1 Лежандрови полиноми

Лежандрови полиноми представљају систем комплетних и ортогоналних полинома који се користе у различитим областима физике, математике и инжењерства. Ови полиноми могу се дефинисати на разне начине, а њихов прорачун помоћу рекурентних формула биће нам посебно користан.

1.1 Дефиниција на основу конструкције ортогоналног система

Један од начина конструкције Лежандрових полинома је на основу услова да ови полиноми чине ортогоналан систем на интервалу $-1 \le x \le 1$. Ако са $P_n(x)$ означимо Лежандров полином реда n, онда услов ортогоналности захтева да важи

$$\int_{x=-1}^{1} P_m(x)P_n(x)\mathrm{d}x = 0, \text{ за } n \neq m. \tag{1.1}$$

Претходни услов одређује полиноме са тачношћу до мултипликативне константе, а ова неодређеност елиминише се дефинисањем вредности полинома у једној конкретној тачки. За Лежандрове полиноме уз услов ортогоналности захтева се и да важи

$$P_n(1) = 1, \tag{1.2}$$

а претходни услов назива се и стандардизацијом. На основу ова два услова Лежандрови полиноми су једнозначно одређени. У наредном пасусу илустрован је принцип како се на основу ова два услова могу одредити Лежандрови полиноми.

Примера ради, прво се може одредити полином нултог реда који, на основу (1.2), мора бити једнак $P_0(x) = 1$. Након тога, Лежандров полином првог реда, који је у општем случају облика $P_1(x) = a_1 x + a_0$, мора бити ортогоналан у односу на $P_0(x)$ и за њега мора да важи $P_1(1) = 1$, што резултује следећим системом једначина по непознатим коефицијентима a_1 и a_0

$$\begin{aligned} &\int_{x=-1}^{1} P_0(x)P_1(x)\mathrm{d}x = \int_{x=-1}^{1} 1\cdot(a_1 x + a_0)\mathrm{d}x = 2a_0 = 0, \\ &P_1(1) = a_1 + a_0 = 1, \end{aligned} \tag{1.3}$$

на основу чега добијамо $a_0 = 0$ и $a_1 = 1$, односно Лежандров полином првог реда постаје $P_1(x) = x$. Лежандров полином другог реда, $P_2(x)$, одређује се на основу услова да је ортогоналан у односу на полиноме $P_0(x)$ и $P_1(x)$ као и на основу услова $P_2(1) = 1$, и тако даље за Лежандрове полиноме виших редова. Полином $P_n(x)$ одређује се на основу услова да мора бити ортогоналан у односу на све полиноме $P_m(x)$, $0 \le m < n$, што резултује са n услова, а са додатним условом $P_n(1) = 1$ укупно имамо $n+1$ услов неопходан за одређивање $n+1$ коефицијента полинома реда n. Овакав начин дефиниције Лежандрових полинома један је од најједноставнијих и најинтуитивнијих, а при томе не захтева теорију диференцијалних једначина.

У најопштијем случају, ортогоналност функција може се увести и тако што се производ тих функција под интегралом, попут израза (1.1), множи (произвољном) тежинском функцијом која зависи од x. Приликом дефинисања услова ортогоналности Лежандрових полинома тежинска функција једнака је јединици, а Лежандрови полиноми чине један од три система класичних ортогоналних полинома. Другом систему ортогоналних полинома припадају Лагерови полиноми, који су ортогонални на полузатвореном интервалу $0 \le x < \infty$, а трећем систему ортогоналних полинома припадају Ермитови полиноми, који су ортогонални на интервалу $-\infty < x < \infty$. Код Лагерових и Ермитових полинома приликом дефиниције ортогоналности за тежинске функције користе се аналитичке функције које обезбеђују конвергенцију свих интеграла. У оквиру овог поглавља доминантно ћемо се бавити Лежандровим полиномима, тако да нећемо детаљније обрађивати Лагерове и Ермитове полиноме.

Други начин дефиниције Лежандрових полинома је на основу генеришуће функције, што превазилази оквире овог курса. Згодно је само напоменути да се на основу ове дефиниције лако може добити рекурентна формула за прорачун Лежандрових полинома, позната и под називом Бонетова рекурентна формула, која ће бити приказана касније. Дефиниција Лежандрових полинома помоћу генеришуће функције директно је повезана са развојем електромагнетских поља по мултиполима, што је

и била основна мотивација за увођење ових полинома 1782. година од стране познатог математичара Адријен-Мари Лежандра (1752-1833).

1.2 Дефиниција на основу диференцијалне једначине

Трећи начин дефиниције Лежандрових полинома је на основу Лежандрове диференцијалне једначине

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\left(1-x^2\right)\frac{\mathrm{d}P_n(x)}{\mathrm{d}x}\right]+n(n+1)P_n(x)=0\,, \tag{1.4}$$

чије решење су Лежандрови полиноми. Иако се нећемо упуштати у детаље у вези са овом диференцијалном једначином, Лежандрова диференцијална једначина појављује се кад год се Лапласова једначина решава методом раздвајања промењивих у сферном координатном систему. Лежандрови полиноми тада се појављују као $P_n(\cos\theta)$, при чему је θ зенитни угао сферног координатног система. Велики број физичких и инжењерских проблема описан је Лапласовом једначином, па се самим тим и Лежандрова диференцијална једначина често јавља у инжењерској пракси.

1.3 Ортогоналност

Стандардизација $P_n(1)=1$ одређује нормализацију Лежандрових полинома (у односу на стандардну L^2 норму на интервалу $-1\le x\le 1$). Пошто су Лежандрови полиноми и ортогонални у односу на исту норму, ова два својства могу се искомбиновати у једну једначину која гласи

$$\int_{x=-1}^{1} P_m(x)P_n(x)\mathrm{d}x=\frac{2}{2n+1}\delta_{mn}\,, \tag{1.5}$$

при чему је δ_{mn} Кронекерова делта функција. Ова нормализација најједноставније се одређује на основу Родригезове формуле, приказане у наставку.

1.4 Родригезова формула и остале експлицитне формуле

Посебно компактан израз за Лежандрове полиноме дат је помоћу Родригезове формуле

$$P_n(x)=\frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(x^2-1\right)^n\,. \tag{1.6}$$

Ова формула омогућава извођење великог броја других израза за полиноме P_n, од којих су неки:

$$P_n(x)=\frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}^2(x-1)^{n-k}(x+1)^k\,, \tag{1.7}$$

$$P_n(x)=\sum_{k=0}^{n}\binom{n}{k}\binom{n+k}{k}\left(\frac{x-1}{2}\right)^k\,, \tag{1.8}$$

$$P_n(x)=\frac{1}{2^n}\sum_{k=0}^{\lfloor n/2\rfloor}(-1)^k\binom{n}{k}\binom{2n-2k}{n}x^{n-2k}\,, \tag{1.9}$$

$$P_n(x)=2^n\sum_{k=0}^{n}x^k\binom{n}{k}\binom{\frac{n+k-1}{2}}{n}\,, \tag{1.10}$$

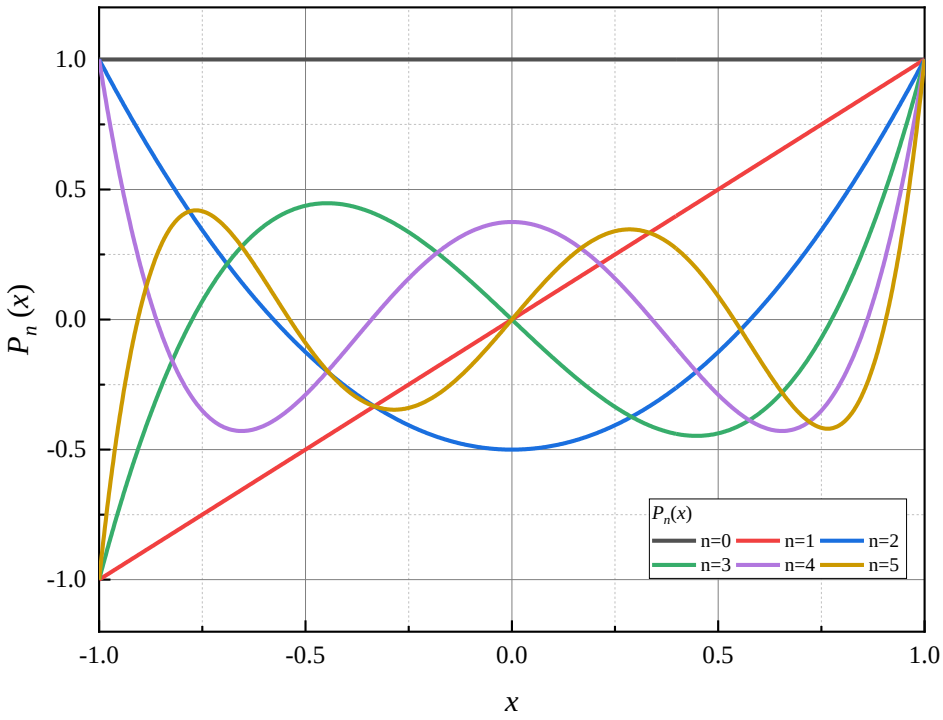
при чему последњи израз, који се може добити непосредно из рекурентних формула, изражава Лежандрове полиноме помоћу монома и уопштеног облика биномијалног коефицијента[1]. У изразу (1.9) $\lfloor n/2\rfloor$ представља највећи природни број c за који важи $c\le n/2$.

Првих једанаест Лежандрових полинома приказано је у (1.11), а првих шест Лежандрових полинома приказано је на слици 1.1.

[1] https://en.wikipedia.org/wiki/Binomial_coefficient

n	$P_n(x)$
0	1
1	x
2	$\frac{1}{2}\left(3x^2-1\right)$
3	$\frac{1}{2}\left(5x^3-3x\right)$
4	$\frac{1}{8}\left(35x^4-30x^2+3\right)$
5	$\frac{1}{8}\left(63x^5-70x^3+15x\right)$
6	$\frac{1}{16}\left(231x^6-315x^4+105x^2-5\right)$
7	$\frac{1}{16}\left(429x^7-693x^5+315x^3-35x\right)$
8	$\frac{1}{128}\left(6435x^8-12012x^6+6930x^4-1260x^2+35\right)$
9	$\frac{1}{128}\left(12155x^9-25740x^7+18018x^5-4620x^3+315x\right)$
10	$\frac{1}{256}\left(46189x^{10}-109395x^8+90090x^6-30030x^4+3465x^2-63\right)$

(1.11)



Слика 1.1. Првих шест Лежандрових полинома.

1.5 Нека својства Лежандрових полинома

Лежандрови полиноми имају дефинисану парност, односно они су или парни или непарни и важи релација

$$P_n(-x)=(-1)^n P_n(x)\,. \tag{1.12}$$

Друго корисно својство је да је средња вредност Лежандрових полинома реда већег од нултог једнака нули, односно да важи

$$\int_{x=-1}^{1} P_n(x)\mathrm{d}x = 0 \text{ за } n \geq 1, \tag{1.13}$$

што се директно може добити на основу ортогоналности полинома $P_n(x)$, $n \geq 1$ у односу на $P_0(x) = 1$. Ово својство је корисно када се користи Лежандров развој функције $f(x)$, $f(x) \approx \tilde{f}(x) = \sum_i a_i P_i$ за апроксимацију произвољне функције или резултата мерења. Тада је средња вредност функције $\tilde{f}(x)$ над интервалом $-1 \leq x \leq 1$ једнака коефицијенту a_0 Лежандровог развоја.

Пошто су диференцијалне једначине и ортогоналност функција независне од мултипликативног коефицијента којим се множи одређена функција, дефиниције Лежандрових полинома су „стандардизоване“ тако што су скалиране да важи

$$P_n(1) = 1. \tag{1.14}$$

Извод Лежандрових полинома у крајњој тачки $x = 1$ је

$$\frac{\mathrm{d}P_n(x=1)}{\mathrm{d}x} = \frac{n(n+1)}{2}. \tag{1.15}$$

За Лежандрове полиноме важи и Аскеј-Гасперова неједнакост која гласи

$$\sum_{j=0}^{n} P_j(x) \geq 0 \text{ за } x \geq -1. \tag{1.16}$$

Рекурентне релације

Као што је напоменуто раније, Лежандрови полиноми задовољавају трочлане рекурентне релације познате под називом Бонетове рекурентне формуле које гласе

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x) \text{ и} \tag{1.17}$$

$$\frac{x^2-1}{n}\frac{\mathrm{d}P_n(x)}{\mathrm{d}x} = xP_n(x) - P_{n-1}(x). \tag{1.18}$$

Приметити да израз (1.18) постаје неупотребљив за прорачун извода Лежандрових полинома на границама посматраног интервала ($|x| = 1$), а уместо њега може се користити алтернативни израз за рекурентан прорачун извода Лежандрових полинома који гласи

$$\frac{\mathrm{d}P_{n+1}(x)}{\mathrm{d}x} = (n+1)P_n(x) + x\frac{\mathrm{d}P_n(x)}{\mathrm{d}x}. \tag{1.19}$$

Релација погодна за интеграцију Лежандрових полинома је

$$(2n+1)P_n(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left(P_{n+1}(x) - P_{n-1}(x)\right). \tag{1.20}$$

На основу релација приказаних до сада може се доказати да важи

$$\frac{\mathrm{d}P_{n+1}(x)}{\mathrm{d}x} = (2n+1)P_n(x) + (2(n-2)+1)P_{n-2}(x) + (2(n-4)+1)P_{n-4}(x) + \ldots, \tag{1.21}$$

односно еквивалентно

$$\frac{\mathrm{d}P_{n+1}(x)}{\mathrm{d}x} = \frac{2P_n(x)}{\|P_n\|^2} + \frac{2P_{n-2}(x)}{\|P_{n-2}\|^2} + \ldots \tag{1.22}$$

при чему је $\|P_n\|$ L^2 норма Лежандровог полинома реда n дефинисана над интервалом $-1 \leq x \leq 1$ као

$$\|P_n\| = \sqrt{\int_{x=-1}^{1} (P_n(x))^2 \mathrm{d}x} = \sqrt{\frac{2}{2n+1}}. \tag{1.23}$$

Нуле

Свих n нула полинома $P_n(x)$ су реалне, међусобно различите, и леже у интервалу $-1 < x < 1$. Додатно, ако их посматрамо као границе које деле интервал $-1 \le x \le 1$ на $n+1$ подинтервал, сваки подинтервал садржаће тачно једну нулу полинома $P_{n+1}(x)$. Ово својство је познато и под називом својство преплитања. На основу својства парности Лежандрових полинома очигледно је да ако је x_k нула полинома $P_n(x)$, онда је и $-x_k$ нула тог полинома. Ове нуле играју важну улогу у нумеричкој интеграцији Гаусовим квадратурним формулама. Посебна квадратурна формула заснована на Лежандровим полиномима позната је под називом Гаус-Лежандрова квадратурна формула.

На основу ових својстава и чињенице да је $P_n(\pm 1) \neq 0$, следи да $P_n(x)$ има $n-1$ локалних екстремума (минимума или максимума) на интервалу $-1 < x < 1$. Еквивалентно, $\mathrm{d}P_n(x)/\mathrm{d}x$ има $n-1$ нулу на интервалу $-1 < x < 1$.

Вредности у карактеристичним тачкама

Парност и стандардизација имплицирају вредности Лежандрових полинома на границама $x = \pm 1$ које износе

$$P_n(1) = 1,\ P_n(-1) = \begin{cases} 1, \text{ за } n = 2m \\ -1, \text{ за } n = 2m+1 \end{cases},\ m \in N_0 . \tag{1.24}$$

У координатном почетку вредности Лежандрових полинома су

$$P_n(0) = \begin{cases} \dfrac{(-1)^m}{4^m}\binom{2m}{m} = \dfrac{(-1)^m}{2^{2m}}\dfrac{(2m)!}{(m!)^2}, & \text{за } n = 2m \\ 0, & \text{за } n = 2m+1 \end{cases},\ m \in N_0 . \tag{1.25}$$

2 Кондициони број

У оквиру нумеричке анализе, кондициони број функције представља меру колико се вредност функције може променити приликом мале промене улазних параметара. Овако уведен кондициони број се користи као мера колико је функција осетљива на промене или грешке улазних параметара, односно колика ће бити грешка вредности функције узрокована грешком у улазним параметрима функције. Са друге стране, у пракси се врло често решава инверзан проблем, односно потребно је одредити аргумент функције за познату вредност функције. Инверзан проблем се (математички) може описати решавањем једначине $f(x) = y$ по x за задато y. У оваквим проблемима кондициони број дефинише се за инверзну функцију $f^{-1}(y) = x$, односно представља меру промене x за мале промене y.

Као што ће бити приказано касније, кондициони број (строго) формално се дефинише као количник (асимптотских) релативних промена излазних параметара и релативних промена улазних параметара у најгорем случају, односно у случају када је овај количник (кондициони број) највећи. У одређеним случајевима користе се и (парцијални) изводи. Када нам функција за коју одређујемо кондициони број није позната у аналитичком облику, улогу аналитичке „функције“ преузимају решења проблема, а улогу „аргумената“ преузимају улазни подаци којима је проблем описан. Пример овакве ситуације је мерење података у (коначном) дискретном броју тачака.

Кондициони број често се примењује и у области линеарне алгебре, а тада су основне дефиниције и процедуре извођења израза за кондициони број релативно једноставне. Кондициони број у оквиру линеарне алгебре, односно кондициони број матрице од посебног нам је значаја, па ће детаљи у вези са овиме бити приказани у оквиру овог поглавља. Видећемо да грешке линеарног проблема могу бити усмерене у различитим правцима у посматраном линеарном простору, а због тога се кондициони број прорачунава у најгорем случају (када је највећи) користећи се теоријом линеарне алгебре. Финално, кондициони број може се дефинисати и за нелинеарне проблеме описане функцијама више променљивих, а такви случајеви нису нам од интереса у овом тренутку па их због тога ни нећемо детаљније разматрати.

За проблем који има мали кондициони број кажемо да је добро условљен (*well-conditioned*), док за проблем са великим кондиционим бројем кажемо да је лоше условљен (*ill-conditioned*). Неформално говорећи, код лоше условљених проблема мале промене улазних параметара (мале промене независних промењивих у случају директних проблема, односно мале промене вредности са „десне стране“ једначине) резултују великим променама излазних параметара (вредности функције у случају директних

проблема, односно независном аргументу у случају инверзних проблема). Ово значи да је у случају лоше условљених проблема (проблема са великим кондиционим бројем) релативно тешко одредити решење са високом тачношћу.

Кондициони број је својство конкретног (посматраног) проблема. Заједно са посматраним проблемом, односно његовим кондиционим бројем, потребно је посматрати и алгоритам који ће се користити за решавање тог проблема. Неки алгоритми имају својство названо стабилност уназад (*backward stability*[2]). Неформално посматрано за алгоритме који су стабилни уназад може се очекивати да ће (релативно) тачно решити добро-условљене проблеме.

2.1 Општа дефиниција у контексту анализе грешке функције

Посматрајмо проблем описан функцијом $f(x)$ чији је улазни параметар x. Када се улазни параметар промени за (малу) вредност δx, вредност функције постаје $f(x+\delta x)$. Апсолутна вредност промене функције је $\|\delta f(x)\| = \|f(x+\delta x) - f(x)\|$, а апсолутна вредност релативне промене функције је $\|f(x+\delta x) - f(x)\| / \|f(x)\| = \|\delta f(x)\| / \|f(x)\|$. На основу овога апсолутна вредност кондиционог броја проблема описаног функцијом $f(x)$ је

$$\lim_{\varepsilon \to 0} \sup_{\|\delta x\| \le \varepsilon} \frac{\|\delta f(x)\|}{\|\delta x\|}, \tag{2.1}$$

при чему је релативни кондициони број

$$\lim_{\varepsilon \to 0} \sup_{\|\delta x\| \le \varepsilon} \frac{\|\delta f(x)\| / \|f(x)\|}{\|\delta x\| / \|x\|}. \tag{2.2}$$

У претходним изразима „sup“ представља ознаку супремума који се дефинише на следећи начин. Горња граница подскупа S делимично уређеног скупа $(P, \le)$ је елемент b скупа P за који важи

$$x \le b, \ \forall x \in S. \tag{2.3}$$

Горња граница b подскупа S назива се супремум подскупа S (или најмања горња граница подскупа S) ако за све горње границе z подскупа S из скупа P важи

$$b \le z \ (b \text{ није веће од било које друге горње границе подскупа } S). \tag{2.4}$$

2.2 Кондициони број у оквиру линеарне алгебре

Кондициони број може се придружити и матрици (A), а кондициони број игра важну улогу приликом решавања система линеарних једначина који се може описати матричном једначином $Ax = b$. У овом случају кондициони број даје процену колика ће бити грешка решења посматраног система једначина (x) уколико постоји мала грешка у улазним подацима (било за вектор слободних коефицијената b, било за матрицу A). Напоменимо да је овај узрок грешке у потпуности независан од алгоритма за решавање система једначина и да представља искључиво својство матрице A, односно овај узрок грешке постојао би и када бисмо имали „савршен“ алгоритам за решавање система линеарних једначина, који не би (додатно) уносио грешке услед заокруживања бројева односно услед коначне тачности представљања бројева у аритметици са помичном тачком и услед коначне тачности спровођења математичких операција.

Грубо говорећи, кондициони број може да се посматра као количник промене решења x и промене слободног вектора b, при чему се промена вектора прорачунава према одређеној норми. Према томе, ако је кондициони број велики, чак и мала грешка у улазним подацима за вектор b може резултовати великом грешком у решењу x. Са друге стране, ако је кондициони број мали, тада грешка решења x неће бити много већа од грешке улазних података за слободни вектор b.

Прецизније говорећи, кондициони број се дефинише као максимални количник релативне грешке x у односу на релативну грешку улазних података за слободни вектор b. Нека је δb вектор грешке улазних података вектора b. Под претпоставком да је A несингуларна матрица, грешка решења

[2] https://en.wikipedia.org/wiki/Numerical_stability

проблема $x = A^{-1}b$ је $\delta x = A^{-1}\delta b$. Однос релативних грешака решења проблема и релативне грешке улазних података вектора b је

$$\frac{\|\delta x\|/\|x\|}{\|\delta b\|/\|b\|} = \frac{\left\|A^{-1}\delta b\right\|/\left\|A^{-1}b\right\|}{\|\delta b\|/\|b\|} = \frac{\left\|A^{-1}\delta b\right\|}{\|\delta b\|}\frac{\|b\|}{\left\|A^{-1}b\right\|}, \tag{2.5}$$

при чему $\|\cdot\|$ представља оператор норме (вектора и матрице), као што ће бити описано у оквиру овог и наредног поглавља.

Максимална вредност претходног количника (за ненулто b и ненулто δb) представља кондициони број $c(A)$ и може се одредити на основу производа две норме оператора као

$$c(A) = \max_{\delta b, b\neq 0}\left\{\frac{\left\|A^{-1}\delta b\right\|}{\|\delta b\|}\frac{\|b\|}{\left\|A^{-1}b\right\|}\right\} = \max_{\delta b\neq 0}\left\{\frac{\left\|A^{-1}\delta b\right\|}{\|\delta b\|}\right\}\max_{b\neq 0}\left\{\frac{\|b\|}{\left\|A^{-1}b\right\|}\right\} = \max_{\delta b\neq 0}\left\{\frac{\left\|A^{-1}\delta b\right\|}{\|\delta b\|}\right\}\max_{x\neq 0}\left\{\frac{\|Ax\|}{\|x\|}\right\} = \left\|A^{-1}\right\|\|A\|. \tag{2.6}$$

Иста дефиниција важи и за сваку конзистентну норму, односно за ону за коју важи

$$c(A) = \left\|A^{-1}\right\|\|A\| \geq \left\|A^{-1}A\right\| = 1. \tag{2.7}$$

Додатни детаљи у вези са кондиционим бројем матрице биће приказани у наредном потпоглављу.

Када је кондициони број једнак јединици (што се може десити само када је матрица A облика скаларног производа линеарне изометрије[3], а јединична матрица је један посебан случај линеарне изометрије), генерално говорећи апроксимација решења није мање тачна од тачности улазних података, под условом да алгоритам решавања проблема не уноси додатне грешке.

Кондициони број може бити и бесконачан, а ово имплицира да је проблем неодређен. Тада не постоји једнозначно дефинисано решење за дефинисану побуду (вектор b), односно матрица A је сингуларна, а у том случају без обзира на избор алгоритма за решавање проблема проблем се не може решити.

Дефиниција кондиционог броја зависи од избора норме, као што се може илустровати кроз наредна два примера.

Ако $\|\cdot\|$ представља стандардну L^2 норму онда је кондициони број матрице A једнак

$$c(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}, \tag{2.8}$$

при чему су $\sigma_{\max}(A)$ и $\sigma_{\min}(A)$ максимална и минимална сингуларна вредност[4] матрице A, респективно. Ако је матрица A нормална[5], тада је кондициони број једнак

$$c(A) = \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}, \tag{2.9}$$

при чему су $\lambda_{\max}(A)$ и $\lambda_{\min}(A)$ максимална и минимална (по модулу) сопствена вредност матрице A. Ако је матрица A јединична, њен кондициони број је $c(A) = 1$, што представља минималну могућу вредност кондиционог броја.

Кондициони број у односу на L^2 норму често се појављује у оквиру нумеричке линеарне алгебре, па му је због тога дат и посебан назив, кондициони број матрице. Детаљи у вези са кондиционим бројем матрице приказани су у следећем поглављу.

Ако $\|\cdot\|$ представља L^∞ норму[6] и матрица A је доње-троугаона несингуларна матрица (то јест, важи $a_{ii} \neq 0$, $\forall i$) онда за кондициони број важи

---

[3] https://en.wikipedia.org/wiki/Isometry#Linear_isometry

[4] https://en.wikipedia.org/wiki/Singular_value_decomposition

[5] https://en.wikipedia.org/wiki/Normal_matrix

[6] https://en.wikipedia.org/wiki/L-infinity

$$c(A) \geq \frac{\max_i \left(|a_{ii}| \right)}{\min_i \left(|a_{ii}| \right)}. \tag{2.10}$$

Кондициони број прорачунат помоћу ове норме генерално је већи од кондиционог броја прорачунатог помоћу L^2 норме. Овај кондициони број генерално може се прорачунати доста лакше (ефикасније) у односу на кондициони број матрице (у односу на L^2 норму), па је то и главни разлог зашто се овакав кондициони број најчешће користи као процена кондиционог броја матрице.

Ако кондициони број није претерано велики у односу на јединицу, матрица је добро условљена, што значи да се инверзна матрица ове матрице може прорачунати релативно тачно. Ако је кондициони број велик, за матрицу се каже да је лоше условљена. Ово у пракси значи да је матрица врло блиска сингуларној матрици, а прорачун њене инверзне матрице односно решавање система линеарних једначина описаног овом матрицом може резултовати великим нумеричким грешкама. Сингуларна матрица има кондициони број који тежи бесконачности.

Као (инжењерска) процена, ако је за проблем $Ax = b$ кондициони број $c(A) = 10^k$, може се очекивати да ће се услед лоше условљености матрице изгубити до k цифара тачности додатно у односу на тачност која ће се изгуби нумерички (због коначне тачности представљања бројева и коначне тачности спровођења математичких операција). Са друге стране, кондициони број не даје максималну вредност грешке услед лоше условљености матрице, већ само даје процену границе те грешке, а стварна вредност грешке зависиће од улазних параметара.

2.3 Кондициони број матрице

У оквиру овог поглавља покушаћемо да измеримо „осетљивост“ проблема описаног једначином

$$Ax = b, \tag{2.11}$$

односно желимо да одговоримо на следеће питање: Ако се A и/или b мало промене, колико ће се променити $x = A^{-1}b$? Пре него кренемо да се бавимо овим питањем, потребан нам је начин да „измеримо“ A и промену ΔA. За норму (дужину) вектора користићемо стандардну (Еуклидску) L^2 норму, а у наставку увешћемо и норму матрице. Видећемо да ће се кондициони број, односно „осетљивост“ матрице A, добити из производа норми матрица A и A^{-1}, као што је и наведено у једначини (2.7). Све матрице које ћемо разматрати у оквиру овог поглавља су квадратне.

2.3.1 Норма матрице и кондициони број матрице

Питање које можемо поставити себи је на који начин се може измерити природна отпорност проблема на грешке (заокруживања) и на који начин се може одлучити да ли је матрица добро условљена или лоше условљена? Ако постоји мала промена b или мала промена A, колика ће бити промена решења x?

Почећемо дискусију од случаја када се у једначини (2.11) вектор b промени тако да његова нова вредност постане $b + \delta b$, при чему δb моделује грешку која постоји у улазним подацима за вектор b. Сматрамо да се приликом тога матрица A не мења, а грешка δb може бити последица грешке мерења експерименталних резултата или последица заокруживања бројева. Претпоставићемо да је δb мало, а сматраћемо да нам правац грешке δb није познат (што одговара практично свим ситуацијама у пракси). Услед грешке δb решење ће се променити са x на $x + \delta x$, што се, полазећи од једначине (2.11) може описати једначином грешке

$$A(x + \delta x) = Ax + A(\delta x) = b + \delta b. \tag{2.12}$$

Одузимањем једначине (2.11) од једначине (2.12) добијамо

$$A(\delta x) = \delta b. \tag{2.13}$$

На основу претходне једначине закључујемо да грешка δb води ка грешци решења $\delta x = A^{-1}(\delta b)$. Према томе постојаће велика промена у решењу x када је A^{-1} „велико“, односно када је A блиска сингуларној матрици. Промена решења x посебно је велика када је δb у правцу који се највише „појачава“ матрицом A^{-1}.

Претпоставимо да је A симетрична матрица[7] и да су јој сопствене вредности позитивне, односно да важи $0 \le \lambda_1 \le \lambda_2 \le ... \le \lambda_n$, а да смо сопствене векторе $x_1, x_2, ..., x_n$ изабрали тако да им је дужина (норма) јединична. Одговарајућа једначина проблема сопствених вредности коју задовољавају сопствене вредности и сопствени вектори је

$$Ax_i = \lambda_i x_i,\ i = 1, 2, ..., n. \tag{2.14}$$

Било који вектор, укључујући и вектор грешке δb, може се представити као линеарна комбинација сопствених вектора[8] x_i, $i = 1, 2, ..., n$. Претпоставимо да је вектор грешке δb облика $\delta b = \varepsilon x_i$, $i = 1, 2, ..., n$, односно да је δb пропорционалан сопственом вектору x_i, при чему је ε произвољан скалар (коефицијент пропорционалности). Тада је, на основу (2.13) и (2.14)

$$\delta x = \frac{\delta b}{\lambda_i} \text{ када је } \delta b = \varepsilon x_i,\ i = 1, 2, ..., n. \tag{2.15}$$

На основу претходног израза очигледно је да је за задату дужину вектора δb, грешка решења δx највећа када је λ_i најмање, односно када је $\lambda_i = \lambda_1$. Према томе највећа грешка решења која потиче од матрице A^{-1} добија се у случају када је δb у правцу првог сопственог вектора x_1, односно за највећу грешку решења δx важи

$$\delta x = \frac{\delta b}{\lambda_1} \text{ када је } \delta b = \varepsilon x_1. \tag{2.16}$$

Ако бисмо нашли L^2 норму претходног израза, добили бисмо да се грешка решења $\|\delta x\|$ добија множењем $\|\delta b\|$ са $1/\lambda_1$ (λ_1 представља најмању сопствену вредност матрице A, а $1/\lambda_1$ представља највећу сопствену вредност матрице[9] A^{-1}). Ово „појачавање“ грешке веће је што је λ_1 ближе нули, односно што је матрица A ближа сингуларној матрици.

Мерење осетљивости искључиво помоћу λ_1, односно само на основу израза (2.16) има озбиљан недостатак. Претпоставимо да смо сваки члан матрице A помножили са 1000. То би одговарало множењу λ_1 са 1000 и матрица би према овоме критеријуму изгледала мање сингуларна. Ово је у супротности са тиме да једноставно скалирање матрице може да од лоше условљене матрице направи добро условљену матрицу. Истина је да би у овоме случају δx било 1000 пута мање, али 1000 пута мање било би и решење $x = A^{-1}b$. Оно што би остало непромењено је релативна грешка $\|\delta x\|/\|x\|$. Према томе дељењем грешке $\|\delta x\|$ са $\|x\|$ нормализујемо проблем и спречавамо да тривијално скалирање проблема утиче на осетљивост. Са друге стране потребно је увести и нормализацију за δb, односно проблем сводимо на поређење релативне промене $\|\delta b\|/\|b\|$ и релативне грешке $\|\delta x\|/\|x\|$, односно посматрамо количник $\dfrac{\|\delta x\|/\|x\|}{\|\delta b\|/\|b\|}$.

Претходни израз може се написати и у следећем облику

$$\frac{\|\delta x\|/\|x\|}{\|\delta b\|/\|b\|} = \frac{\|\delta x\|}{\|\delta b\|}\frac{\|b\|}{\|x\|} \tag{2.17}$$

Што је систем осетљивији овај количник је већи, а ми желимо да нађемо горњу границу за овај количник, зато ћемо посматрати најгори случај (када овај количник има максимум). На основу (2.16) први разломак са десне стране једначине (2.17) максималан је када је $\delta b = \varepsilon x_1$ и то делом одређује најгори случај. Други део који одређује најгори случај је други разломак са десне стране једначине (2.17) за који желимо да утврдимо када ће бити максималан. Тај разломак ће бити највећи када је тачно

---

[7] Симетричне матрице имају строго реалне сопствене вредности, а сопствени вектори су међусобно ортогонални.

[8] Зато што су сопствени вектори управни и „спанују“ односно прожимају читав простор димензије матрице A.

[9] Ако је λ сопствена вредност квадратне матрице A онда је $1/\lambda$ сопствена вредност матрице A^{-1}.

решење x што је могуће мање у односу на тачно b. Према томе да би други разломак са десне стране једначине био максималан, оригинални проблем $Ax = b$ треба да буде у другој екстремној ситуацији у претходни случај, односно вектор b треба да буде у правцу последњег сопственог вектора (који има највећу сопствену вредност), односно треба да важи да је $b = \mu x_n$, при чему је μ произвољан скалар (различит од нуле). Тада је $x = A^{-1}b = b/\lambda_n$ минимално, па је други разломак са десне стране једначине максималан и износи $\|b\|/\|x\| = \lambda_n$.

Према томе комбинација $b = \mu x_n$ и $\delta b = \varepsilon x_1$ чине релативну грешку $\dfrac{\|\delta x\|/\|x\|}{\|\delta b\|/\|b\|}$ највећом могућом. Ово је екстреман случај описан следећом неједначином. За позитивне дефинитне матрице решење $x = A^{-1}b$ и грешка решења $\delta x = A^{-1}\delta b$ увек задовољавају

$$\|x\| \geq \frac{\|b\|}{\lambda_{\max}}, \quad \|\delta x\| \leq \frac{\|\delta b\|}{\lambda_{\min}} \quad \text{и} \quad \frac{\|\delta x\|/\|x\|}{\|\delta b\|/\|b\|} \leq \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{2.18}$$

Количник $c = \lambda_{\max}/\lambda_{\min}$ представља кондициони број позитивне дефинитне матрице A.

Несиметричне матрице

Наша досадашња анализа односи се на симетричне матрице са позитивним сопственим вредностима. Услов да су сопствене вредности позитивне није неопходан, пошто се у претходној анализи λ може заменити са $\|\lambda\|$. Са друге стране, како бисмо могли да разматрамо и несиметричне матрице потребно је извршити веће измене.

За одговарајућу дефиницију кондиционог броја, вратимо се на једначину (2.17). Како бисмо нашли горњу границу релативне грешке $\dfrac{\|\delta x\|/\|x\|}{\|\delta b\|/\|b\|}$ посматрали смо најгори случај када смо узели да је x што је могуће мање, а $b = Ax$ што је могуће веће. Када је матрица A несиметрична, максимум количника $\|b\|/\|x\| = \|Ax\|/\|x\|$ може настати за вектор x који се разликује од сопствених вектора. Овај максимум представља одличну меру „величине“ матрице A, и назива се нормом матрице A.

Норма матрице A је број $\|A\|$ који се дефинише као

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}. \tag{2.19}$$

Другим речима, $\|A\|$ представља горњу границу „појачавачке моћи“ матрице, пошто, на основу дефиниције (2.19) за произвољан вектор x важи

$$\|Ax\| \leq \|A\|\|x\|, \ \forall x. \tag{2.20}$$

Због тога што важи $b = Ax$ и $\delta x = A^{-1}\delta b$, на основу једначине (2.20) добијамо

$$\|b\| \leq \|A\|\|x\| \ \text{и} \ \|\delta x\| \leq \|A^{-1}\|\|\delta b\|. \tag{2.21}$$

Претходна неједначина представља замену за неједначину (2.18) када је A несиметрична матрица. У случају када је A симетрична матрица тада је $\|A\| = \lambda_{\max}$ и $\|A^{-1}\| = 1/\lambda_{\min}$. Према томе адекватна замена за количник $\lambda_{\max}/\lambda_{\min}$ у изразу (2.18) је производ $\|A\|\|A^{-1}\|$ што представља кондициони број у општем случају (несиметричне матрице).

Кондициони број матрице A је $c = \|A\|\|A^{-1}\|$. Однос релативних грешака δx и δb задовољава

$$\frac{\|\delta x\|/\|x\|}{\|\delta b\|/\|b\|} \leq c, \tag{2.22}$$

што се добија директно из неједначина (2.21).

У свим разматрањима до сада сматрали смо да грешка потиче искључиво од грешке улазних података вектора b, односно од вектора грешке δb при чему смо сматрали да не постоје грешке у

улазним подацима матрице A. Посматрајмо сада други случај. Претпоставимо да је промена решења δx узрокована променом (елемената) матрице δA, сматрајући да се b не мења. Показаћемо да тада за релативну грешку важи релација

$$\frac{\|\delta x\|/\|x+\delta x\|}{\|\delta A\|/\|A\|} \le c\,, \tag{2.23}$$

Оно што је посебно занимљиво је да исти кондициони број фигурише и у једначини (2.23) када је промена решења δx условљена само променом матрице δA и у једначини (2.22) када је промена δx условљена само променом δb.

Докажимо неједначину (2.23). На основу једначине $Ax = b$ директно добијамо $(A+\delta A)(x+\delta x) = b$, а након одузимања прве једначине од друге добијамо

$$A\delta x + \delta A(x+\delta x) = 0 \tag{2.24}$$

односно

$$\delta x = -A^{-1}\delta A(x+\delta x). \tag{2.25}$$

На основу дефиниције норме матрице (2.19), множење вектора $(x+\delta x)$ са δA не повећава дужину вектора $(x+\delta x)$ за више од $\|\delta A\|$, а множење новодобијеног вектора са A^{-1} не повећава његову дужину за више од $\|A^{-1}\|$. Према томе $\|\delta x\| \le \|A^{-1}\|\|\delta A\|\|x+\delta x\|$, на основу чега добијамо

$$\frac{\|\delta x\|}{\|x+\delta x\|} \le \|A^{-1}\|\|\delta A\| = c\frac{\|\delta A\|}{\|A\|}. \tag{2.26}$$

На основу неједнакости приказаних до сада, за грешку приликом решавања система једначина можемо закључити да потиче из два извора. Први извор грешке је „природна осетљивост“ проблема на нетачност улазних података, мерена кроз кондициони број c. Други извор је сама грешка δb односно δA. Овакав начин описивања грешке је представљао основу за Вилкинсонову анализу грешака[10]. Пошто решавање система једначина елиминацијом производи матрице L' и U' које у себи садрже одређене грешке (заокруживања), приликом решавања система једначина полази се од матрице $A+\delta A = L'U'$ која у себи садржи грешку, а не полази се од оригиналне (тачне) матрице $A = LU$. У својој анализи Вилкинсон је показао да делимична пивотизација контролише грешку δA, тако да финалној грешци доприноси кондициони број c.

Прорачун норме матрице

Норма матрице A одређује највећу вредност којом се било који вектор (сопствени вектор или било који други вектор) „појачава“ множењем матрицом A, што је очигледно на основу дефиниције норме матрице (2.19). Норма јединичне матрице једнака је 1. Како бисмо израчунали норму матрице, квадрирајмо обе стране израза (2.19) како бисмо дошли до симетричне матрице A^TA

$$\|A\|^2 = \max_{x\neq 0}\frac{\|Ax\|^2}{\|x\|^2} = \max_{x\neq 0}\frac{x^{\mathrm{T}}A^{\mathrm{T}}Ax}{x^{\mathrm{T}}x}. \tag{2.27}$$

На основу овога закључујемо да је $\|A\|$ једнако квадратном корену највеће сопствене вредности матрице $A^{\mathrm{T}}A$, односно да важи $\|A\|^2 = \lambda_{\max}(A^{\mathrm{T}}A)$. Вектор који матрица A „појачава“ највише је сопствени вектор матрице $A^{\mathrm{T}}A$ коме припада највећа сопствена вредност $\lambda_{\max}$, односно важи

$$\frac{x^{\mathrm{T}}A^{\mathrm{T}}Ax}{x^{\mathrm{T}}x} = \frac{x^{\mathrm{T}}(\lambda_{\max}x)}{x^{\mathrm{T}}x} = \lambda_{\max}(A^{\mathrm{T}}A) = \|A\|^2. \tag{2.28}$$

У вези са формулом (2.28) важно је напоменути неколико ствари:

- Норма и кондициони број у пракси не прорачунавају се на овај начин. Уместо тога користе се другачији поступци за њихову процену. Разлог је то што је прорачун

[10] https://en.wikipedia.org/wiki/Wilkinson%27s_polynomial

сопствених вредности матрице $A^{\mathrm{T}}A$ (како би се одредило $\lambda_{\max}$, а касније и $\|A\|^2$) нумерички изузетно захтеван.

- У једначини најмањих квадрата, $A^{\mathrm{T}}Ax = A^Tb$, кондициони број $c\left(A^{\mathrm{T}}A\right)$ једнак је квадрату кондиционог броја $c(A)$. Формирање $A^{\mathrm{T}}A$ јако квари условљеност проблема, па се обично пре прорачуна са матрицом $A^{\mathrm{T}}A$ спроводи неки од поступака ортогонализације (као што је Грам-Шмитов ортогонализациони поступак).
- Сингуларне вредности матрице A (добијене процедуром декомпозиције на сингуларне вредности) су квадратни корени сопствених вредности матрице $A^{\mathrm{T}}A$. На основу тога једначина (2.28) може се написати и у алтернативном облику, $\|A\| = \sigma_{\max}$, одакле се добија (2.8). Ортогоналне матрице U и V остављају дужине вектора непромењене у $\|Ax\| = \left\|U\Sigma V^{\mathrm{T}}x\right\|$. Према томе највећи количник $\|Ax\|/\|x\|$ потиче од највећег елемента σ дијагоналне матрице Σ.

3 Класичне и *Near-Ortho* функције базиса

3.1 Класичне функције базиса

Функције базиса вишег реда које смо користили до сада, а које су биле дефинисане као

$$\alpha_i(u) = \begin{cases} \dfrac{1-u}{2}, & i = 0 \\ \dfrac{1+u}{2}, & i = 1 \\ u^i - 1, & i = 2,4,6,... \\ u^i - u, & i = 3,5,7,... \end{cases}, \quad -1 \le u \le 1, \tag{3.1}$$

надаље ћемо називати класичне функције базиса (*classical basis functions*). Видели смо да ове функције базиса поседују велики број предности у односу на функције базиса нижег реда. За почетак, њихова велика предност је у томе што су у стању да врло ефикасно моделују релативно сложен проблем, односно за жељену тачност решења потребно је значајно мање непознатих него када се користе функција базиса нижег реда. Њихова друга добра особина је то што су класичне функције базиса конструисане тако да је помоћу њих релативно лако задовољити услов континуалности (поља) на споју два елемента. У ту сврху класичне функције базиса су конструисане тако да су само прве две функције базиса (α_0 и α_1) различите од нуле на крајевима посматраног коначног елемента, док су све остале функције базиса (α_i, $i \ge 2$) једнаке нули на крајевима посматраног коначног елемента. На тај начин само коефицијенти развоја уз базисне функције α_0 и α_1 учествују у условима континуалности (поља) на споју два коначна елемента, на основу чега се из неповезаног система једначина формира повезани систем једначина.

Са друге стране, главни недостатак класичних функција базиса је што (по правилу) резултују матрицом система једначина која је слабо условљена, поготову за врло високе редове функција базиса. Овакве матрице (по правилу) имају врло висок кондициони број, што негативно утиче на тачност решења система једначина, а финално (за довољно висок ред функција базиса) може водити и ка неконтролисаном порасту грешке решења са даљим повећањем редова функција базиса.

Висок кондициони број матрице система једначина са собом носи још једну негативну појаву. У случајевима када се систем једначина решава неком од стандардних метода за итеративно решавање система једначина, са порастом кондиционог броја матрице расте и потребан број итерација неопходан за жељену тачност решења система једначина. Самим тим продужава се време (итеративног) решавања система једначина, што је у потпуној супротности са основним мотивом за коришћење итеративних алгоритама за решавање система једначина.

Near-Ortho функције базиса, које ћемо представити у наредном потпоглављу, уведене су тако да отклоне основни недостатка класичних функција базиса, уз задржавање свих предности класичних функција базиса. *Near-Ortho* функције базиса конструисане су тако да резултују матрицом система једначина са (значајно) мањим кондиционим бројем у односу на класичне функције базиса, самим тим *Near-Ortho* функције базиса резултују мањом грешком приликом решавања система једначина

директном методом, мањим бројем итерација када се користе итеративни алгоритми за решавање система једначина и, финално, омогућавају коришћење врло високих редова функција базиса, што води ка изузетно ефикасном нумеричком моделу.

3.2 Near-Ortho

Подсетимо се да смо приликом одређивања напона дуж вода са губицима методом коначних елемената, односно приликом формирања одговарајућег система линеарних једначина прорачунавали два типа интеграла. Први тип интеграла био је облика

$$\int_z \frac{\mathrm{d}\alpha_i(z)}{\mathrm{d}z}\frac{\mathrm{d}\alpha_j(z)}{\mathrm{d}z}\mathrm{d}z\,, \quad 0 \le i, j \le N\,, \tag{3.2}$$

при чему је N био максимални ред функција базиса. Вредности оваквих интеграла фигурисале су у матрици крутости $[S]$. Други тип интеграла био је облика

$$\int_z \alpha_i(z)\alpha_j(z)\mathrm{d}z\,, \quad 0 \le i, j \le N\,, \tag{3.3}$$

а вредности оваквих интеграла фигурисале су у матрици масе $[M]$. Финалну матрицу система једначина добијали смо као $[F]=[S]+R'G'[M]$.

Пракса показује да побољшавањем условљености било које од матрица $[S]$ или $[M]$ по правилу побољшавамо и условљеност финалне матрице $[F]$, што нам је крајњи циљ како бисмо добили што боље условљен систем једначина који ћемо финално решавати. *Near-Ortho* функције базиса уведене су тако да се побољша условљеност матрице масе, $[M]$. Пошто се приликом прорачуна елемената матрице $[M]$ интеграција ионако спроводи у родитељском домену (по u у границама $-1 \le u \le 1$), уместо интеграла (3.3) надаље ћемо посматрати интеграл

$$\int_{u=-1}^{1} \alpha_i(u)\alpha_j(u)\mathrm{d}u\,. \tag{3.4}$$

Код *Near-Ortho* функција базиса желимо да задржимо добро својство класичних функција базиса које се огледало у (релативно) једноставном успостављању континуалности (поља) на споју два коначна елемента. Због тога ћемо за прве две *Near-Ortho* функције базиса узети исте функције базиса као и код класичних, односно, прве две *Near-Ortho* функције базиса су[11]

$$\alpha_i(u) = \begin{cases} \dfrac{1-u}{2}, & i=0 \\ \dfrac{1+u}{2}, & i=1 \end{cases}, \quad -1 \le u \le 1\,. \tag{3.5}$$

У претходном поглављу видели смо да јединична матрица има јединични кондициони број, што је ултимативни циљ са становишта условљености матрице. Због тога, како бисмо побољшали условљеност матрице масе $[M]$, функције базиса изабраћемо тако да добијемо матрицу која по структури што више подсећа на јединичну матрицу, односно желимо да добијемо матрицу $[M]$ која има доминантне елементе на главној дијагонали и што је могуће мање ненултих елементе ван главне дијагонале. Када се подсетимо да су Лежандрови полиноми ортогонални, у ту сврху може се пасти у искушење да за функције базиса вишег реда (α_i, $i \ge 2$) изаберемо Лежандрове полиноме, односно да изаберемо $\alpha_i(u) = P_i(u)$, $i \ge 2$. Са становишта ортогоналности, односно са становишта условљености матрице $[M]$ ово би био одличан избор. Међутим, Лежандрови полиноми на границама коначног елемента ($|u|=1$) су различити од нуле, односно по модулу су једнаки јединици, $|P_n(\pm 1)|=1$, што се може видети и са слике 1.1. Због тога би избор $\alpha_i(u) = P_i(u)$, $i \ge 2$, резултовао тиме да су и ове функције базиса различите од нуле на крајевима коначних елемената, што би значајно отежало успостављање

[11] Постоје функције базиса које поседују виши ниво ортогоналности у односу на *Near-Ortho* функције базиса. Те функције базиса називају се *Max-Ortho* функције базиса, а код њих су прве две функције базиса сложеније у односу на прве две *Near-Ortho* функције базиса. *Max-Ortho* функције базиса превазилазе оквире овог курса.

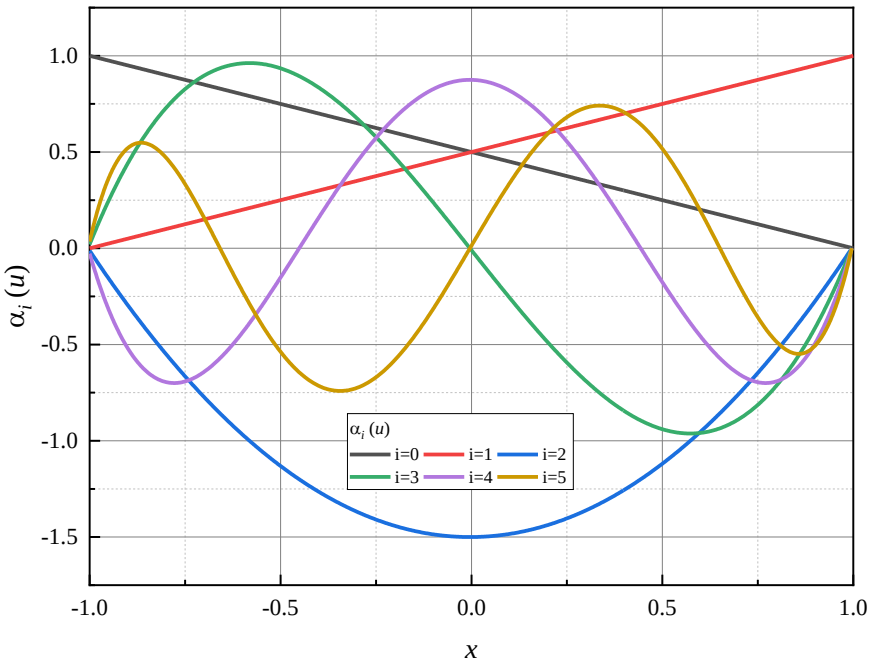
континуалности (поља) на споју два коначна елемента. Због тога функције базиса реда вишег од првог ($\alpha_i(u)$, $i \geq 2$) конструишемо на следећи начин

$$\alpha_i(u) = P_i(u) - P_{i-2}(u), \; i \geq 2, \quad -1 \leq u \leq 1, \tag{3.6}$$

па *Near-Ortho* функције базиса постају

$$\alpha_i(u) = \begin{cases} \dfrac{1-u}{2}, & i = 0 \\ \dfrac{1+u}{2}, & i = 1 \\ P_i(u) - P_{i-2}(u), & i \geq 2 \end{cases}, \quad -1 \leq u \leq 1. \tag{3.7}$$

На слици 3.1 приказано је првих шест *Near-Ortho* функција базиса. Са те слике видимо да само прве две функције базиса имају ненулту вредност на крајевима коначног елемента, што нам је и био циљ.



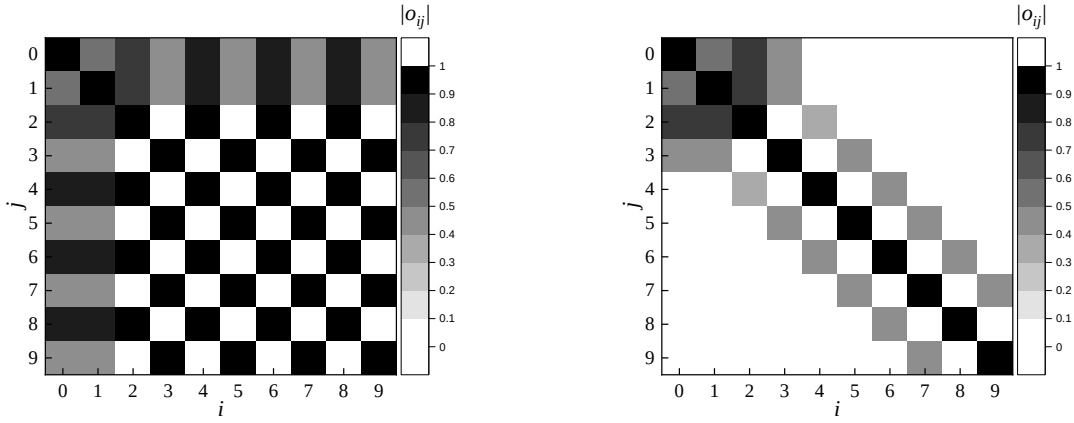
Слика 3.1. Првих шест *Near-Ortho* функција базиса.

Како бисмо стекли бољи увид у структуру матрице масе $[M]$ у случају класичних и *Near-Ortho* функција базиса, уведимо фактор ортогоналности као

$$o_{ij} = \frac{\int_{u=-1}^{1} \alpha_i(u)\alpha_j(u)\mathrm{d}u}{\sqrt{\int_{u=-1}^{1} \alpha_i^2(u)\mathrm{d}u}\sqrt{\int_{u=-1}^{1} \alpha_j^2(u)\mathrm{d}u}}. \tag{3.8}$$

У принципу фактор ортогоналности сличан је елементима који ће се појављивати у матрици масе, односно интегралу у (3.4), а у односу на интеграл (3.4) два интеграла који се појављују у бројиоцу израза (3.8) уведена су само ради скалирања, тако да $|o_{ij}| \leq 1 \;\; \forall i, j$. На слици 3.2(а), у облику матрице, приказан је модул фактора ортогоналности за првих десет класичних функција базиса, а на слици 3.2(б) приказан је исти коефицијент за првих десет *Near-Ortho* функција базиса. Са те слике видимо да матрица модула коефицијента ортогоналности *Near-Ortho* функција базиса по структури више личи на јединичну матрицу него што је то случај за класичне функције базиса. Због тога очекујемо да ће матрица масе бити (доста) боље условљена за *Near-Ortho* функције базиса у односу на класичне функције базиса, односно да ће се ово пренети и на финалну матрицу $[F]$. У наредном потпоглављу, на конкретном примеру

показаћемо да *Near-Ortho* функције базиса резултују матрицом $[F]$ са нижим кондиционим бројем, односно да су постигнути сви циљеви због којих су *Near-Ortho* функције базиса и уведене.



(а) (б)

Слика 3.2. Коефицијент ортогоналности за (а) класичне и (б) *Near-Ortho* функције базиса.

4 Пример примене *Near-Ortho* функција базиса

Пример примене *Near-Ortho* функција базиса из кога се могу сагледати њихове предности у односу на класичне функције базиса приказан је кроз следећи пример.

Пример:

На једном крају вода прикључен је идеалан напонски генератор сталне електромоторне силе $E_0 = 1\,\mathrm{V}$, а на другом крају вод је отворен. Вод је дужине $l = 30\ \mathrm{m}$, подужне отпорности $R' = 9{,}308\ \Omega/\mathrm{m}$ и подужне проводности $G' = 3{,}469\ \mathrm{mS/m}$. Аналитички и Галеркиновим методом коначних елемената вишег реда прорачунати напон дуж вода. Вод је моделован помоћу K униформних коначних елемената, ред функција базиса је N, а нумеричку интеграцију приликом тестирања спровести Гаус-Лежандровом интеграционом формулом у $N_{\mathrm{GL}} = N + 5$ тачака. Прорачун (интеграцију) елемената матрице метода коначних елемената спровести само за један (родитељски) коначни елемент. Полазећи од елемената те матрице, коришћењем вектора повезаности формирати (финалну) матрицу повезаног система једначина реда N_{con} (број непознатих). Реалне бројеве представити у двострукој тачности.

(а) За класичне функције базиса на једном графику приказати фамилију крива корена из средње квадратне грешке напона у функцији броја напознатих, N_{con}, а на другом графику приказати фамилију крива кондиционог броја финалне матрице система једначина у функцији броја непознатих (N_{con}). На сваком графику приказати криве за $K \in \{1, 5, 10, 50, 100\}$, при чему за сваку криву редови функција базиса чине скуп $N \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$. Систем једначина решити директним методом, а корен из средње квадратне грешке прорачунати за $N_{\mathrm{plot}} = 1000$ униформно расподељених тачака дуж вода. Резултате приказати у log-log скали.

(б) Поновити тачку под (а) користећи *Near-Ortho* функције базиса.

(в) За класичне функције базиса из тачке под (а) систем једначина решити итеративним методом[12]. За критеријум конвергенције итеративног метода узети квадратну вредност резидуала $r^2 = 10^{-30}$, а број итерација ограничити на $N_{\mathrm{itmax}} = 10^4$. На једном графику приказати фамилију крива корена из средње квадратне грешке напона у функцији броја напознатих, N_{con}, а на другом укупан број

[12] У случају да се проблем решава помоћу програмског пакета MATLAB, за итеративно решавање једначина користити рутину „x = pcg(A,b,tol,maxit)“, при чему је $\mathrm{tol} = r^2$, а $\mathrm{maxit} = N_{\mathrm{itmax}}$. У случају да се проблем решава помоћу Intel MKL пакета, за итеративно решавање једначина користити рутине „dcg_init“, „dcg_check“ и „dcg“, а подесити и одговарајуће параметре на следећи начин $\mathrm{ipar}[4] = N_{\mathrm{itmax}}$, $\mathrm{ipar}[8] = 1$, $\mathrm{ipar}[9] = 0$ и $\mathrm{dpar}[0] = r^2$.

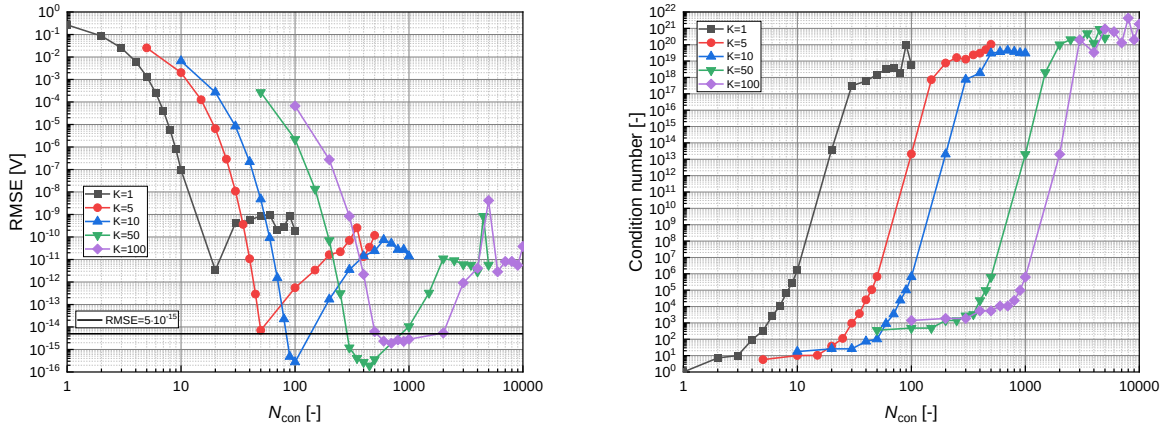
итерација итеративног метода у функцији N_{con}. На сваком графику приказати криве за $K \in \{1, 5, 10, 50, 100\}$, при чему за сваку криву редови функција базиса чине скуп $N \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$. Резултате приказати у log-log скали.

(г) Поновити тачку под (в) користећи *Near-Ortho* функције базиса.

(д) На основу графика из тачака (а)-(г), за оба начина решавања система једначина и за оба типа базисних функција, на једном графику, у функцији броја коначних елемената за $K \in \{1, 5, 10, 50, 100\}$, приказати минималан број непознатих, N_{conmin}, за које је корен из средње квадратне грешке напона мањи од $5 \cdot 10^{-15}$ V. У случајевима када корен из средње квадратне грешке напона не пада испод $5 \cdot 10^{-15}$ V, за N_{conmin} узети број непознатих када корен из средње квадратне грешке напона има минимум. На истом графику приказати и одговарајући корен из средње квадратне грешке напона (график са две вертикалне осе). Резултате приказати у log-log скали. На основу свих приказаних резултата навести главне предности и недостатке *Near-Ortho* функције базиса у односу на класичне функције базиса.

Решење:

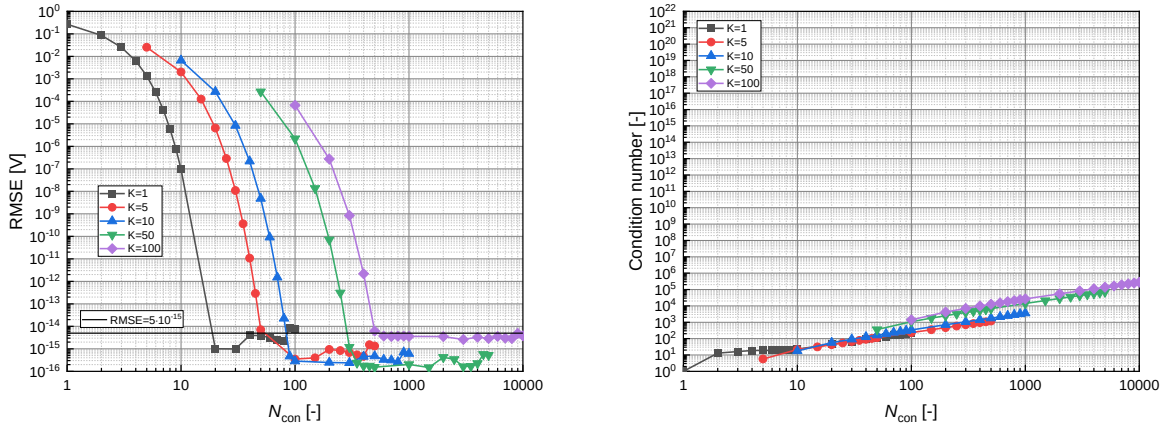
(а) Тражени графици приказани су на слици 4.1.



(а) (б)

Слика 4.1. Класичне функције базиса, директно решавање система једначина.

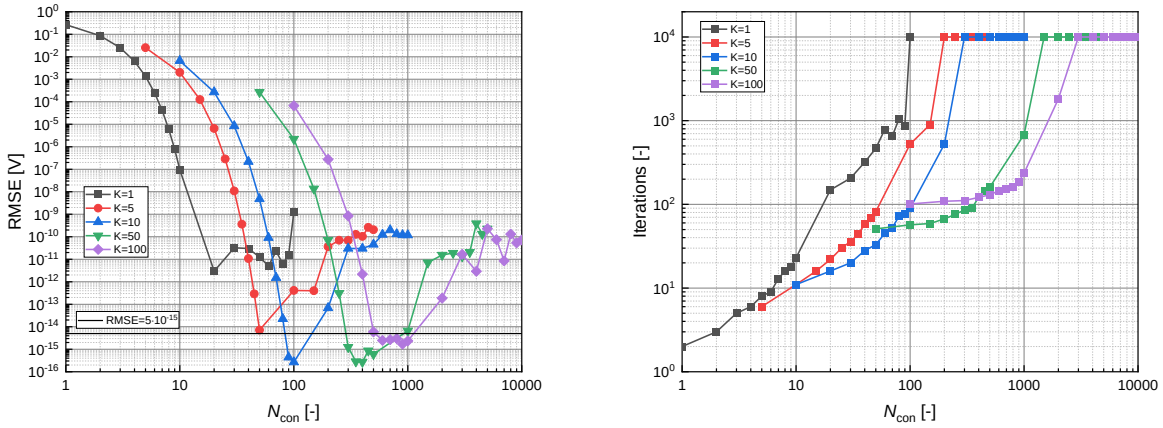
(б) Тражени графици приказани су на слици 4.2.



(а) (б)

Слика 4.2. *Near-Ortho* функције базиса, директно решавање система једначина.

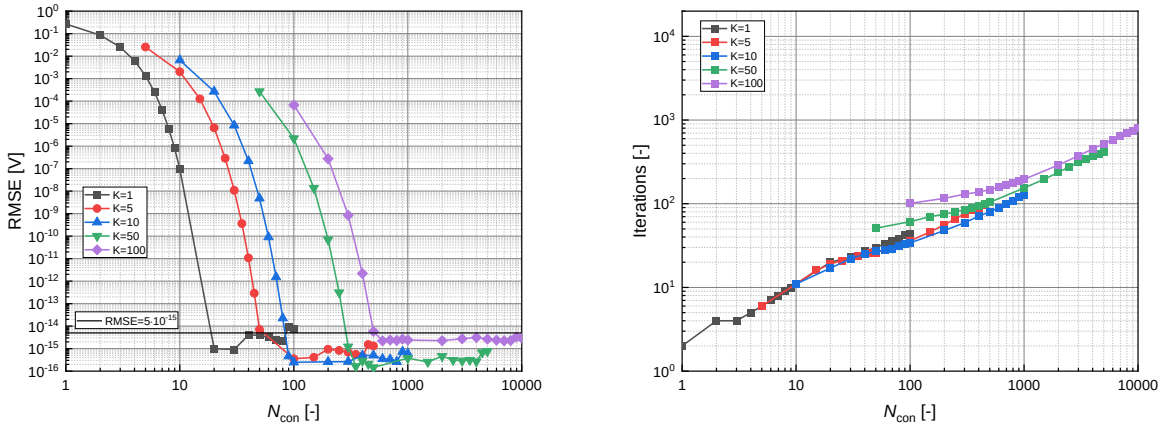
(в) Тражени графици приказани су на слици 4.3.



(а) (б)

Слика 4.3. Класичне функције базиса, итеративно решавање система једначина.

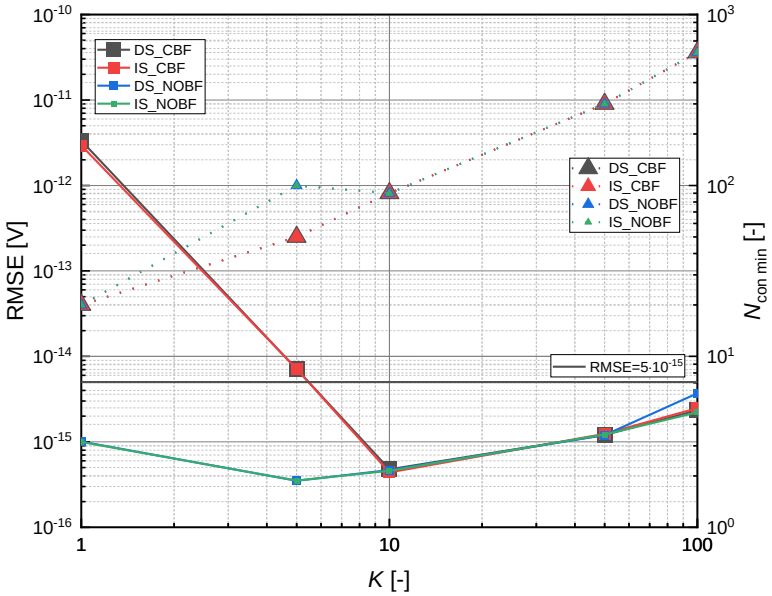
(г) Тражени графици приказани су на слици 4.4.



(а) (б)

Слика 4.4. *Near-Ortho* функције базиса, итеративно решавање система једначина.

(д) Тражени график приказан је на слици 4.5.



Слика 4.5.