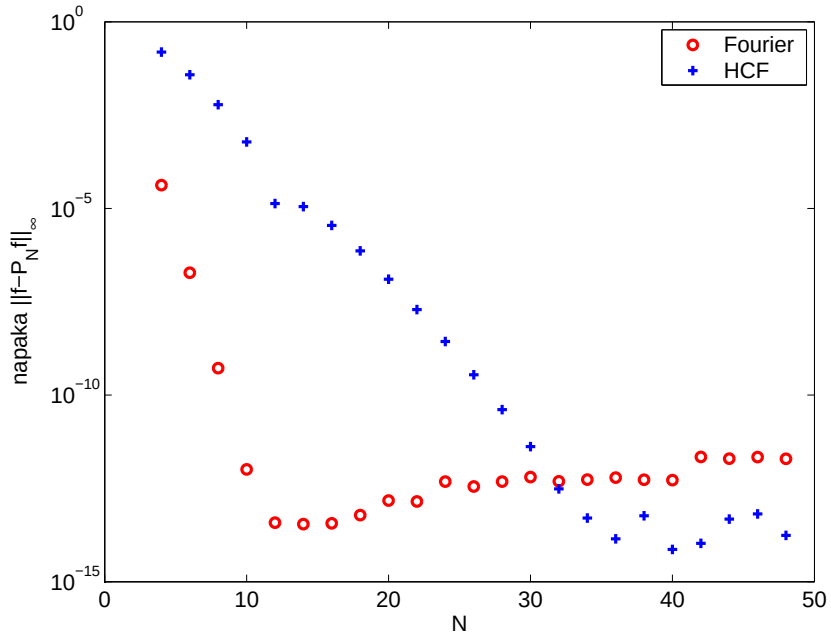


Slika 6.1: Padanje absolutnih vrednosti spektralnih koeficientov a_k (rdeči plusi) in b_k (modri krogci), $1 \le k \le N$ v HCF vrsti za funkciji a) $f(x) = 2x^2 + 3x + 1$ in b) $f(x) = \cos(\cos\frac{\pi x}{2}) + \sin(\sin\frac{\pi x}{2})$. V obeh primerih je $N = 24$, črna črta pa prikazuje napovedano stopnjo padanja E^{-N}.

6.3 Primerjava spektralnih aproksimacij

Če želimo aproksimirati gladko, tj. neskončnokrat zvezno odvedljivo, in periodično funkcijo, je prava izbira, da to storimo z razvojem v trigonometrično Fourierovo vrsto, kjer Fourierove koeficiente v razvoju vrste izračunamo s hitro Fourierovo transformacijo (FFT). To lahko naredimo stabilno in učinkovito, saj potrebujemo le $\mathcal{O}(N \log N)$ operacij za izračun prvih $2N+1$ Fourierovih koeficientov. Analiza konvergence, ki smo jo opisali v razdelku 2.2, kaže na to, da dobimo spektralno konvergenco, tj. eksponentno padanje napake. Podoben rezultat smo dobili v razdelku 6.2 za aproksimacijo funkcije s poldomensko Čebišev-Fourierovo vrsto.

Slika 6.2 prikazuje primerjavo natančnosti aproksimacije, oz. hitrosti konvergence, pri razvoju periodične funkcije $f(x) = \cos(\cos\frac{\pi x}{2})$ v Fourierovo in poldomensko Čebišev-Fourierovo vrsto. Primerjamo maksimalno absolutno vrednost napake glede na število členov N v odrezani vrsti. V obeh primerih opazimo spektralno konvergenco, tj. eksponentno padanje maksimalne napake.



Slika 6.2: Primerjava spektralne konvergence za gladko in periodično funkcijo $f(x) = \cos(\cos\frac{\pi x}{2})$ pri aproksimaciji s Fourierovo (rdeči krogci) in poldomensko Čebišev-Fourierovo vrsto (modri plusi) v odvisnosti od števila členov N v odrezani vrsti.

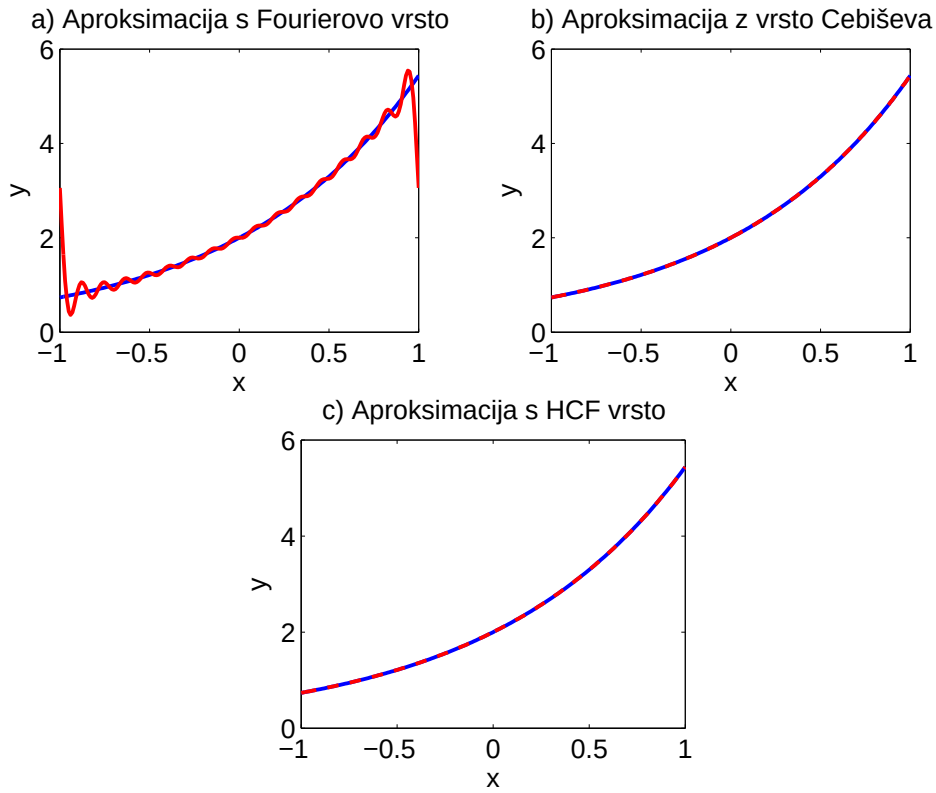
Na sliki 6.2 opazimo tudi, da aproksimacija s Fourierovo vrsto v tem primeru konvergira precej hitreje kot aproksimacija s poldomensko Čebišev-Fourierovo vrsto, kar pomeni, da je za periodične funkcije prava izbira za aproksimacijo še vedno tista s Fourierovo vrsto.

Situacija se popolnoma spremeni, če funkcija bodisi ni gladka, bodisi ni periodična. V tem primeru nastopi Gibbsov fenomen, ki se kaže v odsotnosti konvergence po točkah, počasnem padanju koeficientov Fourierove vrste ter oscilacijah v okolici točk nezveznosti in/ali robov intervala.

Težavam v zvezi z Gibbsovim fenomenom se lahko izognemo, če dano funkcijo aproksimiramo bodisi z vrsto Čebiševa, bodisi s poldomensko Čebišev-Fourierovo vrsto. Pri tem za diskretizacijo intervala $[-1, 1]$, v nasprotju s Fourierovo vrsto, kjer uporabimo ekvidistantne točke, uporabimo točke Čebiševa, ki so gostejše v okolici robnih točk intervala. Gibbsovega fenomena

v tem primeru ni.

Na sliki 6.3 je prikazana neperiodična, toda gladka funkcija $f(x) = 2e^x$, ki jo aproksimiramo z vsemi tremi obravnavanimi vrstami. Aproksimacija s Fourierovo vrsto in s poldomensko Čebišev-Fourierovo vrsto je narejena za odrezno število $N = 16$, aproksimacija z vrsto Čebiševa pa za odrezno število $N = 32$, kar pa v vseh treh primerih pomeni izračun 33 spektralnih koeficientov v razvoju vrste. Na sliki a) (aproksimacija s s Fourierovo vrsto) opazimo oscilacije predvsem v okolici robov intervala, kar pomeni, da v tem primeru dobimo Gibbsov fenomen. Povsem drugačna je situacija na slikah b) (aproksimacija z vrsto Čebiševa) in c) (aproksimacija s poldomensko Čebišev-Fourierovo vrsto), kjer oscilacije izginejo, torej tudi Gibbsovega fenomena ni.

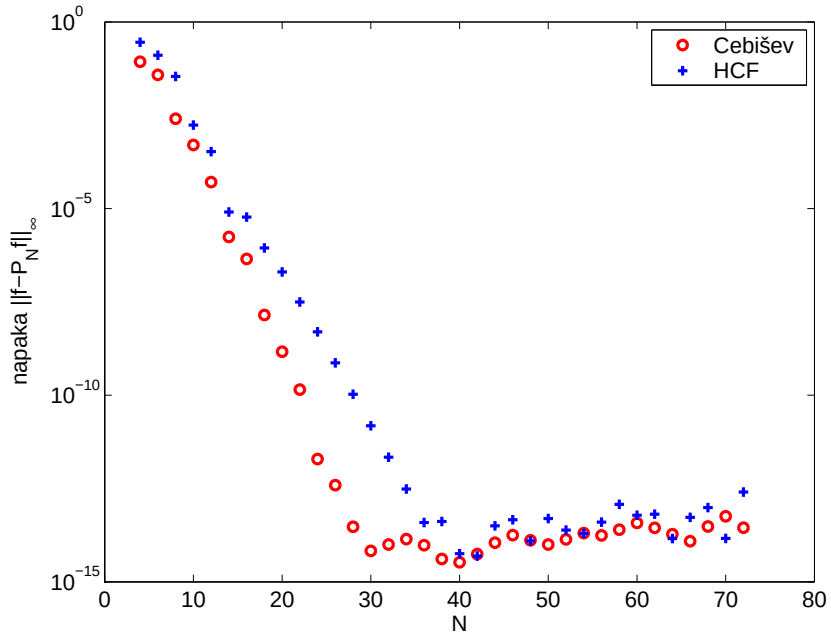


Slika 6.3: Gibbsov fenomen pri aproksimaciji funkcije $f(x) = 2e^x$ (modra črta) s Fourierovo vrsto (rdeča črta) za $N = 16$ na sliki a). Pri aproksimaciji dane funkcije z vrsto Čebiševa (rdeča prekinjena črta) za $N = 32$ na sliki b) in s poldomensko Čebišev-Fourierovo vrsto (rdeča prekinjena črta) za $N = 16$ na sliki c), Gibbsovega fenomena ni. Obe črti se prekrivata.

Za funkcije, ki niso periodične, je tako primernejša aproksimacija z

vrsto Čebiševa ali s poldomensko Čebišev-Fourierovo vrsto. V obeh primerih dobimo za neperiodične gladke funkcije spektralno konvergenco, tj. maksimalna absolutna vrednost napake glede na odrezno število N eksponentno hitro pada. Seveda to velja samo za gladke oz. analitične funkcije. Konvergenca vrste Čebiševa je podrobno opisana v podrazdelku 2.3.4, konvergenca poldomenske Čebišev-Fourierove vrste pa v razdelku 6.2.

Slika 6.4 prikazuje primerjavo natančnosti aproksimacije, oz. hitrosti konvergence, pri razvoju funkcije $f(x) = \cos(2e^x)$ v vrsto Čebiševa in poldomensko Čebišev-Fourierovo vrsto. Primerjamo maksimalno absolutno vrednost napake glede na število členov N v odrezani vrsti. V obeh primerih opazimo spektralno konvergenco, tj. eksponentno padanje maksimalne napake. Opazimo še, da sta aproksimaciji z vrsto Čebiševa in s poldomensko Čebišev-Fourierovo vrsto v tem primeru povsem primerljivi, le da prva konvergira nekoliko hitreje.

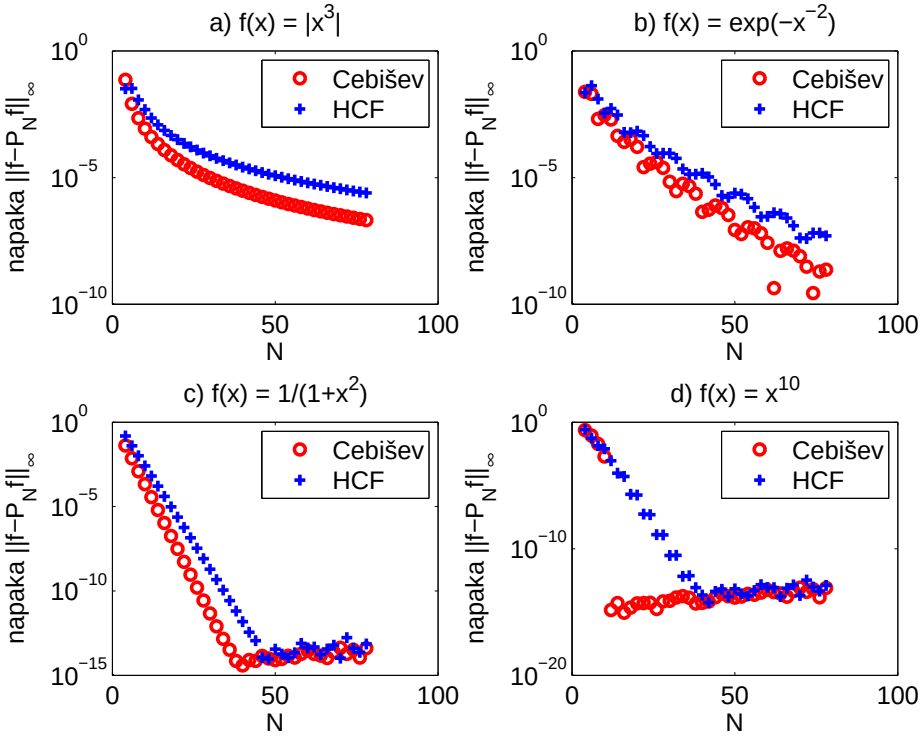


Slika 6.4: Primerjava spektralne konvergence za gladko in neperiodično funkcijo $f(x) = \cos(2e^x)$ pri aproksimaciji z vrsto Čebiševa (rdeči krogci) in s poldomensko Čebišev-Fourierovo vrsto (modri plusi) v odvisnosti od števila členov N v odrezani vrsti.

V primeru, ko funkcije niso gladke, ampak zgolj nekajkrat zvezno odvedljive, konvergenca ni eksponentna, ampak maksimalna absolutna vrednost napake pada s stopnjo gladkosti dane funkcije glede na število členov N v odrezani vrsti. To velja za aproksimacijo funkcije z vsemi tremi obravnavanimi vrstami.

Na sliki 6.5 je prikazana primerjava padanja maksimalne absolutne vrednosti napake glede na število členov N v odrezani vrsti za aproksimacijo dane funkcije z vrsto Čebiševa in poldomensko Čebišev-Fourierovo vrsto. Dane so štiri neperiodične funkcije za katere narašča stopnja gladkosti: a) funkcija $f(x) = |x^3|$ ima tretji odvod z omejeno variacijo, b) funkcija $f(x) = \mathrm{e}^{-x^{-2}}$ je gladka, toda ni analitična, c) funkcija $f(x) = 1/(1+x^2)$ je analitična na okolici intervala $[-1, 1]$, d) funkcija $f(x) = x^{10}$ pa je polinom, ki je za aproksimacijo s polinomi analogen funkciji z omejenim naborom frekvenc za aproksimacijo s Fourierovo vrsto.

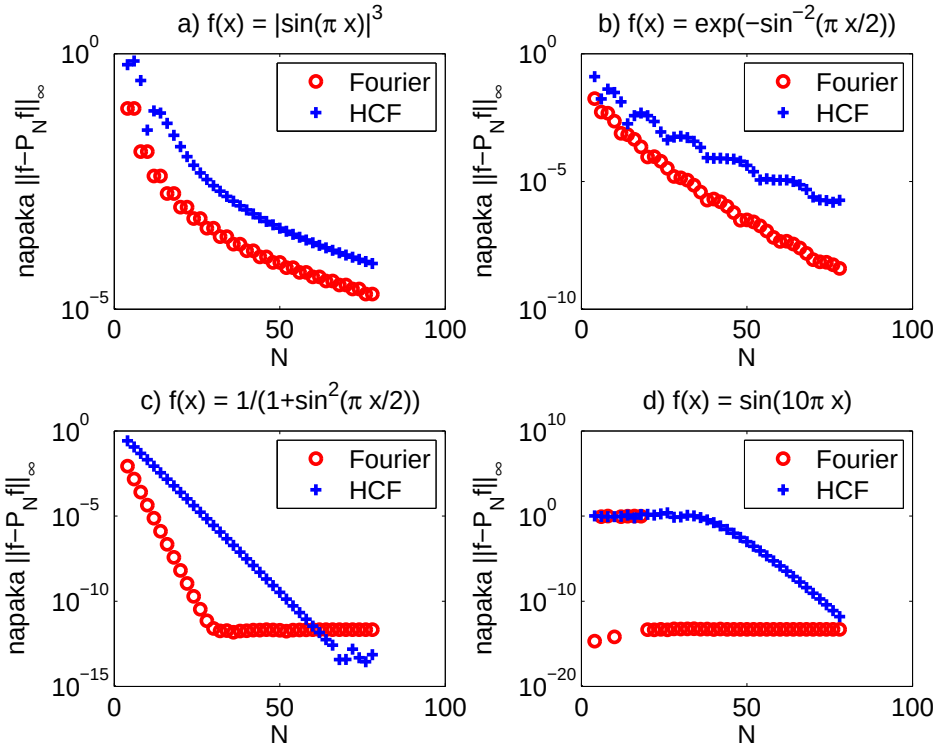
Opazimo zelo podobno obnašanje za oba razreda aproksimacij, razen v zadnjem primeru d) na sliki desno spodaj, kjer je aproksimacija z vrsto Čebiševa bistveno boljša. To ni presenetljivo, saj je vrsta Čebiševa sestavljena iz ortogonalnih polinomov, poldomenska Čebišev-Fourierova vrsta pa iz trigonometričnih funkcij, kar pomeni, da lahko poljuben polinom točno aproksimiramo z vrsto Čebiševa, če je le N dovolj velik.



Slika 6.5: Primerjava hitrosti konvergence za štiri neperiodične funkcije naraščajoče gladkosti: a) $f(x) = |x^3|$, b) $f(x) = \mathrm{e}^{-x^{-2}}$, c) $f(x) = 1/(1+x^2)$ in d) $f(x) = x^{10}$. Maksimalno absolutno vrednost napake pri aproksimaciji z vrsto Čebiševa (rdeči krogci) in s poldomensko Čebišev-Fourierovo vrsto (modri plusi) gledamo v odvisnosti od števila členov N v odrezani vrsti.

Na sliki 6.6 je prikazana primerjava padanja maksimalne absolutne vrednosti napake glede na število členov N v odrezani vrsti za aproksimacijo dane funkcije s Fourierovo in poldomensko Čebišev-Fourierovo vrsto. Dane so štiri periodične funkcije za katere narašča stopnja gladkosti: a) funkcija $f(x) = |\sin x|^3$ ima tretji odvod z omejeno variacijo, b) funkcija $f(x) = \mathrm{e}^{-\sin^{-2}(x/2)}$ je gladka, toda ni analitična, c) funkcija $f(x) = 1/(1+\sin^2(x/2))$ je analitična na okolici intervala $[-1, 1]$ v kompleksni ravnini, d) funkcija $f(x) = \sin(10x)$ pa je funkcija z omejenim naborom frekvenc.

Opazimo, da je aproksimacija s Fourierovo vrsto za periodične funkcije boljša kot aproksimacija s poldomensko Čebišev-Fourierovo vrsto, saj v vseh primerih napaka pada hitreje. To je še posebej vidno na zadnjem primeru d) na sliki desno spodaj, kjer aproksimiramo bazno trigonometrično funkcijo, ki nastopa v razvoju v Fourierovo vrsto. Kljub temu pa tudi napaka aproksimacije s poldomensko Čebišev-Fourierovo vrsto pada dovolj hitro.



Slika 6.6: Primerjava hitrosti konvergence za štiri periodične funkcije naraščajoče gladkosti: a) $f(x) = |\sin x|^3$, b) $f(x) = \mathrm{e}^{-\sin^{-2}(x/2)}$, c) $f(x) = 1/(1+\sin^2(x/2))$ in d) $f(x) = \sin(10x)$. Maksimalno absolutno vrednost napake pri aproksimaciji s Fourierovo (rdeči krogci) in s poldomensko Čebišev-Fourierovo vrsto (modri plusi) gledamo v odvisnosti od števila členov N v odrezani vrsti.

Poglavje 7

Linearni dvotočkovni robni problemi v eni dimenziji

V tretjem poglavju smo opisali kratek uvod v teorijo spektralnih metod, kjer smo obravnavali tako konstrukcijo Fourierovih kot tudi metod Čebiševa. Poleg tega smo opisali osnovna orodja za analizo konvergence in napake. V tem poglavju pa bomo konstruirali nov razred kolokacijskih spektralnih metod za linearne dvotočkovne robne probleme v eni dimenziji, ki jih imenujemo *Čebišev-Fourierove spektralne metode* (ang. *Chebyshev-Fourier spectral methods*). Problemi tipa 1.3 so podrobno obravnavani v različnih monografijah, npr. v U. M. Ascher, R. M. M. Mattheij in R. D. Russell [4] in [5], in so oblike

$$\mathcal{L}u(x) = f(x), \qquad x \in [-1, 1], \tag{7.1}$$

z robnimi pogoji

$$\mathcal{B}u(y) = 0, \qquad y \in \{-1, 1\}, \tag{7.2}$$

kjer je $\mathcal{L}$ linearni diferencialni operator (1.6) in $\mathcal{B}$ par linearnih robnih diferencialnih operatorjev, ki ustrezajo Dirichletovim, Neumannovim ali mešanim (Robinovim) robnim pogojem. V nadaljevanju se omejimo na Dirichletove robne pogoje. Numerično rešitev danega problema iščemo v obliki *poldomenske Čebišev-Fourierove vrste* (6.10)

$$f(x) = \sum_{k=0}^{\infty} a_k T_k^h(\cos \tfrac{\pi x}{2}) + \sum_{k=0}^{\infty} b_k U_k^h(\cos \tfrac{\pi x}{2}) \sin \tfrac{\pi x}{2}, \tag{7.3}$$

kjer za izračun spektralnih koeficientov uporabimo metodo kolokacije. Za konstrukcijo metode potrebujemo dve operatorski matriki, in sicer matriko odvajanja D in multiplikativno matriko F. Izračun elementov teh dveh matrik je podrobno opisan v članku B. Orel in A. Perne [46]. Algoritem za konstrukcijo iskane psevdospektralne metode, ki jo imenujemo *Čebišev-Fourierova kolokacijska metoda* (CFC metoda), pa je opisan v poročilu

B. Orel in A. Perne [45]. V nadaljevanju poglavja bomo obravnavali konstrukcijo te metode ter analizo konvergence in napake. Metodo bomo primerjali na numeričnih zgledih s standardnimi (psevdo)spektralnimi metodami, npr. s *kolokacijsko metodo Čebiševa* (CC metoda).

7.1 Operatorske matrike

Za konstrukcijo iskanih spektralnih Čebišev-Fourierovih metod potrebujemo dve operatorski matriki odvajanja in množenja, ki ju konstruiramo na podoben način, kot smo to opisali pri obravnavi Fourierovih spektralnih metod v podrazdelku 3.4.1 ter spektralnih metod Čebiševa v podrazdelku 3.4.2. Numerično rešitev robnega problema (7.1 – 7.2) iščemo v obliki odrezane poldomenske Čebišev-Fourierove (HCF) vrste.

Matrika odvodov D je transformacijska matrika, ki pretvori spektralne koeficiente a_k in b_k odrezane HCF vrste (7.3) za funkcijo f, v spektralne koeficiente a'_j in b'_j odrezane HCF vrste za prvi odvod funkcije f. Spektralne koeficiente a''_j in b''_j odrezane HCF vrste za drugi odvod dobimo z uporabo matrike D^2.

Multiplikacijska matrika F pa je transformacijska matrika, ki pretvori spektralne koeficiente a_k in b_k odrezane HCF vrste (7.3) za funkcijo f, v spektralne koeficiente $\tilde{a}_j$ in $\tilde{b}_j$ odrezane HCF vrste za produkt odrezane HCF vrste neke znane funkcije g z znanimi koeficienti in odrezane HCF vrste za funkcijo f. V primeru, da je g konstantna funkcija, je ta matrika skalarna, tj. identična matrika pomnožena z nekim skalarjem, sicer pa je polna. V konkretnem primeru je g ena izmed koeficientnih funkcij α, β ali γ, ki nastopajo v diferencialni enačbi (7.1).

7.1.1 Matrika odvodov

Začnimo z odrezano poldomensko Čebišev-Fourierovo vrsto (7.3) za poljubno funkcijo $f \in L^2(-1,1)$

$$f(x) \approx f_N(x) = \sum_{k=0}^{N} a_k T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \qquad (7.4)$$

kjer je N odrezno število, ki določa število členov vrste, tj. pri odreznem številu N ima HCF vrsta $2N+1$ členov. Za dan N potrebujemo $2N+1$ ortogonalnih polinomov: $N+1$ poldomenskih polinomov Čebiševa prve in N poldomenskih polinomov Čebiševa druge vrste.

Najprej izračunamo prvi odvod $\frac{\mathrm{d}f_N}{\mathrm{d}x}$ odrezane HCF vrste f_N, ki je definirana z enačbo (7.4)

$$\begin{aligned}\frac{\mathrm{d}f_N}{\mathrm{d}x}(x) &= -\frac{\pi}{2}\sum_{k=0}^{N} a_k \dot{T}_k^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2} \\ &\quad +\frac{\pi}{2}\sum_{k=0}^{N-1} b_k \left(U_k^h(\cos\tfrac{\pi x}{2})\cos\tfrac{\pi x}{2} - \dot{U}_k^h(\cos\tfrac{\pi x}{2})(1-\cos^2\tfrac{\pi x}{2})\right), \quad (7.5)\end{aligned}$$

kjer s $\dot{T}_k^h(t) = \frac{\mathrm{d}}{\mathrm{d}t}T_k^h(t)$ in $\dot{U}_k^h(t) = \frac{\mathrm{d}}{\mathrm{d}t}U_k^h(t)$ označimo prva odvoda ortogonalnih polinomov T_k^h in U_k^h po spremenljivki $t = \cos\frac{\pi x}{2}$. Polinome $\dot{T}_k^h(t)$ aproksimiramo z odrezano vrsto, sestavljeno iz polinomov U_j^h, polinome $\dot{U}_k^h(1-t^2)$ in $U_k^h(t)t$ pa z odrezano vrsto, sestavljeno iz polinomov T_j^h

$$\dot{T}_k^h(t) = \sum_{j=0}^{k-1} p_j^k U_j^h(t), \quad p_j^k = \tfrac{4}{\pi}\mathcal{L}_U\left(\dot{T}_k^h(t)U_j^h(t)\right), \quad (7.6)$$

$$\dot{U}_k^h(t)t^2 = \sum_{j=0}^{k+1} q_j^k T_j^h(t), \quad q_j^k = \tfrac{4}{\pi}\mathcal{L}_T\left(t^2\dot{U}_k^h(t)T_j^h(t)\right), \quad (7.7)$$

$$\dot{U}_k^h(t) = \sum_{j=0}^{k-1} r_j^k T_j^h(t), \quad r_j^k = \tfrac{4}{\pi}\mathcal{L}_T\left(\dot{U}_k^h(t)T_j^h(t)\right), \quad (7.8)$$

$$U_k^h(t)t = \sum_{j=0}^{k+1} s_j^k T_j^h(t), \quad s_j^k = \tfrac{4}{\pi}\mathcal{L}_T\left(tU_k^h(t)T_j^h(t)\right). \quad (7.9)$$

Koeficiente p_j^k, q_j^k, r_j^k in s_j^k, ki so definirani z enačbami (7.6 – 7.9), izračunamo iz pogojev ortogonalnosti. Prvi odvod (7.5) je z uporabo teh oznak enak

$$\begin{aligned}\frac{\mathrm{d}f_N}{\mathrm{d}x}(x) &= \frac{\pi}{2}\sum_{k=0}^{N-1} b_k \left(\sum_{j=0}^{k+1} q_j^k T_j^h(\cos\tfrac{\pi x}{2}) + \sum_{j=0}^{k+1} s_j^k T_j^h(\cos\tfrac{\pi x}{2}) - \sum_{j=0}^{k-1} r_j^k T_j^h(\cos\tfrac{\pi x}{2})\right) \\ &\quad -\frac{\pi}{2}\sum_{k=0}^{N} a_k \left(\sum_{j=0}^{k-1} p_j^k U_j^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2}\right) \quad (7.10) \\ &= \sum_{j=0}^{N}\left(\frac{\pi}{2}\sum_{k=\max\{j-1,0\}}^{N-1}\left(q_j^k + s_j^k - r_j^k\right) b_k\right) T_j^h(\cos\tfrac{\pi x}{2}) \\ &\quad +\sum_{j=0}^{N-1}\left(\frac{\pi}{2}\sum_{k=j+1}^{N} -p_j^k a_k\right) U_j^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2} \quad (7.11) \\ &= \sum_{j=0}^{N} a_j' T_j^h(\cos\tfrac{\pi x}{2}) + \sum_{j=0}^{N-1} b_j' U_j^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2}. \quad (7.12)\end{aligned}$$

Funkcionala $\mathcal{L}_T$ in $\mathcal{L}_U$ sta definirana z enačbama (5.17) in (5.23). V vrstici (7.11) zamenjamo vrstni red seštevanja, kjer upoštevamo, da so koeficienti r_j^j in r_j^{j-1} enaki 0. Poleg tega so tudi koeficienti q_j^0 enaki 0. Nazadnje v vrstici (7.12) koeficiente odrezane HCF vrste $\frac{\mathrm{d}f_N}{\mathrm{d}x}$ označimo z a'_j in b'_j

$$a'_j = \frac{\pi}{2} \sum_{k=\max\{j-1,0\}}^{N-1} \left(q_j^k + s_j^k - r_j^k\right) b_k, \quad j = 0, 1, \ldots, N, \quad (7.13)$$

$$b'_j = -\frac{\pi}{2} \sum_{k=j+1}^{N} p_j^k a_k, \quad j = 0, 1, \ldots, N-1. \quad (7.14)$$

Enačbe (7.13) in (7.14) so linearne, zato jih lahko predstavimo v matrični obliki

$$\mathbf{u}' = D\,\mathbf{u}, \quad (7.15)$$

kjer je $\mathbf{u} = (a_0, \ldots, a_N, b_0, \ldots, b_{N-1})^T$ vektor spektralnih koeficientov odrezane HFC vrste za funkcijo f in $\mathbf{u}' = \left(a'_0, \ldots, a'_N, b'_0, \ldots, b'_{N-1}\right)^T$ vektor spektralnih koeficientov odvoda odrezane HFC vrste za funkcijo f. Ker so koeficienti a'_j odvisni samo od koeficientov b_k in podobno koeficienti b'_j le od koeficientov a_k, je operatorska matrika odvajanja $D \in \mathbb{R}^{(2N+1)\times(2N+1)}$ bločno antidiagonalna

$$D = \tfrac{\pi}{2} \begin{bmatrix} 0 & H_1 \\ H_2 & 0 \end{bmatrix}, \quad (7.16)$$

kjer je $H_1 \in \mathbb{R}^{(N+1)\times N}$ in $H_2 \in \mathbb{R}^{N\times(N+1)}$. Elementi teh dveh matrik so

$$H_1 = [h_{jk}^1]_{j,k}, \qquad h_{jk}^1 = \begin{cases} q_{j-1}^{k-1} + s_{j-1}^{k-1} - r_{j-1}^{k-1}, & k > j, \\ q_{j-1}^{k-1} + s_{j-1}^{k-1}, & j-1 \le k \le j, \\ 0, & k < j-1, \end{cases} \quad (7.17)$$

$$H_2 = [h_{jk}^2]_{j,k}, \qquad h_{jk}^2 = \begin{cases} -p_{j-1}^{k-1}, & k > j, \\ 0, & k \le j. \end{cases} \quad (7.18)$$

Bloka H_1 in H_2 sta za primer $N = 3$ enaka

$$H_1 = \begin{bmatrix} q_0^0 + s_0^0 & q_0^1 + s_0^1 - r_0^1 & q_0^2 + s_0^2 - r_0^2 \\ q_1^0 + s_1^0 & q_1^1 + s_1^1 & q_1^2 + s_1^2 - r_1^2 \\ 0 & q_2^1 + s_2^1 & q_2^2 + s_2^2 \\ 0 & 0 & q_3^2 + s_3^2 \end{bmatrix}, \; H_2 = \begin{bmatrix} 0 & -p_0^1 & -p_0^2 & -p_0^3 \\ 0 & 0 & -p_1^2 & -p_1^3 \\ 0 & 0 & 0 & -p_2^3 \end{bmatrix}.$$

Operatorska matrika odvajanja D je za primer $N = 3$ enaka

$$D = \begin{bmatrix} 0 & 0 & 0 & 0 & 1.4142 & -2.2706 & 2.8824 \\ 0 & 0 & 0 & 0 & 0.6837 & 4.6970 & -5.9625 \\ 0 & 0 & 0 & 0 & 0 & 1.3170 & 7.8416 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.9619 \\ 0 & -3.6091 & 3.8474 & -4.5910 & 0 & 0 & 0 \\ 0 & 0 & -7.4942 & 7.3713 & 0 & 0 & 0 \\ 0 & 0 & 0 & -11.3192 & 0 & 0 & 0 \end{bmatrix}.$$

Nato izračunamo še drugi odvod $\frac{\mathrm{d}^2 f_N}{\mathrm{d}x^2}$ odrezane HCF vrste f_N, ki je definirana z enačbo (7.4)

$$\frac{\mathrm{d}^2 f_N}{\mathrm{d}x^2} = \sum_{j=0}^{N} a_j'' T_j^h(\cos \tfrac{\pi x}{2}) + \sum_{j=0}^{N-1} b_j'' U_j^h(\cos \tfrac{\pi x}{2}) \sin \tfrac{\pi x}{2}. \tag{7.19}$$

Če označimo $\mathbf{u}$ kot prej in z $\mathbf{u}'' = \left(a_0'', \ldots, a_N'', b_0'', \ldots, b_{N-1}''\right)^T$ vektor spektralnih koeficientov drugega odvoda odrezane HFC vrste za funkcijo f, dobimo linearno zvezo

$$\mathbf{u}'' = D^2\, \mathbf{u}. \tag{7.20}$$

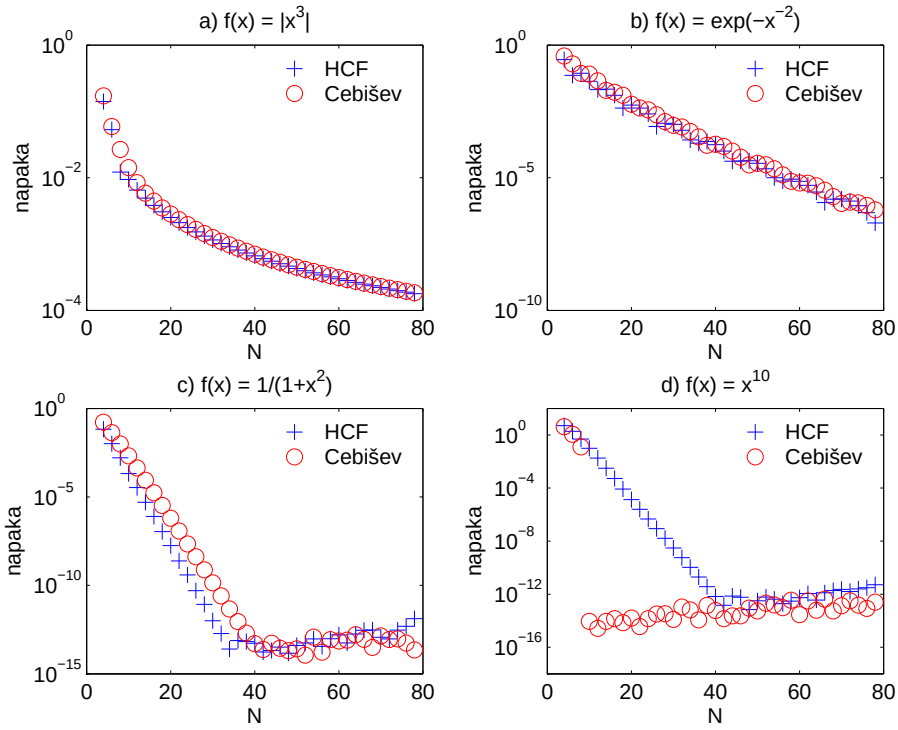
Operatorska matrika odvajanja D^2 je za primer $N = 3$ enaka

$$D^2 = \begin{bmatrix} 0 & -5.1040 & 22.4577 & -55.8569 & 0 & 0 & 0 \\ 0 & -2.4674 & -32.5698 & 98.9745 & 0 & 0 & 0 \\ 0 & 0 & -9.8696 & -79.0531 & 0 & 0 & 0 \\ 0 & 0 & 0 & -22.2066 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2.4674 & -11.8849 & 42.6816 \\ 0 & 0 & 0 & 0 & 0 & -9.8696 & -44.3054 \\ 0 & 0 & 0 & 0 & 0 & 0 & -22.2066 \end{bmatrix}.$$

V naslednjih dveh primerih si oglejmo, kako se povečuje kvaliteta aproksimacije odvoda dane funkcije z uporabo operatorskih matrik odvajanja za Fourierovo vrsto (3.48), vrsto Čebiševa (3.59) in poldomensko Čebišev-Fourierovo vrsto (7.16). V prvem primeru primerjamo neperiodične funkcije, v drugem pa periodične funkcije na intervalu $[-1, 1]$. Oba primera sta vzeta iz knjige L. N. Trefethen [57].

Na sliki 7.1 primerjamo maksimalno absolutno vrednost napake za spektralno odvajanje z matriko odvodov za vrsto Čebiševa in za poldomensko Čebišev-Fourierovo vrsto v odvisnosti od števila členov N za štiri neperiodične funkcije: $f(x) = |x^3|$, $f(x) = \mathrm{e}^{-x^{-2}}$, $f(x) = 1/(1 + x^2)$ in $f(x) = x^{10}$. Pri izbranih funkcijah narašča stopnja gladkosti. Tako ima

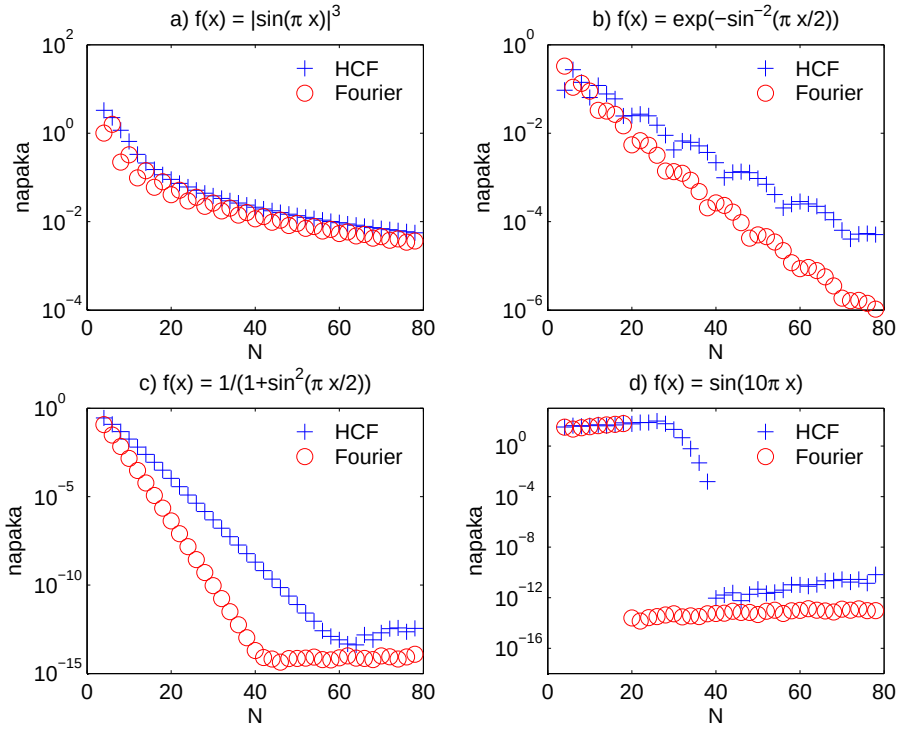
tretji odvod funkcije $f(x) = |x^3|$ omejeno variacijo, funkcija $f(x) = \mathrm{e}^{-x^{-2}}$ je gladka, toda ni analitična, funkcija $f(x) = 1/(1+x^2)$ je analitična na okolici intervala $[-1, 1]$, funkcija $f(x) = x^{10}$ pa polinom, ki je v teoriji spektralnih metod Čebiševa analogen funkciji z omejenim naborom frekvenc (ang. band-limited function) za aproksimacijo s Fourierovo vrsto. Opazimo, da bolj kot je dana funkcija gladka, hitrejša je konvergenca. Poleg tega opazimo tudi, da je natančnost spektralnega odvajanja za obe metodi primerljiva, razen v primeru zadnje funkcije, saj je za polinome primernejša aproksimacija z ortogonalnimi polinomi kot s trigonometričnimi funkcijami.



Slika 7.1: Primerjava spektralnega odvajanja funkcij z uporabo operatorskih matrik odvajanja za vrsto Čebiševa (rdeči krogci) in poldomensko Čebišev-Fourierovo vrsto (modri plusi) za štiri neperiodične funkcije naraščajoče gladkosti: a) $f(x) = |x^3|$, b) $f(x) = \mathrm{e}^{-x^{-2}}$, c) $f(x) = 1/(1+x^2)$ in d) $f(x) = x^{10}$. Primerjamo maksimalno absolutno vrednost napake glede na število členov N.

Na sliki 7.2 primerjamo maksimalno absolutno vrednost napake za spektralno odvajanje z matriko odvodov za Fourierovo in za poldomensko Čebišev-Fourierovo vrsto v odvisnosti od števila členov N za štiri periodične funkcije: $f(x) = |\sin x|^3$, $f(x) = \mathrm{e}^{-\sin^{-2}(x/2)}$, $f(x) = 1/(1+\sin^2(x/2))$ in $f(x) = \sin(10x)$. Pri izbranih funkcijah narašča stopnja gladkosti. Tako

ima tretji odvod funkcije $f(x) = |\sin x|^3$ omejeno variacijo, funkcija $f(x) = \mathrm{e}^{-\sin^{-2}(x/2)}$ je gladka, toda ni analitična, funkcija $f(x) = 1/(1+\sin^2(x/2))$ je analitična na okolici intervala $[-1, 1]$ v kompleksni ravnini, funkcija $f(x) = \sin(10x)$ pa je funkcija z omejenim naborom frekvenc (ang. band-limited function). Opazimo, da s stopnjo gladkosti narašča hitrost konvergence ter da je natančnost spektralnega odvajanja za obe metodi dokaj primerljiva, čeprav je Fourierova metoda večinoma natančnejša.



Slika 7.2: Primerjava spektralnega odvajanja funkcij z uporabo operatorskih matrik odvajanja za Fourierovo (rdeči krogci) in poldomensko Čebišev-Fourierovo vrsto (modri plusi) za štiri periodične funkcije naraščajoče gladkosti: a) $f(x) = |\sin x|^3$, b) $f(x) = \mathrm{e}^{-\sin^2(x/2)}$, c) $f(x) = 1/(1+\sin^2(x/2))$ in d) $f(x) = \sin(10x)$. Primerjamo maksimalno absolutno vrednost napake glede na število členov N.

Podrazdelek zaključimo z opisom numeričnega postopka za izračun koeficientov p_j^k definiranih z enačbo (7.6). Koeficienti so izračunani delno rekurzivno, kjer bazo rekurzije izračunamo z uporabo potenčnih momentov (5.18 – 5.19) in (5.24 – 5.25). Najprej za fiksno vrednost k in izbiro $j = 0, \ldots, k-1$ in $\ell = 0, \ldots, k-j-1$ definiramo koeficiente

$$ {}^{\ell}p_j^k := \tfrac{4}{\pi}\mathcal{L}_U\left(t^{\ell}\dot{T}_k^h(t)U_j^h(t)\right), \tag{7.21}$$

kjer je $p_j^k = {}^0p_j^k$. Predpostavimo, da smo na $k-$tem nivoju, kjer izračunamo in shranimo koeficiente

$$ {}^\ell p_0^k = \tfrac{4}{\pi}\mathcal{L}_U\left(t^\ell \dot{T}_k^h(t)\right).$$

Pri tem uporabimo potenčne momente (5.24) in (5.25) ter primerno orodje za simbolno računanje, npr. programski paket `Mathematica`. Vse preostale koeficiente izračunamo rekurzivno v standardni IEEE aritmetiki z uporabo tričlenske rekurzivne formule (5.1) za poldomenske polinome Čebiševa U_j^h drugega reda

$$ U_j^h(t) = \tfrac{1}{n_{j-1}}\left((t-\alpha_{j-1})U_{j-1}^h(t) - \beta_{j-1}U_{j-2}^h(t)\right), \qquad (7.22)$$

kjer je

$$ n_{j-1} = \left\|(t-\alpha_{j-1})U_{j-1}^h(t) - \beta_{j-1}U_{j-2}^h(t)\right\|.$$

Nato z uporabo enačbe (7.22) dobimo rekurzivne formule

$$\begin{aligned} {}^\ell p_1^k &= \tfrac{1}{n_0}{}^{\ell+1}p_0^k - \tfrac{\alpha_0}{n_0}{}^\ell p_0^k, \quad \ell = 0,\ldots,k-2,\\ {}^\ell p_j^k &= \tfrac{1}{n_{j-1}}{}^{\ell+1}p_{j-1}^k - \tfrac{\alpha_{j-1}}{n_{j-1}}{}^\ell p_{j-1}^k - \tfrac{\beta_{j-1}}{n_{j-1}}{}^\ell p_{j-2}^k, \quad \ell = 0,\ldots,k-j-1.\end{aligned}$$

Podoben postopek uporabimo za izračun koeficientov q_j^k, r_j^k in s_j^k, ki so definirani z enačbami (7.7 – 7.9). Izračun teh koeficientov je tehnične narave, zato v tem delu ni podrobneje opisan. Podrobnosti so opisane v programskih kodah, s katerimi so narejeni numerični zgledi. Opazimo, da je dovolj, da te koeficiente izračunamo samo enkrat in shranimo za vselej.

7.1.2 Multiplikacijska matrika

Začnimo z odrezano poldomensko Čebišev-Fourierovo vrsto (7.4) s spektralnimi koeficienti a_k in b_k za poljubno funkcijo $f \in L^2(-1,1)$. Sledimo oznakam iz podrazdelka 7.1.1, kjer smo z $\mathbf{u} = (a_0,\ldots,a_N,b_0,\ldots,b_{N-1})^T$ označili vektor spektralnih koeficientov HCF vrste. Iščemo spektralne koeficiente $\tilde{a}_j$ in $\tilde{b}_j$ odrezane poldomenske Čebišev-Fourierove vrste za produkt odrezanih HCF vrst f_N za funkcijo f in g_N za neko znano funkcijo $g \in L^2(-1,1)$, ki ima znane spektralne koeficiente c_i in d_i

$$ g(x) \approx g_N(x) = \sum_{i=0}^{N} c_i T_i^h(\cos\tfrac{\pi x}{2}) + \sum_{i=0}^{N-1} d_i U_i^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2}. \qquad (7.23)$$

Najprej zmnožimo odrezani HCF vrsti (7.4) in (7.23) ter produkt odrežemo pri istem odreznem številu N kot obe osnovni vrsti

$$\begin{aligned} f_N(x)g_N(x) &= \left(\sum_{k=0}^{N} a_k T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k U_k^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2}\right)\\ &\quad\cdot\left(\sum_{i=0}^{N} c_i T_i^h(\cos\tfrac{\pi x}{2}) + \sum_{i=0}^{N-1} d_i U_i^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2}\right). \qquad (7.24)\end{aligned}$$

Pri tem uporabimo enakost $\sin^2 \frac{\pi x}{2} = 1 - \cos^2 \frac{\pi x}{2}$ ter označimo $t = \cos \frac{\pi x}{2}$.

Polinome $T_k^h(t)T_i^h(t)$ in $U_k^h(t)U_i^h(t)(1-t^2)$ aproksimiramo z odrezano HCF vrsto, sestavljeno iz polinomov T_j^h, polinome $T_k^h(t)U_i^h(t)$ in $T_i^h(t)U_k^h(t)$ pa z odrezano HCF vrsto, sestavljeno iz polinomov U_j^h

$$T_k^h(t)T_i^h(t) = \sum_{j=0}^{\min\{k+i,N\}} P_j^{ki} T_j^h(t), \quad P_j^{ki} = \tfrac{4}{\pi}\mathcal{L}_T\left(T_k^h(t)T_i^h(t)T_j^h(t)\right), \tag{7.25}$$

$$T_k^h(t)U_i^h(t) = \sum_{j=0}^{\min\{k+i,N-1\}} Q_j^{ki} U_j^h(t), \quad Q_j^{ki} = \tfrac{4}{\pi}\mathcal{L}_U\left(T_k^h(t)U_i^h(t)U_j^h(t)\right), \tag{7.26}$$

$$U_k^h(t)U_i^h(t) = \sum_{j=0}^{\min\{k+i,N\}} R_j^{ki} T_j^h(t), \quad R_j^{ki} = \tfrac{4}{\pi}\mathcal{L}_T\left(U_k^h(t)U_i^h(t)T_j^h(t)\right), \tag{7.27}$$

$$U_k^h(t)U_i^h(t)t^2 = \sum_{j=0}^{\min\{k+i+2,N\}} S_j^{ki} T_j^h(t), \quad S_j^{ki} = \tfrac{4}{\pi}\mathcal{L}_T\left(t^2U_k^h(t)U_i^h(t)T_j^h(t)\right). \tag{7.28}$$

Koeficiente P_j^{ki}, Q_j^{ki}, R_j^{ki} in S_j^{ki}, ki so definirani z enačbami (7.25 – 7.28), določimo iz pogojev ortogonalnosti. Za numeričen izračun teh koeficientov uporabimo podoben postopek kot za izračun koeficientov p_j^k, q_j^k, r_j^k in s_j^k, ki nastopajo v matriki odvodov in so definirani z enačbami (7.6 – 7.9). Predlagan postopek predstavlja kombinacijo direktnega in rekuzivnega pristopa, kjer bazo rekurzije določimo direktno z uporabo potenčnih momentov za poldomenske polinome Čebiševa prve in druge vrste, ki so podani z enačbami (5.18 – 5.19) in (5.24 – 5.25), ter primernega orodja za simbolno računanje, npr. programskega paketa `Mathematica`. Postopek je ponovno zelo tehnične narave, rekurzivne formule pa precej bolj kompliciran kot v primeru koeficientov za matriko odvodov, zato v tem delu niso podrobneje opisane. Podrobnosti najdemo v implementciji programskih kod. Tudi za te koeficiente je dovolj, da jih izračunamo enkrat za vselej.

Z uporabo oznak iz enačb (7.25 – 7.28) je produkt (7.24) enak

$$\begin{aligned} f_N(x)g_N(x) = &\sum_{k=0}^{N}\sum_{i=0}^{N} a_k c_i \sum_{j=0}^{\min\{k+i,N\}} P_j^{ki} T_j^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1}\sum_{i=0}^{N-1} b_k d_i \sum_{j=0}^{\min\{k+i,N\}} R_j^{ki} T_j^h(\cos\tfrac{\pi x}{2}) \\ &- \sum_{k=0}^{N-1}\sum_{i=0}^{N-1} b_k d_i \sum_{j=0}^{\min\{k+i+2,N\}} S_j^{ki} T_j^h(\cos\tfrac{\pi x}{2}) \\ &+ \sum_{k=0}^{N}\sum_{i=0}^{N-1} a_k d_i \sum_{j=0}^{\min\{k+i,N-1\}} Q_j^{ki} U_j^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2} \\ &+ \sum_{k=0}^{N-1}\sum_{i=0}^{N} b_k c_i \sum_{j=0}^{\min\{k+i,N-1\}} Q_j^{ik} U_j^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2} \end{aligned} \tag{7.29}$$

$$
\begin{aligned}
f_N(x)g_N(x) &= \sum_{j=0}^{N}\left(\sum_{k=0}^{N} a_k \left(\sum_{i=\max\{j-k,0\}}^{N} c_i P_j^{ki}\right)\right. \\
&\quad \left. + \sum_{k=0}^{N-1} b_k \left(\sum_{i=\max\{j-k,0\}}^{N-1} d_i R_j^{ki} - \sum_{i=\max\{j-k-2,0\}}^{N-1} d_i S_j^{ki}\right)\right) T_j^h(\cos\tfrac{\pi x}{2}) \\
&\quad + \sum_{j=0}^{N-1}\left(\sum_{k=0}^{N} a_k \left(\sum_{i=\max\{j-k,0\}}^{N-1} d_i Q_j^{ki}\right)\right. \\
&\quad \left. + \sum_{k=0}^{N-1} b_k \left(\sum_{i=\max\{j-k,0\}}^{N} c_i Q_j^{ik}\right)\right) U_j^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2} \qquad (7.30) \\
&= \sum_{j=0}^{N} \tilde{a}_j T_j^h(\cos\tfrac{\pi x}{2}) + \sum_{j=0}^{N-1} \tilde{b}_j U_j^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2}. \qquad (7.31)
\end{aligned}
$$

Funkcionala $\mathcal{L}_T$ in $\mathcal{L}_U$ sta definirana z enačbama (5.17) in (5.23). V vrstici (7.30) zamenjamo vrstni red seštevanja. Nazadnje v vrstici (7.31) koeficiente odrezane HCF vrste za produkt $f_N\, g_N$ označimo z $\tilde{a}_j$ in $\tilde{b}_j$

$$
\begin{aligned}
\tilde{a}_j &= \sum_{k=0}^{N} a_k \left(\sum_{i=\max\{j-k,0\}}^{N} c_i P_j^{ki}\right) + \sum_{k=0}^{N-1} b_k \left(\sum_{i=\max\{j-k,0\}}^{N-1} d_i R_j^{ki} - \sum_{i=\max\{j-k-2,0\}}^{N-1} d_i S_j^{ki}\right), \\
&\qquad j = 0, 1, \ldots, N, \qquad (7.32) \\
\tilde{b}_j &= \sum_{k=0}^{N} a_k \left(\sum_{i=\max\{j-k,0\}}^{N-1} d_i Q_j^{ki}\right) + \sum_{k=0}^{N-1} b_k \left(\sum_{i=\max\{j-k,0\}}^{N} c_i Q_j^{ik}\right), \\
&\qquad j = 0, 1, \ldots, N-1. \qquad (7.33)
\end{aligned}
$$

Enačbe (7.32) in (7.33) so linearne, zato jih lahko predstavimo v matrični obliki

$$
\tilde{\mathbf{u}} = F\,\mathbf{u}, \qquad (7.34)
$$

kjer je $\mathbf{u}$ kot prej vektor spektralnih koeficientov odrezane HFC vrste za funkcijo f in $\tilde{\mathbf{u}} = \left(\tilde{a}_0, \ldots, \tilde{a}_N, \tilde{b}_0, \ldots, \tilde{b}_{N-1}\right)^T$ vektor spektralnih koeficientov $\tilde{a}_j$ in $\tilde{b}_j$ produkta odrezanih HFC vrst za funkciji f in g. Multiplikacijska matrika $F \in \mathbb{R}^{(2N+1)\times(2N+1)}$ je bločna matrika

$$
F = \begin{bmatrix} G_1 & G_2 \\ G_3 & G_4 \end{bmatrix}, \qquad (7.35)
$$

kjer je $G_1 \in \mathbb{R}^{(N+1)\times(N+1)}$, $G_2 \in \mathbb{R}^{(N+1)\times N}$, $G_3 \in \mathbb{R}^{N\times(N+1)}$ in $G_4 \in \mathbb{R}^{N\times N}$. Elementi teh matrik so

$$G_1 = [g^1_{jk}]_{j,k}, \qquad g^1_{jk} = \sum_{i=\max\{j-k,0\}}^{N} c_i P^{k-1,i}_{j-1}, \tag{7.36}$$
$$j = 1,\dots,N+1,\ k = 1,\dots,N+1,$$

$$G_2 = [g^2_{jk}]_{j,k}, \qquad g^2_{jk} = \sum_{i=\max\{j-k,0\}}^{N-1} d_i R^{k-1,i}_{j-1} - \sum_{i=\max\{j-k-2,0\}}^{N-1} d_i S^{k-1,i}_{j-1}, \tag{7.37}$$
$$j = 1,\dots,N+1,\ k = 1,\dots,N,$$

$$G_3 = [g^3_{jk}]_{j,k}, \qquad g^3_{jk} = \sum_{i=\max\{j-k,0\}}^{N-1} d_i Q^{k-1,i}_{j-1}, \tag{7.38}$$
$$j = 1,\dots,N,\ k = 1,\dots,N+1,$$

$$G_4 = [g^4_{jk}]_{j,k}, \qquad g^4_{jk} = \sum_{i=\max\{j-k,0\}}^{N} c_i Q^{i,k-1}_{j-1}, \tag{7.39}$$
$$j = 1,\dots,N,\ k = 1,\dots,N.$$

Bloki G_1, G_2, G_3 in G_4 so za primer $N = 2$ enaki

$$G_1 = \begin{bmatrix} c_0P^{00}_0 + c_1P^{01}_0 + c_2P^{02}_0 & c_0P^{10}_0 + c_1P^{11}_0 + c_2P^{12}_0 & c_0P^{20}_0 + c_1P^{21}_0 + c_2P^{22}_0 \\ c_1P^{01}_1 + c_2P^{02}_1 & c_0P^{10}_1 + c_1P^{11}_1 + c_2P^{12}_1 & c_0P^{20}_1 + c_1P^{21}_1 + c_2P^{22}_1 \\ c_2P^{02}_2 & c_1P^{11}_2 + c_2P^{12}_2 & c_0P^{20}_2 + c_1P^{21}_2 + c_2P^{22}_2 \end{bmatrix},$$

$$G_2 = \begin{bmatrix} d_0(R^{00}_0 - S^{00}_0) + d_1(R^{01}_0 - S^{01}_0) & d_0(R^{10}_0 - S^{10}_0) + d_1(R^{11}_0 - S^{11}_0) \\ -d_0S^{00}_1 + d_1(R^{01}_1 - S^{01}_1) & d_0(R^{10}_1 - S^{10}_1) + d_1(R^{11}_1 - S^{11}_1) \\ -d_0S^{00}_2 - d_1S^{01}_2 & -d_0S^{10}_2 + d_1(R^{11}_2 - S^{11}_2) \end{bmatrix},$$

$$G_3 = \begin{bmatrix} d_0Q^{00}_0 + d_1Q^{01}_0 & d_0Q^{10}_0 + d_1Q^{11}_0 & d_0Q^{20}_0 + d_1Q^{21}_0 \\ d_1Q^{01}_1 & d_0Q^{10}_1 + d_1Q^{11}_1 & d_0Q^{20}_1 + d_1Q^{21}_1 \end{bmatrix},$$

$$G_4 = \begin{bmatrix} c_0Q^{00}_0 + c_1Q^{10}_0 + c_2Q^{20}_0 & c_0Q^{01}_0 + c_1Q^{11}_0 + c_2Q^{21}_0 \\ c_1Q^{10}_1 + c_2Q^{20}_1 & c_0Q^{01}_1 + c_1Q^{11}_1 + c_2Q^{21}_1 \end{bmatrix}.$$

V primeru, da je funkcija f soda, sta bloka G_2 in G_3 ničelna, v primeru, da je funkcija f liha, pa sta bloka G_1 in G_4 ničelna. Če je funkcija f konstantna, je multiplikacijska matrika F skalarna, sicer pa je polna. Operatorska multiplikacijska matrika F je za primer $N = 3$ in $f(x) = x$ (liha

funkcija) enaka

$$F = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.5733 & -0.0672 & 0.0127 \\ 0 & 0 & 0 & 0 & -0.4530 & 0.5356 & -0.0612 \\ 0 & 0 & 0 & 0 & -0.0315 & -0.4437 & 0.5258 \\ 0 & 0 & 0 & 0 & 0.0010 & -0.0316 & -0.4425 \\ 0.5733 & -0.4530 & -0.0315 & 0.0010 & 0 & 0 & 0 \\ -0.0672 & 0.5356 & -0.4437 & -0.0316 & 0 & 0 & 0 \\ 0.0127 & -0.0612 & 0.5258 & -0.4425 & 0 & 0 & 0 \end{bmatrix}.$$

7.2 Konstrukcija Čebišev-Fourierove kolokacijske metode

V tem razdelku bomo konstruirali nov razred Čebišev-Fourierovih spektralnih metod za reševanje linearnih dvotočkovnih robnih problemov v eni dimenziji (7.1 – 7.2). Numerično rešitev iščemo v obliki odrezane poldomenske Čebišev-Fourierove vrste (7.4), ki je sestavljena iz trigonometričnih funkcij, ki so reorganizirane v dve neklasični družini ortogonalnih polinomov, tj. poldomenske polinome Čebiševa prve (5.13 – 5.14) in druge vrste (5.15 – 5.16). Spektralne koeficiente bomo izračunali z metodo kolokacije. Konstrukcija metode je podrobno opisana v poročilu B. Orel in A. Perne [45].

Rešujemo torej problem 1.3, ki ga zapišemo kot linearni dvotočkovni robni problem (3.1)

$$\alpha(x)\frac{\mathrm{d}^2u}{\mathrm{d}x^2} + \beta(x)\frac{\mathrm{d}u}{\mathrm{d}x} + \gamma(x)u = f(x), \quad x \in [-1,1],$$

z Dirichletovimi (3.2) robnimi pogoji $u(-1) = A$ in $u(1) = B$, kjer numerično rešitev iščemo v obliki odrezane poldomenske Čebišev-Fourierove vrste (7.4)

$$u(x) \approx u_N(x) = \sum_{k=0}^{N} a_k T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k U_k^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2}. \qquad (7.40)$$

V odrezano HCF vrsto razvijemo tudi prvi in drugi odvod rešitve

$$\frac{\mathrm{d}u_N}{\mathrm{d}x}(x) = \sum_{k=0}^{N} a_k' T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k' U_k^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2}, \qquad (7.41)$$

$$\frac{\mathrm{d}^2u_N}{\mathrm{d}x^2}(x) = \sum_{k=0}^{N} a_k'' T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k'' U_k^h(\cos\tfrac{\pi x}{2})\sin\tfrac{\pi x}{2}. \qquad (7.42)$$

Z drugimi besedami, iščemo spektralne koeficiente a_k in b_k, tako da bo numerična rešitev u_N zapisana z odrezano vrsto (7.40) čim boljša, tj. želimo,

da bo norma razlike med točno in numerično rešitvijo $\|u - u_N\|$ čim manjša. Pri tem z N označimo odrezno število, kar pomeni, da imamo $2N + 1$ spektralnih koeficientov. Za dano vrednost N uporabimo $2N + 1$ ortogonalnih polinomov: $N + 1$ poldomenskih polinomov Čebiševa prve in N poldomenskih polinomov Čebiševa druge vrste.

Zveza med spektralnimi koeficienti a'_k in b'_k za aproksimacijo prvega odvoda, oz. a''_k in b''_k za aproksimacijo drugega odvoda, ter a_k in b_k je podana z linearnimi zvezami (7.13) in (7.14), oz. v matrični obliki z enačbama (7.15) in (7.20), kjer je

$$\mathbf{u} = (a_0, \ldots, a_N, b_0, \ldots, b_{N-1})^T$$

vektor iskanih spektralnih koeficientov in D operatorska matrika odvajanja, ki je podana z enačbo (7.16). Konstrukcija matrike odvodov je podrobno opisana v podrazdelku 7.1.1.

Začnemo z delitvijo danega intervala $[-1, 1]$ z $2N + 1$ kolokacijskimi vozli, za katere uporabimo točke Čebiševa druge vrste (3.11)

$$x_i = -\cos\left(\frac{\pi i}{2N}\right), \quad i = 0, 1, \ldots, 2N, \tag{7.43}$$

na $2N$ podintervalov

$$-1 = x_0 < x_1 < \cdots < x_{2N} = 1.$$

Pri reševanju robnih problemov z metodo kolokacije privzamemo, da numerična rešitev robnega problema točno zadošča diferencialni enačbi v notranjih kolokacijskih točkah x_i, $i = 1, 2, \ldots, 2N - 1$. Ko vstavimo izraze za odrezane HCF vrste za $\frac{\mathrm{d}^2u}{\mathrm{d}x^2}$ (7.40), $\frac{\mathrm{d}u}{\mathrm{d}x}$ (7.41) in u (7.42) v diferencialno enačbo (3.1), dobimo sistem linearnih enačb, ki mu pravimo kolokacijski sistem

$$\alpha(x_i)\frac{\mathrm{d}^2u_N}{\mathrm{d}x^2}(x_i) + \beta(x_i)\frac{\mathrm{d}u_N}{\mathrm{d}x}(x_i) + \gamma(x_i)u_N(x_i) = f(x_i), \tag{7.44}$$

kjer je $i = 1, 2, \ldots, 2N - 1$. Dodatno imamo še dve enačbi, ki izvirata iz Dirichletovih robnih pogojev (3.2)

$$\sum_{k=0}^{N} a_k T_k^h(0) - \sum_{k=0}^{N-1} b_k U_k^h(0) = A, \tag{7.45}$$

$$\sum_{k=0}^{N} a_k T_k^h(0) + \sum_{k=0}^{N-1} b_k U_k^h(0) = B. \tag{7.46}$$

Nadalje označimo s $C \in \mathbb{R}^{(2N+1)\times(2N+1)}$ kolokacijsko matriko. Elementi te matrike so vrednosti baznih funkcij, tj. poldomenskih polinomov Čebiševa prve in druge vrste, $T_k^h(\cos\frac{\pi\cdot}{2})$ in $U_k^h(\cos\frac{\pi\cdot}{2})\sin\frac{\pi\cdot}{2}$, izračunane

v kolokacijskih točkah. Če množico baznih funkcij označimo s $\{\phi_j\}_{j=0}^{2N}$, so elementi matrike $C = [c_{ij}]_{i,j=1}^{2N+1}$ enaki

$$c_{ij} = \phi_j(x_i). \tag{7.47}$$

Poleg vektorja iskanih spektralnih koeficientov $\mathbf{u}$ označimo z

$$\mathbf{v} = (A, f(x_1), \ldots, f(x_{2N-1}), B)^T$$

vektor funkcijskih vrednosti funkcije desne strani f enačbe (3.1) v notranjih kolokacijskih točkah $x_1, x_2, \ldots, x_{2N-1}$. Prvi in zadnji element vektorja $\mathbf{v}$ sta Dirichletova robna pogoja (3.2) A in B v krajiščih intervala $[-1, 1]$.

Koeficientne funkcije α, β in γ diferencialne enačbe (3.1) v splošnem niso konstantne, pač pa so funkcije neodvisne spremenljivke x. Zato je potrebno te koeficiente aproksimirati z odrezanimi poldomenskimi Čebišev-Fourierovimi vrstami

$$\alpha(x) \approx \alpha_N(x) = \sum_{k=0}^{N} a_k^{\alpha} T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k^{\alpha} U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \tag{7.48}$$

$$\beta(x) \approx \beta_N(x) = \sum_{k=0}^{N} a_k^{\beta} T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k^{\beta} U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \tag{7.49}$$

$$\gamma(x) \approx \gamma_N(x) = \sum_{k=0}^{N} a_k^{\gamma} T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k^{\gamma} U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}. \tag{7.50}$$

Množenje odrezanih HCF vrst α_N, β_N in γ_N z odrezanimi HCF vrstami za numerično rešitev u_N (7.40) ter prvi $\frac{\mathrm{d}u_N}{\mathrm{d}x}$ (7.41) in drugi odvod $\frac{\mathrm{d}^2u_N}{\mathrm{d}x^2}$ (7.42) izvedemo z operatorsko multiplikacijsko matriko, katere konstrukcija je podrobno opisana v podrazdelku 7.1.2.

Z $\mathbf{u}$ kot prej označimo vektor iskanih spektralnih koeficientov a_k in b_k, nato pa z

$$\begin{aligned}
\mathbf{u}_{\alpha} &= \left(a_0^{\alpha}, \ldots, a_N^{\alpha}, b_0^{\alpha}, \ldots, b_{N-1}^{\alpha}\right)^T, \\
\mathbf{u}_{\beta} &= \left(a_0^{\beta}, \ldots, a_N^{\beta}, b_0^{\beta}, \ldots, b_{N-1}^{\beta}\right)^T, \\
\mathbf{u}_{\gamma} &= \left(a_0^{\gamma}, \ldots, a_N^{\gamma}, b_0^{\gamma}, \ldots, b_{N-1}^{\gamma}\right)^T
\end{aligned}$$

označimo še vektorje spektralnih koeficientov za odrezane HCF vrste α_N, β_N in γ_N (7.48 – 7.50). Nadalje označimo z

$$\tilde{\mathbf{u}}_{\alpha} = \left(\tilde{a}_0^{\alpha}, \ldots, \tilde{a}_N^{\alpha}, \tilde{b}_0^{\alpha}, \ldots, \tilde{b}_{N-1}^{\alpha}\right)^T$$

vektor spektralnih koeficientov odrezane poldomenske Čebišev-Fourierove vrste za produkt $\alpha_N \frac{\mathrm{d}^2u_N}{\mathrm{d}x^2}$, z

$$\tilde{\mathbf{u}}_{\beta} = \left(\tilde{a}_0^{\beta}, \ldots, \tilde{a}_N^{\beta}, \tilde{b}_0^{\beta}, \ldots, \tilde{b}_{N-1}^{\beta}\right)^T$$

vektor spektralnih koeficientov za produkt $\beta_N \frac{\mathrm{d}u_N}{\mathrm{d}x}$ ter z

$$\tilde{\mathbf{u}}_\gamma = \left(\tilde{a}_0^\gamma, \ldots, \tilde{a}_N^\gamma, \tilde{b}_0^\gamma, \ldots, \tilde{b}_{N-1}^\gamma\right)^T$$

vektor spektralnih koeficientov za produkt $\gamma_N\, u_N$. Zveze med koeficienti so podane z enačbami

$$\tilde{\mathbf{u}}_\alpha = F_\alpha\, \mathbf{u}'', \quad \tilde{\mathbf{u}}_\beta = F_\beta\, \mathbf{u}', \quad \tilde{\mathbf{u}}_\gamma = F_\gamma\, \mathbf{u}, \tag{7.51}$$

kjer so F_α, F_β in F_γ multiplikacijske matrike, ki so definirane z linearno zvezo (7.35). Nazadnje označimo z $L \in \mathbb{R}^{(2N+1)\times(2N+1)}$ diferencialno operatorsko matriko

$$L = F_\alpha D^2 + F_\beta D + F_\gamma, \tag{7.52}$$

ki pripada diferencialni enačbi (3.1), kjer je D matrika odvodov (7.16), F_α, F_β in F_γ pa so multiplikacijske matrike (7.35).

Linearni sistem enačb (7.44 – 7.46) lahko tedaj zapišemo v matrični obliki

$$U\, \mathbf{u} = \mathbf{v}, \tag{7.53}$$

kjer je matrika $U \in \mathbb{R}^{(2N+1)\times(2N+1)}$ konstruirana takole. Najprej zmnožimo kolokacijsko matriko C in operatorsko matriko L, da dobimo produkt $C\, L$, ki v matrični obliki predstavlja matriko koeficientov sistema linearnih enačb (7.44), nato pa zamenjamo prvo in zadnjo vrstico tako dobljene matrike s prvo in zadnjo vrstico kolokacijske matrike C, da zadostimo Dirichletovima robnima pogojema (7.45) in (7.46).

Rešitev linearnega sistema (7.53) je vektor iskanih spektralnih koeficientov odrezane poldomenske Čebišev-Fourierove vrste za numerično rešitev linearnega dvotočkovnega robnega problema (3.1) in (3.2). Množenje s kolokacijsko matriko C vrne vektor vrednosti numerične rešitve v kolokacijskih vozlih, tj. točkah Čebiševa.

7.3 Analiza napake

V tem razdelku bomo obravnavali analizo napake in red konvergence za nov razred Čebišev-Fourierovih kolokacijskih spektralnih metod, ki smo jih konstruirali v razdelku 7.2. Za dokaz konvergence metode in oceno napake za numerično rešitev se opremo na tehnike za oceno napake, ki jih najdemo v C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang [11] ter v J. Shen, T. Tang in L. Wang [51] in smo jih na kratko predstavili v razdelku 3.6. Ker je poldomenska Čebišev-Fourierova vrsta posplošena trigonometrična vrsta, ki je reorganizirana po (ortogonalnih) poldomenskih polinomih Čebiševa prve in druge vrste, sledimo korakom za oceno napake za Fourierove kolokacijske spektralne metode. Analiza napake in reda konvergence je podrobno opisana v poročilu B. Orel in A. Perne [45].

Omejimo se na reševanje linearnih dvotočkovnih robnih problemov s homogenimi Dirichletovimi robnimi pogoji na intervalu $[-1, 1]$

$$\begin{aligned} \mathcal{L}u &= f, & (7.54)\\ u(-1) = u(1) &= 0, & (7.55) \end{aligned}$$

kjer je, kot običajno, $\mathcal{L}$ linearni diferencialni operator (1.6).

Primeren Hilbertov prostor za analizo napake je $L^2(-1, 1)$, tj. prostor vseh s kvadratom integrabilnih funkcij na intervalu $[-1, 1]$. V tem prostoru je operator $\mathcal{L}$ neomejen. Označimo z

$$(u, v) = \int_{-1}^{1} u(x)v(x)\,\mathrm{d}x \tag{7.56}$$

primeren skalarni produkt v prostoru $L^2(-1, 1)$ in z $\|u\|_{L^2} = (u, u)^{1/2}$ pripadajočo normo. Poleg tega naj bo $T_N \subset L^2(-1, 1)$ podprostor vseh trigonometričnih polinomov stopnje kvečjemu N, ki zadoščajo homogenim robnim pogojem (7.55). Nadalje označimo z $(u, v)_N$ primeren diskretni skalarni produkt s pripadajočo diskretno normo $\|u\|_N = (u, u)_N^{1/2}$. Numerična rešitev $u_N \in T_N$ robnega problema (7.54 – 7.55), kjer spektralne koeficiente izračunamo z metodo kolokacije, zadošča enačbam

$$\begin{aligned} \mathcal{L}_N u_N(x_k) &= f(x_k), & (7.57)\\ u_N(x_0) = u_N(x_{2N}) &= 0, & (7.58) \end{aligned}$$

kjer so vozli x_k, $k = 1, \ldots, 2N-1$, notranje kolokacijske točke (7.43), operator $\mathcal{L}_N$ pa je aproksimacija operatorja $\mathcal{L}$, ki jo dobimo tako, da nadomestimo točne odvode z interpoliranimi odvodi glede na enačbi (7.15) in (7.20) ter da razvijemo koeficiente diferencialne enačbe α, β in γ v pripadajoče odrezane poldomenske Čebišev-Fourierove vrste.

Enačbe (7.54) in (7.55) lahko ekvivalentno zapišemo v šibki obliki kot bilinearno formo

$$(\mathcal{L}u, v) = (f, v), \qquad \forall v \in L^2(-1, 1), \tag{7.59}$$

kjer u zadošča homogenim robnim pogojem (7.55). Kolokacijsko metodo (7.57 – 7.58) lahko tedaj zapišemo kot

$$(\mathcal{L}_N u_N, v)_N = (f, v)_N, \qquad v \in T_N, \tag{7.60}$$

kjer je $u_N \in T_N$. Analiza konvergenčnih lastnosti zahteva obstoj gostega Hilbertovega podprostora v $L^2(-1, 1)$. Primerna izbira je prostor Soboljeva (2.3) za $m = 1$

$$H^1(-1, 1) = \left\{ v \in L^2(-1, 1); \frac{\mathrm{d}v}{\mathrm{d}x} \in L^2(-1, 1) \right\}, \tag{7.61}$$

kjer šibki odvod $\frac{\mathrm{d}v}{\mathrm{d}x}$ pripada prostoru $L^2(-1,1)$. Ta prostor je opremljen z normo Soboljeva (2.5) za $m = 1$

$$\|v\|_{H^1} = \left(\|v\|_{L^2}^2 + \|\tfrac{\mathrm{d}v}{\mathrm{d}x}\|_{L^2}^2\right)^{1/2}, \tag{7.62}$$

za katero velja $\|u\|_{L^2} \leq \|u\|_{H^1}$ za vse $u \in H^1(-1,1)$. Opazimo, da je za vsak $N > 0$, prostor T_N vsebovan v prostoru $H^1(-1,1)$. Poleg tega za analizo konvergence potrebujemo, da operator $\mathcal{L}$, oz. natančneje bilinearna forma $(\mathcal{L}u, v)$, zadošča pogoju pozitivne definitnosti (3.80)

$$\exists \alpha^* > 0 : (\mathcal{L}u, u) \geq \alpha^* \|u\|_{H^1}^2, \quad u \in T_N, \tag{7.63}$$

in pogoju zveznosti (3.79)

$$\exists A > 0 : |(\mathcal{L}u, v)| \leq A \|u\|_{H^1} \|v\|_{H^1}, \quad u, v \in T_N. \tag{7.64}$$

Lema 7.1 *Naj bo $\mathcal{L}$ diferencialni operator (1.6), ki pripada enačbi (7.54), kjer u zadošča homogenim robnim pogojem (7.55). Nadalje privzemimo, da so koeficientne funkcije diferencialne enačbe α, β in γ omejene in strogo pozitivne funkcije na intervalu $[-1,1]$.*

Tedaj bilinearna forma $(\mathcal{L}u, v)$, ki je določena z enačbo (7.59), zadošča pogoju pozitivne definitnosti z

$$\alpha^* = \min_{x \in [-1,1]} \{\alpha(x), \beta(x), \gamma(x)\}. \tag{7.65}$$

in pogoju zveznosti z

$$A = 3 \max_{x \in [-1,1]} \{\alpha(x), \beta(x), \gamma(x)\}. \tag{7.66}$$

Dokaz: Ker je za vsak $u \in H^1(-1,1)$

$$\begin{aligned}
(\mathcal{L}u, u) &= \int_{-1}^{1} (\alpha(x)u'' + \beta(x)u' + \gamma(x)u)\, u \,\mathrm{d}x \\
&= \int_{-1}^{1} \alpha(x)u''u \,\mathrm{d}x + \int_{-1}^{1} \beta(x)u''u \,\mathrm{d}x + \int_{-1}^{1} \gamma(x)u''u \,\mathrm{d}x \\
&\geq \min_{x \in [-1,1]} \{\alpha(x), \beta(x), \gamma(x)\} \left(\int_{-1}^{1} (u')^2 \,\mathrm{d}x + \underbrace{\int_{-1}^{1} u'u \,\mathrm{d}x}_{=0} + \int_{-1}^{1} u^2 \,\mathrm{d}x \right) \\
&= \min_{x \in [-1,1]} \{\alpha(x), \beta(x), \gamma(x)\} \int_{-1}^{1} \left((u')^2 + u^2\right) \mathrm{d}x \\
&= \alpha^* \, \|u\|_{H^1}^2,
\end{aligned}$$

kjer v drugi vrstici uporabimo integracijo po delih (per partes), je pogoj pozitivne definitnosti (7.63) za bilinearno formo (7.59) izpolnjen z

$$\alpha^* = \min_{x\in[-1,1]} \{\alpha(x), \beta(x), \gamma(x)\}.$$

Podobno, ker je

$$\begin{aligned}
|(\mathcal{L}u, v)| &= \left| \int_{-1}^{1} (\alpha(x)u'' + \beta(x)u' + \gamma(x)u)\, v \,\mathrm{d}x \right| \\
&\leq \underbrace{\max_{x\in[-1,1]} \{\alpha(x), \beta(x), \gamma(x)\}}_{\tilde{A}} \left(\left| \int_{-1}^{1} u''v \,\mathrm{d}x \right| + \left| \int_{-1}^{1} u'v \,\mathrm{d}x \right| + \left| \int_{-1}^{1} uv \,\mathrm{d}x \right| \right) \\
&= \tilde{A} \left(\left| \int_{-1}^{1} u'v' \,\mathrm{d}x \right| + \left| \int_{-1}^{1} uv' \,\mathrm{d}x \right| + \left| \int_{-1}^{1} uv \,\mathrm{d}x \right| \right) \\
&\leq \tilde{A} \left(\|u'\|_{L^2} \|v'\|_{L^2} + \|u\|_{L^2} \|v'\|_{L^2} + \|u\|_{L^2} \|v\|_{L^2} \right) \\
&\leq 3\tilde{A} \, \|u\|_{H^1} \|v\|_{H^1},
\end{aligned}$$

z uporabo integracije po delih v drugi vrstici, Cauchy-Schwartzove neenakosti (2.7) v tretji vrstici ter neenakosti $\|u\|_{L^2} \leq \|u\|_{H^1}$ in $\|u'\|_{L^2} \leq \|u\|_{H^1}$ v četrti vrstici, je pogoj zveznosti (7.64) izpolnjen z

$$A = 3 \max_{x\in[-1,1]} \{\alpha(x), \beta(x), \gamma(x)\}.$$

□

Nadalje označimo $e = u_N - R_N u$, kjer je $u \in L^2(-1,1)$ točna in $u_N \in T_N$ numerična rešitev robnega problema (7.54 – 7.55). Pri tem je R_N operator projekcije iz prostora $L^2(-1,1)$ na prostor T_N. Pod pogoji Strangove leme (izrek 3.9), ki smo jim zadostili z dokazom leme 7.1, ima robni problem (7.60) dopustno in enolično numerično rešitev $u_N \in T_N$, ki zadošča neenačbi (3.90)

$$\|u_N\|_{L^2} \leq \frac{1}{\alpha^*} \sup_{0\neq v\in T_N} \frac{|(f,v)_N|}{\|v\|_{L^2}}. \tag{7.67}$$

Poleg tega, ker sta pogoja pozitivne definitnosti in zveznosti izpolnjena, ocena napake za numerično rešitev u_N po Strangovi lemi zadošča neenačbi (3.91)

$$\begin{aligned}
\|u - u_N\|_{H^1} &\leq \|u - R_N u\|_{H^1} + \|e\|_{H^1} \\
&\leq \left(1 + \frac{A}{\alpha^*}\right) \|u - R_N u\|_{H^1} + \frac{1}{\alpha^*} \frac{|(Q_N f, e)_N - (f,e)|}{\|e\|_{H^1}} \\
&\quad + \frac{1}{\alpha^*} \frac{|(\mathcal{L}R_N u, e) - (Q_N \mathcal{L}_N R_N u, e)_N|}{\|e\|_{H^1}}.
\end{aligned} \tag{7.68}$$

Pri tem je Q_N operator projekcije iz prostora $L^2(-1,1)$ na prostor T_N glede na diskretni skalarni produkt in $Q_N v$ je tedaj trigonometrični polinom stopnje N, ki se ujema z v v notranjih kolokacijskih točkah (7.43) in je enak 0 v robnih točkah. Metoda je konvergentna, če vsi trije členi v neenačbi (7.68) konvergirajo proti 0, ko gre N proti ∞. Obe konstanti α^* in A iz leme 7.1 sta neodvisni od odreznega števila N.

Spodnji izrek določa oceno napake in stopnjo konvergence za robne probleme oblike (7.54), kjer so koeficientne funkcije α, β in γ, funkcija desne strani f ter rešitev u zvezno odvedljive do nekega reda m.

Izrek 7.2 *Naj bo $u \in L^2(-1,1)$ točna rešitev robnega problema (7.54) z robnimi pogoji (7.55) ter naj bo $u_N \in T_N$ numerična rešitev dobljena z uporabo razreda Čebišev-Fourierovih kolokacijskih (CFC) metod, ki smo jih konstruirali v razdelku 7.2. Privzemimo, da so koeficientne funkcije α, β in γ, funkcija desne strani f ter rešitev u m-krat zvezno odvedljive. Tedaj je ocena napake za aproksimacijo rešitve za razred CFC metod enaka*

$$\|u - u_N\|_{H^1} \leq \left(1 + \frac{A}{\alpha^*}\right) C_1\, N^{1-m}\, \|u^{(m)}\|_{L^2} + C_2\, N^{2-m}\, \|u^{(m)}\|_{L^2} + D\, N^{1-m}\, \|f^{(m)}\|_{L^2}, \quad (7.69)$$

kjer so konstante C_1, C_2 in D neodvisne od N in m.

Dokaz: Po lemi 7.1 veljata pogoja pozitivne definitnosti (7.63) in zveznosti (7.64). Za dokaz izreka je potrebno dokazati še, da vsi trije členi v neenačbi (7.68) konvergirajo proti 0, ko gre $N \to \infty$. Ker velja neenačba (2.19)

$$\|u - R_N u\|_{H^1} \leq C_1\, N^{1-m} \|u^{(m)}\|_{L^2},$$

kjer je m red gladkosti funkcije desne strani f v diferencialni enačbi (7.54) in rešitve u, in ker velja neenačba

$$\frac{|(Q_N f, e)_N - (f, e)|}{\|e\|_{H^1}} \leq \frac{\|Q_N f - f\|_{H^1} \|e\|_{H^1}}{\|e\|_{H^1}} \leq \tfrac{D}{2}\, N^{1-m}\, \|f^{(m)}\|_{L^2},$$

je gornja zahteva izpolnjena za prva dva člena neenakosti (7.68). Za oceno tretjega člena uporabimo neenakost $\|v\|_{H^1} \leq \|v\|_{H^2}$, ki velja za vsak $v \in H^2(-1,1)$, kjer je

$$H^2(-1,1) = \left\{ v \in L^2(-1,1); v', v'' \in L^2(-1,1) \right\},$$

prostor Soboljeva opremljen z normo

$$\|v\|_{H^2} = \left(\|v\|_{L^2}^2 + \|v'\|_{L^2}^2 + \|v''\|_{L^2}^2 \right)^{1/2},$$

Diferencialni operator $\mathcal{L}$ je v prostoru $H^2(-1,1)$ omejen (glej G. Leoni [42] ali W. P. Ziemer [63]). Velja ocena

$$\begin{aligned}
& \frac{|(\mathcal{L}R_N u, e) - (Q_N \mathcal{L}_N R_N u, e)_N|}{\|e\|_{H^1}} \\
\leq\ & \frac{\|\mathcal{L}R_N u - Q_N \mathcal{L}_N R_N u\|_{H^1} \|e\|_{H^1}}{\|e\|_{H^1}} \\
\leq\ & \|\mathcal{L}R_N u - \mathcal{L}u\|_{H^1} + \|\mathcal{L}u - Q_N \mathcal{L}_N u\|_{H^1} + \|Q_N \mathcal{L}_N u - Q_N \mathcal{L}_N R_N u\|_{H^1} \\
\leq\ & \|\mathcal{L}\|_{H^2} \, \|R_N u - u\|_{H^2} + \|f - Q_N f\|_{H^1} + \|Q_N \mathcal{L}_N\|_{H^2} \, \|u - R_N u\|_{H^2} \\
\leq\ & C_2 \, N^{2-m} \, \|u^{(m)}\|_{L^2} + \tfrac{D}{2} \, N^{1-m} \, \|f^{(m)}\|_{L^2}. \qquad (7.70)
\end{aligned}$$

Vse konstante C_1, C_2 in D so neodvisne od N in m. □

Če so koeficientne funkcije α, β in γ, funkcija desne strani f ter rešitev u gladke ali analitične funkcije na neki domeni, ki vsebuje interval $[-1,1]$, tedaj ima pod pogoji izreka 6.9 Čebišev-Fourierova kolokacijska (CFC) metoda spektralno konvergenco. Dokaz spodnjega izreka, ki določa oceno napake in stopnjo konvergence za robne probleme (7.54) z analitičnimi funkcijami, sledi dokazu izreka 7.2 in ga ne bomo navedli.

Izrek 7.3 *Naj bo $u \in L^2(-1,1)$ točna rešitev robnega problema (7.54) z robnimi pogoji (7.55) ter naj bo $u_N \in T_N$ numerična rešitev dobljena z uporabo razreda Čebišev-Fourierovih kolokacijskih (CFC) metod, ki smo jih konstruirali v razdelku 7.2. Privzemimo, da so koeficientne funkcije α, β in γ, funkcija desne strani f ter rešitev u analitične funkcije na domeni $D(R)$, ki je definirana v izreku 6.9. Tedaj je ocena napake za aproksimacijo rešitve za razred CFC metod enaka*

$$\|u - u_N\| \sim \rho^{-N}, \qquad (7.71)$$

kjer je $\rho = \min(3 + 2\sqrt{2}, 2R + \sqrt{4R^2 - 1})$.

7.4 Numerični primeri

V naslednjih primerih bomo primerjali maksimalno absolutno vrednost napake v odvisnosti od števila členov vrste za numerične rešitve, ki jih dobimo z novim razredom Čebišev-Fourierovih kolokacijskih (CFC) spektralnih metod ter s standardnimi kolokacijskimi spektralnimi metodami Čebiševa (CC).

Splošna ugotovitev je, da so rezultati, ki so dobljeni s CFC metodami, večinoma primerljivi z rezultati, ki so dobljeni s CC metodami, čeprav za marsikateri problem napaka s CC metodo pada hitreje, tj. napaka za manjšo vrednost N doseže strojno natančnost, kot s CFC metodo. Predvsem v primerih, ko je rešitev polinomska, je CC metoda zaradi narave aproksimacije bistveno boljša. V primerih, kjer nastopajo gladke oz. analitične funkcije, je vidna spektralna natančnost za oba tipa metod, tj. napaka pada

eksponentno. V primerih, kjer funkcije niso gladke, ampak le nekajkrat zvezno odvedljive, pa opazimo obnašanje napake kot ga napoveduje izrek 7.2. Napaka s CC metodo pada nekoliko hitreje kot napaka s CFC metodo le v nekaterih primerih, ko rešitev ni gladka funkcija, pač pa le nekajkrat zvezno odvedljiva. Poleg tega dobimo boljše rezultate s CC metodo tudi v primeru, ko je rešitev problema hitro oscilirajoča funkcija.

Računska zahtevnost pri CFC metodah pa je precej večja kot pri CC metodah, saj pri prvih nimamo na voljo podobnega orodja za določitev spektralnih koeficientov kot je FFT oz. DCT pri Fourierovih metodah oz. metodah Čebiševa.

Vsi prikazani primeri predstavljajo neperiodične robne probleme. Numerični zgledi (izračun numeričnih rešitev ter izris slik) so izvedeni z uporabo programskega paketa `Matlab` [43], kjer so ustrezne funkcijske m-datoteke in skripte tako za CFC kot za CC metodo implementirane glede na opisane konstrukcije v razdelkih 7.2 in 3.4. Skozi celoten razdelek uporabljamo okrajšavi $y' = \frac{\mathrm{d}y}{\mathrm{d}x}$ in $y'' = \frac{\mathrm{d}^2y}{\mathrm{d}x^2}$.

Primer 7.4 Za prvi primer vzamemo štiri linearne diferencialne enačbe drugega reda, dve s konstantnimi in dve z nekonstantnimi koeficienti. Pri vseh imamo Dirichletove robne pogoje.

(i) Robni problem s konstantnimi koeficienti

$$y'' - 5y' + 6y = x, \quad y(-1) = \mathrm{e}^{-2} + 2\mathrm{e}^{-3} - \tfrac{1}{36}, \quad y(1) = \mathrm{e}^{2} + 2\mathrm{e}^{3} + \tfrac{11}{36},$$

ki ima rešitev

$$y(x) = \mathrm{e}^{2x} + 2\mathrm{e}^{3x} + \tfrac{x}{6} + \tfrac{5}{36}.$$

(ii) Robni problem s konstantnimi koeficienti in s homogenimi robnimi pogoji

$$-y'' + y' + 2y = x, \qquad y(-1) = 0 = y(1),$$

ki ima rešitev

$$y(x) = \tfrac{x}{2} - \tfrac{1}{4} - \tfrac{\mathrm{e}^2(\mathrm{e}^2+3)}{4(\mathrm{e}^6-1)}\mathrm{e}^{2x} + \tfrac{\mathrm{e}(3\mathrm{e}^4+1)}{4(\mathrm{e}^6-1)}\mathrm{e}^{-x}.$$

(iii) Robni problem z nekonstantnimi koeficienti

$$y'' + xy' + y = x\cos x, \quad y(-1) = -\sin 1, \quad y(1) = \sin 1,$$

ki ima rešitev

$$y(x) = \sin x.$$

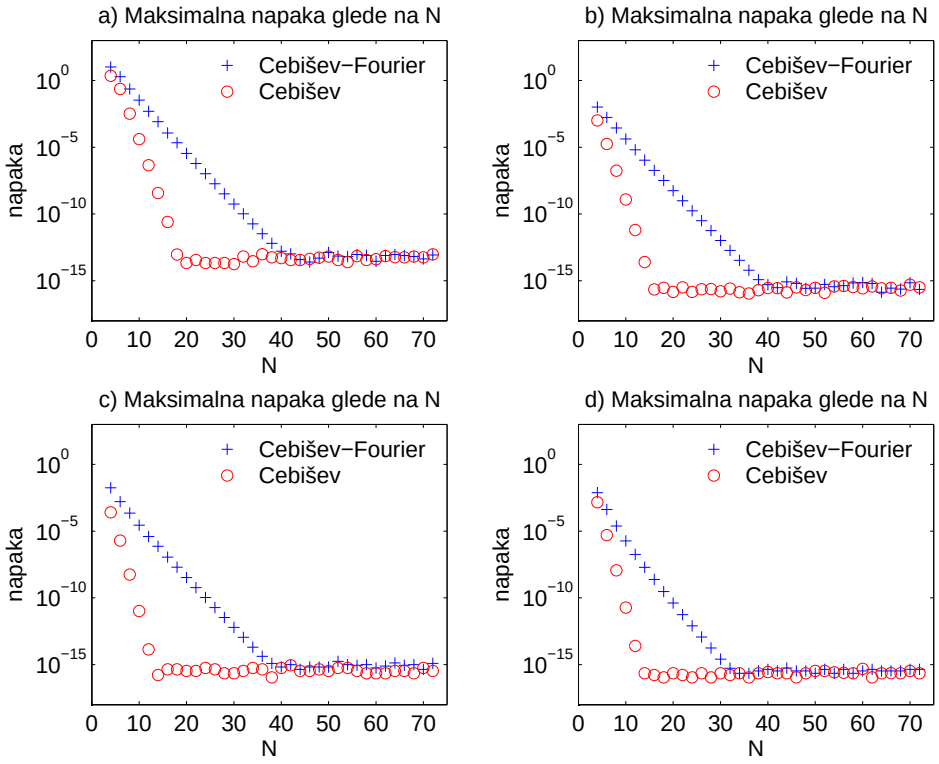
(iv) Robni problem z nekonstantnimi koeficienti

$$y'' + xy' = (2 + x^2)\cos x, \quad y(-1) = \sin 1, \quad y(1) = \sin 1,$$

ki ima rešitev

$$y(x) = x\sin x.$$

Na sliki 7.3 je prikazana primerjava padanja največje absolutne vrednosti napake glede na odrezno število N za numerični rešitvi, ki sta dobljeni s CFC in CC metodo. Na levem zgornjem polju imamo primerjavo za robni problem (i), na desnem zgornjem polju za problem (ii), na levem spodnjem polju za problem (iii), na desnem spodnjem polju pa za problem (iv). Opazimo, da za obe metodi dobimo spektralno konvergenco, saj pada napaka z N eksponentno.



Slika 7.3: Primerjava največjih absolutnih napak numeričnih rešitev glede na število členov vrste N s kolokacijsko metodo Čebiševa (rdeči krogci) in s Čebišev-Fourierovo kolokacijsko metodo (modri plusi). a) Napake za robni problem (i). b) Napake za problem (ii). c) Napake za problem (iii). d) Napake za problem (iv).

Na vseh slikah pa je vidno, da napaka rešitve s kolokacijsko metodo Čebiševa pada precej hitreje kot napaka rešitve s Čebišev-Fourierovo kolokacijsko metodo. To se odraža v tem, da doseže napaka strojno natančnost s prvo metodo pri precej nižji vrednosti za odrezno število N kot napaka z drugo metodo. Obravnavani robni problemi so precej enostavni, njihove točne rešitve pa gladke neperiodične funkcije, katerih razvoj po potencah hitro konvergira. V takih primerih je torej CC metoda učinkovitejša od

CFC metode.

Primer 7.5 Za drugi primer vzamemo štiri linearne diferencialne enačbe drugega reda, prva je Eulerjeva, preostale tri so Airy-jeve, oz. Airy-jevega tipa. Pri vseh imamo Dirichletove robne pogoje.

(i) Eulerjev robni problem

$$x^2y'' - 4xy' + 6y = 0, \quad y(-1) = 0, \quad y(1) = 2,$$

ki ima rešitev

$$y(x) = x^2 + x^3.$$

(ii) Airy-jev robni problem

$$y'' - xy = 0, \qquad y(-1) = 1 = y(1),$$

ki ima rešitev

$$y(x) = \frac{\mathrm{Bi}(-1)-\mathrm{Bi}(1)}{\mathrm{Ai}(1)\mathrm{Bi}(-1)-\mathrm{Ai}(-1)\mathrm{Bi}(1)}\mathrm{Ai}(x) + \frac{\mathrm{Ai}(1)-\mathrm{Ai}(-1)}{\mathrm{Ai}(1)\mathrm{Bi}(-1)-\mathrm{Ai}(-1)\mathrm{Bi}(1)}\mathrm{Bi}(x).$$

kjer sta $\mathrm{Ai}(x)$ in $\mathrm{Bi}(x)$ Airy-jevi funkciji prve in druge vrste (glej M. Abramowitz in I. A. Stegun [1]).

(iii) Airy-jev robni problem

$$y'' - 4096xy = 0, \qquad y(-1) = 1 = y(1),$$

ki ima rešitev

$$y(x) = \frac{\mathrm{Bi}(-16)-\mathrm{Bi}(16)}{\mathrm{Ai}(16)\mathrm{Bi}(-16)-\mathrm{Ai}(-16)\mathrm{Bi}(16)}\mathrm{Ai}(16x) + \frac{\mathrm{Ai}(16)-\mathrm{Ai}(-16)}{\mathrm{Ai}(16)\mathrm{Bi}(-16)-\mathrm{Ai}(-16)\mathrm{Bi}(16)}\mathrm{Bi}(16x).$$

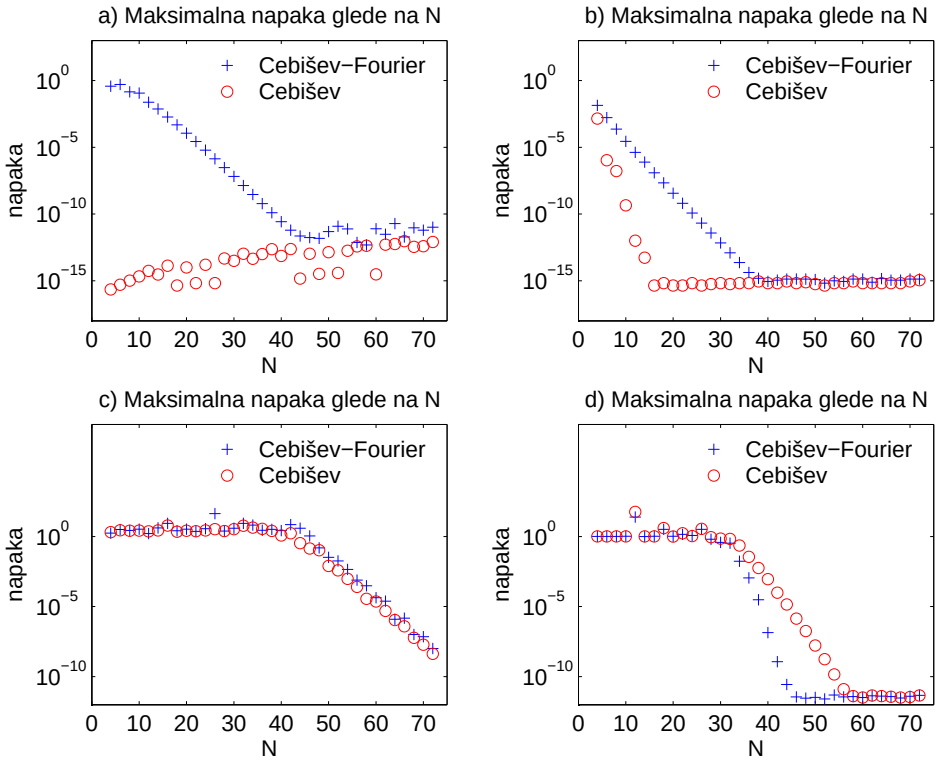
(iv) Airy-jev robni problem

$$y'' - (x - 1000)y = 0, \qquad y(-1) = 1 = y(1),$$

ki ima rešitev

$$\begin{aligned} y(x) \quad = \quad & \frac{\mathrm{Bi}(-1001)-\mathrm{Bi}(-999)}{\mathrm{Ai}(-999)\mathrm{Bi}(-1001)-\mathrm{Ai}(-1001)\mathrm{Bi}(-999)}\mathrm{Ai}(x - 1000) + \\ & \frac{\mathrm{Ai}(-999)-\mathrm{Ai}(-1001)}{\mathrm{Ai}(-999)\mathrm{Bi}(-1001)-\mathrm{Ai}(-1001)\mathrm{Bi}(-999)}\mathrm{Bi}(x - 1000). \end{aligned}$$

Na sliki 7.4 je prikazana primerjava padanja največje absolutne vrednosti napake glede na odrezno število N za numerični rešitvi, ki sta dobljeni s CFC in CC metodo. Na levem zgornjem polju imamo primerjavo za robni problem (i). V tem primeru opazimo, da je metoda CC bistveno boljša, kar pa ni presenetljivo, saj ima Eulerjeva diferencialna enačba, za katero ima karakteristični polinom dve različni pozitivni realni ničli, polinomsko



Slika 7.4: Primerjava največjih absolutnih napak numeričnih rešitev glede na število členov vrste N s kolokacijsko metodo Čebiševa (rdeči krogci) in s Čebišev-Fourierovo kolokacijsko metodo (modri plusi). a) Napake za robni problem (i). b) Napake za problem (ii). c) Napake za problem (iii). d) Napake za problem (iv).

rešitev. Z ortogonalnimi polinomi tako rešitev zelo dobro aproksimiramo že za zelo majhne vrednosti za N, kar pa za aproksimacije s trigonometričnimi funkcijami ne velja. Vseeno pa opazimo, da napaka s CFC metodo pada eksponentno.

Na desnem zgornjem polju imamo primerjavo za robni problem (ii). Podobno kot v zgledih primera 7.4 dobimo za obe metodi spektralno konvergenco, s tem da zopet CC metoda konvergira hitreje od CFC metode. Na levem spodnjem polju imamo primerjavo za robni problem (iii). Opazimo, da začne napaka padati šele od nekega N dalje. Vzrok temu je, da točna rešitev tega problema oscilira. Napaka je za obe numerični rešitvi povsem primerljiva in od nekega N dalje pada eksponentno. Na desnem spodnjem polju imamo primerjavo za robni problem (iv). Kot v prejšnjem primeru je točna rešitev hitro oscilirajoča, kar ima za posledico, da napaka do nekega N ne pada. Ko pa začne padati, se zmanjšuje eksponentno. Opazimo,

da doseže v tem primeru napaka numerične rešitve s CFC metodo strojno natančnost prej kot napaka numerične rešitve s CC metodo.

Primer 7.6 Za tretji in zadnji primer vzamemo štiri linearne diferencialne enačbe drugega reda, katerih točne rešitve niso gladke, ampak so samo nekajkrat zvezno odvedljive. Pri vseh imamo Dirichletove robne pogoje.

(i) Robni problem

$$y'' + |x|y' + y = 6|x| + |x^3| + 3x^3, \quad y(-1) = 1, \quad y(1) = 1,$$

ki ima dvakrat zvezno odvedljivo rešitev

$$y(x) = |x^3|.$$

(ii) Robni problem

$$y'' - 2|x|y' + 3y = 12x|x| + 3x^3|x| - 8x^4, \quad y(-1) = -1, \quad y(1) = 1,$$

ki ima trikrat zvezno odvedljivo rešitev

$$y(x) = x^3|x|.$$

(iii) Robni problem

$$y'' + 2|x|y' - y = 20x^2|x| - x^4|x| + 10x^5, \quad y(-1) = 1, \quad y(1) = 1,$$

ki ima štirikrat zvezno odvedljivo rešitev

$$y(x) = x^4|x|.$$

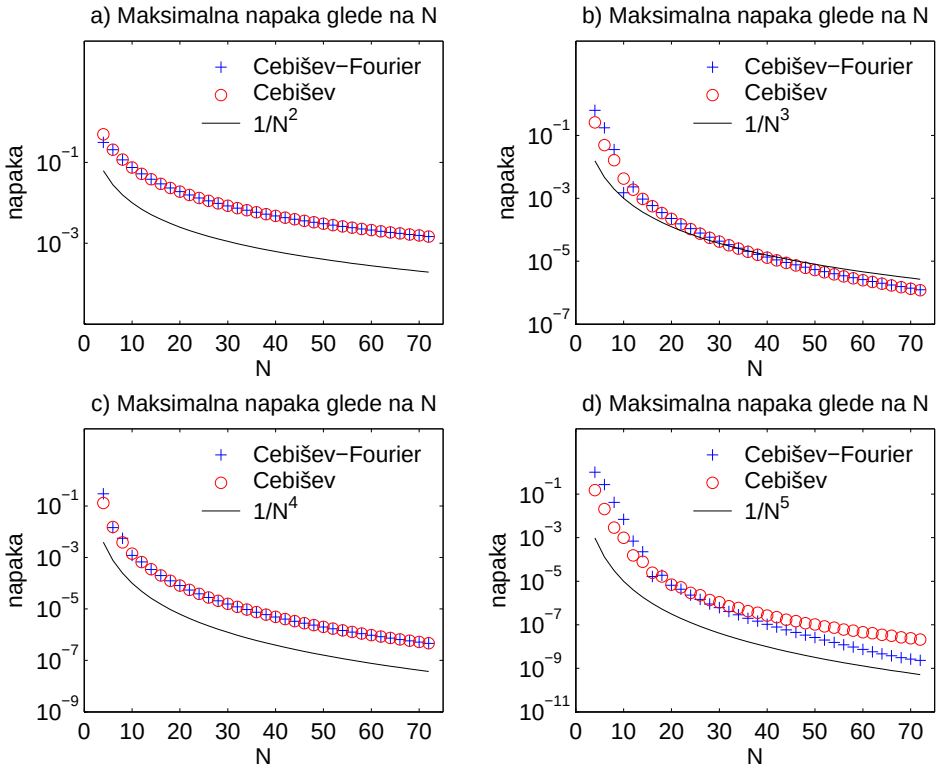
(iv) Robni problem

$$y'' - |x|y' + 2y = 30x^3|x| + 2x^5|x| - 6x^6, \quad y(-1) = -1, \quad y(1) = 1,$$

ki ima petkrat zvezno odvedljivo rešitev

$$y(x) = x^5|x|.$$

Na sliki 7.5 je prikazana primerjava padanja največje absolutne vrednosti napake glede na odrezno število N za numerični rešitvi, ki sta dobljeni s CFC in CC metodo. Koeficienti diferencialne enačbe, funkcija desne strani ter točna rešitev so v vseh primerih zgolj nekajkrat zvezno odvedljive, torej niso gladke. Posledično opazimo, da maksimalna absolutna napaka glede na število členov N ne pada eksponentno, torej v tem primeru nimamo spektralne natančnosti.



Slika 7.5: Primerjava največjih absolutnih napak numeričnih rešitev glede na število členov vrste N s kolokacijsko metodo Čebiševa (rdeči krogci) in s Čebišev-Fourierovo kolokacijsko metodo (modri plusi). Prikazana je tudi referenčna vrednost $1/N^k$ za oceno napake. a) Napake za robni problem (i), $k = 2$. b) Napake za problem (ii), $k = 3$. c) Napake za problem (iii), $k = 4$. d) Napake za problem (iv), $k = 5$.

Na levem zgornjem polju imamo primerjavo za robni problem (i). Opazimo, da je napaka za obe metodi povsem primerljiva in je reda $\mathcal{O}\left(1/N^2\right)$. Na desnem zgornjem polju imamo primerjavo za problem (ii). Tudi tu dobimo primerljivo napako za obe metodi, ki je reda $\mathcal{O}\left(1/N^3\right)$. Na levem spodnjem polju imamo primerjavo za problem (iii). Povsem primerljiva napaka je reda $\mathcal{O}\left(1/N^4\right)$. Na desnem spodnjem polju imamo primerjavo za problem (iv). V tem primeru pa se izkaže, da napaka s CFC metodo pada nekoliko hitreje kot napaka s CC metodo. Obe napaki sta reda $\mathcal{O}\left(1/N^5\right)$. Na primeru robnih problemov, kjer ne nastopajo gladke funkcije, opazimo, da je CFC metoda vsaj tako dobra kot CC metoda, v nekaterih primerih pa celo boljša.

Poglavje 8

Linearni evolucijski problemi

V tem poglavju bomo obravnavali konstrukcijo Čebišev-Fourierovih kolokacijskih spektralnih metod za reševanje (posplošenih) toplotnih enačb paraboličnega (problem 1.4) ter (posplošenih) valovnih enačb hiperboličnega tipa (problem 1.5). Pri teh problemih nastopa ena prostorska in ena časovna spremenljivka, ki ju obravnavamo ločeno. Numerično rešitev problema (4.5 – 4.7) in problema (4.27 – 4.29) tako iščemo v dveh korakih, kot smo to naredili v četrtem poglavju, kjer smo obravnavali konstrukcijo kolokacijskih spektralnih metod Čebiševa.

Za diskretizacijo po prostorski spremenljivki na intervalu $[-1, 1]$ uporabimo kolokacijske spektralne metode, ki smo jih konstruirali v sedmem poglavju, da dobimo začetni problem (4.1) oblike

$$\dot{\mathbf{u}} = \mathbf{f}(t, \mathbf{u}), \quad \mathbf{u}(t_0) = \mathbf{u}_0,$$

ki ga rešimo z uporabo bodisi standardne eksplicitne metode Runge-Kutta četrtega reda (4.2), bodisi eksplicitne Magnusove metode četrtega reda (4.4). V primeru posplošenega hiperboličnega problema (4.27 – 4.29) najprej uvedemo novo spremenljivko, da linearno diferencialno enačbo drugega reda prevedemo na sistem dveh linearnih diferencialnih enačb prvega reda.

Obravnavali bomo zgolj konstrukcijo Čebišev-Fourierovih kolokacijskih (CFC) metod, ne pa tudi analize napake in reda konvergence. To velja tako za CFC metodo za posplošene toplotne enačbe kot tudi za CFC metodo za posplošene valovne enačbe. V obeh primerih ostaja analiza konvergence odprto vprašanje, kar je podlaga za nadaljnje raziskovalno delo na tem področju. Kljub temu pa bomo za oba primera pokazali nekaj numeričnih zgledov, ki potrjujejo spektralno natančnost za gladke oz. analitične funkcije in kažejo na primerljivost Čebišev-Fourierovih kolokacijskih spektralnih metod in kolokacijskih spektralnih metod Čebiševa (CC).

V nadaljevanju poglavja uporabljamo okrajšave za parcialne odvode: $u_x \equiv \frac{\partial u}{\partial x}$, $u_{xx} \equiv \frac{\partial^2 u}{\partial x^2}$, $u_t \equiv \frac{\partial u}{\partial t}$ in $u_{tt} \equiv \frac{\partial^2 u}{\partial t^2}$.

8.1 Konstrukcija Čebišev-Fourierove kolokacijske metode za posplošene toplotne enačbe

Zanimajo nas linearni evolucijski problemi (4.5) v eni dimenziji oblike

$$u_t = \alpha(x,t)u_{xx} + \beta(x,t)u_x + \gamma(x,t)u + \delta(x,t),$$

kjer je $x \in [-1,1]$, $t \geq 0$, koeficientne funkcije α, β, γ in δ pa so v splošnem odvisne od obeh spremenljivk. Poleg enačbe imamo podan tudi začetni pogoj (4.6)

$$u(x,0) = f(x), \quad x \in [-1,1]$$

ter Dirichletove robne pogoje (4.7)

$$u(-1,t) = g(t), \quad u(1,t) = h(t), \quad t \geq 0,$$

ki naj bodo konsistentni: $g(0) = f(-1)$, $h(0) = f(1)$.

Konstrukcijo Čebišev-Fourierove kolokacijske (CFC) metode izvedemo podobno kot konstrukcijo kolokacijske metode Čebiševa. Točno rešitev u diferencialne enačbe (4.5) aproksimiramo z odrezano poldomensko Čebišev-Fourierovo (HCF) vrsto P^N (6.10), ki je razvita po poldomenskih polinomih Čebiševa prve T_k^h in druge U_k^h vrste

$$u(x,t) \approx P^N(x,t) = \sum_{k=0}^{N} a_k(t) T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k(t) U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \quad (8.1)$$

kjer so koeficienti a_k in b_k funkcije odvisne od časovne spremenljivke t in kjer je N odrezno število vrste, ki je povezano s številom iskanih spektralnih koeficientov. Za odrezno število N imamo tako $2N+1$ koeficientov. Parcialne odvode v enačbi (4.5) prav tako aproksimiramo z odrezanimi HCF vrstami tako, da odvajamo vrsto (8.1)

$$u_t(x,t) \approx P_t^N(x,t) = \sum_{k=0}^{N} \dot{a}_k(t) T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} \dot{b}_k(t) U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \quad (8.2)$$

$$u_x(x,t) \approx P_x^N(x,t) = \sum_{k=0}^{N} a_k'(t) T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k'(t) U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \quad (8.3)$$

$$u_{xx} \approx P_{xx}^N(x,t) = \sum_{k=0}^{N} a_k''(t) T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k''(t) U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}. \quad (8.4)$$

Pri tem so koeficienti $\dot{a}_k$ in $\dot{b}_k$ odvodi osnovnih koeficientov a_k in b_k po spremenljivki t, koeficiente a_k, b_k, a_k', b_k', a_k'' in b_k'' pa povezujeta matrični enačbi

$$\mathbf{u}' = D\,\mathbf{u} \quad \text{in} \quad \mathbf{u}'' = D^2\,\mathbf{u}, \quad (8.5)$$

kjer je D operatorska matrika odvajanja (7.16), ki je konstruirana in podrobno opisana v podrazdelku 7.1.1. Spektralne koeficiente odrezanih HCF vrst P^N, P_x^N in P_{xx}^N po vrsti označimo z

$$\begin{aligned}
\mathbf{u}(t) &= (a_0(t), a_1(t), \ldots, a_N(t), b_0(t), b_1(t), \ldots, b_{N-1}(t))^T, \\
\mathbf{u}'(t) &= \left(a_0'(t), a_1'(t), \ldots, a_N'(t), b_0'(t), b_1'(t), \ldots, b_{N-1}'(t)\right)^T, \\
\mathbf{u}''(t) &= \left(a_0''(t), a_1''(t), \ldots, a_N''(t), b_0''(t), b_1''(t), \ldots, b_{N-1}''(t)\right)^T.
\end{aligned}$$

Nato z odrezanimi HCF vrstami aproksimiramo tudi koeficientne funkcije α, β in γ

$$\alpha(x,t) \approx \alpha_N(x,t) = \sum_{k=0}^{N} a_k^\alpha(t) T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k^\alpha(t) U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \tag{8.6}$$

$$\beta(x,t) \approx \beta_N(x,t) = \sum_{k=0}^{N} a_k^\beta(t) T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k^\beta(t) U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \tag{8.7}$$

$$\gamma(x,t) \approx \gamma_N(x,t) = \sum_{k=0}^{N} a_k^\gamma(t) T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k^\gamma(t) U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \tag{8.8}$$

kjer z

$$\begin{aligned}
\mathbf{u}_\alpha(t) &= \left(a_0^\alpha(t), \ldots, a_N^\alpha(t), b_0^\alpha(t), \ldots, b_{N-1}^\alpha(t)\right)^T, \\
\mathbf{u}_\beta(t) &= \left(a_0^\beta(t), \ldots, a_N^\beta(t), b_0^\beta(t), \ldots, b_{N-1}^\beta(t)\right)^T, \\
\mathbf{u}_\gamma(t) &= \left(a_0^\gamma(t), \ldots, a_N^\gamma(t), b_0^\gamma(t), \ldots, b_{N-1}^\gamma(t)\right)^T
\end{aligned}$$

označimo pripadajoče spektralne koeficiente. Poleg tega z

$$\begin{aligned}
\tilde{\mathbf{u}}_\alpha(t) &= \left(\tilde{a}_0^\alpha(t), \ldots, \tilde{a}_N^\alpha(t), \tilde{b}_0^\alpha(t), \ldots, \tilde{b}_{N-1}^\alpha(t)\right)^T, \\
\tilde{\mathbf{u}}_\beta(t) &= \left(\tilde{a}_0^\beta(t), \ldots, \tilde{a}_N^\beta(t), \tilde{b}_0^\beta(t), \ldots, \tilde{b}_{N-1}^\beta(t)\right)^T, \\
\tilde{\mathbf{u}}_\gamma(t) &= \left(\tilde{a}_0^\gamma(t), \ldots, \tilde{a}_N^\gamma(t), \tilde{b}_0^\gamma(t), \ldots, \tilde{b}_{N-1}^\gamma(t)\right)^T
\end{aligned}$$

po vrsti označimo koeficiente produktov $\alpha_N P_{xx}^N$, $\beta_N P_x^N$ in $\gamma_N P^N$, ki nastopajo v aproksimaciji enačbe (4.5) z odrezanimi poldomenskimi Čebišev-Fourierovimi vrstami. Povezava med koeficienti $\tilde{\mathbf{u}}_\alpha$, $\tilde{\mathbf{u}}_\beta$, $\tilde{\mathbf{u}}_\gamma$ in koeficienti $\mathbf{u}$, $\mathbf{u}'$, $\mathbf{u}''$ je podana z matričnimi zvezami

$$\tilde{\mathbf{u}}_\alpha = F_\alpha\, \mathbf{u}'', \quad \tilde{\mathbf{u}}_\beta = F_\beta\, \mathbf{u}' \quad \text{in} \quad \tilde{\mathbf{u}}_\gamma = F_\gamma\, \mathbf{u}, \tag{8.9}$$

kjer so F_α, F_β in F_γ operatorske multiplikacijske matrike (7.35), ki pripadajo posameznim koeficientnim funkcijam. Konstrukcija teh matrik je podrobno opisana v podrazdelku 7.1.2.
Nadalje razvijemo tudi nehomogeni del enačbe δ v odrezano HCF vrsto

$$\delta(x,t) \approx \delta_N(x,t) = \sum_{k=0}^{N} a_k^\delta(t) T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k^\delta(t) U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \tag{8.10}$$

kjer z

$$\mathbf{u}_\delta(t) = \left(a_0^\delta(t), \ldots, a_N^\delta(t), b_0^\delta(t), \ldots, b_{N-1}^\delta(t)\right)^T$$

označimo vektor pripadajočih spektralnih koeficientov.

Za izračun spektralnih koeficientov a_k in b_k uporabimo metodo kolokacije, kjer za kolokacijske vozle izberemo točke Čebiševa $x_i = -\cos\left(\frac{i\pi}{2N}\right)$, $i = 0, 1, \ldots, 2N$, (3.11), da dobimo sistem navadnih diferencialnih enačb prvega reda

$$\begin{aligned} &\sum_{k=0}^{N} \dot{a}_k(t) T_k^h(\cos\tfrac{\pi x_i}{2}) + \sum_{k=0}^{N-1} \dot{b}_k(t) U_k^h(\cos\tfrac{\pi x_i}{2}) \sin\tfrac{\pi x_i}{2} \\ = &\sum_{k=0}^{N} \left(\tilde{a}_k^\alpha(t) + \tilde{a}_k^\beta(t) + \tilde{a}_k^\gamma(t) + a_k^\delta(t)\right) T_k^h(\cos\tfrac{\pi x_i}{2}) \\ &+ \sum_{k=0}^{N} \left(\tilde{b}_k^\alpha(t) + \tilde{b}_k^\beta(t) + \tilde{b}_k^\gamma(t) + b_k^\delta(t)\right) U_k^h(\cos\tfrac{\pi x_i}{2}) \sin\tfrac{\pi x_i}{2} \end{aligned} \tag{8.11}$$

za notranje kolokacijske točke x_i, $i = 1, 2, \ldots, 2N-1$. Naj bo nadalje $C \in \mathbb{R}^{(2N+1)\times(2N+1)}$ kolokacijska matrika vrednosti poldomenskih polinomov Čebiševa v kolokacijskih vozlih, kjer označimo $c(x_i) = \cos\frac{\pi x_i}{2}$ in $s(x_i) = \sin\frac{\pi x_i}{2}$, $i = 1, 2, \ldots, 2N-1$,

$$C^T = \begin{bmatrix} T_0^h(c(x_0)) & T_0^h(c(x_1)) & \cdots & T_0^h(c(x_{2N})) \\ \vdots & \vdots & & \vdots \\ T_N^h(c(x_0)) & T_N^h(c(x_1)) & \cdots & T_N^h(c(x_{2N})) \\ U_0^h(c(x_0))s(x_0) & U_0^h(c(x_1))s(x_1) & \cdots & U_0^h(c(x_{2N}))s(x_{2N}) \\ \vdots & \vdots & & \vdots \\ U_{N-1}^h(c(x_0))s(x_0) & U_{N-1}^h(c(x_1))s(x_1) & \cdots & U_{N-1}^h(c(x_{2N}))s(x_{2N}) \end{bmatrix} \tag{8.12}$$

in naj bo $L \in \mathbb{R}^{(2N+1)\times(2N+1)}$ diferencialna operatorska matrika

$$L = F_\alpha\, D^2 + F_\beta\, D + F_\gamma, \tag{8.13}$$

ki pripada diferencialni enačbi (4.5). Kolokacijska matrika C je obrnljiva. Vse operatorske matrike D, F_α, F_β in F_γ so reda $(2N+1)\times(2N+1)$. Sistem

(8.11) lahko zapišemo v matrični obliki

$$\tilde{C}\ \dot{\mathbf{u}} = \tilde{C}\ L\ \mathbf{u} + \tilde{C}\ \mathbf{u}_\delta, \tag{8.14}$$

kjer z

$$\dot{\mathbf{u}}(t) = \Big(\dot{a}_0(t), \ldots, \dot{a}_N(t), \dot{b}_0(t), \ldots, \dot{b}_{N-1}(t)\Big)^T$$

označimo vektor spektralnih koeficientov odrezane HCF vrste P_t^N in kjer je $\tilde{C} \in \mathbb{R}^{(2N-1)\times(2N+1)}$ matrika C brez prve in zadnje vrstice.

Nazadnje upoštevamo še Dirichletove robne pogoje tako, da sistemu (8.11) dodamo enačbi

$$P^N(-1,t) = \sum_{k=0}^{N} a_k(t) T_k^h(0) - \sum_{k=0}^{N-1} b_k(t) U_k^h(0) = g(t), \tag{8.15}$$

$$P^N(1,t) = \sum_{k=0}^{N} a_k(t) T_k^h(0) + \sum_{k=0}^{N-1} b_k(t) U_k^h(0) = h(t). \tag{8.16}$$

Na ta način iz sistema linearnih diferencialnih enačb dobimo sistem linearnih diferencialno-algebraičnih enačb (DAE), ki ga lahko rešimo z uporabo katere izmed metod za reševanje sistemov DAE. Druga možnost, ki pa ne deluje vedno, zagotovo pa v primeru homogenih robnih pogojev, je, da enačbi (8.15) in (8.16) odvajamo po časovni spremenljivki

$$\sum_{k=0}^{N} \dot{a}_k(t) T_k^h(0) - \sum_{k=0}^{N-1} \dot{b}_k(t) U_k^h(0) = \dot{g}(t), \tag{8.17}$$

$$\sum_{k=0}^{N} \dot{a}_k(t) T_k^h(0) + \sum_{k=0}^{N-1} \dot{b}_k(t) U_k^h(0) = \dot{h}(t), \tag{8.18}$$

in sistemu (8.11) dodamo enačbi (8.17 – 8.18). Sistem (8.14) tako zapišemo v obliki

$$\dot{\mathbf{u}}(t) = U\ \mathbf{u}(t) + \tilde{\mathbf{u}}_\delta(t), \tag{8.19}$$

kjer je matrika U dobljena kot produkt inverzne kolokacijske matrike C^{-1} z matriko $\tilde{C}\ L$ iz sistema (8.14), ki ji dodamo zgoraj in spodaj po eno vrstico samih ničel, vektor $\tilde{\mathbf{d}}$ pa je dobljen tako, da matriko C^{-1} pomnožimo z nehomogenim delom $\tilde{C}\ \mathbf{d}$ sistema (8.14), ki mu zgoraj in spodaj dodamo odvoda robnih pogojev, ki sta podana z enačbama (8.17 – 8.18).

Koeficientne funkcije a_k in b_k nato izračunamo tako, da rešimo sistem linearnih diferencialnih enačb prvega reda (8.19), kjer prvi (homogeni) del sistema rešimo z Magnusovo metodo (4.4) ali z Runge-Kutta metodo (4.2) četrtega reda. Vektor začetnih vrednosti dobimo iz začetnega pogoja (4.6), ko funkcijo f razvijemo v odrezano HCF vrsto

$$f(x) \approx \sum_{k=0}^{N} \lambda_k T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} \mu_k U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \tag{8.20}$$

kjer z $\mathbf{u}_0 = (\lambda_0, \lambda_1, \ldots \lambda_N, \mu_0, \mu_1, \ldots \mu_{N-1})^T$ označimo vektor pripadajočih spektralnih koeficientov.

Konstrukcijo Čebišev-Fourierove kolokacijske (CFC) metode bomo v naslednjih numeričnih zgledih primerjali s kolokacijsko metodo Čebiševa (CC) za nekatere preproste (posplošene) toplotne enačbe.

Primer 8.1 Pokazali bomo tri primere (posplošenih) toplotnih enačb.

(i) Prvi primer je toplotna enačba s homogenimi Dirichletovimi robnimi pogoji in sinusno začetno porazdelitvijo toplote na intervalu $[-1, 1]$

$$\begin{aligned} u_t &= u_{xx}, \\ u(x,0) &= \sin(\pi x), \qquad x \in [-1,1], \\ u(-1,t) &= u(1,t) = 0, \qquad t \geq 0, \end{aligned}$$

ki ima rešitev

$$u(x,t) = \mathrm{e}^{-\pi^2 t} \sin(\pi x).$$

(ii) Drugi primer je posplošena toplotna enačba (paraboličnega tipa) s homogenimi Dirichletovimi robnimi pogoji in sinusno začetno porazdelitvijo toplote na intervalu $[-1, 1]$

$$\begin{aligned} u_t &= u_{xx} + u, \\ u(x,0) &= \sin(2\pi x), \qquad x \in [-1,1], \\ u(-1,t) &= u(1,t) = 0, \qquad t \geq 0, \end{aligned}$$

ki ima rešitev

$$u(x,t) = \mathrm{e}^{-(4\pi^2 - 1)t} \sin(2\pi x).$$

(iii) Tretji primer je toplotna enačba s homogenimi Dirichletovimi robnimi pogoji in sinusno začetno porazdelitvijo toplote na intervalu $[-1, 1]$

$$\begin{aligned} u_t &= \tfrac{1}{64} u_{xx}, \\ u(x,0) &= \sin(8\pi x), \qquad x \in [-1,1], \\ u(-1,t) &= u(1,t) = 0, \qquad t \geq 0, \end{aligned}$$

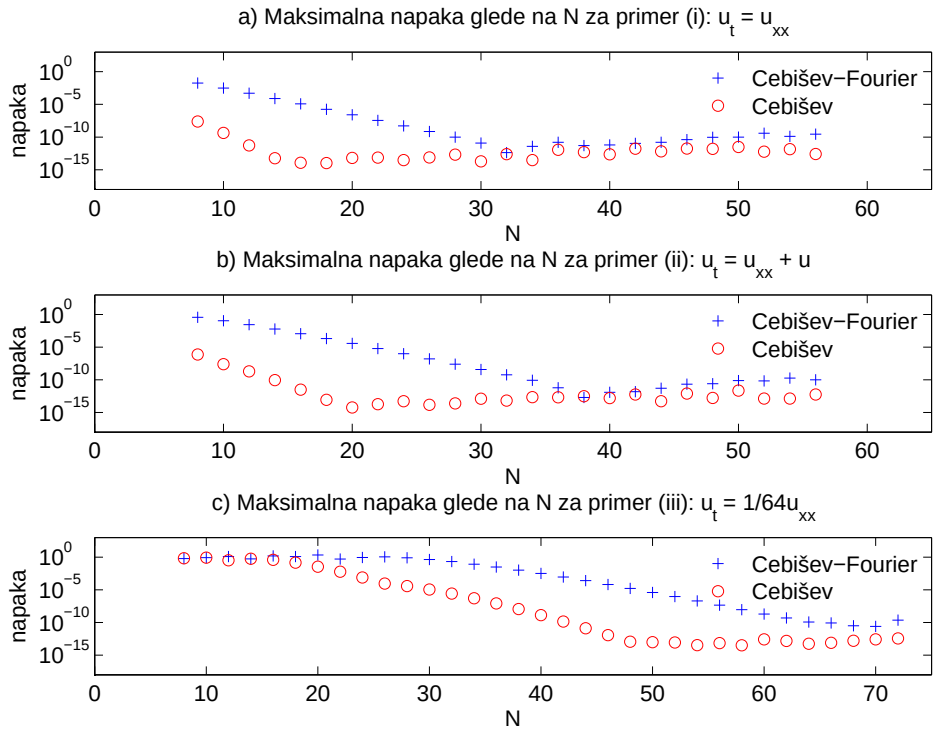
ki ima rešitev

$$u(x,t) = \mathrm{e}^{-\pi^2 t} \sin(8\pi x).$$

Tako začetna porazdelitev toplote kot točna rešitev v tem primeru hitreje oscilirata.

Robni in začetni pogoji so konsistentni. Numerično rešitev iščemo s kolokacijsko metodo Čebiševa (CC) in s Čebišev-Fourierovo kolokacijsko metodo (CFC). Enačbo rešimo za različne vrednosti odreznega števila N ter primerjamo dobljene numerične rešitve s točno ob času $t = 1$. Diskretizacijo

po časovni spremenljivki t napravimo z Magnusovo metodo četrtega reda (4.4). Vsi trije zgledi so zelo preprosti, saj so koeficientne funkcije v vseh primerih konstantne, nehomogeni del je ničelen, robni pogoji pa homogeni. Sistem 8.19 je tedaj homogen, matrika, ki nastopa v Magnusovi metodi, pa konstantna, kar omogoča diskretizacijo po časovni spremenljivki v enem koraku.



Slika 8.1: Primerjava največjih absolutnih napak numeričnih rešitev glede na odrezno število N s kolokacijsko metodo Čebiševa (rdeči krogci) in s Čebišev-Fourierovo kolokacijsko metodo (modri plusi) pri reševanju a) osnovne toplotne enačbe (i) $u_t = u_{xx}$, b) posplošene toplotne enačbe (ii) $u_t = u_{xx} + u$ in c) toplotne enačbe (iii) $u_t = 1/64 u_{xx}$, ki ima hitreje oscilirajočo rešitev.

Slika 8.1 prikazuje primerjavo največjih absolutnih napak numeričnih rešitev glede na odrezno število N z metodama CC in CFC za primer (i) na sliki a) (zgoraj), za primer (ii) na sliki b) (na sredini) ter za primer (iii) na sliki c) (spodaj). Opazimo, da pri obeh metodah maksimalna absolutna vrednost napake pada eksponentno, tj. dobimo spektralno natančnost. V vseh treh primerih, tudi ko točna rešitev hitreje oscilira, napaka pada hitreje z uporabo CC kot z uporabo CFC metode. V primeru (iii) na sliki c) začne napaka padati šele od nekega N dalje za obe metodi.

8.2 Konstrukcija Čebišev-Fourierove kolokacijske metode za posplošene valovne enačbe

Zanimajo nas linearni evolucijski problemi (4.27) v eni dimenziji oblike

$$u_{tt} = \alpha(x,t)u_{xx} + \beta(x,t)u_x + \gamma(x,t)u + \delta(x,t),$$

kjer je $x \in [-1,1]$, $t \geq 0$, koeficientne funkcije α, β, γ in δ pa so v splošnem odvisne od obeh spremenljivk. Poleg enačbe imamo podana tudi začetna pogoja (4.28)

$$u(x,0) = f_1(x), \quad u_t(x,0) = f_2(x), \quad x \in [-1,1]$$

ter Dirichletove robne pogoje (4.29)

$$u(-1,t) = g(t), \quad u(1,t) = h(t), \quad t \geq 0,$$

ki naj bodo konsistentni: $g(0) = f_1(-1)$, $h(0) = f_1(1)$, $g'(0) = f_2(-1)$, $h'(0) = f_2(1)$.

Konstrukcijo psevdospektralne Čebišev-Fourierove (CFC) metode za reševanje posplošenih valovnih enačb (4.27) hiperboličnega tipa izvedemo na podoben način kot za reševanje posplošenih toplotnih enačb (4.5) paraboličnega tipa v razdelku 8.1. Točno rešitev u diferencialne enačbe (4.27) aproksimiramo z odrezano poldomensko Čebišev-Fourierovo (HCF) vrsto P^N (8.1)

$$u(x,t) \approx P^N(x,t) = \sum_{k=0}^{N} a_k(t) T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} b_k(t) U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2},$$

kjer je N odrezno število. Parcialne odvode v enačbi (4.27) prav tako aproksimiramo z odrezanimi HCF vrstami (8.2 – 8.4), kjer poleg naštetih potrebujemo še drugi odvod po t

$$u_{tt}(x,t) \approx P_{tt}^N(x,t) \; = \; \sum_{k=0}^{N} \ddot{a}_k(t) T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} \ddot{b}_k(t) U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}. \tag{8.21}$$

Pri tem so koeficienti $\ddot{a}_k$ in $\ddot{b}_k$ drugi odvodi osnovnih koeficientov a_k in b_k po spremenljivki t. Koeficientne funkcije α, β in γ aproksimiramo z odrezanimi HCF vrstami (8.6 – 8.8), nehomogeni del enačbe δ pa z vrsto (8.10).

Za izračun spektralnih koeficientov a_k in b_k uporabimo metodo kolokacije, kjer za kolokacijske vozle izberemo točke Čebiševa $x_i = -\cos\left(\frac{i\pi}{2N}\right)$, $i = 0, 1, \ldots, 2N$, (3.11), da dobimo sistem navadnih diferencialnih enačb

drugega reda

$$\begin{aligned}
&\sum_{k=0}^{N} \ddot{a}_k(t) T_k^h(\cos\tfrac{\pi x_i}{2}) + \sum_{k=0}^{N-1} \ddot{b}_k(t) U_k^h(\cos\tfrac{\pi x_i}{2}) \sin\tfrac{\pi x_i}{2} \\
= &\sum_{k=0}^{N} \left(\tilde{a}_k^{\alpha}(t) + \tilde{a}_k^{\beta}(t) + \tilde{a}_k^{\gamma}(t) + a_k^{\delta}(t) \right) T_k^h(\cos\tfrac{\pi x_i}{2}) \\
&+ \sum_{k=0}^{N} \left(\tilde{b}_k^{\alpha}(t) + \tilde{b}_k^{\beta}(t) + \tilde{b}_k^{\gamma}(t) + b_k^{\delta}(t) \right) U_k^h(\cos\tfrac{\pi x_i}{2}) \sin\tfrac{\pi x_i}{2} \quad (8.22)
\end{aligned}$$

za notranje kolokacijske točke x_i, $i = 1, 2, \ldots, 2N-1$. Naj bodo C kolokacijska matrika (8.12), $\tilde{C}$ njena podmatrika brez prve in zadnje vrstice ter L diferencialna operatorska matrika (8.13), ki pripada diferencialni enačbi (4.27) in kjer so D, F_α, F_β in F_γ matrike dane z enačbami (8.5) in (8.9). Sistem (8.23) lahko zapišemo v matrični obliki

$$\tilde{C}\ \ddot{\mathbf{u}} = \tilde{C}\ L\ \mathbf{u} + \tilde{C}\ \mathbf{d}, \tag{8.23}$$

kjer z

$$\ddot{\mathbf{u}}(t) = \left(\ddot{a}_0(t), \ldots, \ddot{a}_N(t), \ddot{b}_0(t), \ldots, \ddot{b}_{N-1}(t) \right)^T$$

označimo vektor spektralnih koeficientov odrezane HCF vrste P_{tt}^N.

Nazadnje upoštevamo še Dirichletove robne pogoje tako, da sistemu (8.22) dodamo enačbi (8.15) in (8.16) in dobimo sistem DAE, katerega rešitev obravnavamo enako kot v razdelku 8.1, da dobimo sistem (8.23) v obliki

$$\ddot{\mathbf{u}}(t) = U\ \mathbf{u}(t) + \tilde{\mathbf{u}}_\delta(t). \tag{8.24}$$

Rešitev tega sistema so iskane koeficientne funkcije a_k in b_k. Prvi (homogeni) del sistema (8.24) najprej z uvedbo novih spremenljivk $\mathbf{y}_1 = \mathbf{u}$ in $\mathbf{y}_2 = \dot{\mathbf{u}}$ za katere velja

$$\begin{aligned}
\dot{\mathbf{y}}_1 &= \dot{\mathbf{u}} = \mathbf{y}_2, \\
\dot{\mathbf{y}}_2 &= \ddot{\mathbf{u}} = U\ \mathbf{y}_1,
\end{aligned}$$

prevedemo na sistem linearnih diferencialnih enačb prvega reda

$$\begin{bmatrix} \dot{\mathbf{y}}_1 \\ \dot{\mathbf{y}}_2 \end{bmatrix} = \begin{bmatrix} 0 & I \\ U & 0 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}. \tag{8.25}$$

Sistem (8.25) rešimo z Magnusovo metodo (4.4) ali z Runge-Kutta metodo (4.2) četrtega reda. Vektor začetnih vrednosti dobimo iz začetnih pogojev (4.28), ko funkciji f_1 in f_2 razvijemo v odrezani HCF vrsti

$$f_1(x) \approx \sum_{k=0}^{N} \lambda_k T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} \mu_k U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \tag{8.26}$$

$$f_2(x) \approx \sum_{k=0}^{N} \nu_k T_k^h(\cos\tfrac{\pi x}{2}) + \sum_{k=0}^{N-1} \eta_k U_k^h(\cos\tfrac{\pi x}{2}) \sin\tfrac{\pi x}{2}, \tag{8.27}$$

kjer z $\mathbf{u}_0 = (\lambda_0, \ldots \lambda_N, \mu_0, \ldots \mu_{N-1}, \nu_0, \ldots \nu_N, \eta_0, \ldots \eta_{N-1})^T$ označimo združen vektor pripadajočih spektralnih koeficientov.

Konstrukcijo Čebišev-Fourierove kolokacijske (CFC) metode bomo v naslednjih numeričnih zgledih primerjali s kolokacijsko metodo Čebiševa (CC) za nekatere preproste (posplošene) valovne enačbe.

Primer 8.2 Pokazali bomo tri primere (posplošenih) valovnih enačb.

(i) Prvi primer je valovna enačba s homogenimi Dirichletovimi robnimi pogoji, sinusnim začetnim valom ter ničelno začetno hitrostjo na intervalu $[-1, 1]$

$$\begin{aligned} u_{tt} &= u_{xx}, \\ u(x,0) &= -\tfrac{1}{\pi^2}\sin(\pi x), \qquad x \in [-1,1], \\ u_t(x,0) &= 0, \qquad x \in [-1,1], \\ u(-1,t) &= u(1,t) = 0, \qquad t \geq 0, \end{aligned}$$

ki ima rešitev

$$u(x,t) = -\tfrac{1}{\pi^2}\cos(\pi t)\sin(\pi x).$$

(ii) Drugi primer je posplošena valovna enačba s homogenimi Dirichletovimi robnimi pogoji, sinusnim začetnim valom ter ničelno začetno hitrostjo na intervalu $[-1, 1]$

$$\begin{aligned} u_{tt} &= u_{xx} + 3\pi^2 u, \\ u(x,0) &= \sin(2\pi x), \qquad x \in [-1,1], \\ u_t(x,0) &= 0, \qquad x \in [-1,1], \\ u(-1,t) &= u(1,t) = 0, \qquad t \geq 0, \end{aligned}$$

ki ima rešitev

$$u(x,t) = \cos(\pi t)\sin(2\pi x).$$

(iii) Tretji primer je valovna enačba s homogenimi Dirichletovimi robnimi pogoji, sinusnim začetnim valom ter ničelno začetno hitrostjo na intervalu $[-1, 1]$

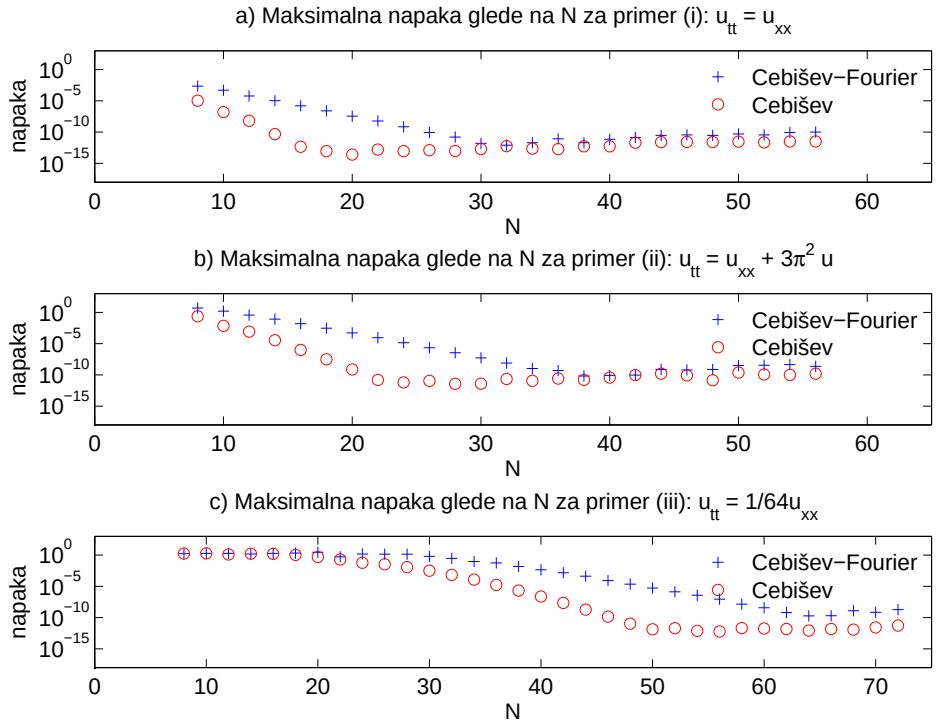
$$\begin{aligned} u_{tt} &= \tfrac{1}{64}u_{xx}, \\ u(x,0) &= \sin(8\pi x), \qquad x \in [-1,1], \\ u_t(x,0) &= 0, \qquad x \in [-1,1], \\ u(-1,t) &= u(1,t) = 0, \qquad t \geq 0, \end{aligned}$$

ki ima rešitev

$$u(x,t) = \cos(\pi t)\sin(8\pi x).$$

Tako začetni val kot točna rešitev v tem primeru hitreje oscilirata.

Robni in začetni pogoji so konsistentni. Numerično rešitev iščemo s kolokacijsko metodo Čebiševa (CC) in s Čebišev-Fourierovo kolokacijsko metodo (CFC). Enačbo rešimo za različne vrednosti odreznega števila N ter primerjamo dobljene numerične rešitve s točno ob času $t = 1$. Diskretizacijo po časovni spremenljivki t napravimo z Magnusovo metodo četrtega reda (4.4). Vsi trije zgledi so zelo preprosti, saj so koeficientne funkcije v vseh primerih konstantne, nehomogeni del je ničelen, robni pogoji pa homogeni. Sistem 8.24 je tedaj homogen, matrika, ki nastopa v Magnusovi metodi, pa konstantna, kar omogoča diskretizacijo po časovni spremenljivki v enem koraku.



Slika 8.2: Primerjava največjih absolutnih napak numeričnih rešitev glede na odrezno število N s kolokacijsko metodo Čebiševa (rdeči krogci) in s Čebišev-Fourierovo kolokacijsko metodo (modri plusi) pri reševanju a) osnovne valovne enačbe (i) $u_{tt} = u_{xx}$, b) posplošene valovne enačbe (ii) $u_{tt} = u_{xx} + 3\pi^2 u$ in c) valovne enačbe (iii) $u_{tt} = 1/64 u_{xx}$, ki ima hitreje oscilirajočo rešitev.

Slika 8.2 prikazuje primerjavo največjih absolutnih napak numeričnih rešitev glede na odrezno število N z metodama CC in CFC za primer (i) na sliki a) (zgoraj), za primer (ii) na sliki b) (na sredini) ter za primer (iii) na sliki c) (spodaj). Opazimo, da pri obeh metodah maksimalna absolutna vrednost napake pada eksponentno, tj. dobimo spektralno natančnost. V

vseh treh primerih, tudi ko točna rešitev hitreje oscilira, napaka pada hitreje z uporabo CC kot z uporabo CFC metode. V primeru (iii) na sliki c) začne napaka padati šele od nekega N dalje za obe metodi.

Primeri, ki so prikazani v tem poglavju, kažejo na to, da se da Čebišev-Fourierovo kolokacijsko metodo uporabiti tudi za reševanje posplošenih toplotnih in valovnih enačb paraboličnega oz. hiperboličnega tipa. Žal pa raziskovalno delo za te tipe linearnih evolucijskih enačb še ni prineslo rezultatov za analizo napake in reda konvergence, kar ostaja odprto vprašanje za v bodoče. Zgledi pa kažejo na spektralno konvergenco za probleme v katerih nastopajo gladke oz. analitične funkcije. V vseh prikazanih primerih pa je hitrost konvergence za kolokacijsko metodo Čebiševa hitrejša kot za Čebišev-Fourierovo kolokacijsko metodo.

Poglavje 9

Sklep

Spektralne metode so poleg metod končnih razlik in metod končnih elementov eden izmed pomembnejših razredov numeričnih metod za reševanje robnih problemov, tako pri navadnih kot tudi pri parcialnih diferencialnih enačbah. Numerično rešitev aproksimiramo s končno (odrezano) funkcijsko vrsto, kjer je potrebno določiti spektralne koeficiente te vrste z eno izmed ustreznih metod, npr. z Galerkinovo metodo, Tau metodo ali metodo kolokacije.

Klasična pristopa sta, da periodične probleme aproksimiramo s Fourierovo vrsto, neperiodične probleme pa z vrsto Čebiševa. V prvem primeru za množico baznih funkcij izberemo trigonometrične funkcije, interval, na katerem iščemo rešitev robnega problema, pa razdelimo z ekvidistantnimi točkami. V drugem primeru za množico baznih funkcij izberemo ortogonalne polinome (najpogosteje polinome Čebiševa prve vrste), interval pa razdelimo s točkami Čebiševa, ki so gosteje razporejene blizu krajišč intervala. S tem se izognemo tako Gibbsovemu kot tudi Rungejevemu fenomenu, ki sta značilna pri uporabi ekvidistantnih točk.

V prvem delu doktorske disertacije smo tako opisali osnovne, predvsem konvergenčne lastnosti Fourierove vrste ter nekaterih osnovnih družin ortogonalnih polinomov (Legendreovi polinomi ter polinomi Čebiševa prve in druge vrste) skupaj s konvergenčnimi lastnostmi vrste Čebiševa. Bistvenega pomena je pojem ortogonalnosti baznih funkcij, kar posledično omogoča izračun spektralnih koeficientov. Poleg tega smo napravili kratek uvod v teorijo spektralnih metod, kjer smo obravnavali tako konstrukcijo metod kot analizo konvergence in napake za te metode. Teorijo smo osvetlili z nekaterimi numeričnimi zgledi uporabe.

Predstavili smo dva standardna pristopa za konstrukcijo spektralnih metod za reševanje modelnega linearnega dvotočkovnega robnega problema z Dirichletovimi robnimi pogoji v eni dimenziji. Ta modelni primer lahko bolj ali manj preprosto posplošimo na Neumannove ali mešane (Robinove) robne pogoje, na linearne diferencialne enačbe višjega reda ali na linearne

robne probleme v dveh ali več dimenzijah. Več dela je s posplošitvami na zahtevnejše nelinearne robne probleme. S tovrstnimi posplošitvami se v tem delu nismo ukvarjali, predstavljajo pa iztočnico za nadaljnje raziskovalno delo. Nekaj časa pa smo posvetili reševanju robnih problemov pri evolucijskih parcialnih diferencialnih enačbah, npr. pri (posplošeni) toplotni ali (posplošeni) valovni enačbi, ki sta paraboličnega oz. hiperboličnega tipa. Opisali smo klasični kolokacijski spektralni metodi Čebiševa (CC) za reševanje teh dveh tipov evolucijskih enačb, kjer smo za diskretizacijo po časovni spremenljivki uporabili bodisi Magnusovo bodisi Runge-Kutta metodo četrtega reda.

Obširni pregled spektralnih metod presega okvirje te doktorske disertacije, zato so predstavljene zgolj osnovne ideje in nekaj rezultatov za njihovo konstrukcijo. Bralec, ki ga tematika spektralnih metod podrobneje zanima, je vabljen, da prebere podrobnosti v širokem spektru literature, ki pokriva to področje, npr. v preglednem članku B. Fornberg in D. M. Sloan [19] ter knjigah J. P. Boyd [7], C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang [10] in [11], B. Fornberg [18], D. Gottlieb in S. A. Orszag [28], B. Mercier [44], J. Shen, T. Tang in L. Wang [51] ter L. N. Trefethen [56] in [57]. Seveda pa je na voljo tudi vrsta originalnih člankov in prispevkov s tega področja.

V drugem delu doktorske disertacije smo najprej definirali in opisali dve neklasični družini ortogonalnih polinomov, tj. poldomenske polinome Čebiševa prve in druge vrste, skupaj s pripadajočo poldomensko Čebišev-Fourierovo (HCF) vrsto. Te polinome je prvi vpeljal D. Huybrechs v članku [35] in so sorodni klasičnim polinomom Čebiševa prve in druge vrste, saj imajo enaki uteži, so pa ortogonalni na krajšem intervalu. Koeficiente v tričlenski rekurzivni formuli, ki predstavlja osnovo za določitev ortogonalnih polinomov, smo izračunali z uporabo stabilnega modificiranega algoritma Čebiševa. Le-ta je v nasprotju z direktnim izračunom rekurzivnih koeficientov stabilen v standardni aritmetiki s plavajočo vejico v dvojni natančnosti (IEEE).

Z uporabo tako definirane poldomenske Čebišev-Fourierove vrste smo nato obravnavali probleme 1.1 – 1.5, ki smo jih zastavili v uvodnem poglavju. Problem 1.1 predstavlja aproksimacijo dane (s kvadratom integrabilne) funkcije z različnimi vrstami (Fourierova vrsta, vrsta Čebiševa, HCF vrsta). Videli smo, da lahko z uporabo HCF vrste aproksimiramo neperiodične funkcije s trigonometrično vrsto, ki je sestavljena iz trigonometričnih sinusnih in kosinusnih funkcij ter sinusnih in kosinusnih funkcij polovičnih kotov. Le-te so organizirane v obliki ortogonalnih poldomenskih polinomov Čebiševa prve in druge vrste. To pomeni, da lahko s primerno reorganizacijo rešujemo neperiodične probleme z orodji, ki so značilna za periodične probleme. Rezultati numeričnih zgledov kažejo na primerljivo padanje napake aproksimacije s poldomensko Čebišev-Fourierovo vrsto glede na napako aproksimacije s Fourierovo vrsto ali z vrsto Čebiševa. Kljub temu pa je iz

primerov razvidno, da je za periodične probleme aproksimacija s Fourierovo vrsto še vedno boljša izbira.

Osrednji del doktorske disertacije pa predstavlja obravnava konstrukcije novega razreda kolokacijskih spektralnih metod za reševanje tako linearnih dvotočkovnih robnih problemov z Dirichletovimi robnimi pogoji v eni dimenziji (problem 1.3), kot tudi linearnih evolucijskih parcialnih diferencialnih enačb, kjer obravnavamo posplošene toplotne enačbe paraboličnega tipa (problem 1.4) in posplošene valovne enačbe hiperboličnega tipa (problem 1.5). Numerično rešitev danih problemov iščemo v obliki HCF vrste, kjer za izračun spektralnih koeficientov uporabimo metodo kolokacije. Bazne funkcije so še vedno trigonometrične sinusne in kosinusne funkcije ter sinusne in kosinusne funkcije polovičnih kotov, ki so organizirane v obliki ortogonalnih poldomenskih polinomov Čebiševa prve in druge vrste.

Podobno kot v primeru aproksimacije se izkaže, da lahko neperiodične probleme rešujemo z uporabo orodij za periodične probleme. Pri reševanju evolucijskih robnih problemov uporabimo spektralne metode samo za diskretizacijo po prostorski spremenljivki, da dobimo sistem navadnih diferencialnih enačb, v katerem kot neodvisna spremenljivka nastopa časovna spremenljivka t. Ta sistem rešimo z uporabo ene izmed metod za reševanje začetnih problemov, npr. s klasično Runge-Kutta metodo četrtega reda, ali z eno izmed metod geometrijske integracije, npr. z Magnusovo metodo četrtega reda.

Poleg konstrukcije razreda kolokacijskih Čebišev-Fourierovih (CFC) spektralnih metod za probleme 1.3 – 1.5 nas zanima tudi analiza konvergence in napake za obravnavane metode. Analizo konvergence smo napravili samo v primeru linearnih dvotočkovnih robnih problemov, ne pa tudi v primeru evolucijskih robnih problemov, kar predstavlja iztočnico za bodoče raziskovalno delo na področju Čebišev-Fourierovih spektralnih metod.

Teoretični izsledki iz analize konvergence za dvotočkovne robne probleme so podkrepljeni s številnimi numeričnimi zgledi, ki kažejo na primerljivo hitrost konvergence za standardni razred CC metod in konstruirani razred CFC metod. V večini prikazanih primerov pada napaka s CC metodo hitreje kot s CFC metodo, kar pa se spremeni v primerih, ko je rešitev robnega problema hitro oscilirajoča ali pa ni gladka. V primeru evolucijskih robnih problemov smo predstavili zgolj nekaj numeričnih zgledov.

Vsi numerični zgledi so bili narejeni z uporabo programskega paketa `Matlab` [43], vendar delujejo tudi v okolju `Octave` [16], nekateri izračuni pa zahtevajo uporabo programskega paketa `Chebfun` [59], ki je združljiv le s prvim od omenjenih orodij. Implementacija primerov je zgrajena na podlagi algoritmov in konstrukcijskih metod, ki so opisane v tem delu. Programska orodja vključujejo tako funkcijske m-datoteke kot tudi skripte. Oboje je bilo uporabljeno za potrebe te doktorske disertacije ter članka in poročila, ki sta podlaga tega dela. Večina slik, ki so vključene v doktorsko disertacijo, je prav tako izrisanih v `Matlabu`, razen nekaterih, ki so izrisane z uporabo

programskega paketa `Mathematica` [62]. Večina programskih kod je delo avtorja, nekatere pa so vzete iz paketov, ki sodijo k uporabljeni literaturi: J. Shen, T. Tang in L. Wang [51] ter L. N. Trefethen [57].

Glavna ugotovitev doktorske disertacije na podlagi konvergenčnih rezultatov in numeričnih zgledov je, da so novi razredi spektralnih metod, ki smo jih konstruirali na podlagi poldomenske Čebišev-Fourierove vrste, vsaj kar zadeva kvaliteto aproksimacije, primerljivi z obstoječimi klasičnimi razredi metod. V primerih, kjer kot koeficienti nastopajo gladke oz. analitične funkcije in je tudi rešitev problema gladka oz. analitična, dobimo spektralno konvergenco, tj. maksimalna absolutna napaka pada eksponentno glede na število členov. To velja tako za kolokacijsko metodo Čebiševa kot za kolokacijsko Čebišev-Fourierovo metodo, vendar je prva v takih primerih praviloma hitrejša. Ko koeficienti in rešitev niso gladke funkcije oz. so samo nekajkrat zvezno odvedljive, se prav tako izkaže, da sta hitrosti konvergence za CC in CFC metodo povsem primerljivi. V nekaterih primerih, npr. pri reševanju Airy-jeve diferencialne enačbe, ki ima hitro oscilirajočo točno rešitev, ter pri nekaterih enačbah, ki imajo le nekajkrat zvezno odvedljivo rešitev, pa se CFC metoda izkaže za uspešnejšo, saj napaka pada hitreje.

Računska zahtevnost razreda CFC metod pa je bistveno slabša, saj za izračun spektralnih koeficientov v poldomenski Čebišev-Fourierovi vrsti nimamo na voljo tako učinkovitega orodja, kot je hitra Fourierova transformacija za izračun spektralnih koeficientov Fourierove vrste ali vrste Čebiševa. Ta tema prav tako predstavlja izziv za bodoče raziskovalno delo.

Nekateri rezultati iz doktorske disertacije so bili objavljeni v članku B. Orel in A. Perne [46], z naslovom *Computations with half-range Chebyshev polynomials* (slo. *Računanje s poldomenskimi polinomi Čebiševa*) ter v poročilu B. Orel in A. Perne [45], z naslovom *Chebyshev-Fourier Spectral Methods for Non-Periodic Boundary Value Problems* (slo. *Čebišev-Fourierove spektralne metode za neperiodične robne probleme*). Rezultati v prvem članku obsegajo konstrukcijo in lastnosti poldomenskih polinomov Čebiševa ter računanje s poldomenskimi Čebišev-Fourierovimi vrstami (konstrukcija operatorskih matrik odvajanja in množenja), v drugem pa konstrukcijo in analizo spektralnih metod za linearne dvotočkovne robne probleme v eni dimenziji.

V doktorski disertaciji ostaja odprto marsikatero vprašanje, ki bo podlaga za nadaljnje raziskovalno delo na področju Čebišev-Fourierovih spektralnih metod. Eno izmed odprtih vprašanj je, ali se da konstruirati orodje za učinkovit izračun spektralnih koeficientov v HCF vrsti, ki bi zmanjšalo računsko zahtevnost tega razreda metod in bi bilo vsaj nekoliko primerljivo s FFT. Poleg tega bo nadaljnje delo osredotočeno na natančnejšo in bolj poglobljeno obravnavo in analizo metod za reševanje nekaterih tipov linearnih evolucijskih robnih problemov paraboličnega in hiperboličnega tipa. Prav tako ostaja odprto vprašanje, kako učinkovito konstruirati metode za stacionarne in evolucijske linearne robne probleme v dveh ali več dimenzijah.

Literatura

[1] M. Abramowitz in I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, National Bureau of Standards Applied Mathematics Series, vol. 55, U.S. Government Printing Office, Washington, D.C., 1964.

[2] B. Adcock, *Spectral Methods and modified Fourier series*, Technical report NA2007/08, DAMTP, University of Cambridge, 2007.

[3] ——, *Univariate modified Fourier methods for second order boundary value problems*, BIT, 49 (2009), str. 249–280.

[4] U. M. Ascher, R. M. M. Mattheij in R. D. Russell, *Numerical solution of boundary value problems for ordinary differential equations*, Prentice Hall Series in Computational Mathematics, Prentice Hall Inc., Englewood Cliffs, NJ, 1988.

[5] ——, *Numerical solution of boundary value problems for ordinary differential equations*, Classics in Applied Mathematics, vol. 13, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1995.

[6] K. Atkinson in W. Han, *Theoretical numerical analysis, A functional analysis framework*, Texts in Applied Mathematics, vol. 39, Springer, Dordrecht, tretja izdaja, 2009.

[7] J. P. Boyd, *Chebyshev and Fourier spectral methods*, Dover Publications Inc., Mineola, NY, druga izdaja, 2001.

[8] ——, *A comparison of numerical algorithms for Fourier extension of the first, second, and third kinds*, J. Comput. Phys., 178 (2002), str. 118–160.

[9] O. P. Bruno, Y. Han in M. M. Pohlman, *Accurate, high-order representation of complex three-dimensional surfaces via Fourier continuation analysis*, J. Comput. Phys., 227 (2007), str. 1094–1125.

[10] C. Canuto, M. Y. Hussaini, A. Quarteroni in T. A. Zang, *Spectral methods in fluid dynamics*, Springer Series in Computational Physics, Springer-Verlag, New York, 1988.

[11] ——, *Spectral methods, Fundamentals in single domains*, Scientific Computation, Springer-Verlag, Berlin, 2006.

[12] J. Céa, *Approximation variationnelle des problèmes aux limites*, Ann. Inst. Fourier (Grenoble), 14 (1964), str. 345–444.

[13] T. S. Chihara, *An introduction to orthogonal polynomials*, Mathematics and its Applications, vol. 13, Gordon and Breach Science Publishers, New York, 1978.

[14] C. W. Clenshaw in A. R. Curtis, *A method for numerical integration on an automatic computer*, Numer. Math., 2 (1960), str. 197–205.

[15] J. W. Cooley in J. W. Tukey, *An algorithm for the machine calculation of complex Fourier series*, Math. Comp., 19 (1965), str. 297–301.

[16] J. W. Eaton, D. Bateman in S. Hauberg, *GNU Octave manual*, Free Software Foundation, Inc., http://www.gnu.org/software/octave/, Boston, MA, tretja izdaja, 2011.

[17] B. Fischer in G. H. Golub, *How to generate unknown orthogonal polynomials out of known orthogonal polynomials*, J. Comput. Appl. Math., 43 (1992), str. 99–115.

[18] B. Fornberg, *A practical guide to pseudospectral methods*, Cambridge Monographs on Applied and Computational Mathematics, vol. 1, Cambridge University Press, Cambridge, 1996.

[19] B. Fornberg in D. M. Sloan, *A review of pseudospectral methods for solving partial differential equations*, Acta Numer., vol. 3, Cambridge Univ. Press, Cambridge, 1994, str. 203–267.

[20] W. Gautschi, *On generating orthogonal polynomials*, SIAM J. Sci. Statist. Comput., 3 (1982), str. 289–317.

[21] ——, *Orthogonal polynomials: applications and computation*, Acta Numer., vol. 5, Cambridge Univ. Press, Cambridge, 1996, str. 45–119.

[22] ——, *Numerical analysis*, Birkhäuser Boston Inc., Boston, MA, 1997.

[23] ——, *Orthogonal polynomials: computation and approximation*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2004.

[24] ——, *Orthogonal polynomials (in Matlab)*, J. Comput. Appl. Math., 178 (2005), str. 215–234.

[25] ——, *Orthogonal polynomials, quadrature, and approximation: computational methods and software (in Matlab)*, Lecture Notes in Math., vol. 1883, Springer, Berlin, 2006, str. 1–77.

[26] W. J. Gibbs, *Fourier's Series*, Nature, 59 (1898-1899), str. 200, 606.

[27] G. H. Golub in J. H. Welsch, *Calculation of Gauss quadrature rules*, Math. Comp., 23 (1969), str. 221–230.

[28] D. Gottlieb in S. A. Orszag, *Numerical analysis of spectral methods: theory and applications*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 26, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1977.

[29] D. Gottlieb in C.-W. Shu, *On the Gibbs phenomenon and its resolution*, SIAM Rev., 39 (1997), str. 644–668.

[30] P. Grandclément, *Introduction to spectral methods*, Technical report, 2006.

[31] P. Grandclément in J. Novak, *Spectral Methods for Numerical Relativity*, Living Rev. Relativity, 12 (2009).

[32] E. Hairer, C. Lubich in G. Wanner, *Geometric numerical integration*, Springer Series in Computational Mathematics, vol. 31, Springer, Heidelberg, druga izdaja, 2006.

[33] E. Hairer, S. P. Nørsett in G. Wanner, *Solving ordinary differential equations. I, Nonstiff problems*, Springer Series in Computational Mathematics, vol. 8, Springer-Verlag, Berlin, druga izdaja, 1993.

[34] P. Henrici, *Applied and computational complex analysis*, Pure and Applied Mathematics, vol. 3, John Wiley & Sons Inc., New York, 1986.

[35] D. Huybrechs, *On the Fourier extension of nonperiodic functions*, SIAM J. Numer. Anal., 47 (2010), str. 4326–4355.

[36] E. Isaacson in H. B. Keller, *Analysis of numerical methods*, John Wiley & Sons Inc., New York, 1966.

[37] A. Iserles, *A first course in the numerical analysis of differential equations*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, druga izdaja, 2009.

[38] A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett in A. Zanna, *Lie-group methods*, Acta Numer., vol. 9, Cambridge Univ. Press, Cambridge, 2000, str. 215–365.

[39] A. Iserles in S. P. Nørsett, *From high oscillation to rapid approximation. I. Modified Fourier expansions*, IMA J. Numer. Anal., 28 (2008), str. 862–887.

[40] A. B. J. Kuijlaars, K. T.-R. McLaughlin, W. Van Assche in M. Vanlessen, *The Riemann-Hilbert approach to strong asymptotics for orthogonal polynomials on* $[-1, 1]$, Adv. Math., 188 (2004), str. 337–398.

[41] P. D. Lax in A. N. Milgram, *Parabolic equations*, Annals of Mathematics Studies, vol. 33, Princeton University Press, Princeton, N. J., 1954, str. 167–190.

[42] G. Leoni, *A first course in Sobolev spaces*, Graduate Studies in Mathematics, vol. 105, American Mathematical Society, Providence, RI, 2009.

[43] MathWorks, *MATLAB Primer R2012b*, The MathWorks, Inc., Natick, MA, http://www.mathworks.com/, devetnajsta izdaja, 2012.

[44] B. Mercier, *An introduction to the numerical analysis of spectral methods*, Lecture Notes in Physics, vol. 318, Springer-Verlag, Berlin, 1989.

[45] B. Orel in A. Perne, *Chebyshev-Fourier Spectral Methods for Non-Periodic Boundary Value Problems*, poročilo, poslano v objavo, 2012.

[46] ——, *Computations with half-range Chebyshev polynomials*, J. Comput. Appl. Math., 236 (2012), str. 1753–1765.

[47] A. Perne, *Metode Liejevih grup in parcialne diferencialne enačbe*, Ljubljana, 2006.

[48] W. H. Press, B. P. Flannery, S. A. Teukolsky in W. T. Vetterling, *Numerical recipes in C, The art of scientific computing*, Cambridge University Press, Cambridge, 1988.

[49] C. Runge, *Über empirische Funktionen und die interpolation zwishen äquidistanten ordinaten*, Z. Math. Phys., 46 (1901), str. 224–243.

[50] R. A. Sack in A. F. Donovan, *An algorithm for Gaussian quadrature given modified moments*, Numer. Math., 18 (1971/72), str. 465–478.

[51] J. Shen, T. Tang in L.-L. Wang, *Spectral methods, Algorithms, analysis and applications*, Springer Series in Computational Mathematics, vol. 41, Springer, Heidelberg, 2011.

[52] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), str. 506–517.

[53] A. Y. Suhov, *A spectral method for the time evolution in parabolic problems*, J. Sci. Comput., 29 (2006), str. 201–217.

[54] G. Szegö, *Orthogonal polynomials*, American Mathematical Society Colloquium Publications, vol. 23, American Mathematical Society, Providence, R.I., 1959.

[55] E. Tadmor, *Filters, mollifiers and the computation of the Gibbs phenomenon*, Acta Numer., 16 (2007), str. 305–378.

[56] L. N. Trefethen, *Finite Difference and Spectral Methods for Ordinary and Partial Differential Equations*, dosegljivo na spletu, 1994.

[57] ——, *Spectral methods in MATLAB*, Software, Environments, and Tools, vol. 10, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[58] ——, *Is Gauss quadrature better than Clenshaw-Curtis?*, SIAM Rev., 50 (2008), str. 67–87.

[59] ——, *CHEBFUN Guide*, http://www2.maths.ox.ac.uk/chebfun/, Oxford, UK, prva popravljena izdaja, 2011.

[60] ——, *Approximation Theory and Approximation Practice*, SIAM, v pripravi, dosegljivo na spletu, 2012.

[61] J. C. Wheeler, *Modified moments and Gaussian quadratures*, 1974, str. 287–296.

[62] S. Wolfram, *The Mathematica Book*, Wolfram Media, Inc., Long Hanborough, Oxfordshire, UK, http://www.wolfram.com/, peta izdaja, 2003.

[63] W. P. Ziemer, *Weakly differentiable functions, Sobolev spaces and functions of bounded variation*, Graduate Texts in Mathematics, vol. 120, Springer-Verlag, New York, 1989.

Izjava o avtorstvu

Spodaj podpisani mag. Andrej Perne, univ. dipl. mat., izjavljam, da sem doktorsko disertacijo z naslovom *Konstrukcija spektralnih metod z neperiodičnimi trigonometričnimi vrstami* (ang. *Construction of spectral methods with non-periodic trigonometric series*) izdelal samostojno pod mentorstvom prof. dr. Bojana Orla.

S svojim podpisom Fakulteti za matematiko in fiziko Univerze v Ljubljani dovoljujem objavo elektronske oblike svojega dela na spletnih straneh.

Ljubljana, 12. 12. 2012

mag. Andrej Perne, univ. dipl. mat.

Zahvala

Doktorska disertacija je plod večletnega raziskovalnega dela na področju spektralnih metod. Da je bila napisana, gre zahvala predvsem mentorju prof. dr. Bojanu Orlu za potrpežljivost, koristne nasvete in dobronamerne spodbude. Zahvaljujem se tudi preostalima članoma komisije prof. dr. Boru Plestenjaku in doc. dr. Emilu Žagarju za konstruktivne in marsikdaj umestne pripombe. Prav tako se zahvaljujem vsem članom seminarja za Numerično analizo, ki so ob poslušanju delov te disertacije prispevali prenekatero pripombo ali nasvet.

Seveda tudi brez širše podpore ne bi šlo. Zato se zahvaljujem dekletu za spodbudo in pomoč, staršem in domačim za podporo ter prijateljem, ki so kakorkoli pripomogli, da je bilo to delo napisano.