Fakulta
jaderná a fyzikálně
inženýrská

VYDAVATELSTVÍ ČVUT

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

ÚVOD DO TEORIE CITLIVOSTI A STABILITY V NUMERICKÉ LINEÁRNÍ ALGEBŘE

RNDr. Jitka Drkošová

Ing. Zdeněk Strakoš, CSc.

RNDr. Jitka Drkošová
Ing. Zdeněk Strakoš, CSc.

ÚVOD DO TEORIE CITLIVOSTI A STABILITY V NUMERICKÉ LINEÁRNÍ ALGEBŘE

1997
Vydavatelství ČVUT

Lektor: doc. RNDr. Emil Humhal, CSc.

ISBN 80-01-01560-2

Obsah

Úvod 2

1 Citlivost úlohy 5

2 Numerická stabilita metody (algoritmu) 7
2.1 Aritmetika s pohyblivou řádovou čárkou 7
2.2 Zaokrouhlovací chyby v aritmetice s konečnou přesností 12
2.3 Přímá a zpětná stabilita 15

3 Citlivost vlastních čísel matic 17
3.1 Schurova dekompozice, spektrální rozklad a Jordanův kanonický tvar . . . 18
3.2 Citlivost vlastních čísel obecných matic 25
3.2.1 Spojitost vlastních čísel 26
3.2.2 Elsnerova a Ostrowského-Elsnerova věta 27
3.2.3 Bauerova-Fikeho a Henriciho věta 35
3.3 Citlivost jednoduchého vlastního čísla 41
3.4 Citlivost vlastních čísel diagonalizovatelných a normálních matic 49
3.5 Příklady 51

4 Citlivost řešení soustav lineárních rovnic 61

5 Odhady chyb a zpětná stabilita 66
5.1 Vlastní čísla 67
5.2 Soustavy lineárních rovnic 69

Závěr 72

Literatura 74

Úvod

Náš výklad zahájíme krátkou úvahou o postupu řešení typického technického, ekonomického, fyzikálního či jiného problému reálného světa za pomoci počítače.

Především potřebujeme problém popsat prostředky daného oboru. Výsledkem je většinou *zjednodušený matematický model.* Zjednodušení znamená vyloučení nepodstatných závislostí a umožňuje sestavit model tak, aby s ním bylo možné dále pracovat. Dopouštíme se tím ovšem první ze série chyb či nepřesností, řekněme *chyby modelu*, protože skutečnost není popsána modelem přesně, ale jen přibližně. *Analýzou matematického modelu* získáme základní informace o hledaném řešení, jeho existenci, jednoznačnosti a dalších vlastnostech. Velmi zřídka jsme však schopni řešení analytickými prostředky také nalézt. Jsou-li například k vytvoření modelu použity diferenciální či integrodiferenciální rovnice a řešení je prvkem nekonečně dimenzionálních prostorů funkcí, je jeho analytické určení možné jen ve velmi jednoduchých či speciálních případech.

Víme-li, že řešení existuje, můžeme se pokusit nalézt jeho *numerickou aproximaci.* Základním krokem je většinou diskretizace nekonečně dimenzionálního problému a jeho převedení na algebraickou, konečně dimenzionální úlohu. Proces diskretizace a s ním spojená *chyba diskretizace* je předmětem podstatné části *numerické analýzy.* Na závěr zbývá numerické řešení konečně dimenzionální algebraické úlohy. Pokud jde o úlohu nelineární, je obvykle tím či oním způsobem linearizována (opět s jistou chybou) a v posledním kroku je numericky řešena *lineární algebraická úloha.* Stejně jako u každého předcházejícího kroku nás musí zajímat nejen vypočtená aproximace řešení, ale i příslušná *numerická chyba.*

Úlohy lineární algebry lze přirozeným způsobem formulovat pomocí matic. Zjednodušeně řečeno, předmětem *numerické lineární algebry* je numerické řešení soustav lineárních rovnic, výpočet vlastních čísel matic, řešení problému nejmenších čtverců a hledání rozkladů matic. Ne všechny tyto úlohy lze řešit přesně (příkladem je určování vlastních čísel matic s dimenzí větší než čtyři). Navíc, často je výhodné hledat pouze vhodnou aproximaci řešení (například iteračním postupem). V této souvislosti hovoříme o *chybě metody.* Výpočty provádíme na počítači s konečnou aritmetikou, což přináší *zaokrouhlovací chyby.* Studium jejich vlivu při řešení úloh lineární algebry je obsahem následujícího textu.

Máme-li řešit některou z úloh numerické lineární algebry, chceme zajisté vybrat co nejlepší metodu. Musí nás přitom zajímat nejen rychlost (počet prováděných operací), ale i odolnost jednotlivých metod vůči šíření zaokrouhlovacích chyb. V této souvislosti mluvíme o *numerické stabilitě metody.* Jak uvidíme v následujících kapitolách, lze k určení chyb řešení způsobených zaokrouhlováním s výhodou použít poznatků o *citlivosti řešené úlohy* vzhledem k malým změnám vstupních dat.

V předloženém textu jsme se pokusili vyložit vznik zaokrouhlovacích chyb v numerických výpočtech, ukázat jejich nebezpečí a naznačit způsob, jakým lze jejich vliv popsat. Od počátku je potřebné mít na paměti, že cílem analýzy citlivosti a analýzy zaokrouhlovacích chyb je rozpoznat, která úloha je snadno a která je obtížně numericky řešitelná a

porozumět numerickému chování jednotlivých metod (zde se však stabilitou jednotlivých metod nebudeme zabývat). V konečném důsledku nám to nejen umožní vybrat vhodnou metodu, ale i určit, jak daleko je námi vypočtené numerické řešení od řešení hledaného.

Tato práce představuje první ze zamýšlené série učebních textů. To, že začínáme se zaokrouhlovacími chybami, je dáno jednak snahou po zdůraznění této mnohdy opomíjené součásti numerických výpočtů, jednak snahou po zaplnění mezery v snadno dostupné literatuře. V dnešní době se často můžeme setkat s bezhlavým používáním počítačového programového vybavení (včetně numerického software) a s jistým trendem k povrchnímu a plytkému posuzování úspěšnosti. Chyby nám v tomto kontextu připadají jako vhodné téma pro začátek. Další části učebních textů budou následovat v intervalech určených zájmem, edičními a časovými možnostmi.

Pokud bude předložený text shledán dobrým a užitečným, je to zásluhou literatury, ze které jsme čerpali. Zdrojem informací nám byly zejména náledující knihy:

[FMC] Watkins, D.S. : Fundamentals of Matrix Computations, J. Willey, N.Y., 1991,

[MPT] Stewart, G.W., Sun, J. : Matrix Perturbation Theory, Academic Press, Boston, 1990,

[ASNA] Higham, N.J. : Accuracy and Stability of Numerical Algorithms, SIAM, Philadelphia, 1996.

Z toho Highamova kniha vyšla bohužel až v době, kdy byl text ze značné míry hotov. Je vhodné zmínit se, že v nejbližší době (začátkem roku 1997) vyjde další monografie

[NLA] Demmel, J.W. : Numerical Linear Algebra, SIAM, Philadelphia, 1997.

Podstatné informace o otázkách lineární algebry a numerických metodách nalezne čtenář v učebních textech:

[LA] Pytlíček, J. : Lineární algebra, FJFI ČVUT, Praha (v přípravě),

[NM] Humhal, E. : Numerická matematika I, FJFI ČVUT, Praha, 1989.

Jako základní knihy pro další studium souvisejících otázek teorie matic a metod numerické lineární algebry doporučujeme:

[SM] Fiedler, M. : Speciální matice a jejich použití v numerické matematice, SNTL, Praha, 1981,

[MC] Golub, G.H., van Loan, C.F. : Matrix Computations (Second Edition), The Johns Hopkins Univ. Press, Baltimore, 1989,

[MA1] Horn, A.G., Johnson, C.R. : Matrix Analysis, Cambridge University Press, Cambridge, 1985,

[MA2] Horn, A.G., Johnson, C.R. : Topics in Matrix Analysis, Cambridge University Press, Cambridge, 1991,

[AEP] Wilkinson, J.H. : Algebraic Eigenvalue Problem, Oxford University Press, London, 1965.

Za pečlivé přečtení rukopisu děkujeme doc. RNDr. Emilu Humhalovi, CSc. a ing. Miroslavu Rozložníkovi. Jejich poznámky a cenné připomínky přispěly ke zlepšení textu.

Za chyby a nedostatky předloženého učebního textu odpovídají plně jeho autoři. Budeme vděčni za jakékoli poznámky a připomínky. Internetové adresy autorů jsou jitka@uivt.cas.cz, strakos@uivt.cas.cz.

Kapitola 1

Citlivost úlohy

Uvažujme některou z úloh numerické lineární algebry, například řešení soustavy lineárních algebraických rovnic či výpočet vlastních čísel matice. Úloha je zadána vstupními daty, tj. hodnotami jednotlivých prvků matice, případně hodnotami prvků pravé strany. Vstupní data bývají většinou zatížena chybami (např. chybami měření nebo některými z chyb zmíněných v úvodu) a naše úloha se tedy většinou liší od té, kterou bychom skutečně chtěli řešit. I když ji vyřešíme přesně, je naše řešení odlišné od řešení skutečně hledaného. Předpokládejme, že chyba vstupních dat není velká. Je rozumné položit si otázku, jak se chyba ve vstupních datech promítne do chyby v řešení. Jinými slovy, ptáme se, jak je úloha *citlivá* na malé změny vstupních dat. *Citlivostí úlohy* budeme tedy rozumět vlastnost určující vliv malých změn (perturbací) vstupních dat na změnu řešení úlohy.

V dalších kapitolách ukážeme, jak lze citlivost popsat a jak se analýza citlivosti řešené úlohy užívá při odhadech velikosti chyby aproximace řešení způsobené zaokrouhlováním při výpočtech v konečné aritmetice (tj. na počítači).

Označme $U(z_1, \ldots, z_m)$ přesné řešení úlohy U se vstupními daty $(z_1, \ldots, z_m)$. Úloha nebude citlivá na změny vstupních dat, pokud změna

$$U(z_1, \ldots, z_m) - U(\tilde{z}_1, \ldots, \tilde{z}_m)$$

bude přiměřená vzdálenosti vstupních dat

$$(z_1, \ldots, z_m) - (\tilde{z}_1, \ldots, \tilde{z}_m).$$

V tomto případě se říká, že úloha je *dobře podmíněná*. Dále se naučíme charakterizovat podmíněnost úloh kvantitativně, tj. velikostí k tomu určených parametrů. Úmyslně zde nezavádíme přesný formální popis, neboť nám jde o pochopení smyslu jednotlivých pojmů. Formální popis bývá závislý na úloze a často obsahuje detaily, které zde nejsou nutné.

Podmíněnost úlohy je dána její základní (např. fyzikální) formulací a matematickým modelem, procesem diskretizace atd. Jak uvidíme dále, pro špatně podmíněnou úlohu (problém je citlivý na malé perturbace vstupních dat) můžeme i při použití velmi kvalitního algoritmu, omezujícího v maximální možné míře vliv zaokrouhlovacích chyb, dostat velkou chybu aproximace řešení. V takovém případě není selhání numerického výpočtu způsobeno špatnou volbou metody (a z toho vyplývajícím zničujícím vlivem zaokrouhlovacích chyb). Problém je v samotné formulaci úlohy. Špatně podmíněné úlohy neumíme často vůbec uspokojivě řešit. Názornou ukázkou špatně podmíněné úlohy je následující příklad.

Příklad 1.1 *Soustava*

$$\begin{aligned} 2x + 6y &= 8 \\ 2x + 6.00001y &= 8.00001 \end{aligned}$$

má řešení $x = 1$ *a* $y = 1$, *soustava s malou relativní změnou v prvku* a_{22} *a* b_2

$$\begin{aligned} 2x + 6y &= 8 \\ 2x + 5.99999y &= 8.00002 \end{aligned}$$

má řešení $x = 10$ *a* $y = -2$. *Matice původní soustavy*

$$\begin{pmatrix} 2 & 6 \\ 2 & 6.00001 \end{pmatrix}$$

má téměř lineárně závislé sloupce (či řádky) a velké číslo podmíněnosti (prvky matice inverzní jsou řádu 10^5*), proto i malá změna vstupních dat způsobila velký rozdíl v řešení.*

V kapitolách 3 a 4 budeme studovat, jak jsou vlastní čísla čtvercové matice citlivá na změny jednotlivých prvků matice a jak je řešení soustavy lineárních algebraických rovnic citlivé na změny prvků matice, či změny prvků pravé strany. V příští kapitole popíšeme vznik zaokrouhlovacích chyb a ukážeme jejich elementární vlastnosti. Zejména však ukážeme souvislost mezi analýzou citlivosti a analýzou numerické stability.

Kapitola 2

Numerická stabilita metody (algoritmu)

V této kapitole se budeme zabývat vlivem zaokrouhlovacích chyb, které vznikají při numerických výpočtech prováděných na počítači v aritmetice s konečnou přesností. Bude nás zajímat, zda je algoritmus *stabilní vůči zaokrouhlovacím chybám*, tj. zda je výsledek výpočtu „dostatečně přesná" aproximace řešení. Nejprve popíšeme vznik zaokrouhlovacích chyb a jejich šíření při provádění elementárních aritmetických operací.

2.1 Aritmetika s pohyblivou řádovou čárkou

Číslo je v počítači zobrazeno jako posloupnost bitů (každý s číselným obsahem 0 nebo 1) konečné délky. Tato délka je pevně stanovena (např. 16, 32, 64 či 128 bitů). Počítač většinou umožňuje několik typů zobrazení čísel, jimž odpovídá několik velikostí paměťových míst. Nás především zajímá, jak jsou v počítači zobrazena reálná čísla.

Je zřejmé, že při zvoleném typu zobrazení a předepsané délce paměťového místa je možno v počítači zobrazit pouze konečný počet čísel. Proto často hovoříme o *konečné aritmetice* či *aritmetice s konečnou přesností*. Množina reálných čísel je v počítači reprezentována svojí konečnou podmnožinou $\mathcal{F} \subset R$, kterou nazýváme soustavou čísel s *pohyblivou řádovou čárkou* (floating point number system). Její prvky lze zapsat ve tvaru

$$y = \pm m \times \beta^{e-t} , \tag{2.1}$$

kde celé číslo β (obvykle $\beta = 2$) je nazýváno *základem*, celé číslo t určuje *přesnost*, celé číslo m pohybující se v rozsahu $0 \leq m < \beta^t - 1$ je nazýváno *mantisou* a celočíselný parametr e *exponentem*. Množina $\mathcal{F}$ je plně určena parametry β, t a horní resp. dolní mezí celočíselného exponentu, $e_{\min} \leq e \leq e_{\max}$.

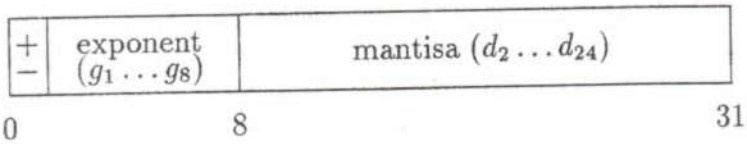
Vztah (2.1) můžeme přepsat do názornějšího tvaru

$$y = \pm\beta^e \left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \ldots + \frac{d_t}{\beta^t} \right) = \pm\beta^e \times 0.d_1 d_2 \ldots d_t , \tag{2.2}$$

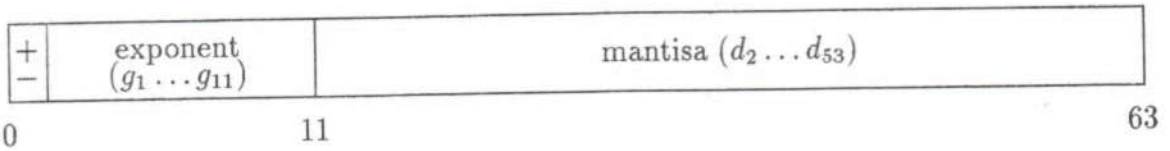
kde každá číslice d_i leží v intervalu $0 \leq d_i \leq \beta - 1$, tj. $0.d_1 d_2 \ldots d_t$ představuje číslo zapsané v číselné soustavě se základem β. Je výhodné uvažovat $m \geq \beta^{t-1}$ pro $y \neq 0$; pak zřejmě $d_1 \neq 0$. Systém dodržující tuto konvenci nazýváme *normalizovaný*. I když se v minulosti používaly různé základy β (do dnešní doby jsou rozšířeny základy $\beta = 2$ a $\beta = 16$) a stále se můžeme setkat s rozličnými hodnotami t, $e_{\min}$ a $e_{\max}$, vývoj spěje k všeobecnému uznání tzv. *IEEE standardní aritmetiky*. Stručně ji vyložíme a další vlastnosti aritmetiky s pohyblivou řádovou čárkou budeme popisovat na tomto příkladu.

IEEE aritmetika používá $\beta = 2$ a rozlišuje dva základní formáty čísel v pohyblivé řádové čárce: čísla s *jednoduchou* a *dvojitou přesností*. V prvém případě je k uložení čísla použito 32, ve druhém 64 bitů. Uložení jednotlivých parametrů je patrné z následujícího schematu:

jednoduchá přesnost



dvojitá přesnost



V případě jednoduché přesnosti je na exponent vyhrazeno 8 bitů, do kterých je možno uložit celé číslo v rozmezí 0 až 255. Řetězce $(0000\ 0000)_2 = 0$ či $(1111\ 1111)_2 = 255$ mají však speciální význam, který bude popsán níže. Zbylá čísla 1 až 254 určují hodnotu veličiny

$$e + 126,$$

tj. hodnota exponentu e se pohybuje v rozmezí

$$e_{\min} = -125 \leq e \leq 128 = e_{\max}.$$

Na mantisu je vyhrazeno zbývajících 23 bitů, přičemž se standardně využívá normalizace; cifra $d_1 = 1$ se přitom nezapisuje. Uložené nenulové číslo v pohyblivé řádové čárce můžeme tedy zapsat ve tvaru

$$y = \pm 2^{(g_1 \ldots g_8)_2 - 126} \times (0.1 d_2 d_3 \ldots d_{24})_2, \tag{2.3}$$

z čehož vyplývá

$$2^{-126} \leq |y| \leq (1 - 2^{-24}) 2^{128} \sim 10^{38}.$$

Čísla v pohyblivé řádové čárce nejsou vzhledem k R rovnoměrně rozložena (viz cvičení 1). Mají-li dvě čísla y_1, y_2 ve vyjádření (2.3) shodný exponent e a jedná-li se o dvě po sobě jdoucí čísla množiny $\mathcal{F}$, $y_2 > y_1$, pak

$$y_2 - y_1 = 2^e \times 2^{-24}.$$

Rozložení čísel množiny $\mathcal{F}$ je charakterizováno pomocí *strojové přesnosti* ε_M, což je vzdálenost čísla 1.0 od nejbližšího vyššího čísla v pohyblivé řádové čárce. Zřejmě platí $\varepsilon_M = 2^1 \times 2^{-24} = 2^{-23}$. Snadno lze ukázat, že vzdálenost libovolného normalizovaného čísla $x \in \mathcal{F}$ od svých sousedů je nejméně $\varepsilon_M|x|/2$ a nejvýše $\varepsilon_M|x|$ (cvičení 2).

Pokud by množina $\mathcal{F}$ obsahovala pouze normalizovaná čísla popsaná (2.3), došlo by k nepříjemnému jevu - zatímco čísla blízká 2^{-126} zprava by byla aproximována s chybou odpovídající počtu bitů mantisy, nejbližší číslo menší než 2^{-126} definované (2.3) je -2^{-126} (dojde k tzv. mezeře v okolí nuly). K odstranění této anomálie obsahuje v IEEE aritmetice množina $\mathcal{F}$ rovněž tzv. čísla *subnormální*, což jsou nenulová nenormalizovaná čísla s exponentem $(0000\ 0000)_2 = 0$, definovaná vztahem

$$y = \pm m \times \beta^{e_{\min}-t}, \quad 0 < m < \beta^{t-1},$$

neboli

$$y = \pm m \times 2^{-149}, \quad 0 < m < 2^{23}.$$

Je-li řetězec mantisy i exponentu nulový, $((d_2 d_3 \ldots d_{24}) = (0 \ldots 0), (g_1 \ldots g_8) = (0 \ldots 0))$, dostaneme reprezentaci čísel ± 0 ($+0$ má odlišnou reprezentaci než -0, avšak samozřejmě je zajištěno, že při srovnání $+0 = -0$). Je-li řetězec exponentu roven $(1111\ 1111)_2 = 255$ a mantisa je nulová, pak zobrazené číslo je definováno jako $\pm\infty$. Je-li řetězec exponentu roven $(1111\ 1111)_2 = 255$ a řetězec mantisy je nenulový (jeho hodnota je libovolná), pak je obsah interpretován jako *NaN* (Not a Number). Shrnutí je uvedeno v následující tabulce:

Tabulka 2.1 IEEE jednoduchá přesnost

řetězec exponentu je:	numerická hodnota uloženého čísla
$(0000\ 0000)_2 = (0)_{10}$	$\pm(0.0 d_2 d_3 \ldots d_{24})_2 \times 2^{-125}$
$(0000\ 0001)_2 = (1)_{10}$	$\pm(0.1 d_2 d_3 \ldots d_{24})_2 \times 2^{-125}$
$\downarrow$	$\downarrow$
$(0111\ 1110)_2 = (126)_{10}$	$\pm(0.1 d_2 d_3 \ldots d_{24})_2 \times 2^{0}$
$(0111\ 1111)_2 = (127)_{10}$	$\pm(0.1 d_2 d_3 \ldots d_{24})_2 \times 2^{1}$
$\downarrow$	$\downarrow$
$(1111\ 1110)_2 = (254)_{10}$	$\pm(0.1 d_2 d_3 \ldots d_{24})_2 \times 2^{128}$
$(1111\ 1111)_2 = (255)_{10}$	$\pm\infty$ pokud $d_2 = d_3 = \ldots = d_{24} = 0$, NaN jinak

Zbývají dvě otázky: jaká je přesnost zobrazení reálného čísla v množině $\mathcal{F}$ a jaká je přesnost provádění elementárních aritmetických operací $+, -, *, /$. Přesnost aproximace je charakterizována *zaokrouhlovací jednotkou* $u = (1/2)\beta^{1-t} = (1/2)2^{-23} = 2^{-24}$. Důkaz následující věty (která je formulována obecně) ponecháme do cvičení:

Věta 2.1 *Nechť $x \in R$ leží mezi nejmenším a největším číslem množiny $\mathcal{F}$. Označíme-li fl zobrazení z R do $\mathcal{F}$, pak platí*

$$fl(x) = x(1+\delta), \quad |\delta| < u, \tag{2.4}$$

kde u je zaokrouhlovací jednotka.

Aritmetické operace $+$, $-$, $*$, $/$ se v IEEE aritmetice v obvyklých případech provádějí tak, jako kdyby byly nejprve provedeny přesně (s nekonečnou přesností) a pak výsledek zaokrouhlen na nejbližší číslo z $\mathcal{F}$ (v případě nerozhodnosti se zaokrouhluje dolů). Jsou-li $x, y \in \mathcal{F}$, pak platí

$$\begin{aligned} fl(x \pm y) &= (x \pm y)(1+\delta_1) \quad |\delta_1| \leq u \\ fl(x * y) &= (x * y)(1+\delta_2) \quad |\delta_2| \leq u \\ fl(x/y) &= (x/y)(1+\delta_3) \quad |\delta_3| \leq u \end{aligned} \tag{2.5}$$

Analogický vztah se obvykle předpokládá i pro operaci odmocnění. Nastane-li vyjímečný případ, je výsledek generován podle tabulky 2.2

Tabulka 2.2 Vyjímky v IEEE aritmetice

typ vyjímky	příklad	výsledek
nedefinované operace	$0/0$, $0 \times \infty$, $\sqrt{-1}$	NaN
přetečení		$\pm\infty$
dělení nenulového čísla nulou		$\pm\infty$
podtečení		subnormální čísla

Přetečením rozumíme případ, kdy je přesný výsledek operace v absolutní hodnotě větší, než největší číslo z $\mathcal{F}$. Podtečením rozumíme případ, kdy je přesný výsledek operace v absolutní hodnotě menší, než nejmenší kladné normalizované číslo.

Vlastnosti aritmetiky s pohyblivou řádovou čárkou jsme vyložili na příkladu IEEE aritmetiky s jednoduchou přesností. Je zřejmé, jak postupovat při odvození charakteristik aritmetiky založené na jiné hodnotě parametrů. Pro doplnění uvádíme porovnání IEEE aritmetiky s jednoduchou a dvojitou přesností.

Tabulka 2.3 IEEE aritmetika s pohyblivou řádovou čárkou

přesnost	počet bitů celkově	mantisa	exponent	zaokrouhlovací jednotka u	rozsah
jednoduchá	32	23(+1)	8	$2^{-24} \sim 5.96 \times 10^{-8}$	$10^{\pm 38}$
dvojitá	64	52(+1)	11	$2^{-53} \sim 1.11 \times 10^{-16}$	$10^{\pm 308}$

Cvičení

1. Vypočtěte a graficky znázorněte na číselné ose prvky množiny čísel s pohyblivou řádovou čárkou pro $\beta = 2$, $t = 3$, $e_{\min} = -1$ a $e_{\max} = 3$.

2. Ukažte, že vzdálenost libovolného normalizovaného čísla x z množiny $\mathcal{F}$ od jeho nejbližšího souseda je nejméně $\varepsilon_M |x|/2$ a nejvýše $\varepsilon_M |x|$.

3. Dokažte větu 2.1.

4. Odvoďte parametry IEEE aritmetiky s pohyblivou řádovou čárkou, dvojitou přesností.

5. Který z následujících výroků je pravdivý v IEEE aritmetice, předpokládáme-li, že a, b jsou normalizovaná čísla v pohyblivé řádové čárce a že nenastane žádná vyjímečná situace?

   (a) $fl(a \text{ op } b) = fl(b \text{ op } a)$ $\quad$ op $= +, *$

   (b) $fl(b - a) = -fl(a - b)$

   (c) $fl(a + a) = fl(2 * a)$

   (d) $fl(0.5 * a) = fl(a/2)$

   (e) $fl((a + b) + c) = fl(a + (b + c))$

   (f) je-li $a \leq b$, pak $a \leq fl((a + b)/2) \leq b$

2.2 Zaokrouhlovací chyby v aritmetice s konečnou přesností

Při povrchním pohledu na vztahy (2.4) a (2.5) by se mohlo zdát, že zaokrouhlovací chyby jsou velmi malé a jejich vliv při provádění numerických výpočtů nebude velký (snad s vyjímkou velkého počtu operací s nějakými extrémními čísly). Ukážeme na několika příkladech, že tento ukvapený závěr je zcela mylný.

Prvním příkladem je tzv. krácení (cancellation), které nastává, odečítáme-li dvě téměř shodná čísla.

Příklad 2.1 *Uvažujeme funkci* $f(x) = (1 - \cos x)/x^2$ *použitou v [ASNA]. Pro* $x = 1.2 \times 10^{-5}$ *je hodnota* $\cos x$ *zaokrouhlená na 10 desetinných míst rovna* $c = 0.9999\ 9999\ 99$, *takže vyčíslením hodnoty* $f(1.2 \times 10^{-5})$ *dostaneme*

$$(1 - c)/x^2 = 10^{-10}/(1.44 \times 10^{-10}) = 0.6944\ldots,$$

což je úplně špatně, neboť $0 \leq f(x) \leq 1/2$ *pro* $x \neq 0$. *Vidíme, že přestože hodnota* $\cos x$ *byla aproximována s přesností na 10 desetinných míst, výsledek výpočtu* $f(x)$ *neaproximuje správnou hodnotu ani s přesností jednoho desetinného místa! Je důležité si uvědomit, že problém není způsoben vlastním odečtením* $1-c$, *to bylo provedeno* přesně. *Problém spočívá v tom, že sama hodnota c byla určena nepřesně a výsledek přesného výpočtu* $1 - c$ *je díky krácení platných cifer stejného řádu, jako je chyba obsažená v c. Tím se významnost nepatrné chyby hodnoty c posunula o 10 řádů a katastrofálně ovlivnila celý další výpočet, byť byl proveden sebepřesněji (často se proto hovoří v této souvislosti o tzv. katastrofickém krácení).*

Pokusíme se krácení popsat pomocí vztahů (2.4) a (2.5). Nechť $\hat{x}$ a $\hat{y}$ jsou dvě čísla zatížená jistou chybou, tj. $\hat{x} = x(1 + \Delta x)$, $\hat{y} = y(1 + \Delta y)$. Předpokládejme, že chyby Δx resp. Δy jsou malé vzhledem k velikosti x resp. y; může jít o chyby způsobené předcházejícím výpočtem nebo třeba o zaokrouhlovací chyby při uložení dat do počítače (pak $\hat{x} = fl(x)$, $\hat{y} = fl(y)$ a $|\Delta x| \leq u$, $|\Delta y| \leq u$). Proveďme *přesný* součet čísel $\hat{x}$ a $\hat{y}$ (čísla mohou mít opačná znaménka, příklad zahrnuje i odečítání):

$$\begin{aligned} \hat{s} = \hat{x} + \hat{y} &= x(1 + \Delta x) + y(1 + \Delta y) \\ &= x + y + x\Delta x + y\Delta y \\ &= (x + y)(1 + \Delta s), \end{aligned}$$

kde

$$\Delta s = \frac{x}{x + y}\Delta x + \frac{y}{x + y}\Delta y.$$

Je zřejmé, že i když hodnoty Δx a Δy jsou malé, není zaručeno, že hodnota Δs bude rovněž malá. Pokud bude $x \gg (x + y)$ a zároveň $\Delta x \neq 0$, nebo $y \gg (x + y)$ a zároveň $\Delta y \neq 0$, bude chyba Δs relativně velká. Znovu vidíme, že krácení není nebezpečné samo o sobě (dojde-li ke krácení při odečtení dvou přesných hodnot, žádná ztráta přesnosti nenastane), ale je nebezpečné tím, že zesiluje vliv předchozích chyb obsažených v datech.

Druhý příklad ukazuje, že i bez krácení, popsaného výše, může dojít při provedení jednoduchého výpočtu k velké chybě.

Příklad 2.2 *Předpokládejme, že chceme nalézt dobrou numerickou aproximaci hodnoty e s použitím vztahu* $e = \lim_{n\to\infty}(1+1/n)^n$, *kde limitu nahradíme prostým výpočtem hodnoty* $f(n) = (1+1/n)^n$ *pro dostatečně velké n. Pro* $n = 10$ *dostaneme v případě výpočtu v IEEE aritmetice s jednoduchou přesností lepší aproximaci čísla e než pro* $n = 10^7$ *(viz cvičení 2)! Příčina je následující. Sčítáme-li* $1+1/n$ *pro* $n \gg 1$, *obsahuje výsledek součtu stále méně a méně informace o čísle n (neboť* $1 \gg 1/n$*). I když provedeme následné umocnění přesně, výsledek je zatížen velkou chybou.*

Posledním příkladem je sčítání řad s kladnými členy.

Příklad 2.3 *Z teorie Fourierových řad je známo, že*

$$\sum_{k=1}^{\infty} k^{-2} = \pi^2/6.$$

Předpokládejme, že tuto identitu neznáme a chceme vypočítat součet řady numericky sčítáním

$$(\ldots((1+2^{-2})+3^{-2})+4^{-2}+\ldots)+m^{-2}),$$

kde m určíme jako nejmenší celé číslo, jehož zahrnutí do výpočtu již nezmění vypočtený součet. Výsledek výpočtu bude překvapivě nepřesný (viz cvičení 3). Příčina je opět zřejmá: řada konverguje velmi pomalu a náš výpočet je prováděn tak, že hodnota přičítaných prvků se stále zmenšuje. Pro jisté m je vypočtený částečný součet $\sum_{k=1}^{m-1} k^{-2}$ *takový, že přičtení* m^{-2} *nezmění jeho hodnotu; zbytek* $\sum_{k=m}^{\infty} 1/k^2$ *je však stále příliš velký. Jak překonat popsanou obtíž? První nápad může být změnit pořadí sčítání (sčítat od nejmenšího prvku k největšímu). Problém ovšem je, že nevíme, kterým prvkem začít. Navíc, uspořádání sčítanců je obecně drahá operace a nelze ji v praktických výpočtech použít. Univerzálním řešením je použití speciálních technik zvyšujících přesnost (samozřejmě na úkor rychlosti). Zvídavého čtenáře odkazujeme na [ASNA], kapitolu 4. Jiným řešením může být použití vhodné identity a řady konvergující podstatně rychleji, viz cvičení 5. V každém případě je vhodné zamyslet se nad konvergencí sčítané řady. Odstrašujícím případem budiž všem eventuální pokus nalézt výše popsaným postupem součet řady* $\sum_{k=1}^{\infty} \frac{1}{k}$.

Cvičení

1. Ukažte, jak je potřeba přepsat uvedené výrazy, aby byl omezen vliv krácení platných cifer

   (a) $\sqrt{x+1} - 1$ pro $x \sim 0$

   (b) $\sin x - \sin y$ pro $x \sim y$

   (c) $x^2 - y^2$ pro $x \sim y$

   (d) $(1 - \cos x)/\sin x$ pro $x \sim 0$

2. Vypočtěte hodnotu výrazu $(1+1/n)^n$ pro $n = 10^1, 10^2, \ldots, 10^7$ a srovnejte výsledky s hodnotou e.

3. Vypočtěte aproximaci součtu nekonečné řady $\sum_{k=1}^{\infty} k^{-2}$ podle příkladu v předcházejícím paragrafu. Určete chybu a udejte, kolik členů řady jste použili.

4. Vypočtěte $\sum_{n=1}^{\infty} \frac{1}{n^2+1}$ s přesností větší než 10^{-6}.
   K určení počtu členů m použijte $\int_m^{\infty} \frac{dx}{x^2+1}$.

5. Vypočtěte $\sum_{n=1}^{\infty} \frac{1}{n^2+1}$ s přesností větší než 10^{-6}.
   Nesčítejte původní řadu, ale použijte identit

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \quad \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}.$$

   K určení počtu členů použijte metodu analogickou cvičení 4. Sčítání provádějte od nejmenších členů k největším.

2.3 Přímá a zpětná stabilita

Vraťme se k úloze U se vstupními daty $(z_1, \ldots, z_m)$ z kapitoly 1. Nechť C označuje algoritmus pro řešení této úlohy. Označme $C(z_1, \ldots, z_m)$ výstup algoritmu C použitého na vstupní data $(z_1, \ldots, z_m)$ při (hypotetickém) výpočtu v přesné aritmetice, předpokládáme $C(z_1, \ldots, z_m) = U(z_1, \ldots, z_m)$. Výsledek odpovídajícího výpočtu v konečné aritmetice označíme jako $fl(C(z_1, \ldots, z_m))$. Zajímá nás chyba výpočtu způsobená zaokrouhlováním v aritmetice s konečnou přesností, tj. rozdíl

$$fl(C(z_1, \ldots, z_m)) - C(z_1, \ldots, z_m). \tag{2.6}$$

Při analýze chyb můžeme postupovat dvěma způsoby.

- *Přímá analýza chyb.* Postupujeme algoritmem a snažíme se odhadnout šíření elementárních zaokrouhlovacích chyb a na základě toho odhadnout přímo velikost výsledné chyby (2.6). Přímé určení odhadu chyby je však možné jen zřídka (v případě jednoduchých výpočtů jako je např. skalární součin vektorů, násobení matice vektorem, apod.).

- *Zpětná analýza chyb.* Hledáme nějaká data $(\tilde{z}_1, \ldots, \tilde{z}_m)$ tak, aby aproximace řešení původní úlohy $U(z_1, \ldots, z_m)$ získaná algoritmem C v konečné aritmetice byla totožná s řešením úlohy $U(\tilde{z}_1, \ldots, \tilde{z}_m)$ získaným algoritmem C při výpočtu v přesné aritmetice. Jinak řečeno: chceme určit vstupní data $(\tilde{z}_1, \ldots, \tilde{z}_m)$ taková, že platí

  $$fl(C(z_1, \ldots, z_m)) = C(\tilde{z}_1, \ldots, \tilde{z}_m).$$

  Cílem zpětné analýzy je interpretovat zaokrouhlovací chyby vzniklé při výpočtu v konečné přesnosti pomocí změn vstupních dat.

Příklad 2.4 *Předpokládejme, že čísla x, y, z jsou zobrazena v konečné aritmetice přesně. Pro součet v aritmetice s pohyblivou řádovou čárkou pak platí*

$$\begin{aligned} fl(fl(x+y)+z) &= [(x+y)(1+\delta_1)+z](1+\delta_2) \\ &= (x+y)(1+\delta_3)+z(1+\delta_2) \\ &= (\tilde{x}+\tilde{y})+\tilde{z}, \end{aligned}$$

kde jsme položili

$$(1+\delta_3) = (1+\delta_2)(1+\delta_1),$$

$|\delta_1| \leq u \ll 1$, $|\delta_2| \leq u \ll 1$. *Zřejmě tedy* $|\delta_3| \sim |\delta_1| + |\delta_2| \ll 1$ *a* $\tilde{x} = x(1+\delta_3)$, $\tilde{y} = y(1+\delta_3)$, $\tilde{z} = z(1+\delta_2)$ *jsou blízké hodnotě x, y a z. Vidíme, že výsledek součtu čísel x, y a z v konečné aritmetice je identický s výsledkem přesného součtu perturbovaných dat $\tilde{x}$, $\tilde{y}$ a $\tilde{z}$.*

Zpětná analýza chyb umožňuje redukovat otázku odhadu chyby řešení na otázku analýzy citlivosti dané úlohy. Pokud je výsledkem zpětné analýzy úloha s perturbovanými daty, pak pro výsledný odhad chyby řešení stačí použít výsledek analýzy citlivosti úlohy U na změny vstupních dat. Formálně zapsáno

$$\begin{aligned} fl(C(z_1,\ldots,z_m)) - C(z_1,\ldots,z_m) &= C(\tilde{z}_1,\ldots,\tilde{z}_m) - C(z_1,\ldots,z_m) \\ &= U(\tilde{z}_1,\ldots,\tilde{z}_m) - U(z_1,\ldots,z_m). \end{aligned}$$

Uvědomme si, že jde o velmi podstatnou věc. Lze totiž oddělit popis chování algoritmu vzhledem k zaokrouhlovacím chybám (popis numerické stability algoritmu) od popisu citlivosti řešené úlohy. Studium zpětné stability umožňuje poznat, kdy za velkou chybu řešení odpovídá špatná volba algoritmu (jeho nestabilita), a kdy je chyba jen nevyhnutelným důsledkem špatných vlastností samotné úlohy. Kapitolu ukončíme zavedením pojmu zpětné stability.

Definice 2.1 *Algoritmus C pro řešení úlohy $U(z_1,\ldots,z_m)$ nazveme* **zpětně stabilním**, *pokud existují data $(\tilde{z}_1,\ldots,\tilde{z}_m)$ tak, že platí*

$$fl(C(z_1,\ldots,z_m)) = C(\tilde{z}_1,\ldots,\tilde{z}_m)$$

a data $(\tilde{z}_1,\ldots,\tilde{z}_m)$ jsou v jistém smyslu blízká původním datům $(z_1,\ldots,z_m)$. Jinými slovy, algoritmus je zpětně stabilní, jestliže se chyby výpočtu způsobené zaokrouhlováním v průběhu algoritmu promítnou do malých změn vstupních dat.

Kapitola 3

Citlivost vlastních čísel matic

Půjde nám o následující otázku: nechť $A = (a_{ij})_{i,j=1,\ldots,N}$ je čtvercová komplexní matice a E je čtvercová matice stejného řádu, jejíž prvky jsou (v jistém smyslu) malé ve srovnání s prvky matice A. Matici E nazveme malou změnou (*perturbací*) matice A. Ptáme se, jaký je vztah mezi spektrem matice A a spektrem *perturbované matice* $A + E$. Jak uvidíme, odpověď závisí na vlastnostech matice A. Nejdříve proto popíšeme vlastnosti některých tříd matic a uvedeme tvrzení, která budeme při studiu citlivosti vlastních čísel potřebovat. Budeme používat následující označení.

Označení 3.1 *Pokud nebude uvedeno jinak, bude pro* $x \in C^N$ *symbol* $\| x \|$ *označovat* **euklidovskou normu** *vektoru*

$$\| x \| = \| x \|_2 = (\sum_{i=1}^{N} |x_i|^2)^{\frac{1}{2}},$$

generovanou skalárním součinem

$$(x, y) = y^* x = \sum_{i=1}^{N} x_i \overline{y}_i, \quad x, y \in C^N.$$

Pro **spektrální poloměr** *matice* $A \in C^{N,N}$ *budeme používat symbolu*

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|,$$

kde $\sigma(A)$ *je* **spektrum** *(tj. množina všech vlastních čísel) matice* A. **Spektrální normu** *(generovanou euklidovskou normou vektoru) pak označíme symbolem* $\| A \|$. *Platí pro ni*

$$\| A \| = \max_{\|x\|=1} \| Ax \| = \| A \|_2 = (\rho(A^* A))^{\frac{1}{2}}.$$

Symbolu A^* *používáme pro matici* **hermitovsky sdruženou** *(***konjugovanou***) s maticí* A,

$$A^* = \bar{A}^T.$$

Číslo podmíněnosti *regulární matice* $A \in C^{N,N}$ *je definováno vztahem*

$$\kappa(A) = \| A \| \| A^{-1} \| .$$

3.1 Schurova dekompozice, spektrální rozklad a Jordanův kanonický tvar

Důležitým nástrojem při studiu vlastních čísel obecné matice jsou podobnostní transformace, tj. transformace typu

$$A \to B = H^{-1}AH,$$

kde H je regulární matice stejného řádu jako matice A a H^{-1} označuje matici inverzní k matici H. Podobnostní transformace zachovává vlastní čísla; cílem je přitom převést původní matici na tvar, z něhož lze vlastní čísla snadno získat (v kterém jsou například totožná s diagonálními prvky). Výsledný tvar matice B je závislý na vlastnostech původní matice A a může být pro různé třídy matic různý. Vlastnosti matice A také charakterizují matici H, která realizuje příslušnou podobnostní transformaci.

Uvažujme případ, kdy matice A je zatížena chybami, ve skutečnosti tedy provádíme podobnostní transformaci matice $A + E$, kde E označuje matici chyb. Co se stane s velikostí chyb při podobnostní transformaci? S použitím odhadu

$$\begin{aligned} \| H^{-1}AH - H^{-1}(A+E)H \| &= \| H^{-1}EH \| \\ &\leq \kappa(H) \| E \| \end{aligned}$$

vidíme, že je-li hodnota $\kappa(H)$ velká, může podobnostní transformace velmi podstatně zvětšit chyby obsažené ve vstupních datech. Ideální by proto bylo používat pouze ty podobnostní tranformace, které nám velikost chyb nezvětší. Příkladem jsou transformace unitární.

Definice 3.1 *Řekneme, že matice $U \in C^{N,N}$ je* **unitární**, *jestliže*

$$U^*U = UU^* = I,$$

kde I je jednotková matice.

Poznámka 3.1 *Všimněme si, že euklidovská norma vektoru a spektrální norma matice jsou invariantní vzhledem k unitárním transformacím. Pro obecnou matici $A \in C^{N,N}$, unitární matici $U \in C^{N,N}$ a vektor $x \in C^N$ platí*

$$\begin{aligned} \|U^*AU\| &= \|A\| \\ \|Ux\| &= \|x\|. \end{aligned}$$

Omezme se na chvíli pouze na unitární podobnostní transformace. Cílem podobnostní transformace by měla být matice v co nejjednodušším tvaru - matice diagonální. Bohužel, ne každou matici lze unitární podobnostní transformací převést na matici diagonální. Každou matici však lze užitím unitárních podobnostních transformací převést na matici horní trojúhelníkovou.

Věta 3.1 *(Schur) Pro libovolnou matici $A \in C^{N,N}$ existuje unitární matice $U \in C^{N,N}$ tak, že matice $R = U^*AU$ je horní trojúhelníková. Matice U může být zvolena tak, aby diagonála matice R obsahovala vlastní čísla matice A v libovolném předepsaném pořadí.*

Důkaz: Důkaz provedeme indukcí podle řádu matice A. Pro matice řádu 1 je platnost tvrzení zřejmá. Předpokládejme, že tvrzení věty platí pro všechny matice až do řádu n včetně. Nechť $A \in C^{n+1,n+1}$, nechť je dáno uspořádání vlastních čísel matice A. Označme λ první vlastní číslo v tomto uspořádání. Bez újmy obecnosti předpokládejme, že příslušný vlastní vektor je normovaný. Tedy platí

$$Ax = \lambda x, \qquad \|x\| = 1.$$

Definujme čtvercovou unitární matici $H \in C^{n+1,n+1}$ následujícím způsobem

$$H = \begin{pmatrix} x & X \end{pmatrix},$$

kde $x \in C^{n+1}$ a $X \in C^{n+1,n}$. Pro matici $A \in C^{n+1,n+1}$ pak platí

$$H^*AH = \begin{pmatrix} x^*Ax & x^*AX \\ X^*Ax & X^*AX \end{pmatrix} = \begin{pmatrix} \lambda & b^* \\ 0 & M \end{pmatrix},$$

neboť X^*Ax je nulový vektor.

H^*AH je horní blokově trojúhelníková matice se čtvercovými diagonálními bloky, tedy množina jejích vlastních čísel je rovna sjednocení množin vlastních čísel těchto bloků. (cvičení 9). Tudíž $\sigma(M) = \sigma(A)\backslash\{\lambda\}$. Podle indukčního předpokladu existuje unitární matice V taková, že V^*MV je horní trojúhelníková s vlastními čísly v předepsaném pořadí. Položíme-li

$$U = \begin{pmatrix} x & XV \end{pmatrix},$$

pak

$$R = U^*AU = \begin{pmatrix} \lambda & b^*V \\ 0 & V^*MV \end{pmatrix}$$

je hledaný rozklad.

□

Definice 3.2 *Rozklad $A = URU^*$ z věty 3.1 budeme nazývat* **Schurovým rozkladem (dekompozicí)** *matice A, matici R nazveme výsledkem* **Schurovy transformace** *matice A.*

Schurova věta je nejen velice silným teoretickým nástrojem, ale má zásadní význam i při praktickém řešení problému vlastních čísel. Určení Schurova rozkladu je cílem QR algoritmu (krásný výklad QR algoritmu nalezne čtenář v [FMC]).

Uvedeme několik důležitých důsledků Schurovy věty. Nejprve připomeneme definici normální matice.

Definice 3.3 *Řekneme, že matice $A \in C^{N,N}$ je* **normální**, *platí-li $A^*A = AA^*$, tj. matice komutuje se svojí maticí hermitovsky sdruženou.*

Věta 3.2 *Nechť matice $A \in C^{N,N}$ je normální. Pak výsledkem její Schurovy transformace je diagonální matice.*

Důkaz: Tvrzení dokážeme opět indukcí podle řádu matice A. Pro $n = 1$ je platnost tvrzení zřejmá. Nechť tvrzení platí až do n včetně. Pro $A \in C^{n+1,n+1}$ označme výsledek Schurovy transformace

$$R = U^*AU = \begin{pmatrix} \rho & r^T \\ 0 & R_1 \end{pmatrix}, \tag{3.1}$$

kde $\rho \in C^1$, $r \in C^n$ a $R_1 \in C^{n,n}$ je horní trojúhelníková matice. Z definice normální matice pak s použitím $U^*U = UU^* = I$ dostaneme

$$R^*R = RR^*,$$

tedy R je rovněž normální matice. Dosazením z výrazu (3.1) můžeme psát

$$\begin{pmatrix} \bar{\rho} & 0 \\ \bar{r} & R_1^* \end{pmatrix} \begin{pmatrix} \rho & r^T \\ 0 & R_1 \end{pmatrix} = \begin{pmatrix} \rho & r^T \\ 0 & R_1 \end{pmatrix} \begin{pmatrix} \bar{\rho} & 0 \\ \bar{r} & R_1^* \end{pmatrix}.$$

Tedy musí platit

$$|\rho|^2 = |\rho|^2 + r^T\bar{r}, \tag{3.2}$$

z čehož plyne $r = 0$. V důsledku toho platí

$$R_1^*R_1 = R_1R_1^*. \tag{3.3}$$

Z rovnice (3.3) vyplývá, že R_1 je normální matice. R_1 je přitom Schurovou dekompozicí sebe sama. S použitím indukčního předpokladu je R_1 a tedy i matice R diagonální.

□

Pro vlastní vektory normální matice platí následující důležitá věta.

Věta 3.3 *Normalizované vlastní vektory normální matice $A \in C^{N,N}$ tvoří ortonormální bazi prostoru C^N.*

Důkaz: Schurovu dekompozici normální matice lze zapsat ve tvaru

$$U^*AU = \Lambda = diag(\lambda_1, \ldots, \lambda_N).$$

Protože U je unitární, je tento zápis ekvivalentní formulaci

$$AU = U\Lambda. \tag{3.4}$$

Označíme-li $u_1, \ldots, u_N$ sloupce matice U, $U = (u_1, \ldots u_N)$, dostáváme z rovnice (3.4)

$$Au_j = \lambda_j u_j \quad j = 1, \ldots, N\ ,$$

neboli $u_1, \ldots, u_N$ jsou vlastní vektory a $\lambda_1, \ldots, \lambda_N$ jsou příslušná vlastní čísla matice A.

□

Poznámka 3.2 *Zřejmě každá matice* $A = U\ diag(\lambda_1, \ldots, \lambda_N)\ U^*$, $UU^* = UU^* = I$, *je maticí normální.*

Důležitou třídou normálních matic jsou matice hermitovské.

Definice 3.4 *Řekneme, že matice* $A \in C^{N,N}$ *je* **hermitovská**, *jestliže platí* $A^* = A$.

Věta 3.4 *Matice* $A \in C^{N,N}$ *je unitární tehdy a jen tehdy, je-li normální a její vlastní čísla leží na jednotkové kružnici. Matice* $A \in C^{N,N}$ *je hermitovská tehdy a jen tehdy, je-li normální a všechna její vlastní čísla jsou reálná.*

Důkaz: Unitární matice je zřejmě normální. Nechť λ je vlastní číslo unitární matice A a x je příslušný vlastní vektor. Pak platí

$$\| x \|^2 = \| Ax \|^2 = (Ax)^*(Ax) = (\lambda x)^*(\lambda x) = |\lambda|^2 \| x \|^2,$$

z čehož plyne $|\lambda| = 1$.
Předpokládejme nyní, že A je normální s vlastními čísly na jednotkové kružnici. Užitím věty 3.2 pak dostaneme:

$$A^*A = U\Lambda^*U^*U\Lambda U^* = I = U\Lambda U^*U\Lambda^*U^* = AA^*.$$

Hermitovská matice je zřejmě normální. Nechť λ je vlastní číslo hermitovské matice A a x je příslušný vlastní vektor

$$\lambda \| x \|^2 = x^*Ax = (Ax)^*x = \bar{\lambda} \| x \|^2,$$

tedy λ musí být reálné číslo.
Předpokládejme nyní, že A je normální a má reálná vlastní čísla. Pak platí:

$$A^* = U\Lambda^*U^* = U\Lambda U^* = A.$$

□

Poznámka 3.3 (Spektrální rozklad) *Pro hermitovskou matici* $A \in C^{N,N}$ *platí*

$$A = U\Lambda U^* = \sum_{i=1}^{N} \lambda_i u_i u_i^*,$$

kde $\lambda_i \in \sigma(A)$ *a* $u_1, \ldots, u_N$ *jsou sloupce unitární matice* $U \in C^{N,N}$ *sestavené z normalizovaných vlastních vektorů matice A. Tento zápis umožňuje zavést pojem funkce hermitovské matice.*

Definice 3.5 *Nechť $A \in C^{N,N}$ je hermitovská matice, $\lambda_1, \ldots, \lambda_N$ její vlastní čísla a $u_1, \ldots, u_N$ odpovídající normalizované vlastní vektory. Je-li Φ reálná funkce reálné proměnné, definujeme funkci $\Phi(A)$ vztahem*

$$\Phi(A) \stackrel{def}{=} \sum_{i=1}^{N} \Phi(\lambda_i) u_i u_i^*. \tag{3.5}$$

Příklad 3.1 *Pro hermitovskou pozitivně semidefinitní matici $A \in C^{N,N}$ můžeme psát*

$$A^{\frac{1}{2}} \stackrel{def}{=} \sum_{i=1}^{N} \lambda_i^{\frac{1}{2}} u_i u_i^* = U \Lambda^{\frac{1}{2}} U^*.$$

Pro obecnější zavedení funkce matice odkazujeme na literaturu popsanou v úvodu.

Viděli jsme, že třída normálních matic je totožná s třídou všech matic, které lze unitární podobnostní transformací převést na diagonální tvar. Nepožadujeme-li, aby matice realizující podobnostní transformaci byla unitární, dostaneme třídu diagonalizovatelných matic.

Definice 3.6 *Matici $A \in C^{N,N}$ nazveme* **diagonalizovatelnou**, *jestliže existuje regulární matice $X \in C^{N,N}$ taková, že $X^{-1}AX = diag(\lambda_1, \ldots, \lambda_N)$.*

Bohužel, ne každá čtvercová matice je diagonalizovatelná. Je-li matice diagonalizovatelná, pak její vlastní vektory tvoří bazi prostoru C^N. Tato baze však může mít špatné numerické vlastnosti, neboť odpovídající matice X může být špatně podmíněná (tj. číslo $\kappa(X) = \| X \| \| X^{-1} \|$ může být velké). Není-li matice diagonalizovatelná, znamená to, že nemá dost vlastních vektorů k vytvoření baze celého prostoru C^N. Jak uvidíme, tento defekt může hrát velmi podstatnou roli. Proto se matice, které nejsou diagonalizovatelné, někdy nazývají *defektními* (defective), zatímco matice diagonalizovatelné se nazývají *jednoduchými* (simple). Můžeme se ptát, lze-li každou matici alespoň aproximovat pomocí matic diagonalizovatelných. Odpověď dává následující věta, která říká, že třída diagonalizovatelných matic je hustá v $C^{N,N}$.

Věta 3.5 *Nechť $A \in C^{N,N}$. Pro každé $\epsilon > 0$ existuje diagonalizovatelná matice $A_\epsilon \in C^{N,N}$ tak, že $\| A - A_\epsilon \| < \epsilon$.*

Důkaz: Uvažujme Schurovu dekompozici matice A, $A = URU^*$. Není-li matice A diagonalizovatelná, musí mít alespoň jedno násobné vlastní číslo (vlastní vektory příslušné různým vlastním číslům jsou lineárně nezávislé). Vlastní čísla matice A leží na diagonále matice R. Stačí tedy nalézt takovou diagonální matici D_ϵ, aby vlastní čísla matice $R_\epsilon = R + D_\epsilon$ byla navzájem různá a $\| D_\epsilon \| < \epsilon$. To je zřejmě vždy možné.

□

Na základě věty 3.5 bychom mohli učinit následující úvahu: při analýze citlivosti se stačí omezit pouze na třídu diagonalizovatelných matic, neboť každou matici vně této třídy

umíme libovolně přesně aproximovat maticí diagonalizovatelnou. Tato úvaha je však, jak uvidíme dále, nesprávná. Existují totiž matice, u nichž i nepatrná změna jejich prvků může vyvolat velmi podstatnou změnu vlastních čísel a vlastních vektorů.

Pro úplnost zbývá uvést, do jakého tvaru (co nejbližšího matici diagonální) lze podobnostní transformací převést obecnou matici. Větu uvádíme bez důkazu (zvídavého a trpělivého čtenáře odkazujeme na [MA]).

Věta 3.6 *(Jordan) Pro každou matici $A \in C^{N,N}$ existuje regulární matice $X \in C^{N,N}$ tak, že platí*

$$X^{-1}AX = diag(J_{n_1}(\lambda_1), J_{n_2}(\lambda_2), \ldots, J_{n_l}(\lambda_l)), \tag{3.6}$$

kde matice na pravé straně je blokově diagonální a $J_{n_k}(\lambda_k) \in C^{n_k,n_k}$ je **Jordanův blok** *tvaru*

$$J_{n_k}(\lambda_k) = \begin{pmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{pmatrix},$$

explicitně neuvedené prvky matice jsou nulové, $n_1+n_2+\ldots+n_l = N$. Pravá strana výrazu (3.6) je nazývána **Jordanovým kanonickým tvarem matice** *A. Je jednoznačně určena až na uspořádání bloků. Vlastní čísla λ_k, $k = 1, \ldots, l$ nemusí být navzájem různá.*

Práce s Jordanovým kanonickým tvarem je obtížná z několika důvodů (například libovolně malé perturbace mohou zcela změnit strukturu Jordanových bloků). Všimněme si, že jedničky na vedlejší diagonále představují vlastně výsledek jisté normalizace. Například transformace jednoduchého bloku

$$\begin{pmatrix} \delta & & \\ & \delta^2 & \\ & & \delta^3 \end{pmatrix}^{-1} \begin{pmatrix} \lambda & 1 & \\ & \lambda & 1 \\ & & \lambda \end{pmatrix} \begin{pmatrix} \delta & & \\ & \delta^2 & \\ & & \delta^3 \end{pmatrix} = \begin{pmatrix} \lambda & \delta & \\ & \lambda & \delta \\ & & \lambda \end{pmatrix}$$

vynásobí vedlejší diagonálu hodnotou δ. Pokud $\delta \to 0$, stává se transformace velmi špatně podmíněnou.

Cvičení

1. Ukažte, že podobnostní transformace zachovává vlastní čísla. Co platí pro vlastní vektory podobných matic?

2. Dokažte, že pro unitární matici $U \in C^{N,N}$ a euklidovskou resp. spektrální normu platí

$$\| Ux \| = \| x \|$$

resp.

$$\| U^*AU \| = \| A \|,$$

kde $x \in C^N$, $A \in C^{N,N}$.

3. Uvědomte si, jaké vztahy platí mezi třídami diagonalizovatelných, normálních, unitárních a hermitovských matic.

4. Proč sloupce unitární matice řádu N tvoří ortonormální bazi v $C^{N,N}$?

5. Nechť $\| x \|_\alpha$ značí libovolnou normu vektoru $x \in C^N$. Chápeme-li matici $A \in C^{N,N}$ jako operátor z C^N do C^N, zavedeme operátorovou normu

$$\| A \|_\alpha = \max_{\|x\|_\alpha = 1} \| Ax \|_\alpha \ .$$

   Uveďte příklady takto definovaných norem v $C^{N,N}$ odlišných od $\| \ . \ \|_2$.

6. Lze každou maticovou normu v $C^{N,N}$ definovat jako operátorovou normu? Uvažte příklad Frobeniovy normy $\| A \|_F = \left(\sum_{i,j=1}^{N} |a_{ij}|^2 \right)^{1/2}$ a volte vhodnou matici A.

7. Maticová norma se nazývá *konzistentní*, pokud $\| AB \|_\alpha \leq \| A \|_\alpha \| B \|_\alpha$ pro libovolné dvě matice $A, B \in C^{N,N}$. Jaké konzistentní maticové normy znáte?

8. Dokažte, že pro libovolnou $A \in C^{N,N}$ a pro spektrální normu platí

$$\rho(A) \leq \| A \| \ . \tag{3.7}$$

   Platí vztah (3.7) i pro jiné maticové normy?

9. Ukažte, že spektrum blokově trojúhelníkové matice je sjednocením spekter diagonálních bloků.

10. Jak vypadají matice, které sčítáme ve výraze (3.5)?

11. Schurovu dekompozici nelze obecně nalézt žádným konečným algoritmem (např. typu Gaussovy eliminace či QR rozkladu). Proč?

3.2 Citlivost vlastních čísel obecných matic

Začneme příkladem ukazujícím význam vlastností matic v teorii citlivosti vlastních čísel.

Příklad 3.2 *Budeme vyšetřovat dvě následující matice A_0, A_1, které mají totožné spektrum $\sigma(A_0) = \sigma(A_1) = \{0\}$,*

$$A_0 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad A_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Uvažujme perturbaci E

$$E = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \epsilon & 0 & 0 & 0 \end{pmatrix},$$

kde ϵ je malé kladné číslo. Vlastní čísla matice $A_0 + E$ se neliší od vlastních čísel matice A_0. Snadno vypočteme spektrum matice $A_1 + E$, $\sigma(A_1 + E) = \{\epsilon^{\frac{1}{4}}, -\epsilon^{\frac{1}{4}}, i\epsilon^{\frac{1}{4}}, -i\epsilon^{\frac{1}{4}}\}$. Zvolme nyní $\epsilon = 10^{-8}$. Zatímco vlastní čísla matice A_0 se od vlastních čísel perturbované matice $A_0 + E$ neliší, u vlastních čísel matice A_1 způsobí stejná perturbace odchylku vlastních čísel řádu 10^{-2}.

V celém zbytku kapitoly 3 budeme používat následující označení.

Označení 3.2 *Nechť matice $A \in C^{N,N}$. Její malou změnu (perturbaci) označme E, $E \in C^{N,N}$, perturbovanou matici budeme značit $\tilde{A}$, $\tilde{A} = A + E$. Vlastní čísla matice A pak budeme označovat $\lambda_1, \ldots, \lambda_N$, vlastní čísla perturbované matice $\tilde{A}$ budeme označovat $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N$, příslušné charakteristické polynomy $\varphi_A(\lambda)$, resp. $\varphi_{\tilde{A}}(\lambda)$.*

3.2.1 Spojitost vlastních čísel

Nejprve ukážeme, že vlastní čísla jsou spojitou funkcí prvků matice.

Věta 3.7 *Nechť matice $A \in C^{N,N}$, λ je její vlastní číslo s algebraickou násobností m. Pak pro každé dostatečně malé $\epsilon > 0$ existuje $\delta > 0$ tak, že pokud je $\| E \| < \delta$, pak kruh*

$$D(\lambda, \epsilon) = \{\zeta \in C; |\zeta - \lambda| \leq \epsilon\}$$

obsahuje právě m vlastních čísel matice $\tilde{A} = A + E$.

Důkaz: Zvolme $\epsilon > 0$ tak, aby $D(\lambda, \epsilon)$ neobsahoval žádné další vlastní číslo matice A. Označme $\eta(\zeta) = \varphi_{\tilde{A}}(\zeta) - \varphi_A(\zeta)$. Hranice kruhu $D(\lambda, \epsilon)$ je kompaktní množina v C, označíme ji ∂D. Charakteristický polynom je spojitou funkcí prvků matice; proto funkce $\eta(\zeta)$ konverguje k nule na kompaktu ∂D pro $\tilde{A} \to A$. Zároveň $\varphi_A(\zeta) \neq 0$ pro $\forall \zeta \in \partial D$; jistě tedy existuje takové číslo $\delta > 0$, že pro $\| E \| < \delta$ platí

$$|\eta(\zeta)| < |\varphi_A(\zeta)| \quad \forall \zeta \in \partial D. \tag{3.8}$$

Nyní použijeme Rouchéovu větu ve tvaru, který je uveden např. v [MPT], str. 167. Funkce φ_A a η jsou analytické v celé množině C. Pak z (3.8) vyplývá, že φ_A a $\varphi_{\tilde{A}} = \eta + \varphi_A$ mají v kruhu $D(\lambda, \epsilon)$ stejný počet nulových bodů.

□

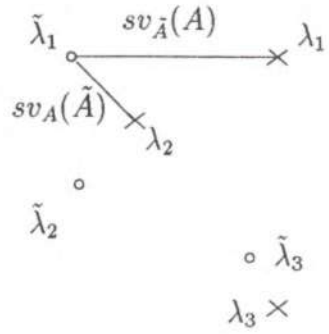
Poznámka 3.4 *Uvědomme si, že věta 3.7 nám nedává žádný kvantitativní odhad pro změnu vlastních čísel. Věta 3.7 neříká ani to, že malá perturbace prvků matice způsobí malou perturbaci vlastních čísel!*

3.2.2 Elsnerova a Ostrowského-Elsnerova věta

Dříve než zformulujeme základní věty teorie citlivosti vlastních čísel pro obecné matice, musíme umět popsat vzájemnou polohu spekter matic A a $\tilde{A}$.

Definice 3.7 *Nechť $A \in C^{N,N}$, $E \in C^{N,N}$, $\tilde{A} = A + E$.* **Spektrální variací** *matice $\tilde{A}$ vzhledem k matici A nazveme*

$$sv_A(\tilde{A}) \stackrel{def}{=} \max_i(\min_j |\tilde{\lambda}_i - \lambda_j|).$$



Obr. 3.1: Spektrální variace $sv_A(\tilde{A})$ a $sv_{\tilde{A}}(A)$

Poznámka 3.5 *Všimněme si, jaký má spektrální variace geometrický význam. Definujeme-li totiž*

$$\tilde{D}_i = \{\zeta; |\zeta - \lambda_i| \leq sv_A(\tilde{A})\}$$

pro $i = 1, \ldots, N$, pak

$$\sigma(\tilde{A}) \subset \bigcup_{i=1}^{N} \tilde{D}_i.$$

Tedy všechna vlastní čísla matice $\tilde{A}$ leží ve sjednocení kruhů se středy ve vlastních číslech matice A a poloměrem $sv_A(\tilde{A})$.

Spektrální variace má velmi nepříjemnou vlastnost. Není symetrická ($sv_A(\tilde{A}) \neq sv_{\tilde{A}}(A)$) a není tudíž metrikou. Obrázek 3.1 je příkladem rozložení spekter matic A a $\tilde{A}$ řádu 3, kdy je $sv_{\tilde{A}}(A) > sv_A(\tilde{A})$. Vlastní čísla matice A jsou označena křížky, vlastní čísla matice $\tilde{A}$ kroužky.

Definice 3.8 *Nechť $A \in C^{N,N}$, $E \in C^{N,N}$, $\tilde{A} = A + E$.* **Hausdorffovou vzdáleností** *spekter matic A a $\tilde{A}$ nazveme*

$$hd(A, \tilde{A}) \stackrel{def}{=} \max(sv_A(\tilde{A}), sv_{\tilde{A}}(A)).$$

Hausdorffova vzdálenost již definuje metriku v $C^{N,N}$, stále však nám nedává názorný pojem vzdálenosti vlastních čísel matic A a $\tilde{A}$. Proto je výhodné používat následující definici.

Definice 3.9 *Nechť $A \in C^{N,N}$, $E \in C^{N,N}$, $\tilde{A} = A + E$.* **Párovou (optimální) vzdáleností** *spekter matic A a $\tilde{A}$ nazveme*

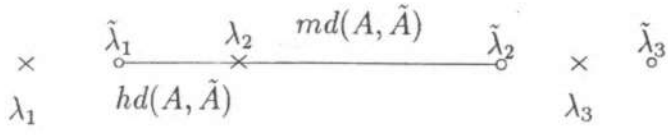
$$md(A, \tilde{A}) \stackrel{def}{=} \min_{\pi}(\max_{i} |\tilde{\lambda}_{\pi(i)} - \lambda_i|),$$

kde π probíhá všechny permutace množiny $\{1, \ldots, N\}$.

Podaří-li se nám ukázat, že párová vzdálenost matic $md(A, \tilde{A})$ je malá, znamená to, že vlastní čísla matic A a $\tilde{A}$ jsou uspořádána v párech tvořených blízkými vlastními čísly. To nám dovoluje názorně si představit změny jednotlivých vlastních čísel. Mezi pojmy definovanými výše platí následující vztah:

$$sv_A(\tilde{A}) \leq hd(A, \tilde{A}) \leq md(A, \tilde{A}).$$

Příklad 3.3 *Situace, kdy $hd(A, \tilde{A}) < md(A, \tilde{A})$ je znázorněna na obrázku 3.2.*



Obr. 3.2: Hausdorffova a optimální vzdálenost spekter matic A a $\tilde{A}$

Dříve než uvedeme Elsnerovu větu, která dává odhad velikosti Hausdorffovy vzdálenosti spekter matic A a $\tilde{A}$, dokážeme pomocné tvrzení.

Lemma 3.1 *(Hadamard) Označme $a_1, \ldots, a_N$ sloupce matice $A \in C^{N,N}$. Pak platí*

$$|det(A)| \leq \Pi_{j=1}^{N} \parallel a_j \parallel$$

a rovnost nastává právě když A má nulový sloupec nebo její sloupce jsou navzájem ortogonální.

Důkaz: Podle věty o QR rozkladu (viz např. [FMC]) víme, že každou komplexní matici lze rozložit na součin matice unitární a matice horní trojúhelníkové. Tedy platí

$$A = UR \; ; \quad U^*A = R, \tag{3.9}$$

kde $U^*U = UU^* = I$ a R je horní trojúhelníková. Označme $r_1, \ldots, r_N$ sloupce matice R a $\rho_{11}, \ldots, \rho_{NN}$ její diagonální prvky. Pro determinant matice A platí

$$|det(A)| = |det(UR)| = |det(R)|.$$

Pro determinant matice R dalšími úpravami dostáváme

$$|det(R)| = \Pi_{j=1}^{N} |\rho_{jj}| \leq \Pi_{j=1}^{N} \parallel r_j \parallel = \Pi_{j=1}^{N} \parallel U^* a_j \parallel = \Pi_{j=1}^{N} \parallel a_j \parallel \tag{3.10}$$

(bylo použito vyjádření (3.9) pro matici R a invariance euklidovské normy vzhledem k násobení unitární maticí).

Důkaz druhé části tvrzení vyplývá přímo ze vztahů (3.10). Pokud má A nulový sloupec, má nulový determinant. Protože poslední výraz v (3.10) je roven nule, zřejmě nastává rovnost. Má-li matice A navzájem ortogonální sloupce, musí být matice R diagonální a v (3.10) opět nastává rovnost. Opačně, aby platila mezi vztahy v (3.10) rovnost, musí být splněno

$$\Pi_{j=1}^{N} |\rho_{jj}| = \Pi_{j=1}^{N} \parallel r_j \parallel .$$

To zřejmě nastane, je-li buď nějaký sloupec r_j nulový (pak je nulový i j-tý sloupec matice A), nebo pokud je $|\rho_{jj}| = \parallel r_j \parallel$ pro všechna $j = 1, \ldots, N$ (sloupce matice A jsou ortogonální).

□

Věta 3.8 *(Elsner) Nechť $A \in C^{N,N}$, $E \in C^{N,N}$ a $\tilde{A} = A + E$. Pak pro Hausdorffovu vzdálenost spekter matic A, $\tilde{A}$ platí*

$$hd(A, \tilde{A}) \leq (\parallel A \parallel + \parallel \tilde{A} \parallel)^{1-\frac{1}{N}} \parallel E \parallel^{\frac{1}{N}} . \tag{3.11}$$

Důkaz: Jelikož pravá strana nerovnosti (3.11) je symetrická v A, $\tilde{A}$, stačí dokázat, že odhad platí pro $sv_A(\tilde{A})$.

Předpokládejme, že $\tilde{\lambda}$ je to vlastní číslo, které realizuje maximum v definici spektrální variace $\tilde{A}$ vzhledem k A. Vezměme příslušný normovaný vlastní vektor x_1 matice $\tilde{A}$ a doplňme ho dalšími vektory $x_2, \ldots, x_N$ tak, aby výsledná matice $X = (x_1, \ldots, x_N)$ byla unitární. Pro N-tou mocninu spektrální variace $\tilde{A}$ vzhledem k A pak platí

$$\begin{aligned} (sv_A(\tilde{A}))^N &\leq \Pi_{l=1}^{N} |\tilde{\lambda} - \lambda_l| = |det(A - \tilde{\lambda} I)| = |det[(A - \tilde{\lambda} I)X]| \\ &\leq \Pi_{i=1}^{N} \parallel (A - \tilde{\lambda} I) x_i \parallel = \parallel (A - \tilde{\lambda} I) x_1 \parallel \Pi_{i=2}^{N} \parallel (A - \tilde{\lambda} I) x_i \parallel \; . \end{aligned} \tag{3.12}$$

Poslední nerovnost ve výraze (3.12) jsme dostali užitím Hadamardova lemmatu. Protože $\tilde{A} = A + E$, $\tilde{\lambda}$ je vlastní číslo matice $\tilde{A}$ příslušné vlastnímu vektoru x_1 a $\| x_1 \| = 1$, platí

$$\| (A - \tilde{\lambda} I)x_1 \| \leq \| E \| .$$

Ostatní členy v součinu na pravé straně výrazu (3.12) odhadneme následujícím způsobem:

$$\| (A - \tilde{\lambda} I)x_i \| \leq \| A - \tilde{\lambda} I \| \leq \| A \| + |\tilde{\lambda}| \leq \| A \| + \| \tilde{A} \|$$

(při odhadech jsme využili nerovnosti $\rho(\tilde{A}) \leq \| \tilde{A} \|$). Dosazením do (3.12) získáme hledaný odhad

$$(sv_A(\tilde{A}))^N \leq \| E \| (\| A \| + \| \tilde{A} \|)^{N-1}.$$

□

Elsnerova věta dává odhad pro Hausdorffovu vzdálenost spekter matic A a $\tilde{A}$. Obdobný odhad odvodíme i pro párovou vzdálenost. Její hodnota vypovídá totiž o vzájemné poloze spekter matic A a $\tilde{A}$ nejvíce. I když znění věty bude až na násobek členem nezávislým na A, E stejné jako u Elsnerovy věty, důkaz je mnohem náročnější. Přejít od Hausdorffovy k párové vzdálenosti není snadné.

Použijeme následující užitečnou techniku. Nechť $A \in C^{N,N}$, $E \in C^{N,N}$ jsou dané matice, $\tilde{A} = A + E$. Budeme se zabývat vlastnostmi matice $A + \tau E$, kde $0 \leq \tau \leq 1$. Označíme

$$\mu \stackrel{def}{=} (2 \max_{\tau \in <0,1>} \| A + \tau E \|)^{1 - \frac{1}{N}}, \quad \gamma = \mu \| E \|^{1/N} .$$

Pak zřejmě platí $\mu \geq (\| A \| + \| \tilde{A} \|)^{1 - \frac{1}{N}}$ a z Elsnerovy věty plyne $sv_A(\tilde{A}) \leq \gamma$. Spektrum matice $\tilde{A}$ jistě leží ve sjednocení kruhů $D_i = \{\zeta \in C; |\zeta - \lambda_i| \leq \gamma\}$, $i = 1, \ldots, N$;

$$\sigma(\tilde{A}) \subset \bigcup_{i=1}^{N} D_i.$$

Dále platí

$$\| A + \tau E \| = \| A + \tau(\tilde{A} - A) \| \leq (1 - \tau) \| A \| + \tau \| \tilde{A} \| \leq \| A \| + \| \tilde{A} \|,$$

z čehož plyne

$$\mu \leq 2 \left(\| A \| + \| \tilde{A} \| \right)^{1 - 1/N}, \quad \gamma \leq 2\, \delta, \quad \delta \equiv \delta(A, \tilde{A}) = \left(\| A \| + \| \tilde{A} \| \right)^{1 - 1/N} \| E \|^{1/N} .$$

Při odhadu optimální vzdálenosti spekter matic A a $\tilde{A}$ využijeme následující důležité tvrzení.

Lemma 3.2 *Nechť libovolné sjednocení m výše popsaných kruhů D_i má s ostatními "kruhy prázdný průnik. Pak toto sjednocení obsahuje právě m vlastních čísel matice $\tilde{A}$.*

Důkaz: Bez újmy obecnosti předpokládejme, že $\bigcup_{i=1}^{m} D_i$ má s kruhy $D_{m+1}, \ldots, D_N$ prázdný průnik. Protože $\bigcup_{i=1}^{m} D_i$ je uzavřená množina, je

$$C \setminus \bigcup_{i=1}^{m} D_i \setminus \bigcup_{i=m+1}^{N} D_i$$

otevřená množina a tudíž $\bigcup_{i=1}^{m} D_i$ je od ostatních disků izolována. Označme

$$\tilde{A}_\tau = \tau\tilde{A} + (1-\tau)A = A + \tau E,$$

kde $\tau \in < 0, 1 >$,

$$D_i^\tau = \{\zeta \in C; |\zeta - \lambda_i| \leq \mu \parallel \tau E \parallel^{\frac{1}{N}}\}.$$

Použitím Elsnerovy věty dostáváme

$$sv_A(\tilde{A}_\tau) \leq \mu \parallel \tau E \parallel^{\frac{1}{N}} = \gamma\, \tau^{\frac{1}{N}}.$$

Dále víme, že

$$\sigma(\tilde{A}_\tau) \subset \bigcup_{i=1}^{N} D_i^\tau.$$

Podle předpokladu je

$$\bigcup_{i=1}^{m} D_i^1 = \bigcup_{i=1}^{m} D_i$$

izolována od ostatních $N-m$ kruhů. Funkce $\gamma\tau^{\frac{1}{N}}$ je pro $\tau \in < 0, 1 >$ monotonně rostoucí. Tedy

$$\bigcup_{i=1}^{m} D_i^\tau \tag{3.13}$$

je izolována od ostatních disků pro každé $\tau \in < 0, 1 >$. Sjednocení $\bigcup_{i=1}^{m} D_i^0$ obsahuje právě m vlastních čísel matice $\tilde{A}_0 = A$. Zkonstruujme posloupnost matic $\tilde{A}_0, \tilde{A}_{\tau_1}, \ldots, \tilde{A}_{\tau_m}, \ldots$, kde $0 < \tau_1 < \ldots < \tau_m < \ldots < 1$ tak, aby

$$\lim_{i\to\infty} A_{\tau_i} = \tilde{A}_1.$$

Protože vlastní čísla jsou spojitou funkcí prvků matice, konvergují i příslušná vlastní čísla

$$\lim_{i\to\infty} \lambda_j(\tilde{A}_{\tau_i}) = \lambda_j(\tilde{A}_1) \text{ pro } j = 1, \ldots, m.$$

A tato limita musí vzhledem k izolovanosti (3.13) ležet v $\bigcup_{i=1}^{m} D_i^1$.

□

Konečně jsme připraveni vyslovit a dokázat slíbenou větu.

Věta 3.9 *(Ostrowski,Elsner) Nechť $A \in C^{N,N}$, $E \in C^{N,N}$ a $\tilde{A} = A + E$. Pak pro párovou vzdálenost spekter matic A, $\tilde{A}$ platí*

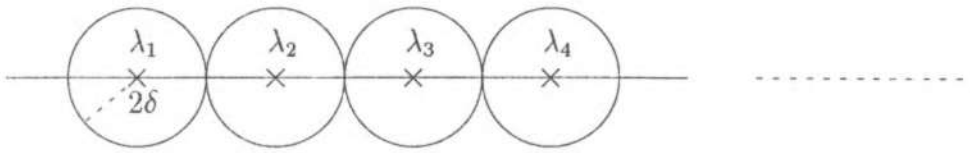
$$md(A, \tilde{A}) \leq (2N-1)(\| A \| + \| \tilde{A} \|)^{1-\frac{1}{N}} \| E \|^{\frac{1}{N}} . \tag{3.14}$$

Důkaz: Použijeme předchozího označení. Nechť $C_1, C_2, \ldots, C_p$ jsou souvislé navzájem disjunktní komponenty sjednocení $\bigcup_{i=1}^{N} D_i$. Podle lemmatu 3.2 obsahuje každá komponenta C_i právě tolik vlastních čísel matice $\tilde{A}$, kolik obsahuje vlastních čísel matice A.

Párová vzdálenost je definována jako maximum ze vzdáleností $|\tilde{\lambda}_{\pi(i)} - \lambda_i|$ při optimálním spárování. Proto stačí uvažovat jen takové permutace π na množině $\{1, \ldots, N\}$, které každému vlastnímu číslu $\lambda_j \in C_l$ přiřadí vlastní číslo $\tilde{\lambda}_{\pi(j)} \in C_l$, tedy páry jsou vytvářeny jen uvnitř jednotlivých komponent, nikoliv mezi čísly v různých komponentách. Bez újmy obecnosti se naše další úvahy budou týkat pouze největší souvislé komponenty označené jako C_1, o níž budeme předpokládat, že je sjednocením kruhů se středy ve vlastních číslech $\lambda_1, \ldots, \lambda_m$,

$$C_1 = \bigcup_{i=1}^{m} \widehat{D_i}.$$

Pokusíme se nalézt takové rozložení vlastních čísel matic A a $\tilde{A}$ v C_1, které je nejhorší možné, tj. kdy nabývá párová vzdálenost na C_1 svého maxima. Snadno nahlédneme, že nejméně příznivý případ pro vzájemnou polohu vlastních čísel $\lambda_1, \ldots, \lambda_m$ a $\tilde{\lambda}_{\pi(1)}, \ldots, \tilde{\lambda}_{\pi(m)}$ nastává, pokud jsou $\lambda_1, \ldots, \lambda_m$ rozloženy na přímce (nikoliv nezbytně na reálné ose) a vzdálenost $|\lambda_i - \lambda_{i-1}| = 2\gamma \leq 4\delta$. Pak je totiž velikost C_1 maximální, jak je naznačeno na obrázku 3.3.

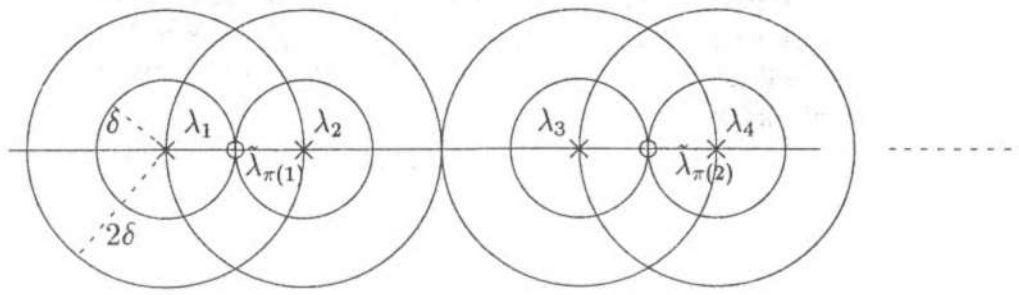


Obr. 3.3: Nejhorší možné rozložení vlastních čísel matice A takové, aby C_1 měla maximální velikost

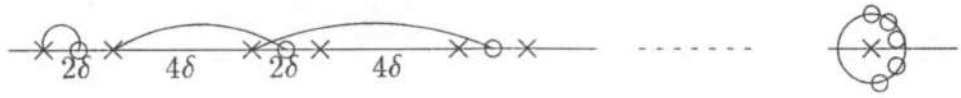
Uvažujme nyní vzájemnou polohu čísel λ_i a $\tilde{\lambda}_{\pi(i)}$. Z Elsnerovy věty víme, že $hd(A, \tilde{A}) \leq \delta$, neboli

$$\begin{aligned} sv_A(\tilde{A}) &\leq \delta \\ sv_{\tilde{A}}(A) &\leq \delta. \end{aligned} \tag{3.15}$$

To znamená, že sjednocení kruhů se středy v λ_i a poloměrem δ musí obsahovat všechna vlastní čísla $\tilde{\lambda}_{\pi(i)}$ a zároveň sjednocení kruhů se středy v $\tilde{\lambda}_{\pi(i)}$ a poloměrem δ musí obsahovat všechna vlastní čísla λ_i. Z hlediska párové vzdálenosti nastane nejhorší případ zjevně tehdy, když rozložení vlastních čísel $\tilde{\lambda}_{\pi(i)}$ bude „nejméně rovnoměrné" (viz obrázek 3.4). Optimální spárování je pro takto rozložená vlastní čísla znázorněno na obrázku 3.5.



Obr. 3.4: Nejhorší možné rozložení spekter matic A a $\tilde{A}$ na komponentě C_1, dovolené Elsnerovou větou



Obr. 3.5: Optimální spárování při nejhorší možné vzájemné poloze vlastních čísel λ_i a $\tilde{\lambda}_{\pi(i)}$

Při tomto spárování musí jedno z vlastních čísel $\tilde{\lambda}_{\pi(k)}$ znázorněných na pravém okraji posledního disku tvořit pár s λ_k, kde $k = \frac{m+1}{2}$ je-li m liché číslo, $k = \frac{m}{2} + 1$ je-li m sudé číslo. V každém případě platí pro vzdálenost λ_k od $\tilde{\lambda}_{\pi(k)}$ odhad

$$|\lambda_k - \tilde{\lambda}_{\pi(k)}| \leq \left\lfloor \frac{m-1}{2} \right\rfloor 2\gamma + \delta \leq (2m-1)\delta,$$

kde symbolem $\lfloor . \rfloor$ značíme celou část příslušného racionálního čísla. Snadno nahlédneme, že uvedený odhad je horním odhadem pro maximální vzdálenost čísel v páru při libovolné permutaci,

$$\min_{\pi} \max_{i=1,\ldots,m} |\lambda_i - \tilde{\lambda}_{\pi(i)}| \leq (2m-1)\delta.$$

□

Základem důkazu věty 3.9 byl neklesající odhad pro spektrální variaci $sv_A(A+\tau E)$. Větu je proto možno formulovat obecněji (důkaz jen sleduje předchozí postup a ponecháme jej čtenáři jako cvičení).

Věta 3.10 *Nechť $A \in C^{N,N}$, $E \in C^{N,N}$ a $\tilde{A} = A + E$. Předpokládejme dále, že $\beta(\tau)$ je pro $\tau \geq 0$ neklesající odhad spektrální variace $sv_A(A + \tau E)$. Pak pro párovou vzdálenost spekter matic A a $\tilde{A}$ platí:*

$$md(A, \tilde{A}) \leq (2N-1)\beta(1). \tag{3.16}$$

Pro úplnost uvádíme, že faktor $(2N-1)$ není optimální a je možné jej nahradit menší hodnotou. Velikost faktoru však pro nás není důležitá. Co je však velmi důležité je fakt, že získané odhady jsou úměrné hodnotě

$$\| E \|^{1/N}.$$

Po obtížné práci je výsledkem depresivně slabý (a často velmi pesimistický) odhad. Je-li například $N = 10^2$ (v praxi se často řeší problémy pro $N \sim 10^3$ až 10^5), a $\| E \| = 10^{-10}$, je výsledný odhad změny vlastních čísel při takto malé perturbaci úměrný $10^{-1/10} \| A \|$ a jeho praktický význam je nepatrný.

Cvičení

1. Ukažte, že $hd(A, \tilde{A})$ definuje metriku v $C^{N,N}$.
2. Jaká je úloha souvislých komponent C_i v důkaze Ostrowského - Elsnerovy věty?
3. Proč jsou v důkaze používány kruhy o poloměru 2δ?
4. K čemu je potřeba neklesající odhad pro $sv_A(\tilde{A}_\tau)$?
5. Dokažte větu 3.10.

3.2.3 Bauerova-Fikeho a Henriciho věta

Rádi bychom dospěli k odhadům citlivosti vlastních čísel, které jsou úměrné nikoliv $\| E \|^{1/N}$, ale pouze $\| E \|$. Není to možné vždy, chceme proto nalézt charakteristiku matice, která bude rozhodujícím způsobem ovlivňovat kvalitu odhadu. Touto charakteristikou bude *odchylka od normality*, studovaná v tomto odstavci.

Další věta má pro nás pouze pomocný charakter, je to však obecná věta velkého významu.

Věta 3.11 *(Bauer-Fike) Nechť $Q \in C^{N,N}$ je regulární matice, $\tilde{A} = A + E$, kde $E \in C^{N,N}$. Předpokládejme, že $\tilde{\lambda}$ je vlastní číslo matice $\tilde{A}$, které není vlastním číslem matice $A \in C^{N,N}$. Pak platí*

$$\| Q^{-1}(A - \tilde{\lambda}I)^{-1}Q \|^{-1} \leq \| Q^{-1}EQ \| . \tag{3.17}$$

Důkaz: Za předpokladů věty platí

$$\begin{aligned} Q^{-1}(\tilde{A} - \tilde{\lambda}I)Q &= Q^{-1}[(A - \tilde{\lambda}I) + E]Q \\ &= Q^{-1}(A - \tilde{\lambda}I)Q\{I + [Q^{-1}(A - \tilde{\lambda}I)^{-1}Q][Q^{-1}EQ]\}. \end{aligned} \tag{3.18}$$

Protože $Q^{-1}(\tilde{A} - \tilde{\lambda}I)Q$ je singulární matice a $Q^{-1}(A - \tilde{\lambda}I)Q$ je regulární matice, musí být matice

$$I + [Q^{-1}(A - \tilde{\lambda}I)^{-1}Q][Q^{-1}EQ]$$

singulární. Musí tedy platit

$$1 \leq \| [Q^{-1}(A - \tilde{\lambda}I)^{-1}Q][Q^{-1}EQ] \| . \tag{3.19}$$

S využitím konzistence maticové normy dostaneme tvrzení věty.

□

Poznámka 3.6 *Dohodneme-li se na označení*

$$\| Q^{-1}(A - \tilde{\lambda}I)^{-1}Q \|^{-1} \stackrel{def}{=} 0 \quad \text{pokud} \quad \tilde{\lambda} \in \sigma(A),$$

pak lze znění Bauer-Fikeho věty formálně rozšířit na všechna vlastní čísla matice $\tilde{A}$.

Poznámka 3.7 *Bauer-Fikeho věta platí i v případě libovolné maticové normy $\| \cdot \|_\alpha$, která splňuje podmínku konzistence, tj. pro niž platí*

$$\| AB \|_\alpha \leq \| A \|_\alpha \| B \|_\alpha ,$$

kde $A, B \in C^{N,N}$. V důkaze se využije faktu, že je-li matice $I + F$ singulární, pak pro libovolnou konzistentní normu platí

$$\| F \|_\alpha \geq 1.$$

Důkaz ponecháme jako cvičení.

Připomeňme, že výsledkem Schurovy transformace normální matice je diagonální matice. Výsledkem Schurovy transformace obecné matice je horní trojúhelníková matice R, přičemž R není určena jednoznačně. Odchylku od normality definujeme následujícím způsobem.

Definice 3.10 *Nechť* $\| \cdot \|_\alpha$ *je norma v* $C^{N,N}$, $A \in C^{N,N}$. *Označme* $\mathcal{U}$ *množinu všech unitárních matic takových, že matice* U^*AU *je horní trojúhelníková. Pro každé* $U \in \mathcal{U}$ *zapišme* $U^*AU = \Lambda_U + R_U$, *kde* Λ_U *je diagonální matice,* R_U *je horní trojúhelníková s nulami na diagonále. Jako* α**-odchylku od normality** *matice* A *pak definujeme číslo*

$$\delta_\alpha(A) \stackrel{def}{=} \min_{U \in \mathcal{U}} \| R_U \|_\alpha .$$

Výpočet odchylky od normality v obecné normě je zřejmě velice obtížný, protože Schurova dekompozice není jednoznačná. Na druhé straně, provádíme-li výpočet ve Frobeniově normě, lze s výhodou využít toho, že tato norma je invariantní vzhledem k unitárním transformacím.

Věta 3.12 *Pro libovolnou matici* $A \in C^{N,N}$ *s vlastními čísly* $\lambda_1, \lambda_2, \ldots, \lambda_N$ *platí*

$$\delta_F(A) = (\| A \|_F^2 - \sum_{i=1}^{N} |\lambda_i|^2)^{\frac{1}{2}}. \tag{3.20}$$

Důkaz: Protože Frobeniova norma

$$\| A \|_F = (\sum_{i=1}^{N} \sum_{j=1}^{N} |a_{ij}|^2)^{\frac{1}{2}}$$

je invariantní vzhledem k unitárním transformacím, s použitím předchozího označení platí

$$\| A \|_F^2 = \| U^*AU \|_F^2 = \| \Lambda_U + R_U \|_F^2 = \sum_{i=1}^{N} |\lambda_i|^2 + \| R_U \|_F^2,$$

kde U je libovolná unitární matice z množiny $\mathcal{U}$.

□

Konečně můžeme vyslovit a dokázat Henriciho větu. Z formálních důvodů budeme ve zbytku části 3.2.3 předpokládat, že pro danou matici A je $\delta_\alpha(A) \neq 0$.

Věta 3.13 *(Henrici) Nechť* $\| \cdot \|_\alpha$ *je norma v* $C^{N,N}$ *taková, že* $\| B \|_\alpha \geq \| B \|$ *pro každou matici* $B \in C^{N,N}$. *Nechť* $A \in C^{N,N}$, *položme* $\tilde{A} = A + E$, *kde* $E \in C^{N,N}$. *Nechť* $\delta_\alpha(A) \neq 0$. *Pak pro každé vlastní číslo* $\tilde{\lambda}$ *matice* $\tilde{A}$ *existuje vlastní číslo* λ *matice* A *tak, že*

$$\frac{\left(\frac{|\tilde{\lambda}-\lambda|}{\delta_\alpha(A)}\right)^N}{1+\left(\frac{|\tilde{\lambda}-\lambda|}{\delta_\alpha(A)}\right)+\ldots+\left(\frac{|\tilde{\lambda}-\lambda|}{\delta_\alpha(A)}\right)^{N-1}} \leq \frac{\| E \|}{\delta_\alpha(A)}. \tag{3.21}$$

Důkaz: Uvažujme libovolné vlastní číslo $\tilde{\lambda}$ matice $\tilde{A}$. Pokud je $\tilde{\lambda}$ zároveň vlastním číslem matice A, je tvrzení triviálně splněno. Uvažujme $\tilde{\lambda} \notin \sigma(A)$. Nechť $U^*AU = \Lambda + R$ je výsledek Schurovy transformace matice A pro nějakou $U \in \mathcal{U}$, matice Λ je diagonální a R horní trojúhelníková s nulovou diagonálou. Z Bauerovy-Fikeho věty (pro $Q \stackrel{def}{=} \tilde{U}$) pak máme

$$\| (\Lambda - \tilde{\lambda}I + R)^{-1} \|^{-1} \leq \| E \|. \tag{3.22}$$

Matici $(\Lambda - \tilde{\lambda}I + R)^{-1}$ lze upravit následujícím způsobem

$$\begin{aligned}(\Lambda - \tilde{\lambda}I + R)^{-1} &= \{(\Lambda - \tilde{\lambda}I)[I - (\Lambda - \tilde{\lambda}I)^{-1}(-R)]\}^{-1} \\ &= [I - (\Lambda - \tilde{\lambda}I)^{-1}(-R)]^{-1}(\Lambda - \tilde{\lambda}I)^{-1}.\end{aligned}$$

Spektrální poloměr matice $(\Lambda - \tilde{\lambda}I)^{-1}(-R)$ je roven nule. To znamená, že rozvoj matice $[I - (\Lambda - \tilde{\lambda}I)^{-1}(-R)]^{-1}$ do Neumannovy řady je konvergentní. Navíc platí, že $R^j = 0$ pro $j \geq N$, rozvoj je tedy konečný,

$$[I - (\Lambda - \tilde{\lambda}I)^{-1}(-R)]^{-1} = I - (\Lambda - \tilde{\lambda}I)^{-1}R + \ldots + (-1)^{N-1}[(\Lambda - \tilde{\lambda}I)^{-1}R]^{N-1}.$$

Označme

$$\omega = \min_{\lambda \in \sigma(A)} |\tilde{\lambda} - \lambda|$$

a odhadněme velikost normy matice $(\Lambda - \tilde{\lambda}I + R)^{-1}$ následujícím způsobem

$$\begin{aligned}\| (\Lambda - \tilde{\lambda}I + R)^{-1} \| \leq \; &\{\| I \| + \| (\Lambda - \tilde{\lambda}I)^{-1} \| \| R \| + \ldots + \\ &+ \| (\Lambda - \tilde{\lambda}I)^{-1} \|^{N-1} \| R \|^{N-1}\} \| (\Lambda - \tilde{\lambda}I)^{-1} \|.\end{aligned}$$

Z definice odchylky od normality a vztahu mezi spektrální normou a normou $\| \cdot \|_\alpha$ máme

$$\| (\Lambda - \tilde{\lambda}I + R)^{-1} \| \leq \omega^{-1}\{1 + \omega^{-1}\delta_\alpha(A) + \ldots + [\omega^{-1}\delta_\alpha(A)]^{N-1}\}. \tag{3.23}$$

Tvrzení věty dostaneme kombinací vztahů (3.22), (3.23) a algebraickou úpravou.

□

Henriciho věta popisuje spojitý přechod mezi dvěma extrémními případy, které mohou pro odhad citlivosti vlastních čísel vzhledem k perturbacím matice nastat. V prvním případě je tento odhad úměrný N-té odmocnině normy matice E, ve druhém pouze velikosti normy matice E. Abychom to nahlédli, budeme se zabývat vlastnostmi reálné funkce reálné proměnné $\Psi(\eta)$ definované vztahem

$$\Psi(\eta) \stackrel{def}{=} \frac{\eta^N}{1 + \eta + \ldots + \eta^{N-1}}, \quad \eta \geq 0. \tag{3.24}$$

Pak pro $\eta \stackrel{def}{=} \frac{|\tilde{\lambda}-\lambda|}{\delta_\alpha(A)}$ platí podle Henriciho věty odhad

$$\Psi(\eta) \leq \frac{\| E \|}{\delta_\alpha(A)}.$$

Všimněme si nyní, jak vypadá $\Psi(\eta)$ pro mezní hodnoty η. Je-li η malé, je jmenovatel ve výraze (3.24) blízký jedné, tudíž $\Psi(\eta) \approx \eta^N$ a tedy asymptotický odhad pro spektrální variaci je

$$\frac{sv_A(\tilde{A})}{\delta_\alpha(A)} \leq \left(\frac{\| E \|}{\delta_\alpha(A)}\right)^{\frac{1}{N}}.$$

Je-li naopak η velké, pak je η^{N-1} nejvýznačnějším členem ve jmenovateli výrazu (3.24) a tudíž $\Psi(\eta) \approx \eta$. Asymptotický odhad pro spektrální variaci má pak tvar

$$\frac{sv_A(\tilde{A})}{\delta_\alpha(A)} \leq \frac{\| E \|}{\delta_\alpha(A)}.$$

Formulujeme-li předchozí úvahy přesně, dostáváme následující důsledek věty 3.13.

Věta 3.14 *S použitím zavedeného označení a za předpokladů věty 3.13 platí.*
Je-li $\frac{\|E\|}{\delta_\alpha(A)} < \frac{1}{N}$, *pak*

$$\frac{sv_A(\tilde{A})}{\delta_\alpha(A)} \leq N^{\frac{1}{N}} \left(\frac{\| E \|}{\delta_\alpha(A)}\right)^{\frac{1}{N}}. \tag{3.25}$$

Je-li $\frac{\|E\|}{\delta_\alpha(A)} > 1$, *pak*

$$sv_A(\tilde{A}) \leq \| E \| + \delta_\alpha(A). \tag{3.26}$$

Důkaz: Pro funkci Ψ definovanou výrazem (3.24) platí

$$\Psi(\eta) < \frac{1}{N} \Rightarrow \eta < 1.$$

Z předpokladu $\frac{\|E\|}{\delta_\alpha(A)} < \frac{1}{N}$ dostáváme $\Psi(\eta) < \frac{1}{N}$ a tedy $\eta < 1$. A protože pro $\eta < 1$ je zřejmě splněno

$$\frac{\eta^N}{N} \leq \frac{\eta^N}{1 + \ldots + \eta^{N-1}},$$

dostaneme

$$\frac{1}{N}\left(\frac{sv_A(\tilde{A})}{\delta_\alpha(A)}\right)^N \leq \Psi\left(\frac{sv_A(\tilde{A})}{\delta_\alpha(A)}\right) \leq \frac{\| E \|}{\delta_\alpha(A)}.$$

Pokud je $\eta > 1$, platí pro funkci Ψ

$$\Psi(\eta) = \frac{\eta}{1 + \eta^{-1} + \ldots + \eta^{-(N-1)}} \geq \eta(1 - \eta^{-1}) = \eta - 1. \tag{3.27}$$

Výrazu (3.27) budeme chtít použít pro proměnnou

$$\zeta = \Psi^{-1}\left(\frac{\| E \|}{\delta_\alpha(A)}\right) \tag{3.28}$$

(všimněme si, že inverzní funkce k funkci Ψ existuje a je monotonní). Nejprve je třeba ověřit, že takto definované ζ je větší než jedna. Použijeme další vlastnosti funkce Ψ, a to $\Psi(\zeta) > 1 \Rightarrow \zeta > 1$. Protože z (3.28) a podle předpokladu

$$\Psi(\zeta) = \frac{\| E \|}{\delta_\alpha(A)} > 1,$$

je i ζ definované výrazem (3.28) větší než jedna, tedy z (3.27) máme

$$\Psi^{-1}\left(\frac{\| E \|}{\delta_\alpha(A)}\right) \leq \Psi\left(\Psi^{-1}\left(\frac{\| E \|}{\delta_\alpha(A)}\right)\right) + 1. \tag{3.29}$$

Z Henriciho věty pak platí

$$\Psi\left(\frac{sv_A(\tilde{A})}{\delta_\alpha(A)}\right) \leq \frac{\| E \|}{\delta_\alpha(A)}. \tag{3.30}$$

Aplikujeme-li na obě strany výrazu (3.30) funkci Ψ^{-1}, pak spolu s užitím (3.29), dostaneme

$$\frac{sv_A(\tilde{A})}{\delta_\alpha(A)} \leq \Psi^{-1}\left(\frac{\| E \|}{\delta_\alpha(A)}\right) \leq \frac{\| E \|}{\delta_\alpha(A)} + 1,$$

čímž je dokázáno (3.26).

□

Protože Ψ je neklesající funkce, je neklesající i Ψ^{-1}. Tedy lze formulovat důsledky věty (3.10) a (3.13) v následujícím tvaru.

Věta 3.15 *Nechť matice $A \in C^{N,N}$, $\tilde{A} = A + E$, kde $E \in C^{N,N}$ a funkce Ψ je definována výrazem (3.24), $\delta_\alpha(A) \neq 0$. Pak platí*

$$md(A, \tilde{A}) \leq (2N - 1)\delta_\alpha(A)\Psi^{-1}\left(\frac{\| E \|}{\delta_\alpha(A)}\right).$$

Důkaz: Podle věty (3.13) platí

$$\frac{sv_A(\tilde{A})}{\delta_\alpha(A)} \leq \Psi^{-1}\left(\frac{\| E \|}{\delta_\alpha(A)}\right)$$

a tedy i

$$sv_A(A + \tau E) \leq \delta_\alpha(A)\Psi^{-1}\left(\frac{\| \tau E \|}{\delta_\alpha(A)}\right),$$

a protože tento odhad je pro $\tau \in < 0, 1 >$ neklesající, lze přímo použít větu 3.10.

□

Stejně jako v Elsnerově a Ostrowského-Elsnerově větě, tak i v Henriciho větě dostáváme pro obecný problém dimenze N odhady, v nichž vystupuje N-tá odmocnina normy matice perturbací $\| E \|$. V další větě ukážeme, že tento odhad lze zlepšit v případě, kdy největší Jordanův blok matice A má řád m, kde $m < N$.

Věta 3.16 *Nechť matice $A \in C^{N,N}$, $\tilde{A} = A + E$, kde $E \in C^{N,N}$ a označme Jordanův kanonický tvar této matice $J = Q^{-1}AQ$. Nechť m je velikost největšího Jordanova bloku v J. Pak pro každé $\tilde{\lambda} \in \sigma(\tilde{A})$ existuje $\lambda \in \sigma(A)$ takové, že*

$$\frac{|\tilde{\lambda} - \lambda|^m}{1 + |\tilde{\lambda} - \lambda| + \ldots + |\tilde{\lambda} - \lambda|^{m-1}} \leq \| Q^{-1}EQ \| . \tag{3.31}$$

Důkaz: Tvrzení se dokazuje zcela analogicky jako Henriciho věta, proto zde důkaz jen naznačíme. Bauerovu-Fikeho větu použijeme v následujícím znění

$$\| Q^{-1}(A - \tilde{\lambda}I)^{-1}Q \|^{-1} \leq \| Q^{-1}EQ \| .$$

Pak máme

$$\begin{aligned} Q^{-1}(A - \tilde{\lambda}I)^{-1}Q &= (Q^{-1}(A - \tilde{\lambda}I)Q)^{-1} = (J - \tilde{\lambda}I)^{-1} = (\Lambda - \tilde{\lambda}I + R)^{-1} \\ &= \{I - (\Lambda - \tilde{\lambda}I)^{-1}R + \\ &+ \ \ldots + (-1)^{N-1}[(\Lambda - \tilde{\lambda}I)^{-1}R]^{N-1}\}(\Lambda - \tilde{\lambda}I)^{-1}. \end{aligned} \tag{3.32}$$

Matice R je horní trojúhelníková část Jordanova kanonického tvaru J s vynulovanou diagonálou. Na její vedlejší diagonále leží buď jedničky nebo nuly a všude jinde jsou nulové prvky. Zřejmě je $\| R \| = 1$. Navíc nejdelší „souvislý pás" nenulových prvků se skládá z $m-1$ jedniček. Proto ve výraze (3.32) budou všechny členy, v nichž se R vyskytuje v mocninách větších nebo rovných m nulové.

□

Cvičení

1. Proč je pro horní trojúhelníkovou matici $R \in C^{N,N}$ s nulovou diagonálou $R^j = 0$ pro $j \geq N$?
2. Dokažte, že je-li $\Psi(\eta) < 1$ pak je $\eta < 1$.
3. Dokažte, že funkce Ψ definovaná vztahem (3.24) je monotonní.
4. Dokažte větu 3.11 pro libovolnou konzistentní maticovou normu $\| . \|_\alpha$.
5. Nechť $\| . \|_\mu$ je libovolná konzistentní norma v $C^{N,N}$. Pak existuje vektorová norma $\| . \|_\nu$ v C^N tak, že $\| Ax \|_\nu \leq \| A \|_\mu \| x \|_\nu$ platí $\forall A \in C^{N,N}$, $\forall x \in C^N$. Dokažte.

3.3 Citlivost jednoduchého vlastního čísla

Poznali jsme, že vlastní čísla matice mohou být velmi citlivá na malé změny prvků matice. Je přirozené ptát se, zda všechna vlastní čísla dané matice jsou stejně citlivá a čím je změna jednotlivých vlastních čísel *podmíněna*. V tomto odstavci se budeme zabývat *podmíněností* jednoduchého vlastního čísla obecné matice. Nejdříve uvedeme a dokážeme Geršgorinovu větu, která má při zkoumání citlivosti jednoduchého vlastního čísla klíčové postavení.

Geršgorinova věta říká, že vlastní čísla dané matice leží ve sjednocení kruhů, které jsou popsány pomocí prvků této matice. Tedy není v pravém slova smyslu větou z teorie citlivosti (žádná matice perturbací zde nevystupuje). Při vhodném užití Geršgorinovy věty však lze získat velmi dobré odhady polohy vlastních čísel perturbované matice.

Věta 3.17 *(Geršgorin) Nechť $A \in C^{N,N}$. Označme součet absolutních hodnot prvků v i-tém řádku s vynecháním prvku na hlavní diagonále α_i, $\alpha_i \stackrel{def}{=} \sum_{j=1,j\neq i}^{N} |a_{ij}|$, $i = 1,\ldots,N$, a*

$$G_i(A) = \{\zeta \in C; |\zeta - a_{ii}| \leq \alpha_i\}.$$

Pak platí

$$\sigma(A) \subset \bigcup_{i=1}^{N} G_i(A),$$

tj. spektrum matice A leží ve sjednocení Geršgorinových kruhů se středy v diagonálních prvcích a poloměry α_i. Pokud je m kruhů $G_i(A)$ izolováno od ostatních $N-m$ kruhů, pak jejich sjednocení obsahuje právě m vlastních čísel matice A.

Důkaz: Použijeme vztah (3.19) z důkazu Bauerovy-Fikeho věty pro speciální volbu příslušných matic a maximovou normu $\| \cdot \|_\infty$, která je definována vztahem $\| B \|_\infty = \max_{1\leq i\leq N} \sum_{j=1}^{N} |b_{ij}|$, $B = (b_{ij})_{i,j=1,\ldots,N} \in C^{N,N}$. V nerovnosti

$$1 \leq \| [Q^{-1}(B - \tilde{\lambda} I)^{-1} Q][Q^{-1} E Q] \|_\infty \tag{3.33}$$

položíme

$$\begin{aligned} Q &\stackrel{def}{=} I \\ B &\stackrel{def}{=} diag(a_{11},\ldots,a_{NN}) \\ B+E &\stackrel{def}{=} A \ (\text{tj. } E = A - diag(a_{11},\ldots,a_{NN})) \\ \tilde{\lambda} &\stackrel{def}{=} \lambda, \end{aligned}$$

kde λ je libovolné vlastní číslo matice A, pro které platí $\lambda \neq a_{ii}$ pro $i = 1,\ldots,N$ (jinak je znění věty triviální). Výsledkem je nerovnost

$$1 \leq \| [diag(a_{11},\ldots,a_{NN}) - \lambda I]^{-1} [A - diag(a_{11},\ldots,a_{NN})] \|_\infty,$$

kterou z definice maximové normy zapíšeme ve tvaru

$$\max_i \frac{\sum_{j=1,j\neq i}^{N} |a_{ij}|}{|a_{ii} - \lambda|} \geq 1.$$

Nechť i_λ je ten řádkový index, v němž se nabývá maxima, tedy platí

$$\sum_{j=1, j\neq i_\lambda}^{N} |a_{i_\lambda j}| \geq |a_{i_\lambda i_\lambda} - \lambda|,$$

což dokazuje první část věty. Důkaz druhé části lze snadno získat pomocí techniky použité při důkazu lemmatu 3.2 a je ponechán čtenáři jako cvičení.

□

Následující příklady jsou ukázkami toho, jak může použití Geršgorinovy věty zlepšit odhady polohy vlastních čísel perturbované matice. Úmyslně volíme elementární příklad matice řádu dvě, neboť nám to umožní názorně vysvětlit důležitou techniku využitou dále v textu.

Příklad 3.4 *Uvažujme matici* $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ *a její perturbaci* $E = \begin{pmatrix} 0 & 10^{-4} \\ 10^{-4} & 0 \end{pmatrix}$. *Perturbovaná matice* $\tilde{A} = A + E$ *je pak ve tvaru*

$$\tilde{A} = \begin{pmatrix} 1 & 10^{-4} \\ 10^{-4} & 2 \end{pmatrix}.$$

Jednoduchým výpočtem určíme spektrum matice $\tilde{A}$, $\sigma(\tilde{A}) = \{\approx 1 - 10^{-8}, \approx 2 + 10^{-8}\}$. *Spektrum matice* A *je zřejmě* $\sigma(A) = \{1, 2\}$. *Ukážeme, jak se liší odhad vzájemné polohy vlastních čísel matic* A *a* $\tilde{A}$ *získaný z Elsnerovy a Geršgorinovy věty.*

Nejprve použijeme Elsnerovu větu. Zřejmě je $\| A \| = 2$ *a* $\| \tilde{A} \| \approx 2 + 10^{-8}$, *tedy pro Hausdorffovu vzdálenost platí* $hd(A, \tilde{A}) \leq (\| A \| + \| \tilde{A} \|)^{1-\frac{1}{N}} \| E \|^{\frac{1}{N}} \approx 2 \times 10^{-2}$. *Tedy podle Elsnerovy věty je změna vlastních čísel omezena hodnotou řádu* 10^{-2}. *Na druhé straně poloměr obou kruhů* $G_i(\tilde{A})$ *z Geršgorinovy věty je* 10^{-4}. *Tento odhad je o dva řády lepší než odhad pro polohu* $\tilde{\lambda}_1, \tilde{\lambda}_2$ *z Elsnerovy věty, ani on však není dostatečně přesný, neboť skutečná vlastní čísla matice* $\tilde{A}$ *mají hodnotu* $\approx 1 - 10^{-8}$, $\approx 2 + 10^{-8}$.

V dalším příkladě ukážeme, jak lze odhad polohy vlastních čísel z příkladu 3.4 zlepšit.

Příklad 3.5 *Uvažujme matice* A *a* E *z příkladu 3.4. Technika pro získání dobrého odhadu polohy vlastních čísel matice*

$$\tilde{A} = \begin{pmatrix} 1 & 10^{-4} \\ 10^{-4} & 2 \end{pmatrix}$$

je založena na Geršgorinově větě a vychází z elementárního poznatku, že podobnostní transformace zachovává vlastní čísla matice. Budeme hledat takovou podobnostní transformaci, aby Geršgorinovy kruhy zůstaly disjunktní a alespoň jeden z nich měl pro transformovanou matici výrazně menší poloměr než pro matici původní.

Proveďme podobnostní transformaci matice $\tilde{A}$ *následujícím způsobem*

$$\tilde{A}(\alpha) = \begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 10^{-4} \\ 10^{-4} & 2 \end{pmatrix} \begin{pmatrix} \alpha^{-1} & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & \alpha 10^{-4} \\ \alpha^{-1} 10^{-4} & 2 \end{pmatrix},$$

kde α je kladné reálné číslo. Ukážeme, jak lze vhodnou volbou parametru α získat dostatečně jemný odhad pro vlastní číslo matice $\tilde{A}$ ležící v kruhu se středem v 1. Podle Geršgorinovy věty leží totiž vlastní čísla matice $\tilde{A}(\alpha)$ ve sjednocení intervalů

$$(1-\alpha 10^{-4}, 1+\alpha 10^{-4}) \cup (2-\alpha^{-1}10^{-4}, 2+\alpha^{-1}10^{-4}).$$

Geršgorinova věta navíc říká, že pokud se tyto dva intervaly neprotnou, obsahuje každý z nich právě jedno vlastní číslo matice $\tilde{A}(\alpha)$. Zvolíme-li tedy α dostatečně malé, ale zároveň tak velké, aby se oba intervaly neprotly (např. $\alpha = 1.01 \times 10^{-4}$), dostaneme velmi dobrý odhad pro vlastní číslo v intervalu $(1-\alpha 10^{-4}, 1+\alpha 10^{-4})$.

Než zformulujeme větu o citlivosti jednoduchého vlastního čísla vzhledem k malým změnám prvků matice, připomeneme pojmy *levý* a *pravý vlastní vektor* a provedeme některé pomocné úvahy, které při důkazu věty budeme potřebovat. Nechť λ je vlastní číslo matice $A \in C^{N,N}$. Pak existuje nenulový vektor $x \in C^N$ tak, že $Ax = \lambda x$. Z vlastností determinantu víme, že

$$\det(A - \lambda I) = \overline{\det(A^* - \bar{\lambda} I)},$$

a protože $\det(A - \lambda I) = 0$, je matice $A^* - \bar{\lambda} I$ singulární. Existuje tedy nenulový vektor $y \in C^N$ tak, že $(A^* - \bar{\lambda} I)y = 0$, neboli

$$y^* A = \lambda y^*.$$

Definice 3.11 *Při výše použitém označení nazýváme vektor x* **pravým** *a vektor y* **levým vlastním vektorem** *matice A příslušným vlastnímu číslu λ.*

Je-li vlastní číslo λ jednoduché (tj. není násobným kořenem charakteristické rovnice), pak levý a pravý vlastní vektor nemohou být vzájemně ortogonální. Při vhodném normování je jejich skalární součin rovný jedné. Ukážeme to následujícím způsobem. Předpokládejme, že $A \in C^{N,N}$ a nechť $J = W^{-1}AW$ je její Jordanův kanonický tvar. Nechť Jordanovy bloky jsou uspořádány tak, že na prvním místě je Jordanův blok obsahující jednoduché vlastní číslo λ,

$$J = \begin{pmatrix} \lambda & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_k \end{pmatrix}, \tag{3.34}$$

kde $J_2, \ldots, J_k$ jsou Jordanovy bloky příslušné ostatním vlastním číslům (nevyznačené prvky matice jsou nulové). Přepíšeme-li Jordanův rozklad do tvaru

$$AW = WJ \tag{3.35}$$

a označíme-li sloupce matice W jako $w_1, \ldots, w_N$, pak pro první sloupec w_1 dostaneme

$$Aw_1 = \lambda w_1.$$

Pro matici J^* pak platí

$$J^* = W^* A^* (W^{-1})^* = \begin{pmatrix} \bar{\lambda} & & & \\ & J_2^* & & \\ & & \ddots & \\ & & & J_k^* \end{pmatrix},$$

neboli

$$A^*(W^{-1})^* = (W^*)^{-1} J^*.$$

Použijeme-li vztah $(W^{-1})^* = (W^*)^{-1}$ a označíme-li sloupce matice $(W^*)^{-1}$ jako $y_1, \ldots, y_N$, opět musí platit

$$A^* y_1 = \bar{\lambda} y_1.$$

Navíc je pro vektory w_1, y_1 splněno

$$y_1^* w_1 = 1.$$

Následuje hlavní věta části 3.3.

Věta 3.18 *Nechť λ je jednoduché vlastní číslo matice $A \in C^{N,N}$, y je příslušný levý a x pravý vlastní vektor. Uvažujme $\tilde{A} = A + E$, $E \in C^{N,N}$. Je-li perturbace E dostatečně malá, pak existuje právě jedno vlastní číslo matice $\tilde{A}$ takové, které lze vyjádřit ve tvaru*

$$\tilde{\lambda} = \lambda + \frac{y^* E x}{y^* x} + O\left(\| E \|^2\right), \tag{3.36}$$

kde $O(\| E \|^2)$ vyjadřuje členy, které lze omezit odhadem $c \| E \|^2$, c je konstanta nezávislá na E.

Důkaz: Základem důkazu je technika vysvětlená v příkladu 3.5; je zde však nutné být pozorný při určení velikosti parametru α podobnostní transformace.

Nechť $\delta > 0$ je vzdálenost jednoduchého vlastního čísla λ od ostatních vlastních čísel matice A. Použijeme předcházejícího označení, tj. $J = W^{-1}AW$, J uvažujeme ve tvaru (3.34). Navíc předpokládejme, že všechny nenulové prvky Jordanovy matice J na vedlejší diagonále jsou rovny $\delta/3$ (toho lze vždy dosáhnout vhodným normováním sloupců matice W). Položme $Y^* \stackrel{def}{=} W^{-1}$, $X \stackrel{def}{=} W$ a označme první sloupec matice X jako x a první sloupec matice Y jako y. Při takto zavedeném označení je x pravý a y levý vlastní vektor příslušný vlastnímu číslu λ a platí $y^*x = 1$. Všimněme si, že při takto normalizovaných vlastních vektorech může být např. hodnota $\| y \|$ značně velká.

Uvažujme nyní, jak vypadají prvky matice $\tilde{J} = Y^*(A+E)X = Y^*AX + Y^*EX$,

$$\tilde{J} = \begin{pmatrix} \lambda + y^*Ex & \epsilon & & \ldots & & \epsilon \\ \epsilon & \mu & \tau & \epsilon & \ldots & \epsilon \\ \vdots & & & & & \vdots \\ & \ddots & \ddots & & & \epsilon \\ & & & & & \tau \\ \epsilon & & \ldots & & \epsilon & \mu \end{pmatrix},$$

kde symbolem ϵ jsou označeny prvky, jejichž absolutní hodnota nepřesáhne $\| Y \| \| E \| \| X \| = \kappa(X) \| E \|$, symbolem μ jsou označeny diagonální prvky jiné než $\lambda + y^*Ex$ a symbolem τ jsou označeny prvky omezené v absolutní hodnotě číslem $\delta/3 + |\epsilon|$. Polohu vlastního čísla ležícího v prvním Geršgorinově kruhu se středem v $\lambda + y^*Ex$ budeme odhadovat užitím Geršgorinovy věty. Přesnost tohoto odhadu zřejmě závisí na hodnotě prvků v prvním řádku matice $\tilde{J}$. Proto se budeme snažit nalézt podobnostní transformaci tak, aby první Geršgorinův kruh byl co možná nejmenší při zaručení jeho disjunktnosti s ostatními kruhy. Použijeme podobnostní transformaci charakterizovanou kladným reálným parametrem α

$$\begin{pmatrix} \alpha & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} \tilde{J} \begin{pmatrix} \alpha & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}^{-1} = \tilde{J}(\alpha).$$

Matice $\tilde{J}(\alpha)$ se od $\tilde{J}$ liší pouze tím, že v prvním řádku má ve sloupcích $2, \ldots, N$ prvky $\alpha\epsilon$ a v prvním sloupci v řádcích $2, \ldots, N$ prvky $\alpha^{-1}\epsilon$. Vlastní čísla matice $\tilde{J}(\alpha)$ jsou totožná s vlastními čísly matice $\tilde{J}$. Podle Geršgorinovy věty je

$$\sigma(\tilde{J}(\alpha)) \subset \bigcup_{i=1}^{N} G_i,$$

kde G_1 má střed v $\lambda + y^*Ex$ a poloměr nepřesahující hodnotu $(N-1)\alpha|\epsilon|$. Ostatní kruhy $G_2, \ldots, G_N$ mají středy v prvcích označených jako μ a poloměry nepřesahující hodnotu $\alpha^{-1}|\epsilon| + \delta/3 + |\epsilon| + (N-3)|\epsilon|$. Abychom pro odhad vlastního čísla matice $\tilde{J}$, které leží v kruhu G_1 mohli použít Geršgorinovu větu, je třeba zaručit, že se kruh G_1 neprotne s žádným jiným (pak je zaručeno, že v G_1 leží právě jedno vlastní číslo matice $\tilde{J}(\alpha)$).

Vzdálenost středu kruhu G_1 od středu libovolného jiného kruhu je v nejhorším případě $\delta - 2|\epsilon|$. Tudíž, aby se kruh G_1 neprotl s žádným jiným, musí být splněno

$$(N-1)\alpha|\epsilon| + \alpha^{-1}|\epsilon| + \frac{\delta}{3} + |\epsilon| + (N-3)|\epsilon| < \delta - 2|\epsilon| \,,$$

což je po jednoduché úpravě

$$(N-1)\alpha|\epsilon| + \alpha^{-1}|\epsilon| + N|\epsilon| < \frac{2}{3}\delta \,. \tag{3.37}$$

Budeme hledat takové podmínky pro parametry $|\epsilon|$ a α, aby bylo splněno (3.37), hodnota $|\epsilon|$ byla co možná největší a hodnota $\alpha|\epsilon|$ přitom co možná nejmenší. Parametr $|\epsilon|$ je omezen hodnotou $\kappa(X) \| E \|$. Nechť je perturbace E matice A tak malá, že platí

$$\frac{2}{3}\delta - N|\epsilon| > \frac{\delta}{2} \,. \tag{3.38}$$

Pak postačující podmínkou pro splnění (3.37) je

$$(N-1)\alpha|\epsilon| + \alpha^{-1}|\epsilon| < \frac{\delta}{2} \,.$$

Nahradíme-li pro zjednodušení hodnotu $N-1$ hodnotou N, pak parametr α musí splňovat

$$\alpha^{-1}|\epsilon| + N\alpha|\epsilon| < \frac{\delta}{2}\,. \tag{3.39}$$

Vynásobením parametrem α dostaneme kvadratickou nerovnost

$$N\alpha^2|\epsilon| - \frac{\delta}{2}\alpha + |\epsilon| < 0\,,$$

která bude jistě splněna, volíme-li

$$\alpha = \frac{4|\epsilon|}{\delta} \tag{3.40}$$

a perturbace E je tak malá, že

$$\frac{16N|\epsilon|^3}{\delta^2} - 2|\epsilon| + |\epsilon| < 0\,,$$

neboli

$$\frac{|\epsilon|}{\delta} < \frac{1}{4\sqrt{N}}\,. \tag{3.41}$$

Dokázali jsme, že pokud je splněno (3.40) a (3.41), platí i (3.37) a tedy kruh G_1 se středem v $\lambda + y^*Ex$ je disjunktní s ostatními kruhy. Poloměr kruhu G_1 je přitom omezen odhadem

$$(N-1)\alpha|\epsilon| \leq \frac{4N|\epsilon|^2}{\delta} = O(\| E \|^2).$$

Kruh G_1 obsahuje tedy právě jedno vlastní číslo matice $\tilde{J}$, označme ho $\tilde{\lambda}$, a platí

$$\tilde{\lambda} = \lambda + y^*Ex + O(\| E \|^2).$$

□

Poznámka 3.8 *Všimněme si, kdy bude člen $O(\| E \|^2)$ v důkaze předešlé věty nevýznamný ve srovnání s ostatními členy. Bude to tehdy, je-li $|\epsilon|/\delta$ dostatečně malé. Pokud je δ malé číslo (tj. jednoduché vlastní číslo λ je špatně separované od ostatních), bude množina matic perturbací, pro něž je odhad (3.36) platný značně omezena (odhad platí jen pro velmi malé perturbace). Velikost čitatele $|\epsilon|$ nezávisí jen na $\| E \|$, ale také na velikosti $\kappa(X)$. Pokud je $\kappa(X) \gg 1$, opět jsou dovoleny jen velmi malé perturbace E. Navíc, malá hodnota δ vede často k velmi špatně podmíněné matici transformace na Jordanův kanonický tvar.*

Poznámka 3.9 *Výraz (3.36) lze napsat i v jiném tvaru, a to*

$$\tilde{\lambda} = \frac{y^*(A+E)x}{y^*x} + O(\| E \|^2).$$

Výraz

$$\frac{y^*(A+E)x}{y^*x} \tag{3.42}$$

připomíná zobecnění Rayleighova kvocientu. Věta 3.18 pak říká, že pro dostatečně malé perturbace je vztah (3.42) aproximací prvního řádu pro perturbaci jednoduchého vlastního čísla.

Odhadněme ze vztahu (3.36) vzdálenost vlastních čísel $\tilde{\lambda}$ a λ,

$$|\tilde{\lambda} - \lambda| \leq \frac{\| y \| \| x \|}{|y^* x|} \| E \| + O(\| E \|^2).$$

Vidíme, že změna $|\tilde{\lambda} - \lambda|$ je úměrná hodnotě $\| E \|$, tato úměrnost je však **podmíněna** velikostí násobitele

$$\nu = \frac{\| y \| \| x \|}{|y^* x|}. \tag{3.43}$$

Definice 3.12 *Nechť λ je jednoduché vlastní číslo matice $A \in C^{N,N}$, y resp. x je příslušný levý resp. pravý vlastní vektor. Pak číslo ν definované vztahem (3.43) nazýváme* **číslem podmíněnosti vlastního čísla** λ.

Všimněme si, že ν je secans úhlu, který svírají vektory x a y. Platí totiž

$$|y^* x| = \| x \| \| y \| \cos\varphi$$

a $\sec\varphi = \frac{1}{\cos\varphi}$. Tedy $\nu = 1$ pokud x a y svírají nulový úhel a hodnota ν roste se zvětšováním úhlu mezi x a y.

Přímým důsledkem věty 3.18 je diferencovatelnost jednoduchého vlastního čísla podle prvků matice.

Věta 3.19 *Nechť λ je jednoduché vlastní číslo matice $A = (a_{ij})_{i,j=1,\ldots,N} \in C^{N,N}$, y je příslušný levý a x pravý vlastní vektor. Pak λ je diferencovatelnou funkcí prvků matice a platí*

$$\frac{\partial \lambda}{\partial a_{ij}} = \frac{(y^* e_i)(e_j^* x)}{y^* x}, \tag{3.44}$$

kde e_k je k-tý vektor standardní euklidovské báze.

Důkaz: Označme $(1)_{ij}$ matici z prostoru $C^{N,N}$, jejímž jediným nenulovým prvkem je jednička na místě (i,j), $(1)_{ij} = e_i e_j^T$. Pak z definice derivace a s použitím věty 3.18 platí

$$\frac{\partial \lambda}{\partial a_{ij}} = \lim_{\gamma \to 0} \frac{\lambda(A + \gamma(1)_{ij}) - \lambda(A)}{\gamma} = \frac{y^*(1)_{ij} x}{y^* x}.$$

□

Obecně však vlastní čísla nejsou diferencovatelnými funkcemi prvků matice, viz cvičení 4.

Cvičení

1. Dokončete důkaz věty 3.17.
2. Ukažte bez využití faktu symetrie, že vlastní čísla matice $\tilde{A}$ v příkladu 3.4 jsou reálná.
3. V příkladu 3.5 odvoďte odhad polohy většího vlastního čísla.
4. Uvažujte pro malé $\epsilon \geq 0$ a $N \geq 2$ matici

$$J(\epsilon) = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & & 1 \\ \epsilon & & & & 0 \end{pmatrix},$$

kde všechny nevyznačené prvky jsou nulové.

(a) Určete všechna vlastní čísla matice $J(\epsilon)$.

(b) S použitím výsledku části 4a ukažte, že vlastní čísla matice $J(0)$ nejsou diferencovatelnými funkcemi prvku $(J(0))_{N1}$.

(c) Ukažte, že vlastní čísla matice $J(0)$ nemají konečná čísla podmíněnosti, tj. neexistuje konečné číslo κ takové, že pro libovolné vlastní číslo $\lambda(J(\epsilon))$ platí

$$|\lambda(J(\epsilon)) - \lambda(J(0))| \leq \kappa\, \epsilon$$

pro všechna dostatečně malá $\epsilon > 0$.

3.4 Citlivost vlastních čísel diagonalizovatelných a normálních matic

Z předcházejícího textu je zřejmé, že citlivost vlastních čísel vzhledem k perturbacím matice je silně závislá na vlastnostech původní matice. V této části ukážeme, jak vypadají odhady polohy vlastních čísel perturbované matice v případě, kdy původní matice je diagonalizovatelná nebo normální. Uvidíme, že v těchto odhadech se již neobjevuje N-tá odmocnina normy matice E, jako tomu bylo v případě obecné matice.

Následující věta dává odhad velikosti spektrální variace a párové vzdálenosti pro diagonalizovatelnou matici.

Věta 3.20 *Nechť $A \in C^{N,N}$ je diagonalizovatelná matice, tj. existuje regulární matice $X \in C^{N,N}$ tak, že $X^{-1}AX = \Lambda = diag(\lambda_1, \ldots, \lambda_N)$. Nechť $\tilde{A} = A + E$, kde $E \in C^{N,N}$, $\| \cdot \|_\alpha$ je libovolná konzistentní maticová norma, pro kterou platí $\| diag(\beta_1, \ldots, \beta_N) \|_\alpha = \max_j |\beta_j|$. Pak platí*

$$sv_A(\tilde{A}) \leq \| X^{-1}EX \|_\alpha \leq \kappa_\alpha(X) \| E \|_\alpha , \tag{3.45}$$

$$md(A, \tilde{A}) \leq (2N-1) \| X^{-1}EX \|_\alpha \leq (2N-1)\, \kappa_\alpha(X) \| E \|_\alpha . \tag{3.46}$$

Důkaz: Uvažujme libovolné vlastní číslo $\tilde{\lambda}$ matice $\tilde{A}$, které není vlastním číslem matice A (jinak nastane triviální případ). Z Bauerovy-Fikeho věty pro $Q \stackrel{def}{=} X$ pak dostáváme

$$\| X^{-1}(A - \tilde{\lambda}I)^{-1}X \|_\alpha^{-1} \leq \| X^{-1}EX \|_\alpha .$$

Využijeme-li toho, že $X^{-1}AX = \Lambda$, dostaneme s použitím předpokladů věty

$$\| (\Lambda - \tilde{\lambda}I)^{-1} \|_\alpha^{-1} = \frac{1}{\max_j \frac{1}{|\lambda_j - \tilde{\lambda}|}} = \min_j |\lambda_j - \tilde{\lambda}| \leq \| X^{-1}EX \|_\alpha,$$

a protože $\tilde{\lambda}$ bylo libovolné vlastní číslo matice $\tilde{A}$, platí

$$\max_i \min_j |\tilde{\lambda}_i - \lambda_j| \leq \| X^{-1}EX \|_\alpha \leq \kappa_\alpha(X) \| E \|_\alpha .$$

Odhad (3.46) pro párovou vzdálenost vyplývá z věty 3.10.

□

Je-li matice A normální, je číslo podmíněnosti každého vlastního čísla rovno jedné a můžeme tedy očekávat, že vlastní čísla nejsou citlivá vzhledem k perturbacím matice. Následující odhad je přímým důsledkem předcházející věty.

Věta 3.21 *Nechť $A \in C^{N,N}$ je normální matice. Položme $\tilde{A} = A + E$, kde $E \in C^{N,N}$. Pak pro párovou vzdálenost spekter matic A a $\tilde{A}$ platí*

$$md(A, \tilde{A}) \leq (2N - 1) \parallel E \parallel .$$

Je na místě poznamenat, že faktor $(2N - 1)$ není ani v tomto případě optimální a lze jej zmenšit. Pro nás jsou však uvedené výsledky dostačující. Také se nebudeme speciálně věnovat maticím hermitovským. Do skript jsme nezařadili ani studium citlivosti vlastních vektorů, jde o náročné téma, které by neúměrně rozšířilo rozsah textu. Zvídavého čtenáře odkazujeme na [MPT].

Cvičení

1. Matice X není pro diagonalizovatelnou matici A určena jednoznačně. Otázkou je, jak normovat sloupce matice X, aby $\kappa(X)$ bylo minimální. Tento velmi obtížný problém se zjednoduší, uvažujeme-li podmíněnost matice měřenou ve Frobeniově normě $\kappa_F(X) = \parallel X \parallel_F \parallel X^{-1} \parallel_F$. Pak platí tvrzení:

   Nechť $X \in C^{N,N}$ je regulární a nechť $Y = (X^{-1})^*$, tj. platí $Y^*X = I$. Pak

   $$\kappa_F(X) \geq \sum_{i=1}^{N} \parallel y_i \parallel \parallel x_i \parallel, \tag{3.47}$$

   kde $x_1, \ldots, x_N$ resp. $y_1, \ldots, y_N$ označují sloupce matic X resp. Y. Rovnost nastává tehdy a jen tehdy, existuje-li konstanta $\alpha \neq 0$ tak, že

   $$\parallel y_i \parallel = \alpha \parallel x_i \parallel, \quad i = 1, \ldots, N.$$

   Dokažte.

2. Nechť A je matice s navzájem různými vlastními čísly, a tudíž diagonalizovatelná, $X^{-1}AX = \Lambda$, $Y = (X^{-1})^*$. Ukažte, že optimální hodnota $\kappa_F(X)$, pro kterou platí rovnost ve vztahu (3.47), je rovna součtu individuálních podmíněností jednotlivých vlastních čísel matice A. Z toho vyplývá, že je-li $\kappa_F(X)$ velké (podstatně větší než N), musí být alespoň jedno vlastní číslo špatně podmíněno (věta 3.20 má rozumný smysl).

3.5 Příklady

Jak jsme viděli, má v teorii citlivosti vlastních čísel matic důležitý význam to, jak velká je pro danou matici odchylka od normality. Pro normální matice dávají odhady polohy vlastních čísel perturbované matice velmi příznivé výsledky. Horší výsledky dostáváme pro diagonalizovatelné matice, u nichž jsou analogické odhady závislé na vlastních vektorech dané matice. Nejhorší je situace pro obecnou matici, kdy ani při malé perturbaci prvků nejsme obecně schopni o změně polohy vlastních čísel říci (z praktického hlediska) nic rozumného.

Pro geometrické zobrazení citlivosti vlastních čísel (a popsání vlivu odchylky od normality) je velice užitečný pojem pseudospektra matice, viz [T1], [T2].

Definice 3.13 *Nechť $A \in C^{N,N}$ a položme $\tilde{A} = A + E$, kde $E \in C^{N,N}$. Pro $\epsilon \geq 0$ definujeme* ϵ-**pseudospektrum** *matice A jako*

$$\sigma_\epsilon(A) = \{ \tilde{\lambda} \in C \; ; \; \tilde{\lambda} \text{ je vlastní číslo matice } \tilde{A} = A + E, \; \| E \| \leq \epsilon \}. \tag{3.48}$$

Ekvivalentní definice pseudospektra může vypadat například takto

$$\sigma_\epsilon(A) = \{ \tilde{\lambda} \in C; \; \| (\tilde{\lambda} I - A)^{-1} \| \geq \epsilon^{-1} \}, \tag{3.49}$$

kde pro $\tilde{\lambda} \in \sigma(A)$ definujeme $\| (\tilde{\lambda} I - A)^{-1} \| \stackrel{def}{=} \infty$.

Důkaz ekvivalence výrazů (3.48) a (3.49) ponecháváme čtenáři jako cvičení.

Poznámka 3.10 *Všimněme si, jak vypadá ϵ-pseudospektrum pro normální matici. Je-li A normální matice s vlastními čísly λ_i, $i = 1, \ldots, N$, platí*

$$\| (\tilde{\lambda} I - A)^{-1} \| = \frac{1}{\min_i |\tilde{\lambda} - \lambda_i|}. \tag{3.50}$$

Pseudospektrum $\sigma_\epsilon(A)$ je pak rovno sjednocení kruhů o poloměru ϵ se středy ve vlastních číslech matice A.

Pro obecnou matici je situace mnohem komplikovanější. Neplatí žádný vztah podobný (3.50). A jak uvidíme v následujících příkladech, může číslo $\| (\tilde{\lambda} I - A)^{-1} \|$ dosahovat velkých hodnot i pro $\tilde{\lambda}$ vzdálená od spektra matice A.

Uvedeme několik příkladů převzatých z [T1]. Všechny matice, jejichž pseudospektra budeme studovat jsou řádu $N = 32$. Pro každou matici uvedeme obrázek se dvěma částmi. V první z nich bude zobrazeno 3200 vlastních čísel pro 100 matic, z nichž každá je ve tvaru $\tilde{A} = A + E$, kde E je náhodně generovaná a $\| E \| = 10^{-3}$. Matice E je generována následujícím způsobem. Nejprve je zkonstruována hustá matice $\bar{E}$, jejímiž prvky jsou náhodné veličiny s komplexním normálním rozdělením se střední hodnotou 0 a standardním rozptylem 1. Pak je vypočtena norma $\| \bar{E} \|$ a E je definována jako $E \stackrel{def}{=} 10^{-3} \bar{E} / \| \bar{E} \|$. Druhá část obrázku zachycuje křivky, které tvoří hranice pro ϵ-pseudospektra $\sigma_\epsilon(A)$, kde za ϵ jsou postupně dosazovány hodnoty $10^{-2}, 10^{-3}, \ldots, 10^{-8}$. Přerušovaná čára (někdy

mimo měřítko obrázku a tudíž nezobrazena) je hranicí **pole hodnot matice A (field of values)** definovaného vztahem

$$\mathcal{F}(A) \stackrel{def}{=} \{\, x^*Ax \,,\ \| x \| = 1 \,\}.$$

Vlastní čísla matice A jsou označena výraznými body. Způsob výpočtu zobrazených hodnot je popsán v [T1]. Grafické zobrazení je provedeno pomocí MATLABu.

1. Jordanův blok

Asi nejznámějším příkladem matice jejíž vlastní čísla jsou citlivá vzhledem ke změnám prvků, je následující Jordanův blok

$$A_1 = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix}.$$

Všechna vlastní čísla matic $A_1 + E$ leží v kruhu o středu 0 a poloměru $(10^{-3})^{\frac{1}{32}} \approx 0.8$. Všimněme si, že většina z nich je umístěna velmi blízko hranici pseudospektra $\sigma_{10^{-3}}(A_1)$. Je to důsledek citlivosti vlastních čísel matice A_1 vzhledem k perturbacím prvků. Hranice pseudospekter $\sigma_\epsilon(A_1)$ tvoří soustředné kruhy se středem v počátku, jak je patrné z obrázku 3.6.

2. Modifikovaný Jordanův blok

Hranice pseudospekter matic mohou být tvořeny křivkami velmi odlišnými od soustředných kruhů. Příkladem je matice A_2

$$A_2 = \begin{pmatrix} 0 & 1 & 1 & & & \\ & 0 & 1 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 0 & 1 & 1 \\ & & & & 0 & 1 \\ & & & & & 0 \end{pmatrix}.$$

Příslušná pseudospektra jsou zobrazena na obrázku 3.7.

3. Wilkinsonova matice

Třetím příkladem je tzv. Wilkinsonova matice, která má tvar

$$A_3 = \begin{pmatrix} \frac{1}{32} & 1 & & & \\ & \frac{2}{32} & 1 & & \\ & & \ddots & \ddots & \\ & & & \frac{31}{32} & 1 \\ & & & & 1 \end{pmatrix}.$$

Tato matice má různá vlastní čísla, tedy je příkladem diagonalizovatelné matice. Spektrum Wilkinsonovy matice $\sigma(A_3)$ je tvořeno jejími diagonálními prvky, tedy je reálné. Na druhé straně, $\sigma_{10^{-3}}(A_3)$ obsahuje velké množství čísel s výraznou imaginární složkou. Podmíněnost matice vlastních vektorů je řádu 10^{22}. Srovnejte obrázek (3.8) s odhadem ve větě 3.20.

4. Frankova matice

Frankova matice je typickým příkladem matice, jejíž některá vlastní čísla jsou špatně podmíněná.

$$A_4 = \begin{pmatrix} 32 & 31 & 30 & 29 & \dots & 2 & 1 \\ 31 & 31 & 30 & 29 & \dots & 2 & 1 \\ & 30 & 30 & 29 & \dots & 2 & 1 \\ & & \ddots & \ddots & \ddots & & \\ & & & & & & \\ & & & & 2 & 2 & 1 \\ & & & & & 1 & 1 \end{pmatrix}.$$

Vlastní čísla Frankovy matice opět leží na reálné ose. Wilkinson ukázal [AEP], že malá vlastní čísla matice A_4 mají velká čísla podmíněnosti, tudíž lze očekávat, že budou citlivá vzhledem k perturbacím prvků matice. Tato vlastnost se výrazně projevuje na tvarech příslušných pseudospekter (viz obrázek 3.9). Srovnejte tyto obrázky s poznámkami za větou 3.18.

5. Náhodná matice

Matice A_6 je náhodně generovaná matice, jejíž prvky jsou náhodné veličiny s komplexním normálním rozdělením se střední hodnotou 0 a standardní odchylkou $N^{-\frac{1}{2}}$. Obrázek pro pseudospektrum $\sigma_{10^{-3}}(A_6)$ je zcela odlišný od příslušných obrázků ve všech zatím uvedených příkladech (viz obrázek 3.10). Místo 3200 bodů je zachyceno pouhých 32. Každý z nich totiž představuje svou stonásobnou kopii. To znamená, že perturbace E s normou řádu 10^{-3} nezmění polohu vlastních čísel matice A_6. Také obrázek hraničních křivek pro vybraná pseudospektra se liší od příslušných obrázků z minulých příkladů. Všechna zobrazená pseudospektra jsou totiž relativně malá. Tedy vlastní čísla náhodně generované matice nejsou citlivá vzhledem k perturbacím prvků matice. Tento závěr ovšem nemá tak optimistické důsledky, jak by se mohlo na první pohled zdát. Matice, se kterými se setkáváme při skutečných výpočtech nejsou náhodné. V aplikacích často vznikají matice, které mají velkou odchylku od normality a tudíž jejich vlastní čísla jsou citlivá vzhledem k perturbacím.

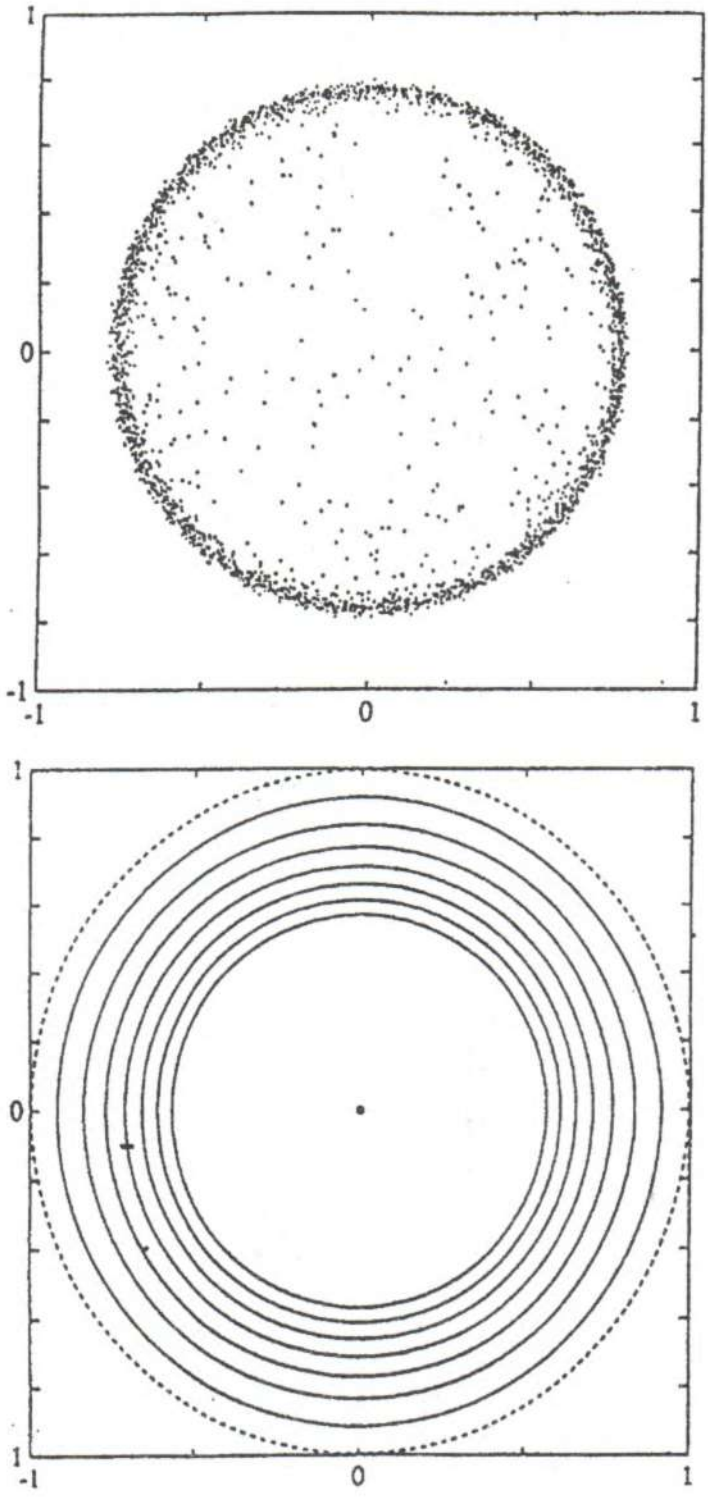
6. Náhodná horní trojúhelníková matice

Nechť matice A_7 je totožná s maticí A_6 s tím rozdílem, že všechny poddiagonální prvky byly nahrazeny nulami. Tato změna má za následek velmi výrazné zvětšení citlivosti vlastních čísel. Příslušná pseudospektra jsou zachycena na obrázku 3.11.

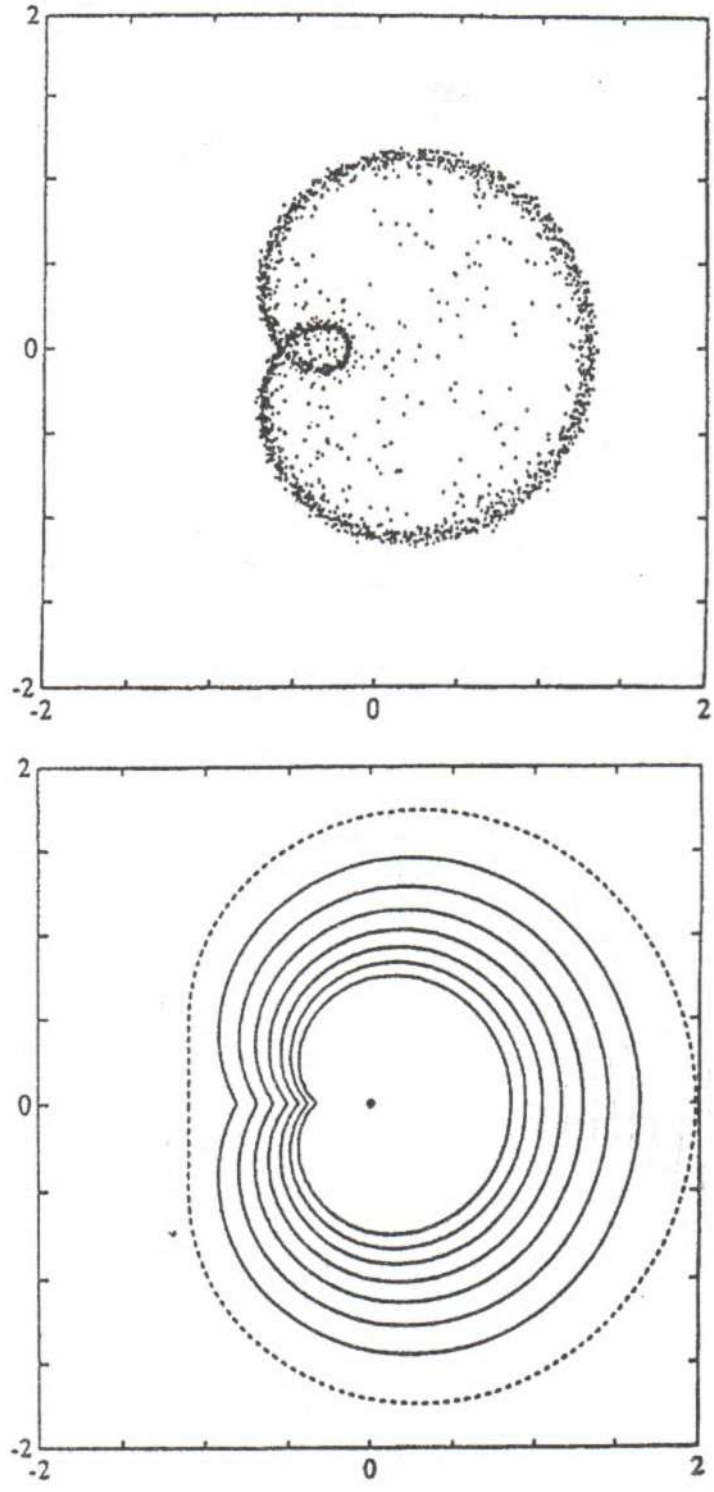
Cvičení

1. Dokažte ekvivalenci definic (3.48) a (3.49).

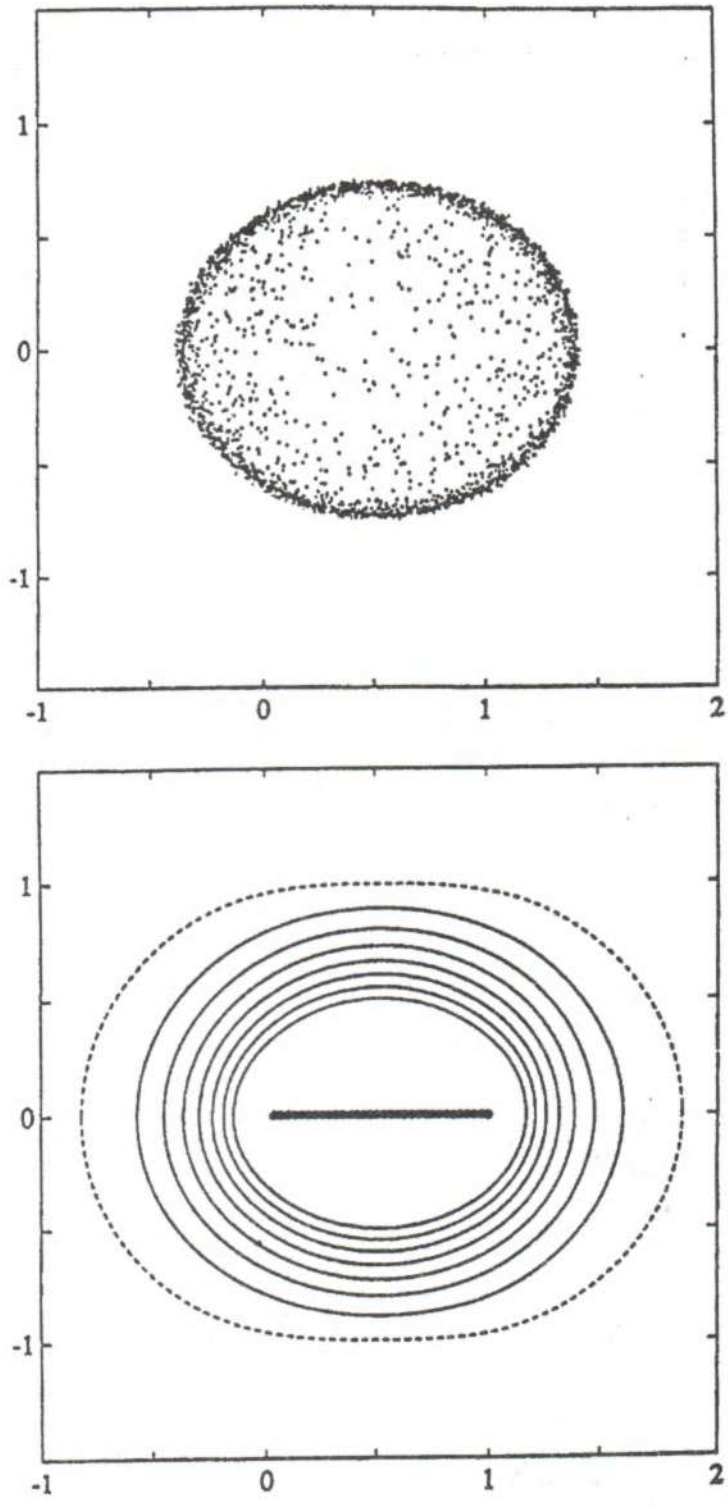
2. S použitím vhodného programového vybavení (např. MATLABu) vypočtěte a graficky znázorněte vzorky pseudospekter matic A_1 až A_6. Řád matic a hodnotu ϵ volte podle potřeby.



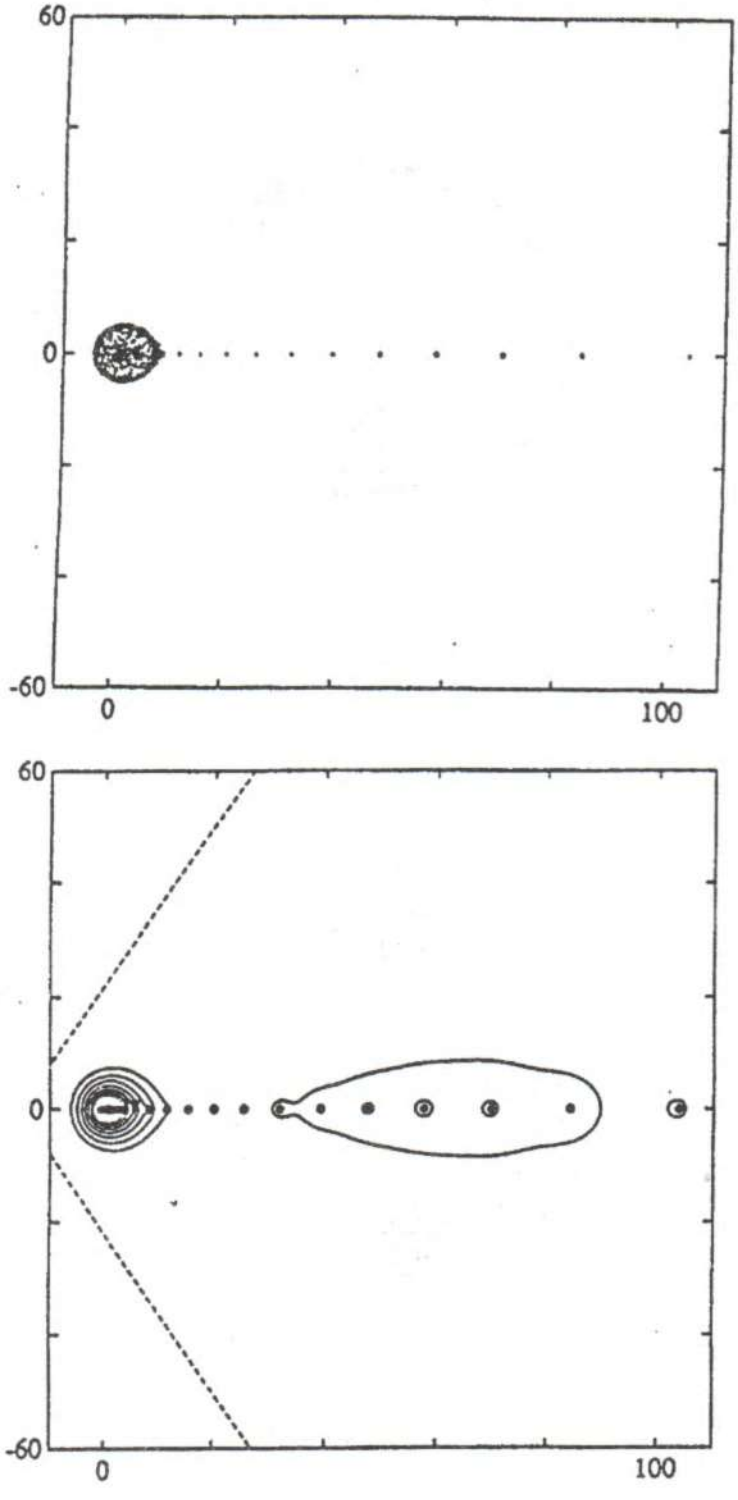
Obr. 3.6: Pseudospektra pro Jordanův blok



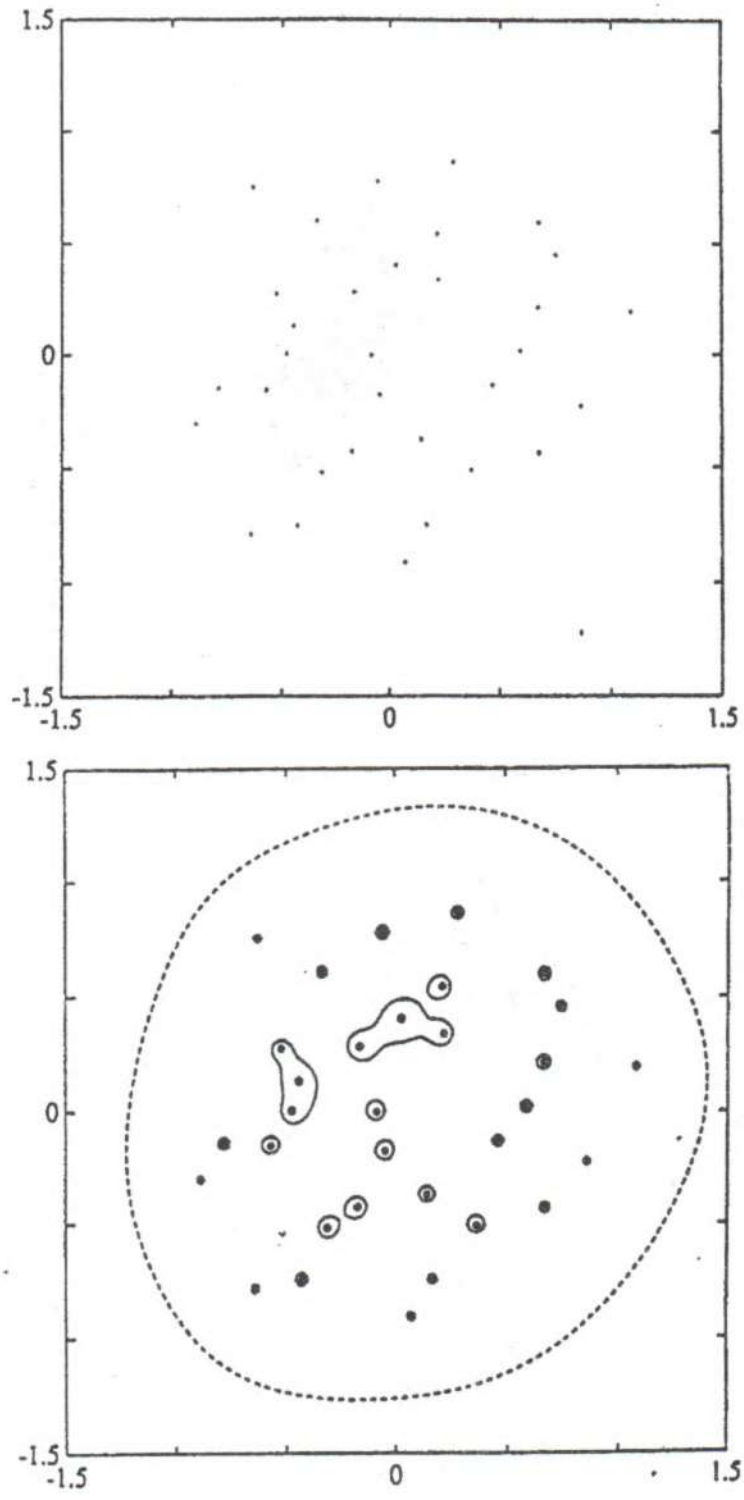
Obr. 3.7: Pseudospektra pro modifikovaný Jordanův blok



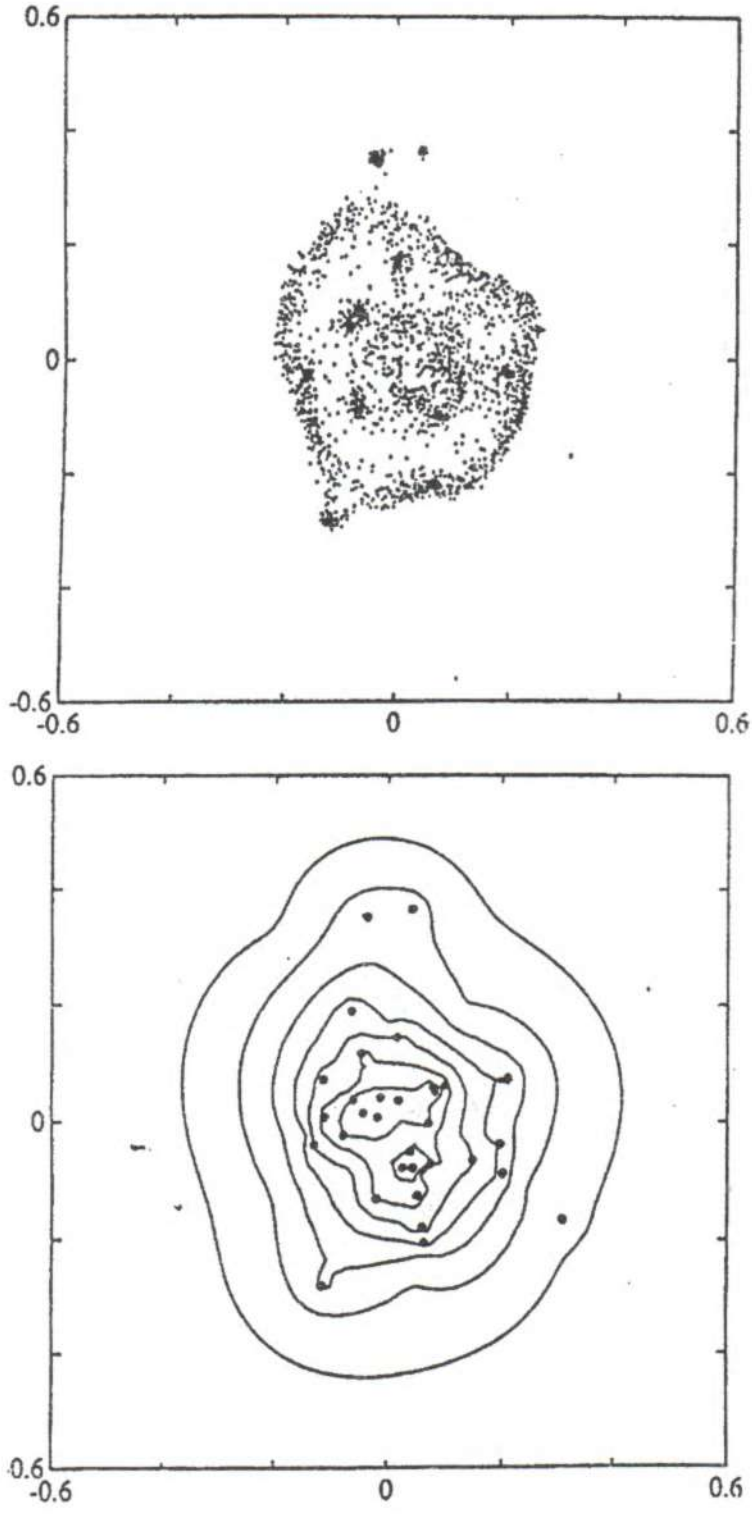
Obr. 3.8: Pseudospektra pro Wilkinsonovu matici



Obr. 3.9: Pseudospektra pro Frankovu matici



Obr. 3.10: Pseudospektra pro náhodnou matici



Obr. 3.11: Pseudospektra pro náhodnou horní trojúhelníkovou matici

Kapitola 4

Citlivost řešení soustav lineárních rovnic

Máme řešit soustavu $Ax = b$, $A \in C^{N,N}$, $b \in C^N$, A je regulární matice. Budeme vyšetřovat, jak je řešení $\tilde{x} = x + \delta x$ soustavy s perturbovanými vstupními daty vzdáleno od řešení x původní soustavy. Budeme se postupně zabývat třemi možnými případy: v prvém je perturbována jen pravá strana soustavy, ve druhém je perturbována jen matice soustavy a ve třetím je perturbována jak pravá strana, tak matice soustavy.

Věta 4.1 *Nechť $A \in C^{N,N}$ je regulární, $b \in C^N$, $b \neq 0$, $\delta b \in C^N$ a platí $Ax = b$, $A(x + \delta x) = b + \delta b$. Pak*

$$\frac{\| \delta x \|}{\| x \|} \leq \kappa(A) \frac{\| \delta b \|}{\| b \|}. \tag{4.1}$$

Důkaz: Z rovnosti $Ax + A\,\delta x = b + \delta b$ dostaneme snadnou úpravou

$$\| \delta x \| \leq \| A^{-1} \| \| \delta b \| .$$

Z rovnice $Ax = b$ je ihned zřejmé

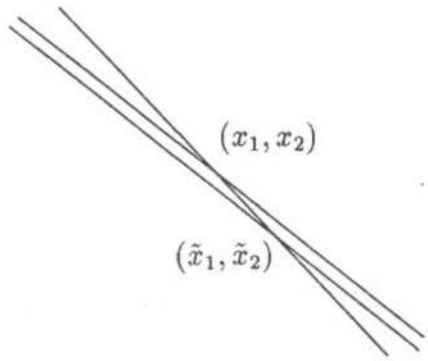
$$\| A \| \| x \| \geq \| b \| .$$

Vydělením obou nerovností dostaneme tvrzení věty.

□

Poznámka 4.1 *Spektrální norma matice je generována euklidovskou normou vektoru, tedy existuje x_0 takové, že $\| A \| = max_{\|x\|=1} \| Ax \| = \| Ax_0 \|$. Pro x_0 nastává ve výraze $\| Ax \| \leq \| A \| \| x \|$ rovnost. Říkáme, že tento odhad je* **ostrý**. *Protože jsme v důkaze věty 4.1 nepoužili jiné nerovnosti, je odhad (4.1) rovněž ostrý.*

Z odhadu (4.1) vyplývá, že pokud je $\kappa(A) \gg 1$ (matice je špatně podmíněná), pak ani při malé perturbaci pravé strany není zaručeno, že řešení $\tilde{x} = x + \delta x$ bude blízké x. Rozhodujícím faktorem je číslo podmíněnosti matice.



Obr. 4.1: Vliv perturbace pravé strany na změnu řešení je-li $\kappa(A) \gg 1$

Příklad 4.1 *Mějme soustavu dvou rovnic o dvou neznámých:*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2. \end{aligned}$$

Je-li $\kappa(A) \gg 1$, jsou vektory (a_{11}, a_{12}) a (a_{21}, a_{22}) téměř lineárně závislé, tedy přímky jimi určené jsou skoro rovnoběžné. Proto například malá perturbace hodnoty b_1 způsobí sice malý posun první přímky, ale ve svém důsledku to znamená velký posun řešení $(\tilde{x}_1, \tilde{x}_2)$ vzhledem k (x_1, x_2) (viz obrázek 4.1).

Uvažujme, že je perturbována jen matice soustavy. Nejprve je třeba nalézt podmínky, za kterých má $(A + \delta A)\tilde{x} = b$ jednoznačné řešení.

Věta 4.2 *Je-li $A \in C^{N,N}$ regulární, $\delta A \in C^{N,N}$ a platí*

$$\frac{\| \delta A \|}{\| A \|} < \frac{1}{\kappa(A)},$$

pak je $A + \delta A$ také regulární.

Důkaz: Podmínku

$$\frac{\| \delta A \|}{\| A \|} < \frac{1}{\kappa(A)} = \frac{1}{\| A \| \| A^{-1} \|}$$

lze vyjádřit ve tvaru $\| A^{-1} \| \| \delta A \| < 1$. Tvrzení věty budeme dokazovat sporem. Je-li $A + \delta A$ singulární, pak existuje nenulový vektor z takový, že $(A + \delta A)z = 0$, tedy $z = -A^{-1}\delta A z$. Odhadem velikostí norem dostaneme

$$\| z \| = \| A^{-1}\delta A z \| \leq \| A^{-1} \| \| \delta A \| \| z \|$$

a vydělením nerovnosti hodnotou $\| z \|$ dostaneme spor.

□

Poznámka 4.2 *Všimněme si, jak souvisí $\kappa(A)$ se vzdáleností A od nejbližší singulární matice: je-li $A+\delta A$ singulární, je $\| \delta A \| / \| A \| \geq 1/\kappa(A)$. Jinými slovy, je-li matice A blízká matici singulární, je špatně podmíněná.*

Poznámka 4.3 *Závěr učiněný v poznámce 4.2 je možné zesílit. Platí totiž následující tvrzení.*

Je-li $A \in C^{N,N}$ regulární, pak existuje matice $\delta A \in C^{N,N}$ tak, že $A + \delta A$ je singulární a

$$\frac{\| \delta A \|}{\| A \|} = \frac{1}{\kappa(A)}.$$

Vidíme, že hodnota $1/\kappa(A)$ nám určuje vzdálenost matice A od množiny singulárních matic. Důkaz uvedeného tvrzení je mimo rámec našeho textu, čtenář jej může nalézt např. v [FMC].

Věta 4.3 *Mějme $A \in C^{N,N}$ regulární, $\delta A \in C^{N,N}$, $b \in C^N$, $b \neq 0$ a nechť je splněno*

$$\frac{\| \delta A \|}{\| A \|} < \frac{1}{\kappa(A)}.$$

Nechť $Ax = b$, $(A + \delta A)(x + \delta x) = b$. Pak platí

$$\frac{\| \delta x \|}{\| x \|} \leq \kappa(A) \frac{\| \delta A \|}{\| A \|} (1 - \kappa(A) \frac{\| \delta A \|}{\| A \|})^{-1}. \tag{4.2}$$

Důkaz: Z rovnice $(A+\delta A)(x+\delta x) = b$ vyjádříme δx a odhadneme velikost jeho normy

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

$$\| \delta x \| \leq \| A^{-1} \| \| \delta A (x + \delta x) \| \leq \| A^{-1} \| \| \delta A \| (\| x \| + \| \delta x \|).$$

Snadnou úpravou dostaneme

$$(1 - \kappa(A) \frac{\| \delta A \|}{\| A \|}) \| \delta x \| \leq \kappa(A) \frac{\| \delta A \|}{\| A \|} \| x \|, \tag{4.3}$$

$1 - \kappa(A)\|\delta A\|/\|A\|$ je podle předpokladu kladné, tedy jím můžeme celou nerovnost vydělit. □

Poznámka 4.4 *Při odvozování odhadu (4.2) byla použita trojúhelníková nerovnost $\| x + \delta x \| \leq \| x \| + \| \delta x \|$. Tento odhad zřejmě není (až na triviální případy) ostrý, neboť δx je dáno perturbací δA a není zaručeno, že je kolineární s x. Odhad (4.2) není proto obecně ostrý.*

Z věty 4.3 je zřejmé, že pokud je A dobře podmíněná a $\| \delta A \| / \| A \|$ je dostatečně malé, pak je $\kappa(A) \| \delta A \| / \| A \| \ll 1$ a tudíž jmenovatel v (4.2) je blízký jedné, takže

$$\| \delta x \| / \| x \| \lessapprox \kappa(A) \| \delta A \| / \| A \| .$$

Pokud je ovšem A špatně podmíněná, je $\| \delta A \| / \| A \| < 1/\kappa(A)$ splněna pouze pro velmi malá δA. Není-li navíc $\| \delta A \| / \| A \| \ll 1/\kappa(A)$, nedostaneme žádný rozumný odhad pro chybu aproximace řešení, protože jmenovatel v (4.2) může být blízký nule.

Příklad 4.2 *Mějme matici $A \in C^{N,N}$ s číslem podmíněnosti $\kappa(A) = 10^6$ a vezměme takovou její perturbaci, pro niž $\| \delta A \| / \| A \| = 2 \times 10^{-7}$. Tedy $\| \delta A \| / \| A \| = 1/(5\ \kappa(A))$ a odhad pro normu chyby řešení (podle věty 4.3) je $\| \delta x \| / \| x \| \leq 1/4$. Navíc, tento odhad není ostrý. Nezaručuje existenci takové perturbace δA, pro niž je $\| \delta x \| / \| x \| = 1/4$. Je varováním, že by v nepříznivém případě k tak velké chybě mohlo dojít.*

Věta 4.4 *Mějme $A \in C^{N,N}$ regulární, $\delta A \in C^{N,N}$, $b \in C^N$, $b \neq 0$, $\delta b \in C^N$ a nechť platí*

$$\frac{\| \delta A \|}{\| A \|} < \frac{1}{\kappa(A)}.$$

Nechť $Ax = b$, $(A + \delta A)(x + \delta x) = b + \delta b$. Pak

$$\frac{\| \delta x \|}{\| x \|} \leq \kappa(A)\left(\frac{\| \delta A \|}{\| A \|} + \frac{\| \delta b \|}{\| b \|}\right)\left(1 - \kappa(A)\frac{\| \delta A \|}{\| A \|}\right)^{-1} \tag{4.4}$$

Důkaz: Z rovnice $(A + \delta A)(x + \delta x) = b + \delta b$ získáme odhad

$$\| \delta x \| \leq \| A^{-1} \| \| \delta b \| + \| A^{-1} \| \| \delta A \| (\| x \| + \| \delta x \|).$$

Po vynásobení pravé strany nerovnosti podílem $\| A \| / \| A \|$, dostaneme

$$\| \delta x \| \leq \kappa(A)\frac{\| \delta b \|}{\| A \|} + \kappa(A)\frac{\| \delta A \|}{\| A \|}(\| x \| + \| \delta x \|).$$

Odhad $\|A\|\|x\| \geq \|b\|$ lze upravit do tvaru $\| x \| / \| b \| \geq 1/ \| A \|$. Dosazením a jednoduchou úpravou získáme tvrzení věty.

□

Všimněme si důležitého poznatku. Zhruba řečeno, jsou-li relativní chyby ve vstupních datech řádu $10^{-\alpha}$ a víme-li, že podmíněnost matice je řádu 10^{β}, můžeme očekávat relativní chybu řešení řádu $10^{-\alpha+\beta}$. Taková úvaha by měla vždy předcházet snaze o numerické řešení problému. Nemá-li řešení rozumný smysl z hlediska primárního problému (fyzikálního, technického, atd.), je třeba hledat chybu ve formulaci úlohy.

Mnohé metody pro řešení lineárních soustav dávají extrémně malá rezidua, často na úrovni strojové přesnosti. Ukážeme, jak je možné získat odhad chyby řešení z vypočteného rezidua. Nechť $\tilde{x}$ je aproximace řešení soustavy $Ax = b$. Označme $\tilde{x} = x + \delta x$. Potom

$$r = b - A\tilde{x} = b - A(x + \delta x) = A\ \delta x,$$

neboli $\delta x = A^{-1}r$. Snadnou úpravou dostaneme

$$\frac{\| \delta x \|}{\| x \|} \leq \kappa(A) \frac{\| r \|}{\| A \| \| x \|} \leq \kappa(A) \frac{\| r \|}{\| b \|}. \tag{4.5}$$

Vidíme, že malé reziduum nezaručuje malou chybu řešení. Důležitým faktorem je zde opět číslo podmíněnosti matice. V následující kapitole uvedeme výsledek umožňující jiný (a poněkud silnější) odhad velikosti chyby na základě spočteného rezidua.

Cvičení

1. Nechť $A \in C^{N,N}$ je regulární matice. Dokažte

$$\kappa(A) = \frac{max\, mag(A)}{min\, mag(A)} \stackrel{def}{=} \frac{\max_{\|x\|=1} \| Ax \|}{\min_{\|x\|=1} \| Ax \|}.$$

2. Nechť $A \in C^{N,N}$ je regulární matice, $a_1, a_2, \ldots, a_N$ jsou její sloupce. Dokažte, že pro libovolné $1 \leq i, j \leq N$ platí

$$\kappa(A) \geq \frac{\| a_i \|}{\| a_j \|}.$$

3. Nechť $A \in C^{N,N}$ je regulární matice. Označme $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_N > 0$ vlastní čísla matic A^*A, AA^*. Dokažte

$$\kappa(A) = \sigma_1/\sigma_N.$$

Kapitola 5

Odhady chyb a zpětná stabilita

Chceme-li odhadovat chyby způsobené zaokrouhlováním (ať už přímo chyby řešení nebo zpětné chyby) a usuzovat z nich na numerickou stabilitu bez znalosti konkrétních vstupních dat (tj. pouze na základě znalosti algoritmu), musíme uvažovat u každé jednotlivé operace ten nejhorší možný případ (největší možnou chybu) a rovněž ten nejhorší možný případ vzájemné kombinace jednotlivých chyb. Při skutečném výpočtu dochází často při vzniku a šíření zaokrouhlovacích chyb k příznivější situaci - chyby často působí proti sobě a výsledná chyba bývá mnohem menší než *a-priori* (bez uvážení vlivu skutečných dat) získaný odhad. Říkáme, že a-priori odhady chyb jsou zpravidla velmi nadhodoceny.

Bylo by proto výhodné získat informaci o chybách až v průběhu či po skončení skutečného výpočtu, tj. *a-posteriori*. Vzhledem k neznalosti řešení ztěží určíme z vypočtených hodnot přímo chybu získané aproximace. Často však lze určit z vypočtených hodnot zpětnou chybu získaného řešení a rovněž získat věrohodnou informaci o podmíněnosti úlohy; s použitím teorie citlivosti nakonec i odhad pro chybu aproximace řešení.

V této kapitole ukážeme, jak lze pomocí rezidua spočítat zpětnou chybu (a tedy analyzovat zpětnou stabilitu) aproximace řešení při výpočtu vlastních čísel matic a při řešení soustavy lineárních rovnic.

5.1 Vlastní čísla

Začneme vyslovením obecné věty, kterou lze snadno dokázat přímým výpočtem.

Věta 5.1 *Nechť* $A \in C^{N,N}$, $X \in C^{N,p}$, $M \in C^{p,p}$, $p \leq N$. *Označme* R *reziduální matici* $R = AX - XM$. *Nechť* $Y^* \in C^{p,N}$ *je matice taková, že* $Y^*X = I$. *Označíme-li*

$$\tilde{A} = A - RY^*, \tag{5.1}$$

pak platí

$$\tilde{A}X = XM. \tag{5.2}$$

Ač je tato věta jednoduchá, její význam je podstatný. Uvažujme $p = 1$. Pak lze znění věty formulovat následovně. Je-li $x \in C^N$, $\| x \| = 1$ a λ je libovolné komplexní číslo, vypočteme reziduum $r = Ax - \lambda x$. Volíme-li $y = x$, pak matice

$$\tilde{A} = A - rx^*$$

má vlastní číslo λ příslušné vlastnímu vektoru x. Neboli, nalezli jsme takovou matici $\tilde{A}$, pro kterou je „získaná" aproximace $\{\lambda, x\}$ přesným vlastním číslem a přesným vlastním vektorem.

Jaký je význam věty 5.1 pro $p > 1$? Pohleďme nyní na matici $\tilde{A}$ jako na lineární operátor z C^N do C^N. Označme $\tilde{\lambda}$ vlastní číslo a $\tilde{z}$ vlastní vektor operátoru $\hat{A}$, který je zúžením operátoru $\tilde{A}$ na podprostor generovaný sloupci matice X. Z podmínky $Y^*X = I$ vyplývá, že matice X i Y mají lineárně nezávislé sloupce. Použijeme-li substituci $\tilde{z} = Xy$, $y \in C^p$, dostaneme jednoduchými úpravami s uvážením (5.2)

$$\tilde{\lambda}Xy = \tilde{\lambda}\tilde{z} = \hat{A}\tilde{z} = \tilde{A}Xy = XMy.$$

Srovnáním prvního a posledního členu a použitím lineární nezávislosti sloupců matice X dostaneme

$$My = \tilde{\lambda}y,$$

čili $\tilde{\lambda}$ je vlastní číslo matice M. Je-li $p \ll N$, lze tedy p vlastních čísel matice $\tilde{A}$ určit s výhodou jako vlastní čísla mnohem menší matice M. Generují-li sloupce matice X invariantní podprostor matice A, tj. platí-li

$$AX = XM,$$

lze uvedený postup použít přímo na matici A.

V závěru této části uvedeme větu, která umožní odhad chyby aproximace vlastního čísla v případě diagonalizovatelné matice.

Věta 5.2 *Nechť $A \in C^{N,N}$ je diagonalizovatelná matice, $X^{-1}AX = \Lambda = diag(\lambda_1, \ldots, \lambda_N)$. Nechť μ je aproximace vlastního čísla a y aproximace odpovídajícího vlastního vektoru matice A, $\| y \| = 1$. Potom existuje $\lambda_i \in \sigma(A)$ takové, že platí*

$$|\lambda_i - \mu| \leq \kappa(X) \| r \|, \quad r = Ay - \mu y. \tag{5.3}$$

Důkaz: Je-li μ vlastním číslem matice A, je tvrzení triviální. V opačném případě s použitím Bauerovy-Fikeho věty dostaneme

$$\| X^{-1}(A - \mu I)^{-1}X \|^{-1} \leq \| X^{-1}EX \|,$$

kde $E = \tilde{A} - A = ry^*$, $\| E \| = \| r \| \| y^* \| = \| r \|$. Uvážíme-li

$$\| X^{-1}(A - \mu I)^{-1}X \|^{-1} = \| [X^{-1}(A - \mu I)X]^{-1} \|^{-1} = \| (\Lambda - \mu I)^{-1} \|^{-1} = \min_{\lambda_j \in \sigma(A)} |\lambda_j - \mu|,$$

dostaneme použitím konzistence spektrální normy tvrzení věty.

□

Je-li matice A normální, je $\kappa(X) = 1$ a odhad chyby aproximace vlastního čísla je dán přímo normou rezidua.

Cvičení

1. Dokažte větu 5.1.
2. Odvoďte tvrzení analogické větě 5.2 s použitím věty 5.1 a výsledků teorie citlivosti.
3. Nechť $A \in C^{N,N}$ je normální, $y \in C^N$, $\| y \| = 1$. Zvolte $\mu = y^*Ay$, vyslovte větu analogickou větě 5.2 a dokažte ji bez použití věty 5.1.

5.2 Soustavy lineárních rovnic

V minulé kapitole jsme ukázali, jak je úloha řešení soustavy $Ax = b$ citlivá vzhledem k perturbacím vstupních dat. K tomu, abychom byli schopni určit odhad chyby spočtené aproximace řešení $\tilde{x}$, musíme tedy nalézt perturbace vstupních dat δA, δb tak, aby platilo $(A + \delta A)\tilde{x} = b + \delta b$. Chceme nalézt co nejmenší perturbace a využít k tomu vypočtených hodnot. To umožňuje následující věta.

Věta 5.3 *(Rigal,Gaches) Nechť $A \in C^{N,N}$ je regulární, $b \in C^N$. Uvažujme soustavu $Ax = b$, $b \neq 0$. Aproximaci jejího řešení označme $\tilde{x}$. Pak existují takové perturbace δA, δb, pro které je $\tilde{x}$ řešením perturbované soustavy*

$$(A + \delta A)\tilde{x} = b + \delta b$$

a platí

$$\min\{\ \nu\ ;\ \frac{\|\ \delta A\ \|}{\|\ A\ \|} \leq \nu;\ \frac{\|\ \delta b\ \|}{\|\ b\ \|} \leq \nu\ \} = \frac{\|\ b - A\ \tilde{x}\ \|}{\|\ A\ \|\|\ \tilde{x}\ \| + \|\ b\ \|} \stackrel{def}{=} \nu_{min}(\tilde{x}). \tag{5.4}$$

Důkaz: Označme $r = b - A\tilde{x}$ a zvolme

$$\delta A = \frac{\|\ A\ \|\|\ \tilde{x}\ \|}{\|\ A\ \|\|\ \tilde{x}\ \| + \|\ b\ \|}\ \frac{r\tilde{x}^*}{\|\ \tilde{x}\ \|^2}. \tag{5.5}$$

Nejdříve ukážeme, že pro takto definované δA platí

$$(A + \delta A)\tilde{x} = b + \delta b. \tag{5.6}$$

Dosazením za δA máme

$$(A + \delta A)\tilde{x} = A\ \tilde{x} + \frac{\|\ A\ \|\|\ \tilde{x}\ \|}{\|\ A\ \|\|\ \tilde{x}\ \| + \|\ b\ \|}\ r\ =\ A\ \tilde{x} + r - \frac{\|\ b\ \|}{\|\ A\ \|\|\ \tilde{x}\ \| + \|\ b\ \|}\ r.$$

Položíme-li

$$\delta b = -\frac{\|\ b\ \|}{\|\ A\ \|\|\ \tilde{x}\ \| + \|\ b\ \|}\ r, \tag{5.7}$$

dostaneme (5.6). Zřejmě $\|\ \delta b\ \| / \|\ b\ \| = \nu_{\min}(\tilde{x})$. Ukážeme, že pro relativní normu perturbace matice $\|\ \delta A\ \| / \|\ A\ \|$ platí totéž. Protože $r x^*$ je matice $N \times N$, platí pro její normu

$$\|\ r\tilde{x}^*\ \| = \max_{\|z\|=1} \|\ r\tilde{x}^* z\ \| = \|\ r\ \| \max_{\|z\|=1} |\tilde{x}^* z|,$$

kde jsme použili základní vlastnosti normy vektoru a faktu, že $\tilde{x}^* z$ je skalár. Ze vztahu $\max_{\|z\|=1} |\tilde{x}^* z| = \|\tilde{x}\|$ konečně dostaneme

$$\|\ \delta A\ \| = \frac{\|\ b - A\ \tilde{x}\ \|}{\|\ A\ \|\ \|\ \tilde{x}\ \| + \|\ b\ \|}\ \|\ A\ \|\ .$$

Nyní zbývá ukázat, že takto definované perturbace jsou skutečně minimální. Důkaz provedeme sporem. Předpokládejme, že existují takové perturbace vstupních dat $\hat{\delta}A$, $\hat{\delta}b$, pro něž je

$$(A+\hat{\delta}A)\tilde{x} = b+\hat{\delta}b, \quad \| \hat{\delta}A \| / \| A \| \leq \nu_{\min}(\tilde{x}), \quad \| \hat{\delta}b \| / \| b \| \leq \nu_{\min}(\tilde{x})$$

a alespoň v jednom případě nastane ostrá nerovnost. Pro $\nu_{\min}(\tilde{x})$ pak máme:

$$\nu_{\min}(\tilde{x}) = \frac{\| b - A\tilde{x} \|}{\| A \| \| \tilde{x} \| + \| b \|} = \frac{\| \hat{\delta}A\ \tilde{x} - \hat{\delta}b \|}{\| A \| \| \tilde{x} \| + \| b \|} \leq \frac{\| \hat{\delta}A \| \| \tilde{x} \| + \| \hat{\delta}b \|}{\| A \| \| \tilde{x} \| + \| b \|} < \nu_{\min}(\tilde{x}),$$

což je spor.

□

Všimněme si, že hodnotu $\nu_{\min}(\tilde{x})$ lze snadno určit. Význam věty 5.3 je podstatný. Aproximace $\tilde{x}$ je určena zpětně stabilním způsobem *tehdy a jen tehdy*, je-li hodnota $\nu_{\min}(\tilde{x})$ malá.

Někdy je výhodné ptát se, jak se zpětná chyba projeví perturbací jednotlivých *prvků* matice a pravé strany. Jsou-li například některé prvky matice či pravé strany nulové, chtěli bychom, aby odpovídající prvky byly nulové i u perturbované matice či pravé strany (dovolujeme změny pouze některých prvků). Analogií věty 5.3 uvažující perturbace po prvcích je následující výsledek.

Věta 5.4 *(Oettli, Prager) Nechť $A \in C^{N,N}$ je regulární, $b \in C^N$, $Ax = b$ a $\tilde{x}$ je aproximace řešení této soustavy. Mějme dále reálnou nezápornou matici $E = (e_{ij})$ řádu N ($e_{ij} \geq 0$ pro $i,j = 1,\ldots,N$) a reálný nezáporný vektor $f = (f_i)$ dimenze N ($f_i \geq 0$ pro $i = 1,\ldots,N$). Označme*

$$\omega_{\min}(\tilde{x}) \stackrel{def}{=} \max_i \frac{|r_i|}{(E|\tilde{x}| + f)_i},$$

kde $r = (r_1,\ldots,r_N)^T = b - A\tilde{x}$, $|\tilde{x}| = (|\tilde{x}_1|,\ldots,|\tilde{x}_N|)^T$ a $\alpha/0$ interpretujeme jako 0 pro $\alpha = 0$ a jako ∞ ve všech ostatních případech. Je-li $\omega_{\min}(\tilde{x}) \neq \infty$, pak existují takové perturbace $\delta A = (\delta a_{ij})$, $\delta b = (\delta b_i)$, pro které je

$$(A+\delta A)\tilde{x} = b+\delta b \tag{5.8}$$

a platí

$$\min\{\ \omega\ ;\ |\delta a_{ij}| \leq \omega\ e_{ij},\ |\delta b_i| \leq \omega\ f_i,\ i,j = 1,\ldots,N\ \} = \omega_{\min}(\tilde{x}).$$

Důkaz: Z definice $\omega_{\min}(\tilde{x})$ platí pro každou složku rezidua $r = b - A\tilde{x}$:

$$|r_i| \leq \omega_{\min}(\tilde{x})\,(E|\tilde{x}| + f)_i, \quad i = 1,\ldots,N.$$

Tedy r lze vyjádřit jako

$$r = D(E|\tilde{x}| + f), \quad D = diag(d_{11},\ldots,d_{NN}),$$

kde pro diagonální matici D platí

$$|D| \leq \omega_{\min}(\tilde{x})\, I, \quad \text{neboli} \quad |d_i| \leq \omega_{\min}(\tilde{x}) \qquad \text{pro } i = 1,\ldots,N.$$

Zvolíme perturbace δA, δb následujícím způsobem

$$\begin{aligned} \delta A &= D\,E\,diag(\frac{\overline{\tilde{x}_1}}{|\tilde{x}_1|},\ldots,\frac{\overline{\tilde{x}_N}}{|\tilde{x}_N|}) \\ \delta b &= -Df, \end{aligned}$$

kde $\overline{\tilde{x}_j}$ je číslo komplexně sdružené k číslu $\tilde{x}_j$. Dosazením zjistíme, že pro takto zvolené perturbace je $\tilde{x}$ řešením (5.8). Je také zřejmé, že pro tyto perturbace platí odhady $|\delta a_{ij}| \leq \omega_{\min}(\tilde{x})\ e_{ij}$, $|\delta b_i| \leq \omega_{\min}(\tilde{x})\ f_i$, $i,j = 1,\ldots,N$.

Zbývá dokázat, že $\omega_{\min}(\tilde{x})$ je optimální. Nechť pro nějaké perturbace $\hat{\delta}A$, $\hat{\delta}b$ a kladné reálné číslo ω je splněno

$$(A + \hat{\delta}A)\tilde{x} = b + \hat{\delta}b$$

$$|\hat{\delta}a_{ij}| \leq \omega\ e_{ij} \quad |\hat{\delta}b_i| \leq \omega\ f_i \quad \text{pro } i,j = 1,\ldots,N.$$

Pro vektor rezidua potom platí

$$|r| = |b - A\tilde{x}| = |\hat{\delta}A\tilde{x} - \hat{\delta}b| \leq |\hat{\delta}A||\tilde{x}| + |\hat{\delta}b| \leq \omega(E|\tilde{x}| + f), \tag{5.9}$$

kde nerovnosti jsou uvažovány po prvcích. Z (5.9) ihned dostáváme

$$\omega \geq \max_{i=1,\ldots,N} \frac{|r_i|}{(E|\tilde{x}| + f)_i} = \omega_{\min}(\tilde{x}).$$

□

Poznámka 5.1 *Zvolíme-li $E = |A|$, tj. $e_{ij} = |a_{ij}|$, $i,j = 1,\ldots,N$, $f = |b|$, dostaneme velikost relativní zpětné chyby po složkách.*

Kombinací věty 5.3 a příslušného odhadu z teorie citlivosti dostaneme odhad pro chybu spočtené aproximace řešení $\| x - \tilde{x} \|$. V našem textu jsme při popisu citlivosti lineárních soustav neuvažovali v dokázaných tvrzeních vliv jednotlivých prvků matice a pravé strany (k větě 5.4 chybí odpovídající věta popisující citlivost). Čtenáře odkazujeme na [MPT].

Cvičení

1. Odvoďte s použitím věty 5.3 odhad pro chybu řešení $\| x - \tilde{x} \|$ a srovnejte výsledek s odhadem v závěru kapitoly 4.

Závěr

Je paradoxem, že s pokrokem technických i programových prostředků roste i tendence k povrchnosti při jejich využívání. V případě numerických výpočtů to znamená přílišné spoléhání se na „schopnosti počítače". V našem textu jsme se pokusili naznačit některé důležité zásady, které by nikdy neměly být opomenuty. Především, chceme-li numericky hledat řešení nějaké úlohy, měli bychom předem vědět, zda tato úloha má matematický smysl a zda je možné numerickým výpočtem dospět k rozumnému řešení. Provádíme-li vlastní numerický výpočet, musí nás zajímat nejen výsledek, ale stejně tak i jeho chyba, lépe řečeno její co nejlepší odhad.

Otázka numerické stability a odhadování chyb je velmi složitá a neexistují zde snadné a jednoduché návody. Vždy je však dobré dodržovat při výběru algoritmu a vytváření programu následující zásady:

1. Vyhýbejte se odečítání hodnot zatížených chybami.
2. Minimalizujte hodnoty mezivýsledků ve srovnání s hodnotou očekávaného výsledku. Velké mezivýsledky vždy hrozí ztrátou přesnosti.
3. Pamatujte, že matematicky ekvivalentní algoritmy nejsou zpravidla numericky ekvivalentní. Hledejte vždy stabilní způsoby řešení a cesty, jak zvýšit přesnost získané aproximace řešení (například metodou iteračního zpřesnění).
4. Transformujete-li úlohu, vyhýbejte se špatně podmíněným transformacím. Kde je to možné, užívejte unitární transformace.

Vždy si buďte vědomi nebezpečí zaokrouhlovacích chyb a numerické nestability. Honba za co nejmenším počtem aritmetických operací a co nejrychlejší paralelní implementací ztratí smysl, produkuje-li náš „superrychlý" algoritmus nesmysly.

Řešíme-li problém vlastních čísel, musíme mít na paměti, že se snažíme spočítat něco, co v principu nelze konečným postupem přesně určit. Navíc, máme-li počítač charakterizovaný zaokrouhlovací jednotkou u, pak v nejlepším případě určíme vlastní čísla matice $A+E$, kde velikost perturbace je řádu u. Výsledek našeho výpočtu je tedy v nejlepším případě vzorek z u-pseudospektra matice A. Je-li matice A normální, je tento vzorek blízký vlastním číslům matice A. Je-li však odchylka od normality matice A velká, jen Bůh ví, co jsme vlastně spočítali. S problémy se můžeme setkat i při řešení soustav lineárních rovnic.

Dobrá metoda nám zaručí malou zpětnou chybu. Dá-li nám rovněž dostatečně přesnou aproximaci řešení, to závisí na podmíněnosti úlohy. V každém případě je velmi žádoucí využívat a-posteriori odhadů chyb vždy, kdy nelze zaručit kvalitu získané aproximace

řešení jiným způsobem. Vzorovou ukázku profesionálního a poučeného přístupu k řešení některých úloh numerické lineární algebry může čtenář nalézt v [LAP].

Otázce numerického programového vybavení a stabilitě jednotlivých numerických metod se, jak doufáme, budeme věnovat v některém z dalších učebních textů.

Literatura

[T1] Trefethen, L.N. : Pseudospectra of matrices, D.F. Griffiths and G.A. Watson, Eds., Numerical Analysis 1991, Longman Sci, Tech. Publ., pp. 243-266, 1992.

[T2] Trefethen, L.N. : Approximation theory and numerical linear algebra, in J.C. Mason and M.G. Cox, eds., Algorithms for Approximation II, Chapman and Hall, London, 1990.

[LAP] Anderson, E., et al. : LAPACK users' guide (Second Edition), SIAM, Philadelphia, 1995.

RNDr. Jitka Drkošová, Ing. Zdeněk Strakoš, CSc.

ÚVOD DO TEORIE CITLIVOSTI A STABILITY V NUMERICKÉ LINEÁRNÍ ALGEBŘE

VydaloVydavatelství ČVUT, Zikova 4, 166 35 Praha 6,
v lednu 1997 jako svou 8752. publikaci.
Vytisklo Ediční středisko ČVUT, Zikova 4, Praha 6.
74 strany, 9 obrázků.
Vydání první. Náklad 100 výtisků. Rozsah 4,95 AA, 5,28 VA.