

Akademie věd České republiky

Ústav informatiky a výpočetní techniky

# Úvod do teorie citlivosti a stability v numerické lineární algebře

Jitka Drkošová

Zdeněk Strakoš

1996

# Obsah

Úvod	1
<b>1 Citlivost úlohy</b>	<b>5</b>
<b>2 Numerická stabilita metody (algoritmu)</b>	<b>7</b>
2.1 Aritmetika s pohyblivou řádovou čárkou . . . . .	7
2.2 Vliv zaokrouhlovacích chyb při výpočtech v aritmetice s konečnou přesností .	10
2.3 Přímá a zpětná stabilita . . . . .	13
<b>3 Citlivost vlastních čísel matic</b>	<b>15</b>
3.1 Schurova dekompozice, spektrální rozklad a Jordanův kanonický tvar . . . .	16
3.2 Citlivost vlastních čísel pro obecné matice . . . . .	22
3.2.1 Spojitost vlastních čísel . . . . .	23
3.2.2 Elsnerova a Ostrowského-Elsnerova věta . . . . .	23
3.2.3 Bauerova-Fikeho a Henriciho věta . . . . .	30
3.3 Citlivost jednoduchého vlastního čísla . . . . .	35
3.4 Citlivost vlastních čísel pro diagonalizovatelné a normální matice . . . . .	43
3.5 Příklady . . . . .	44
<b>4 Citlivost řešení soustav lineárních rovnic</b>	<b>55</b>
<b>5 Odhady chyb a zpětná stabilita</b>	<b>59</b>
5.1 Vlastní čísla . . . . .	59
5.2 Soustavy lineárních rovnic . . . . .	59

# 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 a nezbytné informace o hledaném řešení, jeho existenci, jednoznačnosti a dalších vlastnostech. Velmi zřídka však jsme schopni řešení analytickými prostředky také nalézt. Je-li například problém formulován prostřednictvím diferenciálních či integrodiferenciálních rovnic 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*. Konečně 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 konečném 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, mnohdy je vhodné hledat pouze vhodnou aproximaci řešení (například iteračním postupem). V této souvislosti hovoříme o *chybě metody*. Konečně výpočty provádíme na počítači s konečnou aritmetikou, což přináší *zaokrouhlovací chyby*. Jejich studium je velmi podstatnou součástí numerické lineární algebry.

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, je velmi výhodné studovat přitom *citlivost řešené úlohy* vzhledem k malým změnám vstupních dat.

Předložený text se nezabývá otázkou stability jednotlivých metod (to bude součástí samostatného navazujícího textu). Pokusili jsme se v něm 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 zao-

krouhlovacích chyb je rozpoznat, která úloha je snadno a která úloha je obtížně numericky řešitelná a pochopit numerické chování jednotlivých metod. 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 zao-krouhlovacími chybami, je dáno jednak snahou po zdůraznění této mnohdy opomíjené sou-části 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 posuzo-vání 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 či jeho části shledán dobrým a užitečným, je to zásluhou literatury, ze které jsme čerpali.

Zdrojem informací nám byly zejména následující 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, že v nejbližší době (koncem roku 1996 či začátkem roku 1997) vyjde další monografie

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

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

- [LA] Pytlíček, J.: Lineární algebra, FJFI ČVUT, Praha

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

a jako základní knihu 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: Matrix Computations (Second Edition), The Johns Hopkins Univ. Press, Baltimore, 1989,
- [MA] Horn, A.G., Johnson, C.R.: Matrix Analysis, Cambridge University Press, Cambridge, 1985,
- [MA] 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 chyby a nedostatky předloženého učebního textu odpovídají plně jeho autoři. Budeme vděční za jakékoli poznámky a připomínky. Internetová adresa autorů je [jitka@uivt.cas.cz](mailto:jitka@uivt.cas.cz), [strakos@uivt.cas.cz](mailto: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 se dá 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  $U(z_1, \dots, z_m)$  přesné řešení úlohy  $U$  se vstupními daty  $(z_1, \dots, z_m)$ . Úloha nebude citlivá na změny vstupních dat, pokud změna

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

bude přiměřená vzdálenosti vstupních dat

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

Úmyslně zde nezávádíme 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. V tomto případě se říká, že úloha je *dobře podmíněná*. Dále se naučíme charakterizovat podmíněnost úloh kvantitativně, tj. velikostí k tomu určených parametrů.

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.

### 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  $y = -2$ .*

*Jak uvidíme dále, 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 popíšeme, jak jsou vlastní čísla čtvercových matic citlivá na změny jednotlivých prvků matic 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 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 úíření 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 a několik velikostí k tomu určených paměťových míst. Nás především zajímá, jak jsou v počítači zobrazena reálná čísla.

Je zřejmé, že při zvoleném typu zobrazení a 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  $\mathcal{F} \subset \mathbb{R}$ , kterou nazýváme soustavou čísel s *pohyblivou řádovou čárkou* (floating point number system). Její prvky lze zapsat ve tvaru

$$y = \pm m \times \beta^{e-t} \quad (2.1)$$

kde celé číslo  $\beta$  (obvykle  $\beta = 2$ ) je nazýváno *základnou*, 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  $\beta$ ,  $t$  a horní resp. dolní mezí celočíselného exponentu,  $e_{\min} \leq e \leq e_{\max}$ .

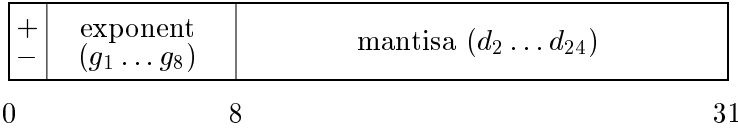
Vztah (2.1) můžeme přepsat do názornějšího tvaru

$$y = \pm \beta^e \left( \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \dots + \frac{d_t}{\beta^t} \right) = \beta^e x 0.d_1 d_2 \dots d_t \quad (2.2)$$

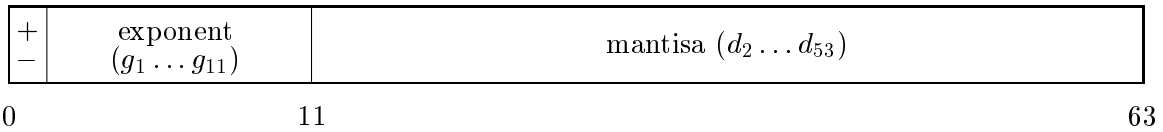
kde každá číslice  $d_i$  leží v intervalu  $0 \leq d_i \leq \beta - 1$ , t.j.  $0.d_1 d_2 \dots d_t$  představuje číslo zapsané v číselné soustavě se základem  $\beta$ . Je výhodné uvažovat  $m \geq \beta^{t-1}$  pro  $y \neq 0$ ; pak zřejmě  $d_1 \neq 0$  pro  $y \neq 0$  a systém dodržující tuto konvenci nazýváme *normalizovaný*. I když se v

minulosti používaly různé základy  $\beta$  (do dnešní doby jsou široce rozšířené základy hodnoty  $\beta = 2$  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 je vyložíme a další vlastnosti aritmetiky s pohyblivou čárkou budeme popisovat na tomto příkladu.

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



dvojitá přesnost



V případě jednoduché přesnosti je na exponent vyhrazeno 8 bitů, do kterých je možno uložit celé číslo v rozmezí  $0 - 255$ . Řetězce  $(0000\ 0000)_2 = 0$  či  $(11111111)_2 = 255$  mají však speciální význam (který bude popsán dále). Zbývající čísla  $1 - 244$  určují hodnotu veličiny

$$e + 126,$$

t.j. hodnota exponentu  $e$  se pohybuje v rozmezí  $-125 \leq e \leq 128$ . 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

$$y = \pm 2^{(g_1 \dots g_8)_2 - 126} x(0.1d_2d_3 \dots d_{24})_2, \quad (2.3)$$

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

Čísla v pohyblivé řádové čárce nejsou vzhledem k  $\mathbb{R}$  rovnoměrně rozložena (viz Cvičení 2.1). Rozdíl 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  $F$ ,  $y_2 > y_1$ , pak  $y_2 - y_1 = 2^e \cdot 2^{-24}$ . Rozložení čísel množiny  $F$  je charakterizováno pomocí **strojové přesnosti**  $\varepsilon_M$ , což je vzdálenost čísla 1.0 od nejbližšího většího čísla v pohyblivé řádové čárce. Zřejmě platí  $\varepsilon_M = 2^1 \cdot 2^{-24} = 2^{-23}$ . Snadno lze ukázat, že vzdálenost libovolného normalizovaného čísla  $x \in F$  od svých sousedů je nejméně  $\varepsilon_M|x|/2$  a nejvýše  $\varepsilon_M|x|$  (cvičení 2).

Pokud by množina  $F$  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  $F$  rovněž tzv. čísla *subnormální*, což jsou nenulová nenormalizovaná čísla s exponentem  $(0000\ 0000)_2 = 0$ , definované vztahem

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

neboli

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

Je-li řetězec mantisy ( $d_2 d_3 \dots d_{2n} = (0 \dots 0)$ ) i exponentu nulový ( $(g_1 \dots g_8) = 0$ ), dostaneme reprezentaci čísel  $\pm 0$  ( $+0$  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 nulový a řetězec mantisy nenulový (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
$(0000 \ 0000)_2 = (0)_{10}$	$\pm(0.d_1 d_2 d_3 \dots d_{23})_2 \times 2^{-126}$
$(0000 \ 0001)_2 = (1)_{10}$	$\pm(0.1d_2 d_3 \dots d_{24})_2 \times 2^{-125}$
↓	↓
$(0111 \ 1110)_2 = (126)_{10}$	$\pm(0.1d_2 d_3 \dots d_{24})_2 \times 2^0$
$(0111 \ 1111)_2 = (127)_{10}$	$\pm(0.1d_2 d_3 \dots d_{24})_2 \times 2^1$
↓	↓
$(1111 \ 1110)_2 = (254)_{10}$	$\pm(0.1d_2 d_3 \dots d_{24})_2 \times 2^{128}$
$(1111 \ 1111)_2 = (255)_{10}$	$\pm\infty$ pokud $d_2 = d_3 = \dots = d_{24} = 0$ , NaN jinak

Zbývají dvě otázky: jaká je přesnost zobrazení reálného čísla v množině  $F$  a jaká je přesnost provádění elementárních aritmetických operací  $+$ ,  $-$ ,  $\times$  a  $/$ . Přesnost aproximace je charakterizována **zaokrouhlovací jednotkou**  $n = 1/2\beta^{1-t} = 1/22^{-23} = 2^{-24}$ . Důkaz následující věty (která je formulována obecně) ponecháme do cvičení:

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

$$fl(x) = x(1 + \delta), \quad |\delta| < n. \quad (2.4)$$

Aritmetické operace  $+$ ,  $-$ ,  $\times$  a  $/$  se v IEEE 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

$$\begin{aligned} fl(x \pm y) &= (x \pm y)(1 + \tilde{\epsilon}_1) & |\tilde{\epsilon}_1| &\leq u \\ fl(xy) &= (xy)(1 + \tilde{\epsilon}_2) & |\tilde{\epsilon}_2| &\leq u \\ fl(x/y) &= (x/y)(1 + \tilde{\epsilon}_3) & |\tilde{\epsilon}_3| &\leq u \end{aligned} \quad (2.5)$$

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

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

Vlastnosti aritmetiky s pohyblivou 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.

přesnost	počet bitů celkově	mantisa	exponent	zaokrouhlovací jednotka $n$	rozsah
jednoduchá	32	23(+1)	8	$2^{-24} \sim 5.96 \times 10^{-8}$	$10^{\pm 38}$
dvojitá	64	52(+1)	11	$2^{-53} \sim 1.11 \times 10^{-6}$	$10^{\pm 308}$

## Cvičení

- Vypočtete a graficky znázorněte na číselné ose prvky množiny čísel s pohyblivou řádovou čárkou pro  $\beta = 2$ ,  $t = 3$ ,  $e_{\min} = -1$  a  $e_{\max} = 3$ .
- Ukažte, že vzdálenost libovolného normalizovaného čísla  $x$  množiny  $\mathcal{F}$  od svého nejbližšího souseda je nejméně  $\varepsilon_M |x|/2$  a nejvýše  $\varepsilon_M |x|$ .
- Dokažte větu 2.1.
- Odvoďte parametry IEEE aritmetiky v pohyblivé řádové čárce, dvojitě přesnosti.
- 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 vyjímecná situace?

- $fl(a \text{ op } b) = fl(b \text{ op } a)$  op = +, \*
- $fl(b - a) = -fl(a - b)$
- $fl(a + a) = fl(2 * a)$
- $fl(0.5 * a) = fl(a/2)$
- $fl((a + b) + c) = fl(a + (b + c))$
- je-li  $a \leq b$ , pak  $a \leq fl((a + b)/2) \leq b$ .

## 2.2 Vliv zaokrouhlovacích chyb při výpočtech 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 výjimkou

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. 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 desítných míst rovno  $c = 0.9999\ 9999\ 99$ , takže vyčíslením hodnoty  $f(1.2 \times 10^{-5})$  dostaneme

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

což je úplně špatně, neboť  $0 \leq f(x) \leq 1/2$  pro  $x \neq 0$ . Vidíme, že i když hodnota  $\cos x$  byla aproximována s přesností na 10 desítných míst, výsledek výpočtu hodnoty  $f(x)$  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 hodnoty  $C$ . Tím se významnost nepatrné chyby hodnoty  $C$  posunula o 10 řádů a katastrofálně ovlivnila celý další výpočet, byť byl proveden sebelepší (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, t.j.  $\hat{x} = x(1 + \Delta x)$ ,  $\hat{y} = y(1 + \Delta y)$ . Předpokládejme, že chyby  $\Delta x$  resp.  $\Delta y$  jsou malé vzhledem k velikosti  $x$  resp.  $y$ ; můžeme 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  $|\Delta x| \leq n$ ,  $|\Delta y| \leq n$ ). Provedme *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 jasné, že i když hodnoty  $\Delta x$  a  $\Delta y$  jsou malé, není zaručeno, že hodnota  $\Delta s$  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  $\Delta s$  relativně velká. Znovu vidíme, že krácení je nebezpečné nikoliv samo o sobě (dojde-li ke krácení při odečtení dvou přesných hodnot, žádná ztráta přesnosti nenastane), ale tím, že zesiluje vliv předchozích chyb, obsažených v datech.

Druhý příklad ukazuje, že i bez krácení popsaného výše můžeme dojít při provedení jednoduchého výpočtu k velké chybě. Předpokládejme, že chceme nalézt dobrou numerickou aproximaci hodnoty  $e$  s použitím vztahu  $e = \lim_{n \rightarrow \infty} (1 + 1/n)^n$ , kde limitu nahradíme prostým výpočtem hodnoty  $f(n) = (1 + 1/n)^n$  pro dostatečně velké  $w$ . Použijeme-li ale hodnotu  $n = 10$ , pak v případě IEEE aritmetiky a jednoduché přesnosti dostaneme lepší aproximaci čísla  $e$  než pro  $n = 10$  (viz cvičení 2.2.2)! Příčina je následující. Sčítáme-li  $1 + 1/n$  pro  $n \gg 1$ , obsahuje výsledek součtu stále méně a méně informace o čísle  $n$  (neboť  $1 \gg 1/w$ ). I když provedeme následné umocnění přesně, výsledek je zatížen velkou chybou.

Posledním příkladem je sčítání řad s kladnými členy. Z teorie Fourierových řad je známo, že

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

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

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

kde  $m$  určíme jako nejmenší celé číslo, jehož zahrnutí do výpočtu nezmění vypočtený součet. Výsledek výpočtu bude překvapivě nepřesný (viz Cvičení 2.2.3). Příčina je opět zřejmá: řada konverguje velmi pomalu a náš výpočet je prováděn tak, že hodnota přičítaných prvků se stále zmenšuje  $\rightarrow$  pro jisté  $m$  pak je vypočtený částečný součet  $\sum_{k=1}^{m-1} k^{-2}$  takový, že přičtení  $m^{-2}$  nezmění jeho hodnotu; zbytek  $\sum_{k=n}^{\infty} 1/k^2$  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 drahé operace a nelze ji v praktických výpočtech použít. Univerzálním řešením je použití speciálních technik zvyšujících přesnost (samozřejmě na úkor rychlosti). Zvidavého čtenáře odkazujeme na [ASNA], kapitolu 4. Jiným řešením může být použití vhodné identity a řady konvergující podstatně rychleji, viz cvičení 3. V každém případě je vhodné zamyslet se nad konvergencí sčítané řady. Odstraňujícím případem budiž všem eventuální pokus nalézt výše popsaným postupem součet řady“  $\sum_{k=1}^{\infty} \frac{1}{k!}$

## Cvičení

- Ukažte, jak je potřeba přepsat uvedené výrazy, aby byl omezen vliv krácení platných cifer
  - $\sqrt{x+1} - 1$  pro  $x \sim 0$
  - $\sin x - \sin y$  pro  $x \sim y$
  - $x^2 - y^2$  pro  $x \sim y$
  - $(1 - \cos x)/\sin x$  pro  $x \sim 0$
- 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.
- Vypočtěte hodnotu výrazu  $(1 + 1/n)^n$  pro  $n = 10^1, 10^2, \dots, 10^7$  a srovnajte s výsledky s hodnotou  $e$ .
- Vypočtěte  $\sum_{n=1}^{\infty} \frac{1}{n^2+1}$  s přesností větší než  $10^{-6}$ .  
K určení počtu členů  $m$  použijte  $\int_m^{\infty} \frac{dx}{x^2+1}$ .
- Vypočtěte  $\sum_{n=1}^{\infty} \frac{1}{n^2+1}$  s přesností větší než  $10^{-6}$ .  
Nesčítejte původní řadu, ale použijte identitu

$$\sum_1^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \quad \sum_1^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}.$$

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

Zcela analogicky se postupuje i při odhadu velikosti chyby pro odčítání.

## 2.3 Přímá a zpětná stabilita

Vraťme se k úloze  $U$  se vstupními daty  $(z_1, \dots, z_m)$  z kapitoly 1. Nechť  $C$  označuje algoritmus pro řešení této úlohy. Označme  $C(z_1, \dots, z_m)$  výstup algoritmu  $C$  použitého na vstupní data  $(z_1, \dots, z_m)$  při (hypotetickém) výpočtu v přesné aritmetice, předpokládáme  $C(z_1, \dots, z_m) = U(z_1, \dots, z_m)$ . Výsledek odpovídajícího výpočtu v konečné aritmetice označíme jako  $fl(C(z_1, \dots, z_m))$ .

Zajímá nás chyba výpočtu způsobená zaokrouhlováním v aritmetice s konečnou přesností, t.j. rozdíl

$$fl(C(z_1, \dots, z_m)) - C(z_1, \dots, z_m). \quad (2.6)$$

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 násobením matice vektorem apod.).

- *Zpětná analýza chyb.* Hledáme taková data  $(\tilde{z}_1, \dots, \tilde{z}_m)$ , aby řešení původní úlohy  $U(z_1, \dots, z_m)$  získané algoritmem  $C$  v konečné aritmetice bylo *totožné* s řešením úlohy  $U(\tilde{z}_1, \dots, \tilde{z}_m)$  získané algoritmem  $C$  při výpočtu v přesné aritmetice. Jinak řečeno: chceme určit vstupní data  $(\tilde{z}_1, \dots, \tilde{z}_m)$  taková, že

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

Vidíme, že cílem zpětné analýzy je interpretovat zaokrouhlovací chyby vzniklé při výpočtu v konečné přesnosti pomocí změn vstupních dat.

**Příklad 2.1** Předpokládejme, že čísla  $x$ ,  $y$ ,  $z$  jsou zobrazena v konečné aritmetice přesně. Pro součet v aritmetice s pohyblivou řádovou čárkou pak platí

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

kde jsme položili

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

$|\delta_1| \leq n \ll 1$ ,  $|\delta_2| \leq n \ll 1$ . Zřejmě tedy  $|\delta_3| \sim |\delta_1| + |\delta_2| \ll 1$  a  $\tilde{x} = x(1 + \delta_3)$ ,  $\tilde{y} = y(1 + \delta_3)$ ,  $\tilde{z} = z(1 + \delta_3)$  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,

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

Uvědomme si, že jde o velmi podstatnou věc. Zpětná stabilita umožňuje oddělit popis chování algoritmu vzhledem k zaokrouhlovacím chybám (popis numerické stability algoritmu) od popisu citlivosti řešené úlohy. Tím je možné poznat, kdy za velkou chybu řešení odpovídá špatná volba algoritmu (jeho nestabilita) a kdy je chyba jen nevyhnutelným důsledkem špatných vlastností samotné úlohy. Kapitulu ukončíme neformální definicí zpětné stability.

**Definice 2.1** Algoritmus  $C$  pro řešení úlohy  $U(z_1, \dots, z_m)$  nazveme zpětně stabilní, pokud platí

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

data  $\tilde{z}_1, \dots, \tilde{z}_m$  jsou v jistém smyslu blízká původním datům  $z_1, \dots, z_m$ . 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  $A$  je čtvercová komplexní matice a  $E$  je čtvercová matice stejného rozměru, 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 podstatným způsobem na vlastnostech matice  $A$ .

Nejdříve proto popíšeme vlastnosti některých tříd matic a uvedeme tvrzení, která budeme při studiu citlivosti vlastních čísel potřebovat.

V dalším budeme používat následující označení.

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

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

*generovanou skalárním součinem*

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

*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 matice  $A$ .*

**Spektrální normu** (generovanou euklidovskou normou vektoru) pak označíme jako  $\|A\|$  a platí pro ni

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

*Symbolu  $A^*$  používáme pro matici hermitovsky sdruženou maticí  $A$ , tedy platí*

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

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

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

### 3.1 Schurova dekompozice, spektrální rozklad a Jordanův kanonický tvar

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

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

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

Uvažujeme případ, kdy matice  $A$  je zatížena chybami ve skutečnosti tedy provádíme podobnostní transformaci matice  $A + \Delta A$ . Co se stane s velikostí chyb při podobnostní transformaci? S použitím odhadu

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

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 transformace, 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

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

Pokusíme se na chvíli omezit 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  $R = U^*AU$  je horní trojúhelníková. Matice  $U$  může být zvolena tak, aby diagonála matice  $R$  obsahovala vlastní čísla matice  $A$  v předepsaném pořadí.

**Důkaz:** Důkaz provedeme indukcí podle dimenze  $n$  matice  $A$ . Pro  $n = 1$  je platnost tvrzení zřejmá. Předpokládejme, že tvrzení věty platí pro všechny matice až do řádu  $n$  včetně. Nechť je dáno uspořádání vlastních čísel matice  $A$  a  $\lambda$  je první vlastní číslo v tomto uspořádání. Bez újmy obecnosti předpokládejme, že příslušný vlastní vektor je normovaný. Tedy platí

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

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

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

kde  $x \in C^{n+1}$  a  $X \in C^{n+1, n}$ .

Pro matici  $A \in C^{n+1, n+1}$  pak platí:

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

nebo  $X^*Ax$  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. (Viz např. [?], str. 37). Tudíž  $\sigma(M) = \sigma(A) \setminus \{\lambda\}$ . Podle indukčního předpokladu existuje unitární matice  $V$  taková, že  $V^H M V$  je horní trojúhelníková s vlastními čísly v předepsaném pořadí. Položíme-li

$$U = \begin{pmatrix} x & X V \end{pmatrix}.$$

Pak

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

je hledaný rozklad. □

**Definice 3.2** Rozklad  $A = URU^*$  budeme nazývat Schurovým rozkladem 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. Její výpočet je předmětem 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  $A^*A = AA^*$ , t.j. matice komutuje se svojí maticí sdruženou.

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

**Důkaz:** Tvzení dokážeme opět indukcí podle dimenze 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  $n + 1$  dimenzionální Schurovy transformace

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

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)

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

Tedy musí platit

$$|\rho|^2 = |\rho|^2 + r^T \bar{r}, \quad (3.2)$$

z čehož plyne  $r = 0$ . Srovnáním bloků (2,2) dostaneme

$$R_1^*R_1 = R_1R_1^*. \quad (3.3)$$

Z rovnice (3.2) vyplývá, že  $R_1$  je normální matice.  $R_1$  je Schurovou dekompozicí sebe sama. S použitím indukčního předpokladu je  $R_1$  a tedy i matice  $R$  je 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 v  $C^N$ .*

**Důkaz:** Schurovu dekompozici normální matice lze zapsat ve tvaru

$$U^*AU = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_N).$$

Protože  $U$  je unitární, je tento zápis ekvivalentní tvaru

$$AU = U\Lambda. \quad (3.4)$$

Označíme-li  $u_1, \dots, u_N$  sloupce matice  $U$ ,  $U = (u_1, u_2 \dots u_N)$  dostáváme z rovnice (3.4)

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

nebo-li  $u_1, \dots, u_N$  jsou vlastní vektory a  $\lambda_1, \dots, \lambda_N$  jsou vlastní čísla matice  $A$ . □

**Poznámka 3.2** *Zjevně každé matice  $A = U \text{diag}(\lambda_1, \dots, \lambda_N) U^*$ ,  $UU^* = UU^* = I$ , je matice normální.*

Připomeneme dvě důležité třídy normálních matic, a to matice unitární a 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  $\lambda$  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  $\lambda$  je vlastní číslo hermitovské matice  $A$  a  $x$  je příslušný vlastní vektor.

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

tedy  $\lambda$  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

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

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

**Definice 3.5** Je-li  $\Phi$  reálná funkce reálné proměnné, definujeme funkci  $\Phi(A)$  vztahem

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

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

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

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

Viděli jsme, že třída normálních matic je shodná s množinou všech matic, které unikátní podobnostní transformací převést na diagonální tvar. Opustíme-li požadavek unitarity 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  $X^{-1}AX = \text{diag}(\lambda_1, \dots, \lambda_N)$ .

Bohužel, ne každá čtvercová matice je diagonalizovatelná. Je-li matice diagonalizovatelná, pak její vlastní vektory tvoří bazi prostoru  $C^N$ . Tato baze však může být špatně podmíněná, t.j. 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  $D_\epsilon$ , aby vlastní čísla matice  $R_\epsilon = R + D$  byla navzájem různá a  $\|D_\epsilon\| < \epsilon$ . To je zřejmě vždy možné.

Mohli bychom zajásat a uvažovat takto: předcházející věta nám umožňuje omezit se při analýze citlivosti pouze na třídu diagonalizovatelných matic, neboť libovolnou matici vně této třídy mohou libovolně přesně aproximovat maticí diagonalizovatelnou.

Bohužel, jak dále uvidíme, naše jásání by bylo velmi předčasné. 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 převést obecnou matici podobnostní transformací. Větu uvádíme bez důkazu (zvědavého 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

$$X^{-1}AX = \text{diag}(J_{n_1}(\lambda_1), J_{n_2}(\lambda_2), \dots, J_{k_l}(\lambda_l)), \quad (3.6)$$

kde matice na pravé straně je blokově diagonální a  $J_k(\lambda_k) \in C^{k,k}$  je **Jordanův blok** ve formě

$$J_{n_2}(\lambda_k) = \begin{pmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{pmatrix}$$

a  $n_1 + n_2 + \dots + n_l = N$  (explicitně neuvedené prvky matice jsou nulové). Pravá strana výrazu (3.6) (**Jordanův kanonický tvar**) je jednoznačně určena až na uspořádání bloků. Vlastní čísla  $\lambda_k$ ,  $k = 1, \dots, l$ , nemusí být navzájem různé.

Z několika důvodů je práce s Jordanovým kanonickým tvarem obtížná (například libovolně malé perturbace můžou zcela změnit strukturu Jordanových bloků). Všimněme si, že jedničky na subdiagonále 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 dubdiagonálu hodnotou  $\delta$ . Pokud  $\delta \rightarrow 0$ , 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 respektive spektrální normu platí

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

respektive

$$\|U^H A U\| = \|A\|,$$

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

3. Uvědomte si, jaké vztahy platí mezi diagonalizovatelnými, normálními, unitárními a hermitovskými maticemi.
4. Proč sloupce unitární matice řádu  $N$  tvoří ortonormální bazi v  $C^{N,N}$ ?
5. Nechť  $\|x\|$  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 operátorskou normu

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

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

6. Lze každou maticovou normu v  $C^{N,N}$  definovat jako operátorskou 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  $\|AB\| \leq \|A\| \|B\|$  pro libovolné dvě matice  $A, B \in C^{N,N}$ . Jaké konsistentní 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 výrazu (3.5)?
11. Schurovu dekompozici nelze obecně nalézt Řádným konečným algoritmem (např. typu Gaussovy eliminace či QR rozkladu). Proč?

### 3.2 Citlivost vlastních čísel pro obecné matice

Začneme příkladem ukazujícím jaký význam mají v teorii citlivosti vlastních čísel vlastnosti matic.

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

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

*Uvažujme perturbační matici  $E$*

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

*kde  $\epsilon$  je malé kladné číslo. Vlastní čísla matice se příliš neliží od vlastních čísel matice  $A_0$ , neboť pro spektrální poloměr platí*

$$\rho(A_0 + E) \leq \|E\| = \epsilon.$$

*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  $A_0 + E$  neliží o víc než  $10^{-8}$ , 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** *Matice  $A$  bude z třídy  $C^{N,N}$ , její malou změnu (perturbaci) budeme značit  $E \in C^{N,N}$ , perturbovanou matici budeme značit  $\tilde{A}$ ,  $\tilde{A} = A + E$ . Vlastní čísla  $A$  pak budeme označovat  $\lambda_1, \dots, \lambda_N$ , vlastní čísla perturbované matice  $\tilde{\lambda}_1, \dots, \tilde{\lambda}_N$ , příslušné charakteristické polynomy  $\varphi_A(\lambda)$ , respektive  $\varphi_{\tilde{A}}(\lambda)$ .*

### 3.2.1 Spojitost vlastních čísel

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

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

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

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

**Důkaz:** Zvolme  $\varepsilon > 0$  tak, aby  $D(\lambda, \varepsilon)$  neobsahoval žádná další vlastní čísla matice  $A$ . Označme  $\eta(\zeta) = \varphi_{\tilde{A}}(\zeta) - \varphi_A(\zeta)$ . Hranice disku  $D$  je kompaktní množina v  $C$ , označíme ji  $\partial D$ . Charakteristický polynom je spojitou funkcí prvků matice; proto funkce  $\eta(\zeta)$  konverguje k nule na kompaktu  $\partial D$  pro  $\tilde{A} \rightarrow A$ . Protože platí  $\varphi_A(\zeta) \neq 0$  pro  $\forall \zeta \in \partial D$ , jistě existuje takové číslo  $\delta > 0$ , že platí

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

Nyní použijeme Rouchého větu. Funkce  $\varphi_A$  a  $\eta$  jsou analytické v celé množině  $C$ . Pak z (3.8) vyplývá, že  $\varphi_A$  a  $\varphi_{\tilde{A}} = \varphi_A + \eta$  mají v kruhu  $D$  stejný počet nulových bodů.

**Poznámka 3.4** *Uvědomme si, že Věta 3.7 nám nedává žádný kvantitativní odhad pro změnu vlastních čísel při dané matici  $A$  a velikosti perturbace  $\|E\|$ . 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 vzdálenost 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*

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

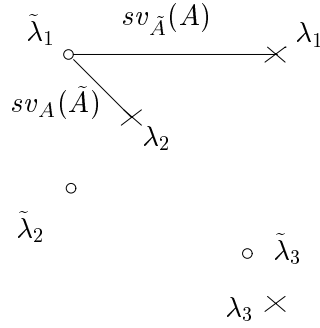
**Poznámka 3.5** *Všimněme si, jaký má spektrální variace geometrický význam. Definujeme-li totiž*

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

*pro  $i = 1, \dots, N$ , 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  $3 \times 3$ , kdy je  $sv_{\tilde{A}}(A) > sv_A(\tilde{A})$ . Vlastní čísla matice  $A$  jsou označené křížky, vlastní čísla matice  $\tilde{A}$  kroužky. Snadné řešení spočívá v symetrizaci.*



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

**Definice 3.8** Necht  $A \in C^{N,N}$ ,  $E \in C^{N,N}$ ,  $\tilde{A} = A + E$ . **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 jiŘ metriku v  $C^{N,N}$ , stále vŮak nám nedává názorový pojem vzdálenosti. Proto se používá následující definice.

**Definice 3.9** Necht  $A \in C^{N,N}$ ,  $E \in C^{N,N}$ ,  $\tilde{A} = A + E$ . **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  $\pi$  probíhá všechny permutace množiny  $\{1, \dots, N\}$ .

Podaří-li se nám ukázat, Ře párová vzdálenost matic  $md(A, \tilde{A})$  je malá, znamená to, Ře je možné z vlastních čísel matice  $A$  a  $\tilde{A}$  vytvořit páry tak, Ře vzdálenost čísel v páru je malá. To nám dovoluje názorně si představit změny individuálních vlastních čísel. Mezi pojmy definovanými výše platí následující vztahy:

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

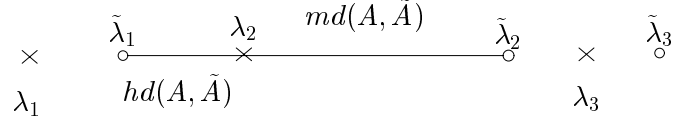
**Příklad 3.3** Příklad, kdy  $hd(A, \tilde{A}) < md(A, \tilde{A})$  je znázorněn na obr. (3.2).

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  $a_1, \dots, a_N$  sloupce matice  $A \in C^{N,N}$ . Pak platí

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

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



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

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

$$A = UR ; \quad U^*A = R, \quad (3.9)$$

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

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

Pro determinant matice  $R$  dalšími úpravami dostáváme

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

(bylo použito vyjádření (3.9) pro matici  $R$  a vlastnost 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

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

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

**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*

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

**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 matice  $\tilde{A}$  a doplňme

ho dalšími vektory  $x_2, \dots, x_N$  tak, aby výsledná matice  $X = (x_1, \dots, x_N)$  byla unitární. Pro  $N$ -tou mocninou spektrální variace  $\tilde{A}$  vzhledem k  $A$  pak platí

$$\begin{aligned} (sv_A(\tilde{A}))^N &\leq \prod_{i=1}^N |\tilde{\lambda} - \lambda_i| = |\det(A - \tilde{\lambda}I)| = |\det[(A - \tilde{\lambda}I)X]| \\ &\leq \prod_{i=1}^N \|(A - \tilde{\lambda}I)x_i\| = \|(A - \tilde{\lambda}I)x_1\| \prod_{i=2}^N \|(A - \tilde{\lambda}I)x_i\| \end{aligned} \quad (3.12)$$

Poslední nerovnost ve výrazu (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. Pokud je  $md(A, \tilde{A})$  malé číslo, znamená to, že vlastní čísla matic  $A$ ,  $\tilde{A}$  jsou uspořádána v párech tvořených blízkými vlastními čísly. I když znění věty bude až na násobek 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

$$\mu \stackrel{\text{def}}{=} \left( 2 \max_{\tau \in \langle 0,1 \rangle} \|A + \tau E\| \right)^{1-\frac{1}{N}}.$$

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

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

z čehož plyne

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

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 disky  $D_{m+1}, \dots, D_N$  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  $\tau \in \langle 0, 1 \rangle$ ,

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

Použitím Elsnerovy věty a při zavedeném označení dostáváme

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

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  $\gamma \tau^{1/N}$  je pro  $\tau \in \langle 0, 1 \rangle$  monotonně rostoucí. Tedy

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

je izolována od ostatních disků pro každé  $\tau \in \langle 0, 1 \rangle$ , což je velmi podstatný závěr. Sjednocení  $\bigcup_{i=1}^m D_i^0$  obsahuje právě  $m$  vlastních čísel matice  $\tilde{A}_0 = A$ . Zkonstruuje posloupnost matic  $\tilde{A}_0, \tilde{A}_{\tau_1}, \dots, \tilde{A}_{\tau_m}, \dots$ ,  $0 < \tau_1 < \dots < 1$  tak, aby

$$\lim_{i \rightarrow \infty} \tilde{A}_{\tau_i} = \tilde{A}_1.$$

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

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

A tato limita musí vzhledem k izolovanosti (3.13) ležet v  $\bigcup_{i=1}^m D_i^0$ . □

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*

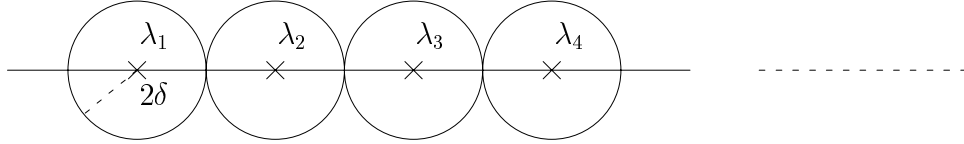
$$md(A, \tilde{A}) \leq (2N - 1)(\| A \| + \| \tilde{A} \|)^{1 - \frac{1}{N}} \| E \|^{1/N}. \tag{3.14}$$

**Důkaz:** Označme  $C_1, C_2, \dots, C_k$  souvislé navzájem disjunktní komponenty sjenocení  $\bigcup_{i=1}^N D_i$ . Podle lemy 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  $\pi$  na množině  $\{1, \dots, N\}$ , které každému vlastnímu číslu  $\lambda_j \in C_l$  přiřadí vlastní číslo  $\tilde{\lambda}_{\pi(j)} \in C_l$ . Spáry jsou vytvářeny jen uvnitř jednotlivých komponent, nikoliv mezi čísly v různých komponentech. Bez újmy obecnosti se naře další úvahy budou týkat pouze největší souvislé komponenty označené jako  $C_1$ , o níž budeme předpokládat, že je sjednocením disků se středy ve vlastních číslech  $\lambda_1, \dots, \lambda_m$ ,

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

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



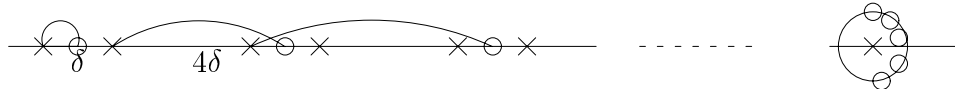
Obrázek 3.3: Nejhorší možné rozložení vlastních čísel matice  $A$  takové, aby  $C_1$  měla maximální rozměr

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

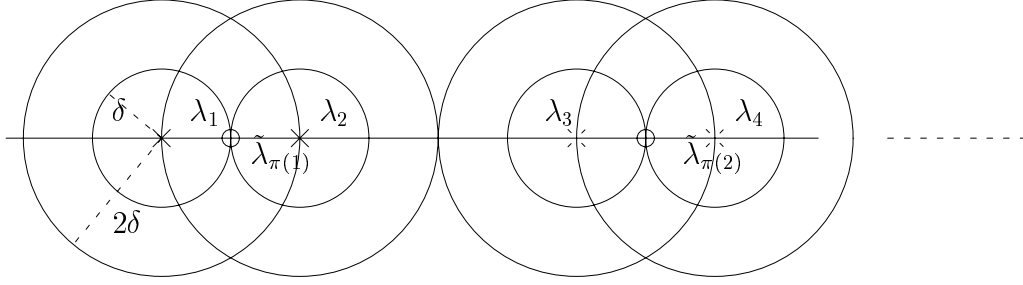
$$\begin{aligned} sv_A(\tilde{A}) &\leq \delta \\ sv_{\tilde{A}}(A) &\leq \delta. \end{aligned} \quad (3.15)$$

To znamená, že sjednocení disků se středy v  $\lambda_i$  a poloměrem  $\delta$  musí obsahovat všechna vlastní čísla  $\tilde{\lambda}_{\pi(i)}$  a zároveň sjednocení disků se středy v  $\tilde{\lambda}_{\pi(i)}$  a poloměrem  $\delta$  musí obsahovat všechna vlastní čísla  $\lambda_i$ . 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 pro takto rozložená vlastní čísla je znázorněno na následujícím schématu.



Při tomto spárování musí jedno z vlastních čísel  $\tilde{\lambda}_{\pi(k)}$  znázorněných na pravém okraji posledního disku tvořit pár s  $\lambda_k$ , kde  $k = \lfloor \frac{m+1}{2} \rfloor$  je-li  $m$  liché,  $k = \frac{m}{2} + 1$  je-li  $m$  sudé číslo



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

(symbolem  $[.]$  zde značíme celou část příslušného racionálního čísla). V každém případě platí pro vzdálenost  $\lambda_k$  od  $\tilde{\lambda}_{\pi(k)}$  odhad

$$|\lambda_k - \lambda_{\pi(k)}| \leq \left\lfloor \frac{m+1}{2} \right\rfloor 2\gamma + \delta = (2m-1)\delta.$$

Snadno nahlédneme, že uvedený odhad je horním odhadem pro maximální vzdálenost čísel v páru při libovolné permutaci,

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

□

Základem důkazu věty 3.14 byl neklesající odhad pro spektrální variaci  $sv_A(A + \tau E)$ . Větu je proto možno formulovat obecněji (důkaz je sleduje předchozí postup o 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:*

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

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) výsledek. Je-li například  $N = 10^2$  (v praxi se často řeší problémy pro  $N \sim 10^3 - 10^5$ ), a  $\|E\| = 10^{-10}$ , je výsledný odhad změny vlastních čísel při takto malé perturbaci úměrný  $10^{-1/10} \|A\|$  a jeho praktická hodnota je nepatrná.

### 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  $2\delta$ ?
4. K čemu je potřeba neklesající odhad pro
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 vřdy, chceme proto namézt charakteristiku matice, která bude rozhodujícím způsobem ovlivňovat kvalitu odhadu. Touto charakteristikou bude odchylka od normality, studované v tomto odstavci.

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

**Věta 3.11** (*Bauer-Fike*) *Nech  $Q \in C^{N,N}$  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*

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

**Důkaz:** Za předpokladů věty platí

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

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

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

singulární. Musí tedy platit

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

S využitím konzistence maticové normy (viz cvičení 7 v části 3.1) dostaneme tvrzení věty.  $\square$

**Poznámka 3.6** *Dohodnem-li se na označení*

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

*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 konsistentní normu platí*

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

*Důkaz ponecháme jako cvičení.*

Připomeňme, Ře Schurova dekompozice normální matice je diagonální matice. Schurova dekompozice obecné matice je horní trojúhelníková matice  $R$ , přičemž  $R$  není určena jednoznačně. Odchylku od normality definujeme následujícím způsobem:

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

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

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

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

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

**Důkaz:** Proto Ře Frobeniova norma

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

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.

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

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

**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  $U^H A U = \Lambda + R$  je Schurova dekompozice matice  $A$  pro nějakou  $U \in \mathcal{U}$ , matice  $\Lambda$  je diagonální a  $R$  horní trojúhelníková s nulovou diagonálou. Z Bauerovy-Fikeho věty pro ( $Q \stackrel{def}{=} U$ ) pak máme

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

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

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

Označme

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

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

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

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

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

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

Henriciho věta dává vlastně 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 proměnné  $\Psi(\eta)$  definované vztahem

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

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 krajní hodnoty  $\eta$ . Je-li  $\eta$  malé, je jmenovatel ve výrazu (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  $\eta$  velké, pak je  $\eta^{N-1}$  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.

#### Důsledek

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

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

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

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

**Důkaz:** Pro funkci  $\Psi$  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

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

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  $\Psi$

$$\Psi(\eta) = \frac{\eta}{1 + \eta^{-1} + \dots + \eta^{-(N-1)}} \geq \eta(1 - \eta^{-1}) = \eta - 1 \quad (3.27)$$

Výrazu (3.27) budeme chtít použít pro proměnou

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

(všimněme si, že inverzní funkce k funkci  $\Psi$  existuje a je monotónní). Nejprve je třeba ověřit, že takto definované  $\zeta$  je většinou neř jedna. Použijeme další vlastnosti funkce  $\Psi$ , 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  $\zeta$  definované výrazem (3.28) většinou neř jedna, tedy z (3.27) máme

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

Z Henriciho věty pak platí

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

Aplikujeme-li na obě strany výrazu (3.30) funkci  $\Psi^{-1}$ , 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).  $\square$

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

**Důsledek** Nechť matice  $A \in C^{N,N}$ ,  $\tilde{A} = A + E$ , kde  $E \in C^{N,N}$  a funkce  $\Psi$  je definována výrazem (3.24). Pak pro  $\eta = \frac{|\tilde{\lambda} - \lambda|}{\delta_\alpha(A)}$  platí

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

**Důkaz:** Podle věty (refhenr) 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 používat větu 3.10.  $\square$

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 má dimenzi  $m$ , kde  $m < N$ .

**Věta 3.14** 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

$$\frac{|\tilde{\lambda} - \lambda|^m}{1 + |\tilde{\lambda} - \lambda| + \dots + |\tilde{\lambda} - \lambda|^{m-1}} \leq \|Q^{-1}EQ\|. \quad (3.31)$$

**Důkaz:** Tvzení 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

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

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

Cvičení

1. Proč je pro horní trojúhelníkovou matici  $R \in C^{N,N}$  s nulovou diagonálou  $R^j = 0$  pro  $j \geq N$ ?
2. Dokařte, že je-li  $\Psi(\eta) < 1$  pak je  $\eta < 1$ .
3. Dokařte, že funkce  $\Psi$  definovaná vztahem (3.24) je monotonní.
4. Dokařte Větu 3.11 pro libovolnou konsistentní maticovou normu  $\| \cdot \|_\alpha$ .
5. Nechť  $\| \cdot \|_m$  je libovolná norma v  $C^{N,N}$ . Existuje taková vektorová norma  $\| \cdot \|_v$  v  $C^N$  tak, že  $\| Ax \| \leq \| A \| \| x \|_v$  platí  $\forall A \in C^{N,N}, \forall x \in C^N$ ? Platí stejné tvrzení pro pro ?? vektorovou normu?

### 3.3 Citlivost jednoduchého vlastního čísla

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 perturbací (řádná perturbovaná matice zde nevystupuje). Přesto lze při vhodném uřítí Geršgorinovy věty získat velmi dobré odhady polohy vlastních čísel perturbované matice.

**Věta 3.15 (Geršgorin)** *Nechť  $A \in C^{N,N}$ ,  $A = (a_{ij})_{i,j=1,\dots,N}$  a označme  $\alpha_i = \sum_{j=1, j \neq i}^N |a_{ij}|$ ,  $i = 1, \dots, N$  a*

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

*Pak platí*

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

*Pokud je  $m$  disků  $G_i(A)$  izolováno od ostatních  $N - m$  disků, pak jejich sjednocení obsahuje právě  $m$  vlastních čísel matice  $A$ .*

**Důkaz:** Použijeme vztah (3.19) z důkazu Bauer-Fikeho věty pro speciální volbu matic  $A$ ,  $\tilde{A}$ ,  $Q$  a maximovou normu. V nerovnosti

$$1 \leq \| Q^{-1}(A - \tilde{\lambda}I)^{-1}Q(Q^{-1}EQ) \| \quad (3.33)$$

tak položíme

$$\begin{aligned} Q &\stackrel{def}{=} I \\ A &\stackrel{def}{=} \text{diag}(a_{11}, \dots, a_{NN}) \\ \tilde{A} &\stackrel{def}{=} A \\ \tilde{\lambda} &\stackrel{def}{=} \lambda, \end{aligned}$$

kde  $\lambda$  je libovolné vlastní číslo matice  $A$ , pro které platí  $\lambda \neq a_{ii}$  pro  $i = 1, \dots, N$  (jinak znění věty vyplývá triviálně). Pro takto zvolené proměnné a maximovou normu ( $\| B \|_{\infty} = \max_i \sum_{j=1}^N |b_{ij}|$  pro  $B \in C^{N,N}$ ) má výraz (3.33) tvar

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

což je z definice maximové normy

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

Nechť  $i_{\lambda}$  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 je zcela analogický důkazu lemmatu 3.2  $\square$

Další příklad je ukázkou toho, že odhad pro polohu vlastních čísel perturbované matice získaný aplikací Geršgorinovy věty může být přesnější než odhad z věty Elsnerovy.

**Příklad 3.4** *Mějme matici*

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

a její perturbaci

$$E = \begin{pmatrix} 0 & 10^{-4} \\ 10^{-4} & 0 \end{pmatrix}.$$

Perturovaná 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 pro vzájemnou polohu vlastních čísel matice  $A$  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 máme

$$\sigma(\tilde{A}) \subset ((1 - 10^{-2}, 1 + 10^{-2}) \cup (2 - 10^{-2}, 2 + 10^{-2})).$$

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, vzhledem k tomu, že skutečná vlastní čísla matice  $\tilde{A}$  mají hodnotu  $\approx 1 - 10^{-8}, \approx 2 + 10^{-8}$ .

V dalším příkladě ukážeme, jak lze odhad pro polohu 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. Umíme vybrat takovou podobnostní transformaci, že ve výsledné matici jsou součty absolutních hodnot mimodiagonálních prvků ve zvoleném řádku menší než tyto součty v odpovídajícím řádku matice  $\tilde{A}$ . Pak dává Geršgorinova věta aplikovaná na tuto matici lepší odhad pro polohu příslušného vlastního čísla matice  $\tilde{A}$ .

Provedme 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  $\alpha$  je nějaký kladný reálný parametr. Ukážeme, jak lze vhodnou volbou parametru  $\alpha$  získat dostatečně jemný odhad pro vlastní číslo matice  $\tilde{A}$  ležící v disku 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 neprotínají, obsahuje každý z nich právě jedno vlastní číslo matice  $\tilde{A}(\alpha)$ . Zvolíme-li tedy  $\alpha$  dostatečně malé, ale zároveň tak velké, aby se oba intervaly neprotýkaly, dostaneme velmi dobrý odhad pro vlastní číslo v intervalu  $(1 - \alpha 10^{-4}, 1 + \alpha 10^{-4})$ .

NeŘ zformulujeme vĕtu o citlivosti jednoduchĕho vlastnĕho ěisla vzhledem k perturbacĕm matice, pŕipomeneme definici levĕho vlastnĕho vektoru a provedeme nĕkterĕ pomocnĕ ũvahy, kterĕ pŕi dŕkazu vĕty 3.16 budeme potŕebovat.

**Definice 3.11** Řekneme, Ře vektor  $y \in C^N$ ,  $y \neq 0$ , pro kterŕ platĕ  $y^H A = \lambda y^H$ , kde  $A \in C^{N,N}$  a  $\lambda$  je nĕjakĕ vlastnĕ ěislo matice  $A$ , je **levŕ vlastnĕ vektor**. (Analogicky hovoŕĕme o vektoru  $x \neq 0$ , pro nĕjŘ je splnĕno  $Ax = \lambda x$  jako o **pravĕm vlastnĕm vektoru**.)

**Poznĕmka 3.8** Levŕ vlastnĕ vektor lze definovat uŘitĕm vlastnostĕ determinantu. Platĕ totiŘ

$$0 = \det(A - \lambda I) = \overline{\det(A^H - \bar{\lambda} I)},$$

levŕ vlastnĕ vektor  $y \neq 0$  je pak definovĕn jako vektor, kterŕ splŕuje rovnost  $A^H y = \bar{\lambda} y$ .

Pŕedpoklĕdejme, Ře  $A \in C^{N,N}$  a nechŕ

$$J = W^{-1} A W \tag{3.34}$$

je jejĕ Jordanŕv kanonickŕ tvar. Nechŕ Jordanovy bloky jsou uspoŕadĕny tak, Ře na prvnĕm mĕstĕ je Jordanŕv blok obsahujĕcĕ jednoduchĕ vlastnĕ ěislo  $\lambda$ .

$$J = \begin{pmatrix} \lambda & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_k \end{pmatrix}, \tag{3.35}$$

kde  $J_2, \dots, J_k$  jsou Jordanovy bloky. Z (3.34) dostĕvĕme

$$A W = W J \tag{3.36}$$

a oznaĕme-li sloupce matice  $W$  jako  $w_1, \dots, w_N$ , pro prvnĕ sloupec matice z rovnosti (3.36) dostaneme

$$A w_1 = \lambda w_1.$$

Pro  $J$  hermitovskŕ sdruŘenou pak platĕ

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

tedy opĕt platĕ

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

Zŕejmĕ je  $(W^{-1})^H = (W^H)^{-1}$  a oznaĕme-li sloupce matice  $(W^H)^{-1}$  jako  $y_1, \dots, y_N$ , opĕt musĕ platit

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

Navĕc je pro vektory  $w_1, y_1$  splnĕno

$$|y_1^H w_1| = 1.$$

**Věta 3.16** *Nechť  $\lambda$  je jednoduché vlastní číslo matice  $A \in C^{N,N}$  s levým vlastním vektorem  $y$  a pravým  $x$ . Pro  $E \in C^{N,N}$  uvažujme  $\tilde{A} = A + E$ . Pak pro každou dostatečně malou perturbaci  $E$  existuje jediné vlastní číslo matice  $\tilde{A}$ , které lze vyjádřit ve tvaru*

$$\tilde{\lambda} = \lambda + \frac{y^H E x}{y^H x} + O(\|E\|^2). \quad (3.37)$$

**Poznámka 3.9** *Symbolu  $O(h^2)$  používáme následujícím způsobem:*

$$x = y + O(h^2) \Leftrightarrow |x - y| \leq Kh^2,$$

kde  $K$  je konstanta nezávislá na  $h$ .

**Důkaz:** Nechť  $\delta > 0$  je vzdálenost jednoduchého vlastního čísla  $\lambda$  od ostatních vlastních čísel matice  $A$ .

Nechť Jordanův kanonický tvar

$$J = W^{-1}AW$$

je definován výrazem (3.35).

Položme

$$(Y')^H \stackrel{\text{def}}{=} W^{-1}$$

$$X' \stackrel{\text{def}}{=} W$$

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  $\lambda$  a platí  $|y'^H x'| = 1$ .

Nyní využijeme toho, že  $J$  je matice speciální struktury, totiž že má na vedlejší diagonále jedničky nebo nuly. Protože pro jednotlivé Jordanovy bloky  $J_i$  platí

$$\begin{pmatrix} \delta & & & \\ & \delta^2 & & \\ & & \ddots & \\ & & & \delta^{l_i} \end{pmatrix}^{-1} J_i \begin{pmatrix} \delta & & & \\ & \delta^2 & & \\ & & \ddots & \\ & & & \delta^{l_i} \end{pmatrix} = \begin{pmatrix} \lambda_i & \delta & & \\ & \lambda_i & \delta & \\ & & \ddots & \ddots \\ & & & \lambda_i \end{pmatrix},$$

máme pro  $J$

$$\begin{pmatrix} \frac{\delta}{3} & & & \\ & (\frac{\delta}{3})^2 & & \\ & & \ddots & \\ & & & (\frac{\delta}{3})^N \end{pmatrix}^{-1} J \begin{pmatrix} \frac{\delta}{3} & & & \\ & (\frac{\delta}{3})^2 & & \\ & & \ddots & \\ & & & (\frac{\delta}{3})^N \end{pmatrix} = J'.$$

Matice  $J'$  má na vedlejší diagonále na místech, kde  $J$  měla jedničky prvky  $\frac{\delta}{3}$ . Všechny ostatní prvky matic  $J$  a  $J'$  jsou totožné.

Označíme-li nyní

$$Y^H = \begin{pmatrix} \frac{\delta}{3} & & & \\ & (\frac{\delta}{3})^2 & & \\ & & \ddots & \\ & & & (\frac{\delta}{3})^N \end{pmatrix}^{-1} (Y')^H,$$

$$X = X' \begin{pmatrix} \frac{\delta}{3} & & & \\ & (\frac{\delta}{3})^2 & & \\ & & \ddots & \\ & & & (\frac{\delta}{3})^N \end{pmatrix},$$

$$y = (\frac{\delta}{3})^{-1} y'$$

$$x = \frac{\delta}{3} x',$$

pak zřejmě platí

$$|y^H x| = 1.$$

Vektory  $y$  respektive  $x$  jsou pravý respektive levý vlastní vektor příslušný jednoduchému vlastnímu číslu  $\lambda$ .

Napišme nyní, jak vypadají prvky matice

$$\tilde{J} = Y^H (A + E) X = Y^H A X + Y^H E X,$$

$$\tilde{J} = \begin{pmatrix} \lambda + y^H E x & \epsilon & & \dots & & \epsilon \\ \epsilon & \mu & \tau & \epsilon & \dots & \epsilon \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \epsilon \\ \vdots & \ddots & \ddots & \ddots & \ddots & \tau \\ \epsilon & & \dots & & \epsilon & \mu \end{pmatrix},$$

kde

$\epsilon$  je označení pro prvky, jejichž absolutní hodnota je menší než  $\|Y\| \|E\| \|X\|$

$\mu$  pro diagonální prvky jiné než  $\lambda + y^H E x$

$$\tau = \begin{cases} 0 \\ \frac{\delta}{3} + \epsilon \end{cases}.$$

Polohu vlastního čísla, které leží v kruhu se středem v  $\lambda + y^H E x$  budeme odhadovat užitím Geršgorinovy věty. Přesnost tohoto odhadu zřejmě závisí na hodnotě prvků v prvním řádku matice  $\tilde{J}$ . Proto nejprve podobnostní transformací převedeme matici  $\tilde{J}$  do tvaru, který užitím Geršgorinovy věty umožní získat relativně jemný odhad pro polohu vlastního čísla  $\tilde{\lambda}$ .

Pro nějaký kladný reálný parametr  $\alpha$  máme

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

a  $\tilde{J}(\alpha)$  se od  $\tilde{J}$  liší pouze tím, že v prvním řádku má ve sloupcích  $2, \dots, N$  prvky  $\alpha\epsilon$  a v prvním sloupci v řádcích  $2, \dots, N$  prvky  $\alpha^{-1}\epsilon$ . A zdůrazněme, že vlastní čísla matice  $\tilde{J}(\alpha)$  jsou totožná s vlastními čísly matice  $\tilde{J}$ , nebo-li

$$\sigma(\tilde{J}) = \sigma(\tilde{J}(\alpha)).$$

Podle Geršgorinovy věty je

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

kde

$G_1$  má střed v  $\lambda + y^H E x$  a poloměr  $(N-1)\alpha|\epsilon|$

$G_2, \dots, G_N$  mají středy v  $\mu$  a poloměr menší než  $\alpha^{-1}|\epsilon| + \frac{\delta}{3} + |\epsilon| + (N-3)|\epsilon|$ .

Abychom pro odhad vlastního čísla matice  $\tilde{J}$ , které leží v disku  $G_1$  mohli použít Geršgorinovu větu, je třeba zaručit, že se disk  $G_1$  neprotne s žádným jiným. (Pak je zaručeno, že v  $G_1$  leží právě jedno vlastní číslo matice  $\tilde{J}(\alpha)$ .)

Vzdálenost středu disku  $G_1$  od středu libovolného jiného disku je v nejhorším případě  $\delta - 2|\epsilon|$ . Tudíž, aby se disk  $G_1$  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

$$(N-1)\alpha|\epsilon| + \alpha^{-1}|\epsilon| + N|\epsilon| < \frac{2}{3}\delta. \quad (3.38)$$

Budeme hledat takové podmínky pro parametry  $\epsilon$  a  $\alpha$ , aby bylo splněno (3.38).

Nechť perturbace  $E$  (na níž závisí hodnota  $\epsilon$ ) matice  $A$  je tak malá, že platí

$$\frac{2}{3}\delta - N|\epsilon| > \frac{\delta}{2} \quad (3.39)$$

a nechť  $\alpha$  je takové, že je navíc splněno

$$\alpha^{-1}|\epsilon| + N\alpha|\epsilon| < \frac{\delta}{2}. \quad (3.40)$$

Jednoduchou úpravou nerovnosti (3.40) a s použitím odhadu pro  $\frac{\delta}{2}$  z výrazu (3.39) dostaneme

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

a ta platí pro

$$\alpha = \frac{4|\epsilon|}{\delta} \quad (3.41)$$

a perturbace  $E$  tak malé, že

$$\frac{16N|\epsilon|}{\delta^2} < 1. \quad (3.42)$$

Dokázali jsme, že pokud je splněno (3.41) a (3.42), platí i (3.38) a tedy disk  $G_1$  se středem v  $\lambda + y^H E x$  je disjunktní s ostatními disky. Poloměr disku  $G_1$  je

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

A tento disk obsahuje právě jedno vlastní číslo matice  $\tilde{J}$  označme ho  $\tilde{\lambda}$ , a platí pro ně

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

□

**Poznámka 3.10** Volbou parametrů v důkaze věty 3.16 kladu podmínky na  $\epsilon$  takové, aby

$$\frac{|\epsilon|}{\delta} \ll 1$$

Pokud je  $\delta$  malé číslo (tj. jednoduché vlastní číslo  $\lambda$  je úpatně separované od ostatních), bude množina perturbací, pro něž je odhad (3.37) platný značně omezená (platí jen pro velmi malé perturbace).

Velikost čitatele  $|\epsilon|$  nezávisí jen na  $\|E\|$ , ale také na velikosti  $\|x\| \|y\|$ . Pokud je  $\|x\| \|y\| \gg 1$ , neumíme zaručit, že  $\frac{|\epsilon|}{\delta}$  bude malé číslo.

**Poznámka 3.11** Výraz (3.37) lze napsat i v jiném tvaru, a to

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

Výraz

$$\frac{y^H (A + E)x}{y^H x}$$

se nazývá **Rayleighův kvocient**.

**Poznámka 3.12** Jiná formulace vztahu (3.37) je například

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

Označíme-li jako

$$\nu = \frac{\|y\| \|x\|}{|y^H x|},$$

pak o  $\nu$  budeme hovořit jako o **čísle podmíněnosti jednoduchého vlastního čísla  $\lambda$** .

Uvědomme si, že  $\nu$  je secans úhlu, který svírají vektory  $x$  a  $y$ . Platí totiž

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

a  $\sec \varphi = \frac{1}{\cos \varphi}$ .

Tedy  $\nu = 1$  pokud  $x$  a  $y$  svírají nulový úhel a roste se zvětšováním úhlu mezi  $x$  a  $y$ . Při tom  $x$  a  $y$  na sebe nemohou být kolmé, protože  $\lambda$  je jednoduché vlastní číslo (viz ??, str. 42).

### 3.4 Citlivost vlastních čísel pro diagonalizovatelné a normální matice

Jak jsme naznačili v úvodu ke kapitole 3, je citlivost vlastních čísel vzhledem k perturbacím matice závislá na speciálních vlastnostech původní matice. V tomto paragrafu 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  $E$ , jak tomu bylo v případě obecné matice.

**Věta 3.17** *Nechť  $A \in C^{N,N}$  je normální matice. Je-li  $\|x\| = 1$ , pak platí*

$$\min_{1 \leq i \leq N} |\lambda_i - x^H A x| \leq \|Ax - (x^H A x)x\|.$$

**Důkaz:** Protože  $A$  je normální matice, existuje unitární matice  $U$  tak, že  $U^H A U = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_N)$  a dostáváme

$$\begin{aligned} \|(A - x^H A x I)x\| &= \|(U \Lambda U^H - x^H A x I)x\| = \|U(\Lambda - x^H A x U^H U)U^H x\| \\ &= \|(\Lambda - x^H A x I)U^H x\| = \sqrt{\sum_{i=1}^N |\lambda_i - x^H A x|^2 |U^H x|^2} \\ &\geq \min_{1 \leq i \leq N} |\lambda_i - x^H A x| \sqrt{\sum_{i=1}^N |U^H x|^2}. \end{aligned}$$

Nyní využijeme definice euklidovské normy a toho, že tato norma je invariantní vzhledem k násobení unitární maticí.  $\square$

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

**Věta 3.18** *Nechť  $A \in C^{N,N}$  je diagonalizovatelná matice, tj. existuje nonsingulární matice  $X \in C^{N,N}$  tak, že  $X^{-1} A X = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_N)$ ,  $\tilde{A} = A + E$ , kde  $E \in C^{N,N}$ . Pak platí*

$$sv_A(\tilde{A}) \leq \|X^{-1} E X\| \leq \|X^{-1}\| \|E\| \|X\| = \kappa(X) \|E\| \quad (3.43)$$

$$md(A, \tilde{A}) \leq (2N - 1) \|X^{-1} E X\| \leq (2N - 1) \kappa(X) \|E\|. \quad (3.44)$$

**Důkaz:** Mějme nějaké vlastní číslo  $\tilde{\lambda}$  matice  $\tilde{A}$ , které není vlastním číslem matice  $A$ . (V opačném případě platí znění věty triviálně.) Z Bauer-Fikeho věty pro  $Q \stackrel{\text{def}}{=} X$  pak dostáváme

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

Využijeme-li toho, že  $X^{-1} A X = \Lambda$ , máme z definice spektrální normy pro diagonální matici

$$\|(\Lambda - \tilde{\lambda})^{-1}\|^{-1} = \frac{1}{\max_j \frac{1}{|\lambda_j - \tilde{\lambda}|}} = \min_j |\lambda_j - \tilde{\lambda}| \leq \|X^{-1} E X\|$$

a protoře  $\tilde{\lambda}$  bylo libovolné vlastní číslo matice  $\tilde{A}$ , platí odhad pro kařdÉ vlastní číslo matice  $\tilde{A}$ , tedy platí

$$\max_i \min_j |\tilde{\lambda}_i - \lambda_j| \leq \| X^{-1} E X \| .$$

DalšÍ nerovnost v (3.43) plyne z vlastnosti konzistence maticové normy.

Odhad (3.44) pro optimální vzdálenost vyplývá z věty 3.10.  $\square$

Přímým důsledkem věty 3.18 je následující odhad optimální vzdálenosti pro normální matici.

**Věta 3.19** *Nech  $A \in C^{N,N}$  je normální matice. Polořme  $\tilde{A} = A + E$ , kde  $E \in C^{N,N}$ . Pak pro optimální vzdálenost je*

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

Označme jako  $V$  matici, jejíř sloupce jsou tvořeny vlastními vektory matice  $A$ . Je-li  $A$  diagonalizovatelná, platí  $V = X$  a číslo podmíněnosti  $\kappa(V) = \| V \| \| V^{-1} \|$  leřÍ v otevřeném intervalu  $(1, \infty)$ . Velikost čísla podmíněnosti  $\kappa(V)$  diagonalizovatelné matice tedy vypovídá o tom, jak je mezi jejími vlastními vektory porušena ortogonalita směrem k lineární závislosti. Na  $\kappa(V)$  můřeme tudíř pohlířet jako na měřítko odchylky od normality. Poznamenejme, ře pro normální matici je  $\kappa(V) = 1$ , pro matici, která není diagonalizovatelná je naopak  $\kappa(V) = \infty$ .

### Cvičení

1. Proč je pro matici  $A$ , která není diagonalizovatelná  $\kappa(V) = \infty$ ?

## 3.5 Příklady

Jak je patrné z tvrzení, která jsme uvedli v předchozích paragrafech, má v teorii citlivosti vlastních čísel matic rozhodující 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 nichř jsou tyto odhady závislé na vlastních vektorech dané matice. NejhoršÍ je situace pro obecnou matici, kdy ani při malé perturbaci prvků nejsme schopni o poloze vlastních čísel říci nic rozumného.

Uvádíme příklady matic, jejichř vlastní čísla jsou (ař na jednu vyjímku) citlivá vzhledem k perturbacím prvků matice. Pro geometrické zobrazení citlivosti vlastních čísel je velice uřitečný pojem pseudospektra matice.

**Definice 3.12** *Nech  $A \in C^{N,N}$  a polořme  $\tilde{A} = A + E$ , kde  $E \in C^{N,N}$ . Pro  $\epsilon \geq 0$  definujeme  $\epsilon$ -pseudospektrum matice  $A$  jako*

$$\sigma_\epsilon(A) = \{ \tilde{\lambda} \in C; \tilde{\lambda} \text{ je vlastní číslo matice } \tilde{A} = A + E \text{ pro nějaké } E, \| E \| \leq \epsilon \}. \quad (3.45)$$

**Poznámka 3.13** *Ekvivalentní definice pseudospektra můře vypadat například takto*

$$\sigma_\epsilon(A) = \{ \tilde{\lambda} \in C; \| (\tilde{\lambda} I - A)^{-1} \| \geq \epsilon^{-1} \}. \quad (3.46)$$

*Pokud je  $\tilde{\lambda}$  vlastním číslem matice  $A$ , definujeme  $\| (\tilde{\lambda} I - A)^{-1} \| \stackrel{def}{=} \infty$ .*

Ekvivalence výrazů (3.45) a (3.46) vyplývá přímo z Bauer-Fikeho věty.

**Poznámka 3.14** *Všimněme si, jak vypadá  $\epsilon$ -pseudospektrum pro normální matici. Je-li totiž  $A$  normální, platí*

$$\|(\tilde{\lambda}I - A)^{-1}\| = \frac{1}{\text{dist}(\tilde{\lambda}, \sigma(A))}, \quad (3.47)$$

kde  $\text{dist}(z, S)$  označuje vzdálenost bodu  $z$  od množiny  $S$ . Tedy pro normální matici je plocha  $\|(\tilde{\lambda}I - A)^{-1}\|$  určena vlastními čísly matice  $A$ . Pseudospektrum  $\sigma_\epsilon(A)$  je pak rovno sjednocení disků o poloměru  $\epsilon$  se středy ve vlastních číslech matice  $A$ .

Pro obecnou matici je situace mnohem komplikovanější. Neplatí žádný vztah podobný (3.47). A jak uvidíme v následujících příkladech, může číslo  $\|(\tilde{\lambda}I - A)^{-1}\|$  dosahovat velkých hodnot i pro  $\tilde{\lambda}$  vzdálená od spektra matice  $A$ .

Všechny matice, jejich pseudospektra budeme studovat, jsou řádu  $N = 32$ .

Vlastní vektory matic jsou normovány a je vypočteno číslo podmíněnosti  $\kappa(V)$  ( $V$  je matice, její sloupce jsou tvořeny vektory matice  $A$ ).

Pro každou matici uvedeme dva obrázky. Na prvním 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  $\tilde{E}$ , jejími prvky jsou náhodné veličiny při komplexním normálním rozdělení se střední hodnotou 0 a standardním rozptylem 1. Pak je vypočtena norma  $\|\tilde{E}\|$  a  $E$  je definována jako  $E \stackrel{\text{def}}{=} 10^{-3} \frac{\tilde{E}}{\|\tilde{E}\|}$ .

Druhý obrázek zachycuje křivky, které tvoří hranice pro  $\epsilon$ -pseudospektra  $\sigma_\epsilon(A)$ , kde za  $\epsilon$  jsou postupně dosazovány hodnoty  $10^{-2}, 10^{-3}, \dots, 10^{-8}$ . Přerušovaná čára (někdy mimo měřítko a tudíž neviditelná) je hranicí pole hodnot matice  $A$  spočtených podle algoritmu ... Vlastní čísla matice  $A$  jsou označena výraznými body.

### 1. Jordanův blok

Začneme asi nejznámějším příkladem matice, její vlastní čísla jsou citlivá vzhledem k perturbacím prvků. Touto maticí je Jordanův blok.

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

Připomeňme, že  $A_1$  má vlastní číslo 0 s algebraickou násobností 32. Označme jako  $\tilde{A}_1 = A_1 + E$ , kde  $E$  je matice dimenze 32, její jediný nenulový prvek je v levém dolním rohu a má hodnotu  $\epsilon$ . Matice  $\tilde{A}_1$  má 32 různých vlastních čísel, která leží na kruhu o poloměru  $\epsilon^{\frac{1}{32}}$ . Tomuto poznatku odpovídá i obrázek (3.5), kde jsou zachycena vlastní čísla ze pseudospektra  $\sigma_{10^{-3}}(A_1)$  pro 100 náhodně generovaných matic  $E$ . Všechna tato čísla leží v disku o 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. Tento jev nesouvisí s tím, že zobrazujeme pouze vlastní čísla pro perturbované matice  $A_1 + E$ , kde za  $E$  bereme náhodné matice s normou  $\|E\| = 10^{-3}$  místo abychom volili  $E$ , pro něž je  $\|E\| \leq 10^{-3}$ .

Hranice pseudospekter  $\sigma_\epsilon(A_1)$  pro různé hodnoty  $\epsilon$  tvoří soustředné disky se středem v počátku, jak je patrné z obrázku (3.5).

## 2. „super“ Jordanův blok

Jak uvidíme hned v dalším příkladě, není z hlediska popisu pseudospektra Jordanův blok typickým reprezentantem na třídě matic, které nejsou normální. Jordanův blok nemá žádný zvláštní význam ani mezi maticemi s násobným vlastním číslem. Hranice pseudospekter těchto matic mohou být totiž tvořeny křivkami odlišnými od soustředných disků. Takovým typickým příkladem je matice  $A_2$

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

Příslušná pseudospektra jsou zobrazena na obrázku (3.5).

## 3. Wilkinsonova matice

Třetí příklad je tzv. Wilkinsonova matice, která má tvar

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

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. Vůimněme si čísla podmíněnosti  $\kappa(V) > 10^{22}$  a srovnáme obrázek (3.5) s odhadem z věty 3.18.

## 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 dokázal, 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 obr. (3.5)). Srovnáme tyto obrázky s poznámkami za větou 3.16.

## 5. ??Lenferinkova-Spijkerova matice

Dalším příkladem je matice, kterou poprvé uvedli autoři Lenferink a Spijker

$$A_5 = \begin{pmatrix} -5 & 2 & & & \\ \frac{1}{2} & -7 & 3 & & \\ & \frac{1}{3} & -9 & 4 & \\ & & \ddots & \ddots & \ddots \\ & & & \frac{1}{31} & -67 & 32 \\ & & & & \frac{1}{32} & -69 \end{pmatrix}.$$

Označme jako  $D$  diagonální matici

$$D = \text{diag}(1, 2!, 3!, \dots, 32!).$$

Pak je matice  $D^{-1}A_5D$  symetrická (všechny prvky na vedlejších diagonálách jsou rovny jedné), tedy její spektrum je tvořeno reálnými čísly. Tudiž i spektrum  $\sigma(A_5)$  je reálné. Avšak již při malé perturbaci prvků matice  $A_5$  se příslušná vlastní čísla vzdalují od reálné osy. Jak je opět patrné z obrázku (3.5) jsou různá vlastní čísla různě citlivá vzhledem k perturbacím prvků. Přitom se zde projevuje jistá pravidelnost.

## 6. Náhodná matice

Na obrázku (3.5) jsou zakreslena příslušná pseudospektra pro matici  $A_6$ . Matice  $A_6$  je náhodně generovaná matice, její prvky jsou náhodné veličiny z komplexního normálního rozdělení se střední hodnotou 0 a standardní hustotou  $N^{-\frac{1}{2}}$ . Navíc pro ni platí,  $\Re \rho(A_6) \approx 1$  a  $\|A_6\| \approx 2$ .

Obrázek pro pseudospektrum  $\sigma_{10^{-3}}(A_6)$  je zcela odlišný od příslušných obrázků ve všech zatím uvedených příkladech. Místo 3200 bodů je zachyceno pouhých 32. Každý z nich totiž představuje svou stonásobnou kopii. To znamená, že perturbace  $E$  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. V aplikacích totiž typicky vznikají matice, které mají velkou odchylku od normality a tudíž jejich vlastní čísla jsou citlivá vzhledem k perturbacím.

## 7. 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 (věnujme si také, jak se změnilo  $\kappa(V)$  oproti předchozímu příkladu).

NO FILE: a1.ps

Obrázek 3.5:

NO FILE: a2.ps

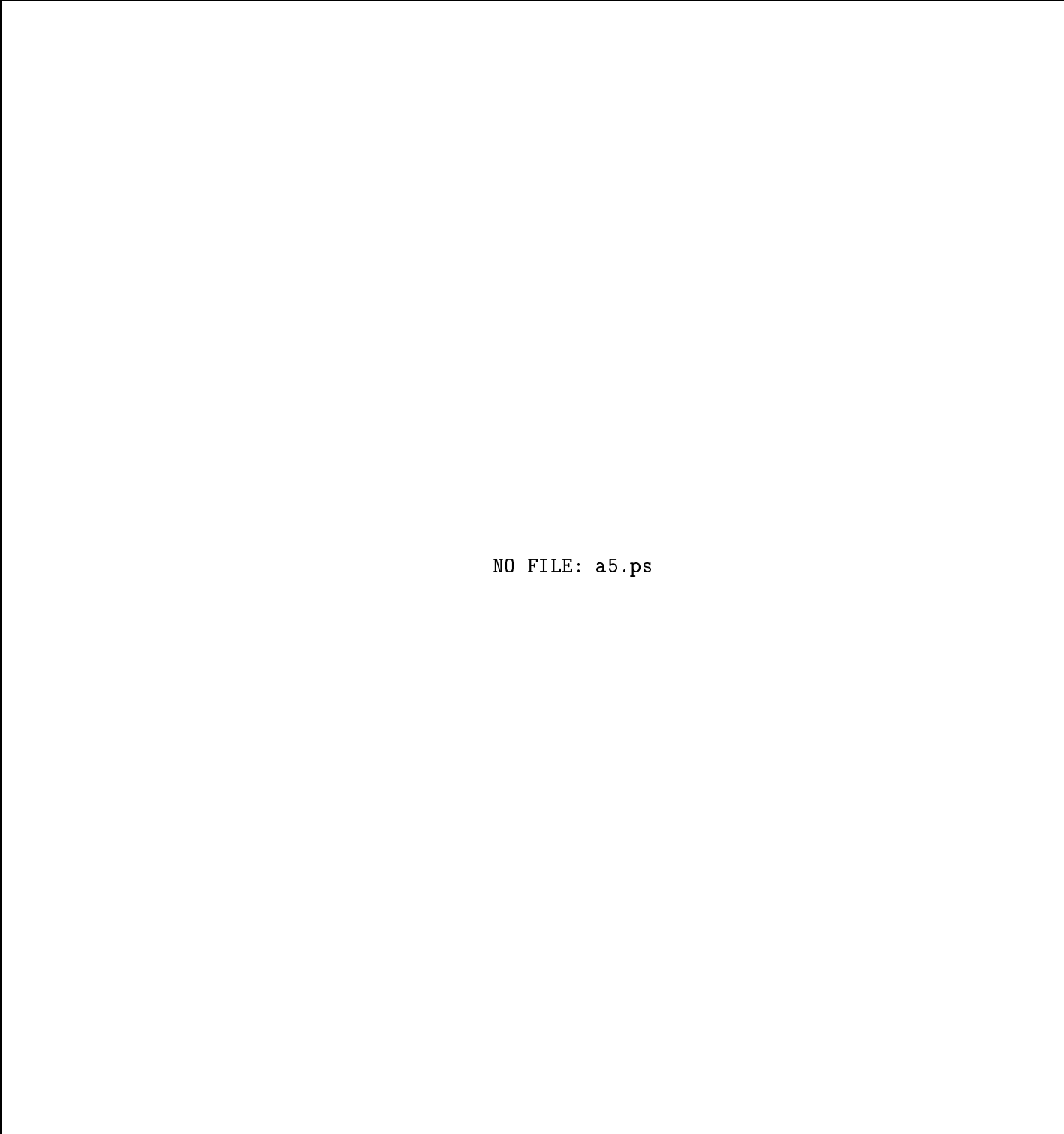
Obrázek 3.6:

NO FILE: a3.ps

Obrázek 3.7:

NO FILE: a4.ps

Obrázek 3.8:



NO FILE: a5.ps

Obrázek 3.9:

NO FILE: a6.ps

Obrázek 3.10:

NO FILE: a7.ps

Obrázek 3.11:

## Kapitola 4

# Citlivost řešení soustav lineárních rovnic

Máme řešit  $Ax = b$ ,  $A \in R^{N,N}$ ,  $b \in R^N$ ,  $A$  nonsingulární. Budeme vyšetřovat, jak souvisí řešení  $\tilde{x} = x + \delta x$  soustavy s perturbovanými vstupními daty a řešení  $x$  původní soustavy. Nalezneme tu vlastnost matice  $A$ , která má pro velikost rozdílu  $\tilde{x} - x$  rozhodující význam. Budeme se postupně zabývat třemi možnými případy: za prvé je perturbována jen pravá strana soustavy, za druhé je perturbována jen matice soustavy a za třetí je perturbována jak pravá strana, tak matice soustavy.

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

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

**Důkaz**

$$Ax + A \delta x = b + \delta b$$

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

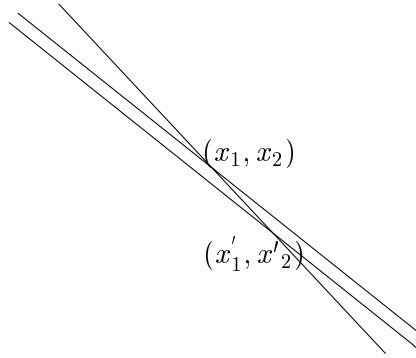
$$\|A\| \|x\| \geq \|b\|$$

$$\frac{\|\delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\delta b\|}{\|b\|} \quad \square$$

**Definice 4.1** *Řekneme, že odhad  $y$  v nerovnosti  $x \leq y$  je ostrý, jestliže se nedá zlepšit.*

**Poznámka 4.1** *Spektrální norma matice je generovaná euklidovskou normou vektoru, tedy  $\exists x_0$  takové, že  $\|A\| = \max_{\|x\|=1} \|Ax\| = \|Ax_0\|$ . Pro  $x_0$  nastává ve výrazu  $\|Ax\| \leq \|A\| \|x\|$  rovnost. Tedy tento odhad nelze zlepšit, a protože jsme v důkaze věty 1 nepoužili jiné nerovnosti, je odhad v (4.1) ostrý.*

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



Obrázek 4.1: Vliv perturbace pravé strany na poruchu řešení je-li  $\kappa(A) \gg 1$

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

$$a_{11}x_1 + a_{12}x_2 = b_1$$

$$a_{21}x_1 + a_{22}x_2 = b_2$$

*Řešení každé z těchto soustav je přímka v  $R^2$ . Řešením první soustavy je přímka kolmá na vektor  $(a_{11}, a_{12})$ , řešením druhé přímka kolmá na vektor  $(a_{21}, a_{22})$ . Řešení soustavy  $(x_1, x_2)$  je v jejich průsečíku.*

*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. Potom i malá perturbace, např. v  $b_1$ , způsobí malý posun první přímky, ale velký posun řešení  $(\tilde{x}_1, \tilde{x}_2)$  vzhledem k  $(x_1, x_2)$  (viz obrázek (4.1)).*

*Na obrázku (4.2) je znázorněna stejná situace pro matici, jejíž číslo podmíněnosti  $\kappa(A) \gg 1$ .*

Další případ, který může nastat, je ten, kdy je perturbována jen matice soustavy. Nejprve je třeba najít podmínky, za kterých má tato úloha vůbec smysl, tj. kdy má  $(A + \delta A)\tilde{x} = b$  jednoznačné řešení.

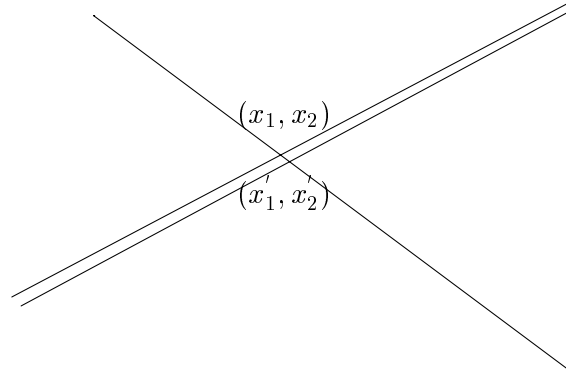
**Věta 4.2** *Je-li  $A \in R^{N,N}$  nonsingulární,  $\delta A \in R^{N,N}$  a platí*

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

*pak je  $A + \delta A$  také nonsingulární.*

**Důkaz** Podmínku

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



Obrázek 4.2: Vliv perturbace pravé strany na poruchu řešení platí-li  $\kappa(A) \gg 1$

lze vyjádřit ve tvaru  $\|A^{-1}\| \|\delta A\| < 1$ . Místo tvrzení věty budeme dokazovat obrácené tvrzení, tj.: je-li  $A + \delta A$  singulární, pak je  $\|A^{-1}\| \|\delta A\| \geq 1$ . Je-li  $A + \delta A$  singulární, pak  $\exists z \neq 0$  tak, že  $(A + \delta A)z = 0$ , tedy  $z = -A^{-1}\delta Az$ . Odhadneme-li tuto normu, máme:

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

Protože jsme předpokládali, že  $z \neq 0$ , je  $\|z\| > 0$  a tedy po vydělení nerovnosti  $\|z\|$  dostaneme tvrzení.  $\square$

**Poznámka 4.2** Všimněme si, jak souvisí  $\kappa(A)$  se vzdáleností  $A$  od nejbližší singulární matice: je-li  $A + \delta A$  singulární, je  $\frac{\|\delta A\|}{\|A\|} \geq \frac{1}{\kappa(A)}$ .

**Věta 4.3** Mějme  $A \in R^{N,N}$  nonsingulární,  $\delta A \in R^{N,N}$  a necht' je splněno

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

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

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

**Důkaz**

$$\begin{aligned} (A + \delta A)(x + \delta x) &= b \\ Ax + A\delta x + \delta A(x + \delta x) &= b \\ \delta x &= -A^{-1}\delta A(x + \delta x) \\ \|\delta x\| &\leq \|A^{-1}\| \|\delta A(x + \delta x)\| \leq \|A^{-1}\| \|\delta A\| (\|x\| + \|\delta x\|) \end{aligned} \quad (4.3)$$

Po vynásobení pravé strany nerovnosti (4.3) podílem  $\frac{\|A\|}{\|\delta A\|}$  a převedení členu s  $\|\delta x\|$  na levou stranu dostáváme

$$(1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}) \|\delta x\| \leq \kappa(A) \frac{\|\delta A\|}{\|A\|} \|x\|, \quad (4.4)$$

$1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}$  je podle předpokladu kladné, tedy jím můžeme celou nerovnost vydělit, aniž by se obrátila, a dostaneme tvrzení.  $\square$

**Poznámka 4.3** Při odvozování odhadu (4.2) byla použita trojúhelníková nerovnost:  $\|x + y\| \leq \|x\| + \|y\|$ . Tento odhad zřejmě není (až na triviální případy) ostrý, tedy ani odhad v (4.2) není ostrý.

Z věty 4.3 je zřejmé, že pokud je  $A$  „dobře“ podmíněná a  $\frac{\|\delta A\|}{\|A\|}$  je dostatečně malé, pak je  $\frac{\|\delta A\|}{\|A\|} \kappa(A) \ll 1$  a tudíž jmenovatel v (4.2) je blízký jedné, takže  $\frac{\|\delta x\|}{\|x\|} \lesssim \kappa(A) \frac{\|\delta A\|}{\|A\|}$ , tedy  $\frac{\|\delta x\|}{\|x\|}$  je malé.

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

**Příklad 4.2** Mějme matici  $A \in R^{N,N}$  s číslem podmíněnosti  $\kappa(A) = 10^6$  a vezměme takovou její perturbaci, pro niž  $\frac{\|\delta A\|}{\|A\|} = 2 \times 10^{-7}$ . Tedy  $\frac{\|\delta A\|}{\|A\|} = \frac{1}{5} (\frac{1}{\kappa(A)})$  a odhad pro normu chyby řešení (podle věty 4.3) je  $\frac{\|\delta x\|}{\|x\|} \leq \frac{\frac{1}{5}}{\frac{4}{5}} = \frac{1}{4}$ .

Tento odhad není ostrý. Nezaručuje existenci takové perturbace  $\delta A$ , pro niž je  $\frac{\|\delta x\|}{\|x\|} = \frac{1}{4}$ . Je pouhým varováním, že se tomuto číslu můžeme libovolně přiblížit.

**Věta 4.4** Mějme  $A \in R^{N,N}$  nonsingulární,  $\delta A \in R^{N,N}$ ,  $b \in R^{N,N}$ ,  $\delta b \in R^{N,N}$  a necht' platí

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

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

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

**Důkaz**

$$(A + \delta A)(x + \delta x) = b + \delta b$$

$$Ax + A\delta x + \delta A(x + \delta x) = b + \delta b$$

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

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

po vynásobení pravé strany nerovnosti podílem  $\frac{\|A\|}{\|\delta A\|}$ , dostáváme

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

Protože  $A$  je nonsingulární, je  $A \neq 0, b \neq 0$ , lze vztah  $\|A\|\|x\| \geq \|b\|$ , upravit jako  $\frac{\|x\|}{\|b\|} \geq \frac{1}{\|A\|}$ . Dosazením a jednoduchou úpravou získáme tvrzení věty.  $\square$

## Kapitola 5

# Odhady chyb a zpětná stabilita

### 5.1 Vlastní čísla

### 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 chybu nějaké spočtené aproximace  $\tilde{x}$ , je třeba ještě provést zpětnou analýzu chyb. To znamená nalézt perturbace vstupních dat  $\delta A$ ,  $\delta b$  tak, aby platilo  $(A + \delta A)\tilde{x} = b + \delta b$ .

To umožňuje následující věta.

**Věta 5.1** (*Rigal, Gaches*) *Nechť  $A \in R^{N,N}$  je nonsingulární,  $b \in R^N$ ,  $Ax = b$  a  $\tilde{x}$  je aproximace řešení této soustavy. Pak existují takové perturbace  $\delta A$ ,  $\delta b$ , pro které je*

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

a platí

$$\nu(\tilde{x}) = \min\left\{\nu; \frac{\|\delta A\|}{\|A\|} \leq \nu; \frac{\|\delta b\|}{\|b\|} \leq \nu\right\} = \frac{\|b - A\tilde{x}\|}{\|A\| \|\tilde{x}\| + \|b\|}.$$

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

$$\delta A = \frac{\|A\| \|\tilde{x}\|}{\|A\| \|\tilde{x}\| + \|b\|} \frac{r\tilde{x}^T}{\|\tilde{x}\|^2}. \quad (5.1)$$

Nejdříve ukážeme, že pro takto definované  $\delta A$  platí

$$(A + \delta A)\tilde{x} = b + \delta b. \quad (5.2)$$

Dosazením za  $\delta A$  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

$$\delta b = -\frac{\|b\|}{\|A\| \|\tilde{x}\| + \|b\|} r, \quad (5.3)$$

dostaneme (5.2).

Dále dokážeme, že pro takto zvolená  $\delta A, \delta b$  je  $\nu(\tilde{x})$  ve tvaru uvedeném v tvrzení věty. Protože  $r\tilde{x}^T$  je matice  $N \times N$ , platí pro její normu:

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

kde jsme použili základní vlastnosti normy vektoru a faktu, že  $\tilde{x}^T z$  je skalár. Z definice skalárního součinu dvou vektorů, které svírají úhel  $\phi$  dostáváme:

$$\max_{\|z\|=1} |\tilde{x}^T z| = \max_{\|z\|=1} \|\tilde{x}\| \|z\| |\cos \phi| = \|\tilde{x}\|.$$

Dosadíme-li za  $\|r\tilde{x}^T\|$  máme pro normy výrazů (5.1) a (5.3)

$$\begin{aligned} \|\delta A\| &= \frac{\|b - A\tilde{x}\|}{\|A\| \|\tilde{x}\| + \|b\|} \|A\| \\ \|\delta b\| &= \frac{\|r\|}{\|A\| \|\tilde{x}\| + \|b\|} \|b\|. \end{aligned}$$

Nyní zbývá ukázat, že takto definované perturbace jsou skutečně minimální. Důkaz provedeme sporem.

Předpokládejme, že existují takové perturbace vstupních dat  $\delta' A, \delta' b$ , pro něž je

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

$$\frac{\|\delta' A\|}{\|A\|} < \nu(\tilde{x}) \quad \frac{\|\delta' b\|}{\|b\|} < \nu(\tilde{x}). \quad (5.4)$$

Pro  $\nu(\tilde{x})$  pak máme:

$$\nu(\tilde{x}) = \frac{\|b - A\tilde{x}\|}{\|A\| \|\tilde{x}\| + \|b\|} = \frac{\|\delta' A\tilde{x} - \delta' b\|}{\|A\| \|\tilde{x}\| + \|b\|} \leq \frac{\|\delta' A\| \|\tilde{x}\| + \|\delta' b\|}{\|A\| \|\tilde{x}\| + \|b\|} < \nu(\tilde{x}),$$

kde jsme poslední nerovnost dostali dosazením (5.4). □

**Poznámka 5.1** • *Věta 5.1 zaručuje existenci  $\nu(\tilde{x})$  (pro dané  $\tilde{x}$  ho umíme spočítat).*

- $\nu(\tilde{x})$  je pro perturbace dat ostrým odhadem.
- Pokud je  $\nu(\tilde{x})$  velké, znamená to, že algoritmus, kterým byla spočtena aproximace řešení  $\tilde{x}$  není zpětně stabilní. Aproximace řešení  $\tilde{x}$  je pak řešením úlohy, která nemusí mít s původní úlohou žádnou souvislost.

Někdy je výhodnější počítat aposteriorní odhad po složkách. Tento odhad dává následující věta.

**Věta 5.2** (Oettli, Prager, 1964) *Nechť  $A \in R^{N,N}$  je nonsingulární,  $b \in R^N$ ,  $Ax = b$  a  $\tilde{x}$  je aproximace řešení této soustavy. Mějme dále reálnou nezápornou matici  $E = (e_{ij})$  ( $e_{ij} \geq 0$  pro  $i, j = 1, \dots, N$ ) a reálný nezáporný vektor  $f = (f_i)$  ( $f_i \geq 0$  pro  $i = 1, \dots, N$ ). Pak existují takové perturbace  $\delta A = (\delta a_{ij})$ ,  $\delta b = (\delta b_i)$ , pro které je*

$$(A + \delta A)\tilde{x} = b + \delta b \quad (5.5)$$

a platí

$$\omega(\tilde{x}) = \min\{\omega; |\delta a_{ij}| \leq \omega e_{ij}, |\delta b_i| \leq \omega f_i, i, j = 1, \dots, N\} = \max_i \frac{|r_i|}{(E|\tilde{x}| + f)_i},$$

kde  $|\tilde{x}| = (|x_1| \dots |x_N|)^T$  předpokládáme, že  $\omega(\tilde{x}) \neq \infty$  a „ $\frac{0}{0}$ “ je interpretováno jako 0.

**Důkaz:** Podle předpokladu pro každou složku vektoru rezidua  $r = b - A\tilde{x}$  platí:

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

Tedy  $r$  lze vyjádřit jako

$$r = D(E|\tilde{x}| + f),$$

kde pro diagonální matici

$$D = \text{diag}(d_{11} \dots d_{NN})$$

je

$$|D| \leq \omega(\tilde{x}) I \quad \text{nebo-li} \quad |d_i| \leq \omega(\tilde{x}) \quad \text{pro } i = 1, \dots, N.$$

Definujme perturbace  $\delta A$ ,  $\delta b$

$$\begin{aligned} \delta A &= D E \text{diag}(\text{sign}\tilde{x}_1, \dots, \text{sign}\tilde{x}_N) \\ \delta b &= -Df. \end{aligned}$$

Dosazením zjistíme, že pro takto zvolené perturbace je  $\tilde{x}$  řešením (5.5). Je také zřejmé, že tyto perturbace realizují  $\omega(\tilde{x})$ .

Ještě zbývá dokázat, že takto definované  $\omega(\tilde{x})$  je optimální. Nechť pro nějaké perturbace  $\delta' A$ ,  $\delta' b$  a kladné reálné číslo  $\omega$  je splněno

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

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

Pro vektor rezidua potom platí

$$|r| = |b - A\tilde{x}| = |\delta' A\tilde{x} - \delta' b| \leq |\delta' A||\tilde{x}| + |\delta' b| \leq \omega(E|\tilde{x}| + f), \quad (5.6)$$

čímž je tvrzení dokázáno, neboť z (5.6) ihned dostáváme

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

□

**Poznámka 5.2** *Zvolíme-li  $E = |A|$ ,  $f = |b|$ , dostaneme velikost relativní zpětné chyby po složkách.*

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 ukázkou profesionálního a poučeného přístupu k řešení některých úloh numerické lineární algebry může čtenář nalézt v [?].

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