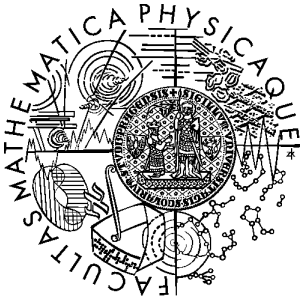
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

# BAKALÁŘSKÁ PRÁCE




Jan Papež

# Odhady energetické a euklidovské normy chyby v metodě konjugovaných gradientů

Katedra numerické matematiky

Vedoucí bakalářské práce: prof. Ing. Zdeněk Strakoš, DrSc.
Studijní program: Matematika, Obecná matematika

2009

Mé upřímné poděkování patří profesoru Strakošovi za trpělivé vedení a poskytnutou literaturu a doktoru Tichému za technickou pomoc.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 27. 7. 2009 Jan Papež

# Obsah

**1 Metoda sdružených gradientů (CG) 7**
1.1 Konzistence CG s Galerkinovou metodou . . . . . 7
1.2 Odvození CG . . . . . 11
1.3 Zápis CG . . . . . 14
1.4 Vlastnosti CG . . . . . 15

**2 CG v souvislostech 18**
2.1 Lanczosova metoda . . . . . 18
2.2 Ortogonální polynomy a Gaussova kvadratura . . . . . 20

**3 Odhady chyb 24**
3.1 Odhady $A$-normy chyby a jejich srovnání . . . . . 24
3.1.1 Dolní odhady . . . . . 24
3.1.2 Horní odhady . . . . . 26
3.2 Odhad euklidovské normy chyby . . . . . 27
3.3 Odhad v PCG . . . . . 27
3.4 Otevřené problémy . . . . . 29

**4 Experimenty 31**
4.1 Stagnace $A$-normy chyby a volba parametru $d$ . . . . . 32
4.2 Adaptivní volba $d$ . . . . . 35
4.2.1 Možné modifikace . . . . . 36
4.3 Rekonstrukce křivky konvergence . . . . . 40

**Použité matice 42**

**Literatura 44**

Název práce: Odhady energetické a euklidovské normy chyby v metodě konjugovaných gradientů
Autor: Jan Papež
Katedra (ústav): Katedra numerické matematiky
Vedoucí bakalářské práce: prof. Ing. Zdeněk Strakoš, DrSc.
e-mail vedoucího: strakos@cs.cas.cz

Abstrakt: Po seznámení s metodou sdružených gradientů (Conjugate Gradient Method - CG) je v předložené práci ukázán význam minimalizace energetické normy chyby, CG je odvozena s využitím minimalizace kvadratického funkcionálu a je ukázána souvislost s Lanczosovou metodou a teorií ortogonálních polynomů. Přehledová část je zakončena odvozením dolních odhadů energetické normy chyby a jejich srovnáním, včetně uvedení dalších souvislostí. Na základě provedených experimentů je v závěru práce navržena heuristika pro adaptivní zpřesňování odhadů ve výpočtech a je sledováno její chování.

Klíčová slova: metoda sdružených gradientů, odhady energetické a euklidovské normy chyby, adaptivní volba parametru odhadu

Title: Estimation of the energy and Euclidean norms of the error in the conjugate gradient method
Author: Jan Papež
Department: Department of Numerical Mathematics
Supervisor: Zdeněk Strakoš
Supervisor's e-mail address: strakos@cs.cas.cz

Abstract: After an introduction to Conjugate Gradient Method (CG) it is demonstrated the importance of the minimalization of the error energy norm, CG is derived using the minimalization of the quadratic functional and the relationship with Lanczos Method and the theory of ortogonal polynoms is presented. Overview part is concluded by derivation of the lower bounds of the error energy norm and its comparison including presentation of other relationships. Based on performed experiments a heuristic for adaptive precising of the estimate in computations is deduced and its behaviour is observed.

Keywords: Conjugate Gradient Method, estimates of the error energy and Euclidean norms, adaptive choice of the estimate parameter

# Úvod

Má práce má dva hlavní cíle. Prvním je seznámit se s metodou sdružených gradientů (Conjugate Gradient Method - CG, [6]) se zaměřením především na možné přístupy k odhadům chyb. Druhým pak navrhnout na základě experimentů některé heuristiky pro použití těchto odhadů v praktických výpočtech.

V první kapitole ukážeme nejprve na příkladě Galerkinovy metody konečných prvků význam minimalizace $A$-normy chyby při řešení soustav lineárních rovnic. Poté odvodíme metodu sdružených gradientů přes minimalizaci kvadratického funkcionálu, která poměrně intuitivně vede na klasické zapsání algoritmu CG. Kromě popisu metody zmíníme některé její vlastnosti, které využijeme v dalších částech práce.

I přes svou zdánlivou jednoduchost propojuje CG mnoho oblastí matematiky a můžeme na ni nahlížet z několika pohledů. Ve druhé kapitole popíšeme vztah CG s Lanczosovou metodou. Obě lze díky ortogonalitě studovat i v souvislosti s ortogonálními polynomy, pro něž definujeme skalární součin prostřednictvím Riemann-Stieltjesova integrálu. Aproximací příslušného Riemann-Stieljesova integrálu Gaussovými kvadraturami pro specifickou volbu integrované funkce byly odhady odvozovány (viz například [2]). Zde uvedeme jeden příklad odhadu založeného na Gaussově kvadratuře, abychom ukázali tento přístup k odhadům v metodě sdružených gradientů a její teoretickou hloubku.

Ve třetí kapitole věnované odhadům $A$-normy chyby provedeme nejprve jejich srovnání (s využitím [9]). Přestože je CG teoreticky metodou finitní, jak uvidíme dále, v každé iteraci dostáváme aproximaci přesného řešení, která je navíc optimální ve smyslu $A$-normy chyby. Proto bývá CG řazena mezi iterační metody a je vhodné sledovat $A$-normu chyby během výpočtu. Bez znalosti přesného řešení samozřejmě nemůžeme tuto normu spočítat. To je důvodem, proč věnujeme pozornost jejím odhadům. Přirozeně od odhadů požadujeme dostatečnou přesnost a malou náročnost výpočtu. V této ka-

pitole se budeme věnovat přesnosti odhadu získaného pomocí předpočítání několika iterací a zamyslíme se nad pevnou i variabilní volbou parametru udávajícího jejich počet. Otázka, jakou volbou zajistit požadovanou přesnost odhadu, zůstává zatím nezodpovězena.

V experimentální části budeme hledat odpověď na výše položenou otázku porovnáváním hodnot různých parametrů (například normy rezidua). Navrhneme heuristiku pro adaptivní volbu parametru a na příkladech budeme sledovat její chování. Zaměříme se i na rekonstrukci křivky konvergence.

Shrnutí práce i směry, kterými by se mohla v budoucnu ubírat má pozornost, budou popsány v závěru.

# Kapitola 1

# Metoda sdružených gradientů (CG)

Ještě než odvodíme a popíšeme metodu sdružených gradientů, uvedeme jako motivaci důkaz konzistence CG s Galerkinovou metodou konečných prvků pro Poissonovu rovnici. Uvidíme, že diskretizace úlohy Galerkinovou metodou vede na soustavu lineárních rovnic se symetrickou pozitivně definitní maticí. Aproximaci řešení této soustavy pak odpovídá funkce, která aproximuje řešení původní Poissonovy rovnice. Ukážeme, že chceme-li minimalizovat energetickou chybu spojité aproximace, musíme minimalizovat $A$-normu chyby diskrétní aproximace řešení soustavy lineárních rovnic. To je (jak uvidíme dále) základní vlastností metody sdružených gradientů, a proto je vhodné použít CG pro řešení této soustavy. V následujících dvou částech postupujeme podle [7], respektive [3].

## 1.1 Konzistence CG s Galerkinovou metodou

Hledejme $u = u(\xi_1, \xi_2)$, kde $\xi_1, \xi_2$ označují prostorové proměnné, takové, že

$$-\Delta u = f \quad \text{v} \quad \Omega \subset \mathbb{R}^2 \,, \tag{1.1}$$

$$u = g_{\mathcal{D}} \quad \text{na} \quad \partial\Omega_{\mathcal{D}} \,, \tag{1.2}$$

$$\frac{\partial u}{\partial n} = g_{\mathcal{N}} \quad \text{na} \quad \partial\Omega_{\mathcal{N}} \,, \tag{1.3}$$

kde $\partial\Omega_{\mathcal{D}} \cup \partial\Omega_{\mathcal{N}} = \partial\Omega$ a $\partial\Omega_{\mathcal{D}}$ a $\partial\Omega_{\mathcal{N}}$ jsou disjunktní. $\partial u/\partial n$ označuje derivaci ve směru normály k hranici $\partial\Omega$. Pro jednoduchost předpokládejme, že $\int_{\partial\Omega_{\mathcal{D}}} ds \neq 0$, a tedy (1.2) – (1.3) nereprezentuje čistě Neumannovu okrajovou podmínku.

V dalším budeme uvažovat řešení $u$ i testovací funkce $v$ vždy z vhodného prostoru funkcí (viz například [1]).

Slabé řešení $u$ úlohy (1.1) – (1.3) splňuje pro každou testovací funkci $v$

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega v\, f + \int_{\partial\Omega_{\mathcal{N}}} v\, g_{\mathcal{N}} \ . \tag{1.4}$$

Uvažováním bilineární formy $a$

$$a(u,v) \equiv (\nabla u, \nabla v) = \int_\Omega \nabla u \cdot \nabla v$$

a lineárního funkcionálu $l$

$$l(v) \equiv (f,v) + (g_{\mathcal{N}}, v)_{\partial\Omega_{\mathcal{N}}} = \int_\Omega v\, f + \int_{\partial\Omega_{\mathcal{N}}} v\, g_{\mathcal{N}}$$

můžeme (1.4) přepsat

$$a(u,v) = l(v) \ . \tag{1.5}$$

Galerkinovu aproximaci (1.5) metodou konečných prvků můžeme stručně charakterizovat takto. Mějme konečně dimenzionální vektorový podprostor $\mathcal{S}_0^h$ prostoru testovacích funkcí a nechť $\phi_1, \ldots, \phi_n$ tvoří jeho bázi. Dále uvažujme funkci $u_{\mathcal{D}}$, která interpoluje na $\Omega_{\mathcal{D}}$ Dirichletovu okrajovou podmínku $g_{\mathcal{D}}$. Potom aproximaci metodou konečných prvků $u_h \in \mathcal{S}_E^h \equiv u_{\mathcal{D}} + \mathcal{S}_0^h$ můžeme zapsat ve tvaru

$$u_h = \sum_{j=1}^{n} \zeta_j \phi_j + u_{\mathcal{D}} \ . \tag{1.6}$$

Dosazením (1.6) do (1.4) a volbou $v = \phi_i,\ i = 1, 2, \ldots, n$ dostáváme soustavu lineárních rovnic

$$Ax = b \ , \tag{1.7}$$

kde

$$
\begin{aligned}
A &= [a_{ij}] \ , \quad a_{ij} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \ , \\
b &= \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \ , \quad b_i = \int_\Omega \phi_i f + \int_{\partial\Omega_\mathcal{N}} \phi_i g_\mathcal{N} - \int_\Omega \nabla\phi_i \cdot \nabla u_\mathcal{D} \ , \\
x &= \begin{bmatrix} \zeta_1 \\ \vdots \\ \zeta_n \end{bmatrix} \ .
\end{aligned}
$$

Za předpokladu $\int_{\partial\Omega_\mathcal{D}} ds \neq 0$ a použitím Poincaré-Friedrichsovy nerovnosti ([3], str. 37) reprezentuje bilineární forma $a(u, v)$ skalární součin na prostoru testovacích funkcí a indukuje na něm *energetickou normu*

$$a(v, v)^{1/2} = \|\nabla v\| \ .$$

Z toho plyne, že matice $A$ je symetrická pozitivně definitní.

Zde připomínáme, že jsme neuvažovali čistě Neumannovu okrajovou podmínku. Navíc budeme dále předpokládat, že Galerkinovské řešení $u_h$ je konformní, neboli že Dirichletovu okrajovou podmínku interpolujeme funkcí $u_\mathcal{D}$ přesně.

Z konstrukce popsané výše je zřejmé, že pro libovolné funkce $\tilde{v}, \tilde{w} \in \mathcal{S}_0^h$ takové, že $\tilde{v} = \Phi\, x, \tilde{w} = \Phi\, y$ pro $\Phi = [\phi_1, \ldots, \phi_n]$, můžeme psát

$$a(\tilde{v}, \tilde{w}) = \int_\Omega \nabla\tilde{v} \cdot \nabla\tilde{w} = (y, Ax) \ .$$

Pak pro libovolnou $\tilde{v} \in \mathcal{S}_0^h$

$$a(\tilde{v}, \tilde{v}) = (x, Ax) = \|x\|_A^2$$

definuje *algebraickou energetickou normu* na $\mathbb{R}^n$.

Za předpokladů na konformitu aproximací, pro každé $\tilde{u}_h \in \mathcal{S}_E^h$ splňuje $u - \tilde{u}_h$ nulovou Dirichletovu okrajovou podmínku. Pro každé Galerkinovo řešení $u_h$, které je dáno (1.6) a (1.7), a pro libovolné $v_h \in \mathcal{S}_0^h$ davají vztahy

$$a(u, v_h) = l(v_h) \quad \text{a} \quad a(u_h, v_h) = l(v_h)$$

tak zvanou *Galerkinovu ortogonalitu*

$$a(u - u_h, v_h) = 0 \ , \tag{1.8}$$

a tedy chyba diskretizace $u - u_h$ je kolmá na podprostor $\mathcal{S}_0^h$ vzhledem k energetickému skalárnímu součinu. Platnost $u_h \in u_\mathcal{D} + \mathcal{S}_0^h$ a (1.8) znamená, že na

$$u - u_h = (u - u_\mathcal{D}) - \sum_{j=1}^{n} \zeta_j \phi_j$$

můžeme pohlížet jako na výsledek ortogonalizace vektoru $u - u_\mathcal{D}$ vzhledem k $\phi_1, \phi_2, \ldots, \phi_n$ v energetickém skalárním součinu. Pak zřejmě musí mít nejmenší normu mezi všemi vektory $u - u_\mathcal{D} - w_h^0$ pro $w_h^0 \in \mathcal{S}_0^h$. Přímým důsledkem konformity aproximací a (1.8) je pak

$$\|\nabla u - \nabla u_h\| = \min_{w_h \in \mathcal{S}_E^h} \|\nabla u - \nabla w_h\| \ . \tag{1.9}$$

Nechť nyní $x = (\zeta_1, \ldots, \zeta_n)^T$ je řešení (1.7) a $x_k \in \mathbb{R}^n$ je jeho aproximace. Z konstrukce a za podmínky konformity

$$u_h^{(k)} = \Phi\, x_k + u_\mathcal{D} \ \in \mathcal{S}_E^h \ , \quad u_h - u_h^{(k)} \in \mathcal{S}_0^h. \tag{1.10}$$

Pak se chyba vypočteného přibližného řešení $u_h^{(k)}$ v energetické normě skládá ze dvou částí:

- z diskretizační chyby $u - u_h$, jejíž energetická norma je dána

$$\|\nabla u - \nabla u_h\|$$

- z algebraické chyby $u_h - u_h^{(k)}$, kterou můžeme vyjádřit pomocí zavedené algebraické energetické normy jako

$$\begin{aligned} \|\nabla u_h - \nabla u_h^{(k)}\|^2 &= \|\nabla(u_h - u_h^{(k)})\|^2 = ((x - x_k), A(x - x_k)) \\ &= \|x - x_k\|_A^2 \, . \end{aligned}$$

A nyní můžeme vyslovit větu, ze které vyplývá konzistence Galerkinovy diskretizace metodou konečných prvků a metody sdružených gradientů.

**Věta 1.** *Nechť $u_h$ je Galerkinova aproximace řešení $u$ úlohy (1.1) – (1.3) a nechť $u_h^{(k)}$ odpovídá (1.10) přibližnému řešení $x_k$ soustavy lineárních rovnic (1.7). Za předpokladu konformity platí*

$$\|\nabla u - \nabla u_h^{(k)}\|^2 = \|\nabla(u - u_h)\|^2 + \|x - x_k\|_A^2 \, .$$

*Důkaz.*

$$\begin{aligned}\|\nabla(u-u_h^{(k)})\|^2 &= a(u-u_h^{(k)},u-u_h^{(k)})\\ &= a(u-u_h+u_h-u_h^{(k)},u-u_h+u_h-u_h^{(k)})\\ &= a(u-u_h,u-u_h)+a(u_h-u_h^{(k)},u_h-u_h^{(k)})\ ,\end{aligned}$$

protože $a(u-u_h,u_h-u_h^{(k)})=0$ z Galerkinovské ortogonality. Věta pak plyne snadným přepsáním druhého členu do tvaru algebraické energetické normy. □

## 1.2 Odvození CG

Uvažujme soustavu lineárních rovnic

$$Ax=b\ ,$$

kde $A\in\mathbb{R}^{n\times n}$ je reálná symetrická pozitivně definitní (SPD) matice a $b\in\mathbb{R}^n$ je vektor pravé strany. Je známo, že danou úlohu lze převést na úlohu minimalizace kvadratického funkcionálu

$$F(x)=\frac{1}{2}(x,Ax)-(x,b)\ .$$

Je-li $A$ SPD, pak nalezení minimalizace kvadratického funkcionálu $F(x)$ je ekvivalentní nalezení řešení rovnice

$$\nabla F(x)=Ax-b=0$$

Definujeme-li $A$-normu

$$\|z\|_A=(z,Az)^{1/2}$$

a uvažujeme-li $x_k$ aproximaci řešení $x$, pak platí

$$\begin{aligned}F(x_k) &= \tfrac{1}{2}\left((x-x_k),A(x-x_k)\right)-\tfrac{1}{2}(x,Ax)=\\ &= \tfrac{1}{2}\|x-x_k\|_A^2-\tfrac{1}{2}\|x\|_A^2\ ,\end{aligned}$$

a tedy minimalizace $F(z)$ na nějakém podprostoru $\mathbb{R}^n$ je ekvivalentní minimalizaci $\|x-z\|_A$ na tomtéž podprostoru.

Mějme počáteční vektor $x_0$ a konstruujme posloupnost aproximací řešení $x$ jednoduchým rekurentním vztahem

$$x_k = x_{k-1} + \gamma_{k-1} p_{k-1} \ , \quad k = 1, 2, \ldots$$

kde $p_{k-1}$ reprezentuje směrový vektor v kroku $k$. Následující aproximace $x_k$ je určena jako bod, ve kterém se minimalizuje $\|x - z\|_A$ na přímce $x_{k-1} + \gamma_{k-1} p_{k-1}$. Jednoduchým výpočtem

$$\|x - x_k\|_A^2 = \|x - x_{k-1}\|_A^2 - 2\gamma_{k-1}(p_{k-1}, r_{k-1}) + \gamma_{k-1}^2 (p_{k-1}, Ap_{k-1}) \ .$$

Derivováním podle $\gamma_{k-1}$

$$0 = -2(p_{k-1}, r_{k-1}) + 2\gamma_{k-1}(p_{k-1}, Ap_{k-1}) \ ,$$

a tedy minima je dosaženo pro

$$\gamma_{k-1} = \frac{(p_{k-1}, r_{k-1})}{(p_{k-1}, Ap_{k-1})} \ . \tag{1.11}$$

Okamžitým důsledkem je ortogonalita rezidua $r_k$ (které se vztahuje k nové aproximaci $x_k$) a směrového vektoru $p_{k-1}$, která v geometrickém smyslu znamená, že gradient $\nabla F(x_k)$ v bodě $x_k$ je kolmý na ekvipotentní plochu určenou vztahem $F(y) = F(x_k)$.

Zbývá určit směrové vektory $p_k$. Přirozenou volbou počátečního směru $p_0$ je $p_0 \equiv r_0 = b - Ax_0$. Pro volbu $p_k = r_k$, kde $r_k$ je $k$-té reziduum $r_k = b - Ax_k$, dostáváme známou metodu největšího spádu (viz například [11]). Ta však zajišťuje jen velmi pomalou konvergenci, protože v každém kroku minimalizuje normu chyby pouze přes jednorozměrný prostor určený reziduem $r_k$. Chceme-li minimalizovat přes větší prostory, musíme při volbě $p_k$ využít informace z více iteračních kroků. Nejjednodušší volbou je generovat nový směr jako kombinaci předchozího směru a (nového) rezidua

$$p_k = r_k + \delta_k p_{k-1}, \text{ pro nějaké } \delta_k \in \mathbb{R} \ .$$

Bez ohledu na to, jak zvolíme skalár $\delta_k$, vlastnost minimalizace $A$-normy na přímce $x_{k-1} + \gamma_{k-1} p_{k-1}$ implikuje

$$(p_k, r_k) = (r_k, r_k) + \delta_k (p_{k-1}, r_k) = (r_k, r_k) = \|r_k\|^2 \ .$$

Důsledkem toho se iterační proces může zastavit pouze při nalezení přesného řešení $x$. Neboť proces končí, je-li $p_k = 0$ nebo $\gamma_k = 0$. V prvním případě

$$0 = (p_k, r_k) = \|r_k\|^2 \ ,$$

a tedy $Ax_k = b$. V druhém případě pak opět platí $0 = (p_k, r_k)$ z (1.11).

Abychom zdůvodnili volbu $\delta_k$, kterou provedeme níže, všimněme si nejprve, že na změnu chyby mezi kroky $k-1$ a $k$

$$x - x_k = (x - x_{k-1}) - \gamma_{k-1} p_{k-1}$$

s hodnotou

$$\gamma_{k-1} = \frac{(p_{k-1}, r_{k-1})}{(p_{k-1}, Ap_{k-1})} = \frac{(p_{k-1}, A(x - x_{k-1}))}{(p_{k-1}, Ap_{k-1})} \ ,$$

můžeme nahlížet jako na $A$-ortogonalizaci chyby $x - x_{k-1}$ vzhledem k $p_{k-1}$. Potom

$$(x - x_{k-1}) = \gamma_{k-1} p_{k-1} + (x - x_k)$$

můžeme interpretovat jako ortogonální rozklad $x - x_{k-1}$ vzhledem ke skalárnímu součinu $(z, Ay), z, y \in \mathbb{R}^n$ a rekurzí dostáváme

$$x - x_0 = \sum_{j=1}^{k} \gamma_{j-1} p_{j-1} + (x - x_k) \ .$$

Nyní *předpokládejme*, že všechny vektory $p_0, p_1, \ldots$ jsou vzájemně ortogonální vzhledem k tomuto skalárnímu součinu (v dalším budeme říkat, že jsou *A-ortogonální* [1]). Tedy platí $(p_i, Ap_j) = 0, \ i \neq j$. Potom

$$x - x_k = (x - x_0) - \sum_{j=1}^{k} \gamma_{j-1} p_{j-1}$$

reprezentuje $A$-ortogonální rozklad $x - x_0$ a důsledkem toho $\|x - x_k\|_A$ je minimální přes všechny aproximace v podprostoru generovaném směrovými vektory $p_0, \ldots, p_{k-1}$, neboli

$$\|x - x_k\|_A = \min_{u \in x_0 + \mathrm{span}\{p_0, \ldots, p_{k-1}\}} \|x - u\|_A \ . \tag{1.12}$$

Navíc za předpokladu $A$-ortogonality $p_j, \ j = 0, 1, \ldots$ plyne $p_n = 0$, a tedy algoritmus nalezne přesné řešení $x$ v nejvýše $n$ krocích.

Protože však máme pouze jeden neurčený parametr $\delta_k$, můžeme se přiblížit globální $A$-ortogonalitě nejvýše tak, že budeme požadovat lokální $A$-ortogonalitu mezi dvěmi následujícími směrovými vektory, to jest

$$(p_{k-1}, Ap_k) = 0 \ ,$$

[1] používá se i termín *sdružené*, odtud název *metoda sdružených gradientů*

což nám dává

$$\delta_k = -\frac{(p_{k-1}, Ar_k)}{(p_{k-1}, Ap_{k-1})} \ .$$

Nyní již máme určený celý algoritmus CG. V části 1.4 o vlastnostech CG dokážeme, že platí

$$(r_i, r_j) = 0 \quad \text{a} \quad (p_i, Ap_j) = 0, \quad i \neq j \ ,$$

pro $r_j \equiv b - Ax_j$. A tedy z lokální $A$-ortogonality $p_k$ a $p_{k-1}$ plyne globální ortogonalita vektorů reziduí i globální $A$-ortogonalita všech směrových vektorů. Platí tedy předpoklad, který jsme vznesli pro odvození (1.12).

Na závěr ještě využijme vztahy

$$(p_{k-1}, r_{k-1}) = \|r_{k-1}\|^2$$

a

$$-Ap_{k-1} = \frac{r_k - r_{k-1}}{\gamma_{k-1}} = \frac{(r_k - r_{k-1})(p_{k-1}, Ap_{k-1})}{(p_{k-1}, r_{k-1})} \ .$$

Pak můžeme psát

$$\delta_k = \frac{\|r_k\|^2}{\|r_{k-1}\|^2} \ , \quad \gamma_{k-1} = \frac{\|r_{k-1}\|^2}{(p_{k-1}, Ap_{k-1})} \ .$$

## 1.3 Zápis CG

Algoritmus CG můžeme shrnout následujícím způsobem, viz [6].
Uvažujme soustavu lineárních rovnic

$$Ax = b \ , \tag{1.13}$$

kde $A \in \mathbb{R}^{n\times n}$ je symetrická pozitivně definitní matice, $b \in \mathbb{R}^n$.

Dáno $x_0,\ r_0 = b - Ax_0,\ p_0 = r_0$ a pro $j = 1, 2, \ldots$

$$\begin{aligned}
&\gamma_{j-1} = (r_{j-1}, r_{j-1})/(p_{j-1}, Ap_{j-1}) \ , \\
&x_j = x_{j-1} + \gamma_{j-1} p_{j-1} \ , \\
&r_j = r_{j-1} - \gamma_{j-1} Ap_{j-1} \ , \\
&\delta_j = (r_j, r_j)/(r_{j-1}, r_{j-1}) \ , \\
&p_j = r_j + \delta_j p_{j-1} \ .
\end{aligned} \tag{1.14}$$

## 1.4 Vlastnosti CG

Při odvozování CG jsme prohlásili, že naše volba $\delta_k$ nám zajistí vzájemnou ortogonalitu vektorů reziduí i vzájemnou $A$-ortogonalitu směrových vektorů. V následující větě shrneme tyto i některé další vlastnosti (viz [6], Theorem 5.1).

**Věta 2.** *Vektory reziduí $r_i$ a směrové vektory $p_j$ získané metodou CG splňují*

$$\begin{aligned}
(r_i, r_j) &= 0 \quad i \neq j \ , && (1.15)\\
(p_i, Ap_j) &= 0 \quad i \neq j \ , && (1.16)\\
(p_i, r_j) &= 0 \quad i < j \ , \quad (p_i, r_j) = \|r_i\|^2 \quad i \geq j \ , && (1.17)\\
(r_i, Ap_i) &= (p_i, Ap_i) \ , \quad (r_i, Ap_j) = 0 \quad i \neq j,\ i \neq j+1 \ . && (1.18)
\end{aligned}$$

*Důkaz.* Z iteračního předpisu pro $p_k$ v (1.14) se ukáže, že

$$p_k = \|r_k\|^2 \sum_{j=0}^{k} \frac{r_j}{\|r_j\|^2} \quad k = 0, 1, 2, \dots \ , \tag{1.19}$$

neboť

$$\begin{aligned}
p_k &= r_k + \delta_k p_{k-1} = r_k + \delta_k (r_{k-1} + \delta_{k-1} p_{k-2}) = \ \dots && (1.20)\\
&= r_k + \delta_k r_{k-1} + \delta_k \delta_{k-1} r_{k-2} + \ \dots \ + \delta_k \cdots \delta_1 r_0
\end{aligned}$$

a pro $k > j+1$ máme z definice $\delta_i$

$$\delta_k \cdots \delta_{j+1} = \frac{(r_k, r_k)}{(r_{k-1}, r_{k-1})} \cdot \frac{(r_{k-1}, r_{k-1})}{(r_{k-2}, r_{k-2})} \cdots \frac{(r_{j+1}, r_{j+1})}{(r_j, r_j)} = \frac{(r_k, r_k)}{(r_j, r_j)} = \frac{\|r_k\|^2}{\|r_j\|^2} \ .$$

Důkaz věty provedeme indukcí. Vektory $r_0$, $p_0 = r_0$ a $r_1$ vztahy splňují, neboť

$$(r_0, r_1) = (p_0, r_1) = (r_0, r_0) - \gamma_0 (r_0, Ap_0) = 0$$

z předpisu pro $r_1$ v (1.14). Předpokládejme, že vztahy (1.15) – (1.18) platí pro $r_0, \dots, r_k$ a $p_0, \dots, p_{k-1}$. Abychom ukázali, že věta zůstane v platnosti i po přidání $p_k$, musíme ukázat, že

$$\begin{aligned}
(r_i, p_k) &= \|r_k\|^2 \quad i \leq k \ , && (1.21)\\
(p_i, Ap_k) &= 0 \qquad i < k \ , && (1.22)\\
(r_k, Ap_i) &= (p_k, Ap_i) \ , \quad i \leq k,\ i \neq k-1 \ . && (1.23)
\end{aligned}$$

Platnost (1.21) plyne z (1.19) a indukčního předpokladu $(r_i, r_j) = 0$ pro $i \neq j$. Pro důkaz (1.22) užijeme iterační předpis pro $r_{i+1}$. Pak

$$(r_{i+1}, p_k) = (r_i, p_k) - \gamma_i (Ap_i, p_k) \ , \quad i < k.$$

Využitím (1.21) můžeme rovnici přepsat

$$\|r_k\|^2 = \|r_k\|^2 - \gamma_i (Ap_i, p_k) \ , \quad i < k.$$

Pokud $\gamma_i > 0$, tedy až do nalezení přesného řešení, (1.22) plyne ze symetrie $A$. Abychom ukázali (1.23), využijeme iteračního předpisu pro $p_{i+1}$ a indukčního předpokladu.

$$(p_k, Ap_i) = (r_k, Ap_i) + \delta_{k-1}(p_{k-1}, Ap_i) = (r_k, Ap_i) \quad \text{pro} \quad i \neq k-1 \ .$$

A tedy věta platí pro vektory $r_0, \ldots, r_k$ a $p_0, \ldots, p_k$. Zbývá nám ukázat, že k nim můžeme přidat i $r_{k+1}$. Proto musíme dokázat, že

$$(r_i, r_{k+1}) = 0 \quad i \leq k \ , \tag{1.24}$$

$$(Ap_i, r_{k+1}) = 0 \quad i < k \ , \tag{1.25}$$

$$(p_i, r_{k+1}) = 0 \quad i \leq k \ . \tag{1.26}$$

Z iteračního předpisu pro $r_{k+1}$ máme

$$(r_i, r_{k+1}) = (r_i, r_k) - \gamma_k (r_i, Ap_k) \ .$$

Jestliže $i < k$, (1.24) plyne z indukčního předpokladu, pro $i = k$ pak z definice $\gamma_k$. Podobně máme pro $i < k$

$$0 = (r_{k+1}, r_{i+1}) = (r_{k+1}, r_i) - \gamma_i (r_{k+1}, Ap_i) = -\gamma_i (r_{k+1}, Ap_i) \ ,$$

a tedy (1.25) platí. Konečně (1.26) plyne z (1.24) a (1.19). □

Z definice $p_j, r_j$ platí, že

$$\begin{aligned} \mathcal{K}_k(A, r_0) &= \text{span}\{r_0, Ar_0, \ldots, A^{k-1} r_0\} \\ &= \text{span}\{p_0, p_1, \ldots, p_{k-1}\} \\ &= \text{span}\{r_0, r_1, \ldots, r_{k-1}\}, \quad \forall k = 0, 1, \ldots \end{aligned}$$

$\mathcal{K}_k(A, r_0)$ se nazývá *k-tý Krylovův podprostor* generovaný $A$ a počátečním reziduem $r_0 = Ax_0 - b$. Vektory reziduí $\{r_0, r_1, \ldots, r_{k-1}\}$ tedy tvoří ortogonální bázi a směrové vektory $\{p_0, p_1, \ldots, p_{k-1}\}$ $A$-ortogonální bázi $\mathcal{K}_k(A, r_0)$. Při odvozování algoritmu CG jsme ukázali, že platí

$$\|x - x_k\|_A = \min_{u \in x_0 + \text{span}\{p_0, \ldots, p_{k-1}\}} \|x - u\|_A \ ,$$

a tedy

$$\|x - x_k\|_A = \min_{u \in x_0 + \mathcal{K}_k(A, r_0)} \|x - u\|_A \ . \tag{1.27}$$

# Kapitola 2

# CG v souvislostech

V této kapitole postupujeme podle [9], kde jsou uvedeny další prameny a odkazy na původní zdroje.

## 2.1 Lanczosova metoda

Jedno z možných použití Lanczosovy metody je nalezení ortonormální báze $\{v_1, v_2, \ldots, v_j\}$ Krylovova prostoru $\mathcal{K}_j(A, r_0)$. Posloupnost ortonormálních vektorů $v_1, v_2, \ldots$ generuje následovně:

Dáno $v_1 = r_0/\|r_0\|,\ \beta_1 \equiv 0$ a pro $j = 1, 2, \ldots$

$$\begin{aligned} \alpha_j &= (Av_j - \beta_j v_{j-1}, v_j) \ , \\ w_j &= Av_j - \alpha_j v_j - \beta_j v_{j-1} \ , \\ \beta_{j+1} &= \|w_j\| \ , \\ v_{j+1} &= w_j/\beta_{j+1} \ . \end{aligned} \tag{2.1}$$

Srovnáním CG (1.14) s (2.1) dostáváme

$$v_{j+1} = (-1)^j \frac{r_j}{\|r_j\|} \tag{2.2}$$

a také vztah mezi koeficienty

$$\begin{aligned} \alpha_j &= \frac{1}{\gamma_{j-1}} + \frac{\delta_{j-1}}{\gamma_{j-2}} \ , \qquad \delta_0 \equiv 0, \quad \gamma_{-1} \equiv 1, \\ \beta_{j+1} &= \frac{\sqrt{\delta_j}}{\gamma_{j-1}} \ , \end{aligned} \tag{2.3}$$

neboť eliminováním $p_{j-1}$ z iteračního předpisu pro $r_j$ v (1.14)

$$-\frac{1}{\gamma_{j-1}}\, r_j = Ar_{j-1} - \left(\frac{1}{\gamma_{j-1}} + \frac{\delta_{j-1}}{\gamma_{j-2}}\right) r_{j-1} + \frac{\delta_{j-1}}{\gamma_{j-2}}\, r_{j-2}$$

a využitím (2.2)

$$\frac{\sqrt{\delta_j}}{\gamma_{j-1}}\, v_{j+1} = Av_j - \left(\frac{1}{\gamma_{j-1}} + \frac{\delta_{j-1}}{\gamma_{j-2}}\right) v_j - \frac{\sqrt{\delta_{j-1}}}{\gamma_{j-2}}\, v_{j-1}\ .$$

Dále označme matici $V_j \equiv [v_1, \ldots, v_j] \in \mathbb{R}^{n\times j}$, která má jako sloupce Lanczosovy vektory $v_j$, a $T_j$ symetrickou tridiagonální matici s kladnými vedlejšími diagonálami

$$T_j = \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_j \\ & & \beta_j & \alpha_j \end{pmatrix} \tag{2.4}$$

Potom můžeme rovnice (2.1) přepsat v maticovém tvaru

$$AV_j = V_j T_j + \beta_{j+1} v_{j+1} e_j^T\,, \tag{2.5}$$

kde $e_j$ je $j$-tý sloupec identické matice $I \in \mathbb{R}^{j\times j}$. Vektory $v_j$ jsou navzájem ortogonální a za předpokladu $\mathcal{K}_n(A, r_0) = \mathbb{R}^n$ platí $v_{n+1} = 0$, a tedy

$$AV_n = V_n T_n\,. \tag{2.6}$$

Protože $x_j \in \mathcal{K}_j(A, r_0)$ a sloupce matice $V_j$ tvoří bázi $\mathcal{K}_j(A, r_0)$, existuje $y_j \in \mathbb{R}^j$ tak, že

$$x_j = x_0 + V_j\, y_j\,, \tag{2.7}$$

Z ortogonality $r_j$ a báze $\{v_1, \ldots, v_j\}$ prostoru $\mathcal{K}_j(A, r_0)$ dostáváme

$$\begin{aligned} 0 = V_j^T r_j = V_j^T(b - Ax_j) = V_j^T(r_0 - AV_j y_j) = \\ = e_1\|r_0\| - V_j^T AV_j y_j = e_1\|r_0\| - T_j y_j\,. \end{aligned}$$

Tedy CG aproximace $x_j$ je určena i vztahem (2.7) a řešením

$$T_j y_j = \|r_0\| e_1\,. \tag{2.8}$$

## 2.2 Ortogonální polynomy a Gaussova kvadratura

Využijeme-li algoritmického zápisu CG (1.14) a ortogonality reziduí, můžeme $j$-tou chybu (resp. reziduum) zapsat jako polynom matice $A$ aplikovaný na počáteční chybu (resp. reziduum), tedy

$$x - x_j = \varphi_j(A)(x - x_0)\,, \quad r_j = \varphi_j(A)(r_0)\,, \quad \varphi_j \in \Pi_j\,, \tag{2.9}$$

kde $\Pi_j$ označuje třídu polynomů $\varphi$ stupně nejvýše $j$ takových, že $\varphi(0) = 1$. Protože matice $A$ je symetrická pozitivně definitní, lze ji rozložit na tvar

$$A = U\Lambda U^T\,, \quad UU^T = U^TU = I\,, \tag{2.10}$$

kde $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ a $U = [u_1, \ldots, u_n]$ je matice obsahující ve sloupcích normalizované vlastní vektory $A$. Použijeme-li (1.27), (2.9) a (2.10), dostáváme

$$\begin{aligned} \|x - x_j\|_A &= \|\varphi_j(A)(x - x_0)\|_A = \min_{\varphi\in\Pi_j} \|\varphi(A)(x - x_0)\|_A \\ &= \min_{\varphi\in\Pi_j} \|\varphi(A)(r_0)\|_{A^{-1}} \\ &= \min_{\varphi\in\Pi_j} \left\{ \sum_{i=1}^{n} \frac{(r_0, u_i)^2}{\lambda_i} \varphi^2(\lambda_i) \right\}^{1/2} . \end{aligned} \tag{2.11}$$

Podobně je Lanczosův vektor $v_{j+1}$ určen nějakým monickým polynomem $\psi_j$ tak, že

$$v_{j+1} = \psi_j(A)v_1 \cdot \frac{1}{\beta_2\beta_3 \ldots \beta_{j+1}}\,. \tag{2.12}$$

Stejně jako v (2.11) s užitím ortogonality $v_{j+1}$ a $v_1, v_2, \ldots, v_j$ je $\psi_j$ určeno podmínkou

$$\|\psi_j(A)v_1\| = \min_{\psi\in\mathcal{M}_j} \|\psi(A)v_1\| = \min_{\psi\in\mathcal{M}_j} \left\{ \sum_{i=1}^{n} (v_1, u_i)^2 \psi^2(\lambda_i) \right\}^{1/2}, \tag{2.13}$$

kde $\mathcal{M}_j$ označuje třídu monických polynomů stupně $j$.

Uvažujeme-li CG nebo Lanczosovu metodu, máme tedy jednoznačně určenou posloupnost monických polynomů $1, \psi_1, \psi_2, \ldots$, které jsou ortogonální vzhledem ke skalárnímu součinu

$$(f, g) = \sum_{i=1}^{n} \omega_i f(\lambda_i) g(\lambda_i)\,,$$

kde váhy $\omega_i$ jsou určeny takto

$$\omega_i = (v_1, u_i)^2\,, \quad \sum_{i=1}^{n} \omega_i = 1\,. \tag{2.14}$$

Budeme-li pro jednodušší zápis předpokládat, že všechna vlastní čísla $A$ jsou různá a uspořádána v rostoucím pořadí (tj. $\lambda_i < \lambda_{i+1}$), můžeme definovat po částech konstantní neklesající distribuční funkci $\omega(\lambda)$ se skoky v bodech $\lambda_1, \ldots, \lambda_n$. Uvažujme k ní odpovídající Riemann-Stieltjesův integrál

$$\int_a^b f(\lambda) d\omega(\lambda) = \sum_{i=1}^{n} \omega_i f(\lambda_i)\,, \tag{2.15}$$

pro $a < \lambda_1 < \ldots < \lambda_n < b$. Pak můžeme (2.13) přepsat jako

$$\psi_j = \arg\min_{\psi \in \mathcal{M}_j} \left\{ \int_a^b \psi^2(\lambda) d\omega(\lambda) \right\}, \quad j = 0, 1, \ldots, n\,.$$

Dříve jsme ukázali, že CG (resp. Lanczosova metoda) začínající s vektorem $\|r_0\| v_1$ (resp. $v_1$) určuje symetrickou tridiagonální matici $T_j$. Podobně jako v (2.10) ji můžeme rozložit na součin

$$T_j = S_j \Theta_j S_j^T\,, \quad S_j^T S_j = S_j S_j^T = I\,,$$

kde $\Theta_j = \mathrm{diag}(\theta_1^{(j)}, \ldots, \theta_j^{(j)}), S_j = [s_1^{(j)}, \ldots, s_j^{(j)}]$.

Budeme-li nyní uvažovat CG aplikovanou na úlohu $T_j y_j = \|r_0\| e_1$ s počátečním reziduem $\|r_0\| e_1$ (respektive Lanczosovu metodu na matici $T_j$ a počáteční vektor $e_1$), můžeme, stejně jako v předešlé úvaze pro $A$, zkonstruovat váhové funkce $\omega_i^{(j)}$ a distribuční funkci $\omega^{(j)}$ tak, že

$$\omega_i^{(j)} = (e_1, s_i^{(j)})^2\,, \quad \sum_{i=1}^{j} \omega_i^{(j)} = 1\,.$$

Potom pro $a < \theta_1^{(j)} < \ldots < \theta_j^{(j)} < b$ je prvních $j$ vektorů z posloupnosti $\{1, \psi_1, \ldots, \psi_n\}$ určeno podmínkou

$$\psi_l = \arg\min_{\psi \in \mathcal{M}_l} \left\{ \int_a^b \psi^2(\lambda) d\omega^{(j)}(\lambda) \right\}, \quad l = 1, 2, \ldots, j\,.$$

Integrál

$$\int_a^b f(\lambda) d\omega^{(j)}(\lambda) = \sum_{i=1}^{j} \omega_i^{(j)} f(\theta_i^{(j)}) \tag{2.16}$$

je pak $j$-tou aproximací integrálu (2.15) pomocí Gaussových kvadratur. Tedy distribuční funkce $\omega^{(1)}(\lambda), \omega^{(2)}(\lambda), \ldots, \omega^{(l)}(\lambda), \ldots$ aproximují původní distribuční funkci $\omega(\lambda)$ ve smyslu Gaussových kvadratur a jsou tedy přesné pro polynomy stupně $2l-1$.

Nyní se podívejme na náš problém z druhé strany. Nechť máme $A$ a $r_0$. Pak Riemann-Stieltjesovy integrály (2.15) a jeho aproximace (2.16) pomocí Gaussových kvadratur jsou pro $j = 1, 2, \ldots, n$ jednoznačně určeny. Pak distribuční funkce $\omega^{(j)}$ jednoznačně určuje matici $T_j$ a ta s (2.7) a (2.8) určuje CG aproximace $x_j$.

Z (2.11) a (2.15) pro funkci $f(\lambda) = \lambda^{-1}$ máme

$$\|x - x_0\|_A^2 = \|r_0\|^2 \sum_{i=1}^{n} \frac{\omega_i}{\lambda_i} = \|r_0\|^2 \int_a^b \lambda^{-1} d\omega(\lambda)\,, \tag{2.17}$$

a použitím (2.6)

$$\|x - x_0\|_A^2 = (r_0, A^{-1} r_0) = \|r_0\|^2 (e_1, T_n^{-1} e_1) \equiv \|r_0\|^2 (T_n^{-1})_{11}\,. \tag{2.18}$$

Tedy

$$\int_a^b \lambda^{-1} d\omega(\lambda) = (T_n^{-1})_{11}\,. \tag{2.19}$$

Zcela analogicky (pro úlohu $T_j y_j = \|r_0\| e_1$ dimenze $j$)

$$\int_a^b \lambda^{-1} d\omega^{(j)}(\lambda) = (T_j^{-1})_{11}\,. \tag{2.20}$$

Označíme-li $R_j(f)$ chybu $j$-té aproximace funkce $f$ Gaussovými kvadraturami, píšeme

$$\int_a^b f(\lambda) d\omega(\lambda) = \int_a^b f(\lambda) d\omega^{(j)}(\lambda) + R_j(f)\,. \tag{2.21}$$

Z předchozího tedy platí pro $f(\lambda) = \lambda^{-1}$

$$\|x - x_0\|_A^2 = \|r_0\|^2 (T_n^{-1})_{11} = \|r_0\|^2 (T_j^{-1})_{11} + \|r_0\|^2 R_j(\lambda^{-1})\,. \tag{2.22}$$

V [5] bylo ukázáno, že

$$R_j(\lambda^{-1}) = \frac{\|x - x_j\|_A^2}{\|r_0\|^2}\,.$$

Potom

$$\|x - x_0\|_A^2 = \|r_0\|^2 (T_j^{-1})_{11} + \|x - x_j\|_A^2\,. \tag{2.23}$$

Použitím (2.8) a úpravami

$$\begin{aligned} \|r_0\|^2 (T_j^{-1})_{11} &= \|r_0\| e_1^T T_j^{-1} e_1 \|r_0\| \\ &= \|r_0\| v_1^T V_j T_j^{-1} e_1 \|r_0\| = (\|r_0\| v_1)^T (V_j T_j^{-1} e_1 \|r_0\|) \\ &= r_0^T (x_j - x_0)\,. \end{aligned} \tag{2.24}$$

Zde jsme využili globální ortonormality $v_1, \ldots, v_j$, díky ní můžeme psát

$$e_1^T = v_1^T V_j\,.$$

Dostáváme tedy vztah pro $A$-normu chyby

$$\|x - x_0\|_A^2 = r_0^T (x_j - x_0) + \|x - x_j\|_A^2\,.$$

# Kapitola 3

# Odhady chyb

## 3.1 Odhady $A$-normy chyby a jejich srovnání

### 3.1.1 Dolní odhady

Použitím Gaussových kvadratur jsme v předchozí kapitole ukázali, že platí

$$\|x - x_0\|_A^2 = \|r_0\|^2 (T_j^{-1})_{11} + \|x - x_j\|_A^2$$

a vyjádřením členu $\|r_0\|^2 (T_j^{-1})_{11}$ s použitím vzájemné ortogonality mezi Lanczosovými vektory $v_1, \ldots, v_j$ a maticového zápisu (2.1)

$$\|x - x_0\|_A^2 = r_0^T (x_j - x_0) + \|x - x_j\|_A^2 \,. \tag{3.1}$$

Jednoduchými algebraickými operacemi lze odvodit podobný, matematicky ekvivalentní vztah (viz [9])

$$\begin{aligned}
(x - x_0)^T A (x - x_0) &= (x - x_j + x_j - x_0)^T A (x - x_0) \\
&= (x - x_j)^T A (x - x_0) + (x_j - x_0)^T A (x - x_0) \\
&= (x - x_j)^T A (x - x_j + x_j - x_0) + (x_j - x_0)^T r_0 \\
&= \|x - x_j\|_A^2 + (x - x_j)^T A (x_j - x_0) + r_0^T (x_j - x_0) \\
&= \|x - x_j\|_A^2 + r_j^T (x_j - x_0) + r_0^T (x_j - x_0) \,,
\end{aligned}$$

a tedy

$$\|x - x_0\|_A^2 = r_j^T (x_j - x_0) + r_0^T (x_j - x_0) + \|x - x_j\|_A^2 \,. \tag{3.2}$$

Už Hesteines a Stiefel v [6] uvedli vztah

$$\|x - x_{i-1}\|_A^2 - \|x - x_i\|_A^2 = \gamma_{i-1} \|r_{i-1}\|^2 \,.$$

Jeho odvození je snadné a využívá pouze lokální $A$-ortogonalitu

$$\begin{aligned}\|x-x_{i-1}\|_A^2 - \|x-x_i\|_A^2 &= \|x-x_i+x_i-x_{i-1}\|_A^2 - \|x-x_i\|_A^2 \\ &= \|x_i-x_{i-1}\|_A^2 + 2(x-x_i)^T A(x_i-x_{i-1}) \\ &= \gamma_{i-1}^2 p_{i-1}^T A p_{i-1} + 2r_i^T(x_i-x_{i-1}) \\ &= \gamma_{i-1}\|r_{i-1}\|^2 \ .\end{aligned}$$

Rekurzí

$$\|x-x_0\|_A^2 = \sum_{i=0}^{j-1} \gamma_i \|r_i\|^2 + \|x-x_j\|_A^2 \ . \tag{3.3}$$

V první kapitole jsme zdůvodnili, proč potřebujeme při výpočtech sledovat $A$-normu chyby $\|x-x_j\|_A$. Z (3.1), (3.2) a (3.3) dostáváme pro přirozené $d$ ekvivalentní vztahy:

$$\|x-x_j\|_A^2 = r_0^T(x_{j+d}-x_j) + \|x-x_{j+d}\|_A^2 \ , \tag{3.4}$$

$$\|x-x_j\|_A^2 = r_0^T(x_{j+d}-x_j) - r_j^T(x_j-x_0) + r_{j+d}^T(x_{j+d}-x_0) + \|x-x_{j+d}\|_A^2 \ , \tag{3.5}$$

$$\|x-x_j\|_A^2 = \sum_{i=j}^{j+d-1} \gamma_i \|r_i\|^2 + \|x-x_{j+d}\|_A^2 \ . \tag{3.6}$$

Víme, že $A$-norma chyby v CG je klesající v každém kroku. Pokud zvolíme $d$ tak, že

$$\|x-x_j\|_A^2 \gg \|x-x_{j+d}\|_A^2 \, , \tag{3.7}$$

můžeme $\|x-x_{j+d}\|_A^2$ na pravé straně (3.4), (3.5) a (3.6) zanedbat a dostaneme dolní odhady pro $\|x-x_j\|_A^2$. Přesnost těchto odhadů je pak dána hodnotou $\|x-x_{j+d}\|_A^2$.

Označme

$$\mu_{j,d} \equiv r_0^T(x_{j+d}-x_j) \ , \tag{3.8}$$

$$\vartheta_{j,d} \equiv r_0^T(x_{j+d}-x_j) - r_j^T(x_j-x_0) + r_{j+d}^T(x_{j+d}-x_0) \ , \tag{3.9}$$

$$\nu_{j,d} \equiv \sum_{i=j}^{j+d-1} \gamma_i \|r_i\|^2 \ . \tag{3.10}$$

Teoreticky (v přesné aritmetice) platí

$$\mu_{j,d} = \vartheta_{j,d} = \nu_{j,d} \ .$$

Odhad $\mu_{j,d}$ je založen na globální ortonormalitě Lanczosových vektorů, která se v konečné aritmetice rychle ztrácí (viz [9], kap. 5, 6). Ve výpočtech se tedy může od ostatních odhadů velmi lišit. A to nikoli kvůli chybám ve výpočtu $\mu_{j,d}$, ale protože neplatí vztahy, které jsme použili při odvození (3.1).

Odhad $\vartheta_{j,d}$ obsahuje oproti $\mu_{j,d}$ navíc členy $r_j^T(x_j - x_0)$ a $r_{j+d}^T(x_{j+d} - x_0)$. Ty jsou v přesné aritmetice rovny nule, ale ve výpočtech mají opravný význam. Na druhou stranu zvyšují jeho náročnost a v aplikacích s proměnlivou volbou parametru $d$ musíme při změně $d$ přepočítávat celý odhad.

Na rozdíl od $\vartheta_{j,d}$ pro výpočet $\nu_{j,d}$ stačí hodnoty, které stejně počítáme během iterací CG ($\gamma_j, \|r_j\|^2$). Při jeho odvození jsme využili pouze lokální ortogonalitu a vidíme, že přepočítání odhadu při změně parametru $d$ je snadným přičítáním k již spočtené hodnotě. Navíc, jak je ukázáno v [9], odhad $\nu_{j,d}$ je numericky stabilní.

### 3.1.2 Horní odhady

Využitím Čebyšovových polynomů lze odvodit (viz například [8])

$$\|x - x_j\|_A \leq 2 \left( \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^j \|x - x_0\|_A \ . \tag{3.11}$$

Přestože je tento odhad často uváděn v souvislosti s metodou sdružených gradientů, konvergenci CG popisuje jen ve výjimečných případech. Pokud by totiž platilo

$$\|x - x_j\|_A \approx 2 \left( \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^j \|x - x_0\|_A \ , \quad j = 1, 2, \ldots n \ ,$$

pak

$$\frac{\|x - x_{j+1}\|_A}{\|x - x_j\|_A} \approx \left( \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right) \ , \quad j = 1, 2, \ldots n - 1 \ ,$$

a tedy konvergence CG by byla lineární. (3.11) totiž využívá pouze informace o krajích spektra $\lambda_{min}, \lambda_{max}$ matice $A$. Pro konvergenci metody sdružených gradientů je však určující i rozložení jednotlivých vlastních čísel $A$, jak můžeme vidět například ve vztahu (2.11).

Další horní odhady jsou uvedeny například v [4] v algoritmu CGQL, kde jsou odvozeny pomocí Gaussových, Gauss-Lobattových a Gauss-Radauových kvadratur.

## 3.2 Odhad euklidovské normy chyby

Přestože jsme doposud zdůrazňovali důležitost $A$-normy chyby, v některých aplikacích (jako například zpracování obrazu) je metoda sdružených gradientů použita pro řešení úlohy, kde potřebujeme minimalizovat $\|x - x_j\|$. Teoreticky v $n$-tém kroku CG dostáváme přesné řešení, pro které samozřejmě platí $\|x - x_n\| = 0$. Důležité je uvědomit si, že CG je navržena tak, aby byla klesající $A$-norma chyby. V článku [6] byl odvozen vztah

$$\|x - x_{j-1}\|^2 - \|x - x_j\|^2 = \frac{\|p_{j-1}\|^2}{\|p_{j-1}\|_A^2} \left( \|x - x_{j-1}\|_A^2 + \|x - x_j\|_A^2 \right) . \quad (3.12)$$

Podobnou úvahou jako v předchozí části pro přirozené $d$ dostáváme

$$\|x - x_j\|^2 = \sum_{i=j}^{j+d-1} \frac{\|p_i\|^2}{\|p_i\|_A^2} (\|x - x_i\|_A^2 + \|x - x_{i+1}\|_A^2) + \|x - x_{j+d}\|^2 , \quad (3.13)$$

za předpokladů

$$\|x - x_j\|^2 \gg \|x - x_{j+d}\|^2 \quad \text{a} \quad \|x - x_{j+d}\|_A^2 \gg \|x - x_{j+2d}\|_A^2 \quad (3.14)$$

a s použitím odhadu $\nu_{i,d}$ pro $\|x - x_i\|_A$

$$\|x - x_j\|^2 \approx \sum_{i=j}^{j+d-1} \frac{\|p_i\|^2}{\|p_i\|_A^2} \left( \gamma_i \|r_i\|^2 + 2 \sum_{k=i+1}^{j+2d-1} \gamma_k \|r_k\|^2 \right) . \quad (3.15)$$

Tento dolní odhad $\|x - x_j\|^2$ tak vyžaduje $2d$ iterací navíc.

## 3.3 Odhad v předpodmíněné metodě sdružených gradientů (PCG)

V této části postupujeme podle [10], kde lze nalézt i důkaz numerické stability níže uvedeného odhadu.

Předpodmíněnou metodou sdružených gradientů rozumíme aplikaci CG na předpodmíněnou úlohu

$$\hat{A}\hat{x} = \hat{b} , \quad (3.16)$$

$$\hat{A} = L^{-1} A L^{-T} , \quad \hat{b} = L^{-1} b , \quad (3.17)$$

kde $L$ je vhodná regulární (dolní trojúhelníková) matice. Předpodmiňovač (angl. *preconditioner*) $M \equiv LL^T$ je volen tak, aby soustava rovnic s maticí $M$ byla jednoduše řešitelná, zatímco matice $L^{-1}AL^{-T}$ by měla zajistit rychlou konvergenci CG (například tím, že je dobře podmíněná nebo má vhodně seskupená vlastní čísla). Algoritmus CG (1.14) pro úlohu $\hat{A}\hat{x} = \hat{b}$ je následující

Dáno $\hat{x}_0,\ \hat{r}_0 = b - \hat{A}\hat{x}_0,\ \hat{p}_0 = \hat{r}_0$ a pro $j = 1, 2, \ldots$

$$\begin{aligned}
&\hat{\gamma}_{j-1} = (\hat{r}_{j-1}, \hat{r}_{j-1})/(\hat{p}_{j-1}, \hat{A}\hat{p}_{j-1}) \ , \\
&\hat{x}_j = \hat{x}_{j-1} + \hat{\gamma}_{j-1}\hat{p}_{j-1} \ , \\
&\hat{r}_j = \hat{r}_{j-1} - \hat{\gamma}_{j-1}\hat{A}\hat{p}_{j-1} \ , \\
&\hat{\delta}_j = (\hat{r}_j, \hat{r}_j)/(\hat{r}_{j-1}, \hat{r}_{j-1}) \ , \\
&\hat{p}_j = \hat{r}_j + \hat{\delta}_j\hat{p}_{j-1} \ .
\end{aligned} \tag{3.18}$$

Označme

$$\begin{aligned}
&\gamma_j \equiv \hat{\gamma}_j \ , \quad \delta_j \equiv \hat{\delta}_j \ , \\
&x_j \equiv L^{-T}\hat{x}_j \ , \quad r_j \equiv L\,\hat{r}_j \ , \quad p_j \equiv L^{-T}\hat{p}_j \ , \quad s_j \equiv L^{-T}L^{-1}r_j = M^{-1}r_j \ .
\end{aligned} \tag{3.19}$$

Pak $x_j$ a $r_j$ odpovídají aproximaci řešení a reziduu původní úlohy $Ax = b$. Algoritmus PCG pro $Ax = b$ můžeme zapsat

Dáno $x_0,\ r_0 = b - Ax_0,\ s_0 = M^{-1}r_0,\ p_0 = r_0$ a pro $j = 1, 2, \ldots$

$$\begin{aligned}
&\gamma_{j-1} = (r_{j-1}, s_{j-1})/(p_{j-1}, Ap_{j-1}) \ , \\
&x_j = x_{j-1} + \gamma_{j-1}p_{j-1} \ , \\
&r_j = r_{j-1} - \gamma_{j-1}Ap_{j-1} \ , \\
&s_j = M^{-1}r_j \ , \\
&\delta_j = (r_j, s_j)/(r_{j-1}, s_{j-1}) \ , \\
&p_j = s_j + \delta_j p_{j-1} \ .
\end{aligned} \tag{3.20}$$

Pro aproximace řešení $\hat{x}_j$ soustavy (3.16) platí vztah (3.6), a tedy

$$\|\hat{x} - \hat{x}_j\|_{\hat{A}}^2 = \sum_{i=j}^{j+d-1} \hat{\gamma}_i \|\hat{r}_i\|^2 + \|\hat{x} - \hat{x}_{j+d}\|_{\hat{A}}^2 \ . \tag{3.21}$$

Použitím (3.19)

$$\|\hat{r}_i\|^2 = r_j^T L^{-T}L^{-1}r_j = r_j^T M^{-1}r_j = (r_j, s_j)$$

a

$$\|\hat{x} - \hat{x}_j\|_{\hat{A}}^2 = (L^T x - L^T x_j)^T L^{-1} A L^{-T} (L^T x - L^T x_j) = \|x - x_j\|_A^2$$

můžeme (3.21) přepsat jako

$$\|x - x_j\|_A^2 = \sum_{i=j}^{j+d-1} \gamma_i(r_i, s_i) + \|x - x_{j+d}\|_A^2 \ .$$

Za předpokladu na pokles normy (3.7) dostáváme dolní odhad pro $A$-normu chyby $j$-té aproximace řešení soustavy $Ax = b$ metodou PCG

$$\|x - x_j\|_A^2 \ \approx \ \hat{\nu}_{j,d} \equiv \sum_{i=j}^{j+d-1} \gamma_i(r_i, s_i) \ . \tag{3.22}$$

Podobně jako u odhadu $\nu_{j,d}$ jsou všechny potřebné hodnoty k dispozici už během iterací PCG.

## 3.4 Otevřené problémy

Naznačená teoretická hloubka metody sdružených gradientů je důvodem, proč stále zůstává mnoho otevřených problémů vztahujících se k CG. Uveďme alespoň ty, ke kterým jsme se dosud v této práci přiblížili.

Popsali jsme několik odhadů, které jsou odvozeny tak, že $\|x - x_j\|_A$ odhadneme pomocí $d$ iterací předpočítaných dopředu. Pro odhad předpokládáme splnění podmínky (3.7), neboli předpokládáme "dostatečný" pokles $A$-normy chyby mezi $j$-tou a $(j + d)$-tou iterací. Nabízí se tedy otázka, jak splnění (3.7) zajistit. V podstatě máme dvě možnosti:

1. volit $d$ velké,
2. v průběhu výpočtu ho měnit.

Například v [10] (str. 21–23) je uvedena úloha dimenze řádu $10^5$, kde v prvních zhruba 2300 iteracích $\|x - x_j\|_A$ prakticky neklesá. Vidíme tedy, že první možnost nemusí zaručit požadovanou přesnost. Adaptivní volba je samozřejmě výhodnější i v případě, že norma chyby klesá prudce a pro splnění podmínky (3.7) stačí pouze malá hodnota $d$. V současnosti však není popsáno žádné kritérium, podle kterého by se $d$ mělo volit. V další kapitole zkusíme na základě experimentů takové navrhnout.

Přestože jsme horní odhady pouze zmínili, nabízí se jejich využití současně s dolními odhady. Pokud by se dolní i horní odhad přibližovaly, získáváme informaci o jejich maximální možné chybě. Bohužel však v konečné aritmetice nemáme zaručeno, že horní odhad je skutečně horním odhadem a neprotne křivku konvergence.

Jednou z výhod odhadu $\nu_{j,d}$ je, že nám umožňuje snadno dopočítat velmi přesný odhad $A$-normy chyby v dřívějších iteracích. Provedeme-li například 600 kroků CG, pak v 300. iteraci jsme schopni dopočítat odhad $A$-normy chyby s hodnotou parametru $d = 300$. (Tato čísla jsou však pouze ilustrativní, v praktických výpočtech s vhodným předpodmíněním provádíme řádově pouze desítky iterací.) Můžeme se ptát na využití tohoto pozorování.

# Kapitola 4

# Experimenty

V experimentální části práce budeme nejprve zjišťovat, jak rozpoznat stagnaci $A$-normy chyby, která způsobuje nesplnění podmínky (3.7). Poté navrhneme heuristiku pro adaptivní volbu parametru $d$ a otestujeme její chování v několika příkladech. Budeme používat odhad $\nu_{j,d}$ pro jeho dobré numerické vlastnosti a snadnou implementaci v aplikacích s proměnlivou hodnotou $d$. Na závěr se zamyslíme nad rekonstrukcí křivky konvergence.

## Zápis podmínky (3.7)

V předchozích kapitolách jsme odvodili několik vztahů, které můžeme zjednodušeně zapsat jako

$$\|x - x_j\|_A^2 = o_{j,d} + \|x - x_{j+d}\|_A^2 \ ,$$

kde $o_{j,d}$ se liší podle toho, jaký vztah využíváme. Za předpokladu (3.7) jsme získali odhady

$$\|x - x_j\|_A^2 \approx o_{j,d} \ .$$

Podmínka (3.7) nám určuje přesnost odhadu, která je rovna $\|x - x_{j+d}\|_A^2$, a v podstatě znamená, že předpokládáme dostatečný pokles normy chyby mezi $j$-tou a $(j+d)$-tou aproximací řešení. Zřejmě ji můžeme zapsat jako

$$\frac{\|x - x_{j+d}\|_A^2}{\|x - x_j\|_A^2} \ll 1 \tag{4.1}$$

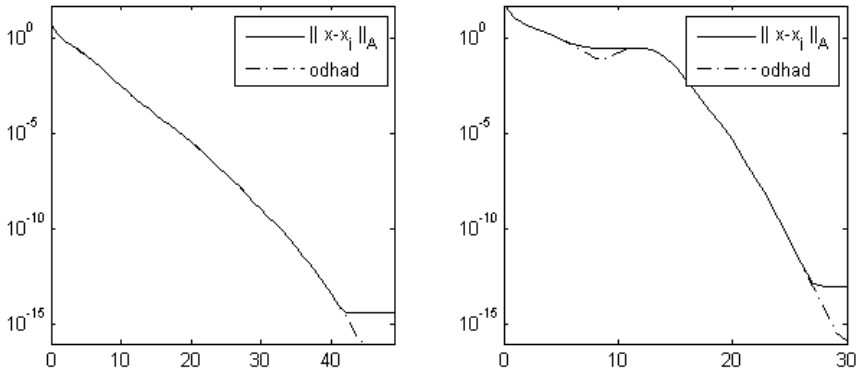
neboli

$$\frac{\|x - x_{j+d}\|_A}{\|x - x_j\|_A} \leq K \ . \tag{4.2}$$

pro předem zvolené $K$, $0 < K \ll 1$.

## 4.1 Stagnace $A$-normy chyby a volba parametru $d$

Problematiku volby parametru $d$ demonstrujme na úvod na dvou jednoduchých příkladech.




Obrázek 4.1: Odhad při klesající a stagnující $A$-normě chyby; matice `strakos`(48, 0.1, 1, 0.99), `matice`(48, 0.1, 100, 0.9)

Na obrázku (4.1) je vykreslen odhad ve dvou různých úlohách dimenze 48 s pevnou volbou parametru $d = 3$. Vlevo vidíme, že tato volba zajistí téměř přesnou aproximaci. To je zajištěno tím, že $A$-norma chyby stále klesá a podmínka (3.7) je splněna. Na grafu vpravo se odhad okolo osmé iterace skutečné hodnotě výrazněji vzdaluje, neboť v tomto úseku $A$-norma chyby stagnuje a (3.7) neplatí. Pro její splnění bychom museli volit $d$ větší, než je délka úseku.

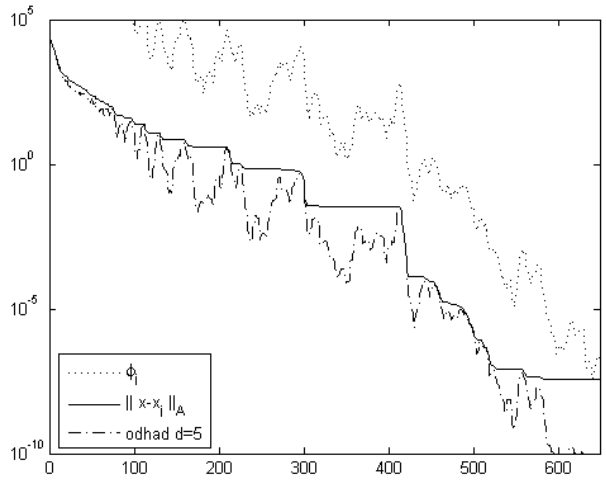
Pokud bychom byli schopni během výpočtu rozpoznat, že $A$-norma chyby stagnuje, pak bychom mohli zvětšováním $d$ dosáhnout přesnějších a spolehlivějších odhadů. Při zastavení CG totiž nemáme jistotu, že odhad odpovídá skutečné chybě vypočtené aproximace. (Pokud před zastavením CG bude $A$-norma chyby stagnovat, pak se může odhad lišit i o několik řádů.)

Zabývejme se nyní tedy tím, jak stagnaci $A$-normy poznat. V algoritmu CG používáme několik veličin, jejichž sledování je jednoduché. Bohužel se ukazuje, že norma rezidua, hodnota $\gamma_j$ i $A$-norma $p_j$ velmi často oscilují, a tedy jejich použití je možné pouze v podobě různých součtů, rozdílů atd. Na mnoha příkladech lze sledovat to, že součty norem reziduí definované

podobně jako odhad $A$-normy $\nu_{j,d}$, tj.

$$\phi_j \equiv \sum_{i=j}^{j+d-1} \|r_i\|,$$

poměrně přesně kopírují odhad $\|x - x_j\|_A$, jen jsou o něco posunuté (většinou větší, viz (4.2)). Tento postřeh nám bohužel neříká nic o chování $A$-normy chyby.

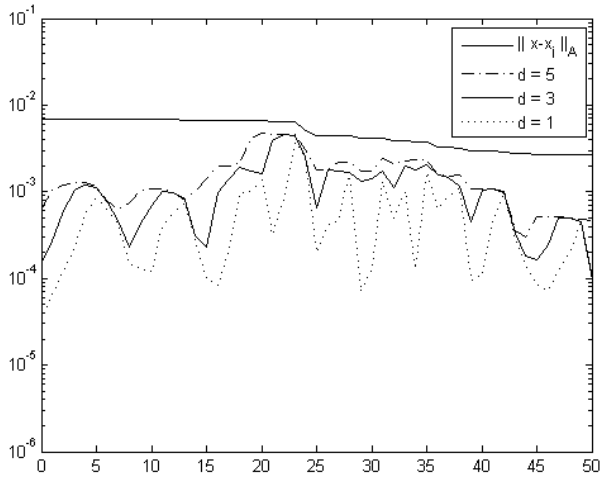



Obrázek 4.2: Chování $\phi_j$; matice `maticestred`(50, 2, 2)

Víme, že $A$-norma chyby je v každé iteraci klesající a jedním z ukazatelů toho, že podmínka (3.7) není splněna, tak může být případ, kdy dolní odhad roste. V každém okamžiku můžeme porovnat nově vypočtený odhad s odhady v předchozích iteracích a případně upravit hodnotu parametru $d$.

Dalším sledováním, které můžeme během výpočtu provádět, je velikost změny odhadu při zvyšování $d$. Z definice odhadu $\nu_{j,d}$ je zřejmé, že při jeho výpočtu můžeme sledovat i "horší" odhady $\nu_{j,d-1}, \nu_{j,d-2}, \ldots$ pro hodnoty parametru $d-1, d-2, \ldots$, jak je naznačeno na obrázku (4.3). Jestliže však v $j$-té iteraci roste rozdíl mezi $\nu_{j,d-i-1}$ a $\nu_{j,d-i}$ , pak roste i hodnota $\gamma_{j+d-i}\|r_{j+d-i}\|^2$, a tedy tato úvaha je ekvivalentní té provedené v předchozím odstavci.

Třetí možností, kterou popíšeme a na jejímž základě navrhneme heuristiku pro adaptivní volbu parametru $d$, je nahrazení hodnot $A$-normy chyby v podmínce (4.2) jejich odhadem. Nejjednodušším způsobem můžeme




Obrázek 4.3: Odhady pro různou hodnotu parametru $d$; matice `bcsstk01`

místo $\|x - x_{j+d}\|_A^2$ uvažovat $\gamma_{j+d}\|r_{j+d}\|^2$ a místo $\|x - x_j\|_A^2$ odhad $\nu_{j,d}$. Tedy chceme

$$\frac{\gamma_{j+d}\|r_{j+d}\|^2}{\nu_{j,d}} \leq H^2 \ , \quad 0 < H \ll 1 \tag{4.3}$$

a $d$ budeme zmenšovat, nebo zvětšovat při splnění, či nesplnění podmínky (4.3), zapsáno algoritmicky

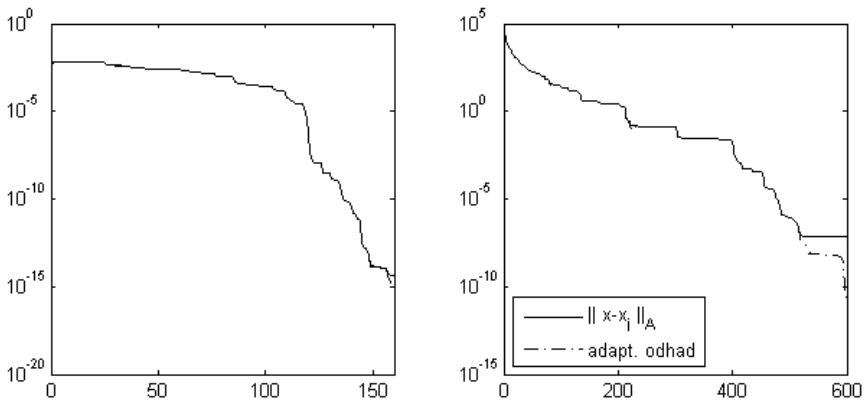
$$\begin{aligned} \textbf{if } \gamma_{j+d}\|r_{j+d}\|^2 \ &> \ \nu_{j,d} \cdot H^2 \ , \\ &\qquad d := d + 1 \\ \textbf{else} \\ &\qquad d := \max\{d - 1, d_{min}\} \\ \textbf{end} \end{aligned} \tag{4.4}$$

kde $d_{min}$ označuje minimální požadovanou hodnotu parametru $d$ (samozřejmě můžeme volit $d_{min} = 1$).

Zde je vhodné si rozmyslet, že tato úvaha má smysl. Je-li v porovnání s $\nu_{j,d}$ hodnota $\gamma_{j+d}\|r_{j+d}\|^2$ "velká", pak můžeme předpokládat, že $A$-norma chyby stagnuje, a je tedy potřeba zvětšovat $d$. (Neboť víme, že platí $\|x - x_{j+d}\|_A^2 > \gamma_{j+d}\|r_{j+d}\|^2$.) Naopak, jestliže je vzhledem k $\nu_{j,d}$ hodnota $\gamma_{j+d}\|r_{j+d}\|^2$ "malá", pak pravděpodobně klesá i $A$-norma chyby a stačí menší hodnota $d$.

## 4.2 Adaptivní volba $d$

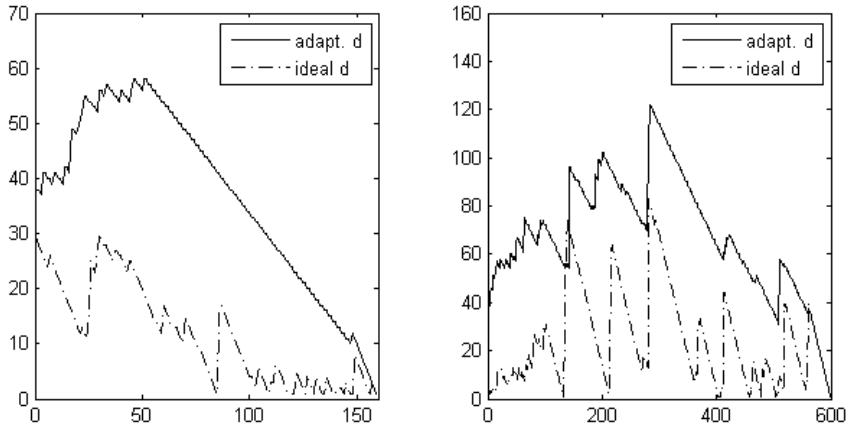
Věnujme se nyní výše popsanému algoritmu pro adaptivní volbu $d$. Problémem je volba parametru $H$. Horní odhad (3.11) jsme sice pro použití v CG nedoporučili, vidíme však z něj, že rychlost konvergence CG závisí na podmíněnosti matice $\kappa(A)$ tak, že dobře podmíněné úlohy konvergují rychle (obrácená implikace obecně neplatí). Nabízí se tedy využít této informace při volbě $H$, abychom zohlednili řešenou úlohu. V úlohách na obrázku (4.4)




Obrázek 4.4: Odhad při adaptivní volbě $d$; matice `bcsstk01`, `maticestred`(50, 2, 2)

jsme volili $H^2 = \kappa(A)^{-1/2}$. Tato volba se ukazuje být poměrně spolehlivá. Na druhou stranu je vykoupena mnohdy zbytečně velkou hodnotou parametru $d$. Za *ideální* hodnotu parametru $d$ označíme nejmenší hodnotu, která pro dané $K$ zajistí splnění (4.2). Přirozeně budeme chtít, aby heuristika (4.4) reagovala i na požadovanou "citlivost" danou $K$, a tedy budeme volit $H^2 = K^2 \cdot \kappa(A)^{-1/2}$. Pak můžeme srovnat adaptivní a ideální hodnotu $d$, viz obrázek (4.5). Z něj vidíme, že adaptivní volba $d$ je téměř vždy zbytečně velká, ale zejména v druhém případě příznivě reaguje na pokles i růst ideální hodnoty $d$.

Pro zmenšení hodnoty $d$ musíme zmírnit podmínku (4.3), neboli zvětšit parametr $H$. Toho lze dosáhnout mnoha způsoby, my v dalším budeme volit $\tilde{H}^2 = K^2 \cdot \kappa(A)^{-1/4}$. Srovnání použití $H$ a $\tilde{H}$ (při volbě $K = 0.4$) a chování heuristiky na několika náhodně vybraných čtvercových SPD maticích můžeme pozorovat na následujících obrázcích. Na grafu vlevo je vykreslen




Obrázek 4.5: Srovnání ideální a adaptivní volby $d$, $K = 0.6$; matice `bcsstk01`, `maticestred`(50, 2, 2)

odhad pro $\tilde{H}$ a skutečná hodnota $\|x - x_j\|_A$. (Použitím $H$ dostáváme přesnější odhad, který ale pro přehlednost nezakreslujeme.) Na pravém grafu pak můžeme srovnat adaptivní a ideální volbu $d$. Kromě úlohy `nos6` (obrázek (4.11)) na nich použitím $\tilde{H}$ dosahujeme přijatelných výsledků.

### 4.2.1 Možné modifikace

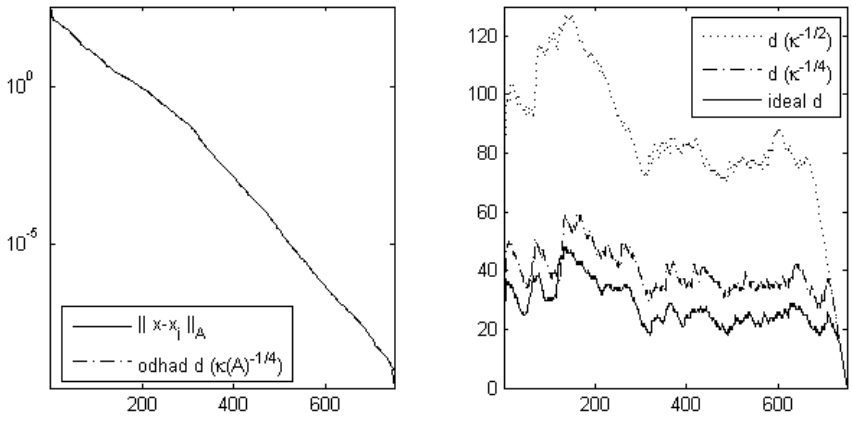
Před odvozením heuristiky (4.4) jsme navrhli nahrazení hodnot $A$-normy chyby v podmínce (4.2) jejich odhady. Pro odvození (4.3) jsme pak využili nejjednodušší odhad $\|x-x_{j+d}\|_A^2$, který samozřejmě lze zpřesnit. Uvažováním přirozeného $l$, můžeme podmínku (4.3) přepsat

$$\frac{\nu_{j+d+l,l}}{\nu_{j,d+l}} \leq H^2 \ , \quad 0 < H \ll 1 \ . \tag{4.5}$$

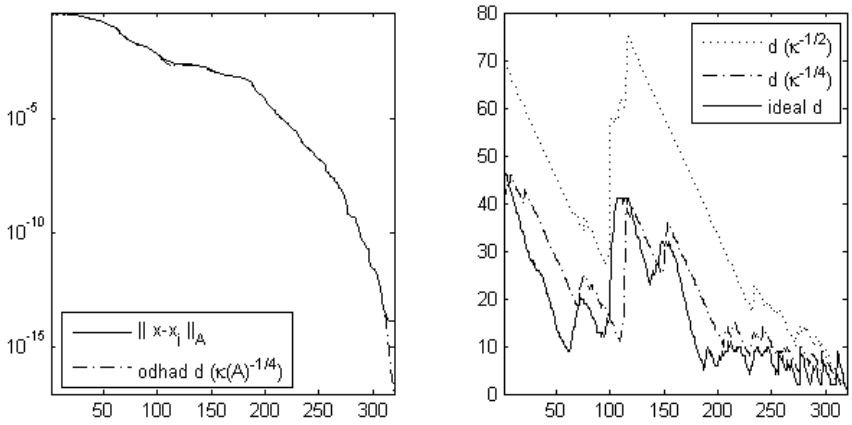
Pro praktické výpočty bude zajímavá volba $l \sim 5 - 10$ .

Také podobně jako v případě $d$ můžeme v průběhu výpočtu měnit i parametr $H$. Například jestliže odhad daný heuristikou (4.4) často roste (a tedy jak víme, neaproximuje $A$-normu chyby přesně), můžeme $H$ zmenšovat.

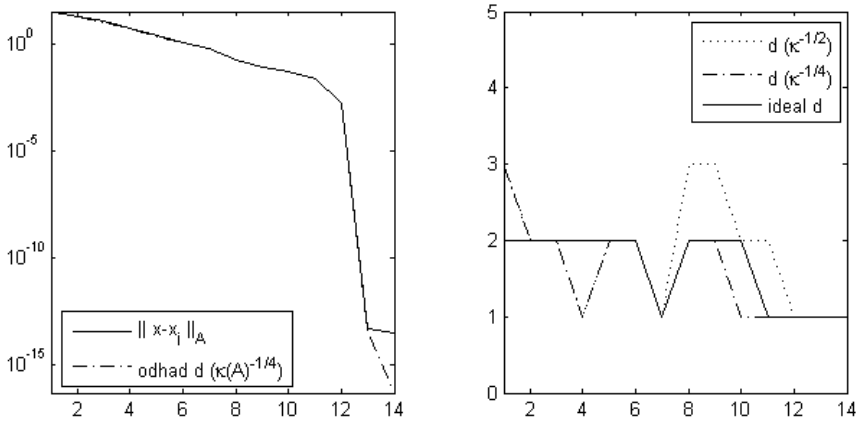
Využití těchto návrhů ponecháváme k dalšímu studiu.



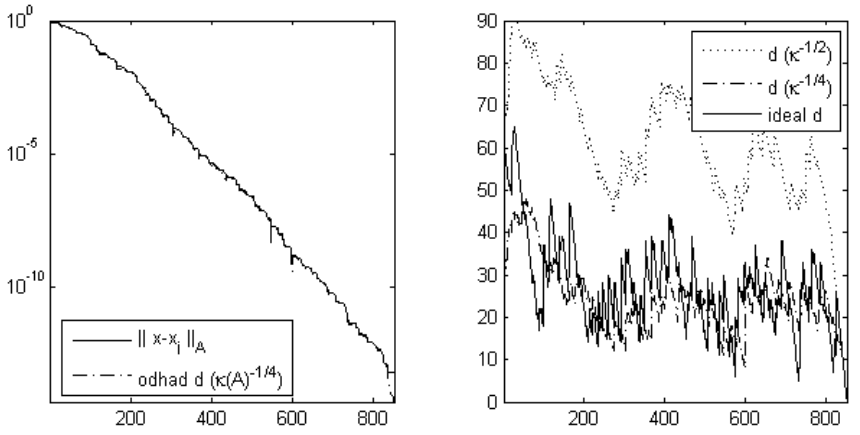

Obrázek 4.6: Srovnání ideální a adaptivní volby $d$; matice `662bus`



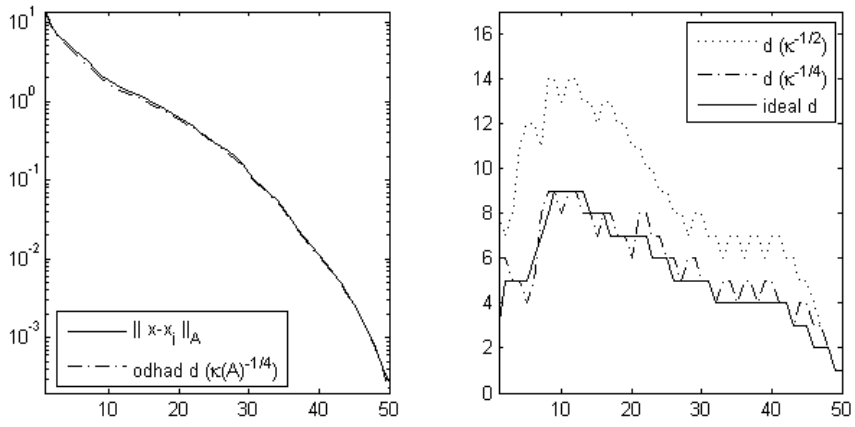

Obrázek 4.7: Srovnání ideální a adaptivní volby $d$; matice `bcsstk05`



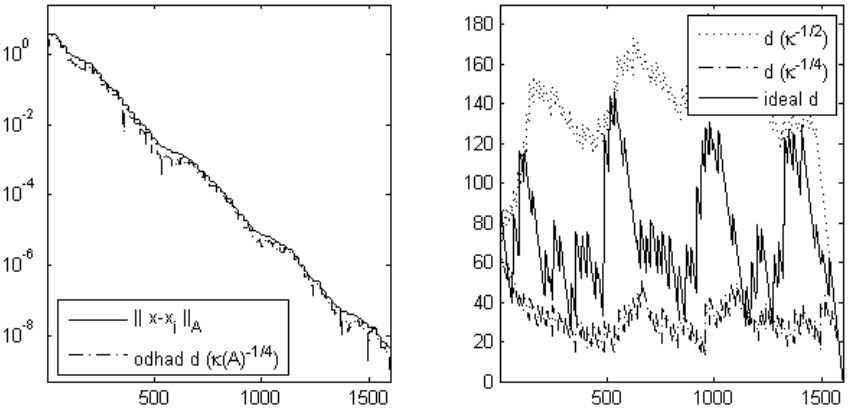

Obrázek 4.8: Srovnání ideální a adaptivní volby $d$; matice `bcsstm02`




Obrázek 4.9: Srovnání ideální a adaptivní volby $d$; matice `bcsstm19`




Obrázek 4.10: Srovnání ideální a adaptivní volby $d$; matice `gr3030`




Obrázek 4.11: Srovnání ideální a adaptivní volby $d$; matice `nos6`

## 4.3 Rekonstrukce křivky konvergence

Ve výpočtech nás často zajímá, jak vypadá křivka konvergence (například počítáme-li více soustav s různou pravou stranou $b$). Samozřejmě můžeme dopočítat $\|\tilde{x} - x_j\|_A\,,\ j = 1, 2, \ldots$ pro $\tilde{x}$ vypočtenou aproximaci přesného řešení $x$, ale existují i jednodušší postupy:

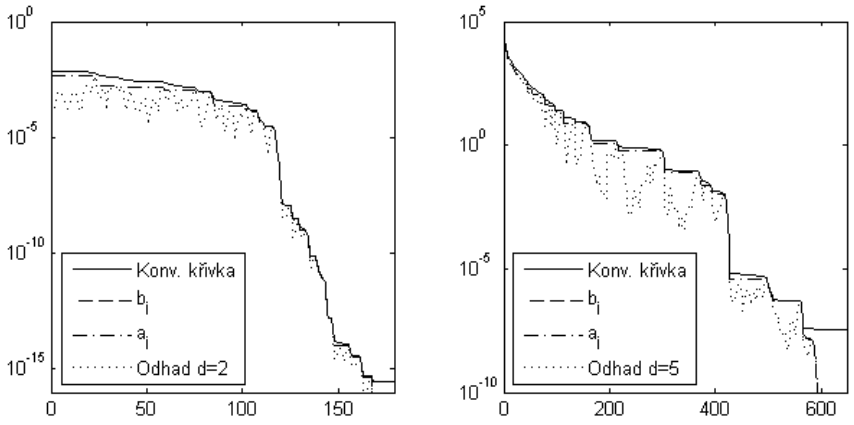
- protože víme, že $A$-norma chyby je nerostoucí, volíme v každé iteraci maximum z následujících odhadů, tedy

$$a_j^2 \equiv \max_{k \geq j} \{\nu_{k,d}\}\ ,$$

nebo

- v $j$-tém kroku CG přičteme hodnotu $\gamma_j \|r_j\|^2$ ke *všem* odhadům v předchozích iteracích, pak

$$b_j^2 \equiv \sum_{i=j}^{n} \gamma_i \|r_i\|^2 \quad (= \nu_{j,n-j})\ .$$




Obrázek 4.12: Rekonstrukce křivky konvergence; matice `bcsstk01`, `maticestred`(50, 2, 2)

Oba postupy nám zřejmě dávají dolní odhad a jejich srovnání můžeme pozorovat na obrázku (4.12). Je vidět, že zatímco $a_j$ se v delších úsecích stagnace $A$-normy chyby výrazněji liší od křivky konvergence, $b_j$ ji aproximuje téměř přesně. Navíc ve srovnání s časovou a výpočetní náročností CG je tato rekonstrukce křivky konvergence prakticky zadarmo.

# Závěr

V práci jsme se seznámili s metodou sdružených gradientů, ukázali význam minimalizace energetické normy chyby a odvodili CG s využitím minimalizace kvadratického funkcionálu. Ukázali jsme souvislosti s Lanczosovou metodou a teorií ortogonálních polynomů a využili je při odvození odhadu energetické normy chyby. Přehledovou část jsme zakončili odvozením dalších dolních odhadů energetické normy chyby, jejich srovnáním a uvedením některých souvislostí.

V experimentální části jsme sice nenalezli odpověď na otázku, jak zajistit splnění podmínky (3.7), ale uvedli jsme heuristiku, která dosahuje v příkladech zajímavých výsledků. Volba parametru $H$, resp. $\tilde{H}$, však není teoreticky zdůvodněna. Použili jsme pro ni číslo podmíněnosti $\kappa(A)$, které obecně v úlohách neznáme, ale jsme ho schopni díky souvislosti CG s Lanczosovou metodou poměrně přesně aproximovat již během několika iterací CG.

Cílem dalšího studia stále zůstává volba parametru $H$, další testování heuristiky (například na mnohodimenzionálních úlohách s předpodmíněním) a především její navržené modifikace. Pozornost si jistě zaslouží i podobnost křivky odhadu a součtu reziduí (které jsme označili jako $\phi_j$).

# Použité matice

Následující funkce generují podle zadaných parametrů diagonální matice s různě uspořádanými vlastními čísly, $n$ označuje dimenzi matice.

- `matice`$(n,\ a,\ b,\ \delta)$

  $\lambda_1 = a,\ \tilde{\lambda}_i = \tilde{\lambda}_{i-1} + \delta^i, \quad i = 2, \ldots, n$
  $\lambda_i = a + (b-a)\tilde{\lambda}_i/\tilde{\lambda}_n, \quad i = 2, \ldots, n$

  Vlastní čísla leží v intervalu $[a, b]$, jsou soustředěna u $b$. Pro větší hodnotu parametru $\delta$ jsou rozptýlena více.

- `maticestred`$(n,\ S,\ \phi)$

  $\lambda_i = -((S-i)^5 \cdot \phi) + S^2, \quad i = 1, 2, \ldots, n$

  Vlastní čísla rostou v páté mocnině jejich indexu. V příkladech používané hodnoty $S$ a $\phi$ zajišťují, že matice je SPD.
  Pro $A =$ `maticestred`$(50,\ 2,\ 2)$ je $\kappa(A) = 2.5e + 08$ .

- `strakos`$(n,\ a,\ b,\ \rho)$: viz například [9], kap. 5

  $\lambda_1 = a,\ \lambda_n = b,$
  $\lambda_i = \lambda_1 + \frac{i-1}{n-1}(\lambda_n - \lambda_1)\rho^{n-i}, \quad i = 2, \ldots, n$

  Vlastní čísla leží v intervalu $[a, b]$ a jsou soustředěna u $a$. Pro větší hodnotu parametru $\rho$ jsou rozptýlena více.
  Pro obrázek (4.1) jsme volili $\rho = 0.99$. Vlastní čísla jsou pak rozložena prakticky rovnoměrně a pokles $A$-normy chyby je stálý a téměř lineární.

Zdrojem následujících matic byl MATRIX MARKET,

`http://math.nist.gov/MatrixMarket/`

- `bcsstk01` : $n = 48$, $\kappa = 1.6e + 06$

- `662bus` : $n = 662$, $\kappa = 8.3e+05$
- `bcsstk05` : $n = 153$, $\kappa = 3.5e+04$
- `bcsstm02` : diagonální, $n = 66$, $\kappa = 8.8$
- `bcsstm19` : diagonální, $n = 817$, $\kappa = 2.3e+05$
- `gr3030` : $n = 900$, $\kappa = 3.8e+02$
- `nos6` : $n = 675$, $\kappa = 8e+06$

U diagonálních matic jsme položili $A \equiv QLQ^T$, kde $Q$ je ortogonální matice získaná Matlab QR-rozkladem náhodné čtvercové matice (v Matlabu `QR(randn(n))`) a $L$ daná diagonální matice.

# Literatura


[1] Brenner S. C., Scott L. R. (1994): *The mathematical theory of finite elements methods*, Springer - Verlag, kap. 0, 1.

[2] Dahlquist G., Golub G. H., Nash S. G. (1978): *Bounds for the error in linear systems*, Springer, str. 154 – 172.

[3] Elman H. C., Silvester D. J., Wathen A. J. (2005): *Finite elements and fast iterative solvers; with application in incompressible fluid dynamics*, Numerical Mathematics and Scientific Computation, Oxford University Press, kap. 2.

[4] Golub G. H., Meurant G. (1997): *Matrices, moments and quadratures II: How to compute the norm of the error in iterative methods*, BIT 37, str. 687 – 705.

[5] Golub G. H., Strakoš Z. (1994): *Estimates in quadratic formulas*, Numer. Algorithms 8, str. 253 – 254.

[6] Hestenes M. R., Stiefel E. (1952): *Methods of conjugate gradient for solving linear systems*, J. Research Nat. Bur. Standarts 49, str. 409 – 435.

[7] Liesen J., Strakoš Z.: *Principles of Krylov subspace methods*, kap. 6, (zatím nepublikováno).

[8] Saad, Y. (2003): *Iterative methods for sparse linear systems*, SIAM, kap. 6.11.

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

[10] Strakoš Z., Tichý P. (2005): *Error estimation in preconditioned conjugate gradients*, BIT Numerical Mathematics 45, str. 789 – 817.

[11] Watkins D. S. (2002): *Fundamentals of Matrix Computation*, John Wiley & Sons, str. 562 – 567.