Pokroky matematiky, fyziky a astronomie

Zdeněk Strakoš

Metoda konjugovaných gradientů jako dobrodružství jdoucí přes staletí

Pokroky matematiky, fyziky a astronomie, Vol. 65 (2020), No. 4, 197–222

Persistent URL: http://dml.cz/dmlcz/148475

Metoda konjugovaných gradientů jako dobrodružství jdoucí přes staletí

Zdeněk Strakoš

Abstrakt. Metoda konjugovaných gradientů a Lanczosova metoda tvoří historický a metodologický základ tzv. metod krylovovských podprostorů pro numerickou aproximaci řešení lineárních rovnic a částečnou aproximaci spektra lineárních operátorů. Ačkoliv jsou v obecném povědomí spojovány především s numerickým řešením velmi rozsáhlých soustav lineárních algebraických rovnic a aproximací vlastních čísel velkých matic, je přirozené uvažovat jejich formulaci v kontextu operátorů na Hilbertových prostorech (konečné či nekonečné dimenze). Ostatně ani v algebraické formulaci nemusí být matice soustavy vůbec sestavována, neboť výpočet používá pouze aplikaci odpovídajícího operátoru na vektor.

Principiální vztah metody konjugovaných gradientů a Lanczosovy metody k problému momentů, k teorii ortogonálních polynomů, Jacobiho matic, řetězových zlomků a Gaussovy kvadratury z nich činí také objekt čistě matematického zájmu. Využití hlubokých matematických souvislostí postupně vedlo k pochopení *adaptivního silně nelineárního chování* obou metod včetně vlivu aritmetiky s konečnou přesností na praktické výpočty. V nejlepším smyslu se zde setkává matematický a informatický pohled.

Příležitost, kterou prolnutí obou oborů přináší, však není využita při zúženém chápání metody konjugovaných gradientů a Lanczosovy metody jako pouhých algoritmických výpočetních nástrojů, které je bohužel až na výjimky rozšířeno napříč literaturou. To má zhoubné důsledky. Pro většinu matematiků (a následně i vědců z jiných oborů, inženýrů a praktiků provádějících výpočty v aplikacích) dominuje v pojetí metody konjugovaných gradientů lineární odhad poklesu velikosti chyby *odpovídající Čebyševově metodě*. Mnoho učebnicových popisů metody konjugovaných gradientů a stejně tak článků na ni odkazujících je pak zatíženo řadou mýtů, nedorozumění i nepřiznaných omylů. Matematicky zcela zmatečné je pojetí metody konjugovaných gradientů pro řešení lineárních rovnic, kdy jsou jednotlivé iterace těsně navzájem provázány podmínkou optimality na podprostorech rostoucí dimenze, jako *zjednodušení* gradientních metod pro řešení nelineárních rovnic.

Příklad metody konjugovaných gradientů a Lanczosovy metody ukazuje, jak krásná a zároveň obtížná i úzká může být cesta k porozumění. Vede nás k trpělivosti a houževnatosti, ale hlavně nás učí pokoře.

1. Úvod

Budeme sledovat pozoruhodný matematický příběh. I matematické příběhy bývají ovlivněny různými nematematickými okolnostmi a zmatky. Dobrým vodítkem při jeho sledování je snaha po rozlišování podstatného a trvale platného od nepodstatného a pomíjejícího. Nebezpečím je povrchnost, které se bohužel i ve vědě příliš často dobrovolně podřizujeme. Metoda konjugovaných gradientů a krylovovské metody obecně nepochybně představují obtížnou matematickou výzvu, které nelze při povrchním přístupu porozumět.

Prof. Ing. Zdeněk Strakoš, DrSc., Katedra numerické matematiky MFF UK, Sokolovská 83, 186 75 Praha 8, e-mail: strakos@karlin.mff.cuni.cz

Ač se budeme snažit o co největší přístupnost textu, budeme nevyhnutelně používat matematická tvrzení a odkazovat na netriviální matematické souvislosti, které budou někdy jen naznačeny. Čtenáře se zájmem o bližší seznámení s tématem odkazujeme v prvé řadě na seminální práce Hestenese, Stiefela a Lanczose z počátku padesátých let [23], [25], [26], [27]. Zdůrazňujeme nutnost chápat je jako těsně vzájemně propojený celek. Opomíjení původní literatury a podstatných souvislostí (z neznalosti i z úmyslu) není novým jevem. S rostoucím informačním zahlcením a naší ztrátou schopnosti rozlišovat je ale nebezpečí následných zmatků mnohem větší než dříve. Návraty k původním pracím a snaha o porozumění jejich hloubce (i jejich možným omylům) jsou důležité pro pochopení podstaty věcí. Jak uvidíme dále, přispívají k řešení otázek, které původní práce ani nemohly v plnosti předjímat.[1]

Ke krylovovským metodám jsem se dostal oklikou přes paralelní počítačové architektury a numerickou nestabilitu způsobenou šířením zaokrouhlovacích chyb, což mne přivedlo ke spolupráci s vynikajícími osobnostmi oboru (A. Greenbaum, G. H. Golub, C. C. Paige, G. Meurant, M. Gutknecht, A. Björck, J. Liesen), za kterou jsem velmi vděčný.

Uvedu některé z publikací, které mohou čtenáři posloužit k dalšímu čtení. Shrnující zpracování oblasti krylovovských metod z různých pohledů lze nalézt například v monografiích [39], [13], [18]. Z publikací, u kterých jsem spoluautorem, budu často odkazovat na knihu [29] a následující texty. Výsledky popisující chování metody konjugovaných gradientů a Lanczosovy metody při výpočtech s konečnou přesností jsou předmětem rozsáhlého přehledu [32] a stručnějšího článku [41]. Formulace metody konjugovaných gradientů na nekonečnědimenzionálních Hilbertových prostorech je v krátké monografii [31] propojena s analýzou modelové úlohy řešení eliptické parciální diferenciální rovnice, s formulací konečnědimenzionální aproximace úlohy a s její transformací (tzv. předpodmíněním) s cílem umožnit rychlejší výpočet. Různé přístupy k paralelním implementacím, poznámky k jejich historii i perspektivám lze nalézt v [7] a [8]. Všechny uvedené práce odkazují na důležitou dřívější literaturu; jen [29] obsahuje téměř 700 citací. Z nich může být dále v textu zmíněn jen nepatrný zlomek a výběr je přirozeně ovlivněn osobním pohledem. Proto doporučujeme čtenáři v případě zájmu srovnání a rozšíření s použitím textů dle vlastního výběru. Čtenáři orientovanému na česky psanou literaturu můžeme nabídnout základní informace (jak obecného charakteru tak o metodě konjugovaných gradientů a o Lanczosově metodě) v učebním textu [11].

Přehledový text k tématu, které bylo ve stejném časopise zpracováno dříve, by měl s užitkem navázat na dřívější publikaci. Nelze se proto vyhnout článku [5], který byl publikován v PMFA u příležitosti padesátého výročí zveřejnění práce Hestenese a Stiefela [23]. Bohužel však zde pozitivní navázání není možné. Předložený text je s [5] v zásadním rozporu jak výkladem myšlenek obsažených v [23] (a [25], [26], [27]), tak výkladem jejich rozvinutí v pozdějších textech a souvislostí s výsledky z jiných matematických oborů důležitých pro pochopení metody konjugovaných gradientů. Předložený text zároveň pojmenovává závažné chyby, zkreslení a omyly rozšířené v literatuře.

[1] Nejde nám zde však o popis historie a ani o vyčerpávající popis všech věcných a časových souvislostí, to ani není v lidských silách.

2. Vymezení úlohy a základ přístupů k jejímu řešení

Předmětem našeho zájmu je numerické řešení soustavy N lineárních algebraických rovnic o N neznámých, které v maticové formě budeme zapisovat

$$Ax = b, \tag{1}$$

kde A je čtvercová matice o N řádcích a sloupcích, x a b jsou sloupcové vektory délky N. Budou nás zajímat rozsáhlé soustavy; v současnosti se běžně na počítačích řeší úlohy s miliony a nezřídka i s miliardami až stovkami miliard proměnných. Z důvodu zjednodušení výkladu a značení budeme uvažovat pouze reálná data a budeme předpokládat, že matice A má lineárně nezávislé sloupce (a tudíž i řádky). Pak má rovnice (1) pro každou pravou stranu právě jedno řešení, které můžeme symbolicky zapsat pomocí inverze matice

$$x = A^{-1}b. \tag{2}$$

Samozřejmě výpočet nekonstruuje (s výjimkou velmi speciálních případů) inverzi matice s následným násobením pravou stranou. Takový postup je obecně nerozumný jak z hlediska výpočetní náročnosti, tak i z hlediska numerické nestability (šíření zaokrouhlovacích chyb). Nejznámnější přístupy k numerickému řešení soustavy (1) (ponecháme stranou teoreticky důležité, ale prakticky stěží použitelné postupy typu Cramerova pravidla) pracují s *rozklady matice* A. Rozlišíme dva případy.

V prvním případě uvažujme rozklad na *součin matic* se speciálními vlastnostmi, které umožňují algoritmicky jednoduché řešení, například

$$A = LU, \tag{3}$$

kde L je dolní trojúheníková a U horní trojúhelníková matice s nulami nad respektive pod hlavní diagonálou. Řešení pak lze symbolicky zapsat

$$x = U^{-1}(L^{-1}b). \tag{4}$$

Zápis $L^{-1}b$ představuje řešení soustavy $Ly = b$ a x je dáno následným řešením soustavy $Ux = y$. Nejde o nic jiného než o základní variantu Gaussovy eliminace, kde je rozklad (3) spolu s řešením soustavy $Ly = b$ proveden přímým eliminačním chodem, a soustava $Ux = y$ představuje zpětnou substituci.

Gaussova eliminace i jiné vhodné přístupy založené na součinovém rozkladu mají velmi podstatnou výhodu ve své *robustnosti*. Jejich chování včetně šíření zakrouhlovacích chyb dobře rozumíme. Díky pracem Goldstina, von Neumanna, Turinga, Forsytha, Wilkinsona a dalších, viz např. [29], historickou poznámku 5.8.2, strany 315–316, umíme indikovat případné numerické nestability a následně je v případe nutnosti ošetřit, byť někdy za značnou cenu. Navíc moderní varianty Gaussovy eliminace a její paralelní implementace umí velmi rychle pracovat i s velkými maticemi se složitou strukturou nenulových prvků. Mají ale také podstatnou nevýhodu. Dokud neprovedeme všechny výpočetní operace symbolicky zapsané v (4), nemáme k dispozici žádnou obecně použitelnou aproximaci řešení.

Termín *aproximace řešení* vyžaduje komentář. Každá rozumná implementace Gaussovy eliminace zaručí velikost (normu) rezidua

$$r = b - A\widehat{x}$$

spočtené aproximace řešení $\widehat{x}$ úměrnou strojové přesnosti počítače. Reziduum na úrovni strojové přesnosti však neznamená zanedbatelnou chybu řešení

$$x - \widehat{x} = A^{-1}r.$$

Zde může být jeden ze zdrojů nedorozumění rozšířeného ve velké části literatury zabývající se odhady chyb v numerickém řešení parciálních diferenciálních rovnic (PDEs), která zanedbává algebraickou chybu a předpokládá (téměř) přesné řešení vzniklých lineárních algebraických rovnic (1).

Použití metod založených na součinovém rozkladu může být pro řadu úloh příliš drahé (pro velmi velké úlohy dokonce nemusí být ani proveditelné). Často je navíc účelné uvažovat pouze přibližné řešení s přesností danou předem uživatelem. V metodách typu Gaussovy eliminace není možné tímto způsobem postupovat.[2]

Jako alternativu proto můžeme zvolit počáteční přiblížení x_0 a hledat posloupnost vektorů x_n, $n = 1, 2, \ldots$ tak, abychom za přijatelnou cenu výpočtu získali aproximaci řešení x_n s přesností předepsanou uživatelem, která často zahrnuje kontext formulace úlohy přesahující algebraická data A, b; viz například diskusi v [8].

Tím se dostáváme k rozkladu matice soustavy na *součet matic* (někdy také nazývaný štěpením), který lze zapsat

$$A = K - Z, \tag{5}$$

kde matice K má lineárně nezávislé sloupce (má tudíž inverzi) a každý prvek matice A je v jednoduchém případě buď součástí K nebo Z (zobecnění není obtížné). Přepsáním rovnice (1) dostaneme

$$Kx = Zx + b = (K - A)x + b.$$

Následně volbou počátečního přiblížení x_0 a výpočtem nové aproximace z levé strany rovnosti (symbolicky zapsaným pomocí inverze matice K) zkonstruujeme iteraci ve dvou ekvivalentních zápisech

$$\begin{aligned} x_n &= K^{-1}Zx_{n-1} + K^{-1}b = x_{n-1} + K^{-1}r_{n-1}, \\ r_{n-1} &= b - Ax_{n-1}, \quad n = 1, 2, \ldots \end{aligned} \tag{6}$$

Volbou štěpení matice A (a případně i dalšími úpravami) dostáváme například metodu prosté iterace, Jacobiho metodu, Gaussovu–Seidelovu metodu, superrelaxační metodu atd.

Všimněme si, že postup výpočtu iterace x_n v (6) nezávisí na tom, jsme-li v první, padesáté či jakékoli jiné iteraci. Můžeme proto očekávat, že po ustálení výpočtu (po překonání tzv. přechodového jevu) bude pokles velikosti chyby v jednotlivých iteracích měřený vhodnou normou blízký konstantě, což je nazýváno lineární rychlostí konvergence (nebo zkráceně lineární konvergencí). Nelze naopak očekávat, že metody odvozené z (6) dají (s výjimkou triviálního případu nebo náhody) přesné řešení. Proto zde hovoříme o *asymptotické lineární konvergenci*. Lze ji popsat jedním číslem, tzv. *asymptotickým konvergenčním faktorem*.

[2]Podrobný výklad s mnoha odkazy na literaturu lze nalézt např. v [31], kapitolách 1, 11 a 12, v článku vyjadřujícím se k různorodosti chápání pojmu ceny iteračních výpočtů [8] a v [29], kapitolách 1 a 5.

Výhodou metod odvozených z (6) je jejich jednoduchost. Nevýhodou je jejich obecná pomalost pramenící z jejich stacionárního charakteru.[3] Předpis (6) založený na fixním rozkladu používá ve všech iteracích stejnou informaci o řešené soustavě a neumožňuje postupné získávání a využití další informace v průběhu výpočtu (neumožňuje *adaptivitu*). Aby byla adaptivita možná, musí se způsob výpočtu jednotlivých iterací měnit s iteračním krokem.

Nabízí se aproximace řešení zapsaného formálně pomocí inverze matice vztahem (2) pomocí polynomu

$$x = A^{-1}b = x_0 + A^{-1}r_0 \approx x_n = x_0 + \phi_{n-1}(A)\, r_0, \quad n = 1, 2, \ldots, \tag{7}$$

kde $r_0 = b - Ax_0$ je počáteční reziduum a $\phi_{n-1}(\lambda)$ je polynom stupně $n-1$. Je zde nutné zdůraznit, že tak jako symbolický zápis (2) neznamená provedení inverze matice a následné násobení vektorem pravé strany, tak ani maticový polynom $\phi_{n-1}(A)$ není konstruován jako aproximace inverze matice A s následným násobením počátečním reziduem. Musí být konstruován ve vztahu k řešení x a musí tedy využít informaci jak o matici A, tak o pravé straně b.

Ze (7) vyplývá, že aproximace řešení x_n hledáme opravou počáteční aproximace x_0 o vektor, který je lineární kombinací počátečního rezidua zobrazeného různými mocninami matice A, leží tedy v tzv. krylovovském prostoru $\mathcal{K}_n(A, r_0)$ posunutém o počáteční aproximaci x_0,

$$x_0 + \mathcal{K}_n(A, r_0) = x_0 + \mathrm{span}\{r_0, Ar_0, \ldots, A^{n-1}r_0\}\,. \tag{8}$$

Chceme-li zkonstruovat efektivní metodu, stojí nyní před námi dvě otázky:

- Pro jaká data (matici A, pravou stranu b a případně pro vhodně zvolenou počáteční aproximaci x_0) bude v krylovovském prostoru $\mathcal{K}_n(A, r_0)$ přijatelné dimenze n dost informace o hledaném řešení $x = A^{-1}b$?
- Jak nalézt aproximaci řešení, která informaci obsaženou v krylovovském prostoru co nejlépe využije?

Oběma otázkami se budeme zabývat dále. Začneme druhou z nich.

3. Odvození metody konjugovaných gradientů

V dalším textu budeme předpokládat, že reálná matice A je symetrická pozitivně definitní (SPD), to jest $A = A^*$ a $z^*Az > 0$ pro každý nenulový vektor z, kde pro reálnou matici značíme hvězdičkou transpozici. Pak má v mnoha fyzikálních aplikacích smysl hledat takové aproximace řešení, které na určeném podprostoru minimalizují tzv. energetický funkcionál[4]

$$J(z) = \tfrac{1}{2} z^*Az - z^*b;$$

[3] Vzhledem ke schopnosti tzv. zhlazování (redukce velikosti složek vektorů odpovídajících ve fourierovském rozkladu vysokým frekvencím) jsou však důležité například jako součást multigridních metod. Mohou být také výhodné ve speciálních případech, kdy je výrazné překonání lineární rychlosti konvergence pro složitější (a výpočetně dražší) metody nedosažitelné.

[4] Minimum funkcionálu $J(z)$ na $\mathcal{R}^N$ je dosaženo v řešení soustavy (1).

viz např. výklad a reference v [31], kapitolách 2 a 12, [29], sekcích 2.5.2 a 2.5.3. Definujeme-li pomocí symetrické pozitivně definitní matice A tzv. energetický skalární součin a jím indukovanou energetickou normu

$$(y,\, z)_a = z^* A y \quad \text{a} \quad \|z\|_a^2 = z^* A z,$$

můžeme pro libovolné z a pro řešení x úlohy (1) psát

$$J(z) = \tfrac{1}{2}(z-x)^* A (z-x) + J(x) = \tfrac{1}{2}\|z-x\|_a^2 + J(x).$$

Vzhledem k tomu, že $J(x)$ je konstanta, znamená minimalizace energetického funkcionálu vůči proměnné z totéž, co minimalizace energetické normy chyby.

Budeme uvažovat metodu hledající aproximace řešení v posunutých krylovovských podprostorech (8) rostoucí dimenze $n = 1, 2, \ldots$ a minimalizující v dané iteraci na daném podprostoru $\mathcal{K}_n(A, r_0)$ energetickou normu chyby $\|x - x_n\|_a$. Užitečný přepis

$$x - x_n = (x - x_0) - (x_n - x_0) = \underbrace{(x-x_0)|_{\mathcal{K}_n} - (x_n - x_0)}_{\in \mathcal{K}_n} + \underbrace{(x-x_0)|_{\mathcal{K}_n^{\perp_a}}}_{\in \mathcal{K}_n^{\perp_a}}$$

dává do vztahu chybu v n-tém kroku a počáteční chybu $x - x_0$, kde na pravé straně jsme použili ortogonální rozklad

$$x - x_0 = (x-x_0)|_{K_n} + (x-x_0)|_{K_n^{\perp_a}}$$

počáteční chyby na část ležící v n-tém krylovovském podprostoru $\mathcal{K}_n(A, r_0)$ a na část na daný podprostor kolmou ve smyslu energetického skalárního součinu. Je zřejmé, že minima energetické normy chyby $\|x - x_n\|_a$ je dosaženo právě tehdy, když

$$x_n - x_0 = (x-x_0)|_{\mathcal{K}_n}, \quad \text{přičemž} \quad x - x_n = (x-x_0)|_{\mathcal{K}_n^{\perp_a}}, \tag{9}$$

neboli právě tehdy, když $(x - x_n) \perp_a \mathcal{K}_n(A, r_0)$. Poslední podmínka je ekvivalentní kolmosti n-tého rezidua $r_n = A(x - x_n)$ na krylovovský prostor $\mathcal{K}_n(A, r_0)$ při použití standardního eukleidovského skalárního součinu, tj.

$$r_n^* w = 0 \quad \text{pro všechna } w \in \mathcal{K}_n(A, r_0). \tag{10}$$

Uvedená podmínka se nazývá *galerkinovskou ortogonalitou* (viz například [38], kapitolu 14) a námi hledaná metoda patří do třídy tzv. galerkinovských metod (podrobnou diskusi včetně zobecnění lze nalézt v [39]) .

Zavedení podmínky optimality k určení aproximace řešení x_n znamená velmi podstatnou nelineární závislost odpovídajícího polynomu $\phi_{n-1}(\lambda)$ a tím i x_n na vstupních datech A, b, x_0. Nejde již jen o mocniny matice A (či jejích částí, které jsou již přítomny v iteračním předpisu (6)). Galerkinovská ortogonalita váže počáteční reziduum a všechny odpovídající mocniny matice A v n-dimenzionální optimalizační úloze minimalizující $\|x - x_n\|_a$ na $\mathcal{K}_n(A, r_0)$.

Otázkou zůstává, jak co nejjednodušeji algoritmicky nalézt danou optimální aproximaci x_n v posunutém krylovovském podprostoru (8), který navíc (nejen z důvodu

Algoritmus 1 Metoda konjugovaných gradientů dle Hestenese a Stiefela (HSCG)

1: **Vstup:** SPD matice $A \in \mathcal{R}^{N\times N}$, pravá strana $b \in \mathcal{R}^N$, počáteční aproximace $x_0 \in \mathcal{R}^N$, maximální počet iterací $nmax$, zastavovací kritérium.
2: **Výstup:** Aproximace řešení x_n po zastavení algoritmu.
3: **Inicializace:** $r_0 = b - Ax_0$, $p_0 = r_0$
4: **for** $n = 1 : nmax$ **do**
5: $\alpha_{n-1} = (r^*_{n-1} r_{n-1})/(p^*_{n-1} A p_{n-1})$
6: $x_n = x_{n-1} + \alpha_{n-1} p_{n-1}, \quad r_n = r_{n-1} - \alpha_{n-1} A p_{n-1}$
7: Vyhodnocení zastavovacího kritéria.
8: $\beta_n = (r^*_n r_n)/(r^*_{n-1} r_{n-1})$
9: $p_n = r_n + \beta_n p_{n-1}$
10: **end for**

numerické nestability v praktickém výpočtu) není možné konstruovat pomocí generujících vektorů $r_0, Ar_0, \ldots, A^{n-1}r_0$. Řešením obou obtíží je konstrukce pomocí vhodně zvolené ortogonální báze.

V první iteraci máme k dispozici počáteční reziduum r_0, které zároveň definuje jednodimenzionální krylovovský podprostor. Zvolíme tedy první směrový vektor rovný počátečnímu reziduu $p_0 = r_0$ a následnou aproximaci řešení nalezneme opravou

$$x_1 = x_0 + \alpha_0 p_0,$$

kde α_0 je zvoleno tak, aby hodnota $\|x - x_1\|_a$ byla minimální.[5] Nové reziduum

$$r_1 = b - Ax_1 = r_0 - \alpha_0 A p_0$$

spolu s předchozím reziduem generuje následný krylovovský podprostor $\mathcal{K}_2(A,r_0)$. Zdánlivě bychom mohli postup zopakovat použitím rezidua r_1 jako nového směrového vektoru. Ihned ale narazíme na obtíž. Při volbě $p_1 = r_1$ by následná *lokální jednorozměrná* minimalizace hodnoty $\|x - x_2\|_a$ prostřednictvím volby α_1 nezaručila dosažení globálního mimima přes dvourozměrný podprostor $\mathcal{K}_2(A, r_0)$. Zvolíme proto směrový vektor p_1 opravou r_1 tak, aby p_0, p_1 rovněž generovaly podprostor $\mathcal{K}_2(A, r_0)$,

$$p_1 = r_1 + \beta_1 p_0,$$

a pokusíme se o vhodnou volbu parametru β_1. Není těžké objevit, že nutnou a postačující podmínkou pro dosažení globálního minima energetické normy chyby v celém podprostoru $\mathcal{K}_2(A, r_0)$ prostřednictvím následné jednorozměrné minimalizace pouze ve směru p_1 je ortogonalita směrových vektorů p_0 a p_1 ve smyslu energetického skalárního součinu. Toho lze v druhé iteraci volbou koeficientu β_1 snadno dosáhnout.

Nechme zatím na chvíli stranou otázku, nakolik půjde ortogonalitu směrových vektorů zaručit při pokračování iterací, a formulujme výpočetní postup s použitím výše

[5] To nastane v bodě x_1, ve kterém je (negativně vzatý) gradient $r_1 = b - Ax_1$ energetického funkcionálu $J(z)$ kolmý na směrový vektor p_0 (ve smyslu standardního eukleidovského skalárního součinu).

popsaných kroků ve formě algoritmu HSCG, kde koeficient β_n určíme z podmínky ortogonality $p_n^* A p_{n-1} = 0$ nového směrového vektoru p_n vůči jeho předchůdci.

Z jednotlivých kroků algoritmu HSCG je zřejmé, že k samovolnému zastavení může dojít buď v případě $r_n = 0$, tj. při dosažení řešení $x_n = x$, anebo v případě $p_n = 0$, kdy $r_n = -\beta_n p_{n-1}$. V důsledku předchozí volby α_{n-1} (viz motivaci popsanou výše pro první iteraci, případně [2], sekci 1.3) je $r_n^* p_{n-1} = 0$ a tudíž opět $r_n^* r_n = 0$. Za chvíli uvidíme, že při přesném výpočtu musí být řešení dosaženo v nejvýše N iteracích. Pouze pro jednoduchost značení a zkrácení délky textu předpokládejme, že k danému šťastnému ukončení dosažením řešení soustavy nedojde dříve. Jde pouze o technický předpoklad a z dalších částí textu bude obecný případ zřejmý.

Bude-li možné ukázat, že *všechny* směrové vektory generované v algoritmu HSCG jsou ortogonální ve smyslu energetického skalárního součinu, pak lze počáteční a n-tou chybu zapsat pomocí vytvořené a-ortogonální báze celého prostoru

$$x - x_0 = \sum_{\ell=0}^{N-1} \alpha_\ell p_\ell = \sum_{\ell=0}^{n-1} \alpha_\ell p_\ell + x - x_n \,, \quad x - x_n = \sum_{\ell=n}^{N-1} \alpha_\ell p_\ell \,.$$

Z ortogonality ve smyslu energetického skalárního součinu je zřejmé, že (ekvivalentní) podmínky optimality (9) a (10) jsou postupně splněny na všech krylovovských podprostorech $\mathcal{K}_n(A, r_0)$, $n = 1, 2, \ldots, N$. Protože $\mathcal{K}_N(A, r_0)$ je roven celému prostoru, musí být $x_N = x$. Navíc je z algoritmu HSCG zřejmé, že jednotlivé krylovovské podprostory jsou generovány nejen směrovými vektory $p_0, p_1, \ldots, p_{N-1}$, ale rovněž rezidui $r_0, r_1, \ldots, r_{N-1}$. Z galerkinovské ortogonality pak ihned plyne, že rezidua musí být navzájem kolmá vzhledem ke standardnímu skalárnímu součinu, $r_j^* r_i = 0$, $i \neq j$.

Vše se tedy redukuje na otázku, zda je možné v algoritmu HSCG zajistit volbou parametrů β_n, $n = 1, 2, \ldots, N-1$, v jednotlivých iteracích ortogonalitu *všech* směrových vektorů, jinými slovy, *zda z lokální ortogonality mezi následnými směrovými vektory vyplývá globální ortogonalita všech směrových vektorů mezi sebou navzájem.* Pro $n = 1$ je vše triviální, neboť směrové vektory jsou pouze dva. Pro $n \geqq 2$ se to však zdá být nemožné, neboť při předepsaném výpočtu nového směrového vektoru dle vztahu $p_n = r_n + \beta_n p_{n-1}$ máme k dispozici pouze jediný volný parametr β_n a máme splnit n podmínek ortogonality vůči všem předchozím směrovým vektorům $p_0, p_1, \ldots, p_{n-1}$. Matematickou indukcí však lze dokázat, že v algoritmu HSCG jsou skutečně všechny směrové vektory ortogonální vzhledem k energetickému skalárnímu součinu, se všemi důsledky popsanými výše.

Je na místě se nad překvapivým (téměř kouzelnickým) vyústěním předcházejících úvah zamyslet. Dotýkáme se zde velmi podstatné skutečnosti. Důkaz indukcí je v pořádku, to však neznamená, že nám umožňuje *porozumět* metodě konjugovaných gradientů reprezentované zde algoritmem HSCG. *Být schopen dokázat matematická tvrzení není zdaleka totéž co porozumět předmětu, ke kterému se vztahují.*

Matematika nabízí, ale také vyžaduje mnohem více, než se zdá z jejího často zdůrazňovaného formalistického modelu definice–hypotéza–věta–důkaz. Krásným čtením na dané téma jsou eseje C. Lanczose *Why Mathematics* [28] a W. Thurstona *On proof and progress in Mathematics* [42].

V našem příběhu jsme ocitli na myšlenkovém rozcestí:

- Můžeme veškeré naše další snažení omezit jen na popisné úvahy, jejichž středem je posloupnost operací definovaná algoritmem HSCG popřípadě jinou algoritmickou realizací metody konjugovaných gradientů. Můžeme algoritmus HSCG také zkoumat pomocí modelů výpočtů blízkých pojetí teoretické informatiky.
- Nebo se můžeme snažit o pochopení matematické podstaty metody, jíž je algoritmus HSCG jednou z možných algoritmických realizací. Pak je vhodné začít s otázkou, proč lokální ortogonalita nového směrového vektoru pouze proti předcházejícímu vektoru v algoritmu HSCG zaručí, že všechny směrové vektory budou navzájem ortogonální. Důkaz matematickou indukcí na danou otázku neodpovídá. Něco velmi podstatného zůstává skryto.

První směr je charakteristický omezením na algoritmické uvažování. Starými básnickými slovy z první kapitoly knihy Geneze vyjádřeno, druhý směr vede do země bohaté rostlinami a tvory různého druhu.

4. Redukce na algoritmické uvažování

Vraťme se na chvíli k polynomiálnímu vyjádření (7). S jeho použitím můžeme zapsat rezidum metody konjugovaných gradientů v n-té iteraci pomocí aproximačního polynomu $\varphi_n^{\mathrm{CG}}(\lambda)$ stupně n

$$r_n = b - Ax_n = r_0 - A\phi_{n-1}(A)r_0 =: \varphi_n^{\mathrm{CG}}(A)r_0, \tag{11}$$

kde

$$\varphi_n^{\mathrm{CG}}(\lambda) = 1 - \lambda\,\phi_{n-1}(\lambda), \quad \varphi_n^{\mathrm{CG}}(0) = 1\,.$$

Polynom $\varphi_n^{\mathrm{CG}}(\lambda)$ závisí díky optimalitě *nelineárně jak na matici A, tak na pravé straně soustavy b*.

Vzhledem k tomu, že metoda konjugovaných gradientů minimalizuje na krylovovských podprostorech energetickou normu chyby, můžeme její hodnotu odhadnout shora s použitím libovolného polynomu $\varphi(\lambda)$ daného stupně, který stejně jako optimální polynom $\varphi_n^{\mathrm{CG}}(\lambda)$ nabývá v nule hodnoty jedna (množinu polynomů stupně n s touto vlastností označíme $\mathcal{P}_n(0)$)

$$\begin{aligned} \|x - x_n\|_a &= \min_{z \in x_0 + \mathcal{K}_n(A, r_0)} \|x - z\|_a = \|\varphi_n^{\mathrm{CG}}(A)(x - x_0)\|_a \qquad (12)\\ &= \min_{\varphi \in \mathcal{P}_n(0)} \|\varphi(A)(x - x_0)\|_a \\ &\leqq \|x - x_0\|_a \min_{\varphi \in \mathcal{P}_n(0)} \|\varphi(A)\|. \qquad (13) \end{aligned}$$

Použití vhodně upraveného Čebyševova polynomu prvního druhu pak vede v několika technických krocích k odhadu relativní hodnoty energetické normy chyby v n-té iteraci

$$\frac{\|x - x_n\|_a}{\|x - x_0\|_a} \leqq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^n, \quad \kappa(A) = \frac{\lambda_N}{\lambda_1}. \tag{14}$$

Podíl největšího a nejmenšího vlastního čísla $\kappa(A) = \lambda_N/\lambda_1$ se nazývá podmíněností SPD matice A. Odhad (14) je v literatuře velmi často spojován s popisem chování metody konjugovaných gradientů, aniž by se autoři odpovídajícím způsobem vypořádali se zjevným zásadním rozporem. Chování metody konjugovaných gradientů závisí nelineárně na vstupních datech A, b, x_0 a vzhledem k optimalitě individuálních iterací na podprostorech rostoucí dimenze lze obecně očekávat zrychlování poklesu chyby (byť k němu nemusí dojít v každé iteraci a může být velmi podstatné nebo jen mírné, v závislosti na řešeném problému). Pokles horního odhadu je však dán stejným multiplikativním faktorem

$$\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}$$

v každé iteraci. Říkáme, že odhad je *lineární*. Navíc odhad nezávisí na pravé straně úlohy ani na počáteční aproximaci řešení, musí tedy platit pro všechny hodnoty b, x_0. V případě numerického řešení okrajových úloh popsaných parciálními diferenciálními rovnicemi diskretizovanými metodou konečných prvků to například znamená, že odhad musí platit pro libovolné vnější síly a okrajové podmínky, což pro použití při řešení konkrétní úlohy vůbec nemusí být výhoda. U matice A odhad závisí pouze na podílu jejího největšího a nejmenšího vlastního čísla.

Navíc nelze pominout následující skutečnost. Chování metody konjugovaných gradientů v praktických výpočtech je ovlivněno (mnohdy velmi výrazně) šířením zaokrouhlovacích chyb (vrátíme se k tomu v sedmé kapitole). Odhad (14) ani jeho případné modifikace nijak neumožňují daný jev popsat. Je příkladem *a priori* přístupu omezeného na nejhorší možný scénář, který obecně nepopisuje konkrétní praktické chování (výjimky jen potvrzují pravidlo). Vše uvedené neznamená, že odhad (14) není užitečný, je-li používán s rozvahou jako to, co je, a ne to, co není, jak tomu v mnoha případech je. Slova klasika platí i zde.

Odhad (14) je v literatuře připisován mnoha autorům. V přímé souvislosti s metodou konjugovaných gradientů se pravděpodobně poprvé objevil v Danielově článku [10], větě 1.2.2, ale to není příliš podstatné, neboť například v Rutishauserově kapitole o teorii gradientních metod, která je součástí [12], se objevil mnohem dříve v souvislosti s metodou založenou na Čebyševových polynomech. Platnost odhadu i pro metodu konjugovaných gradientů je z kontextu zřejmá. Podrobný výklad souvislostí lze nalézt například v [29], sekcích 5.5.2, 5.5.3 a 6.5.2. K odhadu (14) a k omezením jeho použití se vrátíme v následující části článku.

Standardní výklad metody konjugovaných gradientů se v učebnicích i některých monografiích soustředí na odvození algoritmu HSCG a případně také odhadu (14). Na daném algoritmickém základě pak bývají prezentovány různé úvahy a srovnávání, do kterých bývají případně zahrnuty experimenty na jednoduchých modelových úlohách, na jejichž základě jsou činěny dalekosáhlé závěry.[6] Výsledkem je velmi zkreslený a povrchní pohled, který je v různých variacích tak široce rozšířen, že bývá považován za správný a nezpochybnitelný. Je přitom v rozporu s bohatstvím myšlenek původních prací Hestenese, Stiefela a Lanczose z let 1952–1953 citovaných výše.

[6] Jsou důsledkem dvou zásadních metodologických chyb, pro matematiku velmi překvapivých. Nelineární jev je ztotožňován s lineárním popisem a na obecné chování je v pozitivním smyslu usuzováno na základě velmi omezeného výběru velmi jednoduchých a nevhodných modelových příkladů.

Z původních článků (a u některých z nich dokonce i z jejich názvů a abstraktů) je zřejmé, že hlavní výpočetní vlastností metody konjugovaných gradientů je iterační konstrukce posloupnosti aproximací umožňující zastavení kdykoliv je splněno vhodně zvolené kritérium. Dvacet let paradoxně zatěžoval pohled na metodu konjugovaných gradientů elegantní matematický fakt dosažení přesného řešení v konečném počtu iterací. Reid v přednášce v roce 1971 a v návazném článku [37] zdůraznil výhodnost iteračního pohledu (doporučujeme pozornosti záznam otázek vynikajících osobností numerické matematiky a Reidových odpovědí přiložený k textu článku), a tím vrátil metodu konjugovaných gradientů do pozornosti matematiků. Tehdejší neujasněnost je snadno vysvětlitelná historickými okolnostmi. Řešené úlohy souvisely s několika málo základními problémy matematické fyziky, pro které ale nebylo použití metody konjugovaných gradientů nikterak výhodné. Matice soustav byly navíc velmi malé; viz například shrnutí v [29], sekcích 2.5.7 a 5.2.1. Obtížně vysvětlitelné je však přetrvávání různých tabulkových srovnávání s použitím nevhodných triviálních modelových problémů a počtu iterací určovaných vyumělkovanými kritérii pomocí odhadu (14) a jeho případných modifikací do dneška. Je příkladem přenášení zkamenělých pohledů relevantních v počítačovém dávnověku do kontextu, ve kterém zcela ztratily smysl.[7]

Kapitolou samou pro sebe jsou úvahy o tzv. teorii složitosti v souvislosti s metodou konjugovaných gradientů mylně reprezentovanou odhadem (14); viz příklady referencí a stručná diskuse v [8], sekci 3a. Autory jsou matematici a informatici věhlasných jmen. Příslušné publikace přesto o metodě konjugovaných gradientů téměř nic rozumného neříkají a některé z nich obsahují věcně zkreslené až zcela nesmyslné odkazy na původní prameny.

V následujících částech se vydáme do bohatých krajů s hlubokými lesy a vodami.

5. Polynomiální aproximační problém a čebyševovský odhad

Podíváme se blíže na myšlenkový posun, který vedl od polynomiálního vyjádření velikosti chyby v energetické normě (12) k lineárnímu odhadu (14). K tomu nám pomůže změna souřadnic v námi uvažovaném reálném eukleidovském prostoru dimenze N. Namísto standardní eukleidovské báze $e_1, e_2, \ldots, e_N$, kde e_j je sloupcový vektor s jedinou nenulovou složkou rovnou jedné na j-té pozici, $j = 1, \ldots, N$, použijeme ortonormální bázi $q_1, q_2, \ldots, q_N$ sestavenou z normalizovaných vlastních vektorů matice A,

$$Aq_j = \lambda_j q_j, \quad j = 1, \ldots, N.$$

Maticově zapsáno,

$$AQ = Q\Lambda, \quad A = Q\Lambda Q^*, \quad Q^*Q = QQ^* = I, \tag{15}$$

kde Q je ortonormální matice se sloupci $q_1, q_2, \ldots, q_N$, $\Lambda = Q^*AQ$ je diagonální matice sestavená z vlastních čísel matice A a I je jednotková matice. Vztahy (15)

[7] Doporučujeme čtenáři vyhýbat se literatuře týkající se metody konjugovaných gradientů publikované od osmdesátých let, která podobná srovnávání „výpočetní náročnosti“ (a případně i velikosti paměti) obsahuje. Týká se to i hojně čtených a citovaných textů přístupných na internetu, slibujících porozumění metodě konjugovaných gradientů bez „agonizující bolesti“ a námahy. Sliby – chyby. Vyznat se ve složitých věcech a situacích bez bolesti a námahy hledání je možné jen ve špatných pohádkách.

vyjadřují spektrální rozklad symetrické matice, v případě SPD matice navíc víme, že všechna její vlastní čísla jsou kladná. Dosadíme-li spektrální rozklad do polynomiálního vyjádření reziduí (11) a použijeme-li galerkinovskou ortogonalitu (10) ekvivalentní ortogonalitě jednotlivých reziduí, dostaneme pro $i \neq j$

$$\begin{aligned} 0 = r_i^* r_j &= (\varphi_i^{\mathrm{CG}}(A) r_0)^* (\varphi_j^{\mathrm{CG}}(A) r_0) = (\varphi_i^{\mathrm{CG}}(\Lambda) Q^* r_0)^* (\varphi_j^{\mathrm{CG}}(\Lambda) Q^* r_0) \\ &= \|r_0\|^2 \, (\varphi_i^{\mathrm{CG}}(\Lambda) Q^* v_1)^* (\varphi_j^{\mathrm{CG}}(\Lambda) Q^* v_1) \\ &= \|r_0\|^2 \sum_{\ell=1}^{N} \omega_\ell \, \varphi_i^{\mathrm{CG}}(\lambda_\ell) \, \varphi_j^{\mathrm{CG}}(\lambda_\ell), \end{aligned} \tag{16}$$

kde $v_1 = r_0/\|r_0\|$ je normalizované počáteční reziduum.[8] Dospěli jsme ke klíčovému poznatku. Polynomy generující jednotlivá *rezidua jsou ortogonální vzhledem ke skalárnímu součinu definovanému vlastními čísly matice A a kvadráty velikosti projekcí normalizovaného počátečního rezidua do směrů jednotlivých vlastních vektorů*

$$\omega_\ell = (q_\ell^* v_1)^2, \quad \ell = 1, \ldots, N. \tag{17}$$

Tím je dána odpověď na výše položenou otevřenou otázku, proč z lokální ortogonality dvou po sobě následujících směrových vektorů vyplývá, že *všechny* směrové vektory musí být navzájem ortogonální (v energetickém skalárním součinu) a všechna rezidua musí být navzájem ortogonální (vzhledem k eukleidovskému skalárnímu součinu). Jednoduchá konstrukce posloupnosti ortogonálních polynomů pomocí ortogonalizačního procesu známá více než století ihned ukazuje, že *ortogonální polynomy jsou vždy určeny tříčlennou rekurencí*. V algoritmu HSCG taková rekurence odpovídá dvěma provázaným dvoučlenným rekurencím pro rezidua a směrové vektory (pokud bychom algoritmus HSCG přeorganizovali vyjádřením směrových vektorů pomocí reziduí, dostali bychom jeho variantu s tříčlennou rekurencí). Proto *musí stačit* v algoritmu HSCG ortogonalizovat nový směrový vektor jen proti předchozímu. Globální ortogonalita je při přesném výpočtu samozřejmým důsledkem.

Naznačená souvislost s ortogonálními polynomy je pro porozumění metodě konjugovaných gradientů zásadní. Budeme se jí věnovat v následující části textu. V závěru této části se podíváme, co může říci o odhadu (14). Použijeme-li spektrální rozklad (15) v polynomiálním vyjádření velikosti chyby (12), dostaneme

$$\begin{aligned} \|x - x_n\|_a^2 &= \|\varphi_n^{\mathrm{CG}}(A)(x - x_0)\|_a^2 = (A^{-1} \varphi_n^{\mathrm{CG}}(A) r_0)^* (\varphi_n^{\mathrm{CG}}(A) r_0) \\ &= \|r_0\|^2 \sum_{\ell=1}^{N} \frac{\omega_\ell}{\lambda_\ell} \left(\varphi_n^{\mathrm{CG}}(\lambda_\ell)\right)^2 . \end{aligned} \tag{18}$$

Výsledná hodnota je tedy dána váženým součtem kvadrátů hodnot polynomu $\varphi_n^{\mathrm{CG}}(\lambda)$ v jednotlivých bodech spektra λ_ℓ, $\ell = 1, \ldots, N$. Označíme-li $\theta_1^{(n)}, \ldots, \theta_1^{(n)}$ jeho

[8] V polynomiální funkci matice $\varphi(A)$ jsou jednotlivé mocniny argumentu určeny maticovým násobením a výsledek je po následném vynásobení skalárními koeficienty dán součtem po jednotlivých stejnolehlých prvcích. S použitím ortogonálního spektrálního rozkladu tak například platí $A^k = (Q\Lambda Q^*)^k = Q\Lambda^k Q^*$, a pro diagonální matici $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ je $\varphi(\Lambda) = \mathrm{diag}(\varphi(\lambda_1), \ldots, \varphi(\lambda_N))$.

kořeny,[9]

$$\varphi_n^{\mathrm{CG}}(\lambda) = \frac{(\lambda - \theta_1^{(n)}) \ldots (\lambda - \theta_n^{(n)})}{(-1)^n\, \theta_1^{(n)} \ldots\ \theta_n^{(n)}}\ , \tag{19}$$

pak vlastnost minimalizace energetické normy chyby vyžaduje velmi těsnou (nikoliv však jednoduše popsatelnou) vazbu mezi těmito kořeny a *všemi vlastními čísly matice A*; viz například [29], kapitolu 3. Na druhé straně, použijeme-li spektrální rozklad v odhadu (13) a následně nahradíme minimalizační problém na diskrétních bodech spektra čebyševovským minimalizačním problémem na intervalu definovaném extremálními vlastními čísly λ_1 a λ_N, dostaneme

$$\frac{\|x - x_n\|_a}{\|x - x_0\|_a} \leqq \min_{p \in \mathcal{P}_n(0)} \max_{1 \leqq j \leqq N} |p(\lambda_j)| \leqq \min_{p \in \mathcal{P}_n(0)} \max_{\lambda \in [\lambda_1, \lambda_N]} |p(\lambda)|.$$

Čebyševovský odhad (14) je pak dán zjednodušením (s dalším mírným nadhodnocením, zde nepodstaným) čebyševovského mimimalizačního problému; viz například [12], kapitolu II, a podrobný výklad spolu s dalšími odkazy v [29], sekci 5.5.2. Kořeny (upraveného) Čebyševova polynomu použitého k získání (14) nemají naprosto žádný vztah k vlastním číslům matice A v otevřeném intervalu (λ_1, λ_N). Zásadní rozdíl mezi minimalizačním problémem na *diskrétních bodech spektra* a čebyševovským minimalizačním problémem *na intervalu* byl zdůrazněn ve vztahu k metodě konjugovaných gradientů již Lanczosem v roce 1952; viz [26], sekci 5, zejména stranu 46. Daniel přesně vymezil ve výše citovaném článku [10] smysl odhadu (14). Přesto povrchně pragmatický a *zdánlivě* vše řešící přístup, který staví odhad na nesprávné místo a který se míjí s podstatou metody konjugovaných gradientů, naprosto ovládl pole.

Mohlo by se zdát, že jde jen o malicherné trvání na matematické důkladnosti a přesnosti, která zde není potřeba nebo snad může být i překážkou k používání jednoduchého a *prakticky* dostačujícího postupu. Nahrazení skutečného chování metody konjugovaných gradientů čebyševovským odhadem bývá mnohými skutečně považováno za univerzálně použitelný a zcela vyhovující nástroj. Mnozí jejich následovníci si už ani neuvědomí, k jak závažnému pochybení zde dochází. *Metodologická chyba, která pomíjí kontext a používá matematické tvrzení mimo oblast jeho platnosti vymezenou předpoklady vede nevyhnutelně ke zmatku.*

Popisuje-li čebyševovský minimalizační problém a z něj odvozený odhad (14) v určitém *netriviálním* případě chování metody konjugovaných gradientů, znamená to, že rozložení vnitřní části spektra v otevřeném intervalu (λ_1, λ_N) nedovoluje podstatnou adaptivitu. Pak je na místě otázka, zda a proč by měla být metoda konjugovaných gradientů použita. Odpověď může být kladná a vysvětlení přesvědčivé. Ale otázky musí být položeny a zodpovězeny. Povrchnost je nebezpečná tím, že si neklade otázky a uzavírá cestu. K odhadu (14) se vrátíme ještě jednou po části věnované důsledkům šíření zaokrouhovacích chyb (netrpělivého čtenáře zatím odkazujeme na související článek [14]). Je dobré připomenout slova přisuzovaná v Zeidlerově oxfordském průvodci matematikou [46] Einsteinovi (ponecháme citát v anglickém jazyce):

Everything should be made as simple as possible, but not simpler.

[9] Kořeny jsou navzájem různé a leží v intervalu (λ_1, λ_N).

6. Problém momentů a redukce modelu

Naše cesta začala v kontextu počítačového řešení soustav lineárních algebraických rovnic v polovině dvacátého století. Souvislost s ortogonálními polynomy popsaná v předcházející části a vyjádření velikosti chyby pomocí jejich hodnot (18) probouzí zvídavost, zda neexistuje nějaká její dřívější část, která nám je zatím skryta. Posuneme se proto o další století (a v jistém smyslu o mnoho dalších století) zpět. Musíme být velmi struční a mnoho zajímavých souvislostí vynechat – podrobnější výklad by zabral desítky stran (uvedeme odkazy na publikované texty, kde jej lze nalézt spolu s mnoha odkazy a historickými poznámkami).

V roce 1894 byla publikována monumentální Stieltjesova práce věnovaná řetězovým zlomkům [40], která v určitém smyslu dovršila předcházející výsledky mnoha matematiků včetně Eulera, Jacobiho, Darbouxe, Heineho, Christoffela, Čebyševa a Markova; viz například [29], sekce 3.3.5 a 3.3.6 a reference tam uvedené. Řetězový zlomek aproximující nejprve čísla a mnohem později funkce provází matematiku prakticky od jejích počátků, viz [6]. Euler ukázal, jak lze analytickou funkci (vyjádřenou nekonečnou řadou) zapsat pomocí řetězového zlomku. Stieltjes otázku obrátil. Ukázal, za jakých podmínek na (reálné) koeficienty $\gamma_1, \gamma_2, \ldots$ a $\delta_2, \delta_3, \ldots$ řetězový zlomek

$$\mathcal{F}_n(\lambda) \equiv \cfrac{1}{\lambda - \gamma_1 - \cfrac{\delta_2^2}{\lambda - \gamma_2 - \cfrac{\delta_3^2}{\lambda - \gamma_3 - \ldots \cfrac{\ddots}{\lambda - \gamma_{n-1} - \cfrac{\delta_n^2}{\lambda - \gamma_n}}}}} = \frac{\mathcal{R}_n(\lambda)}{\mathcal{P}_n(\lambda)} \tag{20}$$

konverguje pro $n \to \infty$ k analytické funkci proměnné λ. Všimněme si, že n-tý konvergent $\mathcal{F}_n(\lambda)$ není nic jiného než racionální lomená funkce, kde jmenovatel $\mathcal{P}_n(\lambda)$ je polynom stupně n a čitatel $\mathcal{R}_n(\lambda)$ je polynom stupně $n-1$. I když to není zatím zřejmé, již jsme se s nimi setkali, což vysvětlíme níže.

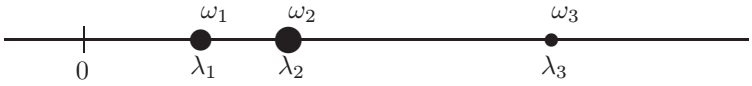
Jako jeden z kroků svého postupu Stieltjes formuloval a vyřešil následující *problém momentů*. Uvažujme nekonečnou posloupnost kladných čísel $m_0, m_1, m_2, \ldots$ Úlohou je nalézt nutné a postačující podmínky pro existenci (nezáporné omezené neklesající) *distribuční funkce* $\omega(\lambda)$ definované na itervalu $[0, +\infty)$ tak, aby pro Riemannův–Stieltjesův integrál monomů platilo

$$\int_0^\infty \lambda^\ell \, d\omega(\lambda) = m_\ell \,, \qquad \ell = 0, 1, 2, \ldots, \tag{21}$$

a určit $\omega(\lambda)$. Bez újmy na obecnosti $\omega(0) = 0$, $\lim\limits_{\lambda\to\infty} \omega(\lambda) = 1$ a $m_0 = 1$. Stieltjes potřeboval zobecnit klasický Riemannův integrál (odtud název Riemannův–Stieltjesův). Zobecněný integrál byl mimo jiné později použit Hilbertem a von Neumannem k integrální reprezentaci samoadjungovaných operátorů na Hilbertově prostoru. Reprezentace operátorů na Hilbertově prostoru pomocí Riemannova–Stieltjesova integrálu

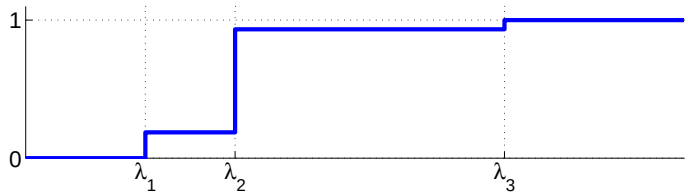
je užitečná nejen pro formulaci matematických základů kvantové mechaniky; viz například [29], závěr sekce 3.4.3 a sekci 3.5, zejména poznámku 3.5.1 a reference tam uvedené. Pokud řešení Stieltjesem formulovaného problému momentů existuje, nazývají se čísla $m_0, m_1, m_2, \ldots$ *momenty distribuční funkce* $\omega(\lambda)$. Všimněme si, že použitá terminologie je obvyklá v teorii pravděpodobnosti a v matematické statistice, což není jen shoda okolností, ale příčinná souvislost. Dané termíny však mají svůj původ v klasické mechanice, jak uvidíme níže. My se zde nemůžeme zabývat obecným řešením Stieltjesova problému momentů. Pro souvislost s metodou konjugovaných gradientů postačí jeho zjednodušená varianta, která je podrobně vyložena v [36].

V předcházející části textu jsme ukázali ortogonalitu polynomů $\varphi_n^{\mathrm{CG}}(\lambda)$ (definujících rezidua v metodě konjugovaných gradientů) vzhledem ke skalárnímu součinu určenému vlastními čísly matice A a vahami ω_ℓ; viz (16)–(17).[10] Můžeme si představit mechanickou analogii, ve které jsou hmotné body o hmotnosti ω_ℓ *distribuovány* na polopřímce představující kladnou reálnou poloosu do bodů odpovídajících vlastním číslům (viz ilustraci pro $N = 3$)



a definovat po částech konstantní distribuční funkci

$$\omega(\lambda) = \begin{cases} 0 & \text{pro } 0 \leqq \lambda < \lambda_1, \\ \displaystyle\sum_{j=1}^{i} \omega_j & \text{pro } \lambda_i \leqq \lambda < \lambda_{i+1},\ \ i = 1, \ldots, N-1, \\ \displaystyle\sum_{j=1}^{N} \omega_j = 1 & \text{pro } \lambda_N \leqq \lambda \end{cases} \tag{22}$$



(opět s ilustrací pro $N = 3$). Její první tři momenty, které lze zapsat vzhledem ke konečnému počtu bodů vzrůstu funkce $\omega(\lambda)$ prostou sumou reprezentující Riemannův–Stieltjesův integrál, nejsou nic jiného, než první tři mechanické momenty (celková distribuovaná hmotnost, poloha těžiště a moment setrvačnosti vzhledem k nule). Stieltjes z analogie s distribucí hmoty na přímce vychází (v obecném případě není ovšem distribuční funkce po částech konstantní) a používá termín *zobecněné mechanické momenty.* Ve zjednodušeném Stieltjesově problému spojeném s řešením soustavy (1) se SPD maticí můžeme momenty s použitím spektrálního rozkladu zapsat vztahem

$$m_\ell = v_1^* A^\ell v_1, \quad \ell = 0, 1, 2, \ldots, 2N-1. \tag{23}$$

[10] Pro zjednodušení značení předpokládejme, že všechna vlastní čísla jsou navzájem různá a všechny váhy jsou nenulové. To odpovídá dřívějšímu technickému předpokladu o dosažení řešení soustavy metodou konjugovaných gradientů (při přesném výpočtu) právě v N iteracích.

Vzhledem ke konečném počtu parametrů distribuční funkce (22) stačí v dalším $2N$ momentů. Pro dané $n < N$ pak zjednodušený Stieltjesův problém momentů spočívá v určení $2n$ kladných čísel

$$\omega_j^{(n)},\ j = 1, \ldots, n,\ \sum_{j=1}^{n} \omega_j^{(n)} = 1, \quad \text{a} \quad 0 < \lambda_1^{(n)} < \lambda_2^{(n)} < \ldots < \lambda_n^{(n)}$$

tak, aby

$$\sum_{j=1}^{n} \omega_j^{(n)} \left(\lambda_j^{(n)}\right)^{\ell} = m_\ell, \quad \ell = 0, 1, \ldots, 2n-1. \tag{24}$$

Jinými slovy, úkolem je nalézt po částech konstantní distribuční funkci $\omega^{(n)}(\lambda)$ s n body vzrůstu $\lambda_j^{(n)}$ a vahami $\omega_j^{(n)}$, $j = 1, \ldots, n$, tak, aby jejích prvních $2n$ momentů bylo totožných s momenty distribuční funkce $\omega(\lambda)$. Soustava rovnic (24) je pozoruhodně nelineární, neboť všechny hodnoty na levé straně jsou neznámé. Následující tvrzení možná překvapí.

Metoda konjugovaných gradientů implicitně řeší v každé iteraci odpovídající zjednodušený Stieltjesův problém momentů. Opačně, známe-li pro nějaké n řešení zjednodušeného Stieltjesova problému momentů, je tím určena n-tá aproximace řešení soustavy (1) *metodou konjugovaných gradientů.*

Abychom podrobněji vysvětlili uvedená tvrzení a ukázali jejich platnost, použijeme matematicky ekvivalentní formulaci metody konjugovaných gradientů a operátorovou (maticovou) formulaci problému momentů. Nejprve zavedeme bázi krylovovských prostorů tvořenou ortonormálními vektory $v_1, v_2, \ldots, v_n$, které jsou v následujícím vztahu s normalizovanými rezidui z algoritmu HSCG,

$$v_j = (-1)^{j-1}\ \frac{r_{j-1}}{\|r_{j-1}\|}.$$

V kontextu aproximace vlastních čísel je tato *Lanczosova báze* počítána po sloupcích pomocí Lanczosova procesu, který v maticovém tvaru můžeme zapsat[11]

$$AV_n = V_n T_n + \delta_{n+1} v_{n+1} e_n^*, \quad n = 1, \ldots, N, \tag{25}$$

kde $V_n = [v_1, \ldots, v_n]$ je matice s N řádky a n sloupci tvořená prvními n Lanczosovými vektory, T_n je třídiagonální matice ortonormalizačních koeficientů uložených po sloupcích, známá pod jménem *Jacobiho matice*,

$$T_n = \begin{pmatrix} \gamma_1 & \delta_2 & & & \\ \delta_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \delta_n \\ & & & \delta_n & \gamma_n \end{pmatrix} = V_n^* A\ V_n \tag{26}$$

[11] Výraz $v_{n+1} e_n^*$ představuje podle pravidel maticového násobení matici s N řádky a n sloupci, ve které je prvních $n-1$ sloupců nulových a poslední sloupec je tvořen násobkem vektoru v_{n+1}.

a e_n je n-tý vektor standardní eukleidovské báze. Pro $n = N$ je $\delta_{N+1} = 0$ a poslední člen Lanczosovy rekurence je nulový. Vyjádříme-li nyní n-tou aproximaci řešení v metodě konjugovaných gradientů pomocí Lanczosovy báze

$$x_n = x_0 + V_n t_n, \tag{27}$$

pak z galerkinovské ortogonality $V_n^* r_n = 0$ rezidua $r_n = b - Ax_n = r_0 - AV_n t_n$ ihned plyne

$$T_n t_n = \|r_0\| e_1. \tag{28}$$

Řešení rovnice (28) spolu se substitucí (27) můžeme nazvat formulací metody konjugovaných gradientů pomocí Lanczosova procesu (zkráceně Lanczosovou formulací).

K Jacobiho matici T_n a k metodě konjugovaných gradientů se můžeme dostat přímo pomocí operátorové formulace problému momentů; viz krásně napsanou a téměř neznámou Vorobjevovu monografii [43], která rozvinula myšlenku zmíněnou dříve Ljusternikem [30] . Hledejme operátor na krylovovském prostoru $\mathcal{K}_n(A, r_0)$ reprezentovaný maticí A_n tak, aby platilo

$$\begin{aligned} A\, r_0 &= A_n r_0, \\ A^2 r_0 &= A_n^2 r_0, \\ &\vdots \\ A^{n-1} r_0 &= A_n^{n-1} r_0, \\ E_n A^n r_0 &= A_n^n r_0, \end{aligned}$$

kde E_n je ortogonální projektor na $\mathcal{K}_n(A,\, r_0)$; viz (8). Můžeme jej zapsat pomocí matice vytvořené z Lanczosovy báze ve tvaru $E_n = V_n V_n^*$. Operátor

$$A_n = E_n A E_n = V_n V_n^* A V_n V_n^* = V_n T_n V_n^*$$

na $\mathcal{K}_n(A,\, r_0)$ za námi použitých předpokladů existuje a je jednoznačně určen (i s triviálním rozšířením na celý $\mathcal{R}^N$). V Lanczosově bázi má operátor A_n matici T_n.

Uvedený postup lze interpretovat jako *redukci (aproximaci) modelu* reprezentovaného původní soustavou (1) s N neznámými na soustavu (28) o n neznámých. Termín *redukce modelu* je známý zejména z mnoha inženýrských aplikací, například z popisu chování dynamických systémů, zpracování signálů a obrazu. Vorobjev formuluje metodu momentů v nekonečnědimenzionálních Hilbertových prostorech a využívá souvislostí se Stieltjesovým problémem momentů. Odkazuje mimo jiné na práce Hestenese, Stiefela a Lanczose a na operátorovou formulaci a chování metody konjugovaných gradientů v nekonečné dimenzi [24].

Jacobiho matice T_n je úhelným kamenem, který spojuje maticovou (operátorovou) formulaci s formulací prostřednictvím ortogonálních polynomů. V Lanzcosově procesu (25) je formou zápisu ortonormalizačních koeficientů. Má však mnohem podstatnější význam. Její charakteristický polynom je roven čitateli ve vyjádření polynomu $\varphi_n^{\mathrm{CG}}(\lambda)$ v (19) a stejně tak jmenovateli $\mathcal{P}_n(\lambda)$ ve vyjádření n-tého konvergentu $\mathcal{F}_n(\lambda)$ řetězového zlomku (20) jako racionální lomené funkce (čitatel $\mathcal{R}_n(\lambda)$ je určen stejnou tříčlennou rekurencí posunutou o jeden krok). Její vlastní čísla jsou zároveň

body vzrůstu $\lambda_j^{(n)}$ distribuční funkce $\omega^{(n)}$, kvadráty prvních elementů odpovídajících normalizovaných vlastních vektorů jsou rovny vahám $\omega_j^{(n)}$; viz (24). Platí tedy $\theta_j^{(n)} = \lambda_j^{(n)}$, $j = 1, \ldots, n$. Jacobiho matice ukazuje, jak nádhernou hloubku může mít zdánlivě velmi jednoduchý matematický objekt. Podrobný výklad s mnoha dalšími souvislostmi lze nalézt v [29], kapitole 3.

Řešení problému momentů souvisí přirozeně s Gaussovou (také v daném kontextu nazývanou Gaussovou–Christoffelovou) kvadraturou Riemannova–Stieltjesova integrálu. Rovná-li se prvních $2n$ momentů distribučních funkcí $\omega(\lambda)$ a $\omega^{(n)}(\lambda)$, musí se rovnat i odpovídající Riemannovy–Stieltjesovy integrály polynomů stupně nejvýše $2n-1$, přičemž integrál odpovídajicí distribuční funkci $\omega^{(n)}(\lambda)$ s n body vzrůstu je sumou používající jen n funkčních hodnot. Zapíšeme-li Gaussovu kvadraturu funkce $f(\lambda)$ rovnicí

$$\int_0^\infty f(\lambda)\, d\omega(\lambda) = \sum_{j=1}^{n} \omega_j^{(n)} f(\lambda_j^{(n)}) + R_n(f)\ ,$$

kde $R_n(f)$ vyjadřuje chybu kvadratury pro danou funkci, lze pro volbu $f(\lambda) = 1/\lambda$ jednotlivé členy vyjádřit následujícím způsobem (viz [29], sekci 3.5)

$$\frac{\|x - x_0\|_a^2}{\|r_0\|^2} = \sum_{j=1}^{n} \omega_j^{(n)} \frac{1}{\lambda_j^{(n)}} + \frac{\|x - x_n\|_a^2}{\|r_0\|^2}\ . \tag{29}$$

Metoda konjugovaných gradientů tedy nejen implicitně určuje uzly a váhy Gaussovy kvadratury, ale její energetická norma chyby je totožná (až na triviální násobek) s odpovídající chybou Gaussovy kvadratury pro funkci $f(\lambda) = 1/\lambda$. Daná vlastnost může být využita pro odhad velikosti chyby v metodě konjugovaných gradientů; viz například [41]. Souvislostí s ortogonálními polynomy, řetězovými zlomky a Gaussovou kvadraturou jsou jasně formulovány již v původních článcích Lanczose, Hestenese a Stiefela z let 1950–1953 citovaných výše. Pokud by vedle častého citování byly také občas čteny, mnohé pozdější texty by byly dávno právem zapomenuty nebo by nebyly vůbec napsány.

V předešlých částech jsme viděli různé stránky a důležité souvislosti metody konjugovaných gradientů. Mnohokráte jsme přitom použili galerkinovskou ortogonalitu, vzájemnou ortogonalitu různých vektorů, jim odpovídajících polynomů a projekcí. Při praktických výpočtech může však být ortogonalita překvapivě rychle ztracena. Důsledky ztráty ortogonality byly pro celou komunitu numerických matematiků po několik desetiletí neřešitelnou výzvou.

7. Jsou hluboké souvislosti a jejich matematická elegance v praktických výpočtech ztraceny?

Důsledky ztráty ortogonality byly diskutovány již v původních článcích [25], [26], [23] s řadou nosných myšlenek, které byly plně rozvinuty později. Tehdejší pohled však vycházel ze snahy upravovat praktický výpočet pokud možno co největším zachováním míry ortogonality mezi rezidui a směrovými vektory, což obvykle vedlo k řádovému

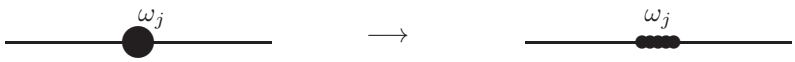
vzrůstu výpočetní náročnosti. Rigorózní analýza mechanismu ztráty ortogonality nebyla považována za proveditelnou.

Zásadní krok v daném směru učinil ve své dizertaci z roku 1971 věnované Lanczosově metodě C. C. Paige. Problémem není hromadění zaokrouhlovacích chyb (ve smyslu jejich sčítání), ale *jejich zesílení příslušným rekurentním výpočtem*[12], analogicky možnému šíření numerické chyby v explicitních diferenčních formulích. Rekurentní výpočet je určen algoritmem realizujícím danou metodu, proto jsou zesílení zaokrouhlovacích chyb a jeho důsledky rigorózně deterministicky analyzovatelné. Pochopení matematického řádu určujícího zesílení zaokrouhlovacích chyb vedlo Paige k následujícímu objevu včetně jeho nádherného rigorózního matematického důkazu:

K významné ztrátě ortogonality může dojít pouze ve směrech odpovídajících vlastním vektorům matice A, jejichž vlastní čísla jsou s velkou přesností numericky aproximována vlastními čísly Jacobiho matic T_n numericky generovaných Lanczosovým procesem.

Důsledkem je fakt, který dobře ilustruje krásu numerické matematiky neredukované na pouhý soubor algoritmických návodů. Hodnoty prvků matice T_n mohou být při numerickém výpočtu od určitého iteračního kroku vzdáleny mnoho řádů od hodnot, které bychom dostali přesným výpočtem. *Přesto je možné určit vlastní čísla matice A pomocí vlastních čísel matic T_n s přesností úměrnou strojové přesnosti počítače. Danou přesnost lze při výpočtu zaručit.* Není nutné (a v praktických výpočtech ani možné) opravovat mezivýsledky, tj. vypočtené prvky matic T_n. Stačí porozumět matematickému řádu, kterému výpočet a tím i šíření zaokrouhlovacích chyb podléhá. To ovšem může být (a zde je) velmi těžké.[13]

Neméně výjimečným je navazující článek [17], ve kterém Greenbaum spojila analýzu Lanczosovy metody s analýzou metody konjugovaných gradientů. Nejprve upravila distribuční funkci $\omega(\lambda)$, viz (22), *nahrazením jednotlivých bodů vzrůstu (hmotných bodů) malými intervaly (těsnými shluky hmotných bodů) při zachování původních vah*



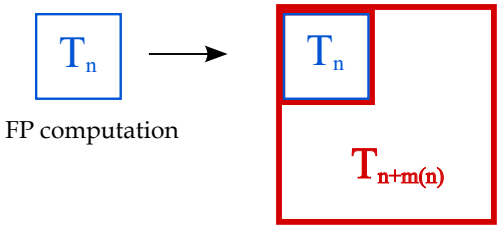
a odhadla velikost odpovídající malé změny v rekurencích Lanczosova procesu a algoritmu HSCG. Potom otázku obrátila a ukázala, jak lze pro Lanczosův proces a pro algoritmus HSCG při výpočtu zatíženém zaokrouhlovacími chybami (který lze interpretovat jako malé změny v zápisu odpovídajících rekurencí) zkonstruovat stejným způsobem upravené distribuční funkce. Nepřesný výpočet zatížený zaokrouhlovacími

[12] Myšlenka šíření a možného zesílení zaokrouhlovacích chyb je přítomna v [23], sekci 8, ale nebyla v prvních pracích jasně odlišena od akumulace. Vzniklá nejasnost přetrvává ve většině literatury dosud.

[13] Výsledky dizertace Paige, které jsou svojí originalitou a významem naprosto výjimečné, byly publikovány v průběhu deseti let(!) se závěrečným článkem [34] z roku 1980. Při dnes platných pravidlech by Paige se svojí prací vůbec nebyl připuštěn k obhajobě, natož aby získal významné místo na univerzitě. Formalizace posuzování kvality s tzv. „objektivizovanými kvantitativními kritérii“, ke kterým jsme v současné vědě dospěli a kterým se klaníme jako zlatému teleti, nejsou známkou pokroku. Jen maskují ztrátu schopnosti kvalitu poznat. Nebo uznat?

chybami pak lze interpretovat jako matematicky přesný výpočet pro upravená vstupní data odpovídající upraveným distribučním funkcím.

Její konstrukce uvažuje prvních n kroků Lanczosova procesu při výpočtu s konečnou přesností. Jeho výsledkem je Jacobiho matice T_n, kde její jednotlivé prvky mohou být mnoho řádů vzdáleny od odpovídajících přesných hodnot. Hledanou distribuční funkci Greenbaum sestrojila s využitím výsledků Paige velmi netriviálním prodloužením Jacobiho matice T_n na potenciálně mnohem větší Jacobiho matici $T_{n+m(n)}$



tak, aby její vlastní čísla byla všechna obsažena v malých intervalech kolem původních vlastních čísel matice A. Jacobiho matice $T_{n+m(n)}$ pak určuje hledanou distribuční funkci, která ovšem závisí na iteračním kroku n. S malou nepřesností výsledné korespondence však můžeme definovat univerzální distribuční funkci nezávislou na n nahrazením individuálních bodů vzrůstu malými intervaly při zachování původních vah; viz [20] a detailní výklad mnoha souvisejících výsledků mnoha autorů v [32] a [29], sekcích 5.6 a 5.9. Opět vidíme, jak velmi myšlenkově i technicky náročná cesta vede nakonec k nádherně jednoduše formulovanému závěru:

K pochopení nepřesného výpočtu zatíženého zaokrouhlovacími chybami stačí uvažovat přesný výpočet s nahrazením původních vlastních čísel shluky velmi blízkých vlastních čísel při zachování velikosti skoků původní distribuční funkce.

Všimněme si, jak důležité byly pro pochopení šíření zaokrouhlovacích chyb ve výpočtech používaných ve 20. a 21. století souvislosti s mnohem dřívějšími myšlenkami spojenými s ortogonálními polynomy, řetězovými zlomky a momenty.

8. Zmatek nad zmatek

Ztráta ortogonality a rovněž lineární nezávislosti generovaných reziduí a směrových vektorů vlivem zaokrouhlovacích chyb a její vliv na zpomalení poklesu velikosti chyby v metodě konjugovaných gradientů jsou zřejmými fakty. Mohou se projevit v závislosti na rozložení vlastních čísel matice soustavy i u velmi malých úloh a již po několika počátečních iteracích; viz [20]. Z předcházející části našeho příběhu víme, jak lze popsat numerické chování metody konjugovaných gradientů pomocí matematické úlohy s distribuční funkcí upravenou nahrazením jednotlivých vlastních čísel velmi těsnými shluky. Mimo vší pochybnost tedy pro matematickou úlohu ekvivalentní přesnému výpočtu (a tím spíše pro nepřesný numerický výpočet) platí:

Má-li SPD matice A úlohy (1) *několik (řekněme t) od sebe vzdálených těsných shluků vlastních čísel, není možné bez další informace obecně očekávat dobrou aproximaci řešení v t iteracích metody konjugovaných gradientů.*

Vše závisí na poloze jednotlivých shluků.[14] Bohužel však je v literatuře rozšířen a stále opakován, ve zřejmém rozporu s uvedeným matematicky dokázaným faktem opakovaně vysvětlovaným po čtyřicet let, chybný argument *t shluků vlastních čísel* = = *dobrá aproximace řešení v t iteracích.* Je používán jako základní pravidlo chování nejen pro metodu konjugovaných gradientů, ale pro krylovovské metody obecně.

Příčiny zmíněného neutěšeného stavu jsou známé a poučné. Několik uznávaných autorů publikovalo uvedený nesmysl v renomovaných časopisech s vágním „heuristickým“ odkazem na to, že základem krylovovských metod je aproximace minimálního polynomu matice soustavy. Z vyjádření chyby aproximace (18) v metodě konjugovaných gradientů skutečně vyplývá, že jsou-li kořeny polynomu $\varphi_n^{\mathrm{CG}}(\lambda)$ rovny všem navzájem různým vlastním číslům matice A, pak jsme dosáhli minimálního polynomu, chyba je nulová a nalezli jsme přesné řešení. V předcházejících iteracích však *aproximace minimálního polynomu* i pro SPD matici zahrnuje veškeré souvislosti popsané výše včetně řešení problému momentů, takže *kvantitativně nelze obecně o velikosti chyby naprosto nic říci.*[15] Ještě mnohem složitější je situace v úlohách, kdy matice soustavy není symetrická. O tom velmi krátce pojednáme v následující kapitole.

9. Co když nemáme k dispozici ortogonální spektrální rozklad?

Nejprve rozšíříme kontext uvažované úlohy (1). Při praktických výpočtech lze jen zřídka použít metodu konjugovaných gradientů (a krylovovské metody obecně) na (diskretizované) úlohy, které přímo modelují studovaný jev. Konvergence by až na triviální případy byla příliš pomalá. Proto je do procesu řešení zahrnuta transformace původní úlohy s cílem dosáhnout takových vlastností transformované matice a pravé strany, které by umožnily krylovovským metodám mnohem rychlejší konvergenci. Daný postup se z historického důvodu poněkud nešťastně nazývá *předpodmiňování*. Zabývá se jím velmi rozsáhlá, různorodá a myšlenkově bohatá oblast výpočtové matematiky, která využívá hluboké teoretické znalosti od teorie grafů přes vlastnosti operátorů a matic k numerické stabilitě. Vedle toho formuluje velmi netriviální heuristiky a používá složité implementační techniky, bez kterých by nebylo možné vytvoření

[14] Velká vlastní čísla λ_j matice A výrazně vzdálená od zbytku spektra jsou tradičně spojována se zrychlováním konvergence metody konjugovaných gradientů, které je vysvětlováno jejich brzkou aproximací vlastními čísly Jacobiho matic (neboli kořeny odpovídajících ortogonálních polynomů; viz (19)) a ztrátou jejich vlivu na chybu aproximace (18). Argument však selhává v přítomnosti zaokrouhlovacích chyb. Je-li velké vlastní číslo λ_j nahrazeno shlukem vlastních čísel, což při přesném výpočtu odpovídá řešení původní úlohy v aritmetice s konečnou přesností, pak nezávisle na vzájemné blízkosti vlastních čísel ve shluku jeho aproximace jedním vlastním číslem Jacobiho matice T_n nic neřeší a během několika iterací je potřeba aproximace téhož shluku dalším vlastním číslem rozšířené Jacobiho matice atd. Proto jsou při nepřesném výpočtu opakovaně generovány aproximační kopie jednotlivých vlastních čísel, což vede ke zpoždění konvergence. Každá kopie navíc stojí právě jednu iteraci, takže očekávané zrychlení může být zčásti nebo i zcela eliminováno v důsledku šíření zaokrouhlovacích chyb; viz [17], stranu 19, [29], sekce 5.9.1 a 5.9.2, a [16], [14].

[15] Bohužel je většina těch, kdo bez rozmyslu papouškují chytlavou tezi, byť by šlo o nesmysl. Úmyslně zde neuvádíme původní publikace, ze kterých se uvedené zmatení rozšířilo (jejich autoři dosáhli významných výsledků a zasloužili se o rozvoj oboru v jiných směrech). Počty pozitivně míněných (někdy i oslavných) citací koncepčně zcela chybných prací se však bohužel pohybují v mnoha stovkách. Vidíme, jak snadno se můžeme dostat i v matematice do doby post-pravdivé, kdy se již nezkoumá rozumnost východiska a logická správnost konstrukce argumentu. Kritické zkoumání je i v matematice nahrazováno *názorem autority.*

efektivních počítačových kódů. Mezi průkopnické práce, které také podstatně přispěly k rychlému rozšíření praktického zájmu o krylovovské metody v osmdesátých letech, jistě patří van der Vorstova dizertace [44] tvořená několika dříve publikovanými články s velmi poučeným komentářem (viz např. poznámku o mylné argumentaci používající shluky vlastních čísel na straně 4) a článek [9] kombinující několik přístupů. Zájemce o základní přehled algebraických přístupů odkazujeme na [3] a zejména na dřívější článek [4], který se sice soustřeďuje na předpodmiňování pomocí konstrukce přibližných inverzí, ale který přesvědčivě ukazuje náročnost problémů, se kterými je nutné se utkat.[16] Širší kontext obsahuje novější přehledový článek [45]. V úvodu zmíněná monografie [31] se soustředí na operátorové předpodmiňování metody konjugovaných gradientů v kontextu numerického řešení okrajových úloh popsaných eliptickými parciálními diferenciálnímu rovnicemi. Nejnovější přehled [35] je organizován po vybraných důležitých aplikacích. Zde se nemůžeme přepodmiňováním šířeji zabývat. Všimneme si pouze jediné otázky, která se vrací zpět k závěru druhé kapitoly našeho textu a která je v celém konceptu předpodmiňování zásadní.

Jaké vlastnosti transformované úlohy umožní metodám krylovovských podprostorů rychlou konvergenci? Víme, že ani pro SPD matici a metodu konjugovaných gradientů nelze dát jednoduchou odpověď s výjimkou případů, kdy lze dosáhnout velmi malé podmíněnosti transformované matice. Cílem předpodmiňování *není obecně snížení čísla podmíněnosti matice soustavy*, instruktivní příklad lze nalézt v [14], viz rovněž [15]. Co tedy? Znovu se v literatuře z temnot vynořuje chybná teze o žádoucím přeskupení vlastních čísel matice do několika těsných od sebe vzdálených shluků. Tato teze je dokonce prezentována *pro obecné matice* i v encyklopedických heslech jako cíl předpodmiňování. Nejde obecně o správný cíl ani pro SPD matice, kde spektrální informace umožňuje popsat chování metody konjugovaných gradientů. Co když nelze sestavit ortonormální bázi prostoru, ve kterém se při numerickém řešení pohybujeme, pomocí vlastních vektorů? Dokonce nemusí jít sestavit *žádnou bázi*, neboť matice soustavy může mít v extrémním případě pouze jediný vlastní vektor. Není tedy možné spojit pro libovolnou úlohu krylovovské metody s konstrukcí distribuční funkce a na jejich chování přímočaře usuzovat ze spektra matice soustavy. Obtížnost problému je patrná z výsledků zveřejněných v [21], [19], [1]. Vedou k parametrizovaným množinám příkladů, pro které libovolně zvolené chování krylovovské metody GMRES nastane při libovolně zvoleném spektru matice soustavy. Jak je tedy možné, že dávno dokázaným výsledkům odporující teze o shlucích vlastních čísel má takovou přitažlivost a životnost?

První příčinou je experimentální zkušenost. V řadě případů lze při shlukovém uspořádání spektra skutečně pozorovat rychlou konvergenci. U pozorování se ale nesmíme zastavit, jako vědci jsme povinni ptát se po příčině, *proč* tomu tak je. Matice může například mít v takovém případě nějakou speciální strukturu invariantních podprostorů, kterou je potřeba zkoumat. To se však neděje. Jde o obtížnou otázku a její otevření by například mohlo zpozdit publikaci článku (nejde o smyšlenku, ale o opakovanou zkušenost z osobních diskusí).

Druhou příčinou je paradoxně výsledek z krátkého a nádherného článku [33]. Autoři ukázali třídu úloh, pro které existuje předpodmínění zaručující velmi malý stupeň

[16] Nejen v daném oboru si svými výsledky získala světové uznání skupina soustředěná kolem M. Tůmy.

minimálních polynomů (a tedy pouze několik vlastních čísel) transformovaných matic. Pak přesný výpočet s použitím vhodné krylovovské metody (ve které nemůže dojít k předčasnému zastavení zvanému breakdown) dá přesné řešení v odpovídajícím počtu několika iterací. Článek je *přirozenou motivací* pro postupy úspěšně používané při praktických výpočtech, což je zcela na místě. Nemůže však být jejich *rigorózním matematickým zdůvodněním*. V praktických výpočtech je z důvodu ceny výpočtu teoretické předpodmiňování zjednodušeno a tím pouze aproximováno (navíc je zde vliv zaokrouhlovacích chyb). Odpovídající transformované matice proto nemají jen několik vlastních čísel, ale v nejlepším případě, který nemusí v závislosti na vlastnostech úlohy nastat, pouze *těsné shluky vlastních čísel*. Celkový počet navzájem různých vlastních čísel je mnohařádově větší než počet shluků a může se pohybovat v řádu miliónů i miliard. *Skutečný stupeň minimálních polynomů odpovídajících výpočtům musí tedy být nejméně stejného řádu jako počet navzájem různých vlastních čísel.* Jsme zpět u výše popsané zmatečnosti argumentace o shlucích a minimálním polynomu. Navíc je zde téměř všudypřítomné zmatení nesprávně dávající do rovnosti počet navzájem různých vlastních čísel a stupeň minimálního polynomu, před čímž důrazně varovali i autoři článku [33].[17] Nic z výše napsaného nesnižuje význam výsledku v [33], který spolu s [21], [19], [1] otevírá prostor pro nové cesty teoretického zkoumání, které se ale zatím bohužel nekoná.

10. Poznámka k formulacím v nekonečnědimenzionálních Hilbertových prostorech

Co z výše napsaného lze přenést do Hilbertových prostorů nekonečné dimenze? Téměř vše, musíme však dát pozor na změnu významu některých pojmů (například pojmů báze, spektra operátoru, omezenosti operátoru, kompaktnosti operátoru, používání duálních prostorů, superlineární konvergence, ...). Také některá odvození a důkazy budou malinko jiné; viz např. [31]. Je pozoruhodné, jak (nejen v kontextu nekonečné dimenze) jsou mnohé dlouho známé výsledky znovu a znovu publikovány jako původní. Například neustále znovuobjevovaná superlineární konvergence byla popsána v krásných textech Karushe [24] a Hayese [22] již v letech 1952 a 1954. Podobně důkaz monotónního poklesu eukleidovské normy chyby v metodě konjugovaných gradientů je obsažen (se samozřejmostí metodické důkladnosti výkladu) již v původním článku [23]. Formulace krylovovských metod v nekonečnědimenzionálních prostorech má dobrý smysl a věnuje se jí v poslední době rostoucí počet publikací. Jde však za rámec a rozsah předloženého textu.

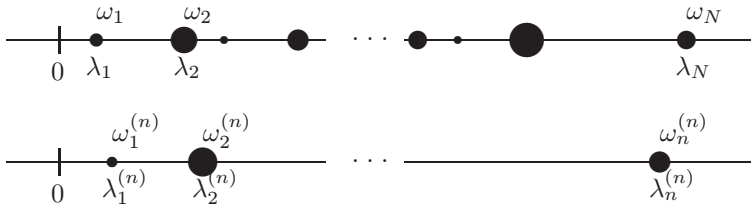
11. Beznadějný boj s větrnými mlýny nebo příležitost?

Metoda konjugovaných gradientů a její příběh, včetně rozvětvujících se příběhů krylovovských metod, nás mohou velmi mnohému naučit.

Matematicky vzato, neměli bychom ani při rutinním numerickém počítání zapomínat, jak souvisí metoda konjugovaných gradientů s úlohou nalézt pro dané $n < N$

[17] Stupeň minimálního polynomu matice je horní mezí počtu navzájem různých vlastních čísel. Opačná nerovnost neplatí. Matice může mít pouze jedno vlastní číslo a stupeň svého minimálního polynomu roven velikosti matice (viz např. Jordanův blok).

distribuční funkci s pouze n body vzrůstu tak, aby *co nejlépe* vystihovala vlastnosti distribuční funkce určené maticí A a normovaným počátečním reziduem, viz obrázek.



Mohou nás ale učit i mnohem důležitějším věcem. Například tomu, že stojí za to kriticky myslet a nejít s většinovým názorem, i když to z pohledu okolím chápaného úspěchu nemusí být výhodné. Stejně tak je dobré být si vědom svých omezení a svých chyb a být vděčný za poučení. Řečeno dávnými slovy ve vzácném překladu, stojí za to *zlomit svoji pýchu a hledat pokoru*. Jsem vděčný všem, které mi bylo dáno v životě potkat a kteří přirozeně umí odlišovat podstatné od nepodstatného a ze kterých vyzařuje radost.

Poděkování. Jsem rovněž velmi vděčný za odborné a lidské společenství, které je prací věnovanou krylovovským metodám ovlivněno a které po mnoho let spoluvytváříme s Mirkem Tůmou, Mirem Rozložníkem a mnoha našimi vzácnými mladšími kolegyněmi a kolegy. Většina z nich významně přispěla k výsledkům týkajícím se krylovovských metod, které jsou ve světě ceněny. Zde je zmíněno jen několik z nich, neboť omezení rozsahu neumožňuje větší tematické rozšíření. Děkuji všem, kteří mi pomohli při psaní předloženého textu.

L i t e r a t u r a

[1] Arioli, M., Pták, V., Strakoš, Z.: *Krylov sequences of maximal length and convergence of GMRES*. BIT 38 (1998), 636–643.

[2] Axelsson, O., Barker, V. A.: *Finite element solution of boundary value problems, theory and computations*. Academic Press, Orlando, FL, 1984.

[3] Benzi, M.: *Preconditioning techniques for large linear systems: a survey*. J. Comput. Phys. 182 (2002), 418–477.

[4] Benzi, M., Tůma, M.: *A comparative study of sparse approximate inverse preconditioners*. Appl. Numer. Math. 30 (1999), 305–340.

[5] Brandts, J., Křížek, M.: *Padesát let metody konjugovaných gradienů aneb zvládnou počítače soustavy miliónů rovnic o miliónech neznámých?* PMFA 47 (2002), 103–113.

[6] Brezinski, C.: *History of continued fractions and Padé approximants*. Springer Series in Computational Mathematics, vol. 12. Springer-Verlag, Berlin, 1991.

[7] Carson, E., Rozložník, M., Strakoš, Z., Tichý, P., Tůma, M.: *The numerical stability analysis of pipelined conjugate gradient methods: historical context and methodology*. SIAM J. Sci. Comput. 40 (2018), A3549–A3580.

[8] Carson, E., Strakoš, Z.: *On the cost of iterative computations*. Philos. Trans. Roy. Soc. A 378 (2020).

[9] Concus, P., Golub, G. H., O'Leary, D. P.: *A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations*. In: J. R. Bunch, D. J. Rose: Sparse Matrix Computations, Academic Press, New York, 2018, 309–332.

[10] Daniel, J. W.: *The conjugate gradient method for linear and nonlinear operator equations*. SIAM J. Numer. Anal. 4 (1967), 10–26.

[11] Duintjer Tebbens, J., Hnětynková, I., Plešinger, M., Strakoš, Z., Tichý, P.: *Analýza metod pro maticové výpočty – základní metody*. MatfyzPress, Praha, 2012.

[12] Engeli, M., Ginsburg, T., Rutishauser, H., Stiefel, E.: *Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems*. Mitt. Inst. Angew. Math. Zürich 8, Birkhäuser, Basel, 1959.

[13] Fischer, B.: *Polynomial based iteration methods for symmetric linear systems*. Wiley-Teubner Series Advances in Numerical Mathematics, John Wiley and Sons, Chichester, 1996.

[14] Gergelits, T., Mardal, K.-A., Nielsen, B. F., Strakoš, Z.: *Laplacian preconditioning of elliptic PDEs: Localization of the eigenvalues of the discretized operator*. SIAM J. Numer. Anal. 57 (2019), 1369–1394.

[15] Gergelits, T., Nielsen, B. F., Strakoš, Z.: *Generalized spectrum of second order differential operators*. SIAM J. Numer. Anal. 58 (2020), 2193–2211.

[16] Gergelits, T., Strakoš, Z.: *Composite convergence bounds based on Chebyshev polynomials and finite precision conjugate gradient computations*. Numer. Algorithms 65 (2014), 759–782.

[17] Greenbaum, A.: *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*. Linear Algebra Appl. 113 (1989), 7–63.

[18] Greenbaum, A.: *Iterative methods for solving linear systems*. Frontiers in Applied Mathematics, vol. 17. SIAM, Philadelphia, PA, 1997.

[19] Greenbaum, A., Pták, V., Strakoš, Z.: *Any nonincreasing convergence curve is possible for GMRES*. SIAM J. Matrix Anal. Appl. 17 (1996), 465–469.

[20] Greenbaum, A., Strakoš, Z.: *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*. SIAM J. Matrix Anal. Appl. 13 (1992), 121–137.

[21] Greenbaum, A., Strakoš, Z.: *Matrices that generate the same Krylov residual spaces*. In: Recent advances in iterative methods. IMA Vol. Math. Appl., vol. 60. Springer, New York, 1994, 95–118.

[22] Hayes, R. M.: *Iterative methods for solving linear problems in Hilbert space*. PhD. Thesis. Univ. of California at Los Angeles, 1954.

[23] Hestenes, M. R., Stiefel, E.: *Methods of conjugate gradients for solving linear systems*. J. Research Nat. Bur. Standards 49 (1952), 409–436.

[24] Karush, W.: *Convergence of a method for solving linear problems*. Proc. Amer. Math. Soc. 3 (1952), 839–851.

[25] Lanczos, C.: *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. J. Research Nat. Bur. Standards 45 (1950), 255–282.

[26] Lanczos, C.: *Solution of systems of linear equations by minimized iterations*. J. Research Nat. Bur. Standards 49 (1952), 33–53.

[27] Lanczos, C.: *Chebyshev polynomials in the solution of large-scale linear systems*. In: Proceedings of the Association for Computing Machinery, Toronto, 1952, Sauls Lithograph Co., Washington, DC, 1953, 124–133.

[28] Lanczos, C.: *Why Mathematics?* Lecture given at the Annual Meeting of the Irish Mathematical Association on October 31, 1966, at Belfield, Dublin.

[29] Liesen, J., Strakoš, Z.: *Krylov subspace methods: Principles and analysis*. Oxford University Press, Oxford, 2013.

[30] Ljusternik, L. A.: *Solution of problems in linear algebra by the method of continued fractions (in Russian)*. Trudy Voronezh. Gos. Inst., Voronezh 2 (1956), 85–90.

[31] Málek, J., Strakoš, Z.: *Preconditioning and the conjugate gradient method in the context of solving PDEs*. SIAM Spotlights, vol. 1. SIAM, Philadelphia, PA, 2015.

[32] Meurant, G., Strakoš, Z.: *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*. Acta Numer. 15 (2006), 471–542.

[33] Murphy, M. F., Golub, G. H., Wathen, A. J.: *A note on preconditioning for indefinite linear systems*. SIAM J. Sci. Comput. 21 (2000), 1969–1972.

[34] Paige, C. C.: *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*. Linear Algebra Appl. 34 (1980), 235–258.

[35] Pearson, J. W. Pestana, J.: *Preconditioned iterative methods for scientific applications*. GAMM-Mitt., to appear (2020).

[36] Pozza, S., Strakoš, Z.: *Algebraic description of the finite Stieltjes moment problem*. Linear Algebra Appl. 561 (2019), 207–227.

[37] Reid, J. K.: *On the method of conjugate gradients for the solution of large sparse systems of linear equations*. In: Large sparse sets of linear equations, Proc. Conf., St. Catherine's Coll., Oxford, 1970, Academic Press, London, 1971, 231–254.

[38] Rektorys, K.: *Variační metody v inženýrských problémech a v problémech matematické fyziky*. SNTL, Praha, 1974.

[39] Saad, Y.: *Iterative methods for sparse linear systems*. 2nd ed., SIAM, Philadelphia, PA, 2003.

[40] Stieltjes, T. J.: *Recherches sur les fractions continues*. Ann. Fac. Sci. Toulouse Sci. Math. Sci. Phys. 8 (1894), J. 1–122. Reprinted in Oeuvres II (P. Noordhoff, Groningen, 1918), 402–566. English translation *Investigations on continued fractions*. in Thomas Jan Stieltjes, Collected Papers, Vol. II, Springer-Verlag, Berlin, 1993, 609–745.

[41] Strakoš, Z., Tichý, P.: *On error estimation in the conjugate gradient method and why it works in finite precision computations*. Electron. Trans. Numer. Anal. 13 (2002), 56–80.

[42] Thurston, W.: *On proof and progress in Mathematics*. Bull. Amer. Math. Soc. 30 (1994), 161–177.

[43] Vorobyev, Yu. V.: *Methods of moments in applied mathematics*. Translated from the Russian original published in 1958 by Bernard Seckler, Gordon and Breach Science Publishers, New York, 1965.

[44] van der Vorst, H. A.: *Preconditioning by incomplete decompositions*. PhD Thesis. University of Utrecht, 1982.

[45] Wathen, A.: *Preconditioning*. Acta Numer. 24 (2015), 329–376.

[46] Zeidler, E.: *Oxford users' guide to mathematics*. Oxford University Press, Oxford, 2004. Translated from the 1996 German original by Bruce Hunt.