Univerzita Karlova v Praze
Matematicko - fyzikální fakulta

Disertační práce

2004 Ctirad Matonoha

Matematicko - fyzikální fakulta
Univerzity Karlovy, Praha

Disertační práce

Numerická realizace metod s lokálně omezeným krokem

Ctirad Matonoha

Školitel: doc. RNDr. Jan Zítko, CSc.
Konzultant: doc. Ing. Ladislav Lukšan, DrSc.
Obor: Vědecko-technické výpočty

Abstrakt

Disertační práce se zabývá odvětvím matematické optimalizace, a sice metodami s lokálně omezeným krokem. Pod pojmem optimalizace rozumíme hledání (lokálního) minima nelineární, tzv. účelové funkce, a to buď bez omezení nebo na množině omezení dané rovnostmi a nerovnostmi. Řeší se rozsáhlé problémy s velmi obecnou strukturou. Metody s lokálně omezeným krokem mají iterační charakter, kde klíčovým problémem je volba směrového vektoru. Ten se hledá opět iteračně, čímž vzniká vnitřní iterační cyklus. Princip spočívá v tom, že se sestrojí jednoduchý model vystihující chování účelové funkce v okolí daného bodu, např. kvadratická aproximace, a hledá se minimum tohoto modelu.

Metody s lokálně omezeným krokem byly vyvinuty, aby se odstranily problémy s nekonvexností účelové funkce. Při vnitřním iteračním cyklu se minimalizuje obecně nekonvexní kvadratická funkce na oblasti (například na kouli) s proměnným poloměrem. Většina metod vede na řešení systému lineárních rovnic, což je neefektivní hlavně pro rozsáhlé optimalizační úlohy. Existují však iterační metody s postupně generovanými Krylovovými podprostory mající speciální strukturu, která nám dovolí řešit tento problém efektivně. Po konečném počtu kroků lze získat téměř optimální řešení, které splňuje podmínky optimality. Jiný přístup při řešení vnitřního cyklu spočívá v transformaci úlohy na parametrizovaný problém vlastních čísel. Hlavní výhodou metod s lokálně omezeným krokem jsou jejich výborné konvergenční vlastnosti.

V této práci je pojednáno o teorii metod s lokálně omezeným krokem. Jsou uvedeny jednotlivé typy těchto metod, jejich základní vlastnosti a algoritmus. Práce obsahuje nové algoritmy pro neomezenou minimalizaci a u každé metody je uveden původní důkaz globální konvergence. Je rovněž provedeno numerické porovnání metod na různých typech testovacích úloh.

Poslední kapitola se zabývá metodami vnitřních bodů nelineárního programování. Metody vnitřních bodů jsou efektivním nástrojem pro řešení obecných problémů nelineárního programování zejména tehdy, je-li úloha velká a strukturovaná (např. řídká nebo rozložitelná). V práci je studována metoda vnitřních bodů, kde se směrové vektory určují použitím metod s lokálně omezeným krokem a je provedena teoretická analýza takto sestrojeného postupu. Sestavené nové algoritmy jsou prověřeny řadou numerických testů.

Obsah

Úvod

Tato práce je věnována metodám s lokálně omezeným krokem pro řešení úloh lokální optimalizace. Pod pojmem optimalizace rozumíme minimalizaci nebo maximalizaci obecné účelové funkce

$$F : \mathbb{R}^n \to \mathbb{R}$$

bez omezení na proměnné nebo s obecnými omezujícími podmínkami ve tvaru rovností a nerovností. Protože maximum funkce F je minimem funkce $-F$, lze se omezit jen na případ minimalizace. Zabýváme se pouze lokální optimalizací, neboť hledání globálních extrémů vyžaduje odlišné prostředky založené spíše na diskrétní matematice a statistice než na numerické analýze. Předpokládáme, že funkce F je dvakrát spojitě diferencovatelná, čímž se vyhýbáme nehladkým úlohám, které lze sice také řešit metodami s lokálně omezeným krokem, ale jejich konkrétní forma a teoretické vlastnosti jsou založeny na odlišném přístupu využívajícím výsledky nehladké analýzy.

K lokální minimalizaci hladké účelové funkce se používají převážně iterační metody a to buď metody spádových směrů nebo metody s lokálně omezeným krokem [34]. Metody spádových směrů jsou implementačně jednodušší, neboť určují směrový vektor pomocí násobení matice vektorem nebo pomocí řešení soustavy lineárních rovnic. Jejich nevýhoda se projeví při použití Newtonovy metody pro minimalizaci nekonvexní funkce, kdy Hessova matice není pozitivně definitní, takže můžeme dostat směrové vektory, které nejsou spádové. Také v případě minimalizace součtu čtverců, kdy je matice normální soustavy rovnic velmi špatně podmíněná nebo dokonce singulární, můžeme dostat nevhodné směrové vektory, buď téměř kolmé ke gradientu minimalizované funkce nebo příliš velké v normě. Pro tyto účely byly vyvinuty metody s lokálně omezeným krokem, kdy je norma směrového vektoru omezená poloměrem oblasti přijatelnosti (*trust region*) a kdy indefinitnost matice modelu není na překážku.

Určení směrového vektoru v metodách s lokálně omezeným krokem je mnohem složitější úloha než ta, která se používá v metodách spádových směrů, neboť je třeba řešit (přesně nebo přibližně) minimalizační úlohu s kvadratickou účelovou funkcí a omezením na normu. Existuje proto celá řada metod, přesných i přibližných, složitých i jednodušších, které byly k tomuto účelu vyvinuty. Tato práce má za cíl vyhodnotit metody s lokálně omezeným krokem vhodné pro řešení rozsáhlých strukturovaných (např. řídkých nebo rozložitelných) úloh a nalézt pro tento případ nové metody, jejichž efektivita by byla vyšší než efektivita metod již publikovaných. V oblasti obecných úloh nelineárního programování (třetí kapitola) se pouštíme do problematiky, která je v počátcích svého rozvoje.

Metody s lokálně omezeným krokem [45] jsou iterační metody k nalezení lokálního minima funkce F bez omezení, případně s omezujícími podmínkami ve tvaru rovností a nerovností. Zvolíme počáteční bod y_0 a jestliže známe bod y_k, $k \geq 0$, a k němu příslušnou funkční hodnotu $F(y_k)$, určíme bod y_{k+1} následujícím způsobem. V bodě y_k

sestrojíme kvadratický model $\tilde{F}_k(x)$ funkce F (např. pomocí Taylorova rozvoje) a najdeme bod x_k, ve kterém dochází k největšímu poklesu funkce $\tilde{F}_k(x)$. Je-li kvadratická funkce $\tilde{F}_k(x)$ nekonvexní, je zdola neomezená, takže pokus najít bod největšího poklesu může buď selhat nebo dávat příliš velký krok x_k a model $\tilde{F}_k(x_k)$ nemusí dobře reprezentovat hodnotu $F(y_k + x_k)$. Proto zvolíme poloměr $\Delta_k > 0$ a bod x_k hledáme pouze v oblasti, kde $\|x_k\| \leq \Delta_k$. Pokud pokles předpověděný modelem odpovídá skutečnému poklesu funkce F v bodě $y_k + x_k$, položíme $y_{k+1} = y_k + x_k$. V opačném případě je krok x_k příliš velký, položíme $y_{k+1} = y_k$ a v další iteraci zmenšíme poloměr Δ_{k+1}. Takto postupujeme dále, dokud nenajdeme bod $y_\star$, který je lokálním minimem funkce F. Tento bod splňuje podmínky optimality, které jsou kriteriem zastavení algoritmu. Metoda s lokálně omezeným krokem tedy vede na řešení podproblému

$$\min_{x \in \mathbb{R}^n} \tilde{F}_k(x) \quad \text{vzhledem k} \quad \|x\| \leq \Delta_k$$

s pevným indexem k, který budeme při řešení tohoto podproblému vynechávat. Zvolíme-li jiný počáteční bod y_0, můžeme dostat jiné lokální minimum.

Jestliže se vyskytují omezení ve tvaru rovností, používají se metody s lokálně omezeným krokem jako efektivní nástroj pro určení směrového vektoru v metodách rekursivního kvadratického programování. V případě výskytu omezení ve tvaru nerovností se často, zejména pro rozsáhlé strukturované úlohy, používají metody vnitřních bodů, kdy se původní úloha převede na posloupnost úloh s pomocnými proměnnými vystupujícími v barierovém členu. Tyto úlohy obsahují pouze omezení ve tvaru rovností. Je tedy logické použít metody s lokálně omezeným krokem i v tomto případě.

První kapitola je úvodem do problematiky. Je zde popsána základní optimalizační metoda, uvedena myšlenka metod s lokálně omezeným krokem a jsou podrobněji studovány vlastnosti těchto metod v případě minimalizace bez omezení. Je uvedena definice obecné metody s lokálně omezeným krokem, popsán obecný algoritmus a stanoveny předpoklady kladené na funkci F, které zaručí globální a superlineární konvergenci tohoto algoritmu.

Metody s lokálně omezeným krokem lze rozdělit na dvě skupiny. Hledá se minimum funkce

$$\psi(x) = \frac{1}{2} x^T \mathbf{A} x + g^T x \quad \text{vzhledem k} \quad \|x\| \leq \Delta.$$

V prvním případě uvažujeme libovolné $x \in \mathbb{R}^n$, tzv. optimální lokálně omezený krok. Ve druhém případě, kdy jde pouze o přibližné řešení, se používají metody Krylovových podprostorů, které generují po částech lineární křivku. Iterační proces se ukončí v takovém bodě, ve kterém tato křivka překročí hranici oblasti zadané omezením $\|x\| \leq \Delta$. Získané minimum funkce $\psi(x)$, ať už optimální nebo přibližné, je označeno jako $x_\star$. V závěru první kapitoly jsou ukázány některé vlastnosti lokálně omezeného kroku.

Ve druhé kapitole se zabýváme jednotlivými metodami vycházejícími z principu lokálně omezeného kroku pro minimalizaci bez omezení [08].

Na začátku kapitoly vycházíme z nutných a postačujících podmínek, které charakterizují optimální lokálně omezený krok a provádíme podrobnou analýzu výpočtu zakončenou algoritmem.

Dále je studována metoda psí nohy (název pochází od jejího autora M.J.D.Powella) a metoda sdružených gradientů pro přibližný výpočet lokálně omezeného kroku. Jelikož metoda psí nohy používá právě jeden krok metody sdružených gradientů, zobecnili jsme tento postup tak, že používáme více kroků. Získané výsledky jsou původní a výsledný algoritmus získaný na základě těchto úvah se ukázal být účinným.

Při použití metod Krylovových podprostorů pro rozsáhlé úlohy dostaneme, jak známo, efektivní algoritmus použitím vhodného předpodmínění. Důkaz globální konvergence i pro tento případ patří mezi původní výsledky práce a je formulován ve větě 2.9.

Kromě minimalizace funkcionálu ψ s maticí $\mathbf{A}$ se v další části druhé kapitoly zabýváme minimalizací funkcionálu $\tilde{\psi}$ s třídiagonální maticí $\mathbf{T}$, která vznikne použitím Lanczosova procesu. Tato metoda je velmi účinná, má však jednu nevýhodu – nelze použít předpodmínění. I když Lanczosův proces lze předpodmiňovat jako každou jinou metodu Krylovových podprostorů, ztrácíme ortogonalitu původních vektorů báze a změní se omezení kvadratického podproblému. Proto jsou navrženy dvě nové původní metody, které jsou kombinací Lanczosova procesu a metody sdružených gradientů.

První metoda spočívá v tom, že se nejprve pomocí omezeného počtu kroků Lanczosova procesu sestaví kvadratický podproblém s třídiagonální maticí, který slouží k přibližnému určení parametru ξ definujícímu optimální lokálně omezený krok. Aproximace optimálního lokálně omezeného kroku se pak získá předpodmíněnou metodou sdružených gradientů aplikovanou na soustavu rovnic s maticí $\mathbf{A} + \xi\mathbf{I}$. Tato metoda byla porovnána s ostatními metodami pro výpočet lokálně omezeného kroku a ukázala se být jednou z nejefektivnějších (výsledky jsou uvedeny v § 2.8).

U druhé metody použijeme nejprve metodu sdružených gradientů, abychom získali třídiagonální matici $\mathbf{T}$ řádu m. Poté na základě hodnoty normy směrového vektoru buď zůstaneme u metody sdružených gradientů nebo řešíme kvadratický podproblém s maticí $\mathbf{T}$. Tuto metodu však nelze předpodmiňovat, neboť bychom změnili omezení kvadratického podproblému.

Jiný přístup k výpočtu optimálního lokálně omezeného kroku spočívá v parametrizaci celé úlohy, která převede výpočet na problém vlastních čísel. Provedli jsme analýzu celého postupu a dokázali globální konvergenci.

Původními výsledky druhé kapitoly jsou kromě tří nových algoritmů také důkazy globální konvergence všech vyšetřovaných metod. Poslední část této kapitoly se zabývá praktickým porovnáním těchto metod. Jsou testovány pomocí několika kolekcí testovacích problémů s velkým počtem proměnných. Hodnotili jsme celkový počet iterací, vyčíslení funkční hodnoty, gradientu a čas výpočtu. Na základě získaných výsledků prezentovaných formou tabulek a grafů je provedeno zhodnocení jednotlivých metod.

Třetí kapitola je věnována použití principu lokálně omezeného kroku v metodách vnitřních bodů pro obecné úlohy nelineárního programování (s omezeními ve tvaru rovností i nerovností). Nerovnosti se odstraní zavedením pomocných proměnných a použitím barierové funkce s parametrem μ, přičemž aproximace řešení dostaneme limitním přechodem pro $\mu \to 0$.

Podle způsobu konstrukce směrových vektorů můžeme metody vnitřních bodů [66] realizovat buď jako metody spádových směrů nebo jako metody s lokálně omezeným krokem. Aplikace metod s lokálně omezeným krokem pro obecná omezení patří mezi původní výsledky. Je vysvětleno, proč se lokálně omezený krok hledá ve dvou krocích jako součet vertikálního a horizontálního kroku (nekompatibilita lineárního omezení a omezení $\|x\| \le \Delta$) a jsou formulovány kvadratické podproblémy pro určení těchto dílčích kroků. Použitím vhodné lineární transformace lze docílit toho, že soustava rovnic použitá pro určení horizontálního kroku je stejná jako předpodmíněná soustava rovnic vyskytující se v metodách spádových směrů.

Dále je navržena nová rozšířená barierová Lagrangeova funkce sloužící k výpočtu poloměru oblasti přijatelnosti a délky kroku. Protože vyžadujeme kladnost pomocných proměnných, které nejsou zahrnuty v omezeních kvadratického podproblému, a také Lagrangeových multiplikátorů odpovídajících omezením ve tvaru nerovností, je třeba omezit příslušné délky kroku pomocí speciální strategie. Dále je popsán efektivní způsob aktualizace poloměru Δ a barierového parametru μ. Závěrem je uveden kompletní algoritmus metody vnitřních bodů s lokálně omezeným krokem pro obecnou úlohu nelineárního programování.

Zcela novou myšlenkou je použití jiných prostředků než pokutových funkcí k posouzení toho, zda je bod získaný v daném iteračním kroku přijatelný. Jde o použití tzv. filtru, jehož význam je popsán v závěru třetí kapitoly.

V poslední části práce jsou metody vnitřních bodů testovány a porovnány pomocí několika kolekcí testovacích problémů s velkým počtem proměnných. Výsledky jsou uvedeny ve formě tabulek a grafů podobným způsobem jako ve druhé kapitole.

Shromáždili a prostudovali jsme 69 prací k uvedené problematice. Kromě nových výsledků zmiňovaných výše jsme se pokusili o stručný přehled všech výsledků, kterých bylo dosaženo v problematice, která byla předmětem disertační práce. Naprogramovali jsme všechny známé metody a porovnali je s postupy, které jsme sami sestavili. Ke všem metodám ve druhé kapitole jsme připojili důkaz globální konvergence a výsledky dosažené v této práci byly použity ve společném článku [36], který byl přijat do tisku a vyjde v časopisu *Numerical Linear Algebra with Applications* v roce 2004.

Kapitola 1

Úloha nelineárního programování

V první kapitole se seznámíme s problémem minimalizace funkce F, definujeme úlohu nelineárního programování a uvedeme podmínky optimality pro nalezení lokálního minima funkce F. Dále definujeme obecně základní optimalizační metodu – iterační metodu sloužící k nalezení tohoto minima a uvedeme příklady těchto metod, mezi které patří metody s lokálně omezeným krokem. Uvedeme základní pojmy, definice, algoritmus a dokážeme globální a superlineární konvergenci této metody. Vyslovíme KKT (Karush, Kuhn, Tucker) podmínky pro nalezení optimálního lokálně omezeného kroku a ukážeme některé jeho vlastnosti.

1.1 Základní optimalizační metoda

Nechť jsou dány (dvakrát) spojitě diferencovatelné funkce

$$F : \mathbb{R}^n \to \mathbb{R}, \quad c_I : \mathbb{R}^n \to \mathbb{R}^{m_I} \quad c_E : \mathbb{R}^n \to \mathbb{R}^{m_E}$$

v proměnné $y \in \mathbb{R}^n$, kde F je funkce, jejíž (lokální) minimum hledáme, c_I a c_E jsou funkce, které reprezentují omezení na hledané minimum a dále

$$I = \{1, \dots, m_I\} \quad \text{a} \quad E = \{m_I + 1, \dots, m_I + m_E = m\}.$$

Mohou též nastat případy $m_I = 0$, $m_E = 0$ i $m = 0$, ale předpokládáme, že $m_E \le n$.

Definice 1.1 *Obecnou úlohou nelineárního programování nazýváme úlohu nalezení bodu $y_\star \in \mathbb{R}^n$ takového, že platí*

$$y_\star = \arg\min_{y \in \mathbb{R}^n} F(y) \tag{1.1}$$

na množině zadané omezeními

$$c_k(y) \le 0,\ k \in I, \quad c_k(y) = 0,\ k \in E, \tag{1.2}$$

kde $c_k(y) \le 0$ je míněno po složkách.

Jestliže $m = 0$, jedná se o neomezenou minimalizaci, v opačném případě jde o minimalizaci s omezeními. Označme

$$g(y) = \left[\frac{\partial F(y)}{\partial y_1}, \dots, \frac{\partial F(y)}{\partial y_n}\right]^T \quad \text{a} \quad \mathbf{G}(y) = \begin{pmatrix} \frac{\partial^2 F(y)}{\partial y_1^2} & \cdots & \frac{\partial^2 F(y)}{\partial y_1 \partial y_n} \\ \vdots & & \vdots \\ \frac{\partial^2 F(y)}{\partial y_n \partial y_1} & \cdots & \frac{\partial^2 F(y)}{\partial y_n^2} \end{pmatrix}$$

gradient a Hessovu matici (matici druhých parciálních derivací) funkce F. Spojitost druhých parciálních derivací implikuje symetrii matice $\mathbf{G}(y)$.

Definice 1.2 *Množinu*

$$\mathcal{P} = \{y \in \mathbb{R}^n : c_I(y) \leq 0,\ c_E(y) = 0\}$$

nazveme přípustnou množinou úlohy (1.1)-(1.2). Jestliže $m = 0$, *pak* $\mathcal{P} = \mathbb{R}^n$.

Funkce $F(y)$ může mít hodně lokálních minim, protože uvažujeme velmi obecné problémy. Je proto obtížné říci něco o řešení problému (1.1)-(1.2). Zaměříme se na hledání lokálního minima.

Definice 1.3 *Bod* $y_\star \in \mathcal{P}$ *je globálním minimem funkce* $F : \mathbb{R}^n \to \mathbb{R}$ *vzhledem k omezením (1.2), jestliže platí*

$$F(y_\star) \leq F(y) \quad \forall y \in \mathcal{P}.$$

Bod $y_\star \in \mathcal{P}$ *je lokálním minimem funkce* $F : \mathbb{R}^n \to \mathbb{R}$, *jestliže existuje číslo* $\varepsilon > 0$ *takové, že*

$$F(y_\star) \leq F(y) \quad \forall y \in \mathcal{B}(y_\star, \varepsilon) \cap \mathcal{P}, \quad \textit{kde} \quad \mathcal{B}(y_\star, \varepsilon) = \{y \in \mathbb{R}^n : \|y - y_\star\| < \varepsilon\}.$$

Jestliže navíc

$$F(y_\star) < F(y) \quad \forall y \in \mathcal{P}, \quad \textit{resp.} \quad \forall y \in \mathcal{B}(y_\star, \varepsilon) \cap \mathcal{P}, \quad y \neq y_\star,$$

pak bod $y_\star \in \mathcal{P}$ *je ostrým globálním, resp. lokálním minimem funkce* F *vzhledem k omezením (1.2). Každé globální minimum je též lokálním minimem.*

K tomu, abychom mohli o bodu y prohlásit, že je lokálním minimem funkce F vzhledem k omezením (1.2), hledáme tzv. optimalizační podmínky, které tento bod splňuje. Pokud se jedná o neomezenou minimalizaci ($m = 0$), jsou tyto podmínky poměrně jednoduché. Následující lemmata stanovují nutné podmínky optimality prvního a druhého řádu (lemma 1.1) a postačující podmínky druhého řádu (lemma 1.2).

Lemma 1.1 *Nechť bod* $y_\star \in \mathbb{R}^n$ *je lokálním minimem funkce* $F : \mathbb{R}^n \to \mathbb{R}$ *a nechť* $F \in \mathcal{C}^1$ *(spojitě diferencovatelná) na* $\mathcal{B}(y_\star, \varepsilon)$. *Pak platí*

$$g(y_\star) = 0.$$

Jestliže navíc $F \in \mathcal{C}^2$ *(dvakrát spojitě diferencovatelná) na* $\mathcal{B}(y_\star, \varepsilon)$, *pak platí*

$$\mathbf{G}(y_\star) \succeq 0,$$

neboli, že matice $\mathbf{G}(y_\star)$ *je positivně semidefinitní.*

Lemma 1.2 *Nechť* $F : \mathbb{R}^n \to \mathbb{R}$, $F \in \mathcal{C}^2$ *na* $\mathcal{B}(y_\star, \varepsilon)$ *a nechť platí*

$$g(y_\star) = 0 \quad \textit{a} \quad \mathbf{G}(y_\star) \succ 0,$$

neboli, že matice $\mathbf{G}(y_\star)$ *je positivně definitní. Pak bod* $y_\star \in \mathbb{R}^n$ *je ostrým lokálním minimem funkce* F.

V případě minimalizace s omezeními ($m > 0$) je věc poněkud složitější. Hledáme podmínky, které jsou nutné nebo postačující k tomu, aby byl bod y, který splňuje tyto podmínky, lokálním minimem funkce F vzhledem k omezením (1.2). Dříve než uvedeme nutné KKT podmínky prvního řádu (věta 1.1) a postačující podmínky druhého řádu (lemma 1.3), zavedeme následující pojmy.

Definice 1.4 *Nechť $\mathcal{P} \subset \mathbb{R}^n$ je přípustná množina úlohy (1.1)-(1.2) a nechť $y \in \mathcal{P}$.*

1. *Množinu*
$$\mathcal{E} = E \cup \{k \in I : c_k(y) = 0\}$$
nazveme množinou indexů aktivních omezení v bodě y.

2. *Řekneme, že v bodě y jsou splněny podmínky regularity, jestliže jsou vektory*
$$\left\{ \nabla c_k(y) = \left(\frac{\partial c_k(y)}{\partial y_1}, \ldots, \frac{\partial c_k(y)}{\partial y_n} \right)^T : k \in \mathcal{E} \right\}$$
lineárně nezávislé.

3. *Jestliže existuje vektor $u \in \mathbb{R}^m$ takový, že platí*
$$g(y) + \sum_{k=1}^{m} u_k \nabla c_k(y) = 0, \tag{1.3}$$
$$c_k(y) = 0, \quad k \in E; \qquad c_k(y) \le 0, \quad u_k \ge 0, \quad u_k c_k(y) = 0, \quad k \in I, \tag{1.4}$$
pak řekneme, že y je stacionárním bodem úlohy (1.1)-(1.2). Podmínky (1.3)-(1.4) se nazývají podmínky optimality a vektor $u \in \mathbb{R}^m$ je Lagrangeův multiplikátor. (Tím, že $y \in \mathcal{P}$, platí již $c_k(y) = 0,\ k \in E$ a $c_k(y) \le 0,\ k \in I$.)

4. *Nechť $y \in \mathcal{P}$ je stacionárním bodem úlohy (1.1)-(1.2). Jestliže $c_k(y) < u_k\ \forall k \in I$, pak řekneme, že v bodě y jsou splněny podmínky striktní komplementarity.*

Věta 1.1 (Karush-Kuhn-Tucker) *Nechť $F : \mathbb{R}^n \to \mathbb{R}$ a $c_k : \mathbb{R}^n \to \mathbb{R},\ 1 \le k \le m$ jsou spojitě diferencovatelné funkce, nechť $y_\star \in \mathbb{R}^n$ je lokálním minimem funkce F na množině zadané omezeními (1.2) a jsou v něm splněny podmínky regularity. Pak $y^\star$ je stacionárním bodem úlohy (1.1)-(1.2).*
Kromě toho, jsou-li funkce F a c_I konvexní, funkce c_E lineární, jsou-li splněny podmínky regularity v bodě $y_\star \in \mathbb{R}^n$ a platí-li (1.3)-(1.4), je bod $y_\star$ globálním minimem funkce F na množině zadané omezeními (1.2).

Důkaz: Viz např. [13], [45].

Lemma 1.3 *Nechť $F : \mathbb{R}^n \to \mathbb{R}$ a $c_k : \mathbb{R}^n \to \mathbb{R},\ 1 \le k \le m$ jsou dvakrát spojitě diferencovatelné funkce, nechť $y \in \mathcal{P}$ je stacionárním bodem úlohy (1.1)-(1.2) a jsou v něm splněny podmínky regularity a striktní komplementarity. Jestliže platí*
$$z^T \Big[\mathbf{G}(y) + \sum_{k=1}^{m} u_k \nabla^2_{yy} c_k(y) \Big] z > 0 \quad \forall z \in \mathbb{R}^n,\ z \ne 0$$
takové, že
$$z^T \nabla c_k(y) = 0 \quad \forall k \in \mathcal{E},$$
pak je bod y ostrým lokálním minimem funkce F na množině zadané omezeními (1.2).

Důkazy lemmat 1.1-1.3 lze najít např. v [21].

V této práci se budeme zabývat iteračními metodami k nalezení lokálního minima $y_\star$ funkce F na množině $\mathcal{P}$. Nejprve uvedeme definici základní metody tohoto typu.

Definice 1.5 *Základní optimalizační metoda je iterační proces, jehož výsledkem je posloupnost bodů* $y_k \in \mathbb{R}^n$, $k \in \mathbb{N}_0$, *taková, že*

$$y_{k+1} = y_k + \alpha_k x_k, \tag{1.5}$$

kde směrový vektor x_k *se určuje na základě hodnot* y_j, $F(y_j)$, $g(y_j)$, $\mathbf{G}(y_j)$, $0 \le j \le k$, *a délka kroku* $\alpha_k > 0$ *se určuje na základě chování funkce* F *v okolí bodu* $y_k \in \mathbb{R}^n$.

Definice 1.6 *Základní optimalizační metoda je globálně konvergentní, jestliže pro libovolný počáteční vektor* $y_0 \in \mathbb{R}^n$ *platí*

$$\liminf_{k\to\infty} \|g(y_k)\| = 0.$$

Pokud je základní optimalizační metoda globálně konvergentní, neznamená to ještě, že $y_k \to y_\star$, ani že $\|g(y_k)\| \to 0$ pro $k \to \infty$. Máme však jistotu, že po dostatečném počtu kroků dostaneme bod y_k takový, že $\|g(y_k)\| \le \varepsilon$ pro $\varepsilon > 0$ jakkoliv malé.

Mezi nejjednodušší a nejznámější optimalizační metody patří tyto metody:

Metoda největšího spádu – je definována vztahy

$$x_k = -g(y_k), \quad \alpha_k = \arg\min_{\alpha \ge 0} F(y_k + \alpha x_k).$$

Výhodou této metody je, že je globálně konvergentní a používá pouze vektory z $\mathbb{R}^n$, což znamená $\mathcal{O}(n)$ paměťových míst a $\mathcal{O}(n)$ operací na iteraci. Nevýhodou je, že používá přesný výběr délky kroku a není superlineárně konvergentní.

Newtonova metoda – je definována vztahy

$$x_k = -\mathbf{G}^{-1}(y_k)\, g(y_k), \quad \alpha_k = 1.$$

Tato metoda je kvadraticky konvergentní a pokud konverguje, stačí k nalezení lokálního minima pouze několik iterací. Dále používá jednoduchý výběr délky kroku. Není však globálně konvergentní a kromě toho je třeba řešit soustavu lineárních rovnic, což znamená $\mathcal{O}(n^2)$ paměťových míst a $\mathcal{O}(n^3)$ operací na iteraci. Navíc je třeba počítat druhé derivace.

Aby se odstranily nevýhody těchto jednoduchých metod, byly vyvinuty důmyslnější a tudíž i složitější metody:

Metody spádových směrů – vyvinuté z metody největšího spádu, které používají nepřesný výběr délky kroku a rychleji konvergují (princip sdružených směrů). Mezi tyto metody patří metoda sdružených gradientů a metoda s proměnnou metrikou.

Metody s lokálně omezeným krokem – vyvinuté z Newtonovy metody, u kterých je zajištěna globální konvergence a je snížen počet operací (nepřesné řešení lineárních rovnic). Tímto způsobem se realizuje Newtonova metoda a Gaussova-Newtonova metoda pro minimalizaci součtu čtverců.

V této práci se budeme zabývat metodami s lokálně omezeným krokem kvůli jejich globální konvergenci při řešení optimalizačních problémů. Budeme přitom uvažovat takové implementace, které dovolí řešit rozsáhlé optimalizační problémy.

K důkazu globální a superlineární konvergence metod s lokálně omezeným krokem budeme používat následující předpoklady na funkci F.

Definice 1.7 *Nechť* $F : \mathbb{R}^n \to \mathbb{R}$ *je dvakrát spojitě diferencovatelná. Řekneme, že*

1. *F je zdola omezená, jestliže platí*

$$F(y) \geq \underline{F} \quad \forall y \in \mathbb{R}^n, \quad \underline{F} \in \mathbb{R}. \tag{1.6}$$

2. *F má omezené druhé derivace, jestliže platí*

$$|z^T \mathbf{G}(y) z| \leq \overline{G}\, \|z\|^2 \quad \forall y, z \in \mathbb{R}^n, \quad \overline{G} > 0. \tag{1.7}$$

Je to ekvivalentní podmínce $\|\mathbf{G}(y)\| \leq \overline{G}\ \forall y \in \mathbb{R}^n$, *protože* $\mathbf{G}(y)$ *je symetrická matice.*

3. *F je stejnoměrně (nebo silně) konvexní, jestliže platí*

$$z^T \mathbf{G}(y) z \geq \underline{G}\, \|z\|^2 \quad \forall y, z \in \mathbb{R}^n, \quad \underline{G} > 0. \tag{1.8}$$

Dále uvedeme vlastnosti konvergentních posloupností.

Definice 1.8 *Posloupnost* $\{y_k\}_{k\in\mathbb{N}_0} \in \mathbb{R}^n$ *konverguje k bodu* $y_\star$ *Q-lineárně, jestliže existuje index* $i \in \mathbb{N}$ *a konstanta* $0 < C < 1$ *tak, že platí*

$$\|y_{k+1} - y_\star\| \leq C\, \|y_k - y_\star\| \quad \forall k \geq i.$$

Posloupnost $\{y_k\}_{k\in\mathbb{N}_0} \in \mathbb{R}^n$ *konverguje k bodu* $y_\star$ *Q-superlineárně, jestliže platí*

$$\lim_{k\to\infty} \frac{\|y_{k+1} - y_\star\|}{\|y_k - y_\star\|} = 0.$$

(Superlineární konvergence implikuje monotonnost posloupnosti $\{\|y_k - y_\star\|\}_{k\in\mathbb{N}_0}$ *počínaje vhodným indexem* $i \in \mathbb{N},\ i \leq k$.*)*
Posloupnost $\{y_k\}_{k\in\mathbb{N}_0} \in \mathbb{R}^n$ *konverguje k bodu* $y_\star$ *Q-kvadraticky, jestliže existuje index* $i \in \mathbb{N}$ *a konstanta* $0 < C < \infty$ *tak, že platí*

$$\|y_{k+1} - y_\star\| \leq C\, \|y_k - y_\star\|^2 \quad \forall k \geq i.$$

Poznámka 1.1 *Při vyšetřování konvergence použijeme věty o střední hodnotě:*

1. $F(y+z) = F(y) + z^T g(y) + \frac{1}{2}\, z^T \mathbf{G}(y + \tilde{\alpha} z) z = F(y) + z^T g(y) + \frac{1}{2}\, z^T \mathbf{G}(y) z + o(\|z\|^2)$
   $\Rightarrow$ *pro funkci F splňující (1.7) a (1.8) platí*

$$z^T g(y) + \frac{1}{2}\, \underline{G}\, \|z\|^2 \leq F(y+z) - F(y) \leq z^T g(y) + \frac{1}{2}\, \overline{G}\, \|z\|^2,$$

   kde $0 \leq \tilde{\alpha} \leq 1$.

2. $g(y+z) = g(y) + \int_0^1 \mathbf{G}(y + \alpha z) z \, d\alpha = g(y) + \mathbf{G}(y) z + o(\|z\|)$
   $\Rightarrow$ *pro funkci F splňující (1.7) a (1.8) platí*

$$\underline{G}\, \|z\| \leq \|g(y+z) - g(y)\| \leq \overline{G}\, \|z\|.$$

Uvedené vztahy plynou z Taylorova rozvoje.

1.2 Metody s lokálně omezeným krokem

Nejprve se budeme zabývat úlohou nalezení minima funkce $F : \mathbb{R}^n \to \mathbb{R}$ v případě, že nemáme žádné omezení tvaru (1.2) – neomezená minimalizace ($m = 0$), [08], [34], [35], [38], [45]. Pro výklad metod s lokálně omezeným krokem zavedeme následující označení:

$$F_k = F(y_k), \quad g_k = g(y_k), \quad \mathbf{G}_k = \mathbf{G}(y_k),$$

definujeme kvadratickou funkci

$$\psi_k(x) = \frac{1}{2}\, x^T \mathbf{A}_k x + g_k^T x,$$

která lokálně aproximuje rozdíl $F(y_k + x) - F(y_k)$, vektor

$$w_k(x) = \frac{\mathbf{A}_k x + g_k}{\|g_k\|}$$

pro přesnost určení směrového vektoru a číslo

$$\varrho_k(x) = \frac{F(y_k + x) - F(y_k)}{\psi_k(x)}$$

pro podíl skutečného a předpověděného poklesu funkce F. Z Taylorova rozvoje funkce F v bodě y_k dostáváme

$$F(y_k + x) = F(y_k) + g_k^T x + \frac{1}{2}\, x^T \mathbf{G}_k x + o(\|x\|^2),$$

takže podle výše uvedených vzorců očekáváme, že matice $\mathbf{A}_k$ bude aproximovat Hessovu matici $\mathbf{G}_k$. Normu uvažujeme euklidovskou.

Definice 1.9 *Základní optimalizační metoda (1.5) je metodou s lokálně omezeným krokem, jestliže se směrové vektory* $x_k \in \mathbb{R}^n$, $k \in \mathbb{N}_0$, *určují tak, že platí*

$$\|x_k\| \le \Delta_k, \tag{1.9}$$

$$\|x_k\| < \Delta_k \quad \Rightarrow \quad \|w_k(x_k)\| = \omega_k \le \overline{\omega}, \tag{1.10}$$

$$-\psi_k(x_k) \ge \underline{\sigma}\, \|g_k\| \min\left\{ \|x_k\|, \frac{\|g_k\|}{\|\mathbf{A}_k\|} \right\}, \tag{1.11}$$

kde $\overline{\omega} < 1$, $0 < \underline{\sigma} < 1$, *a délky kroku* $\alpha_k \ge 0$, $k \in \mathbb{N}_0$, *se vybírají tak, že*

$$\varrho_k(x_k) \le 0 \quad \Rightarrow \quad \alpha_k = 0, \tag{1.12}$$

$$\varrho_k(x_k) > 0 \quad \Rightarrow \quad \alpha_k = 1. \tag{1.13}$$

Posloupnost $\Delta_k > 0$, $k \in \mathbb{N}_0$, *se konstruuje tak, že*

$$\varrho_k(x_k) < \underline{\varrho} \quad \Rightarrow \quad \underline{\beta}\, \|x_k\| \le \Delta_{k+1} \le \overline{\beta}\, \|x_k\|, \tag{1.14}$$

$$\varrho_k(x_k) \ge \underline{\varrho} \quad \Rightarrow \quad \Delta_k \le \Delta_{k+1} \le \overline{\gamma}\, \Delta_k, \tag{1.15}$$

kde $0 < \underline{\beta} \le \overline{\beta} < 1 < \overline{\gamma}$ *a* $0 < \underline{\varrho} < 1$.

Poznámka 1.2 *Aby byla metoda s lokálně omezeným krokem globálně konvergentní, konstruuje se posloupnost* $\{\mathbf{A}_k\}_{k\in\mathbb{N}_0}$ *tak, aby byly matice* $\mathbf{A}_k \neq 0$ *stejnoměrně omezené, tj.*

$$\|\mathbf{A}_k\| \leq M \quad \forall k \in \mathbb{N}_0, \tag{1.16}$$

kde konstanta $M < \infty$ *nezávisí na* $k \in \mathbb{N}_0$. *Globální konvergence platí i za slabších předpokladů, [50]. Stačí, aby platilo*

$$\sum_{k=0}^{\infty} \frac{1}{\|\mathbf{A}_k\|} = \infty.$$

My však budeme vycházet z předpokladu (1.16).

Základní algoritmus metody s lokálně omezeným krokem vypadá následovně. Symbol $\mathbb{R}_S^{n\times n}$ značí prostor symetrických matic řádu $n \times n$.

Algoritmus 1.1 *Metoda s lokálně omezeným krokem.*

Zvolíme $y_0 \in \mathbb{R}^n$, $0 \neq \mathbf{A}_0 \in \mathbb{R}_S^{n\times n}$, $\Delta_0 > 0$, $\varepsilon > 0$, *spočítáme* $F(y_0)$, *položíme* $k = 0$.

1. *Vypočteme gradient* $g(y_k)$. *Je-li* $\|g(y_k)\| < \varepsilon$, *pak* `STOP`.
2. *Určíme vektor* x_k *tak, aby byly splněny podmínky (1.9) až (1.11).*
3. *Položíme* $y_k^+ = y_k + x_k$ *a vypočteme hodnoty* $F(y_k^+)$ *a* $\varrho_k(x_k) = \frac{F(y_k^+)-F(y_k)}{\psi_k(x_k)}$.
4. *Je-li* $\varrho_k(x_k) < \underline{\varrho}$, *určíme* Δ_{k+1} *tak, aby byla splněna podmínka (1.14).*
   Je-li $\varrho_k(x_k) \geq \underline{\varrho}$, *určíme* Δ_{k+1} *tak, aby byla splněna podmínka (1.15).*
5. *Je-li* $\varrho_k(x_k) \leq 0$, *přejdeme na krok 2.*
   Je-li $\varrho_k(x_k) > 0$, *určíme matici* $\mathbf{A}_{k+1} \neq 0$ *tak, aby byla splněna podmínka (1.16), položíme* $y_{k+1} = y_k^+$, $F(y_{k+1}) = F(y_k^+)$, $k := k+1$ *a návrat na krok 1.*

V další části dokážeme globální a superlineární konvergenci metody s lokálně omezeným krokem. Z globální konvergence plyne, že mezi body 5. a 2. nenastane nekonečný cyklus. Budeme používat následující množiny indexů ($\mathbb{N}_0 = \mathbb{N} \cup \{0\}$):

- $\mathbb{N}_1 = \{k \in \mathbb{N}_0 : \|x_k\| < \Delta_k\}$,
- $\mathbb{N}_2 = \{k \in \mathbb{N}_0 : \varrho_k(x_k) > 0\}$,
- $\mathbb{N}_3 = \{k \in \mathbb{N}_0 : \varrho_k(x_k) \geq \underline{\varrho}\}$,
- $\mathbb{N}_4 = \{k \in \mathbb{N}_0 : k \notin \mathbb{N}_1, \, k \in \mathbb{N}_3, \, k \neq 0\}$.

Lemma 1.4 *Aplikujeme-li metodu s lokálně omezeným krokem (1.9)-(1.15) na funkci F, která splňuje podmínku (1.7), platí-li (1.16) a označíme-li* $m_k = \min\{\|g_j\| : 0 \leq j \leq k\}$, *pak existuje konstanta* $c > 0$ *taková, že* $\forall k \in \mathbb{N}_0$ *platí*

$$\|x_k\| \geq c\,\frac{m_k}{M}$$

DůKAZ:

1. Nechť $k \in \mathbb{N}_1$. Pak podle (1.10) platí

$$\Big| \|\mathbf{A}_k x_k\| - \|g_k\| \Big| \le \|\mathbf{A}_k x_k + g_k\| = \|w_k(x_k)\| \|g_k\| \le \overline{\omega}\, \|g_k\|,$$

takže

$$-(\|\mathbf{A}_k x_k\| - \|g_k\|) \le \overline{\omega}\, \|g_k\| \quad \Rightarrow \quad \|\mathbf{A}_k\| \|x_k\| \ge \|\mathbf{A}_k x_k\| \ge (1-\overline{\omega})\, \|g_k\|$$

a tedy

$$\|x_k\| \ge (1-\overline{\omega}) \frac{\|g_k\|}{\|\mathbf{A}_k\|} \ge (1-\overline{\omega}) \frac{m_k}{M} = c_1 \frac{m_k}{M}.$$

2. Nechť $k \notin \mathbb{N}_3$, tedy $\varrho_k(x_k) < \underline{\varrho}$. Pak z $\psi_k(x_k) \le 0$, vztah (1.11), plyne

$$F(y_k + x_k) - F(y_k) = \varrho_k(x_k)\psi_k(x_k) \ge \underline{\varrho}\, \psi_k(x_k) \ge \underline{\varrho}\, (g_k^T x_k - \frac{1}{2} \|\mathbf{A}_k\| \|x_k\|^2)$$

a z věty o střední hodnotě dostaneme $F(y_k + x_k) - F(y_k) \le g_k^T x_k + \frac{1}{2} \overline{G}\, \|x_k\|^2$, což dohromady dává

$$\begin{aligned} g_k^T x_k + \frac{1}{2} \overline{G}\, \|x_k\|^2 \ &\ge\ \underline{\varrho}\, (g_k^T x_k - \frac{1}{2} \|\mathbf{A}_k\| \|x_k\|^2) \quad \Rightarrow \\ \Rightarrow \quad \frac{1}{2} \left(\overline{G} + \underline{\varrho}\, \|\mathbf{A}_k\|\right) \|x_k\|^2 \ &\ge\ \left(\underline{\varrho} - 1\right) g_k^T x_k. \end{aligned}$$

Z (1.11) dostaneme

$$-\underline{\sigma}\, \|g_k\| \min \left\{ \|x_k\|, \frac{\|g_k\|}{\|\mathbf{A}_k\|} \right\} \ge \psi_k(x_k) \ge g_k^T x_k - \frac{1}{2} \|\mathbf{A}_k\| \|x_k\|^2$$

a to spolu s předchozí nerovností, protože $\underline{\varrho} - 1 < 0$, dává

$$\frac{1}{2} \left(\overline{G} + \underline{\varrho}\, \|\mathbf{A}_k\|\right) \|x_k\|^2 \ge \frac{1}{2} \left(\underline{\varrho} - 1\right) \|\mathbf{A}_k\| \|x_k\|^2 - \underline{\sigma} \left(\underline{\varrho} - 1\right) \|g_k\| \min \left\{ \|x_k\|, \frac{\|g_k\|}{\|\mathbf{A}_k\|} \right\},$$

neboli

$$\frac{1}{2} \left(\overline{G} + \|\mathbf{A}_k\|\right) \|x_k\|^2 \ge \underline{\sigma} \left(1 - \underline{\varrho}\right) \|g_k\| \min \left\{ \|x_k\|, \frac{\|g_k\|}{\|\mathbf{A}_k\|} \right\}.$$

Jestliže $\|x_k\| \ge \frac{\|g_k\|}{\|\mathbf{A}_k\|} \ge \frac{m_k}{M}$, jsme hotovi. Nechť tedy $\|x_k\| < \frac{\|g_k\|}{\|\mathbf{A}_k\|}$. Pak můžeme psát

$$\frac{1}{2} \left(\overline{G} \frac{M}{\|\mathbf{A}_0\|} + M \right) \|x_k\|^2 \ge \frac{1}{2} \left(\overline{G} + \|\mathbf{A}_k\|\right) \|x_k\|^2 \ge \underline{\sigma} \left(1 - \underline{\varrho}\right) m_k \|x_k\|$$

a tedy

$$\|x_k\| \ge \left[2\underline{\sigma} \left(1 - \underline{\varrho}\right) \frac{\|\mathbf{A}_0\|}{\overline{G} + \|\mathbf{A}_0\|} \right] \frac{m_k}{M} = c_2 \frac{m_k}{M}.$$

3. Nechť $k = 0$. Pokud $\|g_0\| = 0$, platí zřejmě $\|x_0\| \ge \frac{\|g_0\|}{\|\mathbf{A}_0\|} \ge \frac{m_0}{M}$. Jestliže $\|g_0\| \ne 0$, pak

$$\|x_0\| = \frac{\|x_0\| \|\mathbf{A}_0\|}{\|g_0\|} \frac{\|g_0\|}{\|\mathbf{A}_0\|} \ge \frac{\|x_0\| \|\mathbf{A}_0\|}{\|g_0\|} \frac{m_0}{M} = c_3 \frac{m_0}{M}.$$

4. Nechť $k \in \mathbb{N}_4$. Nechť $j < k$ je největší index takový, pro který platí $j \notin \mathbb{N}_4$ (takový index existuje, neboť $0 \notin \mathbb{N}_4$). Pak postupně podle (1.15), (1.14) a (1.9) dostaneme
$$\|x_k\| = \Delta_k \geq \Delta_{j+1} \geq \min\left\{\Delta_j,\, \underline{\beta}\,\|x_j\|\right\} = \underline{\beta}\,\|x_j\|,$$
takže podle již dokázaného v bodech 1.-3. platí
$$\|x_k\| \geq \underline{\beta}\,\|x_j\| \geq \underline{\beta}\,\min\{c_1, c_2, c_3\}\,\frac{m_k}{M} = c\,\frac{m_k}{M}.$$

□

Věta 1.2 (globální konvergence) *Nechť $\{y_k\}_{k\in\mathbb{N}_0}$ je posloupnost generovaná metodou s lokálně omezeným krokem (1.9)-(1.15) na funkci F, která splňuje podmínky (1.6)-(1.7) a nechť je splněna podmínka (1.16). Pak platí*
$$\liminf_{k\to\infty} \|g_k\| = 0.$$

Důkaz: Předpokládejme existenci $\varepsilon > 0$, že $\|g_k\| \geq \varepsilon\ \forall k \in \mathbb{N}_0$. Pak podle lemmatu 1.4 platí
$$\|x_k\| \geq c\,\frac{\varepsilon}{M} \quad \forall k \in \mathbb{N}_0. \tag{1.17}$$
Nechť $k \in \mathbb{N}_3 \subset \mathbb{N}_2$. Označíme-li $F_k = F(y_k)$, $F_{k+1} = F(y_k + x_k)$, pak podle (1.11) a (1.15) platí
$$\begin{aligned} F_k - F_{k+1} &= -\varrho_k(x_k)\psi_k(x_k) \geq \varrho_k(x_k)\,\underline{\sigma}\,\|g_k\|\min\left\{\|x_k\|,\ \frac{\|g_k\|}{\|\mathbf{A}_k\|}\right\} \geq \\ &\geq \underline{\varrho}\,\underline{\sigma}\,\varepsilon\min\left\{c\,\frac{\varepsilon}{M},\ \frac{\varepsilon}{M}\right\} = \underline{\varrho}\,\underline{\sigma}\,\frac{\varepsilon^2}{M}\,\min\{c, 1\}. \end{aligned}$$
Dále z (1.6) plyne
$$\begin{aligned} F_0 - \underline{F} &\geq F_0 - \lim_{k\to\infty} F_{k+1} = \lim_{k\to\infty}(F_0 - F_{k+1}) = \sum_{k=0}^{\infty}(F_k - F_{k+1}) \geq \\ &\geq \sum_{k\in\mathbb{N}_3}(F_k - F_{k+1}) \geq \underline{\varrho}\,\underline{\sigma}\,\frac{\varepsilon^2}{M}\sum_{k\in\mathbb{N}_3}\min\{c, 1\}. \end{aligned}$$
Je-li množina $\mathbb{N}_3$ nekonečná, dojdeme ihned ke sporu s předpokladem (1.6). Je-li $\mathbb{N}_3$ konečná, pak podle (1.14) platí $\Delta_k \to 0$ pro $k \to \infty$, tedy $\|x_k\| \to 0$ podle (1.9) a to je spor s (1.17). □

Věta 1.3 (superlineární konvergence) *Nechť $\{y_k\}_{k\in\mathbb{N}_0}$ je posloupnost generovaná metodou s lokálně omezeným krokem (1.9)-(1.15) taková, že $y_k \to y_\star$, kde $y_\star$ je lokální minimum funkce F, která splňuje podmínky (1.7)-(1.8) pro*
$$\underline{G} < \lambda_{\min}(\mathbf{G}_\star) \leq \lambda_{\max}(\mathbf{G}_\star) < \overline{G},$$
kde $\lambda_{\min}(\mathbf{G}_\star)$ je nejmenší a $\lambda_{\max}(\mathbf{G}_\star)$ největší vlastní číslo matice $\mathbf{G}_\star = \mathbf{G}(y_\star)$, a nechť je splněna podmínka (1.16). Nechť dále
$$\lim_{k\to\infty}\omega_k = 0, \quad \lim_{k\to\infty}\|u_k\| = 0, \quad \textit{kde} \quad u_k = \frac{1}{\|x_k\|}(\mathbf{A}_k - \mathbf{G}_k)x_k. \tag{1.18}$$
Pak posloupnost y_k konverguje Q-superlineárně k bodu $y_\star$.

DŮKAZ:

1. Podle definice u_k platí

$$x_k^T \mathbf{A}_k x_k = x_k^T \mathbf{G}_k x_k + x_k^T u_k \|x_k\| \geq x_k^T \mathbf{G}_k x_k - \|u_k\| \|x_k\|^2$$

a jelikož $\|u_k\| \to 0$, $\mathbf{G}_k \to \mathbf{G}_\star$ a $\underline{G} < \lambda_{\min}(\mathbf{G}_\star)$, existuje index $k_1 \in \mathbb{N}_0$, že platí

$$x_k^T \mathbf{A}_k x_k \geq \underline{G}\, \|x_k\|^2 \quad \forall k \geq k_1.$$

Z definice $\psi_k(x_k)$ a z (1.11) plyne

$$0 \geq \psi_k(x_k) = \frac{1}{2} x_k^T \mathbf{A}_k x_k + g_k^T x_k \geq \frac{1}{2} \underline{G}\, \|x_k\|^2 - \|g_k\| \|x_k\| \quad \Rightarrow$$

$$\Rightarrow \quad \|g_k\| \geq \frac{1}{2} \underline{G}\, \|x_k\| \quad \forall k \geq k_1.$$

Použijeme-li ještě jednou (1.11), můžeme pro $k \geq k_1$ psát

$$\begin{aligned} -\psi_k(x_k) \ &\geq\ \underline{\sigma}\, \|g_k\| \min\left\{ \|x_k\|,\ \frac{\|g_k\|}{\|\mathbf{A}_k\|} \right\} \geq \frac{1}{2} \underline{\sigma}\, \underline{G}\, \|x_k\|^2 \min\left\{ 1,\ \frac{\|g_k\|}{\|\mathbf{A}_k\| \|x_k\|} \right\} \geq \\ &\geq\ \frac{1}{2} \underline{\sigma}\, \underline{G} \min\left\{ 1,\ \frac{\underline{G}}{2M} \right\} \|x_k\|^2. \end{aligned}$$

2. Podle věty o střední hodnotě platí

$$\begin{aligned} F(y_k + x_k) - F(y_k) \ &=\ g_k^T x_k + \frac{1}{2} x_k^T \mathbf{G}_k x_k + o(\|x_k\|^2) = \\ &=\ \psi_k(x_k) + \frac{1}{2} x_k^T (\mathbf{G}_k - \mathbf{A}_k) x_k + o(\|x_k\|^2), \end{aligned}$$

takže

$$\varrho_k(x_k) = \frac{F(y_k + x_k) - F(y_k)}{\psi_k(x_k)} = 1 + \frac{1}{2} \frac{x_k^T (\mathbf{G}_k - \mathbf{A}_k) x_k + o(\|x_k\|^2)}{\psi_k(x_k)}.$$

Podle bodu 1. však pro $k \geq k_1$ platí

$$\begin{aligned} \left| \frac{x_k^T (\mathbf{G}_k - \mathbf{A}_k) x_k + o(\|x_k\|^2)}{2 \psi_k(x_k)} \right| \ &\leq\ \frac{\|u_k\| \|x_k\|^2 + o(\|x_k\|^2)}{2 |\psi_k(x_k)|} \leq \\ &\leq\ \frac{1}{\underline{\sigma}\, \underline{G} \min\left\{1,\ \frac{\underline{G}}{2M}\right\}} \frac{\|u_k\| \|x_k\|^2 + o(\|x_k\|^2)}{\|x_k\|^2} \ \to\ 0, \end{aligned}$$

neboť $\|u_k\| \to 0$. Tedy $\varrho_k(x_k) \to 1$ a jelikož $\underline{\varrho} < 1$, existuje index $k_2 \geq k_1$, že

$$\varrho_k(x_k) \geq \underline{\varrho} \quad \forall k \geq k_2 \quad \Rightarrow \quad k \in \mathbb{N}_3 \quad \forall k \geq k_2.$$

3. Množina $\mathbb{N}_1$ je nekonečná. Kdyby tomu tak nebylo, muselo by platit $\|x_k\| = \Delta_k$ pro $k \geq k_3 \geq k_2$, přičemž $\Delta_k \geq \Delta_{k_2}\ \forall k \geq k_2$ podle bodu 2. a (1.15). Současně však podle bodu 1. platí $\|x_k\| \leq 2 \frac{\|g_k\|}{\underline{G}}$, takže $y_k \to y_\star$ implikuje $\|g_k\| \to 0$, tedy $\|x_k\| \to 0$ a to je spor. Dále se omezíme pouze na indexy $k \geq k_2$ takové, že $k \in \mathbb{N}_1$ a označíme $w_k \equiv w_k(x_k)$.

4. Platí

$$\mathbf{G}_k x_k = (\mathbf{A}_k x_k + g_k) - (\mathbf{A}_k - \mathbf{G}_k) x_k - g_k = \|g_k\|\, w_k - \|x_k\|\, u_k - g_k,$$

a protože podle (1.10) je $\|w_k\| < 1 \quad (k \in \mathbb{N}_1)$, pak

$$\begin{aligned} \Big(\|\mathbf{G}_k\| + \|u_k\|\Big)\|x_k\| &\geq \Big\|\mathbf{G}_k x_k + \|x_k\|\, u_k\Big\| = \Big\|\|g_k\|\, w_k - g_k\Big\| \geq \\ &\geq \Big|\|w_k\|\|g_k\| - \|g_k\|\Big| = \Big(1 - \|w_k\|\Big)\|g_k\| \quad \Rightarrow \\ \Rightarrow \quad \|x_k\| &\geq \frac{1 - \|w_k\|}{\|\mathbf{G}_k\| + \|u_k\|}\, \|g_k\|. \end{aligned}$$

Jelikož $\|u_k\| \to 0, \ \|w_k\| \to 0$ a $\|\mathbf{G}_k\| \to \|\mathbf{G}_\star\| = \lambda_{\max}(\mathbf{G}_\star) < \overline{G}$, existuje index $k_4 \geq k_2$, že $\|x_k\| \geq \frac{\|g_k\|}{\overline{G}} \ \forall k \geq k_4$.

Podobně platí $x_k = \mathbf{G}_k^{-1}\Big(\|g_k\|\, w_k - \|x_k\|\, u_k - g_k\Big)$, takže za předpokladu, že $\|u_k\|$ je tak malá, že $1 - \|\mathbf{G}_k^{-1}\|\|u_k\| > 0$, platí

$$\|x_k\| \leq \|\mathbf{G}_k^{-1}\|\Big(\|g_k\|\|w_k\| + \|x_k\|\|u_k\| + \|g_k\|\Big) \quad \Rightarrow \quad \|x_k\| \leq \frac{\|\mathbf{G}_k^{-1}\|\,(1 + \|w_k\|)}{1 - \|\mathbf{G}_k^{-1}\|\|u_k\|}\, \|g_k\|$$

a jelikož $\|u_k\| \to 0, \ \|w_k\| \to 0$ a $\|\mathbf{G}_k^{-1}\| \to \|\mathbf{G}_\star^{-1}\| = \frac{1}{\lambda_{\min}(\mathbf{G}_\star)} < \frac{1}{\underline{G}}$, existuje index $k_5 \geq k_4$, že $\|x_k\| \leq \frac{\|g_k\|}{\underline{G}} \ \forall k \geq k_5$. Celkem dostáváme

$$\frac{\|g_k\|}{\overline{G}} \leq \|x_k\| \leq \frac{\|g_k\|}{\underline{G}} \quad \forall k \geq k_5, \ k \in \mathbb{N}_1.$$

5. Podle věty o střední hodnotě platí

$$g_{k+1} - g_k = \mathbf{G}_k x_k + o(\|x_k\|).$$

Označme

$$v_k = \frac{1}{\|g_k\|}\,(g_{k+1} - g_k - \mathbf{A}_k x_k) = \frac{1}{\|g_k\|}\,[(\mathbf{G}_k - \mathbf{A}_k)x_k + o(\|x_k\|)].$$

Pak podle bodu 4. platí

$$\|v_k\| \leq \frac{1}{\|x_k\|\,\underline{G}}\Big[\|(\mathbf{G}_k - \mathbf{A}_k)x_k\| + o(\|x_k\|)\Big] = \frac{\|u_k\|}{\underline{G}} + \frac{o\|x_k\|}{\|x_k\|} \ \to \ 0.$$

Jelikož zároveň $\|w_k\| \to 0$, existuje index $k_6 \geq k_5, \ k_6 \in \mathbb{N}_1$, že

$$\|v_k\| \leq \frac{\underline{G}}{2\overline{G}}, \quad \|w_k\| \leq \frac{\underline{G}}{2\overline{G}} \quad \forall k \geq k_6, \ k \in \mathbb{N}_1.$$

Odtud a z bodu 4. plyne

$$\begin{aligned} \|x_{k+1}\| &\leq \frac{\|g_{k+1}\|}{\underline{G}} \leq \frac{1}{\underline{G}}\Big(\|g_{k+1} - g_k - \mathbf{A}_k x_k\| + \|\mathbf{A}_k x_k + g_k\|\Big) = \\ &= \frac{1}{\underline{G}}\Big(\|v_k\|\|g_k\| + \|w_k\|\|g_k\|\Big) \leq \frac{1}{\underline{G}}\left(\frac{\underline{G}}{2\overline{G}}\,\overline{G}\,\|x_k\| + \frac{\underline{G}}{2\overline{G}}\,\overline{G}\,\|x_k\|\right) = \|x_k\|. \end{aligned}$$

Jelikož $k \in \mathbb{N}_3 \cap \mathbb{N}_1$ podle bodů 2. a 3., platí

$$\|x_{k+1}\| \leq \|x_k\| < \Delta_k \leq \Delta_{k+1} \quad \Rightarrow \quad k + 1 \in \mathbb{N}_1.$$

Indukcí dostaneme $k \in \mathbb{N}_1 \ \forall k \geq k_6$.

6. Podle věty o střední hodnotě platí

$$\|g_k\| \le \overline{G}\,\|y_k - y_\star\|, \qquad \|g_{k+1}\| \ge \underline{G}\,\|y_{k+1} - y_\star\|.$$

Dále $\|v_k\| \to 0$, $\|w_k\| \to 0$ implikuje

$$\frac{\|g_{k+1}\|}{\|g_k\|} \le \frac{\|g_{k+1} - g_k - \mathbf{A}_k x_k\| + \|\mathbf{A}_k x_k + g_k\|}{\|g_k\|} = \|v_k\| + \|w_k\| \;\to\; 0,$$

což dohromady dává

$$\lim_{k\to\infty} \frac{\|y_{k+1} - y_\star\|}{\|y_k - y_\star\|} \le \lim_{k\to\infty} \frac{\overline{G}}{\underline{G}} \frac{\|g_{k+1}\|}{\|g_k\|} = 0.$$

□

Metody s lokálně omezeným krokem lze rozdělit na dvě skupiny. Hledá se minimum kvadratické funkce $\psi_k(x)$ vzhledem k omezení $\|x\| \le \Delta_k$. V prvním případě jde o minimalizaci na celém prostoru, $x \in \mathbb{R}^n$, tzv. optimální lokálně omezený krok, ve druhém případě jde pouze o přibližné řešení, kdy x hledáme na podprostorech prostoru $\mathbb{R}^n$ (např. na Krylovových podprostorech). V další části uvedeme vlastnosti optimálního lokálně omezeného kroku.

Definice 1.10 *Řekneme, že základní optimalizační metoda je metodou s optimálním lokálně omezeným krokem, jestliže používá směrové vektory* $x_k \in \mathbb{R}^n$, $k \in \mathbb{N}_0$, *takové, že*

(1.19) $$x_k = \arg \min_{\|x\| \le \Delta_k} \psi_k(x),$$

přičemž bereme $\|x_k\| = \Delta_k$, *pokud toto minimum není jediné.*

Existují-li dvě různá lokální minima z_1, z_2 funkce ψ_k taková, že $\|z_1\| < \Delta_k$, $\|z_2\| < \Delta_k$, pak platí

$$\mathbf{A}_k z_1 + g_k = 0, \quad \mathbf{A}_k z_2 + g_k = 0 \quad \Rightarrow \quad \mathbf{A}_k(z_1 - z_2) = 0.$$

Tedy $\mathbf{A}_k$ je singulární a jádro $\mathcal{N}(\mathbf{A}_k) \ne \{0\}$. Buď $0 \ne z_3 \in \mathcal{N}(\mathbf{A}_k)$, např. $z_3 = z_1 - z_2$. Protože $\mathbf{A}_k z_3 = 0$ a $g_k^T z_3 = -z_1^T \mathbf{A}_k z_3 = 0$, pak pro libovolný vektor $x_k = z_1 + \kappa z_3$, kde $\kappa \in \mathbb{R}$, platí

$$\begin{aligned}
\psi_k(x_k) &= \frac{1}{2} x_k^T \mathbf{A}_k x_k + g_k^T x_k = \frac{1}{2} (z_1 + \kappa z_3)^T \mathbf{A}_k (z_1 + \kappa z_3) + g_k^T (z_1 + \kappa z_3) = \\
&= \frac{1}{2} z_1^T \mathbf{A}_k z_1 + g_k^T z_1 = \psi_k(z_1).
\end{aligned}$$

Existuje tedy nekonečně mnoho lokálních minim funkce ψ_k vyplňujících lineární varietu, takže můžeme zvolit $\kappa > 0$ takové, aby $\|x_k\| = \Delta_k$.

Věta 1.4 *Směrový vektor určený podle (1.19) vyhovuje podmínkám (1.9) až (1.11) pro* $\overline{\omega} = 0$ *a* $\underline{\sigma} = \frac{1}{2}$, *tedy*

(1.20) $$\psi_k(x_k) \le -\frac{1}{2} \|g_k\| \min\left\{ \|x_k\|, \frac{\|g_k\|}{\|\mathbf{A}_k\|} \right\}.$$

DŮKAZ: Podmínka (1.9) je přímo součástí podmínky (1.19).

1. Nechť $\|x_k\| < \Delta_k$. Pak toto minimum je jediné podle definice, $\mathbf{A}_k$ je positivně definitní, $\psi_k(x)$ je ryze konvexní funkce a $\mathbf{A}_k x_k + g_k = 0$. Takže $w_k(x_k) = 0$ a můžeme položit $\overline{\omega} = 0$. Dále

$$g_k^T \mathbf{A}_k^{-1} g_k \geq \lambda_{\min}(\mathbf{A}_k^{-1}) \|g_k\|^2 = \frac{1}{\lambda_{\max}(\mathbf{A}_k)} \|g_k\|^2 \geq \frac{1}{\|\mathbf{A}_k\|} \|g_k\|^2,$$

takže

$$\psi_k(x_k) = g_k^T x_k + \frac{1}{2} x_k^T \mathbf{A}_k x_k = -g_k^T \mathbf{A}_k^{-1} g_k + \frac{1}{2} g_k^T \mathbf{A}_k^{-1} g_k \leq -\frac{1}{2} \frac{\|g_k\|^2}{\|\mathbf{A}_k\|}.$$

2. Nechť $\|x_k\| = \Delta_k$. Jestliže $g_k^T \mathbf{A}_k g_k > 0$, definujme $\hat{x} = -\frac{g_k^T g_k}{g_k^T \mathbf{A}_k g_k} g_k$.

   (a) Nechť $\|\hat{x}\| < \Delta_k$. Pak

$$\begin{aligned} \psi_k(\hat{x}) &= g_k^T \hat{x} + \frac{1}{2} \hat{x}^T \mathbf{A}_k \hat{x} = -\frac{(g_k^T g_k)^2}{g_k^T \mathbf{A}_k g_k} + \frac{1}{2} \frac{(g_k^T g_k)^2 g_k^T \mathbf{A}_k g_k}{(g_k^T \mathbf{A}_k g_k)^2} = -\frac{1}{2} \frac{(g_k^T g_k)^2}{g_k^T \mathbf{A}_k g_k} \leq \\ &\leq -\frac{1}{2} \frac{\|g_k\|^4}{\|g_k\|^2 \|\mathbf{A}_k\|} = -\frac{1}{2} \frac{\|g_k\|^2}{\|\mathbf{A}_k\|} \end{aligned}$$

   a podle (1.19) je

$$\psi_k(x_k) \leq \psi_k(\hat{x}) \leq -\frac{1}{2} \frac{\|g_k\|^2}{\|\mathbf{A}_k\|}. \tag{1.21}$$

   (b) Nechť $\|\hat{x}\| \geq \Delta_k$. Pak

$$\Delta_k \leq \|\hat{x}\| = \frac{\|g_k\|^3}{g_k^T \mathbf{A}_k g_k} \quad \Rightarrow \quad g_k^T \mathbf{A}_k g_k \leq \frac{\|g_k\|^3}{\Delta_k}.$$

3. Jestliže $g_k^T \mathbf{A}_k g_k \leq 0$, pak platí

$$g_k^T \mathbf{A}_k g_k \leq 0 \leq \frac{\|g_k\|^3}{\Delta_k}.$$

   V případech 2b. a 3. položme $\tilde{x} = -\frac{\Delta_k}{\|g_k\|} g_k$. Platí

$$\begin{aligned} \psi_k(\tilde{x}) &= g_k^T \tilde{x} + \frac{1}{2} \tilde{x}^T \mathbf{A}_k \tilde{x} = -\Delta_k \|g_k\| + \frac{1}{2} \frac{\Delta_k^2}{\|g_k\|^2} g_k^T \mathbf{A}_k g_k \leq \\ &\leq -\Delta_k \|g_k\| + \frac{1}{2} \Delta_k \|g_k\| = -\frac{1}{2} \Delta_k \|g_k\|. \end{aligned}$$

   Protože $\|\tilde{x}\| = \Delta$ a $\|x_k\| = \Delta$, platí

$$\psi_k(x_k) \leq \psi_k(\tilde{x}) \leq -\frac{1}{2} \Delta_k \|g_k\| = -\frac{1}{2} \|x_k\| \|g_k\|. \tag{1.22}$$

Ve všech případech dostáváme

$$\psi_k(x_k) \le -\frac{1}{2}\,\frac{\|g_k\|^2}{\|\mathbf{A}_k\|} \quad \text{nebo} \quad \psi_k(x_k) \le -\frac{1}{2}\,\|x_k\|\|g_k\|,$$

což dohromady dává vztah (1.20). □

Nyní se zaměříme na to, jak nalézt optimální lokálně omezený krok x_k podle definice 1.10. Následující věta stanovuje podmínky, které takové řešení splňuje.

Věta 1.5 *Vektor* $x_k \in \mathbb{R}^n$ *je řešením úlohy (1.19) ve smyslu definice 1.10 právě tehdy, když* $\|x_k\| \le \Delta_k$ *a existuje číslo* $\xi_k \ge 0$ *takové, že matice* $\mathbf{A}_k + \xi_k\mathbf{I}$ *je positivně semidefinitní (*$\mathbf{A}_k + \xi_k\mathbf{I} \succeq 0$*) a platí*

$$(\mathbf{A}_k + \xi_k\mathbf{I})x_k + g_k = 0, \quad (\|x_k\| - \Delta_k)\,\xi_k = 0.$$

DŮKAZ:

1. Nechť x_k je řešením (1.19). Zavedeme Langrangeovu funkci

$$\mathcal{L}(x,\xi) = \psi_k(x) + \frac{\xi}{2}\,(x^T x - \Delta^2).$$

Podle věty 1.1 existuje $\xi_k \in \mathbb{R}$, $\xi_k \ge 0$ takové, že platí

$$(\mathbf{A}_k + \xi_k\mathbf{I})x_k + g_k = 0 \quad \text{a} \quad (\|x_k\| - \Delta_k)\,\xi_k = 0.$$

Zbývá dokázat, že $\mathbf{A}_k + \xi_k\mathbf{I} \succeq 0$. Pro libovolný vektor $x \in \mathbb{R}^n$ platí

$$\begin{aligned}
\psi_k(x) - \psi_k(x_k) &= \frac{1}{2}\,x^T\mathbf{A}_k x + g_k^T x - \frac{1}{2}\,x_k^T\mathbf{A}_k x_k - g_k^T x_k = \\
&= (x_k - x)^T(\mathbf{A}_k + \xi_k\mathbf{I})x_k + \frac{1}{2}\,(x^T\mathbf{A}_k x - x_k^T\mathbf{A}_k x_k) = \\
&= \frac{1}{2}\,(x_k - x)^T(\mathbf{A}_k + \xi_k\mathbf{I})(x_k - x) + \\
&+ \frac{1}{2}\,\big[(x_k - x)^T(\mathbf{A}_k + \xi_k\mathbf{I})(x_k + x) + x^T\mathbf{A}_k x - x_k^T\mathbf{A}_k x_k\big] = \\
&= \frac{1}{2}\,(x_k - x)^T(\mathbf{A}_k + \xi_k\mathbf{I})(x_k - x) + \frac{1}{2}\,\xi_k(x_k^T x_k - x^T x). \qquad (1.23)
\end{aligned}$$

(a) Jestliže $\|x_k\| < \Delta_k$, pak toto x_k je jediné a je lokálním minimem funkce $\psi_k(x)$. Platí $\xi_k = 0$ a $\mathbf{A}_k \succ 0$.

(b) Nechť $\|x_k\| = \Delta_k$. Buď $x^{(i)}$ libovolný vektor takový, že $\|x^{(i)}\| = \|x_k\| = \Delta_k$ a označme $v_i = \frac{1}{\|x_k - x^{(i)}\|}\,(x_k - x^{(i)})$, pokud $x^{(i)} \ne x_k$. Pak podle (1.23) platí

$$0 \le \psi_k(x^{(i)}) - \psi_k(x_k) = \frac{1}{2}\,\|x_k - x^{(i)}\|^2\, v_i^T(\mathbf{A}_k + \xi_k\mathbf{I})v_i. \qquad (1.24)$$

Volíme-li vektory $x^{(i)} \ne x_k$ libovolně na množině $\|x^{(i)}\| = \Delta$, nabývá kosinus

$$\cos\varphi_i = \frac{\left(x_k - x^{(i)}\right)^T x_k}{\|x_k - x^{(i)}\|\|x_k\|} = \frac{v_i^T x_k}{\|x_k\|}$$

libovolných hodnot v intervalu $(0,1\rangle$. Volme nyní posloupnost $\{x^{(i)}\}$ takovou, že $x^{(i)} \to x_k$ tak, aby $\cos\varphi_i \to 0$. Jelikož $v_i^T(\mathbf{A}_k + \xi_k\mathbf{I})v_i \ge 0$ podle (1.24) a posloupnost $\{v_i\}$ konverguje k jistému vektoru v (kolmému na x_k), je též ze spojitosti $v^T(\mathbf{A}_k + \xi_k\mathbf{I})v \ge 0$. Tato nerovnost tedy platí pro všechny vektory z $\mathbb{R}^n$ a tudíž je $\mathbf{A}_k + \xi_k\mathbf{I} \succeq 0$.

2. Nechť $\|x_k\| \le \Delta_k$, nechť $\exists\, \xi_k \ge 0$ takové, že $\mathbf{A}_k + \xi_k \mathbf{I} \succeq 0$ a nechť platí

$$(\mathbf{A}_k + \xi_k \mathbf{I})x_k + g_k = 0, \quad (\|x_k\| - \Delta_k)\, \xi_k = 0.$$

(a) Jestliže $\|x_k\| < \Delta_k$, pak je $\xi_k = 0$ a platí

$$\begin{aligned} \mathbf{A}_k x_k + g_k = 0 \quad &\Leftrightarrow \quad \psi_k'(x_k) = 0 \quad &&\Rightarrow \quad x_k \text{ je kritický bod;} \\ \mathbf{A}_k \succeq 0 \quad &\Leftrightarrow \quad \psi_k''(x_k) \ge 0 \quad &&\Rightarrow \quad \text{v } x_k \text{ je lokální minimum.} \end{aligned}$$

Tedy x_k je řešením (1.19).

(b) Jestliže $\|x_k\| = \Delta_k$, pak podle (1.23) dostaneme pro libovolný vektor x takový, že $\|x\| \le \|x_k\| = \Delta_k$

$$\psi_k(x) - \psi_k(x_k) = \frac{1}{2}\,(x_k - x)^T(\mathbf{A}_k + \xi_k \mathbf{I})(x_k - x) + \frac{1}{2}\,\xi_k(x_k^T x_k - x^T x) \ge 0,$$

což znamená, že v x_k je lokální minimum a platí (1.19). □

Poznámka 1.3 *K důkazu globální konvergence jednotlivých metod ve druhé kapitole použijeme větu 1.2. Vždy budeme předpokládat, že funkce F splňuje podmínky (1.6)-(1.7). Označme dále $x_\star$ spočítané řešení pomocí algoritmu dané metody. Podmínky (1.9)-(1.16) lze splnit algoritmicky, kromě podmínky (1.11), kterou je třeba pro $x_k \equiv x_\star$ ověřit. Ukážeme-li, že $x_\star$ je optimální lokálně omezený krok, tzn. že splňuje podmínky věty 1.5, pak nerovnost (1.11) plyne z věty 1.4. Pokud $x_\star$ není optimální lokálně omezený krok, ale pouze přibližné řešení, musíme pro něj nerovnost (1.11) dokázat. V obou případech bude globální konvergence dané metody plynout z věty 1.2.*

1.3 Vlastnosti optimálního lokálně omezeného kroku

V další části se budeme zabývat podproblémem problému hledání minima funkce F bez omezení, a sice určením lokálně omezeného kroku x_k. Protože k tomu použijeme iterační proces $\{x_{k,i}\}_{i\in\mathbb{N}_0}$, označíme lokálně omezený krok v k-tém kroku jako $x_{k,\star}$. Jelikož se jedná o podproblém a k je pevné, tento index pro jednoduchost vynecháme. Nejprve shrneme celý podproblém. Budeme se snažit najít takové $x_\star$, které splňuje (1.19). Hledáme tedy řešení podproblému

(1.25) $$\min_{x\in\mathbb{R}^n} \psi(x) \quad \text{vzhledem k} \quad \|x\| \le \Delta,$$

kde

(1.26) $$\psi(x) = \frac{1}{2}\, x^T \mathbf{A} x + g^T x,$$

$\mathbf{A}$ je symetrická matice řádu $n \times n$, g je n-dimenzionální vektor, Δ je dané kladné číslo a norma je euklidovská. Řešení $x_\star$ je optimální lokálně omezený krok a platí pro něj věta 1.5. Z ní plyne, že řešit podproblém (1.25) je ekvivalentní nalezení $\xi_\star$ (tzv. Levenbergův-Marquardtův parametr) a řešení lineárního systému

(1.27) $$(\mathbf{A} + \xi \mathbf{I})x = -g$$

pro $\xi = \xi_\star$. Protože chceme $\mathbf{A} + \xi_\star \mathbf{I} \succeq 0$, platí $\xi_\star \ge -\lambda_1$, kde λ_1 je nejmenší vlastní číslo matice $\mathbf{A}$, což spolu s $\xi_\star \ge 0$ dává

$$\xi_\star \in \langle \max\{0, -\lambda_1\}, \infty).$$

Podívejme se, co dále implikuje věta 1.5.

1. Nechť $\mathbf{A} \succ 0$. Pak $\lambda_1 > 0$ a tudíž $\xi_\star \geq 0 > -\lambda_1$. Odpovídající $x_\star$ splňuje rovnici $x_\star = -(\mathbf{A} + \xi_\star \mathbf{I})^{-1} g$. Jestliže $\|x_\star\| < \Delta$, pak je $\xi_\star = 0$ a v bodě $x_\star$ nabývá funkce $\psi(x)$ globálního minima na $\mathbb{R}^n$.
2. Nechť $\mathbf{A} \not\succ 0$, tedy $\lambda_1 \leq 0$. Pak platí $\xi_\star \geq -\lambda_1 \geq 0$. Jestliže je $\xi_\star > -\lambda_1$, pak pro odpovídající $x_\star$, které splňuje $\|x_\star\| = \Delta$, platí opět $x_\star = -(\mathbf{A} + \xi_\star \mathbf{I})^{-1} g$.

Pokud je tedy $\xi_\star > -\lambda_1$, lze $x_\star$ vyjádřit následujícím způsobem. Uvažujme rozklad $\mathbf{A} = \mathbf{QDQ}^T$, kde $\mathbf{Q}$ je ortonormální matice, jejíž sloupce tvoří vlastní vektory matice $\mathbf{A}$ a $\mathbf{D}$ je diagonální matice s vlastními čísly $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ a nechť $\{\lambda_i, q_i\}_{i=1}^n$ jsou vlastní páry matice $\mathbf{A}$. Pak $x_\star$ je řešením (1.25) právě když platí $(\mathbf{QDQ}^T + \xi_\star \mathbf{I}) x_\star = -g$. Označíme-li $\tilde{x}_\star = \mathbf{Q}^T x_\star$, $\tilde{g} = \mathbf{Q}^T g$ a $\tilde{x}_\star^{(i)}$, $\tilde{g}_i$ i-té složky těchto vektorů, pak platí

(1.28) $$(\lambda_i + \xi_\star) \tilde{x}_\star^{(i)} = -\tilde{g}_i, \quad i = 1, \ldots, n \quad \text{a} \quad (\Delta - \|x_\star\|) \xi_\star = 0.$$

Existuje tedy jediný vektor $\tilde{x}_\star$, který má složky

(1.29) $$\tilde{x}^{(i)} = \frac{-\tilde{g}_i}{\lambda_i + \xi_\star}, \quad i = 1, 2, \ldots, n$$

a tudíž existuje jediný vektor $x_\star = \mathbf{Q} \tilde{x}_\star$, který je řešením podproblému (1.25).

Nyní se podívejme na případ, kdy je $\xi_\star = -\lambda_1$. V tomto případě je věc složitější, protože je matice $\mathbf{A} + \xi_\star \mathbf{I}$ singulární a rovnice (1.27) nemá jediné řešení. Označme

$$\mathcal{S}_1 = \{q : \mathbf{A} q = \lambda_1 q, \ q \neq 0\}$$

podprostor všech vlastních vektorů matice $\mathbf{A}$ asociovaných s vlastním číslem λ_1. Platí následující lemma.

Lemma 1.5 *Pro řešení $x_\star$ podproblému (1.25) platí:*

1. $\xi_\star = -\lambda_1$ *nebo* $x_\star \perp \mathcal{S}_1 \quad \Rightarrow \quad g \perp \mathcal{S}_1$,

2. $g \perp \mathcal{S}_1 \quad \Rightarrow \quad \xi_\star = -\lambda_1$ *nebo* $\xi_\star \neq -\lambda_1$ & $x_\star \perp \mathcal{S}_1$.

Důkaz: Tvrzení plynou z rovnic $(\mathbf{A} + \xi_\star \mathbf{I}) x_\star = -g$ a $(\mathbf{A} - \lambda_1 \mathbf{I}) q = 0$ pro $q \in \mathcal{S}_1$:

1. Jestliže $\xi_\star = -\lambda_1$, pak $(\mathbf{A} - \lambda_1 \mathbf{I}) x_\star = -g$. Odtud plyne
$$q^T (\mathbf{A} - \lambda_1 \mathbf{I}) x_\star = -q^T g \quad \& \quad x_\star^T (\mathbf{A} - \lambda_1 \mathbf{I}) q = 0 \quad \Rightarrow \quad q^T g = 0.$$
Jestliže $x_\star \perp q$, pak
$$q^T (\mathbf{A} + \xi_\star \mathbf{I}) x_\star = -q^T g \quad \& \quad x_\star^T (\mathbf{A} - \lambda_1 \mathbf{I}) q = 0 \quad \Rightarrow \quad q^T \mathbf{A} x_\star = -q^T g \quad \& \quad x_\star^T \mathbf{A} q = 0$$
a tedy $q^T g = 0$.
2. Pro $q \in \mathcal{S}_1$ platí
$$0 = q^T g = -q^T (\mathbf{A} + \xi_\star \mathbf{I}) x_\star \quad \& \quad 0 = x_\star^T (\mathbf{A} - \lambda_1 \mathbf{I}) q.$$
Tedy $\xi_\star \, q^T x_\star = -q^T \mathbf{A} x_\star = -\lambda_1 \, x_\star^T q \quad \Rightarrow \quad (\xi_\star + \lambda_1) \, q^T x_\star = 0.$ □

Rozhodující pro vznik singulárního systému je tedy to, zda je $g \perp \mathcal{S}_1$. Pokud ne, nemůže nastat $\xi_\star = \lambda_1$. Naopak, pokud ano, může, ale také nemusí být $\xi_\star = -\lambda_1$.

Definice 1.11 *Jestliže je vektor g kolmý na podprostor $\mathcal{S}_1$, pak řekneme, že pro problém (1.25) nastal „singulární případ“.*

V případě $\xi_\star = -\lambda_1$ lze řešení (1.25) dostat nalezením libovolného řešení rovnice

$$(\mathbf{A} - \lambda_1 \mathbf{I})x = -g,$$

kde $\|x\| \leq \Delta$, a určením vlastního vektoru $q \in \mathcal{S}_1$. Položíme-li

$$x_\star = x + \kappa q, \tag{1.30}$$

kde κ je určeno tak, že $\|x_\star\| = \Delta$, platí

$$(\mathbf{A} + \xi_\star \mathbf{I})x_\star = (\mathbf{A} - \lambda_1 \mathbf{I})(x + \kappa q) = (\mathbf{A} - \lambda_1 \mathbf{I})x = -g$$

a $x_\star$ je řešením podproblému (1.25), protože jsou splněny podmínky věty 1.5.

Na závěr uvedeme několik dalších poznámek o řešení $x_\star$.

Poznámka 1.4 *Řešení $x_\star$ budeme hledat ve tvaru $x_\star = (\mathbf{A} + \xi\mathbf{I})^\dagger w$ pro nějaké w, kde symbol $.^\dagger$ značí Moore-Penroseovu pseudoinverzi (poznámka A.1). Přitom platí*

$$(\mathbf{A} + \xi\mathbf{I})x_\star = -g \quad \Rightarrow \quad (\mathbf{A} + \xi\mathbf{I})(\mathbf{A} + \xi\mathbf{I})^\dagger w = -g \quad \Rightarrow$$
$$\Rightarrow \quad (\mathbf{A} + \xi\mathbf{I})^\dagger w = -(\mathbf{A} + \xi\mathbf{I})^\dagger g \quad \Rightarrow \quad x_\star = -(\mathbf{A} + \xi\mathbf{I})^\dagger g.$$

Lemma 1.6 *Pro řešení $x_\star$ podproblému (1.25) platí:*

1. $g^T x_\star \leq 0$.
2. $\psi(x_\star) \leq 0$.

Důkaz:

1. Protože $\mathbf{A} + \xi_\star \mathbf{I} \succeq 0$, platí
$$(\mathbf{A} + \xi_\star \mathbf{I})x_\star = -g \quad \Rightarrow \quad 0 \leq x_\star^T (\mathbf{A} + \xi_\star \mathbf{I})x_\star = -g^T x_\star.$$
2. Pro $x = 0$ platí $\psi(0) = 0$. Proto pro řešení $x_\star$, ve kterém nabývá funkce ψ nejmenší hodnoty, platí $\psi(x_\star) \leq \psi(0) = 0$. □

Poznámka 1.5 *V definici 1.9 lze místo normy $\|x\| \leq \Delta$ uvažovat také normu*

$$\|x\|_M = \sqrt{x^T \mathbf{M} x} \leq \Delta,$$

kde $\mathbf{M}$ je symetrická a positivně semidefinitní matice. Podmínky pro řešení $x_\star$ ve větě 1.5 potom mají tvar

$$\|x\|_M \leq \Delta, \quad \xi \geq 0, \quad (\mathbf{A} + \xi\mathbf{M})x + g = 0, \quad (\|x\|_M - \Delta)\xi = 0, \quad \mathbf{A} + \xi\mathbf{M} \succeq 0.$$

Poznámka 1.6 *Z (1.27) vidíme, že iterace x závisí na ξ. Budeme proto ve zvláštních případech uvádět $x \equiv x(\xi)$, aby byla vidět závislost x na ξ.*

Kapitola 2

Výpočet lokálně omezeného kroku

V této kapitole se budeme věnovat řešení problému (1.25), tedy nalezení lokálně omezeného kroku $x_\star$. Jedná se o krok 2. algoritmu 1.1 pro minimalizaci funkce $F(y)$ bez omezení. Uvedeme přehled existujících metod (jsou zde též uvedeny nové původní metody vhodné pro řešení rozsáhlých strukturovaných úloh), teoretický rozbor jejich vlastností, algoritmy a důkazy globální konvergence podle poznámky 1.3.

Optimální lokálně omezený krok lze určit pomocí Choleského rozkladu matice v rovnici (1.27) nebo jako lineární kombinaci vlastních vektorů matice $\mathbf{A}$. Nehledáme-li optimální lokálně omezený krok, ale pouze přibližné řešení, pak metody generují po částech lineární křivky. Metoda psí nohy aproximuje minimum funkce ψ na dvourozměrném podprostoru, metoda sdružených gradientů hledá přibližné řešení na Krylovových podprostorech a použití Lanczosova procesu vede na posloupnost aproximací optimálního lokálně omezeného kroku řešením transformovaných třídiagonálních systémů. Optimální lokálně omezený krok lze najít také tak, že problém (1.25) převedeme na parametrizovaný problém vlastních čísel a počítáme vlastní páry jisté matice $\mathbf{B}_\tau$.

2.1 Použití Choleského rozkladu

Nejprve se budeme zabývat metodou, která hledá optimální lokálně omezený krok. Provedeme rozklad $\mathbf{A} = \mathbf{QDQ}^T$, kde $\mathbf{D}$ je diagonální matice s vlastními čísly matice $\mathbf{A}$ a $\mathbf{Q} = (q_1, \ldots, q_n)$ je ortonormální matice, jejíž sloupce tvoří příslušné vlastní vektory. Při řešení rovnice (1.27) provedeme Choleského rozklad $\mathbf{A} + \xi\mathbf{I} = \mathbf{R}^T\mathbf{R}$ a ukážeme, že získané řešení je téměř optimální, [08], [34], [43], [45], [51], [59], [68].

Vektor $x_\star$, který je řešením (1.19), můžeme získat řešením rovnice $(\mathbf{A} + \xi_\star\mathbf{I})x_\star = -g$, známe-li číslo $\xi_\star \geq \max\{0, -\lambda_1\}$, kde λ_1 je nejmenší vlastní číslo matice $\mathbf{A}$, vyhovující podmínkám věty 1.5. Je-li $\xi_\star > \max\{0, -\lambda_1\}$, pak $\mathbf{A} + \xi_\star\mathbf{I}$ je positivně definitní matice a rovnice $(\mathbf{A} + \xi_\star\mathbf{I})x = -g$ má jediné řešení $x_\star$, kde $\|x_\star\| = \Delta$ (věta 1.5). Číslo $\xi_\star$ je řešením rovnice

(2.1) $$\phi(\xi) \equiv \frac{1}{\Delta} - \frac{1}{\|x\|} = 0 \quad \text{pro} \quad \xi > \max\{0, -\lambda_1\},$$

kde x je definováno jako řešení rovnice $(\mathbf{A}+\xi\mathbf{I})x = -g$. Rovnice (2.1) má velmi výhodné vlastnosti z hlediska použití Newtonovy metody, jak dále uvidíme (funkce $\phi(\xi)$ je na intervalu $(-\lambda_1, \infty)$ konvexní, klesající a tudíž existuje jediné $\xi_\star$ takové, že $\phi(\xi_\star) = 0$), oproti rovnici $\tilde{\phi}(\xi) \equiv \|x\| - \Delta = 0$.

Lemma 2.1 *Jestliže* $\xi > \max\{0, -\lambda_1\}$, *pak platí*

$$\phi'(\xi) = -\frac{x^T(\mathbf{A}+\xi\mathbf{I})^{-1}x}{\|x\|^3} \quad a \quad \phi''(\xi) \geq 0. \tag{2.2}$$

DŮKAZ: Protože $x \equiv x(\xi)$, z rovnosti $\phi(\xi) = \frac{1}{\Delta} - \frac{1}{\|x(\xi)\|}$ plyne

$$\phi'(\xi) = \frac{\|x(\xi)\|'}{\|x(\xi)\|^2}.$$

Spočítáme derivaci

$$\|x(\xi)\|' = \left[\sqrt{(x(\xi)^T x(\xi))}\right]' = \frac{x(\xi)^T x'(\xi)}{\|x(\xi)\|}$$

a zderivujeme obě strany rovnosti $(\mathbf{A}+\xi\mathbf{I})x(\xi) = -g$, abychom dostali

$$(\mathbf{A}+\xi\mathbf{I})'x(\xi) + (\mathbf{A}+\xi\mathbf{I})x'(\xi) = 0 \quad \Rightarrow \quad x'(\xi) = -(\mathbf{A}+\xi\mathbf{I})^{-1}x(\xi).$$

Tedy

$$\|x(\xi)\|' = -\frac{x(\xi)^T(\mathbf{A}+\xi\mathbf{I})^{-1}x(\xi)}{\|x(\xi)\|}$$

a celkem

$$\phi'(\xi) = -\frac{x(\xi)^T(\mathbf{A}+\xi\mathbf{I})^{-1}x(\xi)}{\|x(\xi)\|^3}.$$

Nyní využijme rozklad $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T$ a pro $\xi > \max\{0, -\lambda_1\}$ dostaneme

$$\mathbf{Q}^T(\mathbf{A}+\xi\mathbf{I})\mathbf{Q} = \mathbf{D}+\xi\mathbf{I} \quad \Rightarrow \quad (\mathbf{A}+\xi\mathbf{I})^{-1} = \mathbf{Q}(\mathbf{D}+\xi\mathbf{I})^{-1}\mathbf{Q}^T;$$

$$x(\xi) = -(\mathbf{A}+\xi\mathbf{I})^{-1}g = -\mathbf{Q}(\mathbf{D}+\xi\mathbf{I})^{-1}\mathbf{Q}^T g.$$

Tedy

$$\|x(\xi)\| = \sqrt{g^T\mathbf{Q}(\mathbf{D}+\xi\mathbf{I})^{-2}\mathbf{Q}^T g} = \sqrt{\sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_i+\xi)^2}}, \tag{2.3}$$

kde $g = \sum_{i=1}^{n} \vartheta_i q_i$ a dále

$$\phi(\xi) = \frac{1}{\Delta} - \left[\sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_i+\xi)^2}\right]^{-\frac{1}{2}}, \tag{2.4}$$

$$\phi'(\xi) = -\left[\sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_i+\xi)^3}\right]\left[\sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_i+\xi)^2}\right]^{-\frac{3}{2}} \quad \text{a} \tag{2.5}$$

$$\phi''(\xi) = 3\omega(\xi)\left[\sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_i+\xi)^2}\right]^{-\frac{5}{2}}, \quad \text{kde} \tag{2.6}$$

$$\omega(\xi) = \left[\sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_i+\xi)^4}\right]\left[\sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_i+\xi)^2}\right] - \left[\sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_i+\xi)^3}\right]^2 \geq 0 \tag{2.7}$$

podle Schwarzovy nerovnosti. $\square$

Z tohoto lemmatu plyne, že $\phi(\xi)$ klesá monotonně a je konvexní, protože $\phi'(\xi) < 0$ a $\phi''(\xi) \geq 0 \ \forall \xi > \max\{0, -\lambda_1\}$ a $g \neq 0$.

Dále podle (2.3) dostáváme tyto limity:

$$\lim_{\xi \to -\lambda_{1+}} \|x(\xi)\| = \infty, \quad \lim_{\xi \to \infty} \|x(\xi)\| = 0.$$

Platí tedy

$$-\infty \leq \phi(\xi) \leq \frac{1}{\Delta} \tag{2.8}$$

První derivace $\phi'(\xi)$ je rostoucí funkce a je omezená. Vyšetřením limit pro $\xi \to \infty$ a $\xi \to -\lambda_{1+}$ ve vztahu (2.5) odvodíme tyto meze:

$$-\frac{1}{|\vartheta_k|} \leq \phi'(\xi) \leq -\frac{1}{\|g\|}, \tag{2.9}$$

kde $k \in \{1, \dots, n\}$ je nejmenší celé číslo takové, že $\vartheta_k \neq 0$.

1. $\xi \to \infty$:

$$\begin{aligned}
\lim_{\xi \to \infty} \phi'(\xi) &= -\lim_{\xi \to \infty} \left[\sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_i + \xi)^3} \right] \left[\sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_i + \xi)^2} \right]^{-\frac{3}{2}} \cdot \frac{\xi^3}{\xi^3} = \\
&= -\lim_{\xi \to \infty} \underbrace{\left[\sum_{i=1}^{n} \underbrace{\frac{\xi^3}{(\lambda_i + \xi)^3}}_{\to 1} \cdot \vartheta_i^2 \right]}_{\to \|g\|^2} \underbrace{\left[\sum_{i=1}^{n} \underbrace{\frac{\xi^2}{(\lambda_i + \xi)^2}}_{\to 1} \cdot \vartheta_i^2 \right]^{-\frac{3}{2}}}_{\to \|g\|^2} = -\frac{1}{\|g\|}
\end{aligned}$$

2. $\xi \to -\lambda_{1+}$:

   (a) Nechť λ_1 je l-násobné a $k \leq l$.

$$\begin{aligned}
\lim_{\xi \to -\lambda_{1+}} \phi'(\xi) &= -\lim_{\xi \to -\lambda_{1+}} \left[\sum_{i=k}^{l} \frac{\vartheta_i^2}{(\lambda_1 + \xi)^3} + \sum_{i=l+1}^{n} \frac{\vartheta_i^2}{(\lambda_i + \xi)^3} \right] \cdot \\
&\cdot \left[\sum_{i=k}^{l} \frac{\vartheta_i^2}{(\lambda_1 + \xi)^2} + \sum_{i=l+1}^{n} \frac{\vartheta_i^2}{(\lambda_i + \xi)^2} \right]^{-\frac{3}{2}} \cdot \frac{(\lambda_1 + \xi)^3}{(\lambda_1 + \xi)^3} = \\
&= -\lim_{\xi \to -\lambda_{1+}} \left[\underbrace{\sum_{i=k}^{l} \frac{(\lambda_1 + \xi)^3}{(\lambda_1 + \xi)^3} \cdot \vartheta_i^2}_{\to \vartheta_k^2 + \dots + \vartheta_l^2} + \underbrace{\sum_{i=l+1}^{n} \frac{(\lambda_1 + \xi)^3}{(\lambda_i + \xi)^3} \cdot \vartheta_i^2}_{\to 0} \right] \cdot \\
&\cdot \left[\underbrace{\sum_{i=k}^{l} \frac{(\lambda_1 + \xi)^2}{(\lambda_1 + \xi)^2} \cdot \vartheta_i^2}_{\to \vartheta_k^2 + \dots + \vartheta_l^2} + \underbrace{\sum_{i=l+1}^{n} \frac{(\lambda_1 + \xi)^2}{(\lambda_i + \xi)^2} \cdot \vartheta_i^2}_{\to 0} \right]^{-\frac{3}{2}} = \\
&= -\frac{1}{\sqrt{\vartheta_k^2 + \dots + \vartheta_l^2}} \geq -\frac{1}{|\vartheta_k|}
\end{aligned}$$

(b) Nechť $k > l$. Pak $\lambda_k > \lambda_1$ a v limitě můžeme dosadit $\xi := -\lambda_1$.

$$\begin{aligned}
\lim_{\xi \to -\lambda_1+} \phi'(\xi) &= -\lim_{\xi \to -\lambda_1+} \left[\sum_{i=k}^{n} \frac{\vartheta_i^2}{(\lambda_i+\xi)^3}\right] \left[\sum_{i=k}^{n} \frac{\vartheta_i^2}{(\lambda_i+\xi)^2}\right]^{-\frac{3}{2}} = \\
&= -\left[\sum_{i=k}^{n} \frac{\vartheta_i^2}{(\lambda_i-\lambda_1)^3}\right] \left[\sum_{i=k}^{n} \frac{\vartheta_i^2}{(\lambda_i-\lambda_1)^2}\right]^{-\frac{3}{2}} = \\
&= -\left[\sum_{i=k}^{n} \left(\frac{\vartheta_i}{\lambda_i-\lambda_1}\right)^2 \cdot \frac{1}{\lambda_i-\lambda_1}\right] \left[\sum_{i=k}^{n} \left(\frac{\vartheta_i}{\lambda_i-\lambda_1}\right)^2\right]^{-\frac{3}{2}} \geq \\
&\geq -\frac{1}{\lambda_k-\lambda_1} \left[\sum_{i=k}^{n} \left(\frac{\vartheta_i}{\lambda_i-\lambda_1}\right)^2\right]^{-\frac{1}{2}} = \\
&= -\frac{1}{|\vartheta_k|} \frac{|\vartheta_k|}{(\lambda_k-\lambda_1)} \left[\sum_{i=k}^{n} \left(\frac{\vartheta_i}{\lambda_i-\lambda_1}\right)^2\right]^{-\frac{1}{2}} = -\frac{1}{|\vartheta_k|} \frac{|\varrho_k|}{\|\varrho\|} \geq -\frac{1}{|\vartheta_k|},
\end{aligned}$$

kde ϱ je vektor o složkách $(\varrho_k, \ldots, \varrho_n) = \left(\frac{\vartheta_k}{\lambda_k-\lambda_1}, \ldots, \frac{\vartheta_n}{\lambda_n-\lambda_1}\right)$.

Lemma 2.2 *Nechť* $g = \sum_{i=1}^{n} \vartheta_i q_i$ *a* $k \in \{1, \ldots, n\}$ *je nejmenší celé číslo takové, že* $\vartheta_k \neq 0$. *Nechť* $l < n$ *je takové, že* $\lambda_1 = \lambda_2 = \ldots = \lambda_l$. *Pak singulární případ pro problém (1.25) nastává právě když platí* $k > l$.

DŮKAZ: Nechť $g = \sum_{i=1}^{n} \vartheta_i q_i$ a $\mathcal{S}_1 = \{q_1, \ldots, q_l\}$ je podprostor vlastních vektorů asociovaných s l-násobným vlastním číslem λ_1. Platí

$$g^T(q_1, \ldots, q_l) = \sum_{i=1}^{n} \vartheta_i q_i^T (q_1, \ldots, q_l) = (\vartheta_1, \ldots, \vartheta_l).$$

Jestliže nastane singulární případ, tedy $g \perp \mathcal{S}_1$, pak $\vartheta_1 = \ldots = \vartheta_l = 0$ a tudíž je $k > l$. Jestliže naopak je $k > l$, tedy $\vartheta_1 = \ldots = \vartheta_l = 0$, pak je $g \perp \mathcal{S}_1$. $\square$

Nyní použijeme Newtonovu metodu na rovnici (2.1). Abychom mohli spočítat derivace $\phi'(\xi)$ podle lemmatu 2.1, budeme uvažovat Choleského rozklad matice $\mathbf{A} + \xi\mathbf{I} = \mathbf{R}^T\mathbf{R}$, kde $\mathbf{R}$ je horní trojúhelníková matice. Zavedeme-li vektor w, pro který platí $\mathbf{R}^T w = x$, pak z (2.2) plyne

$$\phi'(\xi) = -\frac{\|w\|^2}{\|x\|^3}.$$

Abychom mohli provést Choleského rozklad, musíme zajistit (a to provedeme později), aby $\mathbf{A} + \xi\mathbf{I} \succ 0$, tj. $\xi > -\lambda_1$.

Newtonův krok je definován takto:

$$\xi^+ = \xi - \frac{\phi(\xi)}{\phi'(\xi)},$$

což po dosazení za derivaci dává:

$$\xi^+ = \xi + \frac{\|x\|^3}{\|w\|^2} \cdot \left(\frac{1}{\Delta} - \frac{1}{\|x\|}\right) = \xi + \frac{\|x\|^2}{\|w\|^2} \cdot \left(\frac{\|x\| - \Delta}{\Delta}\right). \tag{2.10}$$

Newtonova metoda má tento tvar:

Algoritmus 2.1 <u>*Newtonova metoda.*</u>

Nechť je dáno $\Delta > 0$ *a* $\xi \geq 0$, *kde* $\mathbf{A} + \xi\mathbf{I} \succ 0$.

1. *Provedeme rozklad* $\mathbf{A} + \xi\mathbf{I} = \mathbf{R}^T\mathbf{R}$.
2. *Řešíme* $\mathbf{R}^T\mathbf{R}x = -g$.
3. *Řešíme* $\mathbf{R}^T w = x$.
4. *Položíme* $\xi := \xi + \left(\frac{\|x\|}{\|w\|}\right)^2 \cdot \left(\frac{\|x\|-\Delta}{\Delta}\right)$.

Nechť pro problém (1.25) nastane singulární případ. Podle definice tedy platí $g \perp \mathcal{S}_1$, odtud je $q_i^T g = \vartheta_i = 0$ pro $i = 1, \ldots, l$, kde l je násobnost vlastního čísla λ_1, a platí

$$\phi(\xi) = \frac{1}{\Delta} - \left[\sum_{i=l+1}^{n} \frac{\vartheta_i^2}{(\lambda_i + \xi)^2}\right]^{-\frac{1}{2}} \quad \text{a} \quad \lim_{\xi \to -\lambda_{1+}} \|x(\xi)\| < \infty.$$

Použijeme-li rozklad $\mathbf{A} = \mathbf{QDQ}^T$, lze rovnici $(\mathbf{A} + \xi\mathbf{I})x = -g$ přepsat na tvar

$$(\mathbf{D} + \xi\mathbf{I})\mathbf{Q}^T x = -\mathbf{Q}^T g.$$

Pokud je $\xi > -\lambda_1$, platí $q_1^T x = 0$. Jestliže však je $\xi = -\lambda_1$, může výraz $q_1^T x$ nabývat libovolných hodnot, matice $\mathbf{A} + \xi\mathbf{I}$ je singulární a funkce $\phi(\xi)$ nabývá v tomto případě více hodnot. Její minimální hodnotu dostaneme pro $q_1^T x = 0$, neboť z ortogonality matice $\mathbf{Q}$ plyne $\|x\| = \|\mathbf{Q}^T x\|$.

Pokud $\lim_{\xi \to -\lambda_{1+}} \phi(\xi) > 0$, má rovnice $\phi(\xi) = 0$ řešení $\xi_\star$ v intervalu $(-\lambda_1, \infty)$, můžeme použít iterační proces (2.10) a hledat řešení problému (1.25) pomocí Choleského rozkladu $\mathbf{A} + \xi\mathbf{I} = \mathbf{R}^T\mathbf{R}$ tak jako v případě, kdy $g \not\perp \mathcal{S}_1$. Pokud $\lim_{\xi \to -\lambda_{1+}} \phi(\xi) \leq 0$, tedy $\|x\| \leq \Delta$, má rovnice $\phi(\xi) = 0$ řešení $\xi_\star = -\lambda_1$ a vektor řešení $x_\star$ je třeba spočítat jako $x_\star = x + \kappa q$, kde

$$(\mathbf{A} - \lambda_1\mathbf{I})(x + \kappa q) = -g, \quad \|x + \kappa q\| = \Delta \quad \text{a} \quad q \in \mathcal{S}_1.$$

Na obrázku 2.1 jsou znázorněny možnosti funkce $\phi(\xi)$ v okolí bodu $-\lambda_1$. V levé části je případ $\vartheta_1 \neq 0$ $(k = 1)$, nenastává tedy singulární případ. Jakmile se ϑ_1 blíží k nule, začne se funkce $\phi(\xi)$ v bodě $-\lambda_1$ spojovat. V pravé části obrázku platí $\vartheta_1 = 0$, $\vartheta_2 \neq 0$ $(k = 2)$, což odpovídá singulárnímu případu. Vpravo nahoře platí $\lim_{\xi \to -\lambda_{1+}} \phi(\xi) \leq 0$, což vede na singulární systém (1.27), protože $\xi_\star = -\lambda_1$. Vpravo dole platí $\lim_{\xi \to -\lambda_{1+}} \phi(\xi) > 0$, existuje tedy $\xi_\star > -\lambda_1$ a systém (1.27) je positivně definitní. Svislá úsečka v bodě $-\lambda_1$ obsahuje hodnoty $\phi(-\lambda_1)$ pro různé hodnoty výrazu $q_1^T x$.

V další části se budeme zabývat hledáním řešení problému (1.25).

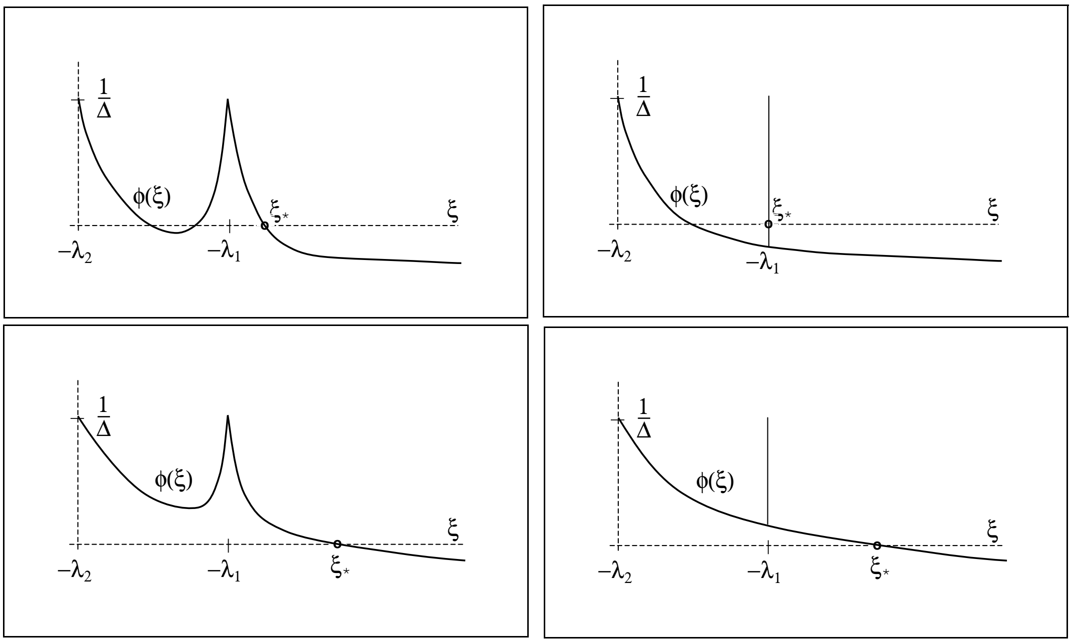
Lemma 2.3 *Nechť je dáno* $\sigma \in (0, 1)$, *nechť* $\mathbf{A} + \xi\mathbf{I} = \mathbf{R}^T\mathbf{R}$, $(\mathbf{A} + \xi\mathbf{I})x = -g$, $\xi \geq 0$ *a nechť vektor* $z \in \mathbb{R}^n$ *splňuje*

$$\|x + z\| = \Delta, \quad \|\mathbf{R}z\|^2 \leq \sigma(\|\mathbf{R}x\|^2 + \xi\Delta^2). \tag{2.11}$$

Pak platí

$$-\psi(x + z) \geq (1 - \sigma)|\psi_\star|, \tag{2.12}$$

kde $\psi_\star = \psi(x_\star)$ *a* $x_\star$ *je řešení problému (1.25).*



Obrázek 2.1: Funkce $\phi(\xi)$ v okolí bodu $-\lambda_1$

DŮKAZ: Pro každé $z \in \mathbb{R}^n$ platí:

$$\begin{aligned}
\psi(x+z) &= g^T(x+z) + \frac{1}{2}(x+z)^T\mathbf{A}(x+z) = \\
&= -x^T(\mathbf{A}+\xi\mathbf{I})(x+z) + \frac{1}{2}(x+z)^T(\mathbf{A}+\xi\mathbf{I})(x+z) - \frac{1}{2}\xi(x+z)^T(x+z) = \\
&= -x^T\mathbf{R}^T\mathbf{R}(x+z) + \frac{1}{2}\|\mathbf{R}x\|^2 + \frac{1}{2}\|\mathbf{R}z\|^2 + x^T\mathbf{R}^T\mathbf{R}z - \frac{1}{2}\xi\|x+z\|^2 = \\
&= \frac{1}{2}\|\mathbf{R}x\|^2 + \frac{1}{2}\|\mathbf{R}z\|^2 - \frac{1}{2}\xi\|x+z\|^2 - \|\mathbf{R}x\|^2 = \\
&= -\frac{1}{2}\left(\|\mathbf{R}x\|^2 + \xi\|x+z\|^2\right) + \frac{1}{2}\|\mathbf{R}z\|^2. \qquad (2.13)
\end{aligned}$$

Dále pro každé z, které splňuje (2.11), platí

$$-\psi(x+z) \geq \frac{1}{2}(\|\mathbf{R}x\|^2 + \xi\Delta^2) - \frac{1}{2}\sigma(\|\mathbf{R}x\|^2 + \xi\Delta^2) = \frac{1}{2}(1-\sigma)(\|\mathbf{R}x\|^2 + \xi\Delta^2).$$

Kromě toho, je-li $\psi_\star = \psi(x+z_\star)$, kde $\|x+z_\star\| \leq \Delta$, pak (2.13) implikuje, že

$$-\psi(x+z_\star) = \frac{1}{2}(\|\mathbf{R}x\|^2 + \xi\|x+z_\star\|^2) - \frac{1}{2}\|\mathbf{R}z_\star\|^2 \leq \frac{1}{2}(\|\mathbf{R}x\|^2 + \xi\Delta^2). \qquad (2.14)$$

Celkem

$$-\psi(x+z) \geq \frac{1}{2}(1-\sigma)(\|\mathbf{R}x\|^2 + \xi\Delta^2) \geq (1-\sigma)(-\psi_\star) = (1-\sigma)|\psi_\star|,$$

neboť $\psi_\star < 0$. □

Důsledkem je, že $|\psi(x+z) - \psi_\star| \leq \sigma|\psi_\star|$. Tedy platí-li (2.11), pak $x+z$ je pro dostatečně malá σ téměř optimální řešení problému (1.25).

Poznámka 2.1 *Označme* $\tilde{x} = x + z$ *a předpokládejme, že platí* $\tilde{x} = \frac{\Delta}{\|x\|} x$. *Jestliže vektor* x *splňuje podmínku*

$$\left(\frac{\Delta}{\|x\|} - 1\right)^2 \leq \sigma\left(1 + \frac{\xi\Delta^2}{\|\mathbf{R}x\|^2}\right), \tag{2.15}$$

pak platí nerovnost (2.11).

DŮKAZ: Platí

$$\begin{aligned}\|\mathbf{R}z\|^2 &= \|\mathbf{R}(\tilde{x} - x)\|^2 = \|\mathbf{R}\Big(\frac{\Delta}{\|x\|}x - x\Big)\|^2 = \|\mathbf{R}\Big(\frac{\Delta}{\|x\|} - 1\Big)x\|^2 = \\ &= \Big(\frac{\Delta}{\|x\|} - 1\Big)^2\|\mathbf{R}x\|^2 \leq \sigma\Big(1 + \frac{\xi\Delta^2}{\|\mathbf{R}x\|^2}\Big)\|\mathbf{R}x\|^2 = \sigma(\|\mathbf{R}x\|^2 + \xi\Delta^2),\end{aligned}$$

což je nerovnost (2.11). □

Vektor $z \in \mathbb{R}^n$ použitý v lemmatu 2.3 nemusí mít nutně tvar $z = \kappa q$ pro $q \in \mathcal{S}_1$. Je však třeba, aby výraz $\|\mathbf{R}z\|^2 = z^T(\mathbf{A} + \xi\mathbf{I})z$ byl co nejmenší, abychom mohli očekávat, že bude splněna nerovnost (2.11). Pro $\xi = -\lambda_1$ a $z = \kappa q$, $q \in \mathcal{S}_1$, pochopitelně platí

$$\|\mathbf{R}z\|^2 = z^T(\mathbf{A} + \xi\mathbf{I})z = \kappa^2 q^T(\mathbf{A} - \lambda_1\mathbf{I})q = 0.$$

Obvykle volíme $z = \kappa v$, kde v je vektor, který je aproximací nějakého vektoru $q \in \mathcal{S}_1$ a $\|v\| = 1$. Lze ho najít vhodnou numerickou metodou, např. inverzní mocninnou metodou. Číslo κ určíme tak, aby platilo $\|x + z\| = \|x + \kappa v\| = \Delta$:

$$\|x + \kappa v\| = \Delta \quad \Rightarrow \quad \kappa^2 + 2\kappa x^T v + \|x\|^2 - \Delta^2 = 0,$$

tedy

$$\begin{aligned}\kappa_{1,2} &= -x^T v \pm \sqrt{(x^T v)^2 + \Delta^2 - \|x\|^2} \cdot \frac{-x^T v \mp \sqrt{(x^T v)^2 + \Delta^2 - \|x\|^2}}{-x^T v \mp \sqrt{(x^T v)^2 + \Delta^2 - \|x\|^2}} = \\ &= \frac{(x^T v)^2 - [(x^T v)^2 + \Delta^2 - \|x\|^2]}{-x^T v \mp \sqrt{(x^T v)^2 + \Delta^2 - \|x\|^2}} = \frac{\Delta^2 - \|x\|^2}{x^T v \pm \sqrt{(x^T v)^2 + \Delta^2 - \|x\|^2}}.\end{aligned}$$

Znaménko ve jmenovateli zvolíme tak, aby byl jmenovatel větší a obešla se možnost dělení malým číslem (když se $\|x\|$ blíží k Δ, je jmenovatel blízko nuly). Odtud plyne

$$\kappa = \frac{\Delta^2 - \|x\|^2}{x^T v + \mathrm{sgn}(x^T v)\sqrt{(x^T v)^2 + \Delta^2 - \|x\|^2}}. \tag{2.16}$$

Test konvergence se provede na základě následujícího lemmatu.

Lemma 2.4 *Nechť je dáno* $\sigma, \sigma_1, \sigma_2 \in (0, 1)$, *iterace* $\{\xi, x\}$, *nechť*

$$\mathbf{A} + \xi\mathbf{I} = \mathbf{R}^T\mathbf{R}, \quad (\mathbf{A} + \xi\mathbf{I})x = -g, \quad \xi \geq 0$$

a nechť $x_\star$ *je řešení (1.25) s funkční hodnotou* $\psi(x_\star) \equiv \psi_\star$.

1. *Jestliže*

(2.17) $$\Big|\Delta - \|x\|\Big| \le \sigma\Delta \quad \textit{nebo} \quad \|x\| \le \Delta \ \ \& \ \ \xi = 0,$$

pak platí

(2.18) $$-\psi(x) \ge (1-\sigma)^2|\psi_\star|$$

a tedy x je aproximační řešení problému (1.25).

2. *Kdykoli je $\|x\| < \Delta$ a $\xi \ne 0$, pak nechť κ a v jsou takové, že $\|x + \kappa v\| = \Delta$. Jestliže*

(2.19) $$\|\mathbf{R}(\kappa v)\|^2 \le \sigma_1(2-\sigma_1)\max\{\sigma_2, \|\mathbf{R}x\|^2 + \xi\Delta^2\},$$

pak platí

(2.20) $$0 \le \psi(x+\kappa v) - \psi_\star \le \sigma_1(2-\sigma_1)\max\{|\psi_\star|, \sigma_2\}$$

a tedy $x + \kappa v$ je aproximační řešení problému (1.25), pro který nastal singulární případ.

DŮKAZ:

1. (a) Vztah (2.13) pro $z = 0$ implikuje, že

$$\begin{aligned} -\psi(x) &= \frac{1}{2}(\|\mathbf{R}x\|^2 + \xi\|x\|^2) \ge \frac{1}{2}(\|\mathbf{R}x\|^2 + (1-\sigma)^2\xi\Delta^2) \ge \\ &\ge \frac{1}{2}(1-\sigma)^2(\|\mathbf{R}x\|^2 + \xi\Delta^2), \end{aligned}$$

neboť $\|\mathbf{R}x\|^2 \ge (1-\sigma)^2\|\mathbf{R}x\|^2$ a podle předpokladu

$$\Big|\Delta - \|x\|\Big| \le \sigma\Delta \quad \Rightarrow \quad \|x\| \ge (1-\sigma)\Delta.$$

Do této poslední nerovnosti dosadíme (2.14) a dostaneme

$$-\psi(x) \ge (1-\sigma)^2(-\psi_\star) = (1-\sigma)^2|\psi_\star|.$$

(b) Pro $\|x\| \le \Delta$ & $\xi = 0$ dosadíme (2.13) a (2.14) a dostaneme

$$-\psi(x) = \frac{1}{2}\|\mathbf{R}x\|^2 \ge -\psi_\star$$

a protože triviálně platí $\psi(x) \ge \psi_\star$, pak $\psi(x) = \psi_\star$.

2. (a) Nechť $\|\mathbf{R}x\|^2 + \xi\Delta^2 > \sigma_2$. Pak jsou splněny předpoklady lemmatu 2.3 pro

$$\sigma = \sigma_1(2-\sigma_1) \in (0,1) \quad \text{a} \quad z = \kappa v$$

a platí

$$-\psi(x+\kappa v) \ge (1-\sigma)|\psi_\star|.$$

Odtud plyne

$$\psi(x+\kappa v) - \psi_\star \le \sigma_1(2-\sigma_1)|\psi_\star|.$$

(b) Nechť $\|\mathbf{R}x\|^2 + \xi\Delta^2 \le \sigma_2$. Jestliže $\psi_\star = \psi(x + z_\star)$, kde $\|x + z_\star\| \le \Delta$, pak (2.13) implikuje, že

$$-\psi_\star = \frac{1}{2}\left(\|\mathbf{R}x\|^2 + \xi\|x + z_\star\|^2\right) - \frac{1}{2}\|\mathbf{R}z_\star\|^2 \le \frac{1}{2}\left(\|\mathbf{R}x\|^2 + \xi\Delta^2\right).$$

Tato nerovnost, (2.13) a (2.19) implikují, že

$$\psi(x + \kappa v) = -\frac{1}{2}\left(\|\mathbf{R}x\|^2 + \xi\|x + \kappa v\|^2\right) + \frac{1}{2}\|\mathbf{R}(\kappa v)\|^2 \le \psi_\star + \frac{1}{2}\sigma_1(2 - \sigma_1)\sigma_2.$$

Odtud plyne

$$\psi(x + \kappa v) - \psi_\star \le \frac{1}{2}\sigma_1(2 - \sigma_1)\sigma_2 \le \sigma_1(2 - \sigma_1)\sigma_2.$$

Spojením obou částí plyne druhá část tvrzení lemmatu. $\square$

Nyní se vrátíme k Newtonově metodě. Z rovnice $\phi(\xi) = 0$, vlastností funkce $\phi(\xi)$ (konvexní a klesající pro $\xi \ge -\lambda_1$) a iteračního procesu (2.10) plynou tyto možnosti (viz obrázek 2.1):

1. Je-li $\xi \in (-\lambda_1, \xi_\star)$, tj. $\phi(\xi) > 0$, pak $\xi^+ \in (-\lambda_1, \xi_\star)$ a Newtonova metoda produkuje monotonně rostoucí posloupnost $\{\xi_i\}_{i\in\mathbb{N}}$ konvergující ke $\xi_\star$.
2. Je-li $\xi \in (\xi_\star, \infty)$, tj. $\phi(\xi) < 0$, pak $\xi^+ < \xi_\star$. Nemusí však být splněna podmínka $\xi^+ > -\lambda_1$, tedy $\mathbf{A} + \xi^+\mathbf{I}$ nemusí být positivně definitní.
3. Je-li $\xi \in (-\infty, -\lambda_1\rangle$, pak $\mathbf{A} + \xi\mathbf{I}$ není positivně definitní, nelze provést Choleského rozklad a musíme zajistit zvětšení ξ^+.

Proto je třeba upravit Newtonovu metodu tak, aby nedošlo k jejímu selhání, aby byla matice $\mathbf{A} + \xi^+\mathbf{I}$ positivně definitní z hlediska provedení Choleského rozkladu. Zavedeme meze ξ_L a ξ_U, pro které platí $\xi_\star \in \langle\xi_L, \xi_U\rangle$, a bod ξ_S jako dolní odhad pro $-\lambda_1$. Položíme $\xi_S := \max_{i=1,\dots,n}\{-a_{ii}\} \le -\lambda_1$, kde a_{ii} jsou diagonální prvky matice $\mathbf{A}$. Hodnoty ξ_L, ξ_U, ξ_S se během výpočtu aktualizují. Z (1.27) plyne $x_\star^T(\mathbf{A} + \xi_\star\mathbf{I})^2 x_\star = \|g\|^2$ a dále

$$\lambda_1\|x_\star\|^2 \le x_\star^T\mathbf{A}x_\star \le \lambda_n\|x_\star\|^2$$

$$\Rightarrow \quad (\lambda_1 + \xi_\star)^2\|x_\star\|^2 \le x_\star^T(\mathbf{A} + \xi_\star\mathbf{I})^2 x_\star \le (\lambda_n + \xi_\star)^2\|x_\star\|^2.$$

- Jestliže je řešení $x_\star$ na hranici, $\|x_\star\| = \Delta$, dostáváme vztah

$$\lambda_1 + \xi_\star \le \frac{\|g\|}{\Delta} \le \lambda_n + \xi_\star.$$

  Spektrum matice $\mathbf{A}$ leží pro jakoukoli multiplikativní maticovou normu v intervalu $\langle -\|\mathbf{A}\|, \|\mathbf{A}\|\rangle$, tedy

$$\frac{\|g\|}{\Delta} - \|\mathbf{A}\| \le \xi_\star \le \frac{\|g\|}{\Delta} + \|\mathbf{A}\|.$$

- Jestliže je $\|x_\star\| < \Delta$, pak $\xi_\star = 0$ a dostaneme

$$\|g\|^2 = x_\star^T\mathbf{A}^2 x_\star \le \|\mathbf{A}\|^2\|x_\star\|^2 < \|\mathbf{A}\|^2\Delta^2 \quad \Rightarrow \quad \frac{\|g\|}{\Delta} - \|\mathbf{A}\| < 0 = \xi_\star < \frac{\|g\|}{\Delta} + \|\mathbf{A}\|.$$

Z těchto úvah dostáváme pro $\xi_\star$ meze $\xi_L \geq 0$ a $\xi_U > 0$:

$$\xi_L = \max\left\{0,\, \xi_S,\, \frac{\|g\|}{\Delta} - \|\mathbf{A}\|\right\}, \quad \xi_U = \frac{\|g\|}{\Delta} + \|\mathbf{A}\| \tag{2.21}$$

a jako počáteční ξ_0 položíme $\xi \equiv \xi_0 = 0$. Kdykoli je však $\xi \leq \xi_L$, položíme

$$\xi = \max\left\{10^{-3}\xi_U,\, \sqrt{\xi_L \xi_U}\right\} \in (\xi_L, \xi_U). \tag{2.22}$$

Je to zvoleno čistě heuristicky, aby platilo $\xi \in (\xi_L, \xi_U)$. V praxi se tato volba ukázala jako velmi efektivní.

Nyní rozebereme všechny možné případy, které mohou nastat pro počáteční a další aktualizovaná ξ (viz obrázek 2.1):

1. $\xi \in (-\lambda_1, \xi_\star)$. V tomto případě je $\mathbf{A} + \xi\mathbf{I} \succ 0$, získáme Choleského rozklad $\mathbf{A} + \xi\mathbf{I} = \mathbf{R}^T\mathbf{R}$ a pomocí něho řešíme rovnici $(\mathbf{A} + \xi\mathbf{I})x = -g$, čímž dostaneme iteraci x a protože $\phi(\xi) > 0$, platí $\|x\| > \Delta$. Testujeme, zda x splňuje podmínky konvergence uvedené v bodě 1. lemmatu 2.4. Pokud ne, položíme $\xi_L = \xi$, podle (2.10) spočítáme $\xi^+ > \xi$, pro které platí $\xi^+ \in (-\lambda_1, \xi_\star)$ a dostaneme posloupnost $\{\xi_i\}_{i\in\mathbb{N}}$, která konverguje monotonně ke $\xi_\star$.

2. $\xi \in (\xi_\star, \infty)$. V tomto případě je opět $\mathbf{A} + \xi\mathbf{I} \succ 0$, čímž získáme Choleského rozklad $\mathbf{A} + \xi\mathbf{I} = \mathbf{R}^T\mathbf{R}$ a řešíme rovnici $(\mathbf{A} + \xi\mathbf{I})x = -g$. Dostaneme iteraci x a protože $\phi(\xi) < 0$, platí $\|x\| < \Delta$. Určíme vektor v, $\|v\| = 1$, který aproximuje vlastní vektor $q \in \mathcal{S}_1$, spočítáme κ podle (2.16) a testujeme podmínky konvergence uvedené v bodech 1. a 2. lemmatu 2.4. Pokud platí (2.19), získáváme přibližné řešení problému (1.25), pro který nastal singulární případ. Nejsou-li podmínky (2.17) a (2.19) splněny, pak postupujeme dále následujícím způsobem. Platí

$$\|\mathbf{R}v\|^2 = v^T(\mathbf{A} + \xi\mathbf{I})v \geq \lambda_1 + \xi \quad \Rightarrow \quad -\lambda_1 \geq \xi - \|\mathbf{R}v\|^2$$

a položíme

$$\xi_S := \max\left\{\xi_S,\, \xi - \|\mathbf{R}v\|^2\right\}, \quad \xi_L := \max\left\{\xi_L, \xi_S\right\}, \quad \xi_U := \xi. \tag{2.23}$$

Pokud je v přesně vektor $q \in \mathcal{S}_1$, platí rovnost

$$-\lambda_1 = \xi - \|\mathbf{R}v\|^2.$$

Newtonovou metodou dále spočítáme $\xi^+ < \xi$ podle vzorce (2.10), pro které platí $\xi^+ \in (-\infty, \xi_\star)$. Je-li $\xi^+ < \xi_L$, položíme podle (2.22)

$$\xi^+ = \max\left\{10^{-3}\xi_U,\, \sqrt{\xi_L \xi_U}\right\} \in (\xi_L, \xi_U) \supseteq \langle -\lambda_1, \xi_U).$$

3. $\xi \in (-\infty, -\lambda_1\rangle$. V tomto případě je $\mathbf{A} + \xi\mathbf{I} \not\succ 0$, získáme pouze částečný rozklad $\mathbf{R}_{k-1}^T\mathbf{R}_{k-1}$ a v k-tém kroku se proces zhroutí. Toto k je nejmenší takové, že hlavní $k \times k$ podmatice matice $\mathbf{A} + \xi\mathbf{I}$ není positivně definitní. Nechť

$$\mathbf{R}_{k-1} = \begin{pmatrix} r_1 & \dots & \dots & \dots & \dots & r \\ 0 & \ddots & & & & \vdots \\ \vdots & \ddots & r_{k-1} & \dots & \dots & r \\ \vdots & & \ddots & 0 & \dots & 0 \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & 0 & 0 \end{pmatrix}, \quad \mathbf{Z}_k = \begin{pmatrix} 0 & \dots & \dots & \dots & \dots & 0 \\ \vdots & & & & & \vdots \\ \vdots & & 0 & \dots & \dots & 0 \\ \vdots & & \vdots & z_k & \dots & z \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & z & \dots & z_n \end{pmatrix},$$

kde r a z značí nenulové prvky a index $\cdot_i$ znamená i-tý prvek na diagonále. Pak platí

$$\mathbf{A} + \xi\mathbf{I} = \mathbf{R}_{k-1}^T \mathbf{R}_{k-1} + \mathbf{Z}_k, \quad \text{kde} \quad z_k \le 0.$$

Definujme vektor

(2.24) $$u = (u_1, \dots, u_{k-1}, 1, 0, \dots, 0)^T \in \mathbb{R}^n,$$

jehož posledních $n - k$ složek je nula, k-tá složka je jednička a je-li $k > 1$, pak prvních $k - 1$ složek u je zvoleno tak, že u je ortogonální na $k - 1$ řádků matice $\mathbf{R}_{k-1}$, tj. $\mathbf{R}_{k-1} u = 0$. Pro takovou volbu u platí

$$u^T(\mathbf{A} + \xi\mathbf{I})u = u^T(\mathbf{R}_{k-1}^T \mathbf{R}_{k-1} + \mathbf{Z}_k)u = z_k \le 0$$

a

$$u^T(\mathbf{A} + \xi\mathbf{I})u \ge (\lambda_1 + \xi)\|u\|^2 \quad \Rightarrow \quad -\lambda_1 \ge \xi - \frac{u^T(\mathbf{A} + \xi\mathbf{I})u}{\|u\|^2}$$

Protože navíc $\mathbf{A} + \xi_\star \mathbf{I} \succeq 0$, pak

$$0 \le u^T(\mathbf{A} + \xi_\star\mathbf{I})u = u^T(\mathbf{A} + \xi\mathbf{I})u + (\xi_\star - \xi)u^T u$$

$$\Rightarrow \quad \xi_\star \ge -\lambda_1 \ge \xi - \frac{u^T(\mathbf{A} + \xi\mathbf{I})u}{\|u\|^2} \ge \xi$$

a položíme

(2.25) $$\xi_L := \xi - \frac{u^T(\mathbf{A} + \xi\mathbf{I})u}{\|u\|^2} \quad \text{a} \quad \xi^+ = \max\left\{10^{-3}\xi_U, \sqrt{\xi_L \xi_U}\right\} \in (\xi_L, \xi_U).$$

Opět se ptáme, zda je $\mathbf{A} + \xi^+\mathbf{I} \succ 0$. Pokud ne, konstruujeme posloupnost $\{\xi_i\}$, která má následující vlastnosti:

(2.26) $$\xi_U > \xi_i > \xi_L \ \ \forall i, \quad \xi_{i+1} > \xi_i, \quad \xi_{i+1} \ge \sqrt{\xi_L \xi_U} \ge \sqrt{\xi_i \xi_U},$$

přičemž v každém kroku aktualizujeme dolní mez $\xi_L \ge \xi_{i+1}$. Posloupnost $\{\xi_i\}$ má limitu, protože je rostoucí, monotonní, shora omezená a platí následující lemma.

Lemma 2.5 *Nechť je posloupnost $\{\xi_i\}$ konstruována podle (2.26). Pak platí*

$$\lim_{i \to \infty} \xi_i = \xi_U.$$

Důkaz: Z 2.26 plyne

$$\xi_{i+1} \ge \sqrt{\xi_i \xi_U} \quad \Rightarrow \quad \frac{\xi_{i+1}}{\xi_i} \ge \sqrt{\frac{\xi_U}{\xi_i}} \quad \Rightarrow \quad 1 = \lim_{i \to \infty} \frac{\xi_{i+1}}{\xi_i} \ge \lim_{i \to \infty} \sqrt{\frac{\xi_U}{\xi_i}}$$

Jelikož $\xi_i < \xi_U \ \forall i$, platí $\lim_{i \to \infty} \xi_i = \xi_U$. □

Protože pro všechna ξ_L platí $\xi_L \le -\lambda_1$, plyne z lemmatu 2.5, že pro $\xi_U > -\lambda_1$ existuje index i takový, pro který je $\xi_{i+1} > -\lambda_1$, tedy že $\mathbf{A} + \xi_{i+1}\mathbf{I} \succ 0$ a můžeme provést Choleského rozklad. Je-li $\xi_U = -\lambda_1$, pak ale $\xi_\star = -\lambda_1$ a z lemmatu plyne, že $\xi_i \to \xi_U = -\lambda_1 = \xi_\star$.

Posloupnost bodů $\{\xi_i\}$ konstruovaná podle bodů 1.-3., kde $\xi_{i+1} \equiv \xi^+$ a $\xi_{i+1} \in (\xi_L, \xi_U)$ má limitu, neboť v každém kroku zmenšujeme interval $\langle \xi_L, \xi_U \rangle$ tak, že $\xi_U - \xi_L \to 0$. Cílem je dostat $\xi_{i+1} \in (-\lambda_1, \xi_\star)$, neboť poté dostaneme monotonně rostoucí posloupnost, která konverguje ke $\xi_\star$.

Poznámka 2.2 *Výpočet nového čísla ξ^+ je kromě Newtonovy metody (2.10) možno provést dalšími jednoduššími, avšak méně výhodnými způsoby, např.:*

1. *Protože funkce $\phi(\xi)$ klesá od $\frac{1}{\Delta}$ k nule, když ξ roste od $-\lambda_1$ ke $\xi_\star$, pak z iteračního procesu (2.10) dostaneme podle (2.8) a (2.9)*

$$\xi^+ = \xi - \frac{\phi(\xi)}{\phi'(\xi)} \geq \xi_L - \frac{1}{\Delta\phi'(\xi)} \geq \xi_L + \frac{\|g\|}{\Delta}$$

   a můžeme jednodušeji položit $\xi^+ = \xi_L + \frac{\|g\|}{\Delta}$.

2. *Máme-li k dispozici ξ_U, pak můžeme položit $\xi^+ = \xi_L + \theta(\xi_U - \xi_L)$, kde např. $\theta \leq 0.1$, aby ξ^+ bylo pokud možno v intervalu $\langle -\lambda_1, \xi_\star \rangle$.*

3. *Jakmile máme položeno ξ_L a ξ_U a víme-li, že $\xi_L > -\lambda_1$, můžeme využít vlastnosti konvexní funkce $\phi(\xi)$. Metodou regula-falsi lze spočítat hodnotu ξ_F, pro kterou platí $\xi_\star < \xi_F < \xi_U$:*

$$\xi_F = \frac{\phi(\xi_L)\xi_U - \phi(\xi_U)\xi_L}{\phi(\xi_L) - \phi(\xi_U)}.$$

   Touto metodou lze rovněž spočítat $\xi_\star$.

Nyní shrneme všechny části a načrtneme algoritmus pro řešení problému (1.25).

Algoritmus 2.2 *Použití Choleského rozkladu pro výpočet lokálně omezeného kroku.*

Zvolíme $\varepsilon, \sigma_1, \sigma_2 \in (0,1)$, $\underline{\delta}, \overline{\delta}$ tak, že $0 < \underline{\delta} < 1 < \overline{\delta}$ (ε je blízko 0, $\underline{\delta}$ a $\overline{\delta}$ blízko 1).

1. *Položíme*

$$\xi_S = \max_{i=1,\dots,n} \{-a_{ii}\}, \quad \xi_L = \max\left\{0, \xi_S, \frac{\|g\|}{\Delta} - \|\mathbf{A}\|\right\}, \quad \xi_U = \frac{\|g\|}{\Delta} + \|\mathbf{A}\|, \quad \xi = 0.$$

2. *Jestliže $\xi > \xi_L$, přejdeme na krok 3. Jinak položíme $\xi = \max\left\{10^{-3}\xi_U, \sqrt{\xi_L \xi_U}\right\}$.*

3. *Je-li $\mathbf{A} + \xi\mathbf{I} \succ 0$, provedeme rozklad $\mathbf{A} + \xi\mathbf{I} = \mathbf{R}^T\mathbf{R}$ a přejdeme na krok 4. Jinak určíme vektor $u \in \mathbb{R}^n$ podle (2.24), provedeme aktualizaci podle (2.25) a opakujeme krok 3.*

4. *Určíme vektor x, který je řešením rovnice $\mathbf{R}^T\mathbf{R}x = -g$.*

   (a) *Jestliže $\|x\| > \overline{\delta}\Delta$, čili $\phi(\xi) > 0$, položíme $\xi_L = \xi$ a přejdeme na krok 6.*

   (b) *Jestliže $\underline{\delta}\Delta \leq \|x\| \leq \overline{\delta}\Delta$, pak položíme $x_\star = x$, $\xi_\star = \xi$ a* `STOP`.

   (c) *Jestliže $\|x\| < \underline{\delta}\Delta$ a $\xi < \varepsilon$, pak $x_\star = x$, $\xi_\star = 0$ a* `STOP`.

   (d) *Jestliže $\|x\| < \underline{\delta}\Delta$ a $\xi > \varepsilon$, čili $\phi(\xi) < 0$, přejdeme na krok 5.*

5. *Určíme vektor* $v \in \mathbb{R}^n$, *který aproximuje vlastní vektor* q_1 *příslušný vlastnímu číslu* λ_1 *a spočítáme číslo* κ *podle (2.16). Jestliže platí nerovnost (2.19), položíme* $x_\star = x + \kappa v$, $\xi_\star = \xi$ *a* `STOP`. *Jinak aktualizujeme meze podle (2.23) a přejdeme na krok 6.*

6. *Řešíme soustavu* $\mathbf{R}^T w = x$ *pro* w, *pomocí Newtonovy metody spočítáme* ξ^+ *podle (2.10), položíme* $\xi = \xi^+$ *a návrat na krok 2.*

Na následujícím velmi jednoduchém příkladě spočítáme jednu iteraci algoritmu 1.1. Pokud bude pro ξ^+ spočtené podle vzorce (2.10) platit $\xi^+ \notin (\xi_L, \xi_U)$, nebudeme uvažovat heuristiku (2.22), ale podle potřeby zvolíme $\xi^+ \in (\xi_L, \xi_U)$, abychom viděli možnosti chování algoritmu 2.2 pro různá ξ.

Příklad 2.1 *Nechť je dána následující dvakrát spojitě diferencovatelná a zdola omezená funkce*

$$F(y) = \frac{1}{2} y_1^2 y_2^2 + y_1^2 + 6y_1y_2 + y_2^2 - 4y_1 + 3y_2,$$

jejíž minimum (bez omezení) hledáme. Nechť je dána k-tá aproximace

$$y^{(k)} = (1,1)^T.$$

Takový vektor může být buď zvolen jako počáteční ($k=0$) *nebo spočítán z předchozí iterace po zvolení počátečního* $y^{(0)} = (0,0)^T$. *Úkolem je nalézt směrový vektor* $x_\star \in \mathbb{R}^2$ *tak, že položíme* $y^{(k+1)} = y^{(k)} + x_\star$. *Určíme tedy kvadratickou aproximaci* $\psi(x)$:

$$F'(y) = (y_1y_2^2 + 2y_1 + 6y_2 - 4,\ y_1^2y_2 + 6y_1 + 2y_2 + 3) \quad \Rightarrow \quad F'(1,1) = (5,12) \equiv g^T;$$

$$F''(y) = \begin{pmatrix} y_2^2 + 2 & 2y_1y_2 + 6 \\ 2y_1y_2 + 6 & y_1^2 + 2 \end{pmatrix} \quad \Rightarrow \quad F''(1,1) = \begin{pmatrix} 3 & 8 \\ 8 & 3 \end{pmatrix} \equiv \mathbf{G} \approx \mathbf{A},$$

takže

$$\psi(x) \equiv \psi(x_1,x_2) = \frac{1}{2} x^T \mathbf{A} x + g^T x = \frac{3}{2} x_1^2 + 8x_1x_2 + \frac{3}{2} x_2^2 + 5x_1 + 12x_2.$$

Postupujeme podle algoritmu 2.2. Vlastní čísla a vlastní vektor odpovídající nejmenšímu vlastnímu číslu matice $\mathbf{A}$ *jsou*

$$\lambda_1 = -5, \quad q_1 = \frac{1}{\sqrt{2}}(1,-1)^T \quad a \quad \lambda_2 = 11 \quad \Rightarrow \quad \|\mathbf{A}\| \geq 11.$$

Nechť aproximace v *vlastního vektoru* q_1 *je*

$$v = \frac{1}{5}(4,-3)^T.$$

Dále máme $\|g\| = 13$ *a zvolíme* $\Delta = 5$, *takže platí*

$$\xi_S = -3, \quad \xi_L = 0, \quad \xi_U \geq 13.6, \quad \textit{počáteční } \ \xi = 0 \notin (\xi_L, \xi_U).$$

Zvolíme proto např. $\xi = 1$ *a jelikož* $\mathbf{A} + 1\mathbf{I} \not\succ 0$, *dostaneme částečný rozklad*

$$\underbrace{\begin{pmatrix} 4 & 8 \\ 8 & 4 \end{pmatrix}}_{\mathbf{A}+\xi\mathbf{I}} = \underbrace{\begin{pmatrix} 2 & 0 \\ 4 & 0 \end{pmatrix}\begin{pmatrix} 2 & 4 \\ 0 & 0 \end{pmatrix}}_{\mathbf{R}_1^T\mathbf{R}_1} + \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & -12 \end{pmatrix}}_{\mathbf{Z}_2}$$

Určíme vektor u podle (2.24), který má druhou složku jedničku a je kolmý na první řádek matice $\mathbf{R}_1$*, tedy*

$$u = (-2, 1)^T.$$

Aktualizujeme

$$\xi_L = 1 + \frac{12}{5} = 3.4$$

a zvolíme $\xi \in (3.4, 13.6)$*, např.* $\xi = 4$*. Avšak opět* $\mathbf{A} + 4\mathbf{I} \not\succ 0$*, takže obdobně jako výše dostaneme*

$$\mathbf{R}_1 = \frac{1}{\sqrt{7}} \begin{pmatrix} 7 & 8 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{Z}_2 = \frac{1}{7} \begin{pmatrix} 0 & 0 \\ 0 & -15 \end{pmatrix}, \quad u = \frac{1}{7}(-8, 7)^T, \quad \xi_L \doteq 4.9$$

a zvolíme $\xi \in (4.9, 13.6)$*. Vidíme, jak* ξ_L *směřuje k hodnotě* $-\lambda_1 = 5$*.*

Nechť nyní $\xi = 13$*. Pak je již* $\mathbf{A} + 13\mathbf{I} \succ 0$*, můžeme provést rozklad a řešíme systém* $(\mathbf{A} + 13\mathbf{I})x = \mathbf{R}^T\mathbf{R}x = -g$*. Dostaneme*

$$\mathbf{R} = \begin{pmatrix} 4 & 2 \\ 0 & \sqrt{12} \end{pmatrix}, \quad x = \frac{1}{24}(2, -19)^T.$$

Protože $\|x\| < \Delta$*, je* $\xi = 13 > \xi_\star$*, takže aktualizujeme meze podle (2.23) takto (předpokládáme, že nerovnost (2.19) není splněna):*

$$\xi - \|\mathbf{R}v\|^2 = 13 - \left[2^2 + \left(\frac{3\sqrt{12}}{5} \right)^2 \right] = 4.68 \leq -\lambda_1,$$

tedy

$$\xi_S = 4.68 \quad \Rightarrow \quad \xi_L = 4.9 \quad a \quad \xi_U = 13.$$

Řešíme soustavu $\mathbf{R}^T w = x$*, dostaneme*

$$w = \left(\frac{1}{48}, -\frac{5}{6\sqrt{12}} \right)^T,$$

takže podle (2.10) platí

$$\xi := \xi + \frac{\|x\|^2}{\|w\|^2} \cdot \left(\frac{\|x\| - \Delta}{\Delta} \right) \doteq 3,86 \notin (\xi_L, \xi_U).$$

Zvolíme tedy $\xi \in (4.9, 13)$*, např.* $\xi = 10$*. Protože* $\mathbf{A} + 10\mathbf{I} \succ 0$*, dostaneme*

$$\mathbf{R} \doteq \frac{1}{9} \begin{pmatrix} 32.45 & 20 \\ 0 & 25.5 \end{pmatrix}, \quad x \doteq (0.3, -1.1)^T, \quad \|x\| = \sqrt{1.3} < \Delta,$$

takže opět $\xi = 10 > \xi_\star$ *a dále*

$$\mathbf{R}v = \frac{1}{45}(69.8, -76.5)^T, \quad w \doteq \left(\frac{1}{12}, -\frac{9}{20} \right)^T$$

a protože $\|x\| < \Delta$*, aktualizujeme*

$$\xi_S = \max\{4.68,\, \xi - \|\mathbf{R}v\|^2\} \doteq 4.7, \quad \xi_L = 4.9, \quad \xi_U = 10, \quad \xi \overset{(2.10)}{\doteq} 5.2 \in (\xi_L, \xi_U).$$

Tentokrát Newtonova metoda poskytla $\xi \in (\xi_L, \xi_U)$ *a protože* $\mathbf{A} + 5.2\mathbf{I} \succ 0$, *dostaneme*

$$\mathbf{R} \doteq \begin{pmatrix} 2.86 & 2.8 \\ 0 & 0.6 \end{pmatrix}, \quad x \doteq (17, -18)^T, \quad w \doteq (5.944, -57.739)^T.$$

Jelikož nyní platí $\|x\| > \Delta$, *nemusíme počítat* $\|\mathbf{R}v\|^2$ *a navíc platí* $\xi = 5.2 \in (-\lambda_1, \xi_\star)$, *takže je od této chvíle záležitost již jednoduchá, protože Newtonova aktualizace (2.10) povede přímo k řešení* $\xi_\star$. *Tedy*

$$\xi_L = \xi = 5.2, \quad \xi \overset{(2.10)}{\doteq} 5.9,$$

takže pro $\mathbf{A} + 5.9\mathbf{I}$ *platí*

$$\mathbf{R} \doteq \begin{pmatrix} 2.98 & 2.68 \\ 0 & 1.31 \end{pmatrix}, \quad x \doteq (3.4, -4.4)^T, \quad w \doteq (1.141, -5.693)^T.$$

Dále

$$\|x\| > \Delta \quad \Rightarrow \quad \xi_L = \xi = 5.9, \quad \xi \overset{(2.10)}{\doteq} 6.$$

Pro hodnotu $\xi = 6$ *dostaneme* $x = (3, -4)^T$ *a protože* $\|x\| = 5$, *získáváme řešení* $x_\star$.

Položíme tedy

$$y^+ = y^{(k)} + (3, -4)^T = (4, -3)^T$$

a pokračujeme v algoritmu 1.1 tím, že testujeme hodnotu podílu skutečného a předpověděného poklesu funkce F :

$$\varrho(x_\star) = \frac{F(y^+) - F(y^{(k)})}{\psi(x_\star)} = \frac{F(4, -3) - F(1, 1)}{\psi(3, -4)} = \frac{0 - 7.5}{-91.5} \doteq 0.082 > 0,$$

takže můžeme položit

$$y^{(k+1)} = (4, -3)^T.$$

Protože však je $\varrho(x_\star)$ *blízko nuly, lze usoudit, že poloměr* $\Delta = 5$ *je příliš velký. Pro další iteraci proto v souladu s podmínkou (1.15) poloměr* Δ *snížíme.*

Funkce F *nabývá globálního minima* $F(y_\star) \doteq -23.3911$ *v bodě* $y_\star \doteq (2.5503, -2.1521)^T$, *takže krok v oblasti o poloměru* $\Delta = 5$ *je skutečně velký. Položíme-li* $k = 0$, *může posloupnost iterací podle algoritmu 1.1 vypadat například tak, jak je uvedeno v tabulce 2.1.*

k	Δ	$\xi_\star$	$x_{\star 1}$	$x_{\star 2}$	$\psi(x_\star)$	$\varrho(x_\star)$	y_1	y_2	$F(y)$	$\|g(y)\|$
0	–	–	–	–	–	–	1	1	7.5	13
1	5	6	3	-4	-91.5	0.082	4	-3	0	34.83
2	3	3.906	1.61	2.55	-34.55	0.118	5.61	-0.45	-4.08	22.33
3	2	0.464	-1.913	-0.583	-12.63	1.16	3.697	-1.033	-18.77	9.07
4	2.5	0	-0.72	-0.65	-3.33	1.2	2.977	-1.683	-22.77	2.6
5	2.5	0	-0.372	-0.375	-0.54	1.1	2.605	-2.058	-23.37	0.56
6	2.5	0	-0.051	-0.09	-0.022	1	2.554	-2.148	-23.391	0.02

Tabulka 2.1: Posloupnost iterací pro příklad 2.1.

Tímto příkladem jsme demonstrovali aplikaci algoritmu 2.2. Na závěr ukážeme, že je tato metoda globálně konvergentní.

Věta 2.1 *Metoda použití Choleského rozkladu, sestavená na základě algoritmu 2.2, je globálně konvergentní.*

DŮKAZ: Použijeme poznámku 1.3. Protože hledáme optimální lokálně omezený krok $x_\star$ podle podmínek věty 1.5, je splněna nerovnost (1.11) podle věty 1.4. Kromě toho používáme matice $\mathbf{A} \equiv \mathbf{A}_k = \mathbf{G}(y_k)$, takže z (1.7) plyne $\|\mathbf{A}_k\| = \|\mathbf{G}(y_k)\| \le \overline{G}$ a je splněna podmínka (1.16). □

2.2 Použití lineární kombinace vlastních vektorů

Známe-li vlastní vektory matice $\mathbf{A}$, lze řešení $x_\star$ problému (1.25) nalézt jako lineární kombinaci těchto vlastních vektorů, [25]. Uvažujme rozklad $\mathbf{A} = \mathbf{QDQ}^T$, kde $\mathbf{D}$ je diagonální matice s prvky $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$ a $\mathbf{Q}$ je matice, jejíž sloupce $q_1, q_2, \ldots, q_n$ jsou ortonormální vlastní vektory $\mathbf{A}$. Definujme $g = \sum_{i=1}^n \vartheta_i q_i$.

Lemma 2.6 *Vektor* $x = -\sum_{i=1}^n c_i q_i$ *je řešením (1.25), je-li vektor* $c = (c_1, \ldots, c_n)^T$ *zvolen následujícím způsobem:*

1. *Nechť* $\lambda_1 > 0$. *Jestliže*
$$\sum_{i=1}^n \frac{\vartheta_i^2}{\lambda_i^2} \le \Delta^2,$$
pak
$$c_i = \frac{\vartheta_i}{\lambda_i}.$$

2. *Jestliže* $\lambda_1 \le 0$, g *je ortogonální na množinu vlastních vektorů odpovídajících* λ_1 *a*
$$\sum_{\lambda_i \ne \lambda_1} \frac{\vartheta_i^2}{(\lambda_i - \lambda_1)^2} \le \Delta^2,$$
pak
$$c_i = \frac{\vartheta_i}{(\lambda_i - \lambda_1)} \quad \textit{pro} \quad \lambda_i \ne \lambda_1,$$
zatímco zbývající c_i *jsou libovolná čísla splňující podmínku*
$$\sum_{\lambda_i = \lambda_1} c_i^2 = \Delta^2 - \sum_{\lambda_i \ne \lambda_1} \frac{\vartheta_i^2}{(\lambda_i - \lambda_1)^2}.$$

3. *Nejsou-li podmínky 1. nebo 2. splněny, pak*
$$c_i = \frac{\vartheta_i}{\lambda_i + \xi},$$
kde $\xi > \max\{0, -\lambda_1\}$ *je zvoleno tak, aby platilo*
$$\sum_{i=1}^n \frac{\vartheta_i^2}{(\lambda_i + \xi)^2} = \Delta^2.$$

Důkaz:

1. Všechna vlastní čísla matice $\mathbf{A}$ jsou kladná, $\mathbf{A}$ je tedy positivně definitní a platí

$$\mathbf{A}x = -\sum_{i=1}^{n} c_i \mathbf{A} q_i = -\sum_{i=1}^{n} c_i \lambda_i q_i = -\sum_{i=1}^{n} \vartheta_i q_i = -g.$$

   Kromě toho,

$$\|x\|^2 = x^T x = \sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} \frac{\vartheta_i^2}{\lambda_i^2} \leq \Delta^2.$$

   Položíme-li $\xi = 0$, je podle věty 1.5 x řešením (1.25).

2. Toto je tzv. singulární případ, $g \perp \mathcal{S}_1$. Pro $\xi = -\lambda_1$ je $\mathbf{A} + \xi\mathbf{I} \succeq 0$ a ϑ_i odpovídající λ_1 jsou podle lemmatu 2.2 rovna nule. Dále

$$\begin{aligned} (\mathbf{A} - \lambda_1 \mathbf{I})x &= -(\mathbf{A} - \lambda_1 \mathbf{I}) \sum_{i=1}^{n} c_i q_i = -\sum_{i=1}^{n} c_i \lambda_i q_i + \sum_{i=1}^{n} c_i \lambda_1 q_i = \\ &= -\sum_{\lambda_i \neq \lambda_1} c_i (\lambda_i - \lambda_1) q_i = -\sum_{\lambda_i \neq \lambda_1} \vartheta_i q_i = -g. \end{aligned}$$

   Mimoto, čísla c_i jsou volena tak, že $\|x\| = \Delta$. Opět jsou splněny předpoklady věty 1.5 a x je řešením (1.25).

3. Pro x platí $\|x\| = \Delta$ a $\mathbf{A} + \xi\mathbf{I} \succeq 0$. Obdobně

$$(\mathbf{A} + \xi\mathbf{I})x = -(\mathbf{A} + \xi\mathbf{I}) \sum_{i=1}^{n} c_i q_i = -\sum_{i=1}^{n} c_i (\lambda_i + \xi) q_i = -\sum_{i=1}^{n} \vartheta_i q_i = -g$$

   a proto je x řešením (1.25). □

Ve třetím případě lze pro ξ získat horní a dolní odhad. Protože $\lambda_i + \xi \geq \lambda_1 + \xi > 0$, platí

$$\Delta^2 = \sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_i + \xi)^2} \leq \sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_1 + \xi)^2} = \frac{\|g\|^2}{(\lambda_1 + \xi)^2}.$$

Odtud plyne horní mez

$$\xi \leq \frac{\|g\|}{\Delta} - \lambda_1 =: \xi_U. \tag{2.27}$$

Dále nechť $\lambda_1 = \lambda_2 = \ldots = \lambda_l$ pro nějaké $l \geq 1$. Potom

$$\Delta^2 = \sum_{i=1}^{n} \frac{\vartheta_i^2}{(\lambda_i + \xi)^2} \geq \sum_{i=1}^{l} \frac{\vartheta_i^2}{(\lambda_i + \xi)^2} = \frac{1}{(\lambda_1 + \xi)^2} \sum_{i=1}^{l} \vartheta_i^2,$$

což vede na dolní mez

$$\xi \geq -\lambda_1 + \frac{1}{\Delta} \sqrt{\sum_{i=1}^{l} \vartheta_i^2} =: \xi_L. \tag{2.28}$$

Nyní přetransformujeme problém (1.25) na ekvivalentní problém, na který snadno aplikujeme lemma 2.6. Matici $\mathbf{A}$ převedeme nejprve na třídiagonální matici pomocí Arnoldiho algoritmu a nakonec provedeme diagonalizaci.

Uvažujme nejprve obecnou matici $\mathbf{M} \in \mathbb{R}^{n\times n}$. Pro její převod na horní Hessenbergův tvar a výpočet ortonormálních vektorů použijeme Arnoldiho algoritmus, který generuje ortonormální vektory $v_1, \ldots, v_k$.

Algoritmus 2.3 *Arnoldiho algoritmus.*

Zvolíme v_1, *kde* $\|v_1\| = 1$, *položíme* $k = 1$ *a provedeme*

1. $h_{j,k} = v_j^T \mathbf{M} v_k, \quad j = 1, 2, \ldots, k.$

2. $y = \mathbf{M} v_k - \sum_{j=1}^{k} h_{j,k} v_j.$

3. $h_{k+1,k} = \|y\|.$

4. *Je-li* $h_{k+1,k} = 0$, *pak* `STOP`.

5. $v_{k+1} = \frac{1}{h_{k+1,k}}\, y.$

6. Položíme $k := k + 1$ *a návrat na krok 1.*

Označme $\mathbf{V}_k = (v_1, \ldots, v_k)$ a

$$\tilde{\mathbf{H}}_k = \begin{pmatrix} h_{1,1} & h_{1,2} & \ldots & \ldots & h_{1,k} \\ h_{2,1} & h_{2,2} & \ldots & \ldots & h_{2,k} \\ & h_{3,2} & \ldots & \ldots & h_{3,k} \\ & & \ddots & & \vdots \\ & & & h_{k,k-1} & h_{k,k} \\ & & & & h_{k+1,k} \end{pmatrix}$$

Matice $\mathbf{V}_k$ je ortonormální a $\tilde{\mathbf{H}}_k \in \mathbb{R}^{(k+1)\times k}$ je horní Hessenbergova matice. Arnoldiho proces lze napsat v maticovém tvaru

$$\mathbf{M}\mathbf{V}_k = \mathbf{V}_{k+1}\tilde{\mathbf{H}}_k. \tag{2.29}$$

Buď $\mathbf{H}_k \in \mathbb{R}^{k\times k}$ čtvercová matice, která vznikne z matice $\tilde{\mathbf{H}}_k$ vynecháním posledního řádku. Pak z ortogonality plyne vztah $\mathbf{V}_k^T \mathbf{M} \mathbf{V}_k = \mathbf{H}_k$. V případě, že $\mathbf{M}$ je naše symetrická matice $\mathbf{A}$, pak je $\mathbf{H}_k$ symetrická Hessenbergova matice, tudíž je třídiagonální, kterou označíme $\mathbf{T}_k$.

Na matici $\mathbf{A}$ aplikujeme Arnoldiho algoritmus, který začíná s vektorem $v_1 = \frac{1}{\|g\|}\, g$, končí pro nějaké $k \le n$ a využijeme vztah $\mathbf{V}_k^T \mathbf{A} \mathbf{V}_k = \mathbf{T}_k$. Označíme-li pro jednoduchost $\mathbf{V} \equiv \mathbf{V}_k$ a $\mathbf{T} \equiv \mathbf{T}_k$, dostaneme substitucí $x = \mathbf{V}\bar{x}$ v (1.25) problém

$$\min_{\bar{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\, \bar{x}^T \mathbf{T} \bar{x} + \bar{g}^T \bar{x} \right\} \quad \text{vzhledem k} \quad \|\bar{x}\| \le \Delta, \tag{2.30}$$

kde $\bar{g} = \mathbf{V}^T g$. Nechť nakonec $\mathbf{T} = \mathbf{B}\mathbf{D}\mathbf{B}^T$ je diagonalizace $\mathbf{T}$, kde $\mathbf{B}$ je ortonormální a $\mathbf{D}$ je diagonální matice. Substituce $\bar{x} = \mathbf{B}\bar{\bar{x}}$ v (2.30) vede na diagonální verzi

$$\min_{\bar{\bar{x}}\in\mathbb{R}^n} \left\{ \frac{1}{2}\, \bar{\bar{x}}^T \mathbf{D} \bar{\bar{x}} + \bar{\bar{g}}^T \bar{\bar{x}} \right\} \quad \text{vzhledem k} \quad \|\bar{\bar{x}}\| \le \Delta, \tag{2.31}$$

kde $\bar{\bar{g}} = \mathbf{B}^T\bar{g} = \mathbf{B}^T\mathbf{V}^T g$. Tento diagonální problém je snadno řešitelný užitím lemmatu 2.6 a mezí ξ_L, ξ_U podle (2.27) a (2.28). Pro rozklad matice $\mathbf{A}$ platí

$$\mathbf{V}^T\mathbf{A}\mathbf{V} = \mathbf{T} = \mathbf{B}\mathbf{D}\mathbf{B}^T \quad \Rightarrow \quad \mathbf{A} = \mathbf{V}\mathbf{B}\mathbf{D}(\mathbf{V}\mathbf{B})^T \equiv \mathbf{Q}\mathbf{D}\mathbf{Q}^T,$$

kde $\mathbf{D}$ je diagonální matice s vlastními čísly $\mathbf{A}$.

Algoritmus 2.4 *Použití vlastních párů pro výpočet lokálně omezeného kroku.*

1. *Položíme* $v_1 = \frac{1}{\|g\|} g$.
2. *Pomocí Arnoldiho algoritmu 2.3 provedeme rozklad* $\mathbf{V}^T\mathbf{A}\mathbf{V} = \mathbf{T}$.
3. *Provedeme rozklad* $\mathbf{T} = \mathbf{B}\mathbf{D}\mathbf{B}^T$.
4. *Položíme* $g := (\mathbf{V}\mathbf{B})^T g \equiv \mathbf{Q}^T g = \sum_{i=1}^k \vartheta_i q_i$.
5. *Spočítáme* ξ_L *a* ξ_U *podle (2.27) a (2.28).*
6. *Pomocí lemmatu 2.6 spočítáme* x *a* ξ.
7. *Položíme* $x_\star = \mathbf{V}\mathbf{B}x \equiv \mathbf{Q}x$ *a* $\xi_\star = \xi$.

K dosažení dané chybové tolerance spočítaného řešení však může být zapotřebí velké k, což způsobuje uložení velké plné matice $\mathbf{V}$ i dlouhý výpočetní čas jedné iterace.

Věta 2.2 *Metoda použití vlastních párů, sestavená na základě algoritmu 2.4, je globálně konvergentní.*

DŮKAZ: Vektor x je řešením diagonálního problému (2.31) a podle lemmatu 2.6 splňuje podmínky věty 1.5 pro matici $\mathbf{D}$ a vektor $\bar{\bar{g}}$. Je tedy optimálním lokálně omezeným krokem, pro který platí nerovnost (1.11) potřebná ke globální konvergenci (poznámka 1.3). Protože však platí $\|x_\star\| = \|x\|$, matice $\mathbf{D}$ má stejná vlastní čísla jako $\mathbf{A}$ a

$$(\mathbf{A} + \xi\mathbf{I})x_\star + g = \mathbf{V}\mathbf{B}(\mathbf{D} + \xi\mathbf{I})(\mathbf{V}\mathbf{B})^T x_\star + g = \mathbf{V}\mathbf{B}[(\mathbf{D} + \xi\mathbf{I})x + \bar{\bar{g}}] = 0,$$

platí totéž i pro $x_\star$. □

2.3 Metoda psí nohy

Výpočet optimálního lokálně omezeného kroku je poměrně náročný a proto byly vyvinuty jednodušší metody, které hledají pouze přibližné řešení. Přesné řešení úlohy (1.25) nahradíme přibližným řešením

$$x = \tau_1 g + \tau_2 \mathbf{A}^{-1} g \tag{2.32}$$

na podprostoru generovaném vektory g a $\mathbf{A}^{-1}g$. Metoda psí nohy, vyvinutá Powellem, nebo modifikovaná metoda psí nohy Dennise a Meie, [08], [11], [34], [35], [45], [49], [60], [68], [69], aproximuje řešení $x_\star$ tak, že generuje po částech lineární křivku $x(\tau)$ s vlastností, že $\psi(x(\tau))$ je monotonně klesající a $\|x(\tau)\|$ je monotonně rostoucí pro $0 \le \tau \le T$. Přibližné řešení (1.25) je dáno takto:

$$\min\{\psi(x(\tau)),\ \|x(\tau)\| \le \Delta,\ 0 \le \tau \le T\}. \tag{2.33}$$

V našich úvahách budeme pracovat s Cauchyho bodem $x_C = \alpha_1 g$ a Newtonovým bodem $x_N = \alpha_2 \mathbf{A}^{-1} g$ a k efektivnímu využití této metody je třeba mít určité předpoklady na matici $\mathbf{A}$.

Předpokládejme nejprve, že $g^T \mathbf{A} g \leq 0$. Pak matice $\mathbf{A}$ není positivně definitní, položíme $x_\star = -\frac{\Delta}{\|g\|} g$ a ukončíme výpočet. Jestliže $g^T \mathbf{A} g > 0$, budeme uvažovat Cauchyho bod $x_C = \alpha_1 g$, bod z 1-dimensionálního podprostoru $\mathrm{sp}\{g\}$, ve kterém nabývá funkce ψ lokálního minima. Dosadíme-li do funkce $\psi(x_C)$, zderivujeme a položíme rovno nule, dostaneme $\alpha_1 = -\frac{g^T g}{g^T \mathbf{A} g}$.

Nechť nyní platí $\|x_C\| \geq \Delta$. Pak položíme $x_\star = -\frac{\Delta}{\|g\|} g$ a ukončíme výpočet, protože nemá význam počítat Newtonův bod x_N, neboť platí $\|x_C\| \leq \|x_N\|$, jak uvidíme dále. V obou těchto případech ve vztahu (2.32) pro $x_\star$ platí $\tau_1 = -\frac{\Delta}{\|g\|}$ a $\tau_2 = 0$.

Nechť tedy $\|x_C\| < \Delta$. Přibližné řešení (1.25) budeme uvažovat ve tvaru

$$x_\star = x_C + \kappa v, \tag{2.34}$$

kde v je konkrétně zvolený vektor a κ je určeno tak, že $\|x_\star\| = \Delta$.

Jestliže není matice $\mathbf{A}$ regulární, nelze spočítat Newtonův bod x_N. V tom případě zvolíme vektor v tak, aby

$$v^T \mathbf{A} v \leq 0, \quad v^T(\mathbf{A} x_C + g) \leq 0, \quad \|v\| = 1$$

(druhá podmínka udává pouze orientaci vektoru v) a $\kappa \geq 0$ určíme tak, aby $\|x_\star\| = \Delta$. Pro takové $x_\star$ platí

$$\begin{aligned} \psi(x_\star) - \psi(x_C) &= \frac{1}{2}(x_C + \kappa v)^T \mathbf{A}(x_C + \kappa v) + g^T(x_C + \kappa v) - \frac{1}{2} x_C^T \mathbf{A} x_C - g^T x_C = \\ &= \kappa x_C^T \mathbf{A} v + \frac{1}{2}\kappa^2 v^T \mathbf{A} v + \kappa g^T v = \kappa v^T(\mathbf{A} x_C + g) + \frac{1}{2}\kappa^2 v^T \mathbf{A} v \leq 0. \end{aligned} \tag{2.35}$$

Je to jediný případ, kdy $x_\star$ nemá tvar (2.32). Vektor v se určí pomocí zobecněného Choleského rozkladu, např. Gilla a Murraye [18] nebo Buncha a Parletta [02].

Nyní se dostáváme k případu, kdy je matice $\mathbf{A}$ regulární a lze spočítat Newtonův bod x_N. Tento bod splňuje $\mathbf{A} x_N + g = 0$, takže $\alpha_2 = -1$. Matice $\mathbf{A}^{-1}$ se nepočítá, provede se opět Choleského rozklad nebo jeho zobecnění. Jestliže platí $\|x_N\| \leq \Delta$, ukončíme výpočet pro $x_\star = x_N$, které je optimálním řešením problému (1.25).

Dostáváme se k poslednímu a nejdůležitějšímu případu, kdy $\|x_C\| < \Delta < \|x_N\|$, kde

$$x_C = -\frac{g^T g}{g^T \mathbf{A} g} g, \quad x_N = -\mathbf{A}^{-1} g \tag{2.36}$$

a $g^T \mathbf{A} g > 0$. Nejprve stanovíme podmínky, které určují, kdy je funkce $\psi(x)$ monotonně klesající na úsečce spojující body x_C a x_N. Libovolný bod na této úsečce označíme

$$x(\vartheta) = x_C + \vartheta(x_N - x_C), \quad \vartheta \in \langle 0, 1 \rangle.$$

Lemma 2.7 *Platí:*

$$\frac{\partial \psi(x(\vartheta))}{\partial \vartheta} = (1 - \vartheta)(x_N - x_C)^T(\mathbf{A} x_C + g). \tag{2.37}$$

DŮKAZ: Protože

$$\psi(x(\vartheta)) = \frac{1}{2}\,(x_C + \vartheta(x_N - x_C))^T\mathbf{A}(x_C + \vartheta(x_N - x_C)) + g^T(x_C + \vartheta(x_N - x_C)),$$

pak

$$\begin{aligned}\frac{\partial\psi(x(\vartheta))}{\partial\vartheta} &= (x_N - x_C)^T\mathbf{A}(x_C + \vartheta(x_N - x_C)) + g^T(x_N - x_C) = \\ &= (x_N - x_C)^T\mathbf{A}(x_C + \vartheta(x_N - x_C)) - (x_N - x_C)^T\mathbf{A}x_N = \\ &= (1-\vartheta)(x_N - x_C)^T\mathbf{A}(x_C - x_N) = (1-\vartheta)(x_N - x_C)^T(\mathbf{A}x_C + g).\end{aligned}$$

□

Definice 2.1 *Definujme číslo*

$$\beta = \frac{(g^Tg)^2}{g^T\mathbf{A}gg^T\mathbf{A}^{-1}g} = \frac{x_C^Tx_C}{x_N^Tx_C} \tag{2.38}$$

Lemma 2.8 *Nechť* $g^T\mathbf{A}g > 0$. *Pak jsou následující podmínky ekvivalentní:*

1. $(x_N - x_C)^Tx_C > 0$,
2. $\beta \in (0, 1)$,
3. $(x_N - x_C)^T\mathbf{A}(x_N - x_C) > 0$,
4. $(x_N - x_C)^T(\mathbf{A}x_C + g) < 0$.

DŮKAZ: Nechť $(x_N - x_C)^Tx_C > 0$. Pak podle (2.36) platí

$$(x_N - x_C)^Tx_C = \frac{g^Tg}{(g^T\mathbf{A}g)^2}\left(g^T\mathbf{A}gg^T\mathbf{A}^{-1}g - (g^Tg)^2\right) > 0 \quad\Rightarrow\quad g^T\mathbf{A}gg^T\mathbf{A}^{-1}g - (g^Tg)^2 > 0.$$

Z této implikace plyne

$$g^T\mathbf{A}gg^T\mathbf{A}^{-1}g > (g^Tg)^2 > 0 \quad\Rightarrow\quad g^T\mathbf{A}^{-1}g > 0.$$

Můžeme tedy vydělit členem $g^T\mathbf{A}gg^T\mathbf{A}^{-1}g$, aniž se změní znaménko nerovnosti a dostaneme

$$0 < \beta < 1.$$

Naopak, jestliže $0 < \beta < 1$, pak platí $(x_N - x_C)^Tx_C > 0$. Podobně

$$\begin{aligned}(x_N - x_C)^T\mathbf{A}(x_N - x_C) &= g^T\mathbf{A}^{-1}g - 2\,\frac{(g^Tg)^2}{g^T\mathbf{A}g} + \frac{(g^Tg)^2}{g^T\mathbf{A}g} = \\ &= \frac{1}{g^T\mathbf{A}g}\,(g^T\mathbf{A}gg^T\mathbf{A}^{-1}g - (g^Tg)^2) > 0 \quad\Leftrightarrow\quad 0 < \beta < 1.\end{aligned}$$

Konečně pro $g = -\mathbf{A}x_N$ máme

$$(x_N - x_C)^T(\mathbf{A}x_C + g) = -(x_N - x_C)^T\mathbf{A}(x_N - x_C),$$

takže všechny podmínky jsou ekvivalentní. □

Důsledek 2.1 *Nechť* $g^T \mathbf{A} g > 0$ *a nechť je splněna libovolná podmínka v lemmatu 2.8. Pak* $g^T \mathbf{A}^{-1} g > 0$, *tedy* $x_N^T x_C > 0$ *a* β *je dobře definováno.*

Z (2.37) a lemmatu 2.8 je patrné, že chování funkce $\psi(x)$ na úsečce $\langle x_C, x_N \rangle$ závisí na skalárním součinu $(x_N - x_C)^T x_C$.

Předpokládejme nejprve, že $(x_N - x_C)^T x_C \leq 0$. Pak $\psi(x)$ není na úsečce $\langle x_C, x_N \rangle$ monotonně klesající a položíme $x_\star$ ve tvaru (2.34) pro $v = -(x_N - x_C)$ a $\kappa > 0$ určíme tak, že $\|x_\star\| = \Delta$. Pro takto zvolené $x_\star$ obdobnou úpravou jako u (2.35) dostaneme

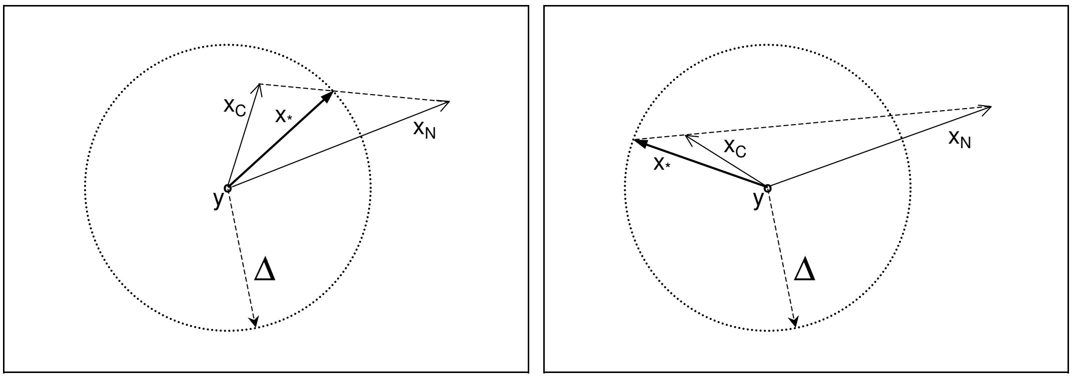
$$\psi(x_\star) - \psi(x_C) = -\kappa (x_N - x_C)^T (\mathbf{A} x_C + g) + \frac{1}{2} \kappa^2 (x_N - x_C)^T \mathbf{A} (x_N - x_C) \leq 0 \tag{2.39}$$

podle lemmatu 2.8 (pozor na ekvivalenci, neboť předpokládáme $(x_N - x_C)^T x_C \leq 0$).

Zbývá uvažovat možnost, kdy $(x_N - x_C)^T x_C > 0$. Funkce $\psi(x)$ je v tomto případě na úsečce $\langle x_C, x_N \rangle$ monotonně klesající a položíme $x_\star$ ve tvaru (2.34) pro $v = x_N - x_C$, které splňuje (2.33). Křivku $x(\tau)$ metody psí nohy pro $T = 2$ definujeme tak, že spojíme body x_C a x_N :

$$x(\tau) = \begin{cases} \tau x_C, & 0 \leq \tau \leq 1; \\ x_C + (\tau - 1)(x_N - x_C), & 1 \leq \tau \leq 2. \end{cases} \tag{2.40}$$

Na obrázku 2.2 jsou znázorněny oba případy:
vlevo $(x_N - x_C)^T x_C > 0$ a vpravo $(x_N - x_C)^T x_C \leq 0$ a způsob výběru řešení $x_\star$.



Obrázek 2.2: Řešení metody psí nohy

Bod x_N je optimálním řešením problému (1.25), pokud je $\|x_N\| \leq \Delta$. Protože však zde je jeho norma větší než Δ, budeme se snažit k tomuto bodu alespoň přiblížit. Pokud je tedy $(x_N - x_C)^T x_C > 0$, provedeme modifikaci metody psí nohy tak, aby bylo zajištěno přiblížení k Newtonovu bodu x_N.

Cauchyho bod je definován tak, že funkce $\psi(x)$ klesá monotonně do bodu x_C. Obdobně pro Newtonův bod platí, že $\psi(x)$ klesá monotonně do bodu x_N. Zavedeme bod γx_N, kde $\gamma \in \langle 0, 1 \rangle$ tak, aby byla splněna podmínka, že funkce $\psi(x)$ je monotonně klesající na úsečce $\langle x_C, \gamma x_N \rangle$. Konstrukcí ukážeme, že existuje celá množina takových čísel γ s výše uvedenou vlastností. Označme

$$x(\vartheta) = x_C + \vartheta (\gamma x_N - x_C), \quad \vartheta \in \langle 0, 1 \rangle$$

libovolný bod na úsečce $\langle x_C, \gamma x_N \rangle$ a hledáme γ takové, že

$$\frac{\partial \psi(x(\vartheta))}{\partial \vartheta} < 0 \quad \forall \vartheta \in (0, 1).$$

Takže

$$\begin{aligned}
x(\vartheta) &= -\frac{g^T g}{g^T \mathbf{A} g}\, g + \vartheta\Big(-\gamma \mathbf{A}^{-1} g + \frac{g^T g}{g^T \mathbf{A} g}\, g\Big) = (\vartheta - 1)\frac{g^T g}{g^T \mathbf{A} g}\, g - \vartheta\gamma \mathbf{A}^{-1} g \\
\frac{\partial x(\vartheta)}{\partial \vartheta} &= \frac{g^T g}{g^T \mathbf{A} g}\, g - \gamma \mathbf{A}^{-1} g \\
\psi(x(\vartheta)) &= \frac{1}{2}\, x(\vartheta)^T \mathbf{A} x(\vartheta) + g^T x(\vartheta),
\end{aligned}$$

což po dosazení a úpravě dává

$$\begin{aligned}
\frac{\partial \psi(x(\vartheta))}{\partial \vartheta} &= x(\vartheta)^T \mathbf{A} \frac{\partial x(\vartheta)}{\partial \vartheta} + g^T \frac{\partial x(\vartheta)}{\partial \vartheta} = \\
&= \left[(\vartheta - 1)\frac{g^T g}{g^T \mathbf{A} g}\, g - \vartheta\gamma \mathbf{A}^{-1} g\right]^T \mathbf{A} \left[\frac{g^T g}{g^T \mathbf{A} g}\, g - \gamma \mathbf{A}^{-1} g\right] + g^T \left[\frac{g^T g}{g^T \mathbf{A} g}\, g - \gamma \mathbf{A}^{-1} g\right] = \\
&= (\vartheta - 1)\left[\frac{(g^T g)^2}{g^T \mathbf{A} g} - \gamma \frac{(g^T g)^2}{g^T \mathbf{A} g}\right] - \vartheta\gamma \frac{(g^T g)^2}{g^T \mathbf{A} g} + \vartheta\gamma^2 g^T \mathbf{A}^{-1} g + \frac{(g^T g)^2}{g^T \mathbf{A} g} - \gamma g^T \mathbf{A}^{-1} g = \\
&= g^T \mathbf{A}^{-1} g \left[(\vartheta - 1)\beta - (\vartheta - 1)\gamma\beta - \vartheta\gamma\beta + \vartheta\gamma^2 + \beta - \gamma\right] = \\
&= g^T \mathbf{A}^{-1} g \left[\vartheta(\gamma^2 - 2\gamma\beta + \beta) + \gamma(\beta - 1)\right].
\end{aligned}$$

Tato funkce je lineární a rostoucí, protože její derivace podle ϑ je

$$\frac{\partial^2 \psi(x(\vartheta))}{\partial \vartheta^2} = g^T \mathbf{A}^{-1} g \left[(\gamma - \beta)^2 + \beta(1 - \beta)\right] > 0,$$

neboť $\beta \in (0,1)$ a $g^T \mathbf{A}^{-1} g > 0$ podle důsledku 2.1. Tedy

$$\frac{\partial \psi(x(\vartheta))}{\partial \vartheta} < \left.\frac{\partial \psi(x(\vartheta))}{\partial \vartheta}\right|_{\vartheta=1} = g^T \mathbf{A}^{-1} g \left[\gamma^2 - (\beta + 1)\gamma + \beta\right]$$

ze spojitosti funkce ψ a hledáme takové $\gamma \in \langle 0, 1\rangle$, pro které je tento výraz nekladný. Snadným výpočtem zjistíme, že $\gamma \in \langle \beta, 1\rangle$. Odtud plyne, že funkce $\psi(x)$ je monotonně klesající podél úsečky od x_C ke γx_N pro všechna

$$\gamma \in \langle \beta, 1\rangle.$$

Dále ukážeme, že norma Cauchyho bodu není větší než γ-násobek normy Newtonova bodu.

Věta 2.3 *Nechť* $g^T \mathbf{A} g > 0$, $\beta > 0$, $\gamma \in \langle \beta, 1\rangle$. *Pak platí*

$$\|x_C\| \leq \|\gamma x_N\|. \tag{2.41}$$

Jestliže $\|x_C\| = \|\gamma x_N\|$, *pak* $x_C = \gamma x_N$.

DŮKAZ: Platí

$$\|x_C\| = \frac{\|g\|^3}{g^T \mathbf{A} g} \cdot \frac{\|g\|\|\mathbf{A}^{-1} g\|}{\|g\|\|\mathbf{A}^{-1} g\|} \leq \frac{\|g\|^4 \|\mathbf{A}^{-1} g\|}{g^T \mathbf{A} g g^T \mathbf{A}^{-1} g} = \beta \|\mathbf{A}^{-1} g\| \leq \gamma \|\mathbf{A}^{-1} g\| = \|\gamma x_N\|.$$

Jestliže $\|x_C\| = \|\gamma x_N\|$, pak jsou všude rovnosti a platí $\gamma = \beta$ a $\|g\|\|\mathbf{A}^{-1} g\| = g^T \mathbf{A}^{-1} g$. Tedy $g = \alpha \mathbf{A}^{-1} g$ pro nějaké $\alpha > 0$, což spolu s $\|x_C\| = \|\gamma x_N\|$ dává $x_C = \gamma x_N$. □

Důsledek 2.2 *Specielně pro* $\gamma = 1$ *platí:*

1. $\|x_C\| \leq \|x_N\|$ *a jestliže* $\|x_C\| = \|x_N\|$, *pak* $x_C = x_N$.

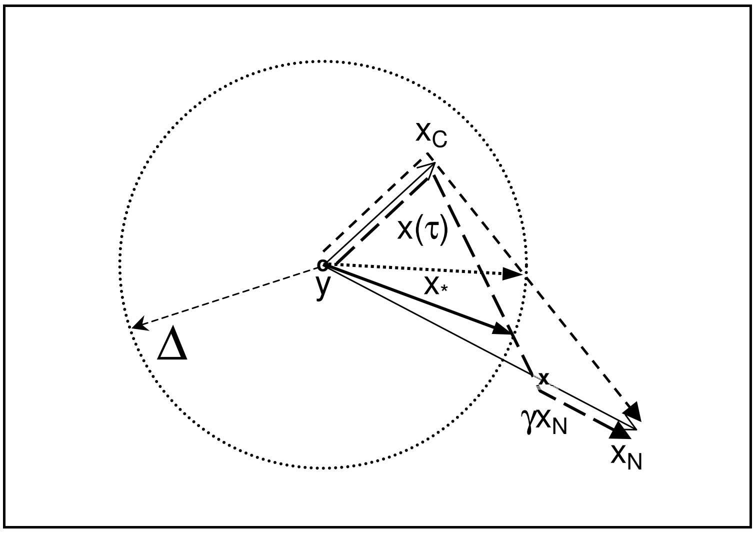
2. $\psi(x)$ *je monotonně klesající od Cauchyho bodu* x_C *k Newtonovu bodu* x_N.

Dále již můžeme definovat křivku modifikované metody psí nohy.

$$x(\tau) = \begin{cases} \tau x_C, & 0 \leq \tau \leq 1; \\ x_C + \frac{\tau - 1}{\gamma}(\gamma x_N - x_C), & 1 \leq \tau \leq 1 + \gamma; \\ (\tau - 1)x_N, & 1 + \gamma \leq \tau \leq 2. \end{cases} \tag{2.42}$$

Jako řešení $x_\star$ opět uvažujeme bod splňující (2.33).

Na obrázku 2.3 jsou znázorněny Cauchyho bod x_C, Newtonův bod x_N, bod γx_N, křivka $x(\tau)$ metody psí nohy (krátce čárkovaná), modifikované metody psí nohy (dlouze čárkovaná) a řešení $x_\star$ metody psí nohy (tečkovaná čára) a modifikované metody psí nohy (plná čára).



Obrázek 2.3: Křivka $x(\tau)$ metody psí nohy

Na závěr shrneme všechny úvahy do algoritmu. Budeme uvažovat obě verze této metody. Jestliže položíme $\gamma = 1$, dostaneme jednoduchou metodu psí nohy, položíme-li $\gamma = \beta$, dostaneme modifikovanou (dvojitou) metodu psí nohy.

Algoritmus 2.5 *Metoda psí nohy pro výpočet lokálně omezeného kroku.*

1. *Spočítáme* $\eta = g^T \mathbf{A} g$. *Jestliže* $\eta \leq 0$, *pak* $x_\star = -\frac{\Delta}{\|g\|} g$ *a* `STOP`.

2. *Určíme vektor* $x_C = -\frac{g^T g}{g^T \mathbf{A} g} g$. *Jestliže* $\|x_C\| \geq \Delta$, *pak* $x_\star = -\frac{\Delta}{\|g\|} g$ *a* `STOP`.

3. *Je-li matice* $\mathbf{A}$ *regulární, určíme vektor* $x_N = -\mathbf{A}^{-1} g$ *a přejdeme na krok 5.*

4. *Určíme vektor* v *tak, že*

$$\|v\| = 1, \quad v^T \mathbf{A} v \leq 0, \quad v^T(\mathbf{A} x_C + g) \leq 0,$$

položíme $x_\star = x_C + \kappa v$, *kde* $\kappa > 0$ *je určeno tak, že* $\|x_\star\| = \Delta$, *a* `STOP`.

5. *Jestliže* $\|x_N\| \le \Delta$, *pak* $x_\star = x_N$ *a* STOP.

6. *Jestliže* $(x_N - x_C)^T x_C \le 0$, *položíme* $x_\star = x_C - \kappa(x_N - x_C)$, *kde* $\kappa > 0$ *je určeno tak, že* $\|x_\star\| = \Delta$, *a* STOP.

7. *Položíme* $\gamma = 1$ *nebo* $\gamma = \frac{(g^T g)^2}{g^T \mathbf{A} g g^T \mathbf{A}^{-1} g} = \frac{x_C^T x_C}{x_N^T x_C}$.

8. *Jestliže* $\|\gamma x_N\| \le \Delta$, *položíme* $x_\star = -\frac{\Delta}{\|x_N\|} x_N$ *a* STOP.

9. *Položíme* $x_\star = x_C + \kappa(\gamma x_N - x_C)$, *kde* $\kappa > 0$ *je určeno tak, že* $\|x_\star\| = \Delta$.

Věta 2.4 *Metoda psí nohy, sestavená na základě algoritmu 2.5, je globálně konvergentní.*

DŮKAZ: Podle poznámky 1.3 stačí ukázat, že pro $x_\star$ je splněna nerovnost (1.11). Důkaz rozložíme na několik částí podle toho, jak algoritmus 2.5 končí pro $x_\star$.

- Je-li $g^T \mathbf{A} g \le 0$ nebo $g^T \mathbf{A} g > 0$ a $\|x_C\| \ge \Delta$, pak $x_\star = -\frac{\Delta}{\|g\|} g$. Pro takové $x_\star$ platí nerovnost (1.11) podle (1.22).
- Nechť tedy $\|x_C\| < \Delta$. Pak pro x_C platí nerovnost (1.11) podle (1.21).
- Není-li $\mathbf{A}$ regulární, pak položíme $x_\star = x_C + \kappa v$ pro jistá κ a v a podle 2.35 platí $\psi(x_\star) - \psi(x_C) \le 0$, takže pro $x_\star$ platí nerovnost (1.11).
- Je-li $\mathbf{A}$ regulární a $\|x_N\| \le \Delta$, pak $x_\star = x_N$ a $x_\star$ je optimální krok, pro který platí (1.11).
- Jestliže $\|x_N\| > \Delta$ a $(x_N - x_C)^T x_C \le 0$, pak $x_\star = x_C + \kappa v$ pro $v = -(x_N - x_C)$ a jisté κ. Pro takové $x_\star$ platí $\psi(x_\star) - \psi(x_C) \le 0$ podle 2.39, takže i (1.11).
- Konečně, pokud $(x_N - x_C)^T x_C > 0$, pak je funkce $\psi(x)$ na úsečce $\langle x_C, \gamma x_N \rangle$ monotonně klesající. Platí tedy $\psi(x_\star) \le \psi(x_C)$, a tím i (1.11). □

Kromě metody psí nohy, která hledá přibližné řešení na dvoudimenzionálním podprostoru, můžeme též uvažovat přesné řešení následující úlohy. Uvažujme opět řešení x ve tvaru (2.32), kde τ_1 a τ_2 jsou určeny tak, že $\|x\| \le \Delta$. Takové x dosadíme do vzorce pro ψ a dostaneme

$$\begin{aligned}\psi(x) &= \frac{1}{2}(\tau_1 g + \tau_2 \mathbf{A}^{-1} g)^T \mathbf{A} (\tau_1 g + \tau_2 \mathbf{A}^{-1} g) + g^T(\tau_1 g + \tau_2 \mathbf{A}^{-1} g) = \\ &= \frac{1}{2}(\tau_1, \tau_2) \begin{pmatrix} g^T \mathbf{A} g & g^T g \\ g^T g & g^T \mathbf{A}^{-1} g \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} + (g^T g,\ g^T \mathbf{A}^{-1} g) \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix}\end{aligned}$$

Dále platí

$$x^T x = (\tau_1 g + \tau_2 \mathbf{A}^{-1} g)^T (\tau_1 g + \tau_2 \mathbf{A}^{-1} g) = (\tau_1, \tau_2) \begin{pmatrix} g^T g & g^T \mathbf{A}^{-1} g \\ g^T \mathbf{A}^{-1} g & (\mathbf{A}^{-1} g)^T (\mathbf{A}^{-1} g) \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix}$$

Vektor $\mathbf{A}^{-1} g$ se spočítá pomocí vhodného rozkladu matice $\mathbf{A}$. Dostáváme úlohu minimalizace funkce

$$\tilde{\psi}(\tau) = \frac{1}{2} \tau^T \tilde{\mathbf{A}} \tau + \tilde{g}^T \tau \quad \text{vzhledem k} \quad \|\tau\|_{\tilde{C}}^2 = \tau^T \tilde{\mathbf{C}} \tau \le \Delta^2,$$

kde

$$(2.43) \quad \tilde{\mathbf{A}} = \begin{pmatrix} g^T\mathbf{A}g & g^Tg \\ g^Tg & g^T\mathbf{A}^{-1}g \end{pmatrix}, \quad \tilde{g} = \begin{pmatrix} g^Tg \\ g^T\mathbf{A}^{-1}g \end{pmatrix}, \quad \tilde{\mathbf{C}} = \begin{pmatrix} g^Tg & g^T\mathbf{A}^{-1}g \\ g^T\mathbf{A}^{-1}g & g^T\mathbf{A}^{-2}g \end{pmatrix},$$

pro vektor $\tau = (\tau_1, \tau_2)^T \in \mathbb{R}^2$, kterou budeme řešit pomocí Choleského rozkladu, algoritmus 2.2.

Řešení $\tau_\star$ je řešením rovnice $(\tilde{\mathbf{A}} + \xi\tilde{\mathbf{C}})\tau = -\tilde{g}$, poznámka 1.5, pro jisté ξ. Položíme tedy

$$\tilde{\phi}(\xi) = \frac{1}{\Delta} - \frac{1}{\|\tau\|_{\tilde{C}}}$$

(platí $\tau \equiv \tau(\xi)$) a postupujeme obdobně jako v důkazu lemmatu 2.1

$$\tilde{\phi}'(\xi) = \frac{\|\tau\|'_{\tilde{C}}}{\|\tau\|^2_{\tilde{C}}}, \qquad \|\tau\|'_{\tilde{C}} = \frac{\tau^T\tilde{\mathbf{C}}^T\tau'}{\|\tau\|_{\tilde{C}}}.$$

Zderivujeme obě strany rovnosti $(\tilde{\mathbf{A}} + \xi\tilde{\mathbf{C}})\tau = -\tilde{g}$ a dostaneme

$$\tau' = -(\tilde{\mathbf{A}} + \xi\tilde{\mathbf{C}})^{-1}\tilde{\mathbf{C}}\tau.$$

Takže za předpokladu, že $\tilde{\mathbf{A}} + \xi\tilde{\mathbf{C}} = \tilde{\mathbf{R}}^T\tilde{\mathbf{R}}$, máme

$$\tilde{\phi}'(\xi) = -\frac{\tau^T\tilde{\mathbf{C}}^T(\tilde{\mathbf{R}}^T\tilde{\mathbf{R}})^{-1}\tilde{\mathbf{C}}\tau}{\|\tau\|^3_{\tilde{C}}}.$$

Jestliže definujeme vektor $\tilde{w} \in \mathbb{R}^2$ tak, že

$$\tilde{\mathbf{R}}^T\tilde{w} = \tilde{\mathbf{C}}\tau,$$

pak Newtonův krok ξ^+ má tvar obdobný vztahu (2.10)

$$(2.44) \qquad \xi^+ = \xi + \frac{\|\tau\|^2_{\tilde{C}}}{\|\tilde{w}\|^2} \cdot \left(\frac{\|\tau\|_{\tilde{C}} - \Delta}{\Delta}\right)$$

Krok 5. algoritmu 2.2 dostane následující podobu. Nechť $\tilde{v} \in \mathbb{R}^2$ je vektor, který aproximuje vlastní vektor $\tilde{q}_1$ příslušný nejmenšímu vlastnímu číslu $\tilde{\lambda}_1$ matice $\tilde{\mathbf{A}}$, pro něhož platí $\|\tilde{v}\|_{\tilde{C}} = \sqrt{\tilde{v}^T\tilde{\mathbf{C}}\tilde{v}} = 1$. Pak pro číslo κ, které splňuje $\|\tau + \kappa\tilde{v}\|_{\tilde{C}} = \Delta$, platí obdoba vztahu (2.16)

$$(2.45) \qquad \kappa = \frac{\Delta^2 - \|\tau\|^2_{\tilde{C}}}{\tau^T\tilde{\mathbf{C}}\tilde{v} + \operatorname{sgn}(\tau^T\tilde{\mathbf{C}}\tilde{v})\sqrt{(\tau^T\tilde{\mathbf{C}}\tilde{v})^2 + \Delta^2 - \|\tau\|^2_{\tilde{C}}}}$$

Algoritmus vypadá takto.

Algoritmus 2.6 *Kombinace metody psí nohy a Choleského rozkladu pro výpočet lokálně omezeného kroku.*

Sestavíme $\tilde{\mathbf{A}}, \tilde{g}, \tilde{\mathbf{C}}$ *podle (2.43) a postupujeme stejně jako v algoritmu 2.2, kde nahradíme veličiny takto:*

$$n = 2, \quad \mathbf{A} \rightsquigarrow \tilde{\mathbf{A}}, \quad g \rightsquigarrow \tilde{g}, \quad x \rightsquigarrow \tau, \quad \|x\| \rightsquigarrow \|\tau\|_{\tilde{C}}, \quad (\mathbf{A}+\xi\mathbf{I} = \mathbf{R}^T\mathbf{R}) \rightsquigarrow (\tilde{\mathbf{A}}+\xi\tilde{\mathbf{C}} = \tilde{\mathbf{R}}^T\tilde{\mathbf{R}}).$$

Číslo κ *počítáme podle (2.45) a* ξ^+ *podle (2.44). Jakmile získáme řešení* $\tau_\star = (\tau_1, \tau_2)^T$, *položíme* $x_\star = \tau_1 g + \tau_2\mathbf{A}^{-1}g$.

Věta 2.5 *Kombinovaná metoda psí nohy a Choleského rozkladu, sestavená na základě algoritmu 2.6, je globálně konvergentní.*

DŮKAZ: Protože se řeší přesně úloha pro $\tilde{\psi}(\tau)$ podle algoritmu 2.2, je $\tau_\star$ optimální krok této modifikované úlohy, takže $x_\star$ je optimální krok na dvoudimenzionálním podprostoru $\mathrm{sp}\{g, \mathbf{A}^{-1}g\}$. Proto pro libovolné x z tohoto podprostoru platí $\psi(x_\star) \le \psi(x)$. Jelikož x_C patří do tohoto podprostoru, platí $\psi(x_\star) \le \psi(x_C)$ a je tedy splněna nerovnost (1.11) podle (1.21). □

2.4 Použití metody sdružených gradientů

V této části se budeme věnovat metodám, které hledají řešení problému (1.25) na podprostorech prostoru $\mathbb{R}^n$. Jako Cauchyho bod je definováno řešení problému

$$\min_{x \in \mathrm{sp}\{g\}} \psi(x) \equiv \frac{1}{2} x^T \mathbf{A} x + g^T x, \quad \|x\| \le \Delta, \tag{2.46}$$

tedy bod, ve kterém nabývá funkce ψ lokálního minima, je z 1-dimensionálního podprostoru $\mathrm{sp}\{g\}$. Stejně tak lze uvažovat minimum ψ na 2-dimensionálním podprostoru, atd. V těchto případech lze řešení snadno nalézt, protože tyto podprostory jsou malé. V obecném problému (1.25) je však hledaný prostor $\mathbb{R}^n$ velký. Budeme proto uvažovat řešení $x_\star$ kompromisního problému

$$\min_{x \in \mathcal{K}_{k+1}} \psi(x), \quad \|x\| \le \Delta, \tag{2.47}$$

kde

$$\mathcal{K}_{k+1} = \mathrm{sp}\{g, \mathbf{A}g, \mathbf{A}^2 g, \ldots, \mathbf{A}^k g\} \tag{2.48}$$

je Krylovův podprostor generovaný vektorem g a maticí $\mathbf{A}$. Nejprve najdeme bázi tohoto podprostoru a řešením problému (2.47) bude lineární kombinace této báze.

Jako první použil myšlenku sdružených gradientů Steihaug, [26], [35], [45], [60], [68]. Stručně ukážeme princip této metody, která generuje po částech lineární křivku, jejíž koncové body jsou iterace sdružených gradientů. Proces ukončíme v tom bodě, ve kterém tato křivka opustí hranici. Metoda je dobře definovaná pro libovolnou matici $\mathbf{A}$ a je vhodná pro rozsáhlé problémy.

Je-li $\mathbf{A}$ positivně definitní a $\|x_\star\| < \Delta$, pak minimum $x_\star$ splňuje rovnici $\mathbf{A}x + g = 0$. Jako residuum označíme vektor $r = \mathbf{A}x + g$.

V následujícím algoritmu uvažujeme tři různá kriteria ukončení iterací:

- Zjistíme-li, že matice $\mathbf{A}$ není positivně definitní, ukončíme proces prodloužením i-té iterace na hranici.
- Je-li norma $(i+1)$-ní iterace větší než Δ, prodloužíme i-tou iteraci na hranici a skončíme.
- Je-li residuum dostatečně malé, skončíme s i-tou iterací jako aproximačním řešením problému (2.47).

Je-li spočtená iterace x_{i+1} uvnitř kruhu o poloměru Δ (její norma je menší než Δ), lze ji přijmout a můžeme pokračovat další iterací.

Algoritmus 2.7 *Metoda sdružených gradientů pro výpočet lokálně omezeného kroku.*

Zvolíme $\varepsilon \in (0,1)$ *a položíme* $x_0 = 0,\ r_0 = g,\ p_0 = -r_0,\ i = 0.$

1. *Spočítáme* $\eta_i = p_i^T \mathbf{A} p_i$. *Je-li* $\eta_i > 0$, *přejdeme na krok 2. Jinak určíme* $\kappa > 0$ *tak, že* $\|x_i + \kappa p_i\| = \Delta$, *položíme* $x_\star = x_i + \kappa p_i$ *a* `STOP`.

2. *Položíme* $\alpha_i = \frac{r_i^T r_i}{\eta_i}$ *a* $x_{i+1} = x_i + \alpha_i p_i$.

   (a) *Je-li* $\|x_{i+1}\| > \Delta$, *určíme* $\kappa > 0$ *tak, že* $\|x_i + \kappa p_i\| = \Delta$, *položíme* $x_\star = x_i + \kappa p_i$ *a* `STOP`.

   (b) *Je-li* $\|x_{i+1}\| = \Delta$, *položíme* $x_\star = x_{i+1}$ *a* `STOP`.

   (c) *Je-li* $\|x_{i+1}\| < \Delta$, *přejdeme na krok 3.*

3. *Spočítáme* $r_{i+1} = r_i + \alpha_i \mathbf{A} p_i$. *Je-li* $\|r_{i+1}\| \le \varepsilon \|g\|$, *položíme* $x_\star = x_{i+1}$ *a* `STOP`. *Jinak přejdeme na krok 4.*

4. *Položíme* $\beta_i = \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i}$, $p_{i+1} = -r_{i+1} + \beta_i p_i$, $i := i+1$ *a návrat na krok 1.*

Po ukončení iterací získáme aproximační řešení $x_\star$ a mohou nastat tyto tři případy:

$$x_\star = \begin{cases} x_i + \kappa p_i, & \kappa > 0, \ \text{pro} \ \eta_i \le 0, \qquad \text{bod 1.} \\ x_i + \kappa p_i, & \kappa > 0, \ \text{pro} \ \eta_i > 0, \qquad \text{bod 2a.} \\ x_{i+1}, & \qquad\qquad\qquad\qquad\quad \text{body 2b. a 3.} \end{cases}$$

Nyní ukážeme, že posloupnost $\{\|x_i\|\}_{i\in\mathbb{N}_0}$ je ostře rostoucí, zatímco posloupnost funkčních hodnot $\{\psi(x_i)\}_{i\in\mathbb{N}_0}$ ostře klesající. Nejprve ukážeme některé vlastnosti metody sdružených gradientů.

Lemma 2.9 *Nechť* $\eta_i = p_i^T \mathbf{A} p_i \neq 0$, $i = 0, 1, \ldots, k$ *a uvažujme iterace generované algoritmem 2.7. Pak platí*

$$p_i^T \mathbf{A} p_j = 0, \quad 0 \le i < j \le k, \tag{2.49}$$

$$r_i^T r_j = 0, \quad 0 \le i < j \le k, \tag{2.50}$$

$$r_i^T p_j = -r_j^T r_j, \quad 0 \le i \le j \le k, \tag{2.51}$$

$$p_i^T p_j = \frac{r_j^T r_j}{r_i^T r_i} p_i^T p_i > 0, \quad 0 \le i \le j \le k, \tag{2.52}$$

$$\psi(x_{i+1}) = \psi(x_i) - \frac{1}{2}\frac{(r_i^T r_i)^2}{\eta_i}, \quad 0 \le i \le k. \tag{2.53}$$

DŮKAZ: Provede se indukcí s využitím vztahů v algoritmu 2.7. □

Věta 2.6 *Nechť* $\{x_j\}_{j=0,1,\ldots,i}$ *jsou iterace generované algoritmem 2.7. Pak:*

1. *Posloupnost* $\{\|x_j\|\}_{j=0,1,\ldots,i}$ *je ostře rostoucí a*

$$\|x_\star\| > \|x_i\|. \tag{2.54}$$

2. Posloupnost $\{\psi(x_j)\}_{j=0,1,\dots,i}$ *je ostře klesající a*

$$\psi(x_\star) < \psi(x_i). \tag{2.55}$$

DŮKAZ:

1. Z kroku 2. algoritmu 2.7 plyne

$$x_j = x_{j-1} + \alpha_{j-1}p_{j-1} = \sum_{k=0}^{j-1} \alpha_k p_k, \quad j = 1, \dots, i \tag{2.56}$$

a

$$\alpha_j > 0, \quad j = 0, 1, \dots, i-1. \tag{2.57}$$

Odtud podle (2.52) je

$$x_j^T p_j = \alpha_0 p_0^T p_j + \dots + \alpha_{j-1} p_{j-1}^T p_j > 0, \quad j = 1, \dots, i. \tag{2.58}$$

Nyní

$$x_{j+1}^T x_{j+1} = (x_j + \alpha_j p_j)^T (x_j + \alpha_j p_j) > x_j^T x_j, \quad j = 0, 1, \dots, i \tag{2.59}$$

podle (2.58) a (2.57). Odtud plyne, že posloupnost $\{\|x_j\|\}_{j=0,1,\dots,i}$ je ostře rostoucí. Je-li $x_\star = x_{i+1}$, pak (2.54) plyne z (2.59). Je-li $x_\star = x_i + \kappa p_i$, pak

$$x_\star^T x_\star = x_i^T x_i + 2\kappa x_i^T p_i + \kappa^2 p_i^T p_i > x_i^T x_i$$

podle (2.58) a toho, že $\kappa > 0$. Odtud plyne (2.54).

2. Je-li $\eta_j > 0$, $j = 0, 1, \dots, i-1$, pak z (2.53) plyne, že posloupnost $\{\psi(x_j)\}_{j=0,1,\dots,i}$ je ostře klesající.
Je-li $x_\star = x_{i+1}$, pak (2.55) plyne odtud. Je-li $x_\star = x_i + \kappa p_i$, pak

$$(\mathbf{A}x_i + g)^T p_i = r_i^T p_i = -r_i^T r_i < 0 \tag{2.60}$$

podle (2.51) a dále

- Je-li $\eta_i > 0$, pak z algoritmu 2.7 plyne $\|x_i\| < \Delta$ a zároveň $\|x_{i+1}\| > \Delta$. Platí tedy $\|x_\star\| = \Delta$ pro $x_\star = x_i + \kappa p_i$, kde $0 < \kappa < \alpha_i$, a z (2.53) plyne $\psi(x_{i+1}) < \psi(x_i)$. Definujme funkci

$$\omega(\kappa) = \psi(x_i + \kappa p_i) - \psi(x_i) = \frac{1}{2}\kappa^2 p_i^T \mathbf{A} p_i + \kappa x_i^T \mathbf{A} p_i + \kappa g^T p_i = \frac{1}{2}\kappa^2 \eta_i - \kappa \|r_i\|^2.$$

Platí $\omega(0) = 0$ a $\omega(\alpha_i) < 0$. Tato funkce je pro $0 < \kappa < \alpha_i$ klesající, minima nabývá pro $\kappa = \frac{\|r_i\|^2}{\eta_i} = \alpha_i$. Tedy $\psi(x_i) > \psi(x_i + \kappa p_i) > \psi(x_{i+1})$.

- Je-li $\eta_i \le 0$, pak z algoritmu 2.7 plyne $\|x_i\| < \Delta$. Nyní

$$\begin{aligned}
\psi(x_i + \kappa p_i) &= \frac{1}{2}(x_i + \kappa p_i)^T \mathbf{A}(x_i + \kappa p_i) + g^T(x_i + \kappa p_i) = \\
&= \psi(x_i) + \kappa x_i^T \mathbf{A} p_i + \frac{1}{2}\kappa^2 p_i^T \mathbf{A} p_i + \kappa g^T p_i = \\
&= \psi(x_i) + \frac{1}{2}\kappa^2 \eta_i + \kappa(\mathbf{A}x_i + g)^T p_i < \psi(x_i)
\end{aligned}$$

podle (2.60).

V obou případech tudíž pro $x_\star = x + \kappa p_i$ dostáváme vztah (2.55). $\square$

V praxi se rovněž zkoušela kombinace metody sdružených gradientů s metodou psí nohy. Ukazuje se, že tyto metody konvergují téměř stejně dobře jako metody s optimálním lokálně omezeným krokem a efektivita těchto metod založených na promítání do podprostoru generovaného vektory g a $\mathbf{A}^{-1}g$ se příliš nezmění, nahradíme-li přesné řešení úlohy (1.25) přibližným řešením $x = \tau_1 g + \tau_2 \mathbf{A}^{-1} g$. Krátce uvedeme obecnější algoritmus.

Zvolíme $m \geq 1$, spočítáme m kroků metody sdružených gradientů a jsou-li všechny iterace uvnitř oblasti, přejdeme na metodu psí nohy. Uvažujme iteraci x_m a Newtonův bod x_n, pro který platí $\mathbf{A}x_n + g = 0$ a uvažujme libovolný bod $x(\vartheta) = x_m + \vartheta(x_n - x_m)$ na úsečce $\langle x_m, x_n \rangle$. Platí tato věta.

Věta 2.7 *Nechť* $p_i^T \mathbf{A} p_i > 0$, $0 \leq i \leq m-1$, *kde* $m \geq 1$ *a uvažujme iterace generované algoritmem 2.7. Nechť* $\|x_m\| < \Delta$ *a* $\mathbf{A}x_m + g = r_m \neq 0$. *Nechť* $x_n \in \mathbb{R}^n$ *je vektor takový, že* $\mathbf{A}x_n + g = 0$. *Pak platí:*

$$\frac{\partial \psi(x(\vartheta))}{\partial \vartheta} = (1 - \vartheta)(x_n - x_m)^T r_m. \tag{2.61}$$

DŮKAZ: Provede se stejně jako u vztahu (2.37), kde za x_C dosadíme x_m. $\square$

Tvrzení této věty tvoří základ následujícího nového algoritmu, který vychází z analýzy metody psí nohy v § 2.3.

Algoritmus 2.8 *Kombinace metody sdružených gradientů a psí nohy pro výpočet lokálně omezeného kroku.*

Zvolíme $\varepsilon \in (0,1)$, $m \geq 1$ *a položíme* $x_0 = 0$, $r_0 = g$, $p_0 = -r_0$, $i = 0$.

1. *Spočítáme* $\eta_i = p_i^T \mathbf{A} p_i$. *Je-li* $\eta_i > 0$, *přejdeme na krok 2. Jinak určíme* $\kappa > 0$ *tak, že* $\|x_i + \kappa p_i\| = \Delta$, *položíme* $x_\star = x_i + \kappa p_i$ *a* `STOP`.

2. *Položíme* $\alpha_i = \frac{r_i^T r_i}{\eta_i}$ *a* $x_{i+1} = x_i + \alpha_i p_i$.

   (a) *Je-li* $\|x_{i+1}\| > \Delta$, *určíme* $\kappa > 0$ *tak, že* $\|x_i + \kappa p_i\| = \Delta$, *položíme* $x_\star = x_i + \kappa p_i$ *a* `STOP`.

   (b) *Je-li* $\|x_{i+1}\| = \Delta$, *položíme* $x_\star = x_{i+1}$ *a* `STOP`.

   (c) *Je-li* $\|x_{i+1}\| < \Delta$, *přejdeme na krok 3.*

3. *Spočítáme* $r_{i+1} = r_i + \alpha_i \mathbf{A} p_i$. *Je-li* $\|r_{i+1}\| \leq \varepsilon \|g\|$, *položíme* $x_\star = x_{i+1}$ *a* `STOP`. *Jinak přejdeme na krok 4.*

4. *Jestliže* $i + 1 = m$, *přejdeme na krok 5. Jinak položíme* $\beta_i = \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i}$, $p_{i+1} = r_{i+1} + \beta_i p_i$, $i := i + 1$ *a návrat na krok 1.*

5. *Řešíme soustavu* $\mathbf{A}x_n + g = 0$. *Jestliže* $\|x_n\| \leq \Delta$, *položíme* $x_\star = x_n$ *a* `STOP`.

6. *Jestliže* $(x_n - x_m)^T r_m \geq 0$, *položíme* $x_\star = x_m - \kappa(x_n - x_m)$, *kde* $\kappa > 0$ *je určeno tak, že* $\|x_\star\| = \Delta$, *a* `STOP`.

7. *Jestliže* $(x_n - x_m)^T r_m < 0$, *položíme* $x_\star = x_m + \kappa(x_n - x_m)$, *kde* $\kappa > 0$ *je určeno tak, že* $\|x_\star\| = \Delta$, *a* `STOP`.

Obvykle volíme $m \leq 5$. Pro $m = 1$ dostaneme jednoduchou metodu psí nohy.

Podle poznámky 1.3 jsou obě tyto metody globálně konvergentní, jestliže $x_\star$ splňuje nerovnost (1.11).

Věta 2.8 *Nechť $\{x_j\}_{j=0,\dots,i}$ jsou iterace generované algoritmy 2.7 a 2.8 a nechť platí $\eta_j = p_j^T \mathbf{A} p_j > 0$, $j = 0, \dots, i$. Pak je pro $x_\star$ splněna nerovnost (1.11).*

Důkaz: Z věty 2.6 a vztahu (2.53) plyne

$$\begin{aligned} -\psi(x_\star) &\geq -\psi(x_i) > \dots > -\psi(x_1) = -\psi(x_0) + \frac{1}{2}\,\frac{(r_0^T r_0)^2}{\eta_0} = \frac{1}{2}\,\frac{(r_0^T r_0)^2}{p_0^T \mathbf{A} p_0} \geq \\ &\geq \frac{1}{2}\,\frac{\|g\|^4}{\|g\|^2 \|\mathbf{A}\|} = \frac{1}{2}\,\|g\|\,\frac{\|g\|}{\|\mathbf{A}\|} \geq \frac{1}{2}\,\|g\| \min\left\{\|x_\star\|, \frac{\|g\|}{\|\mathbf{A}\|}\right\}, \end{aligned}$$

kde $\underline{\sigma} = \frac{1}{2}$. □

Důsledek 2.3 *Použití metody sdružených gradientů, resp. její kombinace s metodou psí nohy, sestavené na základě algoritmu 2.7, resp. 2.8, je globálně konvergentní metodou.*

2.5 Předpodmíněná metoda sdružených gradientů

Pro velké řídké systémy je vhodné předpodmínit metodu sdružených gradientů užitím symetrické a positivně definitní matice $\mathbf{C} \in \mathbb{R}^{n \times n}$. Jestliže použijeme metodu sdružených gradientů, algoritmus (2.7), na minimalizaci kvadratické funkce

$$\tilde{\psi}(\tilde{x}) = \frac{1}{2}\,\tilde{x}^T \tilde{\mathbf{A}} \tilde{x} + \tilde{g}^T \tilde{x}, \quad \|\tilde{x}\| \leq \Delta,$$

kde

(2.62) $$\tilde{\mathbf{A}} = \mathbf{C}^{-\frac{1}{2}} \mathbf{A} \mathbf{C}^{-\frac{1}{2}}, \quad \tilde{g} = \mathbf{C}^{-\frac{1}{2}} g, \quad \tilde{x} = \mathbf{C}^{\frac{1}{2}} x,$$

a položíme-li

$$x_i = \mathbf{C}^{-\frac{1}{2}} \tilde{x}_i, \quad r_i = \mathbf{C}^{\frac{1}{2}} \tilde{r}_i, \quad p_i = \mathbf{C}^{-\frac{1}{2}} \tilde{p}_i,$$

dostaneme předpodmíněnou metodu sdružených gradientů. Matice $\mathbf{C} \approx \mathbf{A}$ se volí tak, aby $\tilde{\mathbf{A}}$ byla lépe podmíněná než $\mathbf{A}$ a aby $\mathbf{C}$ byla snadno invertovatelná. Nejčastěji se provede neúplný rozklad $\mathbf{A} \approx \mathbf{R}^T \mathbf{R}$ a položíme $\mathbf{C} = \mathbf{R}^T \mathbf{R}$, kde horní trojúhelníková matice $\mathbf{R}$ má stejnou strukturu nenulových prvků jako matice $\mathbf{A}$.

Použití předpodmiňovače $\mathbf{C}$ změní omezení $\|x\| \leq \Delta$ na $\|x\|_C = \sqrt{x^T \mathbf{C} x} \leq \Delta$, což je třeba algoritmicky ošetřit.

Algoritmus 2.9 *<u>Předpodmíněná metoda sdružených gradientů pro výpočet lokálně omezeného kroku.</u>*

Zvolíme $\varepsilon \in (0, 1)$ a položíme $x_0 = 0$, $r_0 = g$, $p_0 = -\mathbf{C}^{-1} r_0$, $i = 0$.

1. *Spočítáme $\eta_i = p_i^T \mathbf{A} p_i$. Je-li $\eta_i > 0$, přejdeme na krok 2. Jinak určíme $\kappa > 0$ tak, že $\|x_i + \kappa p_i\| = \Delta$, položíme $x_\star = x_i + \kappa p_i$ a* `STOP`.

2. *Položíme $\alpha_i = \frac{r_i^T \mathbf{C}^{-1} r_i}{\eta_i}$ a $x_{i+1} = x_i + \alpha_i p_i$.*

(a) *Je-li* $\|x_{i+1}\| > \Delta$, *určíme* $\kappa > 0$ *tak, že* $\|x_i + \kappa p_i\| = \Delta$, *položíme* $x_\star = x_i + \kappa p_i$ *a* `STOP`.

(b) *Je-li* $\|x_{i+1}\| = \Delta$, *položíme* $x_\star = x_{i+1}$ *a* `STOP`.

(c) *Je-li* $\|x_{i+1}\| < \Delta$, *přejdeme na krok 3.*

3. *Spočítáme* $r_{i+1} = r_i + \alpha_i \mathbf{A} p_i$. *Je-li* $\|r_{i+1}\| \le \varepsilon \|g\|$, *položíme* $x_\star = x_{i+1}$ *a* `STOP`. *Jinak přejdeme na krok 4.*

4. *Položíme* $\beta_i = \frac{r_{i+1}^T \mathbf{C}^{-1} r_{i+1}}{r_i^T \mathbf{C}^{-1} r_i}$, $p_{i+1} = -\mathbf{C}^{-1} r_{i+1} + \beta_i p_i$, $i := i+1$ *a návrat na krok 1.*

Je nutno poznamenat, že posloupnost norem $\{\|x_j\|\}$ nemusí být monotonně rostoucí. To platí pro posloupnost $\{\|\tilde{x}_j\|\} = \{\|x_j\|_C\}$. Výpočet ukončíme, jakmile bod x_{i+1} překročí hranici Δ, tj. $\|x_{i+1}\| \ge \Delta$. Získané přibližné řešení $x_\star$ proto nemusí být tak dobré jako u nepředpodmíněné metody sdružených gradientů, kde normy iterací rostou monotonně, avšak použitím předpodmínění docílíme zrychlení konvergence.

K důkazu globální konvergence této metody použijeme poznámku 1.3.

Věta 2.9 *Nechť* $\mathbf{C}$ *je matice předpodmínění a* $\kappa(\mathbf{C})$ *její číslo podmíněnosti (poznámka A.3). Jestliže je splněna podmínka*

$$\kappa(\mathbf{C}) \le \overline{C} < \infty,$$

pak pro $x_\star$ *platí nerovnost (1.11).*

DŮKAZ: Nerovnost, kterou máme dokázat, platí pro $\tilde{\psi}(\tilde{x}_\star)$ s konstantou $\underline{\tilde{\sigma}} \in (0,1)$, věta 2.8. Ze vztahů (2.62) plyne

$$\psi(x) = \tilde{\psi}(\tilde{x}), \quad \|\tilde{\mathbf{A}}\| \le \|\mathbf{C}^{-\frac{1}{2}}\|^2 \|\mathbf{A}\|, \quad \|g\| \le \|\mathbf{C}^{\frac{1}{2}}\| \|\tilde{g}\|, \quad \|x\| \le \|\mathbf{C}^{-\frac{1}{2}}\| \|\tilde{x}\|,$$

takže

$$\begin{aligned}
-\psi(x_\star) &= -\tilde{\psi}(\tilde{x}_\star) \ge \underline{\tilde{\sigma}}\, \|\tilde{g}\| \min\left\{ \|\tilde{x}_\star\|, \frac{\|\tilde{g}\|}{\|\tilde{\mathbf{A}}\|} \right\} \ge \\
&\ge \underline{\tilde{\sigma}} \frac{\|g\|}{\|\mathbf{C}^{\frac{1}{2}}\|} \min\left\{ \frac{\|x_\star\|}{\|\mathbf{C}^{-\frac{1}{2}}\|}, \frac{\|g\|}{\|\mathbf{C}^{\frac{1}{2}}\| \|\mathbf{C}^{-\frac{1}{2}}\|^2 \|\mathbf{A}\|} \right\} = \\
&= \underline{\tilde{\sigma}} \frac{\|g\|}{\kappa(\mathbf{C}^{\frac{1}{2}})} \min\left\{ \|x_\star\|, \frac{\|g\|}{\kappa(\mathbf{C}^{\frac{1}{2}}) \|\mathbf{A}\|} \right\} \ge \\
&\ge \underline{\tilde{\sigma}} \frac{\|g\|}{\kappa(\mathbf{C}^{\frac{1}{2}})} \min\left\{ \frac{\|x_\star\|}{\kappa(\mathbf{C}^{\frac{1}{2}})}, \frac{\|g\|}{\kappa(\mathbf{C}^{\frac{1}{2}}) \|\mathbf{A}\|} \right\} = \frac{\underline{\tilde{\sigma}}}{\kappa(\mathbf{C})} \|g\| \min\left\{ \|x_\star\|, \frac{\|g\|}{\|\mathbf{A}\|} \right\}
\end{aligned}$$

a protože je $1 \le \kappa(\mathbf{C}) \le \overline{C} < \infty$, položením $\underline{\sigma} = \frac{1}{\overline{C}} \underline{\tilde{\sigma}}$ dostaneme $0 < \underline{\sigma} < 1$. □

Důsledek 2.4 *Předpodmíněná metoda sdružených gradientů, sestavená na základě algoritmu 2.9, je globálně konvergentní.*

2.6 Použití Lanczosovy metody

V této části použijeme pro vytvoření báze Krylovova podprostoru $\mathcal{K}_{k+1}$ pomocí metody sdružených gradientů myšlenku Lanczosovy metody, [08], [23], která generuje ortonormální bázi $\mathcal{K}_{k+1} = \mathrm{sp}\{g, \mathbf{A}g, \mathbf{A}^2 g, \ldots, \mathbf{A}^k g\} = \mathrm{sp}\{q_0, q_1, \ldots, q_k\}$.

Algoritmus 2.10 *Metoda sdružených gradientů (CGM).*

Položíme $r_0 = g,\ p_0 = -r_0$ *a pro* $j = 0, 1, \ldots, k-1$ *provedeme:*

1. $\alpha_j = \frac{r_j^T r_j}{p_j^T \mathbf{A} p_j}, \quad r_{j+1} = r_j + \alpha_j \mathbf{A} p_j,$

2. $\beta_j = \frac{r_{j+1}^T r_{j+1}}{r_j^T r_j}, \quad p_{j+1} = -r_{j+1} + \beta_j p_j$

Algoritmus 2.11 *Lanczosova metoda (LM).*

Položíme $t_0 = g,\ q_{-1} = 0$ *a pro* $j = 0, 1, \ldots, k$ *provedeme:*

1. $\gamma_j = \|t_j\|, \quad q_j = \frac{1}{\gamma_j} t_j,$

2. $\delta_j = q_j^T \mathbf{A} q_j, \quad t_{j+1} = \mathbf{A} q_j - \delta_j q_j - \gamma_j q_{j-1}$

Metoda CGM končí pro $r_{k+1} = 0$ a v $\mathcal{K}_{k+1}$ generuje $\mathbf{A}$-ortogonální bázi $p_0, p_1, \ldots, p_k$, zatímco metoda LM končí pro $t_{k+1} = 0$ a v $\mathcal{K}_{k+1}$ generuje ortonormální bázi $q_0, q_1, \ldots, q_k$. Lanczosovu iteraci lze psát v maticovém tvaru

$$\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_k \mathbf{T}_k + \gamma_{k+1} q_{k+1} e_{k+1}^T, \tag{2.63}$$

kde $\mathbf{Q}_k = (q_0, \ldots, q_k)$, $\mathbf{Q}_k^T \mathbf{Q}_k = \mathbf{I}_{k+1}$ a $\mathbf{T}_k \in \mathbb{R}^{(k+1)\times(k+1)}$ je třídiagonální matice

$$\mathbf{T}_k = \begin{pmatrix} \delta_0 & \gamma_1 & & \\ \gamma_1 & \delta_1 & \ddots & \\ & \ddots & \ddots & \gamma_k \\ & & \gamma_k & \delta_k \end{pmatrix} \tag{2.64}$$

Dále platí vztahy

$$\mathbf{Q}_k^T \mathbf{A} \mathbf{Q}_k = \mathbf{T}_k, \quad \mathbf{Q}_k^T g = \gamma_0 e_1 \quad \text{a} \quad g = \gamma_0 q_0. \tag{2.65}$$

První rovnost plyne z (2.63). Dále $q_0 = \frac{1}{\gamma_0} t_0 = \frac{1}{\|g\|} g$ a proto je $q_0^T g = \|g\| = \gamma_0$ a $q_i^T g = 0\ \forall i \neq 0$. Tedy $\mathbf{Q}_k^T g = \gamma_0 e_1$. Tím jsou dokázány druhá a třetí rovnost (2.65).

Protože residua $\{r_k\}$ jsou ortogonální, vztah (2.50), můžeme Lanczosovy vektory q_i získat z iterací sdružených gradientů pomocí vztahu

$$q_i = \sigma_i \frac{1}{\|r_i\|} r_i, \quad i = 0, 1, \ldots, k, \quad \text{kde} \quad \sigma_i = \pm 1, \tag{2.66}$$

jak lze nalézt např. v [19] a platí následující lemma.

Lemma 2.10 *Uvažujme algoritmy CGM a LM a nechť* $\alpha_j \neq 0,\ j = 0, \ldots, k$. *Pak platí*

$$\sigma_i = -\sigma_{i-1}\ \mathrm{sgn}(\alpha_{i-1}), \quad \sigma_0 = +1$$

a matici $\mathbf{T}_k$ *lze vyjádřit následujícím způsobem*

$$\mathbf{T}_k = \begin{pmatrix} \frac{1}{\alpha_0} & \frac{\sqrt{\beta_0}}{|\alpha_0|} & & & \\ \frac{\sqrt{\beta_0}}{|\alpha_0|} & \frac{1}{\alpha_1} + \frac{\beta_0}{\alpha_0} & \frac{\sqrt{\beta_1}}{|\alpha_1|} & & \\ & \frac{\sqrt{\beta_1}}{|\alpha_1|} & \ddots & \ddots & \\ & & \ddots & \ddots & \frac{\sqrt{\beta_{k-1}}}{|\alpha_{k-1}|} \\ & & & \frac{\sqrt{\beta_{k-1}}}{|\alpha_{k-1}|} & \frac{1}{\alpha_k} + \frac{\beta_{k-1}}{\alpha_{k-1}} \end{pmatrix} \tag{2.67}$$

Důkaz: Použijeme algoritmy CGM a LM a využijeme $\mathbf{A}$-ortogonalitu směrů $\{p_i\}$ a ortogonalitu residuí $\{r_i\}$. Z (2.65) plyne $q_0 = \frac{1}{\|g\|} g = \frac{1}{\|r_0\|} r_0$ a tedy $\sigma_0 = +1$. Označíme-li dále

$$\mathbf{P}_k = (p_0, \ldots, p_k), \quad \mathbf{R}_k = (r_0, \ldots, r_k), \quad \mathbf{S}_k = \mathrm{diag}(\sigma_0\|r_0\|, \ldots, \sigma_k\|r_k\|) \quad \text{a}$$

$$\mathbf{B}_k = \begin{pmatrix} -1 & \beta_0 & & \\ & \ddots & \ddots & \\ & & \ddots & \beta_{k-1} \\ & & & -1 \end{pmatrix},$$

pak lze psát $\mathbf{R}_k = \mathbf{Q}_k\mathbf{S}_k$ a $\mathbf{R}_k = \mathbf{P}_k\mathbf{B}_k$. Protože $\alpha_i = \frac{r_i^T r_i}{p_i^T \mathbf{A} p_i}$ a směry $\{p_i\}$ jsou $\mathbf{A}$-ortogonální, platí

$$\mathbf{P}_k^T \mathbf{A} \mathbf{P}_k = \mathrm{diag}\left(\frac{r_0^T r_0}{\alpha_0}, \ldots, \frac{r_k^T r_k}{\alpha_k}\right) = \mathbf{S}_k\, \mathrm{diag}\left(\frac{1}{\alpha_0}, \ldots, \frac{1}{\alpha_k}\right) \mathbf{S}_k.$$

Dále $\beta_i = \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i}$ a tudíž

$$\mathbf{S}_k\mathbf{B}_k\mathbf{S}_k^{-1} = \begin{pmatrix} -1 & \sigma_0\sigma_1\sqrt{\beta_0} & & \\ & \ddots & \ddots & \\ & & \ddots & \sigma_{k-1}\sigma_k\sqrt{\beta_{k-1}} \\ & & & -1 \end{pmatrix}$$

Nyní dosadíme za $\mathbf{T}_k$ a dostaneme

$$\begin{aligned} \mathbf{T}_k &= \mathbf{Q}_k^T \mathbf{A} \mathbf{Q}_k = \mathbf{S}_k^{-1}\mathbf{R}_k^T \mathbf{A} \mathbf{R}_k \mathbf{S}_k^{-1} = \mathbf{S}_k^{-1}\mathbf{B}_k^T \mathbf{P}_k^T \mathbf{A} \mathbf{P}_k \mathbf{B}_k \mathbf{S}_k^{-1} = \\ &= \mathbf{S}_k^{-1}\mathbf{B}_k^T \mathbf{S}_k\, \mathrm{diag}\left(\frac{1}{\alpha_0}, \ldots, \frac{1}{\alpha_k}\right) \mathbf{S}_k \mathbf{B}_k \mathbf{S}_k^{-1} = \\ &= \begin{pmatrix} \frac{1}{\alpha_0} & -\sigma_0\sigma_1\frac{\sqrt{\beta_0}}{\alpha_0} & & \\ -\sigma_0\sigma_1\frac{\sqrt{\beta_0}}{\alpha_0} & \frac{1}{\alpha_1} + \frac{\beta_0}{\alpha_0} & \ddots & \\ & \ddots & \ddots & -\sigma_{k-1}\,\sigma_k\frac{\sqrt{\beta_{k-1}}}{\alpha_{k-1}} \\ & & -\sigma_{k-1}\,\sigma_k\frac{\sqrt{\beta_{k-1}}}{\alpha_{k-1}} & \frac{1}{\alpha_k} + \frac{\beta_{k-1}}{\alpha_{k-1}} \end{pmatrix} \end{aligned} \tag{2.68}$$

Získáváme výrazy pro δ_i, $i = 0, \ldots, k$, a protože $\gamma_i = \|t_i\| \geq 0$, vychází srovnáním prvků matic (2.64) a (2.68), že $\gamma_i = \frac{\sqrt{\beta_{i-1}}}{|\alpha_{i-1}|}$ a $\sigma_{i-1}\,\sigma_i\ \mathrm{sgn}(\alpha_{i-1}) = -1$, $i = 1, \ldots, k$. Tím jsme převedli matici $\mathbf{T}_k(\delta, \gamma)$ na matici $\mathbf{T}_k(\alpha, \beta)$. □

Nyní se vrátíme k problému (1.25) a uvažujme řešení z Krylovova podprostoru $\mathcal{K}_{k+1}$: $x \in \mathcal{K}_{k+1} = \{x \in \mathbb{R}^n : x = \mathbf{Q}_k h\}$, tedy řešení je tvaru

$$x_k = \mathbf{Q}_k h_k, \tag{2.69}$$

kde x_k řeší problém

$$\min_{x \in \mathcal{K}_{k+1}} \left\{ \frac{1}{2}\, x^T \mathbf{A} x + g^T x \right\} \quad \text{vzhledem k} \quad \|x\| \leq \Delta. \tag{2.70}$$

Ze vztahů (2.65) plyne po dosazení, že vektor h_k řeší problém

$$\min_{h \in \mathbb{R}^{k+1}} \left\{ \frac{1}{2}\, h^T \mathbf{T}_k h + (\gamma_0 e_1)^T h \right\} \quad \text{vzhledem k} \quad \|h\| \leq \Delta. \tag{2.71}$$

Podle věty 1.5 platí

$$(\mathbf{T}_k + \xi_k \mathbf{I}_{k+1}) h_k = -\gamma_0 e_1, \tag{2.72}$$

kde $\mathbf{T}_k + \xi_k \mathbf{I}_{k+1} \succeq 0$, $\xi_k \geq 0$ a $\xi_k(\|h_k\| - \Delta) = 0$. Odtud a ze vztahů (2.65) plyne

$$\mathbf{Q}_k^T(\mathbf{A} + \xi_k \mathbf{I}_n) x_k = \mathbf{Q}_k^T(\mathbf{A} + \xi_k \mathbf{I}_n)\mathbf{Q}_k h_k = (\mathbf{T}_k + \xi_k \mathbf{I}_{k+1}) h_k = -\gamma_0 e_1 = -\mathbf{Q}_k^T g$$

a dále

$$\xi_k(\|x_k\| - \Delta) = \xi_k(\|h_k\| - \Delta) = 0.$$

Vidíme, že x_k je Galerkinova aproximace k $x_\star$ z prostoru generovaného maticí $\mathbf{Q}_k$. Ptáme se tedy, jak dobrá tato aproximace je, konkrétně jaká je chyba $(\mathbf{A} + \xi_k \mathbf{I}_n) x_k + g$.

Věta 2.10 *Platí:*

$$(\mathbf{A} + \xi_k \mathbf{I}_n) x_k + g = \gamma_{k+1} \cdot e_{k+1}^T h_k \cdot q_{k+1}$$

a tedy

$$\|(\mathbf{A} + \xi_k \mathbf{I}_n) x_k + g\| = \gamma_{k+1} \cdot |e_{k+1}^T h_k|. \tag{2.73}$$

Důkaz: Použijeme vztahy (2.63), (2.72) a (2.65):

$$\begin{aligned}
\mathbf{A} x_k &= \mathbf{A}\mathbf{Q}_k h_k = \mathbf{Q}_k \mathbf{T}_k h_k + \gamma_{k+1}(e_{k+1}^T h_k) q_{k+1} = \\
&= -\mathbf{Q}_k(\xi_k h_k + \gamma_0 e_1) + \gamma_{k+1}(e_{k+1}^T h_k) q_{k+1} = \\
&= -\xi_k \mathbf{Q}_k h_k - \gamma_0 \mathbf{Q}_k e_1 + \gamma_{k+1}(e_{k+1}^T h_k) q_{k+1} = \\
&= -\xi_k x_k - \gamma_0 q_0 + \gamma_{k+1}(e_{k+1}^T h_k) q_{k+1} = -\xi_k x_k - g + \gamma_{k+1}(e_{k+1}^T h_k) q_{k+1}.
\end{aligned}$$

Odtud již snadno plyne vztah (2.73). □

Tedy k tomu, abychom změřili chybu, stačí znát číslo γ_{k+1} a poslední složku vektoru h_k. Chyba bude malá, bude-li malé alespoň jedno z těchto dvou čísel.

Nyní se zaměříme na problém (2.71) a uvedeme někeré vlastnosti Lanczosovy metody. Začneme jednoduchou definicí.

Definice 2.2 *Řekneme, že symetrická třídiagonální matice je degenerovaná (reducibilní, rozložitelná), jestliže alespoň jeden prvek mimo diagonálu je nulový. V opačném případě řekneme, že je nedegenerovaná (ireducibilní, nerozložitelná).*

Lemma 2.11 *Nechť obecná třídiagonální matice* $\mathbf{T}$ *je nedegenerovaná a* v *je vlastní vektor matice* $\mathbf{T}$*. Pak první složka* v *je nenulová.*

DŮKAZ: Uvažujme vztah $\mathbf{T}v = \mu v$ pro nějaké vlastní číslo μ matice $\mathbf{T}$. Nechť první složka $v^{(1)}$ vektoru v je nulová. Protože $\mathbf{T}$ je nedegenerovaná, je nutně druhá složka $v^{(2)}$ rovna nule. Obdobně dostaneme $v^{(3)}$ rovno nule, atd. Tedy $v^{(i)} = 0 \ \forall i$, což je spor. □

Lemma 2.12 *Nechť matice* $\mathbf{T}_k$ *je nedegenerovaná. Pak singulární případ nemůže nastat pro problém (2.71).*

DŮKAZ: Nechť singulární případ nastane. Pak podle definice singulárního případu (definice 1.11) je vektor $\gamma_0 e_1$ ortogonální na v_k, což je vlastní vektor odpovídající nejmenšímu vlastnímu číslu μ_k matice $\mathbf{T}_k$. Platí tedy

$$0 = \gamma_0 e_1^T v_k = \gamma_0 v_k^{(1)} + 0 \cdot v_k^{(2)} + \ldots \quad \Rightarrow \quad v_k^{(1)} = 0,$$

protože γ_0 je norma nenulového vektoru g. První složka vlastního vektoru je nulová a to je spor. □

Důsledek 2.5 *Nechť matice* $\mathbf{T}_{n-1}$ *je nedegenerovaná. Pak singulární případ nemůže nastat pro původní problém (1.25).*

DŮKAZ: Je-li $k = n - 1$, pak sloupce matice $\mathbf{Q}_{n-1}$ tvoří bázi prostoru $\mathbb{R}^n$. Problémy (1.25) a (2.70) jsou proto identické a mezi (2.70) a (2.71) existuje regulární transformace. Tvrzení důsledku plyne z předchozí věty v případě $k = n - 1$. □

Důsledek 2.6 *Jestliže pro problém (1.25) nastane singulární případ, pak se Lanczosova třídiagonální matice* $\mathbf{T}_{n-1}$ *degeneruje v nějakém bodě* $k \le n - 1$*.*

Lemma 2.13 *Nechť* $\mathbf{T}_k$ *je nedegenerovaná,* h_k *a* ξ_k *splňují (2.72) a* $\mathbf{T}_k + \xi_k \mathbf{I}_{k+1} \succeq 0$*. Pak* $\mathbf{T}_k + \xi_k \mathbf{I}_{k+1} \succ 0$*.*

DŮKAZ: Sporem: nechť je $\mathbf{T}_k + \xi_k \mathbf{I}_{k+1}$ singulární. Pak existuje nenulový vlastní vektor v_k, pro který je $(\mathbf{T}_k + \xi_k \mathbf{I}_{k+1})v_k = 0$. To spolu s (2.72) implikuje, že

$$0 = v_k^T(\mathbf{T}_k + \xi_k \mathbf{I}_{k+1})h_k = -\gamma_0 v_k^T e_1 = -\gamma_0 v_k^{(1)}.$$

Protože γ_0 je norma vektoru g, který je nenulový, je nutně $v_k^{(1)} = 0$, což je spor. Matice $\mathbf{T}_k + \xi_k \mathbf{I}_{k+1}$ je tedy regulární, podle předpokladu positivně semidefinitní, tudíž je positivně definitní. □

Pokud je tedy matice $\mathbf{T}_k$ nedegenerovaná, má rovnice (2.72) jediné řešení $h_k \in \mathbb{R}^{k+1}$.

Lemma 2.14 *Nechť* $e_{k+1}^T h_k = 0$*. Pak matice* $\mathbf{T}_k$ *je degenerovaná.*

DŮKAZ: Nechť $\mathbf{T}_k$ je nedegenerovaná. Podle předpokladu je $h_k^{(k+1)} = 0$. Protože $\mathbf{T}_k$ je nedegenerovaná, tak z $(k+1)$-ní složky rovnice (2.72) plyne, že $h_k^{(k)} = 0$. Obdobně dostaneme $h_k^{(k-1)} = 0$, atd. až ze druhé složky téže rovnice vyplyne $h_k^{(1)} = 0$. Ale to je spor, neboť současně z první složky plyne, že $h_k^{(1)} \neq 0$. Tedy $\mathbf{T}_k$ je degenerovaná. □

Nyní přistoupíme k řešení problému (1.25).

- Nechť nenastane singulární případ. Dokud jsou matice $\mathbf{T}_i,\ i \geq 1$ nedegenerované, platí $h_i^{(i+1)} \neq 0$ podle lemmatu 2.14. Jestliže existuje k takové, že se $\mathbf{T}_k$ degeneruje, tedy že $\gamma_{k+1} = 0$, pak získáváme řešení problému (1.25), věta 2.11.
- Jestliže nastane singulární případ, pak se matice $\mathbf{T}_k$ pro nějaké $k \leq n-1$ podle důsledku 2.6 degeneruje a má l-blokově diagonální tvar

$$\mathbf{T}_k = \begin{pmatrix} \mathbf{T}_{k_1} & & & \\ & \mathbf{T}_{k_2} & & \\ & & \ddots & \\ & & & \mathbf{T}_{k_l} \end{pmatrix} \tag{2.74}$$

  kde $\gamma_{k_i+1} = 0$ pro $i = 1, \ldots, l$. Pro řešení problému (1.25) platí věta 2.12.

Věta 2.11 *Nechť pro problém (1.25) nenastane singulární případ, nechť* $\gamma_{k+1} = 0$ *a nechť matice* $\mathbf{T}_k$ *je nedegenerovaná. Pak* x_k *řeší problém (1.25).*

DŮKAZ: Protože nenastane singulární případ, platí $g \not\perp \mathcal{S}_1$, kde $\mathcal{S}_1$ je prostor vlastních vektorů asociovaných s nejmenším vlastním číslem λ_1 matice $\mathbf{A}$. Podle lemmatu A.1 je λ_1 též nejmenším vlastním číslem matice $\mathbf{T}_k$. Vektor x_k nyní splňuje podmínky věty 1.5:

- (2.72) $\Rightarrow$ $\mathbf{T}_k + \xi_k \mathbf{I}_{k+1} \succeq 0 \Rightarrow \xi_k \geq -\lambda_1 \Rightarrow \mathbf{A} + \xi_k \mathbf{I}_n \succeq 0$, navíc $\xi_k \geq 0$;
- (2.69) $\Rightarrow$ $\|x_k\| = \|h_k\| \Rightarrow \xi_k(\|x_k\| - \Delta) = \xi_k(\|h_k\| - \Delta) = 0$;
- (2.73) $\Rightarrow$ $\|(\mathbf{A} + \xi_k \mathbf{I}_n)x_k + g\| = 0 \Rightarrow (\mathbf{A} + \xi_k \mathbf{I}_n)x_k + g = 0$. □

Žádané řešení tedy dostaneme z prvního a jediného nedegenerovaného bloku Lanczosovy třídiagonální matice $\mathbf{T}_k$. Zbývá uvažovat možnost, že se matice $\mathbf{T}_k$ degeneruje na l bloků tvaru (2.74), tedy že pro problém (1.25) nastal singulární případ. Předpokládejme, že poslední blok $\mathbf{T}_{k_l}$ je první, jehož nejmenší vlastní číslo se rovná nejmenšímu vlastnímu číslu λ_1 matice $\mathbf{A}$. Nechť dále $[h_{k_1}, \xi_{k_1}]$ je řešení problému (2.71), kde namísto $\mathbf{T}_k$ je její první nedegenerovaný blok $\mathbf{T}_{k_1}$. Pak lze uvažovat dva případy.

Věta 2.12 *Nechť pro problém (1.25) nastane singulární případ a* $\mathbf{T}_k$ *má blokový tvar (2.74). Pak*

1. *Je-li* $-\lambda_1 \leq \xi_{k_1}$, *pak řešení* $[x_k, \xi_k]$ *problému (1.25) je dáno takto:*

   $$x_k = \mathbf{Q}_{k_1} h_{k_1}, \quad \xi_k = \xi_{k_1},$$

   kde $\mathbf{Q}_{k_1} = (q_0, \ldots, q_{k_1}) \in \mathbb{R}^{n \times (k_1+1)}$ *a* $h_{k_1} \in \mathbb{R}^{k_1+1}$ *řeší positivně definitní systém* $(\mathbf{T}_{k_1} + \xi_{k_1} \mathbf{I}_{k_1+1}) h_{k_1} = -\gamma_0 e_1$.

2. *Je-li* $-\lambda_1 > \xi_{k_1}$, *pak řešení* $[x_k, \xi_k]$ *problému (1.25) je dáno takto:*

   $$x_k = \mathbf{Q}_k h_k, \quad \xi_k = -\lambda_1,$$

   kde h_k *má* l *vektorových složek* $h_k = (h^T, 0, \ldots, 0, \kappa z^T)^T$ *odpovídajících blokům* $\mathbf{T}_{k_1}, \ldots, \mathbf{T}_{k_l}$, h *je řešení regulárního třídiagonálního systému*

   $$(\mathbf{T}_{k_1} - \lambda_1 \mathbf{I}_{k_1+1}) h = -\gamma_0 e_1,$$

   z *je vlastní vektor podmatice* $\mathbf{T}_{k_l}$ *odpovídající nejmenšímu vlastnímu číslu* λ_1 *a* κ *je určeno tak, že* $\|h_k\| = \Delta$.

DŮKAZ:

1. Protože $-\lambda_1 \leq \xi_{k_1}$, je $\mathbf{A} + \xi_{k_1}\mathbf{I}_n \succeq 0$. Řešme problém (2.71) pro $k = k_1$. Zde nenastává singulární případ, protože $\mathbf{T}_{k_1}$ je nedegenerovaná matice. Přitom platí

$$\xi_{k_1} \geq 0, \quad \xi_{k_1}(\|h_{k_1}\| - \Delta) = 0 \quad \text{a} \quad \mathbf{T}_{k_1} + \xi_{k_1}\mathbf{I}_{k_1+1} \succeq 0.$$

Podle lemmatu 2.13 je $\mathbf{T}_{k_1} + \xi_{k_1}\mathbf{I}_{k_1+1} \succ 0$ a proto je řešení $h_{k_1} \in \mathbb{R}^{k_1+1}$, které splňuje $\|h_{k_1}\| \leq \Delta$, jediné. Protože $\|x_k\| = \|h_{k_1}\|$, platí

$$\|x_k\| \leq \Delta \quad \text{a} \quad \xi_{k_1}(\|x_k\| - \Delta) = 0.$$

Konečně podle věty 2.10 je

$$\|(\mathbf{A} + \xi_{k_1}\mathbf{I}_n)x_k + g\| = 0$$

a x_k tedy společně s $\xi_k = \xi_{k_1}$ splňuje podmínky věty 1.5.

2. Systém $(\mathbf{T}_{k_1} - \lambda_1\mathbf{I}_{k_1+1})h = -\gamma_0 e_1$ je regulární, protože podle předpokladu patří λ_1 do l-tého bloku $\mathbf{T}_{k_l}$, takže v prvním bloku $\mathbf{T}_{k_1}$ jsou vlastní čísla $\lambda_i > \lambda_1$ a proto je $\mathbf{T}_{k_1} - \lambda_1\mathbf{I}_{k_1+1} \succ 0$. Kromě toho, jelikož $[h_{k_1}, \xi_{k_1}]$ je řešení problému (2.71) pro $k = k_1$, pak pro $-\lambda_1 > \xi_{k_1}$ plyne z rovnic

$$(\mathbf{T}_{k_1} - \lambda_1\mathbf{I}_{k_1+1})h = -\gamma_0 e_1 \quad \text{a} \quad (\mathbf{T}_{k_1} + \xi_{k_1}\mathbf{I}_{k_1+1})h_{k_1} = -\gamma_0 e_1,$$

že $\|h\| < \|h_{k_1}\| \leq \Delta$ podle lemmatu A.2. Existuje tedy κ takové, že

$$\|x_k\|^2 = \|h_k\|^2 = \|h\|^2 + \kappa^2\|z\|^2 = \Delta^2.$$

Navíc platí

$$\xi_{k_1} \geq 0 \quad \Rightarrow \quad -\lambda_1 > 0.$$

Matice $\mathbf{T}_{k_1}$ je první celý nedegenerovaný blok matice $\mathbf{T}_k$ a proto je $\gamma_{k_1+1} = 0$. Obdobně $\mathbf{T}_{k_l}$ je poslední blok, takže $\gamma_{k_l+1} = 0$. Napíšeme-li $x = \mathbf{Q}_{k_1}h$ a $y = \mathbf{Q}_{k_l}z$, pak z (2.63) plyne

$$\begin{aligned} \mathbf{A}x &= \mathbf{A}\mathbf{Q}_{k_1}h = \mathbf{Q}_{k_1}\mathbf{T}_{k_1}h = \mathbf{Q}_{k_1}(\lambda_1 h - \gamma_0 e_1) = \lambda_1 x - g, \\ \mathbf{A}y &= \mathbf{A}\mathbf{Q}_{k_l}z = \mathbf{Q}_{k_l}\mathbf{T}_{k_l}z = \mathbf{Q}_{k_l}\lambda_1 z = \lambda_1 y. \end{aligned}$$

Tedy

$$\begin{aligned} (\mathbf{A} - \lambda_1\mathbf{I}_n)x_k &= (\mathbf{A} - \lambda_1\mathbf{I}_n)\mathbf{Q}_k h_k = (\mathbf{A} - \lambda_1\mathbf{I}_n)(\mathbf{Q}_{k_1}h + \kappa\mathbf{Q}_{k_l}z) = \\ &= (\mathbf{A} - \lambda_1\mathbf{I}_n)(x + \kappa y) = \mathbf{A}x + \kappa\mathbf{A}y - \lambda_1 x - \kappa\lambda_1 y = \\ &= \lambda_1 x - g + \kappa\lambda_1 y - \lambda_1 x - \kappa\lambda_1 y = -g \end{aligned}$$

a x_k společně s $\xi_k = -\lambda_1$ splňuje všechny podmínky věty 1.5. □

Srovnáme-li to s Newtonovou metodou, konkrétně s obrázkem 2.1, vidíme, že výsledek 1. odpovídá singulárnímu případu na obrázku vpravo dole, kde $\xi_\star \geq -\lambda_1$. Naopak výsledek 2. je ekvivalentní singulárnímu případu znázorněnému na obrázku vpravo nahoře, kde platí $\xi_\star = -\lambda_1$ a pro řešení $x_\star \equiv x_k$ se vztahuje analýza singulárního případu uvedená v § 1.3, tedy obdoba vztahu $x_\star = x + \kappa q$ pro $q \in \mathcal{S}_1$.

Nyní uvedeme algoritmus této metody. K použití věty 2.12 však potřebujeme spočítat nejmenší vlastní číslo matice $\mathbf{A}$, což je obtížné pro rozsáhlé problémy. Tvar řešení uvedený v této větě je užitečný jen z teoretického hlediska. Hledáme-li pouze aproximaci lokálně omezeného kroku, vystačíme s řešením v jednodušším tvaru. K vytvoření Lanczosových vektorů použijeme metodu sdružených gradientů (algoritmus 2.10). Na rozdíl od předchozího přístupu (algoritmy 2.7 a 2.8) připustíme též možnost $p_k^T \mathbf{A} p_k \le 0$. Jakmile bude ovšem tento skalární součin malý, přejdeme k Lanczosově metodě (algoritmus 2.11). Zavedeme parametr INT, který bude určovat, zda je řešení uvnitř oblasti či nikoli. Protože dále potřebujeme ukládat matici $\mathbf{Q}_k$, je z praktického hlediska vhodné uvažovat maximální počet sloupců této matice. Zvolíme pevné m, které určuje, jaký maximální počet ortonormálních vektorů q_k připustíme. Obvykle volíme $m = 100$.

Algoritmus 2.12 *Použití Lanczosovy metody pro výpočet lokálně omezeného kroku.*

Zvolíme $\varepsilon_{sg}, \varepsilon \in (0,1)$, *např.* 10^{-6}, $m > 0$ *a položíme*

$$x_0 = 0,\ r_0 = g,\ \gamma_0 = \|g\|,\ q_0 = \frac{1}{\|g\|} g,\ \sigma = 1,\ p_0 = -r_0,\ \text{INT} = \text{„}\mathit{true}\text{“},\ \text{k} = 0.$$

1. *Je-li* $|p_k^T \mathbf{A} p_k| \le \varepsilon_{sg}$, *přejdeme na krok 8. Jinak položíme* $\alpha_k = \frac{r_k^T r_k}{p_k^T \mathbf{A} p_k}$.
2. *Je-li* $k = 0$, *pak* $\delta_0 = \frac{1}{\alpha_0}$, *jinak* $\delta_k = \frac{1}{\alpha_k} + \frac{\beta_{k-1}}{\alpha_{k-1}}$.
3. *Je-li* INT = „*true*“, *ale* $\Big\{\alpha_k \le 0$ *nebo* $\|x_k + \alpha_k p_k\| \ge \Delta\Big\}$, *pak* INT = „*false*“.
4. *Je-li* INT = „*true*“, *pak* $x_{k+1} = x_k + \alpha_k p_k$, *jinak řešíme třídiagonální problém (2.71) pro* h_k.
5. *Položíme* $r_{k+1} = r_k + \alpha_k \mathbf{A} p_k$, $\beta_k = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$ *a* $\gamma_{k+1} = \frac{\sqrt{\beta_k}}{|\alpha_k|}$.
6. *Je-li* INT = „*true*“, *pak: Je-li* $\|r_{k+1}\| \le \varepsilon$, *položíme* $x_\star = x_{k+1}$ *a* `STOP`.
   Je-li $k + 1 = m$, *položíme* $x_\star = \mathbf{Q}_k h_k$ *a* `STOP`.
   Je-li INT = „*false*“, *pak: Je-li* $\gamma_{k+1} \cdot |e_{k+1}^T h_k| \le \varepsilon$, *položíme* $x_\star = \mathbf{Q}_k h_k$ *a* `STOP`.
7. *Položíme* $\sigma := -\sigma\,\mathrm{sgn}(\alpha_k)$, $q_{k+1} = \sigma \frac{1}{\|r_{k+1}\|} r_{k+1}$, $p_{k+1} = -r_{k+1} + \beta_k p_k$, $k := k + 1$ *a návrat na krok 1.*
8. *Položíme* INT = „*false*“.
9. *Spočítáme* $\delta_k = q_k^T \mathbf{A} q_k$.
10. *Je-li* $k = 0$, *pak položíme* $t_1 = \mathbf{A} q_0 - \delta_0 q_0$, *jinak* $t_{k+1} = \mathbf{A} q_k - \delta_k q_k - \gamma_k q_{k-1}$.
11. *Spočítáme* $\gamma_{k+1} = \|t_{k+1}\|$ *a řešíme třídiagonální problém (2.71) pro* h_k.
12. *Je-li* $\gamma_{k+1} \cdot |e_{k+1}^T h_k| \le \varepsilon$ *nebo* $k + 1 = m$, *položíme* $x_\star = \mathbf{Q}_k h_k$ *a* `STOP`.
13. *Položíme* $q_{k+1} = \frac{1}{\gamma_{k+1}} t_{k+1}$, $k := k + 1$ *a návrat na krok 9.*

Pokud potřebujeme spočítat $\|x_k + \alpha_k p_k\|$, lze to provést rekurentně. Jelikož

$$\|x_k + \alpha_k p_k\|^2 = \|x_k\|^2 + 2\alpha_k x_k^T p_k + \alpha_k^2 \|p_k\|^2,$$

lze použít následující lemma.

Lemma 2.15 *Platí*

1. $x_k^T p_k = \beta_{k-1} \left[x_{k-1}^T p_{k-1} + \alpha_{k-1} \|p_{k-1}\|^2 \right]$;
2. $\|p_k\|^2 = \|r_k\|^2 + \beta_{k-1}^2 \|p_{k-1}\|^2$.

Všechny hodnoty vystupující v rekurenci jsou známé z předchozí iterace. Lze tedy snadno spočítat $\|x_k + \alpha_k p_k\|$.
DŮKAZ: Využijeme ortogonalitu residuí $\{r_i\}$.

1. $x_k^T p_k = (x_{k-1} + \alpha_{k-1} p_{k-1})^T (-r_k + \beta_{k-1} p_{k-1}) = \beta_{k-1} \left[x_{k-1}^T p_{k-1} + \alpha_{k-1} \|p_{k-1}\|^2 \right]$, neboť
   (a) p_{k-1} je lineární kombinací $r_{k-1}, r_{k-2}, \dots, r_0$ a proto $p_{k-1}^T r_k = 0$;
   (b) x_{k-1} je lineární kombinací $p_{k-2}, \dots, p_0$, tedy $r_{k-2}, \dots, r_0$ a proto $x_{k-1}^T r_k = 0$.
2. $\|p_k\|^2 = p_k^T p_k = (-r_k + \beta_{k-1} p_{k-1})^T (-r_k + \beta_{k-1} p_{k-1}) = \|r_k\|^2 + \beta_{k-1}^2 \|p_{k-1}\|^2$, neboť opět $p_{k-1}^T r_k = 0$. □

Nakonec zbývá uvažovat problém (2.71) pro h_k. Aplikujeme algoritmus 2.2 na rovnici

$$\phi(\xi_k) \equiv \frac{1}{\Delta} - \frac{1}{\|h_k\|} = 0,$$

kde h_k řeší systém (2.72). Protože matice $\mathbf{T}_k$ je třídiagonální, jedná se o jednoduchý problém. Choleského rozklad matice $\mathbf{T}_k + \xi_k \mathbf{I}_{k+1}$ provedeme ve tvaru

$$\mathbf{T}_k + \xi_k \mathbf{I}_{k+1} = \mathbf{B}\mathbf{D}\mathbf{B}^T,$$

kde $\mathbf{B}$ je jednotková horní bidiagonální a $\mathbf{D}$ je diagonální matice. Výpočet (2.72) se tím zjednoduší o jeden maticovo-vektorový součin:

$$\mathbf{B}\mathbf{D}\mathbf{B}^T h_k = -\gamma_0 e_1 \quad \Leftrightarrow \quad \mathbf{D}\mathbf{B}^T h_k = -\gamma_0 \mathbf{B}^{-1} e_1 = -\gamma_0 e_1 \quad \Leftrightarrow \quad \mathbf{B}^T h_k = -\gamma_0 \mathbf{D}^{-1} e_1,$$

neboť $\mathbf{B}^{-1}$ má v prvním sloupci stejně jako $\mathbf{B}$ vektor e_1 a proto platí $\mathbf{B}^{-1} e_1 = e_1$.

Globální konvergence této metody plyne z toho, že hledáme optimální lokálně omezený krok redukovaného problému (2.71) pro h_k.

Věta 2.13 *Metoda sestavená na základě algoritmu 2.12 je globálně konvergentní.*

DŮKAZ: Dokážeme nerovnost (1.11) pro $x_\star$, která podle poznámky 1.3 stačí ke globální konvergenci uvedené metody. Je-li řešení uvnitř oblasti, $\|x_\star\| < \Delta$, tj. INT = „true“, platí $p_k^T \mathbf{A} p_k > 0$, metoda je totožná s metodou sdružených gradientů, algoritmus 2.7, a globální konvergence plyne z věty 2.8. Je-li INT = „false“, pak řešíme problém (2.71) pro h_k a řešení $x_\star$ je tvaru $x_\star = \mathbf{Q}_k h_k$. Protože se problém (2.71) řeší pomocí algoritmu 2.2, je h_k optimální lokálně omezený krok pro problém (2.71) a tudíž platí

$$-\tilde{\psi}(h_k) \geq \underline{\tilde{\sigma}} \, \|\gamma_0 e_1\| \min \left\{ \|h_k\|, \frac{\|\gamma_0 e_1\|}{\|\mathbf{T}_k\|} \right\},$$

kde
$$\tilde{\psi}(h) = \frac{1}{2}\, h^T \mathbf{T}_k h + (\gamma_0 e_1)^T h,$$
takže pro $x_\star = \mathbf{Q}_k h_k$ máme $\|x_\star\| = \|h_k\|$ a podle (2.65) $\|\gamma_0 e_1\| = \|g\|$, $\|\mathbf{T}_k\| \le \|\mathbf{A}\|$, tedy

$$\begin{aligned} -\psi(x_\star) &= -\frac{1}{2}\, h_k^T \mathbf{Q}_k^T \mathbf{A} \mathbf{Q}_k h_k + g^T \mathbf{Q}_k h_k = -\tilde{\psi}(h_k) \ge \\ &\ge \underline{\tilde{\sigma}}\, \|\gamma_0 e_1\| \min\left\{ \|h_k\|, \frac{\|\gamma_0 e_1\|}{\|\mathbf{T}_k\|} \right\} \ge \underline{\sigma}\, \|g\| \min\left\{ \|x_\star\|, \frac{\|g\|}{\|\mathbf{A}\|} \right\}, \end{aligned}$$

kde $\underline{\sigma} = \underline{\tilde{\sigma}}$. □

Uvedeme ještě dvě nové modifikace použití Lanczosovy metody, které se v praxi osvědčily, kombinované s metodou sdružených gradientů a Choleského rozkladem.

V prvním případě zvolíme pevné m (obvykle malé) a spočítáme m kroků Lanczosovy metody (algoritmus 2.11). Získáme třídiagonální matici $\mathbf{T}_{m-1}$ řádu m a pomocí Choleského rozkladu matice $\mathbf{T}_{m-1} + \xi_{m-1}\mathbf{I}_m$ řešíme problém (2.71) pro $h_{m-1} \in \mathbb{R}^m$, abychom získali parametr $\xi_\star > 0$. Tato hodnota nám pro přibližný výpočet řešení problému (1.25) postačí a řešíme rovnici $(\mathbf{A} + \xi_\star \mathbf{I})x + g = 0$ metodou sdružených gradientů. Dostaneme lepší aproximaci lokálně omezeného kroku, nežli v případě $\xi = 0$. Obdobně jako u algoritmu 2.9 budeme uvažovat předpodmínění s maticí $\mathbf{C} \in \mathbb{R}^n$, která je symetrická, positivně definitní, snadno invertovatelná a platí $\mathbf{C} = \mathbf{R}^T\mathbf{R}$, kde $\mathbf{R}^T\mathbf{R}$ je neúplný rozklad matice $\mathbf{A} + \xi_\star \mathbf{I}$ a $\mathbf{R}$ má nenulové prvky pouze tam, kde jsou nenulové prvky matice $\mathbf{A} + \xi_\star \mathbf{I}$. Tuto variantu nazveme LCG, Lanczos – sdružené gradienty pro výpočet lokálně omezeného kroku.

Algoritmus 2.13 *Kombinovaná metoda LCG pro výpočet lokálně omezeného kroku.*

Zvolíme $\varepsilon \in (0,1)$, $\omega \in (0,1\rangle$, $\bar{\xi} > 0$, $m, m_1, m_2 \in \mathbb{N}$.

1. *Provedeme* m *kroků Lanczosova algoritmu 2.11, dostaneme matici* $\mathbf{T}$ *řádu* m *a položíme* $\xi = 0$, $i = 1$.

2. *Jestliže* $\xi \ge \bar{\xi}$, *položíme* $\xi = \bar{\xi}$ *a přejdeme na krok 6.*

3. *Je-li* $\mathbf{T} + \xi\mathbf{I} \succ 0$, *provedeme rozklad* $\mathbf{T} + \xi\mathbf{I} = \mathbf{B}\mathbf{D}\mathbf{B}^T$ *a přejdeme na krok 4. Jinak zvětšíme* ξ *a opakujeme krok 3.*

4. *Řešíme rovnici* $\mathbf{B}\mathbf{D}\mathbf{B}^T h + \gamma_0 e_1 = 0$. *Jestliže* $\|h\| \le \Delta$ *nebo* $i = m_1$, *přejdeme na krok 6.*

5. *Řešíme rovnici* $\mathbf{B}w = h$, *položíme* $\xi := \xi + \frac{\|h\|^2}{w^T\mathbf{D}^{-1}w} \cdot \frac{\|h\| - \omega\Delta}{\omega\Delta}$, $i := i + 1$ *a návrat na krok 2.*

6. *Položíme* $x_0 = 0$, $r_0 = g$, $p_0 = -\mathbf{C}^{-1} r_0$, $i = 0$.

7. *Spočítáme* $\eta_i = p_i^T(\mathbf{A} + \xi\mathbf{I})p_i$. *Je-li* $\eta_i \le 0$, *určíme* $\kappa > 0$ *tak, že* $\|x_i + \kappa p_i\| = \Delta$, *položíme* $x_\star = x_i + \kappa p_i$ *a* `STOP`.

8. *Položíme* $\alpha_i = \frac{r_i^T \mathbf{C}^{-1} r_i}{\eta_i}$ *a* $x_{i+1} = x_i + \alpha_i p_i$. *Je-li* $\|x_{i+1}\| \ge \Delta$, *určíme* $\kappa > 0$ *tak, že* $\|x_i + \kappa p_i\| = \Delta$, *položíme* $x_\star = x_i + \kappa p_i$ *a* `STOP`.

9. Spočítáme $r_{i+1} = r_i + \alpha_i(\mathbf{A} + \xi\mathbf{I})p_i$. *Je-li* $\|r_{i+1}\| \le \varepsilon\|g\|$ *nebo* $i+1 = n + m_2$, *položíme* $x_\star = x_{i+1}$ *a* `STOP`.

10. Položíme $\beta_i = \frac{r_{i+1}^T\mathbf{C}^{-1}r_{i+1}}{r_i^T\mathbf{C}^{-1}r_i}$, $p_{i+1} = -\mathbf{C}^{-1}r_{i+1} + \beta_i p_i$, $i := i+1$ *a návrat na krok 7.*

Obvykle volíme $\varepsilon = 10^{-6}$, $\omega = 0.9$, $\bar{\xi} = 100$, $m = 5$, $m_1 = 10$, $m_2 = 3$. Provedeme tedy m kroků Lanczosovy metody, poté provedeme nejvýše m_1 kroků Choleského rozkladu pro nalezení parametru $\xi_\star$ a nakonec řešíme rovnici $(\mathbf{A} + \xi_\star\mathbf{I})x + g = 0$ metodou sdružených gradientů a to tak, že pokud nedostaneme řešení po n krocích (např. kvůli špatné podmíněnosti soustavy), provedeme navíc m_2 kroků. U Choleského rozkladu můžeme uvažovat meze ξ_L a ξ_U, ale rovněž jednoduchá mez $\bar{\xi}$ pro výpočet aproximace $\xi_\star$ dostačuje. Při aktualizaci ξ volíme místo Δ hodnotu $\omega\Delta$, která se v praxi ukázala z hlediska konvergence jako výhodnější. Vynecháme-li kroky 1.-5. a položíme $\xi_\star = 0$, dostaneme metodu sdružených gradientů uvedenou jako algoritmus 2.7.

Věta 2.14 *Kombinovaná metoda LCG, sestavená na základě algoritmu 2.13, je globálně konvergentní.*

Důkaz: Ukážeme, že $x_\star$ splňuje nerovnost (1.11). Uvažujme nejprve tuto metodu bez předpodmínění, tj. $\mathbf{C} = \mathbf{I}$. Protože aplikujeme metodu sdružených gradientů na matici $\mathbf{A} + \xi\mathbf{I}$, podle věty 2.8 platí

$$-\psi(x_\star) \ge \underline{\tilde{\sigma}}\,\|g\| \min\left\{\|x_\star\|, \frac{\|g\|}{\|\mathbf{A} + \xi\mathbf{I}\|}\right\},$$

kde $\underline{\tilde{\sigma}} \in (0,1)$. Číslo ξ se určuje řešením úlohy (2.71) pomocí Choleského rozkladu matice $\mathbf{T} + \xi\mathbf{I}$, takže platí

$$\xi \le \frac{\gamma_0}{\Delta} + \|\mathbf{T}\| \le \frac{\|g\|}{\Delta} + \|\mathbf{A}\|$$

podle (2.21), (2.65) a toho, že $\gamma_0 = \|g\|$. Odtud máme

$$\frac{\|\mathbf{A} + \xi\mathbf{I}\|}{\|g\|} \le \frac{2\|\mathbf{A}\|}{\|g\|} + \frac{1}{\Delta} \le 2\max\left\{\frac{2\|\mathbf{A}\|}{\|g\|}, \frac{1}{\Delta}\right\} \quad \Rightarrow \quad \frac{\|g\|}{\|\mathbf{A} + \xi\mathbf{I}\|} \ge \min\left\{\frac{\|g\|}{4\|\mathbf{A}\|}, \frac{\Delta}{2}\right\},$$

takže

$$\begin{aligned} -\psi(x_\star) &\ge \underline{\tilde{\sigma}}\,\|g\| \min\left\{\|x_\star\|, \frac{\|g\|}{4\|\mathbf{A}\|}, \frac{\Delta}{2}\right\} \ge \underline{\tilde{\sigma}}\,\|g\| \min\left\{\frac{\|x_\star\|}{2}, \frac{\|g\|}{4\|\mathbf{A}\|}\right\} \ge \\ &\ge \frac{1}{4}\,\underline{\tilde{\sigma}}\,\|g\| \min\left\{\|x_\star\|, \frac{\|g\|}{\|\mathbf{A}\|}\right\} \end{aligned}$$

Položíme-li $\underline{\sigma} = \frac{1}{4}\underline{\tilde{\sigma}}$, platí $0 < \underline{\sigma} < 1$. Jestliže uvažujeme tuto metodu s předpodmíněním, pak je rovněž globálně konvergentní podle věty 2.9, kde $\underline{\sigma} = \frac{1}{4\kappa(\mathbf{C})}\underline{\tilde{\sigma}}$. □

Druhá modifikace spočívá v tom, že použijeme metodu sdružených gradientů ke generování Lanczosových vektorů, vztah (2.66). Zvolíme pevné m (obvykle malé) a spočítáme m kroků metody sdružených gradientů (algoritmus 2.10). Tím získáme třídiagonální matici $\mathbf{T}_{m-1}$ řádu m a dále postupujeme následujícím způsobem. Jestliže $\|x_m\| < \Delta$, pokračujeme metodou sdružených gradientů až do konce, protože již neaktualizujeme matici $\mathbf{T}_{m-1}$. Jestliže $\|x_m\| \ge \Delta$, řešíme přibližně problém (2.71) pro $h_{m-1} \in \mathbb{R}^m$ pomocí Choleského rozkladu matice $\mathbf{T}_{m-1} + \xi_{m-1}\mathbf{I}_m$ a položíme $x_\star = \mathbf{Q}_{m-1}h_{m-1}$. U této metody nelze uvažovat předpodmínění, neboť bychom dostali nesprávnou třídiagonální matici $\mathbf{T}_{m-1}$, tedy jiné koeficienty α_i a β_i. Tuto variantu nazveme CGL, sdružené gradienty – Lanczos pro výpočet lokálně omezeného kroku.

Algoritmus 2.14 *Kombinovaná metoda CGL pro výpočet lokálně omezeného kroku.*

Zvolíme $\varepsilon \in (0,1)$, $\omega \in (0,1\rangle$, $\bar{\omega} > 1$, $m, m_1, m_2 \in \mathbb{N}$.

1. *Položíme* $x_0 = 0$, $r_0 = g$, $p_0 = -r_0$, $q_0 = \frac{1}{\|g\|}\, g$, $\sigma = 1$, $i = 0$.

2. *Spočítáme* $\eta_i = p_i^T \mathbf{A} p_i$. *Je-li* $\eta_i > 0$, *přejdeme na krok 3. Jinak určíme* $\kappa > 0$ *tak, že* $\|x_i + \kappa p_i\| = \Delta$, *položíme* $x_\star = x_i + \kappa p_i$ *a* STOP.

3. *Položíme* $\alpha_i = \frac{r_i^T r_i}{\eta_i}$ *a* $x_{i+1} = x_i + \alpha_i p_i$. *Jestliže*
$$\|x_{i+1}\| \geq \Delta \quad a \quad i+1 > m \qquad nebo \qquad \|x_{i+1}\| \geq \bar{\omega}\Delta,$$
určíme $\kappa > 0$ *tak, že* $\|x_i + \kappa p_i\| = \Delta$, *položíme* $x_\star = x_i + \kappa p_i$ *a* STOP. *Jinak spočítáme* $r_{i+1} = r_i + \alpha_i \mathbf{A} p_i$ *a přejdeme na krok 4.*

4. *Pokud* $\|x_{i+1}\| \geq \Delta$: *jestliže* $i+1 = m$ *nebo* $\|r_{i+1}\| \leq \varepsilon\|g\|$, *přejdeme na krok 7.* *Pokud* $\|x_{i+1}\| < \Delta$: *jestliže* $i+1 = n + m_2$ *nebo* $\|r_{i+1}\| \leq \varepsilon\|g\|$, *položíme* $x_\star = x_{i+1}$ *a* STOP.

5. *Spočítáme* $\beta_i = \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i}$ *a* $p_{i+1} = -r_{i+1} + \beta_i p_i$. *Je-li* $i+1 < m$, *položíme* $\sigma := -\sigma\,\mathrm{sgn}(\alpha_i)$ *a* $q_{i+1} = \sigma \frac{1}{\|r_{i+1}\|}\, r_{i+1}$.

6. *Je-li* $i = 0$, *pak* $\delta_0 = \frac{1}{\alpha_0}$, *jinak* $\delta_i = \frac{1}{\alpha_i} + \frac{\beta_{i-1}}{\alpha_{i-1}}$ *a* $\gamma_i = \frac{\sqrt{\beta_{i-1}}}{|\alpha_{i-1}|}$. *Položíme* $i := i+1$ *a návrat na krok 2.*

7. *Je-li* $i = 0$, *pak* $\delta_0 = \frac{1}{\alpha_0}$, *jinak* $\delta_i = \frac{1}{\alpha_i} + \frac{\beta_{i-1}}{\alpha_{i-1}}$ *a* $\gamma_i = \frac{\sqrt{\beta_{i-1}}}{|\alpha_{i-1}|}$. *Položíme* $\gamma_0 = \|g\|$, $\xi = 0$, $i = 1$.

8. *Je-li* $\mathbf{T} + \xi\mathbf{I} \succ 0$, *provedeme rozklad* $\mathbf{T} + \xi\mathbf{I} = \mathbf{B}\mathbf{D}\mathbf{B}^T$ *a přejdeme na krok 9. Jinak zvětšíme* ξ *a opakujeme krok 8.*

9. *Řešíme rovnici* $\mathbf{B}\mathbf{D}\mathbf{B}^T h + \gamma_0 e_1 = 0$. *Jestliže buď* $\|h\| \leq \Delta$ *nebo* $i = m_1$, *položíme* $x_\star = \mathbf{Q}h$ *a* STOP.

10. *Řešíme rovnici* $\mathbf{B}w = h$, *položíme* $\xi := \xi + \frac{\|h\|^2}{w^T \mathbf{D}^{-1} w} \cdot \frac{\|h\| - \omega\Delta}{\omega\Delta}$, $i := i+1$ *a návrat na krok 8.*

Obvykle volíme $\varepsilon = 10^{-6}$, $\omega = 0.9$, $\bar{\omega} = 100$, $m = 10$, $m_1 = 10$, $m_2 = 3$. Provedeme tedy nejméně m kroků metody sdružených gradietů, i když překročíme hranici (skončíme pouze v případě velkého překročení hranice, parametr $\bar{\omega}$). To je rozdíl oproti algoritmu 2.7, kde končíme vždy při překročení hranice. Potom se rozhodujeme, zdali zůstaneme u metody sdružených gradietů, $\|x_m\| < \Delta$, nebo přejdeme na Lanczosovu metodu, $\|x_m\| \geq \Delta$. V prvním případě provedeme nejvýše $n+m_2$ iterací a algoritmus je totožný s algoritmem 2.7. Ve druhém případě ukončíme Lanczosovu metodu nejvýše po m_1 iteracích a řešení h, i když nesplňuje podmínku $\|h\| \leq \Delta$, použijeme k výpočtu přibližného řešení $x_\star$. Při použití Choleského rozkladu pro aktualizaci ξ opět zvolíme místo Δ hodnotu $\omega\Delta$.

Věta 2.15 *Kombinovaná metoda CGL, sestavená na základě algoritmu 2.14, je globálně konvergentní.*

DŮKAZ: Nechť nejprve platí $\|x_m\| < \Delta$. Pak je algoritmus 2.14 totožný s algoritmem 2.7 a platí věta 2.8. Jestliže $\|x_m\| \geq \Delta$, řešíme problém (2.71) pro h_{m-1} s maticí $\mathbf{T}_{m-1}$ a položíme $x_\star = \mathbf{Q}_{m-1} h_{m-1}$. Platí tedy věta 2.13. □

Všechny tři algoritmy 2.12, 2.13 a 2.14, které jsme zde uvedli, hledají přibližné řešení $x_\star$ lokálně omezeného kroku na Krylovově podprostoru $\mathcal{K}_{k+1}$ pro nějaké k a v praxi se ukázaly (hlavně algoritmus 2.13) jako efektivní.

V poslední části této kapitoly, § 2.7, se zaměříme na trochu složitější metodu, ve které budeme hledat $x_\star$ opět na celém prostoru $\mathbb{R}^n$ a toto $x_\star$ je optimálním lokálně omezeným krokem.

2.7 Parametrizovaný problém vlastních čísel

V této části převedeme problém hledání minima funkce ψ na parametrizovaný problém vlastních čísel, [08], [46], [54], [55], [56], [57], [58]. Jak je patrné z (1.25)-(1.26), pro $g = 0$ získáváme problém nalezení nejmenšího vlastního čísla a odpovídajícího vlastního vektoru matice $\mathbf{A}$, čímž získáme řešení. Jestliže $g \neq 0$, přidáním nového parametru τ převedeme problém (1.25) na nový problém, nalezení nejmenšího vlastního páru jisté matice $\mathbf{B}_\tau$. Cílem bude nalézt takový parametr $\tau_\star$, který zajistí, že vlastní vektor bude obsahovat řešení problému (1.25). Nebudeme potřebovat maticové rozklady, výsledný algoritmus používá pouze násobení maticí. Pro nalezení vlastních párů použijeme implicitně restartovanou Lanczosovu metodu, [29].

2.7.1 Struktura problému

Nechť je dán problém (1.25)-(1.26). Pro řešení $x_\star$ platí věta 1.5, je tedy řešením (1.25) právě když platí

(2.75) $$\|x_\star\| \leq \Delta, \quad (\mathbf{A} + \xi_\star \mathbf{I})x_\star + g = 0, \quad \xi_\star \geq 0, \quad (\|x_\star\| - \Delta)\xi_\star = 0, \quad \mathbf{A} + \xi_\star \mathbf{I} \succeq 0,$$

tedy $\xi_\star \geq -\lambda_1$, kde λ_1 je nejmenší vlastní číslo matice $\mathbf{A}$. Zavedeme parametr $\tau \in \mathbb{R}$ a sestrojíme matici

(2.76) $$\mathbf{B}_\tau = \begin{pmatrix} \tau & g^T \\ g & \mathbf{A} \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)},$$

pro kterou platí

(2.77) $$\frac{\tau}{2} + \psi(x) = \frac{1}{2}(1, x^T)\mathbf{B}_\tau \begin{pmatrix} 1 \\ x \end{pmatrix}.$$

Problém (1.25) můžeme přepsat na problém

(2.78) $$\min_{y \in \mathbb{R}^{n+1}} \frac{1}{2} y^T \mathbf{B}_\tau y \quad \text{vzhledem k} \quad y^T y \leq 1 + \Delta^2, \quad e_1^T y = 1,$$

kde $y = (1, x^T)^T$, $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^{n+1}$, což dává podnět k tomu, že řešení problému (1.25) lze získat nalezením nejmenšího vlastního čísla $-\xi$ a jemu odpovídajícího vlastního vektoru $(1, x^T)^T$ matice $\mathbf{B}_\tau$ pro vhodné τ. Rovnice

(2.79) $$\begin{pmatrix} \tau & g^T \\ g & \mathbf{A} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ x \end{pmatrix} = (-\xi) \begin{pmatrix} 1 \\ x \end{pmatrix}$$

je ekvivalentní zápisu

(2.80) $$\tau + \xi = -g^T x, \quad (\mathbf{A} + \xi\mathbf{I})x + g = 0.$$

Nechť $\lambda_1, \lambda_2, \ldots, \lambda_n$ jsou vlastní čísla matice $\mathbf{A}$ a $q_1, q_2, \ldots, q_n$ příslušné ortonormální vlastní vektory (q_j odpovídá λ_j), které tvoří bázi prostoru $\mathbb{R}^n$. Nechť $g = \sum_{j=1}^n \overline{\vartheta}_j q_j$ a $\xi \neq -\lambda_j \ \forall j$. Pak pro $j = 1, \ldots, n$ platí

$$\mathbf{A}q_j = \lambda_j q_j \quad \Rightarrow \quad (\mathbf{A} + \xi\mathbf{I})q_j = (\lambda_j + \xi)q_j \quad \Rightarrow \quad (\mathbf{A} + \xi\mathbf{I})^{-1}\overline{\vartheta}_j q_j = \frac{\overline{\vartheta}_j}{\lambda_j + \xi}\, q_j.$$

Sečteme přes j :

$$(\mathbf{A} + \xi\mathbf{I})^{-1} \sum_{j=1}^n \overline{\vartheta}_j q_j = \sum_{j=1}^n \frac{\overline{\vartheta}_j}{\lambda_j + \xi}\, q_j$$

a dostaneme

$$\begin{aligned} g^T(\mathbf{A} + \xi\mathbf{I})^{-1}g &= \left(\sum_{j=1}^n \overline{\vartheta}_j q_j^T\right)(\mathbf{A} + \xi\mathbf{I})^{-1}\left(\sum_{j=1}^n \overline{\vartheta}_j q_j\right) = \\ &= \left(\sum_{j=1}^n \overline{\vartheta}_j q_j^T\right)\sum_{j=1}^n \frac{\overline{\vartheta}_j}{\lambda_j + \xi}\, q_j = \sum_{j=1}^n \frac{\overline{\vartheta}_j^2}{\lambda_j + \xi}. \end{aligned}$$

Přečíslujme nyní vlastní čísla matice $\mathbf{A}$ tak, že $\lambda_1 < \lambda_2 < \ldots < \lambda_m$ jsou navzájem různá, $m \leq n$. Nechť $q_1, \ldots, q_{i_1}$ odpovídají λ_1, $q_{i_1+1}, \ldots, q_{i_2}$ odpovídají λ_2, atd. až $q_{i_{m-1}+1}, \ldots, q_{i_m}$ odpovídají λ_m. Pak

$$g = \sum_{j=1}^n \overline{\vartheta}_j q_j = \sum_{j=1}^{i_1} \overline{\vartheta}_j q_j + \sum_{j=i_1+1}^{i_2} \overline{\vartheta}_j q_j + \ldots + \sum_{j=i_{m-1}+1}^{i_m} \overline{\vartheta}_j q_j.$$

Označíme-li nyní

$$\vartheta_1^2 = \sum_{j=1}^{i_1} \overline{\vartheta}_j^2, \quad \vartheta_2^2 = \sum_{j=i_1+1}^{i_2} \overline{\vartheta}_j^2, \quad \ldots, \quad \vartheta_m^2 = \sum_{j=i_{m-1}+1}^{i_m} \overline{\vartheta}_j^2,$$

pak platí

$$-g^T x = g^T(\mathbf{A} + \xi\mathbf{I})^{-1}g = \sum_{j=1}^n \frac{\overline{\vartheta}_j^2}{\lambda_j + \xi} = \sum_{j=1}^m \frac{\vartheta_j^2}{\lambda_j + \xi}$$

a

$$x^T x = g^T(\mathbf{A} + \xi\mathbf{I})^{-2}g = \sum_{j=1}^n \frac{\overline{\vartheta}_j^2}{(\lambda_j + \xi)^2} = \sum_{j=1}^m \frac{\vartheta_j^2}{(\lambda_j + \xi)^2},$$

kde ϑ_j^2 je součet čtverců rozvojových koeficientů vektoru g v bázi vlastních vektorů, odpovídající všem vlastním vektorům asociovaným s λ_j. Jestliže $\xi = -\lambda_i$ pro nějaké i, pak pro $j \neq i$ platí

$$(\mathbf{A} - \lambda_i\mathbf{I})q_j = (\lambda_j - \lambda_i)q_j \quad \Rightarrow \quad q_j = (\mathbf{A} - \lambda_i\mathbf{I})^\dagger(\mathbf{A} - \lambda_i\mathbf{I})q_j = (\lambda_j - \lambda_i)(\mathbf{A} - \lambda_i\mathbf{I})^\dagger q_j$$

a tedy obdobně pro případ $x = -(\mathbf{A} - \lambda_i \mathbf{I})^\dagger g$ platí

$$-g^T x = g^T (\mathbf{A} - \lambda_i \mathbf{I})^\dagger g = \sum_{j=1,\, j\neq i}^{m} \frac{\vartheta_j^2}{\lambda_j - \lambda_i}$$

a

$$x^T x = g^T (\mathbf{A} - \lambda_i \mathbf{I})^\dagger (\mathbf{A} - \lambda_i \mathbf{I})^\dagger g = \sum_{j=1,\, j\neq i}^{m} \frac{\vartheta_j^2}{(\lambda_j - \lambda_i)^2}.$$

Nechť $\mathcal{S}_j$ značí množiny vlastních vektorů asociovaných s vlastním číslem λ_j matice $\mathbf{A}$, tedy

$$\mathcal{S}_j = \{q : \mathbf{A}q = \lambda_j q\}, \quad j = 1, \ldots, m.$$

Jestliže je $g \perp \mathcal{S}_k$ pro nějaká $k \in \{1, \ldots, m\}$ (např. $k = 1$ pro $\xi_\star = -\lambda_1$, lemma 1.5), pak jsou odpovídající $\vartheta_k = 0$. Definujeme-li množinu indexů $\mathcal{M}$ jako

$$\mathcal{M} = \{1, \ldots, m\} - \{j : \xi = -\lambda_j\} - \{j : g \perp \mathcal{S}_j\},$$

pak platí

$$-g^T x = \sum_{j \in \mathcal{M}} \frac{\vartheta_j^2}{\lambda_j + \xi} \qquad \text{a} \qquad x^T x = \sum_{j \in \mathcal{M}} \frac{\vartheta_j^2}{(\lambda_j + \xi)^2} \tag{2.81}$$

Označme $-\xi_1(\tau)$ nejmenší vlastní číslo matice $\mathbf{B}_\tau$. Z Cauchyho věty (věta A.1) plyne, že vlastní čísla matice $\mathbf{A}$ oddělují vlastní čísla matice $\mathbf{B}_\tau$. Proto je $\xi_1(\tau) \geq -\lambda_1$. To znamená, že matice $\mathbf{A} + \xi_1(\tau)\mathbf{I}$ je nezávisle na hodnotě $\tau \in \mathbb{R}$ vždy positivně semidefinitní. Definujme dále funkci $\phi(\xi)$:

$$\phi(\xi) \stackrel{\text{def}}{=} \sum_{j \in \mathcal{M}} \frac{\vartheta_j^2}{\lambda_j + \xi} = -g^T x \quad \Rightarrow \quad \phi'(\xi) = -\sum_{j \in \mathcal{M}} \frac{\vartheta_j^2}{(\lambda_j + \xi)^2} = -x^T x. \tag{2.82}$$

Definiční obor funkce $\phi(\xi)$ je podle (2.81) množina $\mathbb{R} - \{-\lambda_j : j \in \mathcal{M}\}$. Nyní uděláme shrnutí celého postupu hledání řešení problému (1.25).

1. Najdeme nejmenší vlastní číslo $-\xi_1(\tau)$ a odpovídající vlastní vektor y matice $\mathbf{B}_\tau$ pro dané τ.

2. Normalizujeme vektor y tak, aby měl první složku rovnou jedné: $y = (1, x^T)^T$. Případ, kdy je první složka rovna nule, je popsán níže.

3. Vypočítáme hodnotu $\phi'(\xi_1(\tau))$.

4. Podaří-li se nám najít takové $\tau_\star$, že odpovídající x splňuje
$$-x^T x = \phi'(\xi_1(\tau_\star)) = -\Delta^2, \quad \text{kde} \quad \tau_\star + \xi_1(\tau_\star) = -g^T x = \phi(\xi_1(\tau_\star)),$$
pak jsou splněny podmínky pro řešení, konkrétně
$$(\mathbf{A} + \xi_1(\tau_\star)\mathbf{I})x + g = 0, \quad (\|x\| - \Delta)\xi_1(\tau_\star) = 0 \quad \text{a} \quad \mathbf{A} + \xi_1(\tau_\star)\mathbf{I} \succeq 0.$$

5. Jestliže $\xi_1(\tau_\star) \geq 0$, pak $x_\star \equiv x$ je řešením problému (1.25) na hranici s odpovídajícím $\xi_\star \equiv \xi_1(\tau_\star)$.

6. Pokud během iteračního procesu hledání $\tau_\star$ nastane případ $-\xi_1(\tau) > 0$, pak z Cauchyho věty plyne, že matice $\mathbf{A}$ je positivně definitní a příslušné řešení $x_\star$ uvnitř oblasti lze získat např. metodou sdružených gradientů, algoritmus 2.7.

Poznamenáváme, že nejen ξ, ale i x závisí na τ.

2.7.2 Singulární případ

V první kapitole, § 1.3, jsme definovali singulární případ, který nastává, jestliže je vektor g kolmý na prostor $\mathcal{S}_1$, tedy prostor vlastních vektorů asociovaných s nejmenším vlastním číslem λ_1 matice $\mathbf{A}$.

Podívejme se nyní na případ, kdy všechny vlastní vektory odpovídající nejmenšímu vlastnímu číslu $-\xi_1(\tau)$ mají první složku nulovou a nemohou tedy být normalizovány tak, aby ji měly rovnu jedné. Níže uvedené lemma říká, že tato situace nastává právě v singulárním případě.

Lemma 2.16 *Nechť* $\tau \in \mathbb{R}$ *a* $1 \le j \le m$. *Pak* $g \perp \mathcal{S}_j$ *právě tehdy, když pro každé* $q \in \mathcal{S}_j$ *je* $\{\lambda_j, (0, q^T)^T\}$ *vlastní pár matice* $\mathbf{B}_\tau$.

Důkaz: Tvrzení lemmatu plyne z toho, že $g \perp \mathcal{S}_j$ a $\mathbf{A}q = \lambda_j q$ je ekvivalentní zápisu

$$\begin{pmatrix} \tau & g^T \\ g & \mathbf{A} \end{pmatrix} \begin{pmatrix} 0 \\ q \end{pmatrix} = \lambda_j \begin{pmatrix} 0 \\ q \end{pmatrix}$$

□

Dále ukážeme, že v singulárním případě existuje jistá kritická hodnota $\tilde{\tau}_1$ taková, že pro $\tau > \tilde{\tau}_1$ mají všechny vlastní vektory odpovídající $-\xi_1(\tau)$ nulovou první složku. Ukážeme také, že pro každé τ existuje vždy vlastní vektor matice $\mathbf{B}_\tau$ (ne nutně odpovídající $-\xi_1(\tau)$), který může být normalizován tak, aby měl první složku rovnou jedné. Jestliže $g \not\perp \mathcal{S}_1$ nebo $g \perp \mathcal{S}_1$ a $\tau \le \tilde{\tau}_1$, pak tento vlastní vektor odpovídá $-\xi_1(\tau)$. Jestliže $g \perp \mathcal{S}_1$ a $\tau > \tilde{\tau}_1$, pak tento vlastní vektor odpovídá druhému nejmenšímu vlastnímu číslu $-\xi_2(\tau)$.

Nechť $g \perp \mathcal{S}_1$ a $\mathcal{Z}_1(\tau)$ je vlastní podprostor $\mathbf{B}_\tau$ odpovídající λ_1 (ovšem ne nutně odpovídající $-\xi_1(\tau)$). Pak podle lemmatu 2.16 je množina $\{(0, q^T) : q \in \mathcal{S}_1\}$ podmnožinou $\mathcal{Z}_1(\tau)$. Oba podprostory mají stejnou dimenzi pro všechna τ až na jediné. Následující lemma říká, že existuje právě jedno τ takové, že $\dim \mathcal{Z}_1(\tau) = \dim \mathcal{S}_1 + 1$.

Lemma 2.17 *Nechť* $1 \le j \le m$, $g \perp \mathcal{S}_j$ *a nechť* $p_j = -(\mathbf{A} - \lambda_j \mathbf{I})^\dagger g$. *Pak* $\{\lambda_j, (1, p_j^T)^T\}$ *je vlastní pár* $\mathbf{B}_\tau$ *právě když* $\tau = \tilde{\tau}_j = \lambda_j - g^T p_j$. *Kromě toho,* $\forall q \in S_j$ *platí*

$$(1, p_j^T)^T \perp (0, q^T)^T.$$

Důkaz: Nechť $\tau = \tilde{\tau}_j$. Pak

$$\begin{pmatrix} \tilde{\tau}_j & g^T \\ g & \mathbf{A} \end{pmatrix} \begin{pmatrix} 1 \\ p_j \end{pmatrix} = \begin{pmatrix} \tilde{\tau}_j + g^T p_j \\ g + \mathbf{A} p_j \end{pmatrix} = \lambda_j \begin{pmatrix} 1 \\ p_j \end{pmatrix},$$

neboť podle definice je $\tilde{\tau}_j + g^T p_j = \lambda_j$ a

$$(\mathbf{A} - \lambda_j \mathbf{I}) p_j = -(\mathbf{A} - \lambda_j \mathbf{I})(\mathbf{A} - \lambda_j \mathbf{I})^\dagger g = -g$$

podle poznámky A.2. Jestliže naopak je $\{\lambda_j, (1, p_j^T)^T\}$ vlastní pár $\mathbf{B}_\tau$, tzn.

$$\begin{pmatrix} \tau & g^T \\ g & \mathbf{A} \end{pmatrix} \begin{pmatrix} 1 \\ p_j \end{pmatrix} = \lambda_j \begin{pmatrix} 1 \\ p_j \end{pmatrix},$$

pak odtud plyne přímo $\tau = \lambda_j - g^T p_j = \tilde{\tau}_j$. Konečně, jestliže $q \in \mathcal{S}_j$, pak

$$\begin{aligned} q^T p_j &= -q^T (\mathbf{A} - \lambda_j \mathbf{I})^\dagger g = -q^T (\mathbf{A} - \lambda_j \mathbf{I})^\dagger (\mathbf{A} - \lambda_j \mathbf{I})(\mathbf{A} - \lambda_j \mathbf{I})^\dagger g = \\ &= q^T (\mathbf{A} - \lambda_j \mathbf{I})^\dagger (\mathbf{A} - \lambda_j \mathbf{I}) p_j = q^T (\mathbf{A} - \lambda_j \mathbf{I})(\mathbf{A} - \lambda_j \mathbf{I})^\dagger p_j = 0, \end{aligned}$$

protože $(\mathbf{A} - \lambda_j \mathbf{I}) q = 0$. □

Nyní uvedeme hlavní výsledek obou lemmat.

Věta 2.16 *Nechť*

$$1 \leq j \leq m, \quad g \perp S_j, \quad \mathcal{Z}_j(\tau) = \{z \in \mathbb{R}^{n+1} : \mathbf{B}_\tau z = \lambda_j z\} \quad a \quad p_j = -(\mathbf{A} - \lambda_j \mathbf{I})^\dagger g.$$

Jestliže $\tilde{\tau}_j = \lambda_j - g^T p_j$, *pak* $\dim \mathcal{Z}_j(\tilde{\tau}_j) = \dim \mathcal{S}_j + 1$ *a pro každou jinou hodnotu* τ *je* $\dim \mathcal{Z}_j(\tau) = \dim \mathcal{S}_j$. *Kromě toho, je-li* m_j *násobnost* λ_j *a* $\{q_1, \ldots, q_{m_j}\}$ *je ortonormální báze pro* $\mathcal{S}_j$, *pak*

$$\left\{ \begin{pmatrix} 1 \\ p_j \end{pmatrix}, \begin{pmatrix} 0 \\ q_1 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ q_{m_j} \end{pmatrix} \right\}, \quad \textit{resp.} \quad \left\{ \begin{pmatrix} 0 \\ q_1 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ q_{m_j} \end{pmatrix} \right\}$$

je ortogonální báze pro $\mathcal{Z}_j(\tilde{\tau}_j)$, *resp.* $\mathcal{Z}_j(\tau)$, $\tau \neq \tilde{\tau}_j$.

Důkaz: Protože $g \perp \mathcal{S}_j$, je podle lemmatu 2.16 $\{\lambda_j, (0, q^T)^T\}$ $\forall q \in \mathcal{S}_j$ vlastní pár matice $\mathbf{B}_\tau$, kde $\mathbf{A}q = \lambda_j q$. Nechť m_j je násobnost λ_j a označme $\{q_1, \ldots, q_{m_j}\}$ bázi $\mathcal{S}_j$. Lemma 2.17 říká, že navíc $\{\lambda_j, (1, p_j^T)^T\}$ je vlastní pár $\mathbf{B}_\tau$ právě když $\tau = \tilde{\tau}_j$ a $(1, p_j^T)^T \perp (0, q^T)^T$ $\forall q \in S_j$. Jestliže tedy $\tau \neq \tilde{\tau}_j$, platí $\dim \mathcal{Z}_j(\tau) = \dim \mathcal{S}_j = m_j$ a

$$\left\{ \begin{pmatrix} 0 \\ q_1 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ q_{m_j} \end{pmatrix} \right\}$$

je báze pro $\mathcal{Z}_j(\tau)$. Pokud $\tau = \tilde{\tau}_j$, platí $\dim \mathcal{Z}_j(\tilde{\tau}_j) = \dim \mathcal{S}_j + 1 = m_j + 1$ a

$$\left\{ \begin{pmatrix} 1 \\ p_j \end{pmatrix}, \begin{pmatrix} 0 \\ q_1 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ q_{m_j} \end{pmatrix} \right\}$$

je báze $\mathcal{Z}_j(\tilde{\tau}_j)$. □

Následující věta říká, že vždy existuje vlastní vektor matice $\mathbf{B}_\tau$, který můžeme normalizovat tak, aby měl první složku rovnou jedné a charakterizuje vlastní číslo, které tomuto vlastnímu vektoru odpovídá.

Věta 2.17 *Nechť* $\xi(\tau)$ *je největší řešení rovnice* $\phi(\xi) = \tau + \xi$ *pro libovolné* $\tau \in \mathbb{R}$. *Pak* $-\xi(\tau)$ *je vlastním číslem* $\mathbf{B}_\tau$ *a odpovídající vlastní vektor má první složku nenulovou.*

Důkaz: Nechť nejprve je g ortogonální na podprostory $\{\mathcal{S}_1, \ldots, \mathcal{S}_l : 1 \leq l < m.\}$ Pak z (2.82) plyne

$$\phi(\xi) = \sum_{j=1}^{m} \frac{\vartheta_j^2}{\lambda_j + \xi} = \sum_{j=l+1}^{m} \frac{\vartheta_j^2}{\lambda_j + \xi},$$

neboť $\vartheta_1 = \ldots = \vartheta_l = 0$. Nechť $\xi(\tau)$ je největší řešení rovnice $\phi(\xi) = \tau + \xi$. Podle (2.82) je funkce $\phi(\xi)$ ostře klesající na celém svém definičním oboru a funkce $\omega(\xi) \stackrel{\text{def}}{=} \tau + \xi$ je rostoucí přímka. Proto je $\xi(\tau) \in (-\lambda_{l+1}, \infty)$ jediné. Protože je $g \perp \{\mathcal{S}_1, \ldots, \mathcal{S}_l\}$, platí $g \in \mathcal{R}(\mathbf{A} + \xi\mathbf{I})$ pro $\xi \in (-\lambda_{l+1}, \infty)$. Odtud plyne $g \in \mathcal{R}(\mathbf{A} + \xi(\tau)\mathbf{I})$ a $(\mathbf{A} + \xi(\tau)\mathbf{I})p(\tau) = -g$ pro $p(\tau) = -(\mathbf{A} + \xi(\tau)\mathbf{I})^\dagger g$. Ukážeme, že $\{-\xi(\tau), (1, p(\tau))^T\}$ je vlastní pár $\mathbf{B}_\tau$. Jelikož

$$\tau + g^T p(\tau) = \tau - g^T(\mathbf{A} + \xi(\tau)\mathbf{I})^\dagger g = \tau - \phi(\xi(\tau)) = -\xi(\tau)$$

podle definice $\xi(\tau)$ a

$$g + \mathbf{A}p(\tau) = -(\mathbf{A} + \xi(\tau)\mathbf{I})p(\tau) + \mathbf{A}p(\tau) = -\xi(\tau)p(\tau),$$

pak

$$\begin{pmatrix} \tau & g^T \\ g & \mathbf{A} \end{pmatrix} \begin{pmatrix} 1 \\ p(\tau) \end{pmatrix} = \begin{pmatrix} \tau + g^T p(\tau) \\ g + \mathbf{A} p(\tau) \end{pmatrix} = \begin{pmatrix} -\xi(\tau) \\ -\xi(\tau) p(\tau) \end{pmatrix} = -\xi(\tau) \begin{pmatrix} 1 \\ p(\tau) \end{pmatrix}.$$

Jestliže g není ortogonální na $\mathcal{S}_1$, pak $\xi(\tau) \in (-\lambda_1, \infty)$ a $\mathbf{A} + \xi(\tau)\mathbf{I}$ je regulární matice. Předchozí důkaz tedy platí pro $p(\tau) = -(\mathbf{A} + \xi(\tau)\mathbf{I})^{-1} g$. □

Označme $-\xi_j(\tau)$, $j = 1, 2, \ldots, n+1$ vlastní čísla $\mathbf{B}_\tau$ v neklesajícím pořadí. Je-li g ortogonální na vlastní podprostory $\mathcal{S}_1, \ldots, \mathcal{S}_l$ odpovídající nejmenším l různým vlastním číslům $\lambda_1, \ldots, \lambda_l$, pak následující lemma 2.19 charakterizuje $l+1$ nejmenších vlastních čísel $\mathbf{B}_\tau$. Není-li g ortogonální na $\mathcal{S}_1$, pak toto lemma charakterizuje nejmenší vlastní číslo $\mathbf{B}_\tau$. Nejprve uvedeme jeden pomocný výsledek.

Lemma 2.18 *Nechť* $g \perp \{\mathcal{S}_i, \mathcal{S}_j\}$ *pro* $i < j$ *a nechť funkce* $\phi(\xi)$ *je definována pro všechna* $\xi \in \langle -\lambda_j, -\lambda_i \rangle$. *Pak platí* $\tilde{\tau}_i < \tilde{\tau}_j$.

DŮKAZ: Číslo $\tilde{\tau}_j$ s příslušným vektorem p_j je definováno v lemmatu 2.17, takže

$$\tilde{\tau}_j = \lambda_j - g^T p_j = \lambda_j + g^T (\mathbf{A} - \lambda_j \mathbf{I})^\dagger g = \lambda_j + \phi(-\lambda_j).$$

Protože funkce ϕ je monotonně klesající v $\langle -\lambda_j, -\lambda_i \rangle$ a $\lambda_i < \lambda_j$, platí $\tilde{\tau}_i < \tilde{\tau}_j$. □

Lemma 2.19 *Nechť* $\{-\xi(\tau), (1, p(\tau)^T)^T\}$ *je vlastní pár matice* $\mathbf{B}_\tau$ *daný větou 2.17, tedy* $\xi(\tau)$ *je největší řešení rovnice* $\phi(\xi) = \tau + \xi$, *a nechť*

$$\tilde{\tau}_j = \lambda_j - g^T p_j, \qquad kde \quad p_j = -(\mathbf{A} - \lambda_j \mathbf{I})^\dagger g.$$

Jestliže $g \not\perp \mathcal{S}_1$, *pak je* $-\xi_1(\tau) = -\xi(\tau)$, *neboli,* $-\xi(\tau)$ *je nejmenší vlastní číslo* $\mathbf{B}_\tau$. *Jestliže* $1 \le l < m$ *a* $g \perp \{\mathcal{S}_1, \ldots, \mathcal{S}_l\}$, *pak platí*

1. *Je-li* $\tau = \tilde{\tau}_j$, $j \in \{1, \ldots, l\}$, *pak* $-\xi_i(\tau) = \lambda_i$, $i = 1, \ldots, l$, *navíc* $-\xi_j(\tau) = -\xi(\tau)$ *a* $-\xi_{l+1}(\tau)$ *je druhý největší kořen rovnice* $\phi(\xi) = \tau + \xi$.

2. *Je-li* $\tau < \tilde{\tau}_1$, *pak* $-\xi_1(\tau) = -\xi(\tau)$ *a* $-\xi_i(\tau) = \lambda_{i-1}$, $i = 2, \ldots, l+1$.

3. *Je-li* $\tilde{\tau}_{j-1} < \tau < \tilde{\tau}_j$, $2 \le j \le l$, *pak* $-\xi_i(\tau) = \lambda_i$, $i = 1, \ldots, j-1$, $-\xi_j(\tau) = -\xi(\tau)$ *a* $-\xi_i(\tau) = \lambda_{i-1}$, $i = j+1, \ldots, l+1$.

4. *Je-li* $\tau > \tilde{\tau}_l$, *pak* $-\xi_i(\tau) = \lambda_i$, $i = 1, \ldots, l$ *a* $-\xi_{l+1}(\tau) = -\xi(\tau)$.

DŮKAZ: Je to přímý důsledek Cauchyho věty A.1, lemmat 2.16, 2.17, věty 2.17 a vlastností funkcí $\phi(\xi)$ a $\omega(\xi) \stackrel{\text{def}}{=} \tau + \xi$. Jestliže $g \not\perp \mathcal{S}_1$, pak $\phi(\xi)$ není definovaná pro $\xi = -\lambda_1$. Pro bod $\xi(\tau)$ z věty 2.17 tedy platí $\xi(\tau) > -\lambda_1$ a protože $-\xi(\tau)$ je vlastním číslem $\mathbf{B}_\tau$, je podle Cauchyho věty nejmenším vlastním číslem, tedy $-\xi(\tau) = -\xi_1(\tau)$. Nechť nyní $g \perp \{\mathcal{S}_1, \ldots, \mathcal{S}_l\}$. Funkce $\phi(\xi)$ je v tomto případě spojitá v $(-\lambda_{l+1}, \infty)$ a podle lemmatu 2.16 jsou $\lambda_1, \ldots, \lambda_l$ vlastní čísla $\mathbf{B}_\tau$, jejichž vlastní vektory mají nulovou první složku. Dále

1. Jestliže $\tau = \tilde{\tau}_j$, $j \in \{1, \ldots, l\}$, pak podle lemmatu 2.17 je λ_j vlastní číslo $\mathbf{B}_\tau$, jehož vlastní vektor má první složku nenulovou. Existují tedy dva vlastní vektory, které jsou na sebe kolmé, takže λ_j je dvojnásobné vlastní číslo $\mathbf{B}_\tau$. Navíc tento vlastní vektor je ten z věty 2.17, takže platí $-\xi(\tau) = \lambda_j$. Podle Cauchyho věty má $\mathbf{B}_\tau$ další vlastní číslo $-\xi_{l+1}(\tau)$ větší než λ_l, kde $\xi_{l+1}(\tau)$ je druhý největší kořen rovnice $\phi(\xi) = \tau + \xi$.

2. Jelikož pro $\tau = \tilde{\tau}_1$ má rovnice $\phi(\xi) = \tau + \xi$ řešení $\xi(\tau) = -\lambda_1$, bod 1., pak pro $\tau < \tilde{\tau}_1$ je $\xi(\tau) > -\lambda_1$. Platí tedy $-\xi_1(\tau) = -\xi(\tau)$, $-\xi_2(\tau) = \lambda_1, \ldots, -\xi_{l+1}(\tau) = \lambda_l$.

3. Jelikož bodu $\tau = \tilde{\tau}_{j-1}$ odpovídá řešení $\xi(\tau) = -\lambda_{j-1}$ a bodu $\tau = \tilde{\tau}_j$ odpovídá řešení $\xi(\tau) = -\lambda_j$, pak pro $\tilde{\tau}_{j-1} < \tau < \tilde{\tau}_j$ odpovídá řešení $-\lambda_j < \xi(\tau) < -\lambda_{j-1}$. Navíc $-\xi(\tau)$ je vlastní číslo $\mathbf{B}_\tau$, takže $-\xi_i(\tau) = \lambda_i$, $i = 1, \ldots, j-1$, $-\xi_j(\tau) = -\xi(\tau)$ a $-\xi_i(\tau) = \lambda_{i-1}$, $i = j+1, \ldots, l+1$.

4. Obdobně jako v bodu 2. pro $\xi(\tau)$ platí $\xi(\tau) < -\lambda_l$ a proto $-\xi_i(\tau) = \lambda_i$, $i = 1, \ldots, l$ a $-\xi_{l+1}(\tau) = -\xi(\tau)$.

Ve všech uvedených případech z Cauchyho věty plyne, že další vlastní číslo $\mathbf{B}_\tau$ splňuje $-\xi_{l+2}(\tau) > \lambda_l$. $\square$

Nechť například pro $n = 4$ je $l = 2$, tedy $g \perp \{\mathcal{S}_1, \mathcal{S}_2\}$ a λ_1, λ_2 jsou jednonásobná vlastní čísla matice $\mathbf{A}$ s vlastními vektory q_1, q_2. Pak podle lemmatu 2.16 jsou $\{\lambda_1, (0, q_1^T)^T\}$ a $\{\lambda_2, (0, q_2^T)^T\}$ vlastní páry matice $\mathbf{B}_\tau$ $\forall \tau \in \mathbb{R}$. Navíc, pokud $\tau = \tilde{\tau}_j$, $j \in \{1, 2\}$, je podle lemmatu 2.17 $\{\lambda_j, (1, p_j^T)^T\}$ pro $p_j = -(\mathbf{A} - \lambda_j \mathbf{I})^\dagger g$ také vlastním párem matice $\mathbf{B}_\tau$ a platí $(1, p_j^T)^T \perp (0, q_j^T)^T$. Dále podle věty 2.17 je $-\xi(\tau)$, kde $\xi(\tau)$ je největší řešení rovnice $\phi(\xi) = \tau + \xi$, vlastním číslem matice $\mathbf{B}(\tau)$, jehož vlastní vektor má první složku nenulovou. Pro $\tau = \tilde{\tau}_j$, $j \in \{1, 2\}$, je tedy $\xi(\tau) = -\lambda_j$. Na obrázku 2.4 je uvedeno všech pět možností výskytu parametru τ a řešení $\xi(\tau)$. Tlustá křivka je funkce $\phi(\xi)$ a šikmé přímky jsou funkce $\omega(\xi) = \tau + \xi$ pro různá τ :

- přímka 1 odpovídá případu, kdy $\tau < \tilde{\tau}_1$ a tedy $\xi(\tau) > -\lambda_1$,
- přímka 2 odpovídá případu, kdy $\tau = \tilde{\tau}_1$ a tedy $\xi(\tau) = -\lambda_1$,
- přímka 3 odpovídá případu, kdy $\tilde{\tau}_1 < \tau < \tilde{\tau}_2$ a tedy $-\lambda_2 < \xi(\tau) < -\lambda_1$,
- přímka 4 odpovídá případu, kdy $\tau = \tilde{\tau}_2$ a tedy $\xi(\tau) = -\lambda_2$,
- přímka 5 odpovídá případu, kdy $\tau > \tilde{\tau}_2$ a tedy $\xi(\tau) < -\lambda_2$.

Na obrázku 2.5 je znázorněno rozložení vlastních čísel matic $\mathbf{A}$ a $\mathbf{B}(\tau)$. Zajímá nás vlastní číslo $-\xi(\tau)$, protože odpovídající vlastní vektor lze normalizovat tak, aby měl první složku rovnou jedné.

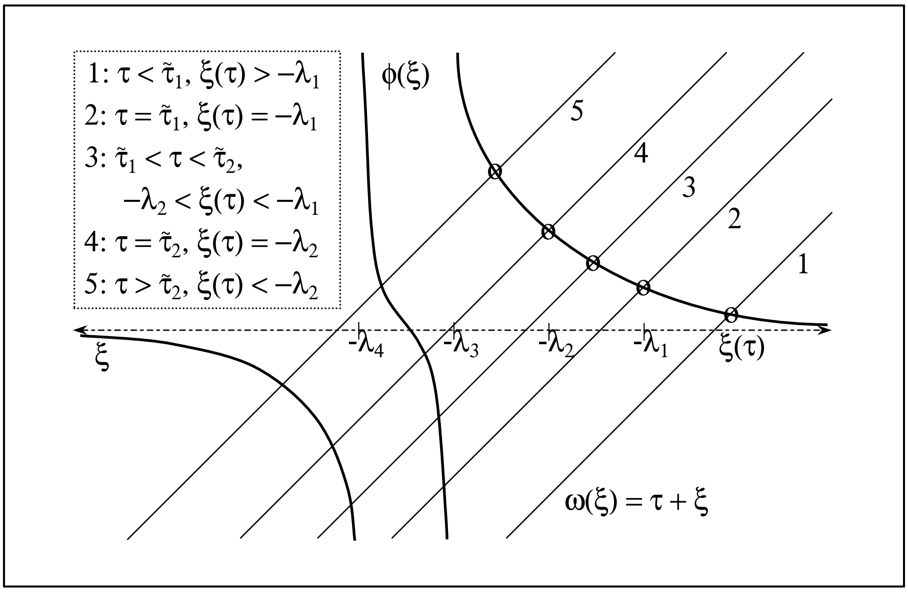
Matice $\mathbf{B}(\tau)$ má tedy v tomto případě ($g \perp \{\mathcal{S}_1, \mathcal{S}_2\}$) první tři nejmenší vlastní čísla λ_1, λ_2 a $-\xi(\tau)$. Toto $-\xi(\tau)$ je mezi těmito třemi jediné, jehož vlastní vektor má první složku nenulovou. K tomu, abychom měli jistotu, že dostaneme vlastní vektor, který lze normalizovat, bychom měli spočítat tři (obecně $l+1$) nejmenší vlastní čísla nebo zmenšit parametr τ tak, aby jeho hodnota byla menší než $\tilde{\tau}_1$. Pak stačí pouze jedno nejmenší vlastní číslo.

Příklad 2.2 *Nechť je dána matice*

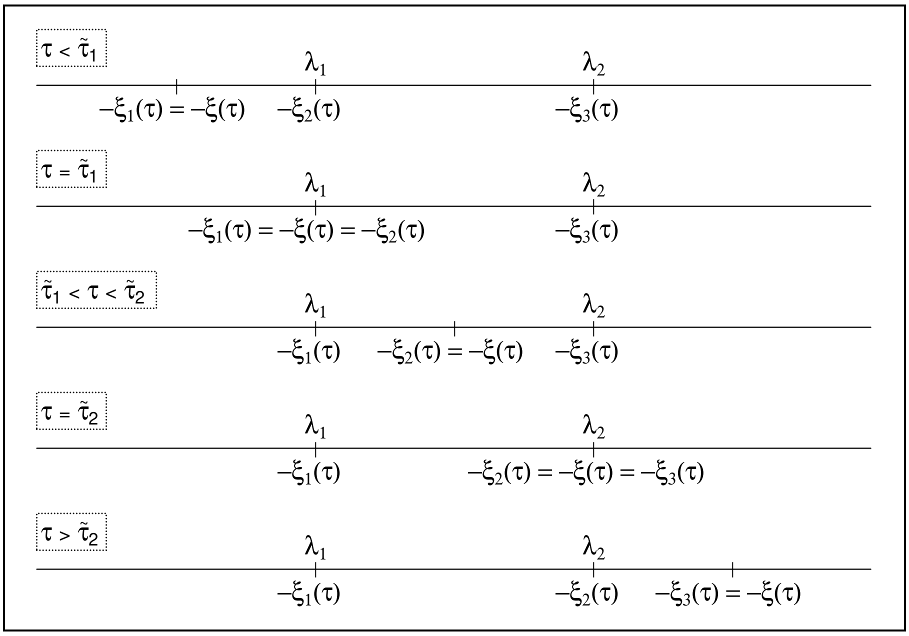
$$\mathbf{A} = \begin{pmatrix} 3 & 2 \\ 2 & 0 \end{pmatrix},$$

která má vlastní čísla a odpovídající ortonormální vlastní vektory

$$\lambda_1 = -1, \quad q_1 = \frac{1}{\sqrt{5}}(-1, 2)^T; \qquad \lambda_2 = 4, \quad q_2 = \frac{1}{\sqrt{5}}(2, 1)^T,$$



Obrázek 2.4: Řešení $\xi(\tau)$ rovnice $\phi(\xi) = \tau + \xi$



Obrázek 2.5: Rozložení vlastních čísel matic $\mathbf{A}$ a $\mathbf{B}_\tau$

a nechť je dán vektor

$$g = (2,1)^T.$$

Protože je g kolmý na q_1, nastává singulární případ. Podívejme se, jak vypadají vlastní čísla $\tilde{\xi}_i \equiv -\xi_i,\ i = 1,2,3$ (bez uspořádání) a vlastní vektory matice $\mathbf{B}_\tau$. Podle (2.76) je

$$\mathbf{B}_\tau = \begin{pmatrix} \tau & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 0 \end{pmatrix}$$

a charakteristický polynom této matice má tvar

$$\tilde{\xi}^3 - (\tau+3)\,\tilde{\xi}^2 + (3\tau - 9)\,\tilde{\xi} + 4\tau - 5 = 0.$$

Tato rovnice má jeden kořen $\tilde{\xi}_1 = -1$ a to pro libovolné τ, takže $-\xi_1 = -1$ je vlastním číslem $\mathbf{B}_\tau$. Další kořeny jsou (po vydělení členem $\tilde{\xi} + 1$) řešením rovnice

$$\tilde{\xi}^2 - (\tau+4)\,\tilde{\xi} + 4\tau - 5 = 0$$

a tedy

$$-\xi_2 \equiv \tilde{\xi}_2 = \frac{\tau + 4 - \sqrt{\tau^2 - 8\tau + 36}}{2}, \quad -\xi_3 \equiv \tilde{\xi}_3 = \frac{\tau + 4 + \sqrt{\tau^2 - 8\tau + 36}}{2}.$$

Protože platí

$$\tau^2 - 8\tau + 36 > 0 \quad \forall \tau,$$

je $-\xi_2 < -\xi_3$. Dále zjistíme, že $\forall\tau$ platí $-\xi_2 < 4$ a $-\xi_3 > 4$, což odpovídá Cauchyho větě, takže 4 není vlastním číslem $\mathbf{B}_\tau$.

Vraťme se k vlastnímu číslu $-\xi_1 = -1$ a najděme příslušný jednotkový vlastní vektor $(a,b,c)^T$. Platí

$$\begin{pmatrix} \tau & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = - \begin{pmatrix} a \\ b \\ c \end{pmatrix} \quad \Rightarrow \quad \left. \begin{array}{r} (\tau+1)a + 2b + c = 0 \\ a + 2b + c = 0 \end{array} \right\} \quad \Rightarrow \quad \tau a = 0.$$

Je-li $\tau \neq 0$, platí $a = 0$ a vlastní vektor příslušný vlastnímu číslu -1 má tvar

$$\frac{1}{\sqrt{5}\,b}\,(0, b, -2b)^T.$$

Podle lemmatu (2.16) je opravdu $\{\lambda_1, (0, q_1^T)^T\}$ vlastní pár matice $\mathbf{B}_\tau$. Je-li $\tau = 0$, pak navíc platí $-\xi_2 = -1$ a $-\xi_3 = 5$, tedy -1 je dvojnásobné vlastní číslo, jehož druhý vlastní vektor je kolmý na vektor $(0, q_1^T)^T$.

Analyzujme nyní lemma 2.17. Moore-Penroseova pseudoinverze $(\mathbf{A} - \lambda_1 \mathbf{I})^\dagger$ singulární matice

$$\mathbf{A} - \lambda_1 \mathbf{I} = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix},$$

pro kterou platí podmínky uvedené v poznámce A.1, má tvar

$$(\mathbf{A} - \lambda_1 \mathbf{I})^\dagger = \frac{1}{25}\,(\mathbf{A} - \lambda_1 \mathbf{I}).$$

Pro vektor p_1, který splňuje $(\mathbf{A} - \lambda_1 \mathbf{I})p_1 = -g$, platí

$$p_1 = -(\mathbf{A} - \lambda_1 \mathbf{I})^\dagger g = -\frac{1}{25}\,(\mathbf{A} - \lambda_1 \mathbf{I})g = -\frac{1}{5}\,(2,1)^T.$$

Tento vektor je skutečně kolmý na vektor q_1 a dále

$$\tilde{\tau}_1 = \lambda_1 - g^T p_1 = -1 + \frac{1}{5}(2,1)(2,1)^T = 0.$$

Takže pro $\tau = \tilde{\tau}_1 = 0$ má matice $\mathbf{B}_\tau$ dvojnásobné vlastní číslo $-\xi_1 = \lambda_1 = -1$ a dá se ověřit, že $(1, p_1^T)^T$ je druhý příslušný vlastní vektor.

Přejdeme dále k větě 2.17. Z tvarů vektorů g, q_1, q_2 a toho, že $g \perp q_1$ plyne

$$g = \vartheta_1 q_1 + \vartheta_2 q_2 \quad \Rightarrow \quad \vartheta_1 = 0, \quad \vartheta_2 = \sqrt{5}$$

a tudíž podle (2.82)

$$\phi(\xi) = \frac{5}{4+\xi},$$

takže řešení $\xi(\tau)$ rovnice $\phi(\xi) = \tau + \xi$ splňuje

$$\frac{5}{4+\xi} = \tau + \xi \quad \Rightarrow \quad \xi^2 + (\tau + 4)\,\xi + 4\tau - 5 = 0,$$

což je ale až na znaménko u lineárního členu charakteristický polynom $\mathbf{B}_\tau$ vydělený členem $\xi + 1$. Proto je $-\xi(\tau)$ vlastním číslem matice $\mathbf{B}_\tau$. Pro $\tau = 0$ je $-\xi(\tau) = -1 = \lambda_1$ a již víme, že příslušný vlastní vektor je $(1, p_1^T)^T$, má tedy první složku nenulovou. Zbývá ukázat rozdělení vlastních čísel podle lemmatu 2.19. Dá se ukázat, že platí

$$\begin{aligned} \tau < 0 \quad &\Rightarrow \quad -\xi_2 < -1, \quad -\xi_3 < 5; \\ \tau > 0 \quad &\Rightarrow \quad -\xi_2 > -1, \quad -\xi_3 > 5, \end{aligned}$$

takže celkem dostáváme hodnoty uvedené v tabulce 2.2, kde $-\xi_1 \leq -\xi_2 < -\xi_3$:

τ	$-\xi_1$	$-\xi_2$	$-\xi_3$
< 0	< -1	-1	$\in (4,5)$
0	-1	-1	5
> 0	-1	$\in (-1,4)$	> 5

Tabulka 2.2: Vlastní čísla matice $\mathbf{B}_\tau$ pro příklad 2.2.

Jestliže tedy máme matici $\mathbf{B}_\tau$, spočítáme její, pokud možno nejmenší, vlastní číslo $-\xi(\tau)$ takové, že příslušný vlastní vektor lze normalizovat na tvar $(1, x^T)^T$. Tím získáme obecně k-tou iteraci $\xi_k = -\xi(\tau)$ a $x_k = x$.

Další lemma ukazuje, jak lze spočítat téměř optimální řešení problému (1.25) v singulárním případě.

Lemma 2.20 *Nechť $g \perp \mathcal{S}_1$ a $p = -(\mathbf{A} - \lambda_1 \mathbf{I})^\dagger g$. Jestliže $\lambda_1 \leq 0$ a $\|p\| \leq \Delta$, pak se řešení (1.25) skládají z množiny $\{x : x = p + q,\ q \in \mathcal{S}_1,\ \|x\| = \Delta\}$ s $\xi_\star = -\lambda_1$.*

Důkaz: Napíšeme řešení ve tvaru $x_\star = p + \kappa q$ a ukážeme, že splňuje vztahy (2.75). Předně platí $\xi_\star = -\lambda_1 \geq 0$ a $\mathbf{A} + \xi_\star \mathbf{I} \succeq 0$. Číslo κ určíme tak, aby $\|x_\star\| = \Delta$. Konečně platí

$$(\mathbf{A} + \xi_\star \mathbf{I}) x_\star = (\mathbf{A} - \lambda_1 \mathbf{I})(p + \kappa q) = -(\mathbf{A} - \lambda_1 \mathbf{I})(\mathbf{A} - \lambda_1 \mathbf{I})^\dagger g + \kappa(\mathbf{A} - \lambda_1 \mathbf{I}) q = -g.$$

První sčítanec je roven $-g$ podle poznámky A.2, protože $g \perp \mathcal{S}_1$ a druhý sčítanec je nula, protože $\{\lambda_1, q\}$ je vlastní pár matice $\mathbf{A}$. $\square$

Jak zjistíme, že nastal singulární případ? Nechť $(\nu, u^T)^T$ je jednotkový vlastní vektor matice $\mathbf{B}_\tau$ odpovídající nejmenšímu vlastnímu číslu $-\xi_1(\tau)$. Tedy

$$\begin{pmatrix} \tau & g^T \\ g & \mathbf{A} \end{pmatrix} \begin{pmatrix} \nu \\ u \end{pmatrix} = -\xi_1(\tau) \begin{pmatrix} \nu \\ u \end{pmatrix} \quad \Rightarrow \quad (\mathbf{A} + \xi_1(\tau)\mathbf{I})\, u = -\nu g,$$

odtud

$$\frac{\|(\mathbf{A} + \xi_1(\tau)\mathbf{I})\, u\|}{\|u\|} = \frac{|\nu|\|g\|}{\sqrt{1-\nu^2}},$$

neboť $\nu^2 + \|u\|^2 = 1$. Platí proto implikace

$$|\nu|\|g\| \le \varepsilon\sqrt{1-\nu^2} \quad \Rightarrow \quad \|(\mathbf{A} + \xi_1(\tau)\mathbf{I})\, u\| \le \varepsilon\|u\|.$$

Zjišťujeme, že $\{-\xi_1(\tau), u\}$ aproximuje vlastní pár matice $\mathbf{A}$ a podle lemmatu 2.16 je $g \perp \mathcal{S}_1$.

Pro singulární případ potřebujeme alternativní způsob, jak definovat k-tou iteraci. Abychom definovali bod $\{\xi_k, x_k\}$, spočítáme dva nejmenší vlastní páry matice $\mathbf{B}_{\tau_k}$:

$$\{-\xi_1(\tau_k), (\nu_1, u_1^T)^T\} \quad \text{a} \quad \{-\xi_2(\tau_k), (\nu_2, u_2^T)^T\}.$$

Jsou-li obě $|\nu_1|$ i $|\nu_2|$ malá, tj. jestliže

$$|\nu_j|\|g\| \le \varepsilon\sqrt{1-\nu_j^2}, \quad j = 1, 2,$$

pak zmenšíme parametr τ_k, abychom se přiblížili kritické hodnotě $\tilde{\tau}_1$, při které existuje vlastní vektor matice $\mathbf{B}_{\tau_k}$, který má první složku nenulovou (lemma 2.19). Tento vlastní vektor bude odpovídat buď prvnímu ($\tau_k \le \tilde{\tau}_1$) nebo druhému ($\tilde{\tau}_1 < \tau_k \le \tilde{\tau}_2$) nejmenšímu vlastnímu číslu matice $\mathbf{B}_{\tau_k}$. Připomínáme, že případ $\nu_1 = \nu_2 = 0$ může nastat jen když je $g \perp \{\mathcal{S}_1, \mathcal{S}_2\}$. Bude-li tedy τ_k blízko jedné z kritických hodnot, bude platit buď

$$|\nu_1|\|g\| > \varepsilon\sqrt{1-\nu_1^2} \quad \text{nebo} \quad |\nu_2|\|g\| > \varepsilon\sqrt{1-\nu_2^2}.$$

Tedy po možné redukci parametru τ_k lze iteraci $\{\xi_k, x_k\}$ definovat takto:
Jestliže $|\nu_1|\|g\| \le \varepsilon\sqrt{1-\nu_1^2}$, položíme $\xi_k = -\xi_2(\tau_k)$ a $x_k = \frac{1}{\nu_2} u_2$, jinak položíme $\xi_k = -\xi_1(\tau_k)$ a $x_k = \frac{1}{\nu_1} u_1$.

2.7.3 Quasi-optimální řešení

Následující věta stanovuje, že za jistých podmínek dává posledních n složek speciální lineární kombinace vlastních vektorů $\mathbf{B}_\tau$ téměř optimální řešení pro (1.25). Lemma 2.21 poskytuje podmínky, za jakých lze spočítat tuto lineární kombinaci a lemma 2.22 ukazuje, jak ji spočítat.

Věta 2.18 *Nechť $-\xi_1(\tau)$ je nejmenší vlastní číslo $\mathbf{B}_\tau$ s odpovídajícím vlastním vektorem $v_1 = (\nu_1, u_1^T)^T$. Nechť $-\xi_i(\tau)$ je jakékoli ze zbývajících n vlastních čísel $\mathbf{B}_\tau$ s odpovídajícím vlastním vektorem $v_i = (\nu_i, u_i^T)^T$. Definujme matice*

$$\mathbf{V} = (v_1, v_i) \in \mathbb{R}^{(n+1)\times 2}, \quad \mathbf{U} = (u_1, u_i) \in \mathbb{R}^{n\times 2}$$

a předpokládejme $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, neboli, že v_1 a v_i jsou ortonormální. Nechť $\varepsilon > 0$. Jestliže existuje vektor $y = (y_1, y_2)^T \in \mathbb{R}^2$, $\|y\| = 1$ tak, že

1. $(e_1^T \mathbf{V} y)^2 = \frac{1}{1+\Delta^2}$,

2. $[\xi_1(\tau) - \xi_i(\tau)]\, y_2^2\, (1+\Delta^2) \le -2\,\varepsilon\, \psi(\tilde{x})$ *pro* $\tilde{x} = \frac{1}{e_1^T \mathbf{V} y}\, \mathbf{U} y$,

pak $\psi(x_\star) \le \psi(\tilde{x}) \le \frac{1}{1+\varepsilon}\, \psi(x_\star)$, *kde* $x_\star$ *je optimální krok problému (1.25), ve smyslu definice 1.10, na hranici s* $\psi(x_\star) \le 0$.

DŮKAZ: Protože $x_\star$ je řešení na hranici, platí $\psi(x_\star) \le \psi(x)\ \forall x \in \mathbb{R}^n$ taková, že $\|x\| = \Delta$. K tomu, aby $\psi(x_\star) \le \psi(\tilde{x})$, stačí ukázat, že $\|\tilde{x}\| = \Delta$. Takže

$$\mathbf{V} y = \begin{pmatrix} \nu_1 & \nu_i \\ u_1 & u_i \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \nu_1 y_1 + \nu_i y_2 \\ \mathbf{U} y \end{pmatrix}; \quad e_1^T \mathbf{V} y = \nu_1 y_1 + \nu_i y_2$$

$$\text{(2.83)} \qquad \Rightarrow \quad \frac{1}{e_1^T \mathbf{V} y}\, \mathbf{V} y = \left(\frac{\nu_1 y_1 + \nu_i y_2}{\nu_1 y_1 + \nu_i y_2},\ \frac{1}{e_1^T \mathbf{V} y}\, (\mathbf{U} y)^T \right)^T = (1, \tilde{x}^T)^T.$$

Protože

$$\|\mathbf{V} y\|^2 = y^T \mathbf{V}^T \mathbf{V} y = y^T y = 1,$$

pak podle předpokladu 1. máme

$$1 + \|\tilde{x}\|^2 = \|(1, \tilde{x}^T)\|^2 = \frac{1}{(e_1^T \mathbf{V} y)^2}\, \|\mathbf{V} y\|^2 = \frac{1}{(e_1^T \mathbf{V} y)^2} = 1 + \Delta^2 \quad \Rightarrow \quad \|\tilde{x}\| = \Delta.$$

Nyní dokážeme druhou část nerovnosti. Protože $\|x_\star\| = \Delta$, pak $\|(1, x_\star^T)\|^2 = 1 + \Delta^2$ a z (2.77) plyne

$$\text{(2.84)} \qquad \tau + 2\psi(x_\star) = (1, x_\star^T) \mathbf{B}_\tau (1, x_\star^T)^T \ge -\xi_1(\tau) \|(1, x_\star^T)\|^2 = -\xi_1(\tau)(1+\Delta^2).$$

Nyní z $\mathbf{B}_\tau v_j = -\xi_j(\tau) v_j$, $j = 1, i$, kde $\mathbf{V}^T \mathbf{V} = \mathbf{I}$, z předpokladu 1. a vztahu (2.83) plyne

$$\begin{aligned} \tau + 2\psi(\tilde{x}) &= (1, \tilde{x}^T) \mathbf{B}_\tau (1, \tilde{x}^T)^T = \frac{1}{(e_1^T \mathbf{V} y)^2}\, y^T \mathbf{V}^T \mathbf{B}_\tau \mathbf{V} y = \\ \text{(2.85)} \qquad &= (1+\Delta^2)(y_1 v_1 + y_2 v_i)^T \mathbf{B}_\tau (y_1 v_1 + y_2 v_i) = -(1+\Delta^2)\left(\xi_1(\tau) y_1^2 + \xi_i(\tau) y_2^2\right). \end{aligned}$$

Dále $\|y\| = 1$ implikuje $y_1^2 = 1 - y_2^2$ a tudíž podle (2.84)

$$\begin{aligned} \tau + 2\psi(\tilde{x}) &= -(1+\Delta^2)\left(\xi_1(\tau) + [\xi_i(\tau) - \xi_1(\tau)]\, y_2^2\right) \le \\ &\le \tau + 2\psi(x_\star) + (1+\Delta^2)[\xi_1(\tau) - \xi_i(\tau)]\, y_2^2. \end{aligned}$$

Jestliže je splněn předpoklad 2., pak

$$\tau + 2\psi(\tilde{x}) + 2\varepsilon\psi(\tilde{x}) \le \tau + 2\psi(x_\star)$$

a odtud již plyne druhá nerovnost tvrzení věty. □

Z této věty přímo plyne, že

$$\text{(2.86)} \qquad 0 \le \psi(\tilde{x}) - \psi(x_\star) \le \frac{\varepsilon}{1+\varepsilon} |\, \psi(x_\star)|,$$

což implikuje, že za podmínek věty 2.18 bude $\psi(\tilde{x})$ libovolně blízko hodnotě $\psi(x_\star)$. Takové $\tilde{x}$ nazveme quasi-optimální řešení pro problém (1.25). Protože uvažujeme dvě nejmenší vlastní čísla matice $\mathbf{B}_\tau$, položíme $i = 2$.

Poznámka 2.3 *Pro $\tilde{x}$ platí:*

$$\mathbf{U}y = (u_1, u_2)\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = y_1u_1 + y_2u_2 \quad \Rightarrow \quad \tilde{x} = \frac{1}{\nu_1 y_1 + \nu_2 y_2}\,(y_1u_1 + y_2u_2).$$

Nyní následují podmínky pro výpočet vektoru y ve větě 2.18.

Lemma 2.21 *Nechť $v_j = (\nu_j, u_j^T)^T$, kde $\nu_j \in \mathbb{R}$, $u_j \in \mathbb{R}^n$ pro $j = 1, 2$. Definujme matice $\mathbf{V} = (v_1, v_2)$ a $\mathbf{U} = (u_1, u_2)$ a předpokládejme, že $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. Jestliže existuje $\beta > 0$ takové, že $\|\mathbf{V}^T e_1\|^2 \geq \frac{1}{\beta}$, pak existuje vektor $y \in \mathbb{R}^2$, $y \neq 0$, který splňuje*

$$\|\mathbf{V}y\|^2 = \beta(e_1^T\mathbf{V}y)^2. \tag{2.87}$$

DŮKAZ: Podle vztahu (2.87) platí

$$y^Ty = y^T\mathbf{V}^T\mathbf{V}y = \|\mathbf{V}y\|^2 = \beta(e_1^T\mathbf{V}y)^2 = \beta(y^T\mathbf{V}^Te_1e_1^T\mathbf{V}y),$$

což je ekvivalentní výrazu

$$y^T(\mathbf{I} - \beta\mathbf{V}^Te_1e_1^T\mathbf{V})y = 0. \tag{2.88}$$

Rovnice (2.88) má netriviální řešení jen když je matice $\mathbf{M} = \mathbf{I} - \beta\mathbf{V}^Te_1e_1^T\mathbf{V} \in \mathbb{R}^{2\times 2}$ indefinitní nebo singulární. Budeme tedy studovat vlastní čísla matice $\mathbf{M}$ a výpočtem zjistíme, že oba vlastní páry jsou dány takto:

$$\{1 - \beta e_1^T\mathbf{V}\mathbf{V}^Te_1\,,\, \mathbf{V}^Te_1\} \quad \text{a} \quad \{1, w\}, \quad \text{kde} \quad w \perp \mathbf{V}^Te_1 \quad \text{je libovolný vektor.}$$

Matice $\mathbf{M}$ je tedy indefinitní nebo singulární, jestliže

$$1 - \beta e_1^T\mathbf{V}\mathbf{V}^Te_1 \leq 0 \quad \Leftrightarrow \quad e_1^T\mathbf{V}\mathbf{V}^Te_1 = \|\mathbf{V}^Te_1\|^2 \geq \frac{1}{\beta}$$

a za tohoto předpokladu existuje vektor $y \in \mathbb{R}^2$, $y \neq 0$, který splňuje (2.87). □

Poznámka 2.4 *Platí*

$$\mathbf{V}^Te_1 = \begin{pmatrix} \nu_1 & u_1^T \\ \nu_2 & u_2^T \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = (\nu_1, \nu_2)^T \quad \Rightarrow \quad \|\mathbf{V}^Te_1\|^2 = \nu_1^2 + \nu_2^2.$$

Je-li splněn předpoklad lemmatu 2.21 pro $\beta = 1 + \Delta^2$, tzn. $\nu_1^2 + \nu_2^2 \geq \frac{1}{1+\Delta^2}$, pak existuje vektor y, $\|y\| = 1$, že $\|\mathbf{V}y\|^2 = (1 + \Delta^2)(e_1^T\mathbf{V}y)^2$. Protože však pro $\|y\| = 1$ platí $\|\mathbf{V}y\|^2 = 1$, je splněna podmínka 1. věty 2.18. K tomu, aby $\tilde{x}$ bylo quasi-optimální řešení problému (1.25), stačí ověřit podmínku 2. Následující lemma poskytuje způsob výpočtu vektoru y (používáme označení $s = \mathbf{V}^Te_1$).

Lemma 2.22 *Nechť $\beta > 0$, $s \in \mathbb{R}^2$ a $\|s\|^2 \geq \frac{1}{\beta}$. Rovnice*

$$y^T\left(\mathbf{I} - \beta ss^T\right)y = 0 \tag{2.89}$$

v proměnné $y \in \mathbb{R}^2$ má dvě netriviální řešení, je-li matice $\mathbf{M} = \mathbf{I} - \beta ss^T$ indefinitní a jedno netriviální řešení, je-li $\mathbf{M}$ singulární.

DŮKAZ: Nechť $\mathbf{P} \in \mathbb{R}^{2\times 2}$ je taková, že

$$\mathbf{P}^T s = \|s\|\, e_1 \quad \text{a} \quad \mathbf{P}^T\mathbf{P} = \mathbf{I}. \tag{2.90}$$

Aplikujme tuto ortogonální transformaci na matici $\mathbf{M}$:

$$\mathbf{P}^T\mathbf{M}\mathbf{P} = \mathbf{P}^T\left(\mathbf{I} - \beta s s^T\right)\mathbf{P} = \mathbf{I} - \beta\|s\|^2 e_1 e_1^T.$$

Položíme-li $w = \mathbf{P}^T y$, lze rovnici (2.89) přepsat tímto způsobem:

$$y^T\left(\mathbf{I} - \beta s s^T\right) y = w^T\mathbf{P}^T\left(\mathbf{I} - \beta s s^T\right)\mathbf{P}w = w^T\left(\mathbf{I} - \beta\|s\|^2 e_1 e_1^T\right) w = w^T\begin{pmatrix} -\varrho & 0 \\ 0 & 1 \end{pmatrix} w = 0,$$

kde $-\varrho = 1 - \beta\|s\|^2$. Netriviální řešení rovnice (2.89) jsou rovny $y = \mathbf{P}w$, kde w je dáno takto:

1. Je-li matice $\mathbf{M}$ indefinitní, tzn. je-li $\varrho > 0$, pak

$$0 = w^T\begin{pmatrix} -\varrho & 0 \\ 0 & 1 \end{pmatrix} w = -\varrho w_1^2 + w_2^2.$$

   Řešením této rovnice je $w_1 = 1$ a $w_2 = \pm\sqrt{\varrho}$, tedy řešením (2.89) jsou dva vektory

$$w = (1, \pm\sqrt{\varrho})^T.$$

2. Je-li matice $\mathbf{M}$ singulární, tzn. je-li $\varrho = 0$, pak

$$0 = w^T\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} w = w_2^2.$$

   Řešením je $w_1 = 1$ a $w_2 = 0$, tudíž řešením (2.89) je jediný vektor $w = e_1$. $\square$

Poznámka 2.5 *Jestliže* $s = \mathbf{V}^T e_1 = (\nu_1, \nu_2)^T \in \mathbb{R}^2$, *tedy* $\|s\|^2 = \nu_1^2 + \nu_2^2$, *a* $\beta = 1 + \Delta^2$, *pak z výše uvedených lemmat spočítáme vektor* y. *Matice* $\mathbf{P}$, *která splňuje (2.90), má tvar*

$$\mathbf{P} = \frac{1}{\sqrt{\nu_1^2 + \nu_2^2}}\begin{pmatrix} \nu_1 & \pm\nu_2 \\ \nu_2 & \mp\nu_1 \end{pmatrix}$$

a dále

1. *Je-li*

$$\varrho = (1 + \Delta^2)(\nu_1^2 + \nu_2^2) - 1 > 0 \quad \Rightarrow \quad \nu_1^2 + \nu_2^2 > \frac{1}{1 + \Delta^2},$$

   je matice $\mathbf{M} = \mathbf{I} - \beta s s^T$ *indefinitní, vektor*

$$w = \left(1, \pm\sqrt{(1 + \Delta^2)(\nu_1^2 + \nu_2^2) - 1}\right)^T,$$

   takže rovnice (2.89) má 2 řešení $y = (y_1, y_2)^T = \mathbf{P}w$, *pro která platí*

$$\|y\| = \sqrt{y^T y} = \sqrt{w^T\mathbf{P}^T\mathbf{P}w} = \sqrt{w^T w} = \sqrt{(1 + \Delta^2)(\nu_1^2 + \nu_2^2)},$$

   takže po vydělení této normy dostaneme

$$y_1 = \frac{\nu_1 \pm \nu_2\sqrt{(1 + \Delta^2)(\nu_1^2 + \nu_2^2) - 1}}{(\nu_1^2 + \nu_2^2)\sqrt{1 + \Delta^2}}, \quad y_2 = \frac{\nu_2 \mp \nu_1\sqrt{(1 + \Delta^2)(\nu_1^2 + \nu_2^2) - 1}}{(\nu_1^2 + \nu_2^2)\sqrt{1 + \Delta^2}}.$$

2. Pokud

$$\varrho = 0 \quad \Rightarrow \quad \nu_1^2 + \nu_2^2 = \frac{1}{1+\Delta^2},$$

je matice $\mathbf{M} = \mathbf{I} - \beta s s^T$ *singulární, vektor*

$$w = (1,0)^T$$

a rovnice (2.89) má 1 řešení, kde y má jednodušší tvar

$$y_1 = \frac{\nu_1}{\sqrt{\nu_1^2 + \nu_2^2}}, \quad y_2 = \frac{\nu_2}{\sqrt{\nu_1^2 + \nu_2^2}}.$$

Poznámka 2.6 *Pokud jsou splněny podmínky pro quasi-optimální řešení*

$$x_\star \equiv \tilde{x} = \frac{1}{\nu_1 y_1 + \nu_2 y_2}(y_1 u_1 + y_2 u_2),$$

pak hledáme takové $\xi_\star$*, pro které platí (2.79) pro* $x = \tilde{x}$*, kde* $\|\tilde{x}\| = \Delta$*. Jestliže vynásobíme druhou rovnici* $\tilde{x}$*, přičteme k první a použijeme vztah (2.85), dostaneme*

$$\xi_\star = -\frac{\tau + 2g^T\tilde{x} + \tilde{x}^T\mathbf{A}\tilde{x}}{1+\Delta^2} = -\frac{\tau + 2\psi(\tilde{x})}{1+\Delta^2} = \xi_1(\tau)y_1^2 + \xi_2(\tau)y_2^2.$$

2.7.4 Interpolační schema

Vztah (2.82) definuje funkci $\phi(\xi)$. Nyní chceme najít takový parametr τ, pomocí kterého bude odpovídající dvojice iterací $\{\xi, x\}$ splňovat vztahy

$$\phi(\xi) = \tau + \xi, \quad \phi'(\xi) = -\Delta^2, \quad \text{kde} \quad \phi(\xi) = -g^T x, \quad \phi'(\xi) = -x^T x \quad \text{a} \quad (\mathbf{A} + \xi\mathbf{I})x + g = 0.$$

Z (2.80) plyne, že zbývá zajistit splnění podmínky $x^T x = \Delta^2$. Zkonstruujeme funkci $\hat{\phi}(\xi)$, která interpoluje ϕ a její derivaci ϕ' ve dvou vhodně zvolených bodech. Z interpolační funkce $\hat{\phi}$ určíme $\hat{\xi}$ splňující

$$\hat{\phi}'(\hat{\xi}) = -\Delta^2.$$

Toto $\hat{\xi}$ a hodnotu $\hat{\phi}(\hat{\xi})$ použijeme k výpočtu nového τ splňujícího $\tau + \hat{\xi} = \hat{\phi}(\hat{\xi})$ a spočítáme nové iterace $\{\xi, x\}$ z matice $\mathbf{B}_\tau$.

V první iteraci potřebujeme jednobodové interpolační schema, neboť máme k dispozici pouze zvolené τ_0 a odpovídající iteraci $\{\xi_0, x_0\}$. Podle vztahu (2.81) zvolíme interpolační funkci

$$\hat{\phi}(\xi) = \frac{\vartheta^2}{\lambda + \xi}.$$

Protože platí $\tau_0 + \xi_0 = -g^T x_0$ pro $(\mathbf{A} + \xi_0\mathbf{I})x_0 + g = 0$, vede požadavek

$$\hat{\phi}(\xi_0) = \phi(\xi_0) = -g^T x_0 \quad \text{a} \quad \hat{\phi}'(\xi_0) = \phi'(\xi_0) = -x_0^T x_0$$

na přímé určení koeficientů ϑ^2 a λ :

$$\hat{\phi}(\xi_0) = \frac{\vartheta^2}{\lambda + \xi_0} = -g^T x_0 \quad \text{a} \quad \hat{\phi}'(\xi_0) = \frac{-\vartheta^2}{(\lambda+\xi_0)^2} = -x_0^T x_0 \quad \Rightarrow \quad -\frac{g^T x_0}{\lambda + \xi_0} = x_0^T x_0 \quad \Rightarrow$$

$$(2.91) \;\Rightarrow\quad \lambda = -\xi_0 - \frac{g^T x_0}{x_0^T x_0} = \frac{x_0^T \mathbf{A} x_0}{x_0^T x_0} \geq \lambda_1 \quad \text{a} \quad \vartheta^2 = -g^T x_0\,(\lambda + \xi_0) = \frac{(g^T x_0)^2}{x_0^T x_0}$$

Dále spočítáme $\hat{\xi}$ takové, že $\hat{\phi}'(\hat{\xi}) = -\Delta^2$:

$$\hat{\phi}'(\hat{\xi}) = \frac{-\vartheta^2}{(\lambda+\hat{\xi})^2} = -\Delta^2 \quad\Rightarrow\quad \frac{(g^T x_0)^2}{\|x_0\|^2}\,\frac{1}{(\lambda+\hat{\xi})^2} = \Delta^2 \quad\Rightarrow\quad |\lambda + \hat{\xi}| = \frac{g^T x_0}{\|x_0\|\,\Delta} \quad\Rightarrow$$

$$(2.92) \qquad \Rightarrow\quad \hat{\xi} = -\lambda \pm \frac{g^T x_0}{\|x_0\|\,\Delta} = \xi_0 + \frac{g^T x_0}{x_0^T x_0} \pm \frac{g^T x_0}{\|x_0\|\,\Delta}$$

a nakonec dosadíme spočtené hodnoty (2.91) a (2.92) pro $\hat{\xi}, \vartheta^2, \lambda$ do vzorce pro τ_1 :

$$\begin{aligned}
\tau_1 &= -\hat{\xi} + \hat{\phi}(\hat{\xi}) = -\hat{\xi} + \frac{\vartheta^2}{\lambda+\hat{\xi}} = \\
&= -\xi_0 - \frac{g^T x_0}{x_0^T x_0} \mp \frac{g^T x_0}{\|x_0\|\,\Delta} + \frac{(g^T x_0)^2}{x_0^T x_0}\,\frac{1}{\lambda + \left(-\lambda \pm \frac{g^T x_0}{\|x_0\|\,\Delta}\right)} = \\
&= \tau_0 + g^T x_0 - \frac{g^T x_0}{\|x_0\|^2} \mp \frac{g^T x_0}{\|x_0\|\,\Delta} \pm \frac{g^T x_0}{\|x_0\|}\,\Delta = \\
&= \tau_0 - \frac{g^T x_0}{\|x_0\|} \cdot \left(-\|x_0\| + \frac{1}{\|x_0\|} \pm \frac{1}{\Delta} \mp \Delta\right) = \\
(2.93) \qquad &= \tau_0 + \frac{\tau_0 + \xi_0}{\|x_0\|} \cdot \frac{\Delta \pm \|x_0\|}{\Delta} \cdot \left(\frac{1}{\|x_0\|} \mp \Delta\right)
\end{aligned}$$

Existují tedy dvě možnosti volby τ_1, přičemž volíme buď obě horní nebo obě dolní znaménka. Z matice $\mathbf{B}_{\tau_1}$ spočítáme první iteraci ξ_1, x_1.

V dalších krocích interpolačního schematu pro $k \geq 1$ použijeme poslední dvě po sobě jdoucí iterace

$$\xi_{k-1}, \quad \xi_k, \quad x_{k-1}, \quad x_k$$

a hodnoty

$$\phi(\xi_{k-1}) = \tau_{k-1} + \xi_{k-1}, \qquad \phi(\xi_k) = \tau_k + \xi_k, \qquad \phi'(\xi_{k-1}) = -\|x_{k-1}\|^2, \qquad \phi'(\xi_k) = -\|x_k\|^2.$$

Definujeme

$$(2.94) \qquad \hat{\phi}(\xi) = \frac{\vartheta^2}{\lambda + \xi} + \eta$$

pro jisté η a spočítáme $\hat{\xi}$ tak, aby $\hat{\phi}'(\hat{\xi}) = -\Delta^2$. Pro funkci $\frac{1}{\sqrt{-\phi'(\xi)}}$ použijeme Lagrangeův interpolační vzorec [52]

$$\frac{1}{\Delta} = \frac{1}{\sqrt{-\phi'(\xi_{k-1})}} \cdot \frac{\xi_k - \hat{\xi}}{\xi_k - \xi_{k-1}} + \frac{1}{\sqrt{-\phi'(\xi_k)}} \cdot \frac{\hat{\xi} - \xi_{k-1}}{\xi_k - \xi_{k-1}}$$

Dosazením a výpočtem dostaneme

$$\begin{aligned}
\frac{1}{\Delta} &= \frac{1}{\|x_{k-1}\|} \cdot \frac{\xi_k - \hat{\xi}}{\xi_k - \xi_{k-1}} + \frac{1}{\|x_k\|} \cdot \frac{\hat{\xi} - \xi_{k-1}}{\xi_k - \xi_{k-1}} = \frac{(\xi_k - \hat{\xi})\|x_k\| + (\hat{\xi} - \xi_{k-1})\|x_{k-1}\|}{(\xi_k - \xi_{k-1})\,\|x_k\|\|x_{k-1}\|} = \\
&= \frac{\xi_k\|x_k\| - \xi_{k-1}\|x_{k-1}\| - (\|x_k\| - \|x_{k-1}\|)\,\hat{\xi}}{(\xi_k - \xi_{k-1})\,\|x_k\|\|x_{k-1}\|} \quad\Rightarrow
\end{aligned}$$

$$\Rightarrow \qquad \frac{(\xi_k - \xi_{k-1})\,\|x_k\|\|x_{k-1}\|}{\Delta} = \xi_k\|x_k\| - \xi_{k-1}\|x_{k-1}\| - (\|x_k\| - \|x_{k-1}\|)\,\hat{\xi} \qquad \Rightarrow$$

$$\Rightarrow \qquad (\|x_k\| - \|x_{k-1}\|)\,\hat{\xi} = \xi_k\|x_k\| - \xi_{k-1}\|x_{k-1}\| - \frac{(\xi_k - \xi_{k-1})\,\|x_k\|\|x_{k-1}\|}{\Delta},$$

tedy

$$\hat{\xi} = \frac{\xi_k\|x_k\| - \xi_{k-1}\|x_{k-1}\|}{\|x_k\| - \|x_{k-1}\|} - \frac{(\xi_k - \xi_{k-1}\|)\,\|x_k\|\|x_{k-1}\|}{\Delta\,(\|x_k\| - \|x_{k-1}\|)} = \tag{2.95}$$

$$= \frac{\xi_{k-1}\|x_{k-1}\|\,(\|x_k\| - \Delta) + \xi_k\|x_k\|\,(\Delta - \|x_{k-1}\|)}{\Delta\,(\|x_k\| - \|x_{k-1}\|)} \tag{2.96}$$

Nyní najdeme hodnoty ϑ^2 a λ, které splňují požadovanou rovnost $\hat{\phi}'(\hat{\xi}) = \frac{-\vartheta^2}{(\lambda+\hat{\xi})^2} = -\Delta^2$. Z této rovnice vyjádříme $\hat{\xi} = -\lambda \pm \frac{\vartheta}{\Delta}$, porovnáme s (2.95) a dostaneme

$$\lambda = -\frac{\xi_k\|x_k\| - \xi_{k-1}\|x_{k-1}\|}{\|x_k\| - \|x_{k-1}\|} \quad \text{a} \quad \vartheta^2 = \frac{(\xi_k - \xi_{k-1})^2\|x_k\|^2\|x_{k-1}\|^2}{(\|x_k\| - \|x_{k-1}\|)^2} \tag{2.97}$$

Pro výpočet η využijeme toho, že známe hodnoty $\phi(\xi_{k-1})$ a $\phi(\xi_k)$ a položíme nejprve

$$\eta_j = \phi(\xi_j) - \frac{\vartheta^2}{\lambda + \xi_j}, \quad j = k-1, k$$

podle (2.94), protože v uzlových bodech ξ_{k-1}, ξ_k jsou funkční hodnoty interpolační funkce $\hat{\phi}(\xi)$ a interpolované funkce $\phi(\xi)$ rovny. Podle Lagrangeova interpolačního vzorce [52] platí

$$\eta = \frac{\xi_k - \hat{\xi}}{\xi_k - \xi_{k-1}}\,\eta_{k-1} + \frac{\hat{\xi} - \xi_{k-1}}{\xi_k - \xi_{k-1}}\,\eta_k$$

Nakonec provedeme aktualizaci parametru τ pomocí vzorce $\tau_{k+1} = -\hat{\xi} + \hat{\phi}(\hat{\xi})$, kde využijeme vztahy $\tau_j = -\xi_j + \phi(\xi_j)$, $j = k-1, k$. Přitom použijeme označení

$$\omega_1 = \frac{\xi_k - \hat{\xi}}{\xi_k - \xi_{k-1}}, \quad \omega_2 = \frac{\hat{\xi} - \xi_{k-1}}{\xi_k - \xi_{k-1}} \quad \Rightarrow \quad \omega_1 + \omega_2 = 1.$$

Tedy po dosazení vztahů (2.96) a (2.97) pro $\hat{\xi}, \vartheta^2, \lambda$ dostaneme

$$\begin{aligned}
\tau_{k+1} &= -\hat{\xi} + \hat{\phi}(\hat{\xi}) = -\hat{\xi} + \frac{\vartheta^2}{\lambda + \hat{\xi}} + \eta = -\hat{\xi} + \frac{\vartheta^2}{\lambda + \hat{\xi}} + \omega_1\eta_{k-1} + \omega_2\eta_k = \\
&= -\hat{\xi} + \frac{\vartheta^2}{\lambda + \hat{\xi}} + \omega_1\left(\phi(\xi_{k-1}) - \frac{\vartheta^2}{\lambda + \xi_{k-1}}\right) + \omega_2\left(\phi(\xi_k) - \frac{\vartheta^2}{\lambda + \xi_k}\right) = \\
&= \underbrace{-\hat{\xi} + \omega_1(\tau_{k-1} + \xi_{k-1}) + \omega_2(\tau_k + \xi_k)}_{\mathcal{A}_1} + \vartheta^2 \underbrace{\left(\frac{1}{\lambda + \hat{\xi}} - \frac{\omega_1}{\lambda + \xi_{k-1}} - \frac{\omega_2}{\lambda + \xi_k}\right)}_{\mathcal{A}_2};
\end{aligned}$$

$$\begin{aligned}
\mathcal{A}_1 &= \omega_1\tau_{k-1} + \omega_2\tau_k - \hat{\xi} + \frac{\xi_k - \hat{\xi}}{\xi_k - \xi_{k-1}}\,\xi_{k-1} + \frac{\hat{\xi} - \xi_{k-1}}{\xi_k - \xi_{k-1}}\,\xi_k = \\
&= \omega_1\tau_{k-1} + \omega_2\tau_k - \hat{\xi} + \frac{\hat{\xi}\,\xi_k - \hat{\xi}\,\xi_{k-1}}{\xi_k - \xi_{k-1}} = \omega_1\tau_{k-1} + \omega_2\tau_k;
\end{aligned}$$

$$
\begin{aligned}
\mathcal{A}_2 &= \frac{1}{-\frac{\xi_k\|x_k\|-\xi_{k-1}\|x_{k-1}\|}{\|x_k\|-\|x_{k-1}\|}+\hat{\xi}} - \frac{\omega_1}{-\frac{\xi_k\|x_k\|-\xi_{k-1}\|x_{k-1}\|}{\|x_k\|-\|x_{k-1}\|}+\xi_{k-1}} - \frac{\omega_2}{-\frac{\xi_k\|x_k\|-\xi_{k-1}\|x_{k-1}\|}{\|x_k\|-\|x_{k-1}\|}+\xi_k} = \\
&= (\|x_k\|-\|x_{k-1}\|)\cdot \\
&\cdot \left(\frac{1}{\|x_k\|(\hat{\xi}-\xi_k)+\|x_{k-1}\|(\xi_{k-1}-\hat{\xi})} - \frac{\omega_1}{\|x_k\|(\xi_{k-1}-\xi_k)} - \frac{\omega_2}{\|x_{k-1}\|(\xi_{k-1}-\xi_k)}\right) = \\
&= \frac{\|x_k\|-\|x_{k-1}\|}{\xi_k-\xi_{k-1}}\left(\frac{-1}{\omega_1\|x_k\|+\omega_2\|x_{k-1}\|}+\frac{\omega_1}{\|x_k\|}+\frac{\omega_2}{\|x_{k-1}\|}\right) = \\
&= \frac{\|x_k\|-\|x_{k-1}\|}{\xi_k-\xi_{k-1}}\left(\frac{(\omega_1\|x_k\|+\omega_2\|x_{k-1}\|)(\omega_1\|x_{k-1}\|+\omega_2\|x_k\|)-\|x_k\|\|x_{k-1}\|}{\|x_k\|\|x_{k-1}\|(\omega_1\|x_k\|+\omega_2\|x_{k-1}\|)}\right) = \\
&= \frac{\|x_k\|-\|x_{k-1}\|}{\xi_k-\xi_{k-1}}\cdot\frac{\omega_1\omega_2(\|x_k\|^2+\|x_{k-1}\|^2)+\|x_k\|\|x_{k-1}\|\overbrace{(\omega_1^2+\omega_2^2-1)}^{=-2\omega_1\omega_2}}{\|x_k\|\|x_{k-1}\|(\omega_1\|x_k\|+\omega_2\|x_{k-1}\|)} = \\
&= \frac{\|x_k\|-\|x_{k-1}\|}{\xi_k-\xi_{k-1}}\cdot\frac{\omega_1\omega_2(\|x_k\|-\|x_{k-1}\|)^2}{\|x_k\|\|x_{k-1}\|(\omega_1\|x_k\|+\omega_2\|x_{k-1}\|)};
\end{aligned}
$$

Celkem

$$
\begin{aligned}
\tau_{k+1} &= \omega_1\tau_{k-1}+\omega_2\tau_k + \\
&+ \frac{(\xi_k-\xi_{k-1})^2\|x_k\|^2\|x_{k-1}\|^2}{(\|x_k\|-\|x_{k-1}\|)^2}\cdot\frac{1}{\xi_k-\xi_{k-1}}\cdot\frac{\omega_1\omega_2(\|x_k\|-\|x_{k-1}\|)^3}{\|x_k\|\|x_{k-1}\|(\omega_1\|x_k\|+\omega_2\|x_{k-1}\|)} = \\
&= \omega_1\tau_{k-1}+\omega_2\tau_k+\frac{\omega_1\omega_2\|x_k\|\|x_{k-1}\|(\|x_k\|-\|x_{k-1}\|)}{\omega_1\|x_k\|+\omega_2\|x_{k-1}\|}\cdot(\xi_k-\xi_{k-1}). \qquad (2.98)
\end{aligned}
$$

Dostáváme matici $\mathbf{B}_{\tau_{k+1}}$ a spočítáme iteraci ξ_{k+1}, x_{k+1}.

Pomocí těchto dvou interpolačních schemat, jednobodového pro $k=0$ a dvoubodového pro $k \geq 1$ konstruujeme posloupnost $\{\tau_k\}$. Nyní chceme, aby se tato posloupnost blížila k optimálnímu parametru $\tau_\star = -\xi_\star - g^T x_\star$. Zkonstruujeme proto horní a dolní odhady τ_L a τ_U, které budeme v každé iteraci upravovat tak, aby se délka intervalu $\langle\tau_L, \tau_U\rangle$ zkracovala. Počáteční meze stanovíme takto:

1. Protože $\|x_\star\| \leq \Delta$ a $\xi_\star \geq -\lambda_1$, získáme horní mez
$$\tau_\star = -\xi_\star - g^T x_\star \leq \lambda_1 + \|g\|\,\Delta.$$

2. Dolní mez stanovíme takto:

   (a) Jestliže $\xi_\star = -\lambda_1$, pak ze vztahu $g^T x_\star \leq 0$, lemma 1.6, plyne
   $$\tau_\star = -\xi_\star - g^T x_\star \geq -\xi_\star = \lambda_1.$$

   (b) Nechť tedy $\xi_\star > -\lambda_1$.

      i. Pokud $\tau_\star \geq \lambda_1$, máme dolní mez, a to opět λ_1.

      ii. Pokud $\tau_\star < \lambda_1$, pak postupujeme následujícím způsobem. Víme, že $-\xi_\star$ je nejmenší vlastní číslo matice $\mathbf{B}_{\tau_\star}$, jehož odpovídající vlastní vektor $(\nu, u^T)^T$ má první složku nenulovou, $x_\star = \frac{1}{\nu}u$, kde $\|x_\star\| \leq \Delta$, a protože libovolný diagonální prvek matice není menší než nejmenší vlastní číslo téže matice, tzn. $-\xi_\star \le \tau_\star$, pak platí

$$\tau_\star + \xi_\star = -g^T x_\star = x_\star^T(\mathbf{A} + \xi_\star \mathbf{I})x_\star \ge (\lambda_1 + \xi_\star)\Delta^2 \ge (\lambda_1 - \tau_\star)\Delta^2$$

a použitím vztahu (2.81)

$$\tau_\star + \xi_\star = \sum_{j \in \mathcal{M}} \frac{\vartheta_j^2}{\lambda_j + \xi_\star} \le \sum_{j \in \mathcal{M}} \frac{\vartheta_j^2}{\lambda_j - \tau_\star} \le \frac{\|g\|^2}{\lambda_1 - \tau_\star}.$$

Odtud plyne dolní mez na $\tau_\star$

$$(\lambda_1 - \tau_\star)\Delta^2 \le \frac{\|g\|^2}{\lambda_1 - \tau_\star} \quad \Rightarrow \quad \lambda_1 - \tau_\star \le \frac{\|g\|}{\Delta} \quad \Rightarrow \quad \lambda_1 - \frac{\|g\|}{\Delta} \le \tau_\star < \lambda_1.$$

Pro $\tau_\star$ jsme tedy celkově dostali tyto meze [53]

$$\lambda_1 - \frac{\|g\|}{\Delta} \le \tau_\star \le \lambda_1 + \|g\|\,\Delta.$$

Výpočet λ_1 může být náročný a proto ho nahradíme jednoduchým horním odhadem. Můžeme ho získat buď tak, že vezmeme nejmenší diagonální prvek matice $\mathbf{A}$ nebo spočítáme Rayleighův podíl $\frac{w^T\mathbf{A}w}{w^Tw}$, kde w je libovolný vektor. Položíme tedy

$$\lambda_U = \min\{a_{ii} : i = 1, \ldots, n\} \quad \text{nebo} \quad \lambda_U = \frac{w^T\mathbf{A}w}{w^Tw}$$

a platí $\tau_\star \le \lambda_U + \|g\|\,\Delta = \tau_U$. Protože má platit $\xi_\star \ge 0$, tedy pro nejmenší vlastní číslo $-\xi_\star$ matice $\mathbf{B}_{\tau_\star}$ má platit $-\xi_\star \le 0$, nesmí být matice $\mathbf{B}_{\tau_\star}$ positivně definitní. Zvolíme proto počáteční $\tau_0 \le 0$ tak, že položíme

$$\tau_0 = \min\{0, \tau_U\}.$$

Z matice $\mathbf{B}_{\tau_0}$ spočítáme její nejmenší vlastní číslo $-\xi_1(\tau_0)$ a položíme

$$\lambda_L = -\xi_1(\tau_0).$$

Z Cauchyho věty A.1 plyne $\lambda_L \le \lambda_1$ a získáme mez $\tau_\star \ge \lambda_L - \frac{\|g\|}{\Delta}$. Použitím tohoto prostého schematu můžeme upravit horní a dolní odhady pro $\tau_\star$:

$$\tau_L \equiv \lambda_L - \frac{\|g\|}{\Delta} \le \tau_\star \le \lambda_U + \|g\|\,\Delta \equiv \tau_U. \tag{2.99}$$

Meze λ_L, λ_U aktualizujeme v každé iteraci použitím vlastního páru $\{-\xi_1(\tau_k), (\nu_1, u_1^T)^T\}$ matice $\mathbf{B}_{\tau_k}$ následujícím způsobem:

$$\lambda_L = \max\{\lambda_L,\ -\xi_1(\tau_k)\},$$

$$\lambda_U = \min\left\{\lambda_U,\ \frac{u_1^T\mathbf{A}u_1}{u_1^Tu_1}\right\}, \quad \text{kde} \quad \frac{u_1^T\mathbf{A}u_1}{u_1^Tu_1} = -\frac{u_1^T[\xi_1(\tau_k)u_1 + \nu_1 g]}{u_1^Tu_1} = -\xi_1(\tau_k) - \nu_1\frac{g^Tu_1}{u_1^Tu_1}.$$

V každé iteraci aktualizujeme jednu z mezí τ_L nebo τ_U, takže vždy zkracujeme délku intervalu $\langle\tau_L, \tau_U\rangle$. Je-li $\|x_k\| > \Delta$, pak chceme dosáhnout, aby platilo $\|x_{k+1}\| < \|x_k\|$.

Vektor x_k splňuje druhý vztah (2.80), tj. $(\mathbf{A}+\xi_k\mathbf{I})x_k = -g$. K tomu, aby platilo $\|x_{k+1}\| < \|x_k\|$, stačí podle lemmatu A.2, abychom k diagonále matice $\mathbf{A}+\xi_k\mathbf{I}$ přičetli kladné číslo κ. Tedy pro x_{k+1}, které splňuje $(\mathbf{A}+\xi_k\mathbf{I}+\kappa\mathbf{I})x_{k+1} = -g$, platí nerovnost $\|x_{k+1}\| < \|x_k\|$. Ale $\mathbf{A}+\xi_k\mathbf{I}+\kappa\mathbf{I} = \mathbf{A}+(\xi_k+\kappa)\mathbf{I} \equiv \mathbf{A}+\xi_{k+1}\mathbf{I}$, tzn. $\xi_k < \xi_{k+1}$, tedy nejmenší vlastní číslo $-\xi_{k+1}$ matice $\mathbf{B}_{\tau_{k+1}}$ musí být menší než nejmenší vlastní číslo $-\xi_k$ matice $\mathbf{B}_{\tau_k}$. To je dosaženo v případě, že platí $\tau_{k+1} < \tau_k$ (důsledek A.1). Proto položíme $\tau_U := \tau_k$. Obdobně pro $\|x_k\| < \Delta$ položíme $\tau_L := \tau_k$, čímž se interval $\langle \tau_L, \tau_U\rangle$ zmenšuje v každé iteraci.

Mějme čísla $\tau_L \le \tau_\star \le \tau_U$, příslušná nejmenší vlastní čísla $-\xi_L \le -\xi_\star \le -\xi_U$ matic $\mathbf{B}_{\tau_L}, \mathbf{B}_{\tau_\star}, \mathbf{B}_{\tau_U}$ a odpovídající vlastní vektory $(1, x_L^T)^T, (1, x_\star^T)^T, (1, x_U^T)^T$. Pak z rovnice $(\mathbf{A}+\xi\mathbf{I})x = -g$ plyne $\|x_L\| \le \|x_\star\| \le \|x_U\|$, kde $\|x_\star\| \le \Delta$.

Jestliže lze normalizovat vlastní vektor odpovídající až druhému nejmenšímu vlastnímu číslu, pak aktualizujeme jen horní mez, abychom v souladu s analýzou uvedenou v § 2.7.2 dosáhli čísla $\tau_{k+1} \le \tilde{\tau}_1$, pro které lze normalizovat vlastní vektor odpovídající prvnímu nejmenšímu vlastnímu číslu.

V případě, že τ_{k+1}, spočítané interpolačními schematy (2.93) nebo (2.98), nepatří do daného intervalu $\langle \tau_L, \tau_U\rangle$, položíme jednoduše $\tau_{k+1} := \frac{\tau_L+\tau_U}{2}$, jako střed tohoto intervalu.

2.7.5 Algoritmus

Dříve než uvedeme algoritmus této metody, určíme kriteria zastavení. V každé iteraci kontrolujeme řešení na hranici, řešení uvnitř oblasti a quasi-optimální řešení. Můžeme též skončit, jestliže dosáhneme maximálního předepsaného počtu iterací nebo pokud je délka intervalu $\langle \tau_L, \tau_U\rangle$ pro optimální $\tau_\star$ příliš malá. Za daných tolerancí

$$\varepsilon_\Delta,\ \varepsilon_\tau,\ \varepsilon_\nu,\ \varepsilon_Q \in (0,1) \quad \text{a} \quad \varepsilon_I \in \langle 0,1)$$

zavedeme následující kriteria zastavení algoritmu.

Nechť $\{-\xi_j(\tau_k), (\nu_j, u_j^T)^T\}_{j=1,2}$ jsou dva nejmenší vlastní páry matice $\mathbf{B}_{\tau_k}$ a $\{\xi_k, x_k\}$ jsou dané iterace, $\xi_k = \xi_j(\tau_k)$, $x_k = \frac{1}{\nu_j}u_j$ pro j takové, že $|\nu_j|\|g\| > \varepsilon_\nu\sqrt{1-\nu_j^2}$. Pak:

1. Řešení na hranici – Jestliže je splněna podmínka

$$\Big|\|x_k\| - \Delta\Big| \le \varepsilon_\Delta\Delta \quad \text{a} \quad \xi_1(\tau_k) \ge 0,$$

položíme $x_\star = x_k$ s odpovídajícím $\xi_\star = \xi_k$.

2. Řešení uvnitř oblasti – Jestliže je splněna podmínka

$$\|u_1\| < |\nu_1|\,\Delta \quad \text{a} \quad \xi_1(\tau_k) \le \varepsilon_I,$$

je řešení (1.25) uvnitř oblasti, matice $\mathbf{A}$ je positivně definitní a toto řešení splňuje lineární systém $\mathbf{A}x = -g$. Přirozenou volbou pro řešení této soustavy je např. metoda sdružených gradientů, algoritmus 2.7. Odpovídající $\xi_\star = 0$.

3. Quasi-optimální řešení – Jestliže je splněna podmínka

$$\nu_1^2 + \nu_2^2 \ge \frac{1}{1+\Delta^2},$$

spočítáme vektor $y \in \mathbb{R}^2$ podle poznámky 2.5 a vektor $\tilde{x} \in \mathbb{R}^n$ podle poznámky 2.3. Jestliže tato y a $\tilde{x}$ splňují podmínku 2. ve větě 2.18, tedy

$$[\xi_1(\tau) - \xi_2(\tau)]\, y_2^2\, (1 + \Delta^2) \leq -2\, \varepsilon_Q\, \psi(\tilde{x}),$$

získáváme quasi-optimální řešení problému (1.25) a položíme $x_\star = \tilde{x}$ s odpovídajícím $\xi_\star = \xi_1(\tau_k) y_1^2 + \xi_2(\tau_k) y_2^2$.

4. Interval je příliš malý – Jestliže je splněna podmínka

$$|\tau_U - \tau_L| \leq \varepsilon_\tau \max\{|\tau_L|, |\tau_U|\},$$

pak položíme $\xi_\star = \xi_1(\tau_k)$. Nemáme-li řešení na hranici, pak jsme v singulárním případě, τ_k je od $\tilde{\tau}_1 = \tau_\star$ vzdáleno o ε_τ a platí $-\xi_1(\tau_k) = \lambda_1$ (lemma 2.16). Podle lemmatu 2.19 a věty 2.17 má pro $\tau_k = \tilde{\tau}_1$ vlastní vektor příslušný λ_1 první složku nenulovou. Platí tedy $|\nu_1| > \varepsilon_\nu$ a položíme $\bar{x} = \frac{1}{\nu_1} u_1$. Protože v tomto případě platí $\|\bar{x}\| < \Delta$, spočítáme řešení $x_\star$ jako $x_\star = \bar{x} + \kappa q$, kde q je vlastní vektor asociovaný s nejmenším vlastním číslem matice $\mathbf{A}$ (lemma 2.20) a κ určíme tak, aby $\|x_\star\| = \Delta$. Platí pro něj obdoba vztahu (2.16)

$$\kappa = \frac{\Delta^2 - \|\bar{x}\|^2}{\bar{x}^T q + \operatorname{sgn}(\bar{x}^T q)\sqrt{(\bar{x}^T q)^2 + \Delta^2 - \|\bar{x}\|^2}}.$$

Vektor q je k dispozici, protože matice $\mathbf{B}_{\tau_k}$ má dvojnásobné vlastní číslo λ_1 s vlastními vektory $(\nu_1, u_1^T)^T$ a $(0, q^T)^T$, kde q je též vlastní vektor matice $\mathbf{A}$.

Nyní spojíme všechny úvahy dohromady a uvedeme algoritmus.

Algoritmus 2.15 *Parametrizovaný problém vlastních čísel pro výpočet lokálně omezeného kroku.*

Zvolíme $\varepsilon_\Delta,\ \varepsilon_\tau,\ \varepsilon_\nu,\ \varepsilon_Q \in (0,1)$ *a* $\varepsilon_I \in \langle 0,1)$.

1. *Položíme* $\lambda_U = \min\{a_{ii} : i = 1, \ldots, n\}$ *nebo* $\lambda_U = \frac{w^T \mathbf{A} w}{w^T w}$ *pro libovolné* $w \in \mathbb{R}^n$.

2. *Položíme* $\tau_U = \lambda_U + \|g\|\,\Delta,\ \tau_0 = \min\{0, \tau_U\}$ *a* $k = 0$.

3. *Spočítáme dva nejmenší vlastní páry matice* $\mathbf{B}_{\tau_k}$:

$$\left\{-\xi_1(\tau_k), (\nu_1, u_1^T)^T\right\} \quad a \quad \left\{-\xi_2(\tau_k), (\nu_2, u_2^T)^T\right\}.$$

4. *Pokud* $k = 0$, *položíme* $\lambda_L = -\xi_1(\tau_0)$ *a* $\tau_L = \lambda_L - \frac{\|g\|}{\Delta}$.

5. *Jestliže*

$$|\nu_i|\|g\| \leq \varepsilon_\nu \sqrt{1 - \nu_i^2}, \quad i = 1, 2 \quad a \quad |\tau_U - \tau_L| > \varepsilon_\tau \max\{|\tau_L|, |\tau_U|\},$$

položíme $\tau_U = \tau_k,\ \tau_k = \frac{\tau_L + \tau_U}{2}$, *spočítáme*

$$\left\{-\xi_1(\tau_k), (\nu_1, u_1^T)^T\right\} \quad a \quad \left\{-\xi_2(\tau_k), (\nu_2, u_2^T)^T\right\}$$

a opakujeme krok 5.

6. *Jestliže* $|\nu_1| \|g\| > \varepsilon_\nu \sqrt{1-\nu_1^2}$, *položíme*

$$\xi_k = \xi_1(\tau_k) \quad a \quad x_k = \frac{1}{\nu_1} u_1.$$

Pokud $\|x_k\| < \Delta$, *aktualizujeme* $\tau_L = \tau_k$, *jinak* $\tau_U = \tau_k$ *a přejdeme na krok 8.*

7. *Jestliže* $|\nu_2| \|g\| > \varepsilon_\nu \sqrt{1-\nu_2^2}$, *položíme*

$$\xi_k = \xi_2(\tau_k) \quad a \quad x_k = \frac{1}{\nu_2} u_2.$$

Aktualizujeme $\tau_U = \tau_k$.

8. *Pro iteraci* x_k *testujeme stopující kriteria algoritmu podle bodů uvedených výše.*

9. *Položíme* $\lambda_L = \max\{\lambda_L, \; -\xi_1(\tau_k)\}$, $\lambda_U = \min\left\{\lambda_U, \; -\xi_1(\tau_k) - \nu_1 \frac{g^T u_1}{u_1^T u_1}\right\}$ *a*

$$\tau_L = \max\left\{\tau_L, \; \lambda_L - \frac{\|g\|}{\Delta}\right\}, \qquad \tau_U = \min\{\tau_U, \; \lambda_U + \|g\| \Delta\}.$$

10. *Je-li* $k = 0$, *určíme* τ_{k+1} *podle interpolačního schematu (2.93), jinak podle (2.98).*

11. *Jestliže* $\tau_{k+1} \notin \langle \tau_L, \tau_U \rangle$, *položíme* $\tau_{k+1} = \frac{\tau_L + \tau_U}{2}$.

12. *Položíme* $k := k + 1$ *a návrat na krok 3.*

Na závěr dokážeme globální konvergenci.

Věta 2.19 *Parametrizovaný problém vlastních čísel, sestavený na základě algoritmu 2.15, je globálně konvergentní metodou.*

DŮKAZ: Cílem algoritmu 2.15 je nalézt takové $\tau_\star$, že pro vlastní vektor $(1, x_\star^T)^T$ příslušný nejmenšímu vlastnímu číslu $-\xi_1(\tau_\star) \leq 0$ matice $\mathbf{B}_{\tau_\star}$ platí $\|x_\star\| = \Delta$. Pokud během výpočtu zjistíme, že je matice $\mathbf{A}$ positivně definitní, $-\xi_1(\tau) > 0$, přejdeme na algoritmus 2.7, u kterého je dokázána globální konvergence, věta 2.8. Jestliže dostaneme quasi-optimální řešení $\tilde{x}$, je podle věty 2.18 o ε vzdáleno od optimálního lokálně omezeného kroku. Skončíme-li s $\tau_\star$, pak platí $\xi_\star = \xi_1(\tau_\star) \geq 0$, $\|x_\star\| = \Delta$, $\mathbf{A} + \xi_\star \mathbf{I} \succeq 0$, neboť $\xi_\star \geq -\lambda_1$, z (2.80) plyne $(\mathbf{A} + \xi_\star \mathbf{I}) x_\star + g = 0$, a protože také platí $(\|x_\star\| - \Delta)\xi_\star = 0$, je $x_\star$ podle věty 1.5 optimálním lokálně omezeným krokem. □

2.8 Numerické výsledky

Algoritmy popsané v této kapitole byly implementovány v prostředí optimalizačního systému UFO [39] a testovány pomocí dvou kolekcí rozsáhlých a strukturovaných testovacích problémů [42] (vždy 22 optimalizačních problémů bez omezujících podmínek s 1000 nebo 5000 proměnnými). Jednotlivé testované algoritmy jsou uvedeny v tabulce 2.3.

Tyto algoritmy byly použity jako součást programů realizujících Newtonovu metodu (MN) pro minimalizaci obecné účelové funkce [10] a dvě modifikace Gaussovy-Newtonovy metody pro minimalizaci součtu čtverců, kombinace s Newtonovou metodou (MG) a s Marwilovou metodou s proměnnou metrikou (GM), [33].

U metody sdružených gradientů, algoritmus (2.7), a kombinované metody LCG(m), algoritmus (2.13), byly použity tři typy předpodmínění:

Zkratka	Použitá metoda	Algoritmus
CHDM	Optimální krok s použitím Choleského rozkladu	2.2
DLM-1	Jednoduchá metoda psí nohy	2.5 pro $\gamma = 1$
DLM-2	Modifikovaná (dvojitá) metoda psí nohy	2.5 pro $\gamma = \beta$
DLCHM	Optimální krok na dvojrozměrném podprostoru	2.6
CGM	Metoda sdružených gradientů (CG)	2.7
PCGM	Předpodmíněná metoda sdružených gradientů	2.9
CGDLM(m)	Kombinace metody CG s metodou psí nohy	2.8 pro m
LM(m)	Optimální krok s použitím Lanczosovy metody	2.12 pro m
LCGM(m)	Kombinace Lanczosovy metody s metodou CG	2.13 pro m
CGLM(m)	Kombinace metody CG s Lanczosovou metodou	2.14 pro m

Tabulka 2.3: Přehled testovaných metod pro neomezenou minimalizaci.

1. P = 0 : Bez předpodmínění, $\mathbf{C} = \mathbf{I}$.
2. P = −1 : S předpodmíněním pomocí matice $\mathbf{C}$.
3. P = +1 : S předpodmíněním pomocí matice $\mathbf{C}$. Před zahájením iteračního procesu testujeme, zda řešení soustavy $\mathbf{C}w = -g$ splňuje podmínku $\|\mathbf{A}w + g\| \le \varepsilon\|g\|$. Je-li tato podmínka splněna, položíme $x_\star = w$ a metodu sdružených gradientů vynecháme.

Výsledky jsou uvedeny v tabulkách 2.4-2.7, kde jednotlivé sloupce znamenají:

- Metoda – použitá metoda podle tabulky 2.3
- P – typ předpodmínění: $0, -1, +1$
- NIT – celkový počet hlavních iterací (y_k v algoritmu 1.1)
- NFV – celkový počet vyčíslení hodnoty funkce F
- NFG – celkový počet vyčíslení gradientu funkce F
- NCG – celkový počet iterací metody sdružených gradientů (vnitřní iterace)
- T – celkový čas

Kromě metody CGDLM(5) byla rovněž testována metoda CGDLM(3) a CGDLM(8), ale rozdíly byly nepodstatné, CGDLM(5) vyšla nejlépe. Rovněž metoda LM(100) se ukázala nejlepší, pro např. $m = 50$ či $m = 500$ vyšly horší výsledky. Obdobně LCGM(5) byla účinnější ve srovnání s LCGM(3) a LCGM(8). Buď bylo potřeba více iterací nebo se spotřebovalo více strojového času. Konečně totéž lze říci i o metodě CGLM(10).

Z výsledků lze vyčíst známý fakt, že metoda sdružených gradientů bez předpodmínění není pro rozsáhlé strukturované úlohy efektivní. S předpodmíněním dochází k výraznému poklesu počtu iterací. Také časově vychází předpodmínění většinou lépe, což je patrné hlavně u problémů s 5000 proměnnými.

Mezi nejlepší metody patří výpočet optimálního lokálně omezeného kroku s použitím Choleského rozkladu. Tento algoritmus urychluje konvergenci optimalizačních metod (menší počet hlavních iterací ve srovnání s ostatními algoritmy), není však nejrychlejší. Další účinnou metodou je kombinace metody sdružených gradientů s metodou psí nohy. Tato metoda patří mezi nejrychlejší s malým počtem hlavních iterací, nicméně musíme vzít do úvahy ještě iterace metody sdružených gradientů. Velmi dobře vychází také nová kombinovaná metoda LCGM(5) s předpodmíněním, kde se sníží počet iterací metody sdružených gradientů a i časově je tato metoda výhodná. Další metody, tedy metoda psí nohy, prostá metoda sdružených gradientů a kombinovaná metoda CGLM(10), jsou méně efektivní. Za pozornost stojí též předpodmíněná metoda sdružených gradientů PCGM pro 5000 proměnných. Nejslabší metodou je patrně výpočet optimálního lokálně omezeného kroku založený na použití Lanczosovy metody LM(100).

Pro rozsáhlé strukturované úlohy je prakticky nepoužitelná metoda založená na použití lineární kombinace vlastních vektorů, algoritmus 2.4, neboť používá Arnoldiho proces, takže dostaneme velkou hustou matici $\mathbf{V}$.

Také byl předběžně testován výpočet optimálního lokálně omezeného kroku založený na parametrizovaném problému vlastních čísel, algoritmus 2.15. Tato metoda sice dokáže spočítat vše s vysokou přesností, avšak spotřebuje velmi mnoho strojového času pro výpočet nejmenšího vlastního čísla a odpovídajícího vlastního vektoru matice $\mathbf{B}_\tau$ pomocí programů knihovny ARPACK [29]. Ke zlepšení nevedlo ani použití jiné numerické metody, např. metody inverzních iterací. Dá se tedy předpokládat, že efektivita metod s optimálním lokálně omezeným krokem klesá, pokud potřebujeme znát vlastní čísla obecných matic.

Na závěr je nutno poznamenat, že nelze obecně rozhodnout, které metody jsou nejlepší, protože pro některé úlohy může být ta či ona metoda dobrá, avšak pro jinou úlohu může dávat horší výsledky.

Graficky jsou výsledky testů znázorněny na obrázcích 2.6-2.9.

Informace o systému UFO a testovaných příkladech lze získat na adrese

http://www.cs.cas.cz/~luksan/test.html

Metoda	P	NIT	NFV	NFG	NCG	T
CHDM	0	1 918	1 955	8 797	0	4.65
DLM-1	0	2 515	2 716	11 859	0	4.42
DLM-2	0	2 411	2 577	11 395	0	4.50
DLCHM	0	2 514	2 721	11 939	0	4.84
CGM	0	3 329	3 784	16 456	53 573	8.20
PCGM	-1	2 631	2 823	13 019	910	5.14
PCGM	+1	2 626	2 818	12 994	56	5.24
CGDLM(5)	0	2 292	2 456	10 673	12 203	4.61
LM(100)	0	3 107	3 444	15 306	55 632	8.53
LCGM(5)	0	3 003	3 332	14 702	52 831	8.89
LCGM(5)	-1	1 999	2 046	9 201	1 161	4.25
LCGM(5)	+1	2 000	2 047	9 206	304	4.30
CGLM(10)	0	3 126	3 498	15 418	59 639	8.37

Tabulka 2.4: Metoda MN pro neomezenou minimalizaci, $n = 1000$.

Metoda	P	NIT	NFV	NFG	NCG	T
CHDM	0	4 278	4 345	4 567	0	9.78
DLM-1	0	7 045	7 201	7 467	0	12.92
DLM-2	0	6 462	6 596	6 872	0	12.17
DLCHM	0	5 771	5 903	6 132	0	11.50
CGM	0	4 606	4 897	4 916	60 119	12.55
PCGM	-1	5 449	5 572	5 884	5 529	10.74
PCGM	+1	5 476	5 600	5 898	4 581	10.84
CGDLM(5)	0	4 006	4 149	4 285	23 823	8.56
LM(100)	0	4 681	4 929	4 989	64 829	12.41
LCGM(5)	0	4 573	4 855	4 885	60 846	13.70
LCGM(5)	-1	6 335	6 411	6 599	7 350	13.06
LCGM(5)	+1	6 337	6 442	6 601	6 046	13.34
CGLM(10)	0	5 034	5 286	5 344	73 493	13.08

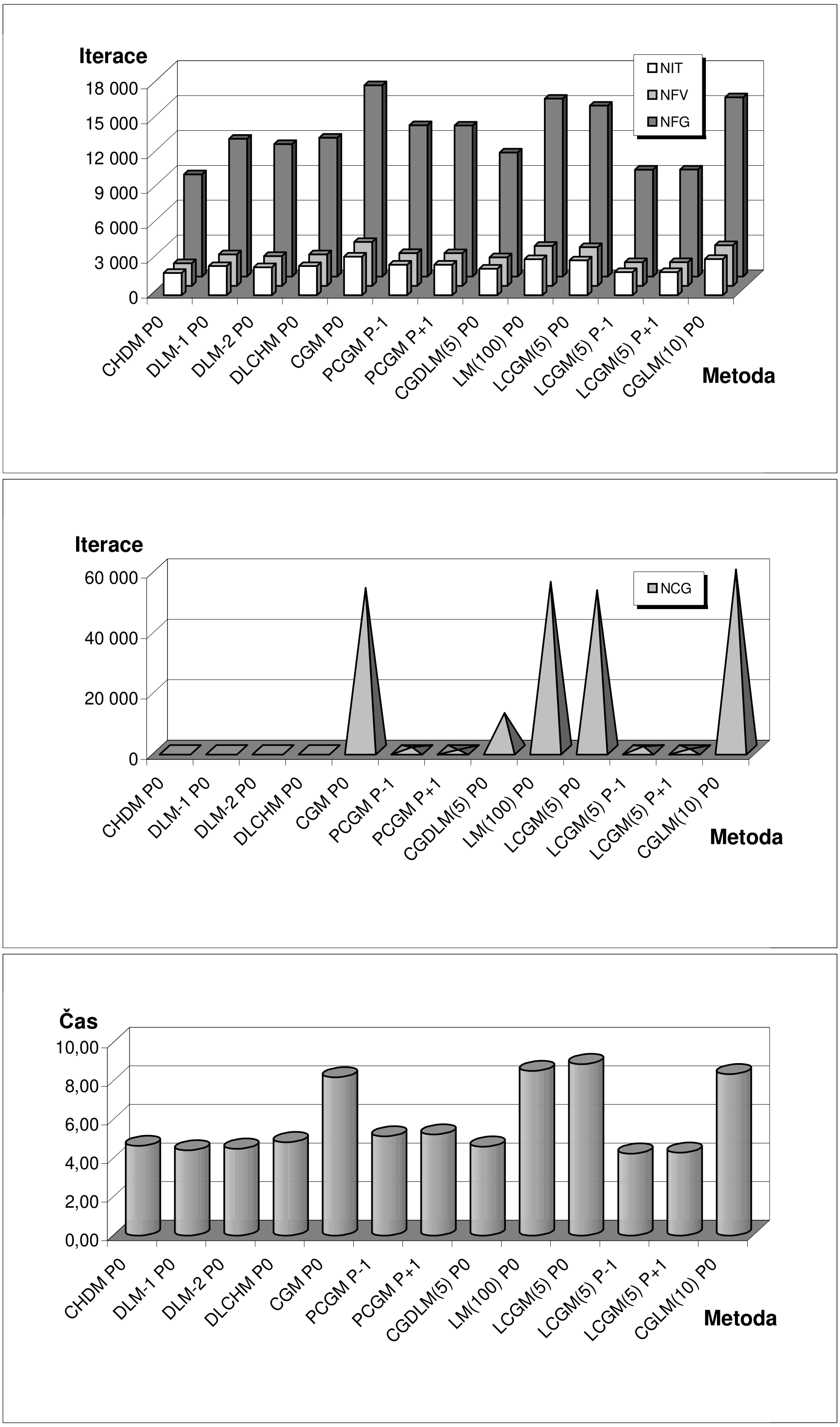
Tabulka 2.5: Metoda MG pro minimalizaci součtu čtverců, $n = 1000$.

Metoda	**P**	**NIT**	**NFV**	**NFG**	**NCG**	**T**
CHDM	0	4 108	4 242	4 129	0	8.84
DLM-1	0	5 731	5 898	5 751	0	10.05
DLM-2	0	6 370	6 504	6 391	0	10.84
DLCHM	0	5 719	5 913	5 740	0	10.24
CGM	0	4 965	5 317	4 987	62 837	12.25
PCGM	-1	7 639	7 851	7 659	8 445	17.03
PCGM	+1	7 569	7 778	7 589	8 386	16.56
CGDLM(5)	0	3 957	4 104	3 978	23 463	8.06
LM(100)	0	5 076	5 426	5 097	71 035	12.97
LCGM(5)	0	4 718	5 116	4 737	64 384	13.59
LCGM(5)	-1	6 065	6 183	6 087	9 188	12.66
LCGM(5)	+1	5 905	6 053	5 925	7 872	12.33
CGLM(10)	0	4 986	5 312	5 007	75 463	12.45

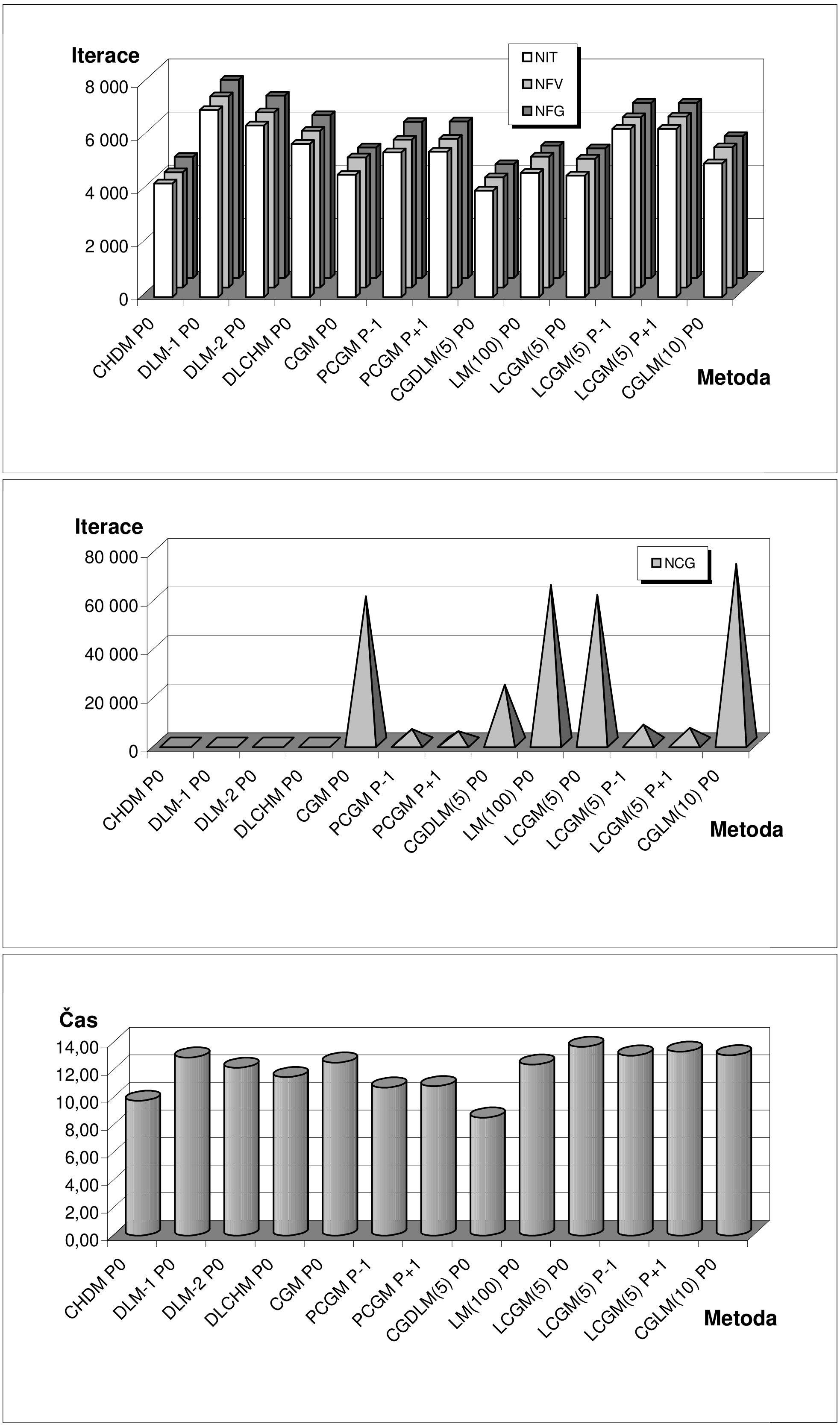
Tabulka 2.6: Metoda GM pro minimalizaci součtu čtverců, $n = 1000$.

Metoda	**P**	**NIT**	**NFV**	**NFG**	**NCG**	**T**
CHDM	0	8 391	8 566	35 824	0	2:02.44
DLM-1	0	9 657	10 133	42 425	0	1:55.77
DLM-2	0	9 717	10 195	42 452	0	1:52.20
DLCHM	0	9 625	10 150	42 260	0	1:56.05
CGM	0	16 894	19 163	83 933	358 111	6:04.42
PCGM	-1	10 600	11 271	50 385	3 767	2:25.42
PCGM	+1	10 599	11 269	50 382	83	2:26.88
CGDLM(5)	0	8 938	9 276	39 032	47 236	2:02.84
LM(100)	0	14 679	16 383	71 483	366 695	6:41.45
LCGM(5)	0	14 906	16 751	72 727	355 106	6:26.30
LCGM(5)	-1	8 347	8 454	35 939	4 329	1:48.87
LCGM(5)	+1	8 346	8 454	35 933	624	1:49.67
CGLM(10)	0	15 655	17 723	76 696	394 060	6:30.89

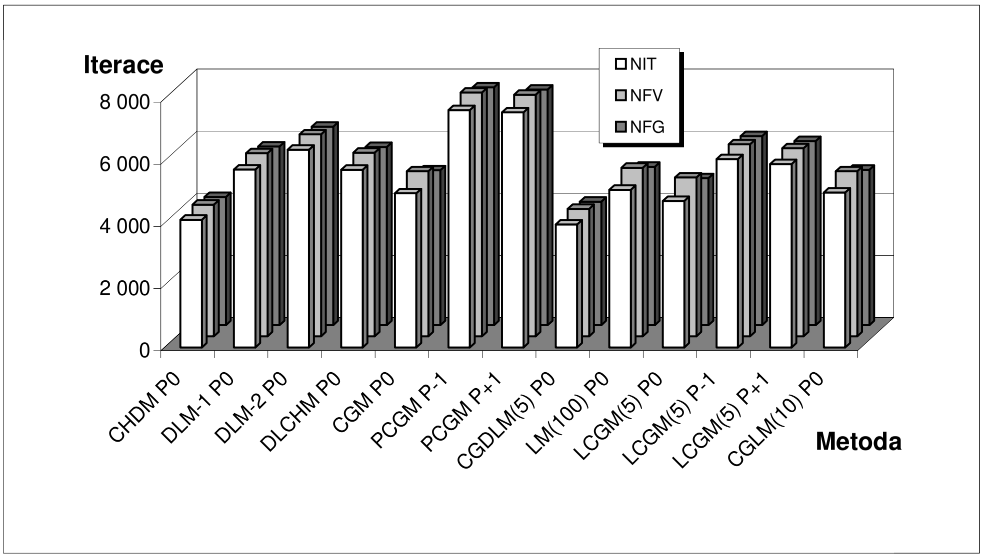
Tabulka 2.7: Metoda MN pro neomezenou minimalizaci, $n = 5000$.

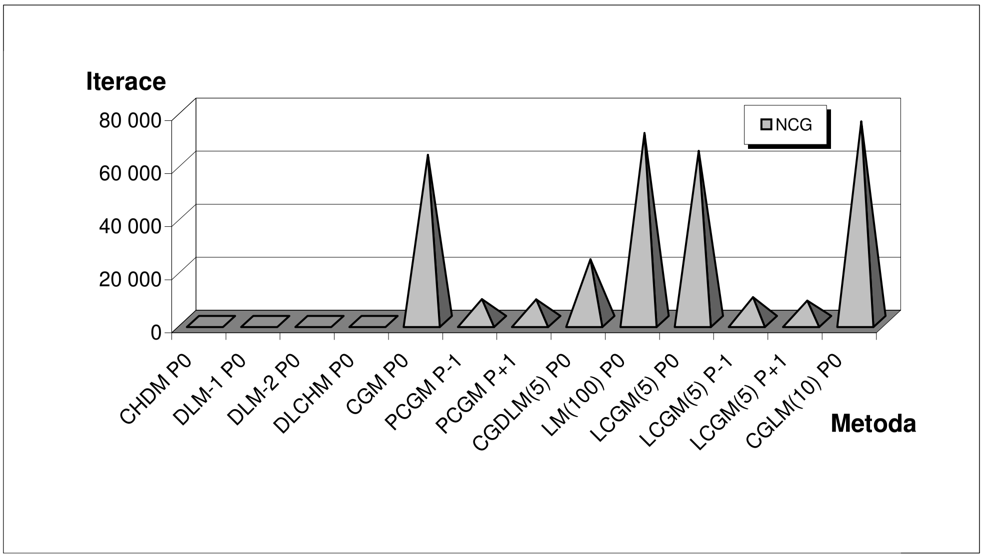


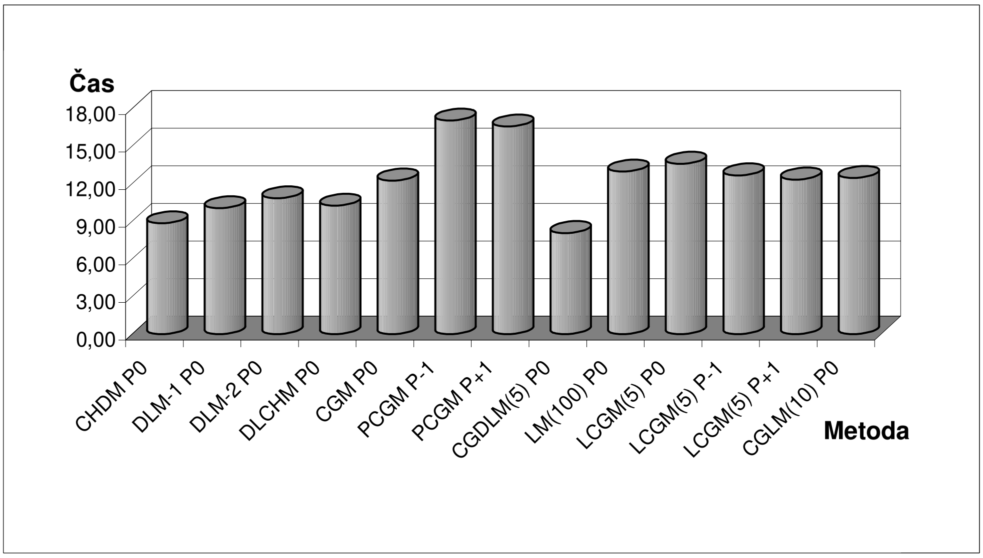
Obrázek 2.6: Metoda MN pro neomezenou minimalizaci, $n = 1000$.



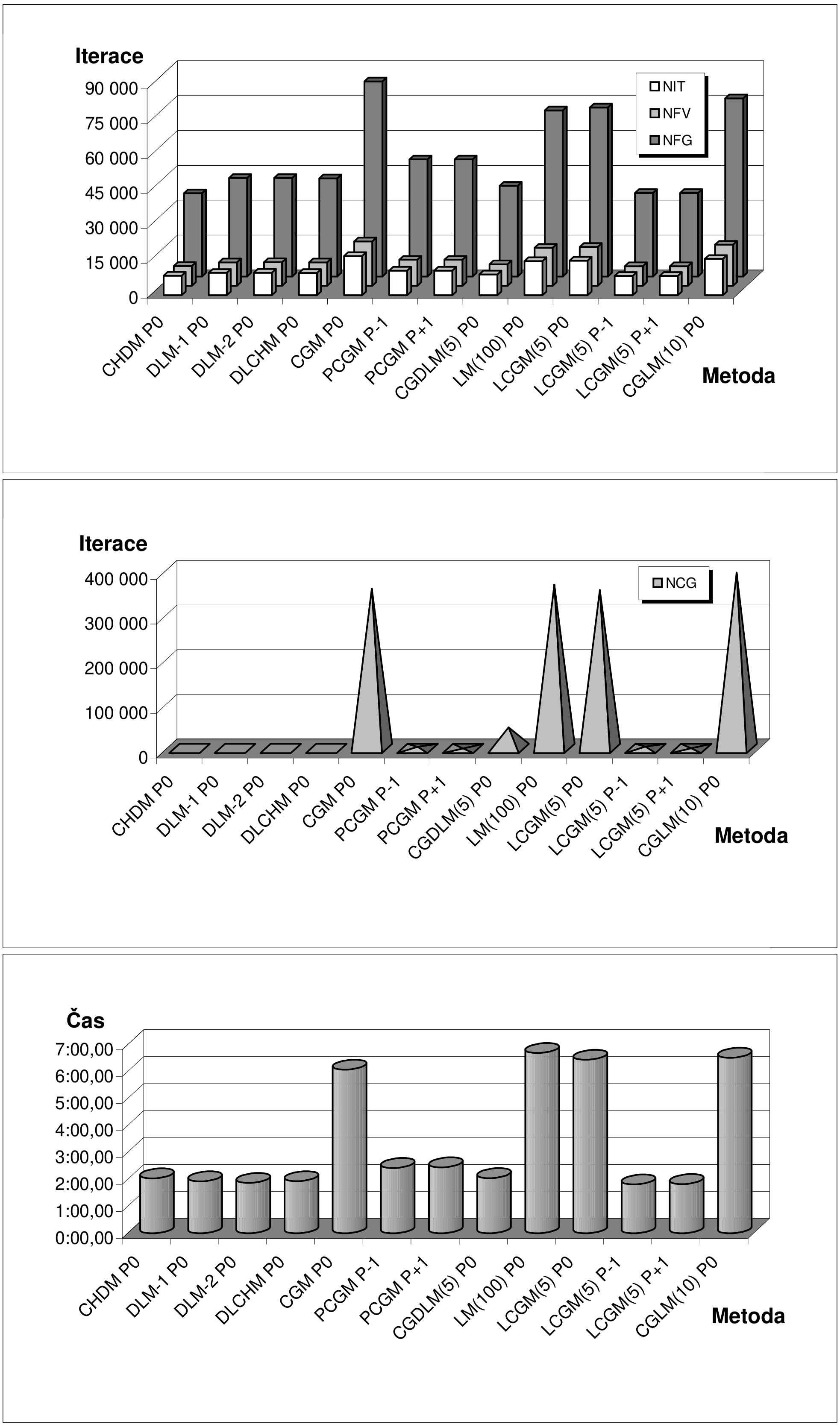
Obrázek 2.7: Metoda MG pro minimalizaci součtu čtverců, $n = 1000$.







Obrázek 2.8: Metoda GM pro minimalizaci součtu čtverců, $n = 1000$.



Obrázek 2.9: Metoda MN pro neomezenou minimalizaci, $n = 5000$.

Kapitola 3

Metody vnitřních bodů pro minimalizaci s omezeními

Problémem minimalizace funkce $F(y)$ na množině omezení $c(y) = 0$ se zabývají práce [40], [61], přičemž použití metody s lokálně omezeným krokem nalezneme v [27], [41], [48]. V jiných pracech najdeme jak omezení $c_E(y) = 0$, tak omezení $c_I(y) \leq 0$, [08], [17], [37], [45], [67]. O metodě vnitřního bodu pojednávají práce [03], [04], [36], [62], [65], [66]. Jiný typ omezení $l \leq y \leq u$ je vyšetřen v [06], [07], [09], [30], [31], [32].

V této kapitole pojednáme o problému minimalizace funkce F na množině obecných omezení ve tvaru rovností i nerovností a navážeme na výše uvedené práce. Zformulujeme ekvivalentní problém, ve kterém se vyskytují pouze omezení s rovnostmi a odvodíme soustavu lineárních rovnic pro nalezení směrových vektorů. Ukážeme, že řešení této soustavy je ekvivalentní úloze nalezení lokálně omezeného kroku kvadratické funkce s lineárním omezením. Sestrojíme novou původní iterační metodu, sloužící k jejímu řešení. Dále uvedeme kompletní algoritmus metody vnitřních bodů s lokálně omezeným krokem pro obecnou úlohu nelineárního programování, ve kterém použijeme novou speciální rozšířenou Lagrangeovu funkci.

V závěru kapitoly použijeme k posouzení přijatelnosti získaného kroku namísto pokutové funkce jiný prostředek, tzv. filtr, [01], [05], [14], [15], [16], [22], [24], [44], [63], [64], a sestrojíme nový algoritmus metody vnitřních bodů používající princip filtru.

3.1 Metody vnitřních bodů

Budeme se zabývat obecnou úlohou nelineárního programování, t.j. problémem nalezení minima funkce f na množině dané omezeními ve tvaru rovností a nerovností

$$f(y) \;\rightarrow\; \min, \quad \text{vzhledem k} \quad c_I(y) \leq 0, \quad c_E(y) = 0, \tag{3.1}$$

kde

$$f : \mathbb{R}^n \rightarrow \mathbb{R}, \quad c_I : \mathbb{R}^n \rightarrow \mathbb{R}^{m_I}, \quad c_E : \mathbb{R}^n \rightarrow \mathbb{R}^{m_E}$$

jsou dvakrát spojitě diferencovatelné funkce ($c_I \leq 0$ je myšleno po složkách),

$$I = \{1, \ldots, m_I\}, \quad E = \{m_I + 1, \ldots, m_I + m_E = m\}$$

a předpokládáme, že $m_E \leq n$. Označíme-li

$$c(y) = [c_I^T(y), c_E^T(y)]^T = [c_1(y), \ldots, c_m(y)]^T \in \mathbb{R}^m,$$

pak podle věty 1.1 existuje vektor Lagrangeových multiplikátorů

$$u_\star = (u_1, \dots, u_m)^T \in \mathbb{R}^m$$

tak, že řešení $y_\star$ splňuje rovnost

$$\nabla f(y_\star) + \sum_{k=1}^{m} u_k \nabla c_k(y_\star) = 0, \tag{3.2}$$

$$c_k(y_\star) = 0, \quad k \in E; \qquad c_k(y_\star) \le 0, \quad u_k \ge 0, \quad u_k c_k(y_\star) = 0, \quad k \in I, \tag{3.3}$$

kde

$$\nabla f(y_\star) = \left(\frac{\partial f(y_\star)}{\partial y_1}, \dots, \frac{\partial f(y_\star)}{\partial y_n}\right)^T, \quad \nabla c_k(y_\star) = \left(\frac{\partial c_k(y_\star)}{\partial y_1}, \dots, \frac{\partial c_k(y_\star)}{\partial y_n}\right)^T.$$

Tento problém je obtížně řešitelný z důvodu výskytu nerovností $c_I(y) \le 0$. Abychom tyto nerovnosti odstranili, zavedeme vektor pomocných proměnných

$$s \equiv s_I = (s_1, \dots, s_{m_I})^T \in \mathbb{R}^{m_I}$$

a převedeme problém (3.1) na úlohu s rovnostmi a jednoduchými nerovnostmi

$$f(y) \ \rightarrow \ \min, \quad c_I(y) + s = 0, \quad s \ge 0, \quad c_E(y) = 0. \tag{3.4}$$

Podstata metod vnitřních bodů spočívá v náhradě omezení $s \ge 0$ přidáním logaritmického barierového členu s konstantou $\mu > 0$ k funkci f. Definujeme-li vektor

$$\begin{aligned} h(y,s) &= [h_I^T(y,s), h_E^T(y,s)]^T = [h_1(y,s), \dots, h_m(y,s)]^T = \\ &= [c_1(y) + s_1, \dots, c_{m_I}(y) + s_{m_I}, c_{m_I+1}(y), \dots, c_m(y)]^T \in \mathbb{R}^m, \end{aligned}$$

dostaneme úlohu

$$F(y,s) \ \rightarrow \ \min, \quad h(y,s) = 0, \tag{3.5}$$

kde

$$F(y,s) = f(y) - \mu e^T \ln(\mathbf{S}_I) e = f(y) - \mu \sum_{i=1}^{m_I} \ln(s_i), \tag{3.6}$$

$e = (1, \dots, 1)^T$ a $\mathbf{S}_I = \operatorname{diag}(s_1, \dots, s_{m_I})$, která má pouze omezení ve tvaru rovností. Logaritmická barierová funkce vyžaduje, aby platilo $s_i > 0 \ \forall i \in I$. Tyto složky však mohou být libovolně malé, takže dostaneme nerovnost $c_i(y) \le 0$ s libovolnou přesností, přestože v metodách vnitřních bodů nikdy nedosáhneme přesné rovnosti $c_i(y) = 0$, $i \in I$. Dále předpokládáme, že $\mu \to 0$, čímž dostaneme v limitním případě řešení původní úlohy (3.1). Cílem našich dalších úvah bude tedy nalezení řešení modifikovaného problému (3.5), jehož řešení se pro $\mu \to 0$ blíží k řešení původního problému (3.1). Zkonstruujeme algoritmus, využívající princip lokálně omezeného kroku, pro nalezení řešení problému (3.5) s pevným $\mu > 0$.

Jsou-li splněny podmínky regularity, definice 1.4, bod 2., pak řešení $y_\star, s_\star$ problému (3.5) splňuje následující KKT podmínky (nutné podmínky pro extrém). Buď

$$\mathcal{L}(y,s,u) = F(y,s) + u^T h(y,s) = f(y) - \mu e^T \ln(\mathbf{S}_I) e + u^T h(y,s)$$

Lagrangeova funkce problému (3.5) s multiplikátory $u = [u_I^T, u_E^T]^T = (u_1, \dots, u_m)^T \in \mathbb{R}^m$ a $\mathbf{U}_I = \operatorname{diag}(u_1, \dots, u_{m_I})$. Označme Jacobiho matici

$$
\begin{aligned}
[\mathbf{A}_I(y), \mathbf{A}_E(y)] &= \nabla_y h(y,s) = [\nabla_y h_I(y,s), \nabla_y h_E(y,s)] = \\
&= [\nabla_y h_1(y,s), \dots, \nabla_y h_m(y,s)] = [\nabla_y c_1(y), \dots, \nabla_y c_m(y)] = \\
&= \begin{pmatrix} \frac{\partial c_1(y)}{\partial y_1} & \dots & \frac{\partial c_m(y)}{\partial y_1} \\ \vdots & & \vdots \\ \frac{\partial c_1(y)}{\partial y_n} & \dots & \frac{\partial c_m(y)}{\partial y_n} \end{pmatrix} \in \mathbb{R}^{n \times m}
\end{aligned}
$$

a dále

$$
\begin{aligned}
g_y(y,s,u) &= \nabla_y \mathcal{L}(y,s,u) = \nabla_y f(y) + [\mathbf{A}_I(y), \mathbf{A}_E(y)]\, u \\
g_s(y,s,u) &= \nabla_s \mathcal{L}(y,s,u) = -\mu \mathbf{S}_I^{-1} e + u_I = -\mu \mathbf{S}_I^{-1} e + \mathbf{U}_I e \\
\mathbf{G}_{yy}(y,s,u) &= \nabla^2_{yy} \mathcal{L}(y,s,u) = \nabla^2_{yy} f(y) + \sum_{k=1}^{m} u_k \nabla^2_{yy} c_k(y) \\
\mathbf{G}_{ss}(y,s,u) &= \nabla^2_{ss} \mathcal{L}(y,s,u) = \mu \mathbf{S}_I^{-2}
\end{aligned}
$$

gradienty a Hessovy matice funkce $\mathcal{L}$. Pak existuje vektor $u_\star \in \mathbb{R}^m$, že platí (věta 1.1)

$$
\nabla_y \mathcal{L}(y_\star, s_\star, u_\star) = 0, \quad \nabla_s \mathcal{L}(y_\star, s_\star, u_\star) = 0, \quad \nabla_u \mathcal{L}(y_\star, s_\star, u_\star) = 0. \tag{3.7}
$$

Hledáme tedy řešení $y_\star, s_\star, u_\star$ soustavy rovnic

$$
g_y(y,s,u) = 0, \quad g_s(y,s,u) = 0, \quad h(y,s) = 0. \tag{3.8}
$$

Základní metody pro řešení problému (3.5) jsou iterační a jejich iterační krok má tvar

$$
y^+ = y + \alpha_y d_y, \quad s^+ = s + \alpha_s d_s, \quad u^+ = u + \alpha_u d_u, \tag{3.9}
$$

kde $d_y \in \mathbb{R}^n$, $d_s \in \mathbb{R}^{m_I}$, $d_u \in \mathbb{R}^m$ jsou směrové vektory a $\alpha_y, \alpha_s, \alpha_u > 0$ jsou délky kroku. Pro definici logaritmické barierové funkce je nutná podmínka $s_i^+ > 0\ \forall i \in I$ a z rovnice $g_s = 0$ plyne, že požadujeme rovněž nerovnost $u_i^+ > 0\ \forall i \in I$. Obě nerovnosti lze zajistit vhodným výběrem délky kroku. Pro nalezení směrových vektorů použijeme metodu odvozenou z Newtonovy metody aplikovanou na nelineární KKT systém (3.8), kde $\alpha_y = \alpha_s = \alpha_u = 1$.

Budeme uvažovat dva přístupy. Primární formulace vznikne použitím Newtonovy metody na soustavu (3.8). Jestliže nejprve vynásobíme druhou rovnici (3.8) maticí $\mathbf{S}_I$

$$
g_s = -\mu \mathbf{S}_I^{-1} e + \mathbf{U}_I e = 0 \quad \Rightarrow \quad \mathbf{S}_I g_s = -\mu e + \mathbf{S}_I \mathbf{U}_I e = 0
$$

a teprve na tuto výslednou soustavu aplikujeme Newtonovu metodu, dostaneme tzv. primárně-duální formulaci. Ta je výhodnější, vede na efektivnější algoritmy. Jedním z důvodů je ten fakt, že když jde μ k nule, tak pravá strana μe primárně-duální formulace jde rovněž k nule. Naopak člen $\mu \mathbf{S}_I^{-1} e$ v primární formulaci může mít složky daleko od nuly, pokud platí $s_i \to 0$. Po aplikaci Newtonovy metody dostaneme rovnice uvedené v tabulce 3.1 – liší se pouze druhá.

Předpokládejme, že $d_u = \left[d_{u_I}^T, d_{u_E}^T\right]^T$ a nechť matice $\mathbf{B}_{yy}$ aproximuje Hessovu matici $\mathbf{G}_{yy}$, neboť v praxi místo druhých derivací počítáme jejich aproximace pomocí diferencí.

Primární formulace	Primárně-duální formulace
$\mathbf{G}_{yy}d_y + [\mathbf{A}_I, \mathbf{A}_E]\, d_u = -g_y$	$\mathbf{G}_{yy}d_y + [\mathbf{A}_I, \mathbf{A}_E]\, d_u = -g_y$
$\mathbf{G}_{ss}d_s + [\mathbf{I}, 0]\, d_u = -g_s$	$\mathbf{U}_I d_s + \mathbf{S}_I d_{u_I} = -\mathbf{S}_I g_s$
$[\mathbf{A}_I, \mathbf{A}_E]^T d_y + \left[d_s^T, 0\right]^T = -h$	$[\mathbf{A}_I, \mathbf{A}_E]^T d_y + \left[d_s^T, 0\right]^T = -h$

Tabulka 3.1: Primární a primárně-duální formulace

Abychom výslednou soustavu lépe vyškálovali, vynásobíme obě druhé rovnice jistou diagonální maticí $\mathbf{D}_I \in \mathbb{R}^{m_I}$ a abychom dostali symetrickou soustavu, nahradíme neznámou d_s výrazem $\mathbf{D}_I^{-1} d_s$. Dostaneme tuto primární, resp. primárně-duální iterační metodu se soustavou lineárních rovnic pro neznámé $d_y, d_s, d_{u_I}, d_{u_E}$:

$$\begin{pmatrix} \mathbf{B}_{yy} & 0 & \mathbf{A}_I & \mathbf{A}_E \\ 0 & \mu \mathbf{D}_I \mathbf{S}_I^{-2} \mathbf{D}_I & \mathbf{D}_I & 0 \\ \mathbf{A}_I^T & \mathbf{D}_I & 0 & 0 \\ \mathbf{A}_E^T & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} d_y \\ \mathbf{D}_I^{-1} d_s \\ d_{u_I} \\ d_{u_E} \end{pmatrix} = - \begin{pmatrix} g_y \\ \mathbf{D}_I g_s \\ h_I \\ h_E \end{pmatrix}$$

resp.

$$\begin{pmatrix} \mathbf{B}_{yy} & 0 & \mathbf{A}_I & \mathbf{A}_E \\ 0 & \mathbf{D}_I \mathbf{S}_I^{-1} \mathbf{U}_I \mathbf{D}_I & \mathbf{D}_I & 0 \\ \mathbf{A}_I^T & \mathbf{D}_I & 0 & 0 \\ \mathbf{A}_E^T & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} d_y \\ \mathbf{D}_I^{-1} d_s \\ d_{u_I} \\ d_{u_E} \end{pmatrix} = - \begin{pmatrix} g_y \\ \mathbf{D}_I g_s \\ h_I \\ h_E \end{pmatrix}$$

Jestliže pro primární, resp. primárně-duální formulaci položíme

$$\mathbf{D}_I = \frac{1}{\sqrt{\mu}} \mathbf{S}_I, \quad \text{resp.} \quad \mathbf{D}_I = (\mathbf{S}_I \mathbf{U}_I^{-1})^{\frac{1}{2}} \tag{3.10}$$

(v praxi lze volit i jiná vyjádření matice $\mathbf{D}_I$), má výsledný systém tento tvar

$$\begin{pmatrix} \mathbf{B} & \mathbf{A} \\ \mathbf{A}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} d \\ d_u \end{pmatrix} = - \begin{pmatrix} g \\ h \end{pmatrix}, \tag{3.11}$$

kde

$$d, g \in \mathbb{R}^{n+m_I}, \quad \mathbf{B} \in \mathbb{R}^{(n+m_I)\times(n+m_I)}, \quad \mathbf{A} \in \mathbb{R}^{(n+m_I)\times(m_I+m_E)}, \quad d_u, h \in \mathbb{R}^{m_I+m_E},$$

přičemž

$$d = \begin{pmatrix} d_y \\ \mathbf{D}_I^{-1} d_s \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_{yy} & 0 \\ 0 & \mathbf{I} \end{pmatrix}, \quad g = \begin{pmatrix} g_y \\ \mathbf{D}_I g_s \end{pmatrix} \tag{3.12}$$

$$d_u = \begin{pmatrix} d_{u_I} \\ d_{u_E} \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} \mathbf{A}_I & \mathbf{A}_E \\ \mathbf{D}_I & 0 \end{pmatrix}, \quad h = \begin{pmatrix} h_I \\ h_E \end{pmatrix} \tag{3.13}$$

a pro primární, resp. primárně-duální formulaci je

$$\mathbf{D}_I g_s = \frac{1}{\sqrt{\mu}} \mathbf{S}_I \mathbf{U}_I e - \sqrt{\mu}\, e, \quad \text{resp.} \quad \mathbf{D}_I g_s = (\mathbf{S}_I \mathbf{U}_I)^{\frac{1}{2}} e - \mu (\mathbf{S}_I^{-1} \mathbf{U}_I^{-1})^{\frac{1}{2}} e. \tag{3.14}$$

Systém (3.11) má dimenzi $n+2m_I+m_E$. Tuto velikost lze zmenšit částečnou eliminací. Z druhé rovnice (3.8) plyne

$$-\mu \mathbf{S}_I^{-1} e + \mathbf{U}_I e = 0 \quad \Rightarrow \quad \mathbf{S}_I \mathbf{U}_I e = \mu e$$

a jestliže $\mu \to 0$, pak pro libovolný index $i \in I$ platí buď $u_i \to 0$ nebo $s_i \to 0$. Množinu omezení s nerovnostmi rozdělíme na aktivní a neaktivní podmnožinu. Podle bodu 1. definice 1.4 jsou aktivní ta omezení, která splňují $c_i(y) = 0,\ i \in I \cup E$. Pro odpovídající $i \in I$ tedy platí $s_i = 0$. Jelikož uvažujeme $s_i > 0\ \forall i \in I$, zobecníme definici aktivních a neaktivních omezení takto (ε_I je vhodná konstanta):

- Jestliže platí $s_i \le \varepsilon_I u_i,\ i \in I$, nazveme příslušná omezení aktivní a označíme je symbolem $\hat{.}$ spolu s příslušnými veličinami, tedy např. $\hat{c}_I(y), \hat{s}_I, \hat{h}_I, \hat{u}_I$. Jsou to ta omezení, pro která je $c_i(y),\ i \in I$, blízko nuly, přičemž $\hat{c}_I \in \mathbb{R}^{\hat{m}_I}$.
- Jestliže platí $s_i > \varepsilon_I u_i,\ i \in I$, nazveme příslušná omezení neaktivní a označíme je symbolem $\check{.}$ spolu s příslušnými veličinami, tedy např. $\check{c}_I(y), \check{s}_I, \check{h}_I, \check{u}_I$. Jsou to ta omezení, pro která je $u_i,\ i \in I$, blízko nuly, přičemž $\check{u}_I \in \mathbb{R}^{\check{m}_I}$, kde $\hat{m}_I + \check{m}_I = m_I$.

Soustava (3.11) má tedy tvar

$$\begin{pmatrix} \mathbf{B}_{yy} & 0 & 0 & \hat{\mathbf{A}}_I & \check{\mathbf{A}}_I & \mathbf{A}_E \\ 0 & \mathbf{I} & 0 & \hat{\mathbf{D}}_I & 0 & 0 \\ 0 & 0 & \mathbf{I} & 0 & \check{\mathbf{D}}_I & 0 \\ \hat{\mathbf{A}}_I^T & \hat{\mathbf{D}}_I & 0 & 0 & 0 & 0 \\ \check{\mathbf{A}}_I^T & 0 & \check{\mathbf{D}}_I & 0 & 0 & 0 \\ \mathbf{A}_E^T & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} d_y \\ \hat{\mathbf{D}}_I^{-1} \hat{d}_s \\ \check{\mathbf{D}}_I^{-1} \check{d}_s \\ \hat{d}_{u_I} \\ \check{d}_{u_I} \\ d_{u_E} \end{pmatrix} = - \begin{pmatrix} g_y \\ \hat{\mathbf{D}}_I \hat{g}_s \\ \check{\mathbf{D}}_I \check{g}_s \\ \hat{h}_I \\ \check{h}_I \\ h_E \end{pmatrix} \tag{3.15}$$

Z této soustavy vyloučíme neaktivní omezení. To samozřejmě není nutné, ale pro některé typy úloh to je dobré. Např. v primárně-duální formulaci je $\mathbf{D}_I = (\mathbf{S}_I \mathbf{U}_I^{-1})^{\frac{1}{2}}$. Prvky této matice mohou být neomezené, protože $u_i \to 0$, jestliže je i-té omezení $c_i(y)$ v bodě y neaktivní. Parametrem ε_I lze ovlivňovat míru toho, která omezení nazveme neaktivní. Z páté rovnice tedy plyne

$$\check{d}_s = -\check{\mathbf{A}}_I^T d_y - \check{h}_I = -\left(\check{\mathbf{A}}_I^T d_y + \check{h}_I\right) \tag{3.16}$$

a ze třetí rovnice

$$\check{\mathbf{D}}_I \check{d}_{u_I} = -\check{\mathbf{D}}_I^{-1} \check{d}_s - \check{\mathbf{D}}_I \check{g}_s \quad \Rightarrow \quad \check{d}_{u_I} = -\check{\mathbf{D}}_I^{-2} \check{d}_s - \check{g}_s.$$

Po dosazení (3.10), (3.14) a úpravě dostaneme pro primární metodu

$$\begin{aligned} \check{d}_{u_I} &= \mu \check{\mathbf{S}}_I^{-2} \left(\check{\mathbf{A}}_I^T d_y + \check{c}_I + \check{s}_I\right) - \sqrt{\mu}\, \check{\mathbf{S}}_I^{-1} \left(\frac{1}{\sqrt{\mu}} \check{\mathbf{S}}_I \check{\mathbf{U}}_I e - \sqrt{\mu}\, e\right) = \\ &= \mu \check{\mathbf{S}}_I^{-2} \left(\check{\mathbf{A}}_I^T d_y + \check{c}_I\right) + 2\mu \check{\mathbf{S}}_I^{-1} e - \check{\mathbf{U}}_I e \end{aligned} \tag{3.17}$$

a pro primárně-duální metodu

$$\begin{aligned} \check{d}_{u_I} &= \left(\check{\mathbf{S}}_I \check{\mathbf{U}}_I^{-1}\right)^{-1} \left(\check{\mathbf{A}}_I^T d_y + \check{c}_I + \check{s}_I\right) - \left(\check{\mathbf{S}}_I \check{\mathbf{U}}_I^{-1}\right)^{-\frac{1}{2}} \left[\left(\check{\mathbf{S}}_I \check{\mathbf{U}}_I\right)^{\frac{1}{2}} e - \mu \left(\check{\mathbf{S}}_I^{-1} \check{\mathbf{U}}_I^{-1}\right)^{\frac{1}{2}} e\right] = \\ &= \check{\mathbf{S}}_I^{-1} \check{\mathbf{U}}_I \left(\check{\mathbf{A}}_I^T d_y + \check{c}_I\right) + \mu \check{\mathbf{S}}_I^{-1} e. \end{aligned} \tag{3.18}$$

Do první rovnice (3.15) dosadíme za $\check{d}_{u_I}$ a dostaneme pro primární metodu

$$\mathbf{B}_{yy} d_y + \hat{\mathbf{A}}_I \hat{d}_{u_I} + \check{\mathbf{A}}_I \left[\mu \check{\mathbf{S}}_I^{-2} (\check{\mathbf{A}}_I^T d_y + \check{c}_I) + 2\mu \check{\mathbf{S}}_I^{-1} e - \check{\mathbf{U}}_I e\right] + \mathbf{A}_E d_{u_E} = -g_y,$$

neboli

$$\left(\mathbf{B}_{yy} + \mu\check{\mathbf{A}}_I\check{\mathbf{S}}_I^{-2}\check{\mathbf{A}}_I^T\right) d_y + \hat{\mathbf{A}}_I\hat{d}_{u_I} + \mathbf{A}_E d_{u_E} = -\left(g_y + \mu\check{\mathbf{A}}_I\check{\mathbf{S}}_I^{-2}\check{c}_I + 2\mu\check{\mathbf{A}}_I\check{\mathbf{S}}_I^{-1}e - \check{\mathbf{A}}_I\check{\mathbf{U}}_I e\right)$$

a pro primárně-duální metodu

$$\mathbf{B}_{yy}d_y + \hat{\mathbf{A}}_I\hat{d}_{u_I} + \check{\mathbf{A}}_I\left[\check{\mathbf{S}}_I^{-1}\check{\mathbf{U}}_I\left(\check{\mathbf{A}}_I^T d_y + \check{c}_I\right) + \mu\check{\mathbf{S}}_I^{-1}e\right] + \mathbf{A}_E d_{u_E} = -g_y,$$

neboli

$$\left(\mathbf{B}_{yy} + \check{\mathbf{A}}_I\check{\mathbf{S}}_I^{-1}\check{\mathbf{U}}_I\check{\mathbf{A}}_I^T\right) d_y + \hat{\mathbf{A}}_I\hat{d}_{u_I} + \mathbf{A}_E d_{u_E} = -\left(g_y + \check{\mathbf{A}}_I\check{\mathbf{S}}_I^{-1}\check{\mathbf{U}}_I\check{c}_I + \mu\check{\mathbf{A}}_I\check{\mathbf{S}}_I^{-1}e\right).$$

Po této eliminaci obsahuje systém (3.11) pouze aktivní omezení a má tento tvar

$$(3.19)\qquad \begin{pmatrix} \bar{\mathbf{B}}_{yy} & 0 & \hat{\mathbf{A}}_I & \mathbf{A}_E \\ 0 & \mathbf{I} & \hat{\mathbf{D}}_I & 0 \\ \hat{\mathbf{A}}_I^T & \hat{\mathbf{D}}_I & 0 & 0 \\ \mathbf{A}_E^T & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} d_y \\ \hat{\mathbf{D}}_I^{-1}\hat{d}_s \\ \hat{d}_{u_I} \\ d_{u_E} \end{pmatrix} = -\begin{pmatrix} \bar{g}_y \\ \hat{\mathbf{D}}_I\hat{g}_s \\ \hat{h}_I \\ h_E \end{pmatrix}$$

kde pro primární metodu je

$$\bar{\mathbf{B}}_{yy} = \mathbf{B}_{yy} + \mu\check{\mathbf{A}}_I\check{\mathbf{S}}_I^{-2}\check{\mathbf{A}}_I^T, \qquad \bar{g}_y = g_y + \mu\check{\mathbf{A}}_I\check{\mathbf{S}}_I^{-2}\check{c}_I + 2\mu\check{\mathbf{A}}_I\check{\mathbf{S}}_I^{-1}e - \check{\mathbf{A}}_I\check{\mathbf{U}}_I e,$$

$$\hat{\mathbf{D}}_I = \frac{1}{\sqrt{\mu}}\hat{\mathbf{S}}_I, \qquad \hat{\mathbf{D}}_I\hat{g}_s = \frac{1}{\sqrt{\mu}}\hat{\mathbf{S}}_I\hat{\mathbf{U}}_I e - \sqrt{\mu}\, e$$

a pro primárně-duální metodu

$$\bar{\mathbf{B}}_{yy} = \mathbf{B}_{yy} + \check{\mathbf{A}}_I\check{\mathbf{S}}_I^{-1}\check{\mathbf{U}}_I\check{\mathbf{A}}_I^T, \qquad \bar{g}_y = g_y + \check{\mathbf{A}}_I\check{\mathbf{S}}_I^{-1}\check{\mathbf{U}}_I\check{c}_I + \mu\check{\mathbf{A}}_I\check{\mathbf{S}}_I^{-1}e,$$

$$\hat{\mathbf{D}}_I = (\hat{\mathbf{S}}_I\hat{\mathbf{U}}_I^{-1})^{\frac{1}{2}}, \qquad \hat{\mathbf{D}}_I\hat{g}_s = (\hat{\mathbf{S}}_I\hat{\mathbf{U}}_I)^{\frac{1}{2}}e - \mu(\hat{\mathbf{S}}_I^{-1}\hat{\mathbf{U}}_I^{-1})^{\frac{1}{2}}e$$

Smysl eliminace neaktivních omezení je ten, že matice $\bar{\mathbf{B}}_{yy}$ a vektor $\bar{g}_y$ jsou omezené (předpokládáme, že původní matice $\mathbf{B}_{yy}$ a vektor g_y jsou omezené) a dimenze soustavy (3.19) je $n + 2\hat{m}_I + m_E$. Oproti systému (3.11) je tedy snížena o počet neaktivních omezení.

3.2 Použití metod s lokálně omezeným krokem

Metody vnitřních bodů lze realizovat buď jako metody spádových směrů nebo jako metody s lokálně omezeným krokem. V prvním případě je směrový vektor řešením systému (3.19) a délku kroku volíme tak, aby došlo k poklesu vhodné pokutové funkce. My použijeme druhý případ. Protože řešené podúlohy budou obsahovat pouze aktivní veličiny, budeme v § 3.2 symbol $\hat{}$ vynechávat (to odpovídá případu $\varepsilon_I = \infty$).

Uvažujme problém minimalizace kvadratické funkce vzhledem k lineárnímu omezení:

$$(3.20)\qquad \psi(d) = \frac{1}{2}d^T\mathbf{B}d + g^T d \;\rightarrow\; \min, \quad \mathbf{A}^T d + h = 0,$$

kde

$$d \in \mathbb{R}^{n+m_I}, \quad \mathbf{B} \in \mathbb{R}^{(n+m_I)\times(n+m_I)}, \quad g \in \mathbb{R}^{n+m_I}, \quad \mathbf{A} \in \mathbb{R}^{(n+m_I)\times(m_I+m_E)}, \quad h \in \mathbb{R}^{m_I+m_E},$$

přičemž

$$d = \begin{pmatrix} d_y \\ \mathbf{D}_I^{-1} d_s \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \bar{\mathbf{B}}_{yy} & 0 \\ 0 & \mathbf{I} \end{pmatrix}, \quad g = \begin{pmatrix} \bar{g}_y \\ \mathbf{D}_I g_s \end{pmatrix}, \tag{3.21}$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_I & \mathbf{A}_E \\ \mathbf{D}_I & 0 \end{pmatrix}, \quad h = \begin{pmatrix} h_I \\ h_E \end{pmatrix} \tag{3.22}$$

Označíme-li $d_u = [d_{u_I}^T, d_{u_E}^T]^T \in \mathbb{R}^{m_I+m_E}$ Lagrangeův multiplikátor, má Lagrangeova funkce pro problém (3.20) tvar

$$\mathcal{L}(d, d_u) = \psi(d) + d_u^T(\mathbf{A}^T d + h).$$

Dále použijeme větu 1.1. Pro řešení $d_\star$ problému (3.20) existuje vektor $d_{u\star} \in \mathbb{R}^{m_I+m_E}$ takový, že platí

$$\nabla_d \mathcal{L}(d_\star, d_{u\star}) = 0, \quad \nabla_{d_u} \mathcal{L}(d_\star, d_{u\star}) = 0.$$

Pokud použijeme iterační metodu podobně jako u původního problému (3.5) a aplikujeme-li na tyto rovnice Newtonovu metodu, dostaneme soustavu (3.19). Vidíme, že oba problémy – nalezení směrových vektorů pro problém (3.5) a problém (3.20) – jsou ekvivalentní, neboť vedou na stejnou soustavu rovnic, kterou můžeme napsat v obecném tvaru

$$\begin{pmatrix} \mathbf{B} & \mathbf{A} \\ \mathbf{A}^T & 0 \end{pmatrix} \cdot \begin{pmatrix} d \\ d_u \end{pmatrix} = - \begin{pmatrix} g \\ h \end{pmatrix} \tag{3.23}$$

Původní problém nalezení (aktivních) směrových vektorů jsme tedy převedli na problém minimalizace kvadratické funkce, na který aplikujeme metodu s lokálně omezeným krokem. Po přidání omezení

$$\|d\| \le \Delta \tag{3.24}$$

nám vznikne vedle lineárního omezení z (3.20) ještě další podmínka. Obě podmínky však mohou být nekompatibilní (norma řešení (3.20), kterou neznáme, může být větší než Δ), jak je vidět na obrázku 3.1. Z tohoto důvodu budeme lokálně omezený krok hledat ve dvou krocích jako součet vertikálního a horizontálního kroku d_V a d_H. Tato myšlenka pochází od Byrda a Omojokuna [12] a jejím smyslem je udělat obě omezení kompatibilní při současném zajištění dostatečného poklesu funkce $\psi(d)$. Budeme tedy uvažovat řešení $d_\star$ ve tvaru $d_\star = d_V + d_H$.

Uvažujme nejprve tuto minimalizační úlohu

$$\|\mathbf{A}^T d + h\| \to \min, \quad \|d\| \le \delta\Delta \tag{3.25}$$

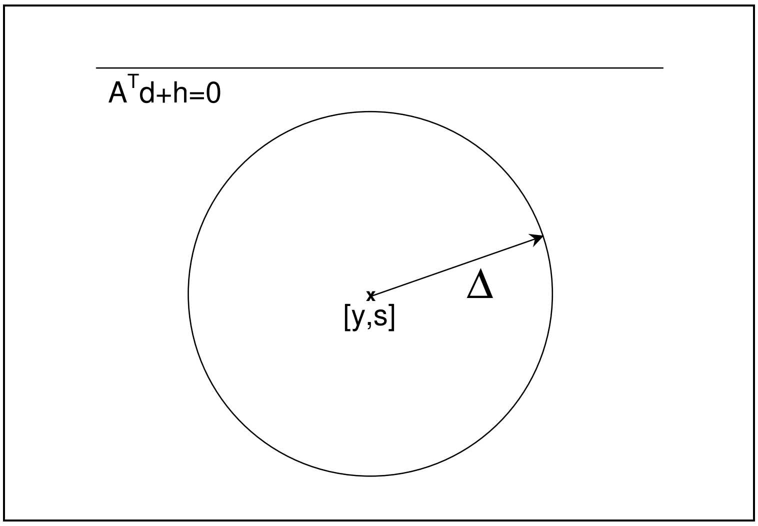
pro $0 < \delta < 1$ (např. $\delta = 0.8$). Protože platí

$$\|\mathbf{A}^T d + h\|^2 = d^T \mathbf{A}\mathbf{A}^T d + 2h^T \mathbf{A}^T d + h^T h,$$

je úloha (3.25) ekvivalentní úloze

$$\psi_V(d) = \frac{1}{2} d^T \mathbf{A}\mathbf{A}^T d + h^T \mathbf{A}^T d \to \min, \quad \|d\| \le \delta\Delta. \tag{3.26}$$

Zde se nevyskytuje lineární omezení, řešení tohoto problému najdeme metodou s lokálně omezeným krokem a označíme ho d_V jako vertikální krok směrového vektoru d. Tento krok tedy minimalizuje normu lineárního omezení v oblasti menší než Δ.



Obrázek 3.1: Nekompatibilita obou omezení

K získání d_V jako řešení problému (3.26) použijeme metodu psí nohy, § 2.3, kde místo $\mathbf{A}$ uvažujeme $\mathbf{A}\mathbf{A}^T$ a místo g uvažujeme $\mathbf{A}h$. Cauchyho d_C a Newtonův d_N krok spočítáme podle (2.36):

$$d_C = -\frac{h^T\mathbf{A}^T\mathbf{A}h}{h^T\mathbf{A}^T\mathbf{A}\mathbf{A}^T\mathbf{A}h}\,\mathbf{A}h = -\frac{\|\mathbf{A}h\|^2}{\|\mathbf{A}^T\mathbf{A}h\|^2}\,\mathbf{A}h.$$

Newtonův krok d_N splňuje normální rovnici $\mathbf{A}\mathbf{A}^T d_N + \mathbf{A}h = 0$. Dále předpokládáme, že matice $\mathbf{A}_E$ má lineárně nezávislé sloupce a tudíž i matice $\mathbf{A} = \begin{pmatrix} \mathbf{A}_I & \mathbf{A}_E \\ \mathbf{D}_I & 0 \end{pmatrix}$ má lineárně nezávislé sloupce. Tedy $\mathbf{A}w = 0 \;\Leftrightarrow\; w = 0$. Existuje proto inverze $(\mathbf{A}^T\mathbf{A})^{-1}$. Nechť

$$\mathbf{Z} \in \mathbb{R}^{(n+m_I)\times(n-m_E)} \tag{3.27}$$

je matice, jejíž sloupce tvoří bázi nulového prostoru matice $\mathbf{A}^T$, takže $\mathbf{A}^T\mathbf{Z} = 0$ a uvažujme Newtonův krok ve tvaru $d_N = \mathbf{A}u + \mathbf{Z}v$. Pak platí

$$0 = \mathbf{A}\mathbf{A}^T(\mathbf{A}u + \mathbf{Z}v) + \mathbf{A}h = \mathbf{A}(\mathbf{A}^T\mathbf{A}u + h) \quad \Rightarrow$$

$$\Rightarrow \quad \mathbf{A}^T\mathbf{A}u + h = 0 \quad \Rightarrow \quad u = -(\mathbf{A}^T\mathbf{A})^{-1}h.$$

Za vektor v zvolíme nulu, aby norma vektoru d_N byla minimální. Cauchyho a Newtonův krok mají tedy tvar

$$d_C = -\frac{\|\mathbf{A}h\|^2}{\|\mathbf{A}^T\mathbf{A}h\|^2}\,\mathbf{A}h, \quad d_N = -\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}h. \tag{3.28}$$

Nyní ukážeme vlastnosti těchto kroků, které použijeme v algoritmu 2.5. Platí

$$0 \le \|\mathbf{A}^T d_C + h\|^2 = \|\mathbf{A}^T d_C\|^2 + 2h^T\mathbf{A}^T d_C + \|h\|^2 = \frac{\|\mathbf{A}h\|^4}{\|\mathbf{A}^T\mathbf{A}h\|^2} - \frac{2\|\mathbf{A}h\|^4}{\|\mathbf{A}^T\mathbf{A}h\|^2} + \|h\|^2$$

$$\Leftrightarrow \quad \|h\|^2 \ge \frac{\|\mathbf{A}h\|^4}{\|\mathbf{A}^T\mathbf{A}h\|^2},$$

což není nic jiného než Schwarzova nerovnost. Platí-li ovšem rovnost

$$0 = \|\mathbf{A}^T d_C + h\|^2 \quad \Leftrightarrow \quad \|\mathbf{A}^T \mathbf{A} h\| \|h\| = \|\mathbf{A} h\|^2, \tag{3.29}$$

musí být vektory h a $\mathbf{A}^T \mathbf{A} h$ lineárně závislé, takže

$$\mathbf{A}^T \mathbf{A} h = \alpha h \quad \Rightarrow \quad (\mathbf{A}^T \mathbf{A})^{-1} h = \frac{1}{\alpha}\, h,$$

kde

$$\alpha = \frac{\|\mathbf{A}^T \mathbf{A} h\|}{\|h\|} = \frac{\|\mathbf{A}^T \mathbf{A} h\|^2}{\|h\| \|\mathbf{A}^T \mathbf{A} h\|} = \frac{\|\mathbf{A}^T \mathbf{A} h\|^2}{\|\mathbf{A} h\|^2}$$

podle (3.29). Odtud

$$d_N = -\mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} h = -\frac{1}{\alpha}\, \mathbf{A} h = -\frac{\|\mathbf{A} h\|^2}{\|\mathbf{A}^T \mathbf{A} h\|^2}\, \mathbf{A} h = d_C.$$

Takže

$$d_C \neq d_N \quad \Rightarrow \quad 0 < \|\mathbf{A}^T d_C + h\|^2 \quad \Rightarrow \quad \|h\|^2 > \frac{\|\mathbf{A} h\|^4}{\|\mathbf{A}^T \mathbf{A} h\|^2}.$$

Odtud plyne

$$\|d_C\| = \frac{\|\mathbf{A} h\|^3}{\|\mathbf{A}^T \mathbf{A} h\|^2} \cdot \frac{\|\mathbf{A} h\| \|\mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} h\|}{\|\mathbf{A} h\| \|\mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} h\|} \leq \frac{\|\mathbf{A} h\|^4 \|d_N\|}{\|\mathbf{A}^T \mathbf{A} h\|^2 \|h\|^2} < \|d_N\|$$

a dále

$$(d_N - d_C)^T d_C = \frac{\|\mathbf{A} h\|^2}{\|\mathbf{A}^T \mathbf{A} h\|^2} \|h\|^2 - \frac{\|\mathbf{A} h\|^6}{\|\mathbf{A}^T \mathbf{A} h\|^4} = \frac{\|\mathbf{A} h\|^2}{\|\mathbf{A}^T \mathbf{A} h\|^2} \left(\|h\|^2 - \frac{\|\mathbf{A} h\|^4}{\|\mathbf{A}^T \mathbf{A} h\|^2} \right) > 0,$$

takže iterace podle algoritmu 2.5 mají tento jednoduchý tvar:

- jestliže $\|d_C\| \geq \delta\Delta$, položíme

$$d_V = \frac{\delta\Delta}{\|d_C\|}\, d_C$$

- jestliže $\|d_N\| \leq \delta\Delta$, položíme

$$d_V = d_N$$

- ve zbývajícím případě $\|d_C\| < \delta\Delta < \|d_N\|$ položíme

$$d_V = d_C + \kappa(d_N - d_C),$$

  kde $\kappa > 0$ je zvoleno tak, aby $\|d_V\| = \delta\Delta$.

Lemma 3.1 *Pro vertikální krok d_V platí* $\|\mathbf{A}^T d_V + h\| < \|h\|$.

DŮKAZ: Provede se dosazením všech tří případů pro d_V a využitím vztahu

$$\|\mathbf{A}^T d_V + h\|^2 = \|\mathbf{A}^T d_V\|^2 + 2h^T \mathbf{A}^T d_V + \|h\|^2.$$

□

Nechť d_V je získaný vertikální krok. Nyní přeformulujeme problém (3.20)-(3.24) takto:

$$\psi_H(d) = \frac{1}{2}\, d^T \mathbf{B} d + g^T d \;\rightarrow\; \min, \quad \mathbf{A}^T d = \mathbf{A}^T d_V, \quad \|d\| \le \Delta. \tag{3.30}$$

Tato nová formulace obsahuje kompatibilní omezení, neboť volba $d = d_V$ splňuje obě podmínky. Celkový krok d problému (3.30) lze napsat ve tvaru $d = d_V + d_H$. Protože požadujeme $\mathbf{A}^T d = \mathbf{A}^T d_V$, musí krok d_H splňovat $\mathbf{A}^T d_H = 0$. Je-li $\mathbf{Z}$ matice z (3.27), pak $d_H = \mathbf{Z} d_Z$ pro nějaké $d_Z \in \mathbb{R}^{n-m_E}$ a protože d_V leží v oboru hodnot matice $\mathbf{A}$ podle (3.28), tj. $d_V = \mathbf{A} w$ pro nějaký vektor w, a $\mathbf{A}^T \mathbf{Z} = 0$, platí

$$d_V^T d_H = w^T \mathbf{A}^T \mathbf{Z} d_Z = 0.$$

Odtud dostaneme podle Pythagorovy věty

$$\Delta^2 \ge \|d\|^2 = \|d_V\|^2 + \|d_H\|^2 \quad \Rightarrow \quad \|d_H\| = \|\mathbf{Z} d_Z\| \le \sqrt{\Delta^2 - \|d_V\|^2}.$$

Dosadíme do (3.30):

$$\begin{aligned} \psi_H(d) &= \frac{1}{2}(d_V + \mathbf{Z} d_Z)^T \mathbf{B}(d_V + \mathbf{Z} d_Z) + g^T (d_V + \mathbf{Z} d_Z) = \\ &= \frac{1}{2}\, d_Z^T \mathbf{Z}^T \mathbf{B} \mathbf{Z} d_Z + (\mathbf{B} d_V + g)^T \mathbf{Z} d_Z + \frac{1}{2}\, d_V^T \mathbf{B} d_V + g^T d_V \end{aligned}$$

a dostaneme pro d_Z úlohu najít minimum kvadratické funkce

$$\psi_Z(d) = \frac{1}{2}\, d^T \mathbf{B}_Z d + g_Z^T d \;\rightarrow\; \min, \quad \|\mathbf{Z} d\| \le \bar{\Delta}, \tag{3.31}$$

kde

$$\mathbf{B}_Z = \mathbf{Z}^T \mathbf{B} \mathbf{Z}, \quad g_Z = \mathbf{Z}^T (\mathbf{B} d_V + g), \quad \bar{\Delta} = \sqrt{\Delta^2 - \|d_V\|^2},$$

a jejíž řešení $d_H = \mathbf{Z} d_Z$ nazveme horizontálním krokem směrového vektoru d. Zde se opět nevyskytuje lineární omezení podobně jako u vertikálního kroku, avšak oblast přípustných d_Z nemusí být sférická. Použijeme proto předpodmíněnou metodu sdružených gradientů, § 2.5, pro $\mathbf{C} = \mathbf{Z}^T \mathbf{Z}$. Z (3.27) plyne, že problém (3.31) je velikosti $n - m_E$ a iterace vypadají podle algoritmu 2.9 takto:

$$d_Z = 0, \qquad r_Z = g_Z, \qquad \tilde{r}_Z = (\mathbf{Z}^T \mathbf{Z})^{-1} r_Z, \qquad p_Z = -\tilde{r}_Z,$$

$$\eta = p_Z^T \mathbf{B}_Z p_Z, \qquad \alpha = \frac{1}{\eta}\, r_Z^T \tilde{r}_Z, \qquad d_Z^+ = d_Z + \alpha p_Z,$$

$$r_Z^+ = r_Z + \alpha \mathbf{B}_Z p_Z, \qquad \tilde{r}_Z^+ = (\mathbf{Z}^T \mathbf{Z})^{-1} r_Z^+, \qquad \beta = \frac{(r_Z^+)^T \tilde{r}_Z^+}{r_Z^T \tilde{r}_Z}, \qquad p_Z^+ = -\tilde{r}_Z^+ + \beta p_Z.$$

Tyto iterace zahrneme do původního problému (3.20) pro $d = d_V + \mathbf{Z} d_Z$ a místo omezení $\|\mathbf{Z} d_Z\| \le \bar{\Delta}$ budeme uvažovat omezení v původním tvaru $\|d\| \le \Delta$. Z rekurence pro d_Z plyne

$$d^+ = d_V + \mathbf{Z} d_Z^+ = d_V + \mathbf{Z} d_Z + \alpha \mathbf{Z} p_Z \equiv d + \alpha p,$$

takže můžeme položit $p = \mathbf{Z}p_Z$. Dále

$$r_Z = \mathbf{B}_Z d_Z + g_Z = \mathbf{Z}^T\mathbf{B}\mathbf{Z}d_Z + \mathbf{Z}^T(\mathbf{B}d_V + g) = \mathbf{Z}^T(\mathbf{B}d - \mathbf{B}d_V + \mathbf{B}d_V + g) = \mathbf{Z}^T r$$

a pro předpodmíněná residua zavedeme vektor $\tilde{r} = \mathbf{Z}\tilde{r}_Z$. V takto upravené metodě sdružených gradientů se vyskytuje násobení

$$\tilde{r} = \mathbf{Z}\tilde{r}_Z = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T r \stackrel{\text{def}}{=} \mathbf{P}_Z r,$$

kde

$$\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T.$$

Nyní je třeba se zbavit matice $\mathbf{Z}$, která se obtížně určuje, [20]. Projekci $\mathbf{P}_Z$ můžeme nahradit ekvivalentní projekcí

$$\mathbf{P}_A = \mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T,$$

která nevyžaduje matici $\mathbf{Z}$ a použijeme násobení $\tilde{r} = \mathbf{P}_A r$. Že se jedná o ekvivalentní matice, plyne ze zjištění, že

$$\mathbf{P}_Z\mathbf{A} = \mathbf{P}_A\mathbf{A} = 0 \quad \text{a} \quad \mathbf{P}_Z\mathbf{Z} = \mathbf{P}_A\mathbf{Z} = \mathbf{Z},$$

takže pro libovolný vektor w, který lze napsat ve tvaru $w = \mathbf{A}w_1 + \mathbf{Z}w_2$, platí

$$\mathbf{P}_Z w = \mathbf{Z}w_2 = \mathbf{P}_A w.$$

Matice $\mathbf{A}$ má lineárně nezávislé sloupce, proto je $\mathbf{A}^T\mathbf{A}$ regulární a existuje inverze.

Iterace pro původní problém (3.20) mají tedy tvar

$$\text{počáteční} \quad d_Z = 0 \quad \Rightarrow \quad \text{počáteční} \quad d = d_V, \qquad r = \mathbf{B}d + g, \qquad \tilde{r} = \mathbf{P}_A r,$$

$$p = \mathbf{Z}p_Z = -\mathbf{Z}\tilde{r}_Z = -\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T r = -\mathbf{P}_Z r = -\mathbf{P}_A r = -\tilde{r},$$

$$\eta = p_Z^T\mathbf{B}_Z p_Z = p_Z^T\mathbf{Z}^T\mathbf{B}\mathbf{Z}p_Z = p^T\mathbf{B}p, \qquad \alpha = \frac{1}{\eta} r_Z^T\tilde{r}_Z = \frac{1}{\eta} r^T\mathbf{Z}\tilde{r}_Z = \frac{1}{\eta} r^T\tilde{r},$$

$$d^+ = d + \alpha p, \qquad r^+ = r + \alpha\mathbf{B}p, \qquad \tilde{r}^+ = \mathbf{P}_A r, \qquad \beta = \frac{(r_Z^+)^T\tilde{r}_Z^+}{r_Z^T\tilde{r}_Z} = \frac{(r^+)^T\tilde{r}^+}{r^T\tilde{r}},$$

$$p^+ = \mathbf{Z}p_Z^+ = -\mathbf{Z}\tilde{r}_Z^+ + \beta\mathbf{Z}p_Z = -\tilde{r}^+ + \beta p.$$

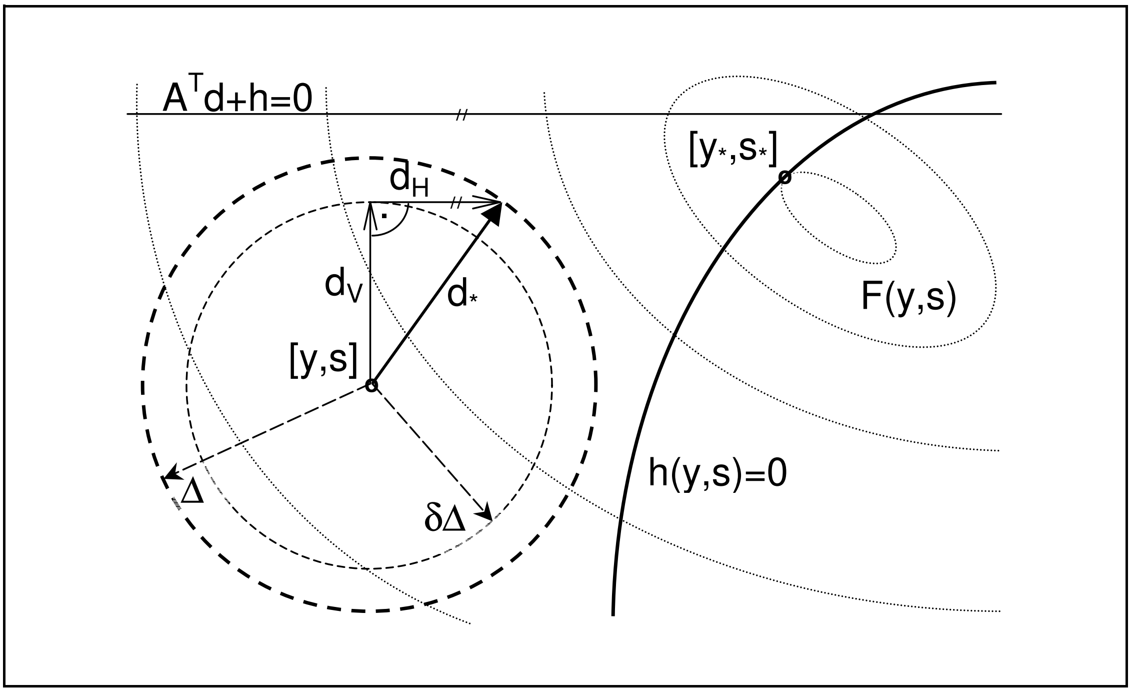
Z metody sdružených gradientů ovšem nelze získat Lagrangeovy multiplikátory d_u, takže je musíme vyjádřit jinak. Z první rovnice (3.23) plyne

$$\mathbf{A}d_u = -(g + \mathbf{B}d) = -r \quad \Rightarrow \quad d_u = -(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T r$$

jako řešení úlohy nejmenších čtverců. Navíc odtud plyne

$$\tilde{r} = \mathbf{P}_A r = [\mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T]r = r + \mathbf{A}d_u.$$

Na obrázku 3.2 je znázorněn způsob výpočtu směrového vektoru $d_\star$ pomocí vertikálního a horizontálního kroku. Vidíme vrstevnice funkce $F(y,s)$, omezení $h(y,s) = 0$ a řešení $[y_\star, s_\star]$. Vyjdeme z bodu $[y,s]$ a hledáme $d_\star = [d_y^T, (\mathbf{D}_I^{-1}d_s)^T]^T$. Protože může dojít k nekompatibilitě omezení $\mathbf{A}^T d + h = 0$ a $\|d\| \le \Delta$, uvažujeme $d_\star$ ve tvaru



Obrázek 3.2: Vertikální a horizontální krok

$d_\star = d_V + d_H$, kde d_V řeší úlohu (3.25) a $d_H = \mathbf{Z} d_Z$ úlohu (3.31). Vektor $d_\star$ však obsahuje jen aktivní omezení, $d_\star = [d_y^T, (\mathbf{D}_I^{-1} \hat{d}_s)^T]^T$, musíme proto dopočítat hodnoty $\check{d}_s$ pro neaktivní omezení podle 3.16. Položíme $d_s = (\hat{d}_s^T, \check{d}_s^T)^T$, $y^+ = y + \alpha_y d_y$, $s^+ = s + \alpha_s d_s$. Totéž provedeme s Lagrangeovými multiplikátory, $u^+ = u + \alpha_u d_u$, kde vektor d_u je rovněž rozdělen na $\hat{d}_u$ a $\check{d}_u$. Délky kroků α_s, α_u volíme tak, aby $s^+ > 0$, $u^+ > 0$.

Nyní zkompletujeme všechny uvedené úvahy pro $d_\star$ do nového algoritmu, který slouží k výpočtu lokálně omezeného kroku (*trust region method*) vzhledem k lineárnímu omezení (*constrained*), úloha (3.20).

Algoritmus 3.1 *Constrained trust region method.*

1. *Zvolíme* $\delta \in (0,1)$, *např.* $\delta = 0.8$ *a* $\varepsilon \in (0,1)$.

2. *Spočítáme* $d_C = -\frac{\|\mathbf{A}h\|^2}{\|\mathbf{A}^T\mathbf{A}h\|^2} \mathbf{A}h$ *a* $d_N = -\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}h$.

3. (a) *Jestliže* $\|d_C\| \geq \delta\Delta$, *položíme* $d = \frac{\delta\Delta}{\|d_C\|} d_C$;

   (b) *Pokud* $\|d_C\| < \delta\Delta < \|d_N\|$, *položíme* $d = d_C + \kappa(d_N - d_C)$, *kde* κ *je určeno tak, že* $\|d\| = \delta\Delta$;

   (c) *Ve zbývajícím případě* $\|d_N\| \leq \delta\Delta$ *položíme* $d = d_N$.

4. *Položíme* $r = \mathbf{B}d + g$, $d_u = -(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T r$, $\tilde{r} = r + \mathbf{A}d_u$, $p = -\tilde{r}$.

5. *Spočítáme* $\eta = p^T\mathbf{B}p$. *Je-li* $\eta > 0$, *přejdeme na krok 6. Jinak určíme* $\kappa > 0$ *tak, že* $\|d + \kappa p\| = \Delta$, *položíme* $d_\star = d + \kappa p$ *a přejdeme na krok 10.*

6. *Položíme* $\alpha = \frac{r^T\tilde{r}}{\eta}$ *a* $d^+ = d + \alpha p$.

(a) *Je-li* $\|d^+\| > \Delta$, *určíme* $\kappa > 0$ *tak, že* $\|d + \kappa p\| = \Delta$, *položíme* $d_\star = d + \kappa p$ *a přejdeme na krok 10.*

(b) *Je-li* $\|d^+\| = \Delta$, *položíme* $d_\star = d^+$ *a přejdeme na krok 10.*

(c) *Je-li* $\|d^+\| < \Delta$, *přejdeme na krok 7.*

7. *Spočítáme* $r^+ = r + \alpha \mathbf{B} p$, $d_u^+ = -(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T r^+$. *Je-li* $\|r^+\| \le \varepsilon \|g\|$, *položíme* $d_\star = d^+$, $d_{u\star} = d_u^+$ *a* `STOP`. *Jinak přejdeme na krok 8.*

8. *Položíme* $\tilde{r}^+ = r^+ + \mathbf{A} d_u^+$, $\beta = \frac{(r^+)^T \tilde{r}^+}{r^T \tilde{r}}$, $p^+ = -\tilde{r}^+ + \beta p$.

9. *Odstraníme index* $.^+$ *a návrat na krok 5.*

10. *Spočítáme* $r_\star = r + \kappa \mathbf{B} p$ *a položíme* $d_{u\star} = -(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T r_\star$.

3.3 Algoritmus

Metody s lokálně omezeným krokem vedou na krok $y^+ = y + \alpha_y d_y$, $s^+ = s + \alpha_s d_s$, kde buď $\alpha_y = 1$ a $\alpha_s \in (0, 1\rangle$ je takové, že $s^+ > 0$ (tzv. přijatelný krok) nebo $\alpha_y = \alpha_s = 0$ (tzv. nulový krok). Kromě toho volíme $u^+ = u + \alpha_u d_u$, kde $\alpha_u \in (0, 1\rangle$ je takové, že $u^+ > 0$. Vyjádřeno po složkách má tedy platit $s_i + \alpha_s d_{si} > 0$, $u_i + \alpha_u d_{ui} > 0 \ \forall i$. Spočítáme tedy podíly $-\frac{s_i}{d_{s\,i}}$, $-\frac{u_i}{d_{u\,i}}$, definujeme veličiny

$$\tilde{\alpha}_s = \vartheta \min_{i \in I,\, d_{s\,i} < 0} \left(-\frac{s_i}{d_{si}} \right), \qquad \tilde{\alpha}_u = \vartheta \min_{i \in I,\, d_{u\,i} < 0} \left(-\frac{u_i}{d_{ui}} \right), \tag{3.32}$$

kde $0 < \vartheta < 1$ je koeficient blízký k jednotce, a položíme

$$\alpha_y = \alpha, \quad \alpha_s = \min\{\alpha, \tilde{\alpha}_s\}, \quad \alpha_u = \min\{\alpha, \tilde{\alpha}_u\}.$$

Přijatelný krok znamená, že buď dojde k poklesu hodnoty $F(y^+, s^+)$ oproti $F(y, s)$ nebo ke zlepšení splnění omezení, tzn. $h(y^+, s^+)$ je blíže nule než $h(y, s)$. Za tím účelem definujeme následující pokutovou funkci $P(\alpha)$ s koeficientem $\sigma > 0$:

$$P(\alpha) = F(y + \alpha_y d_y, s + \alpha_s d_s) + (u + d_u)^T h(y + \alpha_y d_y, s + \alpha_s d_s) + \frac{\sigma}{2} \|h(y + \alpha_y d_y, s + \alpha_s d_s)\|^2.$$

Pro vektor u^+ uvažujeme plný krok $u + d_u$, aby v derivaci $P'(\alpha)$ nevystupoval člen s α_u a tato derivace se dala vyjádřit pomocí vztahů (3.12)-(3.13). Spočítáme skutečný a předpověděný pokles funkce $P(\alpha)$. Skutečný pokles je definován jako rozdíl $P(1) - P(0)$. Jestliže

$$Q(\alpha) = P(0) + \alpha P'(0) + \frac{\alpha^2}{2} d^T \mathbf{B} d$$

je kvadratická aproximace funkce $P(\alpha)$, pak je rozdíl

$$Q(1) - Q(0) = P'(0) + \frac{1}{2} d^T \mathbf{B} d$$

předpověděný pokles funkce $P(\alpha)$. Abychom rozhodli, zda je krok přijatelný či nikoli, vytvoříme podíl skutečného a předpověděného poklesu. Jestliže

$$\varrho \stackrel{\text{def}}{=} \frac{P(1) - P(0)}{Q(1) - Q(0)} > 0,$$

je krok přijatelný a poloměr Δ můžeme zvětšit. V opačném případě je krok nepřijatelný (nulový) a poloměr Δ je třeba snížit. Aby $\varrho > 0$, je nutnou podmínkou pro aplikaci metod s lokálně omezeným krokem splnění nerovnosti $Q(1) - Q(0) < 0$.

Věta 3.1 *Nechť*

$$r_u = \mathbf{A}^T d + h, \quad \theta = h^T(h - r_u) \quad a \quad \gamma = d^T g + d^T \mathbf{A} d_u.$$

Jestliže

$$\theta > 0 \quad a \quad \sigma > \frac{\frac{1}{2}\, d^T \mathbf{B} d + \gamma}{\theta},$$

pak platí nerovnost $Q(1) - Q(0) < 0$.

Důkaz: Zderivujeme funkci $P(\alpha)$:

$$\begin{aligned} P'(\alpha) &= d_y^T \nabla_y F(y + \alpha_y d_y, s + \alpha_s d_s) + d_s^T \nabla_s F(y + \alpha_y d_y, s + \alpha_s d_s) + \\ &+ (u + d_u)^T \left[d_y^T \nabla_y h(y + \alpha_y d_y, s + \alpha_s d_s) + d_s^T \nabla_s h(y + \alpha_y d_y, s + \alpha_s d_s) \right]^T + \\ &+ \sigma h^T(y + \alpha_y d_y, s + \alpha_s d_s) \cdot \\ &\cdot \left[d_y^T \nabla_y h(y + \alpha_y d_y, s + \alpha_s d_s) + d_s^T \nabla_s h(y + \alpha_y d_y, s + \alpha_s d_s) \right]^T \end{aligned}$$

$$\begin{aligned} \Rightarrow \quad P'(0) &= d_y^T \nabla_y f(y) - \mu d_s^T \mathbf{S}_I^{-1} e + (u + d_u)^T \left(d_y^T [\mathbf{A}_I(y), \mathbf{A}_E(y)] + d_s^T [\mathbf{I}, 0] \right)^T + \\ &+ \sigma h^T(y, s) \left(d_y^T [\mathbf{A}_I(y), \mathbf{A}_E(y)] + d_s^T [\mathbf{I}, 0] \right)^T . \end{aligned}$$

Dále dosadíme vztahy (3.12)-(3.13):

$$\begin{aligned} P'(0) &= \begin{pmatrix} d_y \\ \mathbf{D}_I^{-1} d_s \end{pmatrix}^T \begin{pmatrix} \nabla_y f(y) + [\mathbf{A}_I, \mathbf{A}_E]\, u \\ -\mu \mathbf{D}_I \mathbf{S}_I^{-1} e + [\mathbf{D}_I, 0]\, u \end{pmatrix} + \begin{pmatrix} d_y \\ \mathbf{D}_I^{-1} d_s \end{pmatrix}^T \begin{pmatrix} \mathbf{A}_I & \mathbf{A}_E \\ \mathbf{D}_I & 0 \end{pmatrix} d_u + \\ &+ \sigma \begin{pmatrix} d_y \\ \mathbf{D}_I^{-1} d_s \end{pmatrix}^T \begin{pmatrix} \mathbf{A}_I & \mathbf{A}_E \\ \mathbf{D}_I & 0 \end{pmatrix} h(y, s) = d^T g + d^T \mathbf{A} d_u + \sigma\, d^T \mathbf{A} h. \end{aligned} \tag{3.33}$$

Celkem platí

$$Q(1) - Q(0) = \frac{1}{2}\, d^T \mathbf{B} d + d^T g + d^T \mathbf{A} d_u + \sigma\, d^T \mathbf{A} h = \frac{1}{2}\, d^T \mathbf{B} d + \gamma + \sigma\, d^T \mathbf{A} h.$$

Dále

$$\mathbf{A}^T d + h = \begin{pmatrix} \hat{\mathbf{A}}_I^T & \hat{\mathbf{D}}_I & 0 \\ \check{\mathbf{A}}_I^T & 0 & \check{\mathbf{D}}_I \\ \mathbf{A}_E^T & 0 & 0 \end{pmatrix} \begin{pmatrix} d_y \\ \hat{\mathbf{D}}_I^{-1} \hat{d}_s \\ \check{\mathbf{D}}_I^{-1} \check{d}_s \end{pmatrix} + \begin{pmatrix} \hat{h}_I \\ \check{h}_I \\ h_E \end{pmatrix} = \hat{\mathbf{A}}^T \hat{d} + \hat{h},$$

protože druhá rovnice je rovna nule podle (3.16) – neaktivní omezení eliminujeme ze soustavy (3.15) a počítáme přesně. Ve vertikálním kroku hledáme vektor d_V, který minimalizuje normu $\|\hat{\mathbf{A}}^T \hat{d} + \hat{h}\|$ v oblasti $\|\hat{d}\| \le \delta\Delta < \Delta$, platí tedy $\|\hat{\mathbf{A}}^T d_V + \hat{h}\| < \|\hat{h}\|$ podle lemmatu 3.1. V horizontálním kroku platí $\hat{\mathbf{A}}^T \hat{d} + \hat{h} = \hat{\mathbf{A}}^T d_V + \hat{h}$ a proto

$$\|r_u\| = \|\mathbf{A}^T d + h\| = \|\hat{\mathbf{A}}^T \hat{d} + \hat{h}\| = \|\hat{\mathbf{A}}^T d_V + \hat{h}\| < \|\hat{h}\| \le \sqrt{\|\hat{h}\|^2 + \|\check{h}\|^2} = \|h\|.$$

Odtud plyne

$$|h^T r_u| \le \|h\|\|r_u\| = h^T h \, \frac{\|r_u\|}{\|h\|} < h^T h \quad \Rightarrow \quad \theta = h^T(h - r_u) > 0.$$

Dále platí

$$d^T \mathbf{A} h = h^T(\mathbf{A}^T d + h - h) = h^T(r_u - h) = -\theta$$

a tedy

$$Q(1) - Q(0) = \frac{1}{2}\, d^T \mathbf{B} d + \gamma - \sigma\,\theta.$$

Nyní pro libovolné spočítané vektory d a d_u stačí zvolit hodnotu $\sigma > \frac{\frac{1}{2} d^T \mathbf{B} d + \gamma}{\theta}$ takovou, aby platilo $Q(1) - Q(0) < 0$. □

V praxi se zkoušely i jiné pokutové funkce, např.

$$P_1(\alpha) = F(y + \alpha_y d_y, s + \alpha_s d_s) + \sigma \|h(y + \alpha_y d_y, s + \alpha_s d_s)\|.$$

Pro tuto funkci lze dokázat globální konvergenci algoritmu, pokud se parametr σ během iteračního procesu nezmenšuje. Tato monotonní strategie však může vést na nevhodně velké hodnoty σ, které mohou zpomalit konvergenci. U funkce $P(\alpha)$ můžeme volit malé hodnoty parametru σ, proto je tato pokutová funkce vhodná pro výpočty. Pro tuto funkci však nelze globální konvergenci v obecném případě dokázat.

Parametr μ měníme v každé iteraci. Většina implementací metod vnitřních bodů volí hodnotu μ tak, že $0 < \mu < \frac{s^T u_I}{m_I}$, tedy $\mu = \nu \frac{s^T u_I}{m_I}$ pro $\nu \in (0,1)$. V praxi se ukázalo, že algoritmus pracuje nejlépe tehdy, když jdou složky $s_i u_i$ stejnoměrně k nule. K tomuto účelu zavedeme veličinu

$$\omega = \frac{\min_{i \in I}\{s_i u_i\}}{s^T u_I / m_I}$$

Zřejmě platí $0 < \omega \le 1$ a $\omega = 1$ právě když je $s_i u_i$ konstantní $\forall i \in I$. Nyní použijeme následující heuristiku pro volbu ν a tedy μ zavedenou v [65], která se v praxi ukázala jako velmi efektivní:

$$\mu = \nu \, \frac{s^T u_I}{m_I}, \quad \text{kde} \quad \nu = 10^{-1} \cdot \min\left\{\frac{1 - \omega}{20\omega}, 2\right\}^3$$

Je nutno poznamenat, že pomalý pokles μ může značně zvýšit celkový počet iterací a naopak rychlý pokles μ může vést na selhání metody.

Poloměr Δ měníme v závislosti na hodnotě podílu ϱ. Je-li ϱ blízko nuly, Δ zmenšíme, je-li ϱ blízko jedničky nebo větší než jedna, Δ zvětšíme.

Nyní již můžeme uvést kompletní algoritmus pro řešení obecného nelineárního problému s omezeními ve tvaru rovností a nerovností metodou vnitřních bodů (*interior point method*) realizovanou jako metody s lokálně omezeným krokem (*trust region method*).

Algoritmus 3.2 *Trust region interior point method.*

Data: $f : \mathbb{R}^n \to \mathbb{R}$, $c = [c_I, c_E] : \mathbb{R}^n \to \mathbb{R}^m$.
Zvolíme $y \in \mathbb{R}^n$, $s \in \mathbb{R}^{m_I}$, $u \in \mathbb{R}^m$, $\mu > 0$, $\vartheta \in (0,1)$, $\varepsilon, \varepsilon_I \in (0,1)$, *např.*

$$s = u = 10^{-1} \cdot (1, \ldots, 1)^T, \quad \mu = 1, \quad \vartheta = 0.95, \quad \varepsilon = 10^{-6}, \quad \varepsilon_I = 10^{-1}.$$

Zvolíme $0 < \underline{\delta} < 1 < \overline{\delta}$ *a* $0 < \underline{\varrho} < \overline{\varrho} < 1$.

1. *Určíme* $\mathbf{A}_I(y)$, $\mathbf{A}_E(y)$, $g(y,s,u)$, $h(y,s)$.

2. *Jestliže* $(\mu < \varepsilon \;\;\&\;\; \|g\| < \varepsilon \;\;\&\;\; \|h\| < \varepsilon)$, *pak* `STOP`.

3. *Pomocí diferencí spočítáme druhé derivace funkcí* $f(y)$ *a* $c(y)$ *a položíme* $\Delta = \|g\|$.

4. *Určíme aktivní a neaktivní omezení* $(s_i \le \varepsilon_I u_i,$ *resp.* $s_i > \varepsilon_I u_i)$.

5. *Pomocí algoritmu 3.1 spočítáme vektory* $d_y, \hat{\mathbf{D}}_I^{-1}\hat{d}_s, \hat{d}_{u_I}, d_{u_E}$.

6. *Vypočítáme vektory* $\check{d}_s, \check{d}_{u_I}$, *tím získáme* d_s, d_u *a položíme* $d = [d_y^T, (\mathbf{D}_I^{-1}d_s)^T]^T$.

7. *Zvolíme* $\sigma > 0$ *tak, aby* $\sigma > -\frac{\frac{1}{2}d^T\mathbf{B}d + d^Tg + d^T\mathbf{A}d_u}{d^T\mathbf{A}h}$.

8. *Podle (3.32) určíme maximální délku kroku* $\tilde{\alpha}_s > 0$ *tak, aby* $s + \tilde{\alpha}_s d_s > 0$ *a položíme* $\alpha_s = \min\{1, \tilde{\alpha}_s\}$.

9. *Vypočítáme podíl* $\varrho = \frac{P(1) - P(0)}{Q(1) - Q(0)}$.

10. *Je-li* $\varrho \le 0$, *zvolíme* $\Delta < \|d\|$ *a návrat na krok 5.*

11. *Podle (3.32) určíme maximální délku kroku* $\tilde{\alpha}_u > 0$ *tak, aby* $u + \tilde{\alpha}_u d_u > 0$ *a položíme* $\alpha_u = \min\{1, \tilde{\alpha}_u\}$.

12. *Položíme* $y := y + d_y$, $s := s + \alpha_s d_s$, $u := u + \alpha_u d_u$.

13. *Je-li* $\varrho < \underline{\varrho}$, *položíme* $\Delta := \underline{\delta}\,\|d\|$, *je-li* $\overline{\varrho} < \varrho$, *položíme* $\Delta := \overline{\delta}\,\Delta$.

14. *Spočítáme* $\omega = \frac{\min_{i \in I}\{s_i u_i\}}{s^T u_I / m_I}$, $\nu = 10^{-1} \cdot \min\left\{\frac{1-\omega}{20\omega}, 2\right\}^3$ *a položíme* $\mu = \nu\,\frac{s^T u_I}{m_I}$.

15. *Návrat na krok 1.*

3.4 Použití filtru

Jestliže podle postupu výše spočítáme kandidáta na nové přiblížení, tedy krok

$$y^+ = y + \alpha_y d_y, \quad s^+ = s + \alpha_s d_s,$$

pak zpravidla používáme jistou pokutovou funkci, pomocí které zjišťujeme, zda je tento krok přijatelný, tzn. dojde buď k poklesu hodnoty funkce F nebo ke zlepšení splnění omezení h. Vraťme se zpět k problému (3.5), tj. problému

$$F(y,s) \;\to\; \min, \quad h(y,s) = 0.$$

Řešení tohoto problému vyžaduje dva cíle. Minimalizovat funkci F a splnit jisté podmínky. Označíme-li $\chi(y,s) = \|h(y,s)\|$, pak v podstatě hledáme řešení problému

$$F(y,s) \;\to\; \min, \quad \chi(y,s) = 0. \tag{3.34}$$

Jedná se tedy o minimalizaci dvou funkcí, přičemž funkce $\chi(y,s)$ má přednost, neboť hledáme takové řešení, ve kterém je tato funkce rovna nule.

Fletcher a Leyffer [15] poskytli techniku, která se vyhýbá určitým obtížím při použití klasické pokutové funkce, např. aktualizaci pokutového parametru σ. Přijatelnost kroku je určena srovnáním chování omezení a hodnoty funkce F s předchozími iteracemi sdruženými v tzv. filtru. Nová iterace je přijatelná, jestliže se ve srovnání se všemi iteracemi uloženými v daném filtru dostatečně zlepší omezení nebo hodnota funkce F.

Budeme používat označení

$$F^{\mu_k} \equiv F(y,s) = f(y) - \mu_k e^T \ln(\mathbf{S}_I)e, \quad \chi \equiv \chi(y,s)$$

pro obecný bod (y,s) a

$$F_i^{\mu_k} \equiv F(y_i,s_i) = f(y_i) - \mu_k e^T \ln(\mathbf{S}_I)_i\, e, \quad \chi_i \equiv \chi(y_i,s_i)$$

pro i-tou iteraci (y_i,s_i), kde $(\mathbf{S}_I)_i$ značí matici $\mathbf{S}_I$ v iteraci i.

Definice 3.1 *Řekneme, že iterace (y_i,s_i) a příslušná dvojice $(F_i^{\mu_k},\chi_i)$ dominují nad iterací (y_j,s_j) a dvojicí $(F_j^{\mu_k},\chi_j)$ právě když platí*

$$F_i^{\mu_k} \le F_j^{\mu_k} \quad \& \quad \chi_i \le \chi_j, \quad i \ne j,$$

což je ekivalentní zápisu

$$\max\{F_i^{\mu_k} - F_j^{\mu_k}, \chi_i - \chi_j\} \le 0,$$

pro nějaké pevné μ_k.

Tato definice říká, že iterace (y_i,s_i) je nejméně tak dobrá jako iterace (y_j,s_j) (vzhledem k oběma mírám) a tedy iterace (y_j,s_j) ztrácí význam. Požadavkem pro přijetí nové iterace tedy je, aby nad ní nedominovala některá z předchozích iterací. Stačí si tedy pamatovat ty iterace, nad kterými nedominuje žádná jiná. K tomuto účelu vytvoříme strukturu zvanou filtr, kterou použijeme jako kriterium pro přijetí či odmítnutí iterace.

Definice 3.2 *Filtr $\mathcal{F}_k$ je množina dvojic $\{(F_i^{\mu_k},\chi_i) : i \in \{0,\dots,k\}\}$ takových, že žádná dvojice nedominuje nad jinou. Řekneme, že dvojice $(F_{k+1}^{\mu_k},\chi_{k+1})$ je přijatelná do filtru $\mathcal{F}_k$, jestliže nad ní nedominuje žádná jiná ve filtru $\mathcal{F}_k$. Množina $\mathcal{I}_k$ je množina indexů $i \in \{0,\dots,k\}$ takových, že $(F_i^{\mu_k},\chi_i) \in \mathcal{F}_k$. Množina*

$$\mathcal{D}_k = \{(F^{\mu_k},\chi) : \; F^{\mu_k} > F_i^{\mu_k} \;\text{ a }\; \chi > \chi_i \;\text{ pro nějaké }\; i \in \mathcal{I}_k\} \subset \mathbb{R}^2$$

je množina všech nepřijatelných dvojic v iteraci $k+1$.

V praxi to znamená, že pokud máme v k-té iteraci filtr $\mathcal{F}_k$, ptáme se, zda můžeme přijmout dvojici $(F_{k+1}^{\mu_k},\chi_{k+1})$, tedy $(k+1)$-ní iteraci. Pokud nad touto dvojicí nedominuje žádná jiná z filtru $\mathcal{F}_k$, přijmeme tuto novou iteraci a do filtru $\mathcal{F}_k$ přidáme dvojici $(F_{k+1}^{\mu_k},\chi_{k+1})$. V opačném případě dvojici $(F_{k+1}^{\mu_k},\chi_{k+1})$ do filtru $\mathcal{F}_k$ nepřidáme. Iterace $k+1$ je tedy přijatelná, jestliže platí

$$(F_{k+1}^{\mu_k},\chi_{k+1}) \notin \mathcal{D}_k.$$

Protože při přechodu ke $(k+1)$-ní iteraci aktualizujeme parametr μ_k na μ_{k+1}, nahradíme hodnoty $F_i^{\mu_k}$ aktualizovanými hodnotami $F_i^{\mu_{k+1}}$ pro $i \in \mathcal{I}_k$ a eventuelně $i = k+1$. Dále odstraníme ty dvojice $(F_i^{\mu_{k+1}},\chi_i)$, nad kterými dominuje jiná dvojice ve filtru. Tím dostaneme nový filtr $\mathcal{F}_{k+1}$ a nové množiny $\mathcal{I}_{k+1}$, $\mathcal{D}_{k+1}$.

Budeme používat tři různé fitry:

Barierový filtr – dvojice $(F_{k+1}^{\mu_k}, \chi_{k+1})$ je přijatelná do filtru $\mathcal{F}_k$, jestliže platí

$$F_{k+1}^{\mu_k} < F_i^{\mu_k} \quad \text{nebo} \quad \chi_{k+1} < \chi_i \quad \forall i \in \mathcal{I}_k.$$

Fletcherův-Leyfferův filtr – dvojice $(F_{k+1}^{\mu_k}, \chi_{k+1})$ je přijatelná do filtru $\mathcal{F}_k$, jestliže platí

$$f(y_{k+1}) < f(y_i) \quad \text{nebo} \quad \chi_{k+1} < \chi_i \quad \forall i \in \mathcal{I}_k.$$

V tomto případě neuvažujeme ve funkci F barierový člen.

Markovův filtr – dvojice $(F_{k+1}^{\mu_k}, \chi_{k+1})$ je přijatelná do filtru $\mathcal{F}_k$, jestliže platí

$$F_{k+1}^{\mu_k} < F_k^{\mu_k} \quad \text{nebo} \quad \chi_{k+1} < \chi_k.$$

Zde uvažujeme ke srovnání pouze předchozí iteraci, $\mathcal{I}_k = \{k\}$.

Pokud v iteraci $k+1$ pro nějakou délku kroku α_k nedostaneme dvojici přijatelnou do $\mathcal{F}_k$, můžeme buď použít pokutovou funkci nebo zmenšovat délku kroku α_k tak, aby příslušná dvojice $(k+1)$-ní iterace byla přijatelná do filtru. Tuto druhou variantu jsme vyzkoušeli v numerických testech.

Výše uvedenou jednoduchou definici přijetí či odmítnutí iterace je třeba upravit. Především zamezíme tomu, aby se do filtru dostaly dvojice, které jsou sice přijatelné do filtru, ale jsou libovolně blízko hranice tohoto filtru. Iterace $k+1$ je přijatelná, jestliže platí

$$\max\{\mathcal{A}_i,\ \chi_i - \chi_{k+1}\} > \varepsilon_F \chi_i \quad \forall i \in \mathcal{I}_k, \tag{3.35}$$

kde $\varepsilon_F > 0$ je nějaká konstanta a

- $\mathcal{A}_i = F_i^{\mu_{k+1}} - F_{k+1}^{\mu_{k+1}}$ pro barierový filtr;
- $\mathcal{A}_i = f(y_i) - f(y_{k+1})$ pro Fletcherův-Leyfferův filtr;
- $\mathcal{A}_i = F_k^{\mu_{k+1}} - F_{k+1}^{\mu_{k+1}}$ pro Markovův filtr.

Dále zavedeme horní mez na funkci $\chi(y,s)$ v podobě konstanty $\overline{\chi} \geq \chi_0$ a uvažujeme pouze takové body (y,s), pro které platí

$$\chi(y,s) \leq \overline{\chi}.$$

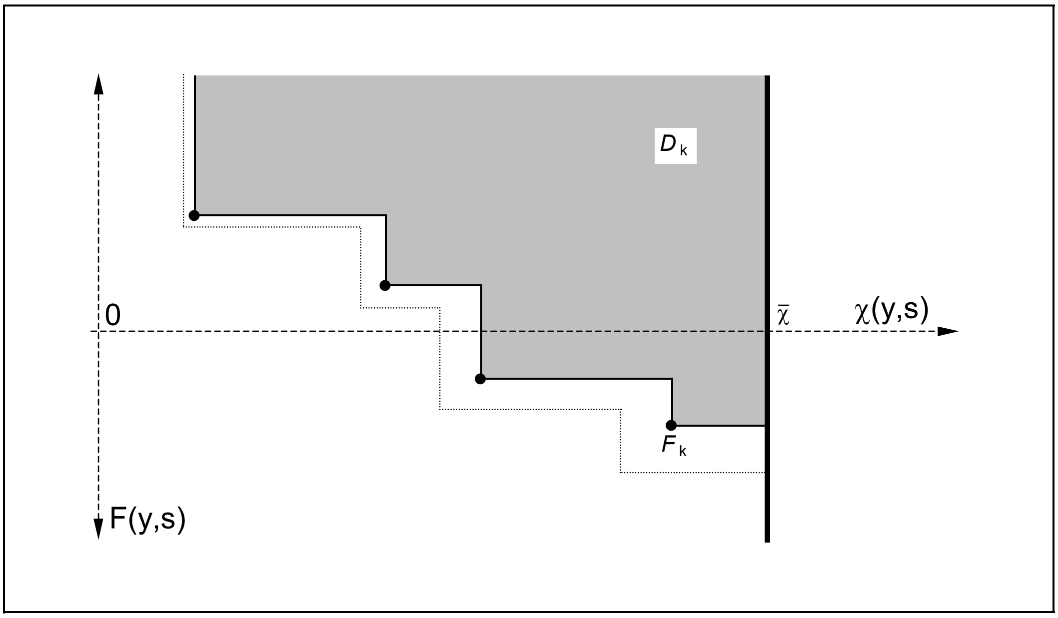
Inicializace filtru je taková, že po zvolení počáteční iterace (y_0, s_0) definujeme počáteční filtr jako

$$\mathcal{F}_0 = \{(F_0^{\mu_0}, \chi_0)\}.$$

Máme-li iteraci (y_k, s_k), která splňuje podmínku

$$\|\chi_k\| \leq \varepsilon_X,$$

kde $\varepsilon_X > 0$ je nějaká konstanta, lze požadovat, aby došlo také k dostatečnému poklesu v příslušné funkční hodnotě. Konečně můžeme stanovit, jaký maximální počet dvojic budeme do filtru ukládat, protože tento počet může být značně velký. V numerických testech jsme zvolili 50 dvojic. Po překročení tohoto počtu začneme z filtru odstraňovat nejstarší iterace.



Obrázek 3.3: Filtr $\mathcal{F}_k$ v iteraci k.

Na obrázku 3.3 je znázorněn filtr $\mathcal{F}_k$ v iteraci k. Černé puntíky jsou dvojice, které tvoří filtr, vybarvená oblast je množina $\mathcal{D}_k$ nepřijatelných dvojic v iteraci $k+1$, tlustá čára vpravo je horní mez $\overline{\chi}$ a tečkovaná křivka je upravená oblast přijatelnosti dvojice podle (3.35).

Nyní uvedeme algoritmus. Protože princip filtru lze snadněji implementovat ve spojení s metodou spádových směrů, použijeme pro určení směrových vektorů tuto metodu namísto metody s lokálně omezeným krokem. Protože provedeme porovnání principu filtru s klasickým postupem, uvedeme nejprve algoritmus používající pokutovou funkci, jehož detailní popis je uveden v [36].

Algoritmus 3.3 *Penalty line search interior point method.*

Data: $f : \mathbb{R}^n \to \mathbb{R}$, $c = [c_I, c_E] : \mathbb{R}^n \to \mathbb{R}^m$.
Zvolíme $y \in \mathbb{R}^n$, $s \in \mathbb{R}^{m_I}$, $u \in \mathbb{R}^m$, $\mu, \sigma, \overline{\Delta} > 0$, $\vartheta, \beta, \varepsilon, \varepsilon_I \in (0,1)$, *např.*

$$s = u = 10^{-1} \cdot (1, \ldots, 1)^T, \quad \mu = \sigma = 1, \quad \overline{\Delta} = 1000,$$

$$\vartheta = 0.95, \quad \beta = 0.5, \quad \varepsilon = 10^{-6}, \quad \varepsilon_I = 10^{-1}.$$

1. *Určíme* $\mathbf{A}_I(y)$, $\mathbf{A}_E(y)$, $g(y,s,u)$, $h(y,s)$.
2. *Jestliže* $(\mu < \varepsilon \;\;\&\;\; \|g\| < \varepsilon \;\;\&\;\; \|h\| < \varepsilon)$, *pak* `STOP`.
3. *Pomocí diferencí spočítáme druhé derivace funkcí* $f(y)$ *a* $c(y)$.
4. *Položíme* $\mathbf{D}_I = \mathbf{I}$, *určíme aktivní a neaktivní omezení* $(s_i \le \varepsilon_I u_i$, *resp.* $s_i > \varepsilon_I u_i)$ *a sestavíme lineární systém (3.19).*
5. *Určíme vektory* $d_y, \hat{d}_s, \hat{d}_{u_I}, d_{u_E}$ *nepřesným řešením systému (3.19) pomocí předpodmíněné metody sdružených gradientů.*

6. *Vypočteme vektory* $\check{d}_s, \check{d}_{u_I}$, *tím získáme* d_s, d_u *a položíme* $d = [d_y^T, d_s^T]^T$.

7. *Spočítáme směrovou derivaci* $P'(0)$ *podle (3.33).*

8. *Je-li* $P'(0) \geq 0$, *provedeme restart, tj. položíme* $\mathbf{B}_{yy} = \mathbf{I}$, *a návrat na krok 5.*

9. *Podle (3.32) určíme maximální délky kroku* $\tilde{\alpha}_s, \tilde{\alpha}_u > 0$ *tak, aby* $s + \tilde{\alpha}_s d_s > 0$ *a* $u + \tilde{\alpha}_u d_u > 0$ *a položíme* $\tilde{\alpha} = \min\{1, \frac{\overline{\Delta}}{\|d_y\|}\}$.

10. *Najdeme nejmenší číslo* $l \geq 0$ *takové, že* $P(\alpha) < P(0)$, *kde* $\alpha = \beta^l \tilde{\alpha}$, $\alpha_y = \alpha$, $\alpha_s = \min\{\alpha, \tilde{\alpha}_s\}$, $\alpha_u = \min\{\alpha, \tilde{\alpha}_u\}$.

11. *Položíme* $y := y + \alpha_y d_y$, $s := s + \alpha_s d_s$, $u := u + \alpha_u d_u$.

12. *Spočítáme* $\omega = \frac{\min_{i \in I}\{s_i u_i\}}{s^T u_I / m_I}$, $\nu = 10^{-1} \cdot \min\left\{\frac{1-\omega}{20\omega}, 2\right\}^3$ *a položíme* $\mu = \nu \frac{s^T u_I}{m_I}$.

13. *Návrat na krok 1.*

V případě použití principu filtru použijeme pokutovou funkci pouze k provádění restartů. Zavedeme označení

$$F_\alpha^\mu = F(y + \alpha_y d_y, s + \alpha_s d_s), \quad \chi_\alpha = \|h(y + \alpha_y d_y, s + \alpha_s d_s)\|$$

a pro jednoduchost vynecháme index k.

Algoritmus 3.4 *Filter line search interior point method.*

Data: $f : \mathbb{R}^n \to \mathbb{R}$, $c = [c_I, c_E] : \mathbb{R}^n \to \mathbb{R}^m$.
Zvolíme $y \in \mathbb{R}^n$, $s \in \mathbb{R}^{m_I}$, $u \in \mathbb{R}^m$, $\mu, \sigma, \overline{\Delta} > 0$, $\vartheta, \beta, \varepsilon, \varepsilon_I, \varepsilon_F \in (0, 1)$, *např.*

$$s = u = 10^{-1} \cdot (1, \ldots, 1)^T, \quad \mu = \sigma = 1, \quad \overline{\Delta} = 1000,$$

$$\vartheta = 0.95, \quad \beta = 0.5, \quad \varepsilon = 10^{-6}, \quad \varepsilon_I = 10^{-1}, \quad \varepsilon_F = 10^{-4}.$$

Položíme $\chi = \|h(y, s)\|$ *a* $\mathcal{F} = \{(F(y, s), \chi)\}$.

1. *Určíme* $\mathbf{A}_I(y)$, $\mathbf{A}_E(y)$, $g(y, s, u)$, $h(y, s)$.

2. *Jestliže* $(\mu < \varepsilon$ & $\|g\| < \varepsilon$ & $\|h\| < \varepsilon)$, *pak* `STOP`.

3. *Pomocí diferencí spočítáme druhé derivace funkcí* $f(y)$ *a* $c(y)$.

4. *Položíme* $\mathbf{D}_I = \mathbf{I}$, *určíme aktivní a neaktivní omezení* $(s_i \leq \varepsilon_I u_i$, *resp.* $s_i > \varepsilon_I u_i)$ *a sestavíme lineární systém (3.19).*

5. *Určíme vektory* $d_y, \hat{d}_s, \hat{d}_{u_I}, d_{u_E}$ *nepřesným řešením systému (3.19) pomocí předpodmíněné metody sdružených gradientů.*

6. *Vypočteme vektory* $\check{d}_s, \check{d}_{u_I}$, *tím získáme* d_s, d_u *a položíme* $d = [d_y^T, d_s^T]^T$.

7. *Spočítáme směrovou derivaci* $P'(0)$ *podle (3.33).*

8. *Je-li* $P'(0) \geq 0$, *provedeme restart, tj. položíme* $\mathbf{B}_{yy} = \mathbf{I}$, *a návrat na krok 5.*

9. *Podle (3.32) určíme maximální délky kroku* $\tilde{\alpha}_s, \tilde{\alpha}_u > 0$ *tak, aby* $s + \tilde{\alpha}_s d_s > 0$ *a* $u + \tilde{\alpha}_u d_u > 0$ *a položíme* $\tilde{\alpha} = \min\{1, \frac{\Delta}{\|d_y\|}\}$.

10. *Najdeme nejmenší číslo* $l \geq 0$ *takové, že* $(F_\alpha^\mu, \chi_\alpha) \notin \mathcal{D}$, *podmínka (3.35), kde* $\alpha = \beta^l \tilde{\alpha}, \quad \alpha_y = \alpha, \quad \alpha_s = \min\{\alpha, \tilde{\alpha}_s\}, \quad \alpha_u = \min\{\alpha, \tilde{\alpha}_u\}$.

11. *Položíme* $y := y + \alpha d_y, \ s := s + \alpha_s d_s, \ u := u + \alpha_u d_u$.

12. *Spočítáme* $\omega = \frac{\min_{i \in I}\{s_i u_i\}}{s^T u_I / m_I}, \ \nu = 10^{-1} \cdot \min\left\{\frac{1-\omega}{20\omega}, 2\right\}^3$ *a položíme* $\mu = \nu \frac{s^T u_I}{m_I}$.

13. *Aktualizujeme filtr* $\mathcal{F}$ *a návrat na krok 1.*

Aktualizace filtru $\mathcal{F}$ znamená, že do něj přidáme novou dvojici, přepočítáme hodnoty v něm obsažených dvojic po aktualizaci parametru μ, odstraníme ty dvojice, nad kterými dominuje nějaká jiná dvojice a pokud je celkový počet dvojic obsažených ve filtru větší než jistá hodnota (zvolili jsme 50), odstraníme nejstarší dvojici.

Poznamenejme, že oba uvedené algoritmy se liší pouze v krocích 10.-11., kde se provádí výběr délky kroku. Porovnání obou algoritmů je uvedeno v § 3.5.

3.5 Numerické výsledky

Algoritmy popsané v této kapitole byly implementovány v prostředí optimalizačního systému UFO [39] a testovány pomocí tří kolekcí rozsáhlých a strukturovaných testovacích problémů [36] (vždy 18 optimalizačních problémů s omezeními ve tvaru rovností a nerovností). Tyto problémy vznikly úpravou problémů s omezeními ve tvaru rovností uvedených v [42]. Rovnosti $c(y) = 0$ jsou v první množině nahrazeny nerovnostmi $c(y) \geq 0$ a ve druhé nerovnostmi $c(y) \leq 0$. Třetí množina obsahuje omezení $-1 \leq y \leq 1$ a $-1 \leq c(y) \leq 1$. Pro první a druhou množinu jsme z testů vynechali osmý problém, neboť spotřeboval více než polovinu celkového strojového času. Ve všech případech uvažujeme 1000 proměnných. Použité algoritmy jsou uvedeny v tabulce 3.2. U algoritmu 3.2 byly pro definici oblasti přijatelnosti vyzkoušeny dvě normy, jednak standardní euklidovská norma $\|d\|_2 \leq \Delta$ a také norma $\|d\|_1 = \sum_i |d_i| \leq \Delta$. Pro test jsme použili primárně-duální formulaci, protože primární formulace dávala horší výsledky. U algoritmu 3.3 byl rovněž testován případ, kdy se k výběru délky kroku nepoužívá žádná pokutová funkce, tzn. krok $y + d_y, \ s + \alpha_s d_s > 0, \ u + \alpha_u d_u > 0$ přijmeme vždy bez ohledu na splnění jakéhokoli kriteria.

Výsledky jsou uvedeny v tabulkách 3.3-3.5, kde jednotlivé sloupce znamenají

- Metoda – použitý algoritmus podle tabulky 3.2
- NIT – celkový počet hlavních iterací (pro y_k, s_k)
- NFV – celkový počet vyčíslení hodnoty funkce F
- NFG – celkový počet vyčíslení gradientu funkce F
- NCG – celkový počet iterací metody sdružených gradientů (vnitřní iterace)
- NR – celkový počet restartů

Zkratka	Použitý algoritmus	Pokutová funkce nebo filtr
TRIPM-1	3.2 s $\|d\|_1$	$P(\alpha)$
TRIPM-2	3.2 s $\|d\|_2$	$P(\alpha)$
PLSIPM-N	3.3	žádná
PLSIPM-P	3.3	$P(\alpha)$
FLSIPM-M	3.4	Markovův
FLSIPM-FL	3.4	Fletcherův-Leyfferův
FLSIPM-B	3.4	barierový

Tabulka 3.2: Přehled testovaných metod pro minimalizaci s omezeními.

- T – celkový čas
- NF – celkový počet problémů z dané množiny, které se nepodařilo vyřešit

Graficky jsou výsledky testů znázorněny na obrázcích 3.4-3.6. Z výsledků lze vyčíst, že nový algoritmus 3.2, metoda vnitřních bodů s lokálně omezeným krokem, není nejefektivnější. Je to způsobeno tím, že se pokutová funkce používá nejen k rozhodování o úspěšnosti kroku, ale také k určování poloměru oblasti přijatelnosti, což může vést k příliš rychlému zmenšování tohoto poloměru. Jinými slovy, kvadratický podproblém používající pokutovou funkci $P(\alpha)$ není vhodný k řízení délky kroku tak, jako podproblém používaný v případě minimalizace bez omezení a je třeba nalézt vhodnější model (některé jiné modely byly již vyzkoušeny, ale výsledky jsou ještě horší). To se týká zejména algoritmů používajících princip filtru, kdy kvadratický model nemůže vycházet z pokutové funkce. To je však otázka dalšího výzkumu. Porovnáme-li jednotlivé použité normy, nelze jednoznačně určit, který případ je lepší. Vidíme, že euklidovská norma potřebuje sice více hlavních iterací, vyžaduje však méne strojového času. Horší je pouze u třetí množiny příkladů. Navíc je třeba poznamenat, že metody vnitřních bodů s lokálně omezeným krokem nedokázaly spočítat vždy jeden příklad z dané kolekce.

Překvapivě dobře vychází metoda, při které nepoužíváme pokutovou funkci, což je vidět hlavně u třetí množiny příkladů, kde je malý celkový počet všech iterací i celkový strojový čas. U druhé množiny příkladů však dochází k výraznému nárůstu počtu iterací sdružených gradientů.

Velmi dobře si vedou metody používající princip filtru. Ve všech případech jsou nejrychlejší a přijatelný je i celkový počet hlavních iterací. Všimněme si výrazně nižšího celkového počtu iterací metody sdružených gradientů u první kolekce příkladů oproti metodám používajícím pokutovou funkci.

Co se týká typu filtru, nejsou patrné žádné větší rozdíly. Můžeme tedy používat nejjednodušší z nich, Markovův filtr, i když u třetí kolekce příkladů se spotřebovalo více strojového času, než u zbylých dvou typů filtru.

Závěrem je třeba upozornit, že algoritmy 3.2 a 3.4 jsou nové, takže je potřeba je otestovat na dalších příkladech a provést jejich možná vylepšení.

Informace o systému UFO a testovaných příkladech lze získat na adrese

http://www.cs.cas.cz/~luksan/test.html

Metoda	NIT	NFV	NFG	NCG	NR	T	NF
TRIPM-1	1 106	1 171	8 522	26 060	0	10.22	1
TRIPM-2	1 344	1 431	11 995	18 188	0	9.31	1
PLSIPM-N	567	567	4 137	24 969	20	4.94	0
PLSIPM-P	550	593	3 936	21 806	14	4.68	0
FLSIPM-M	608	650	4 326	13 152	9	3.97	0
FLSIPM-FL	602	720	4 280	13 336	8	3.93	0
FLSIPM-B	604	705	4 346	13 231	11	4.08	0

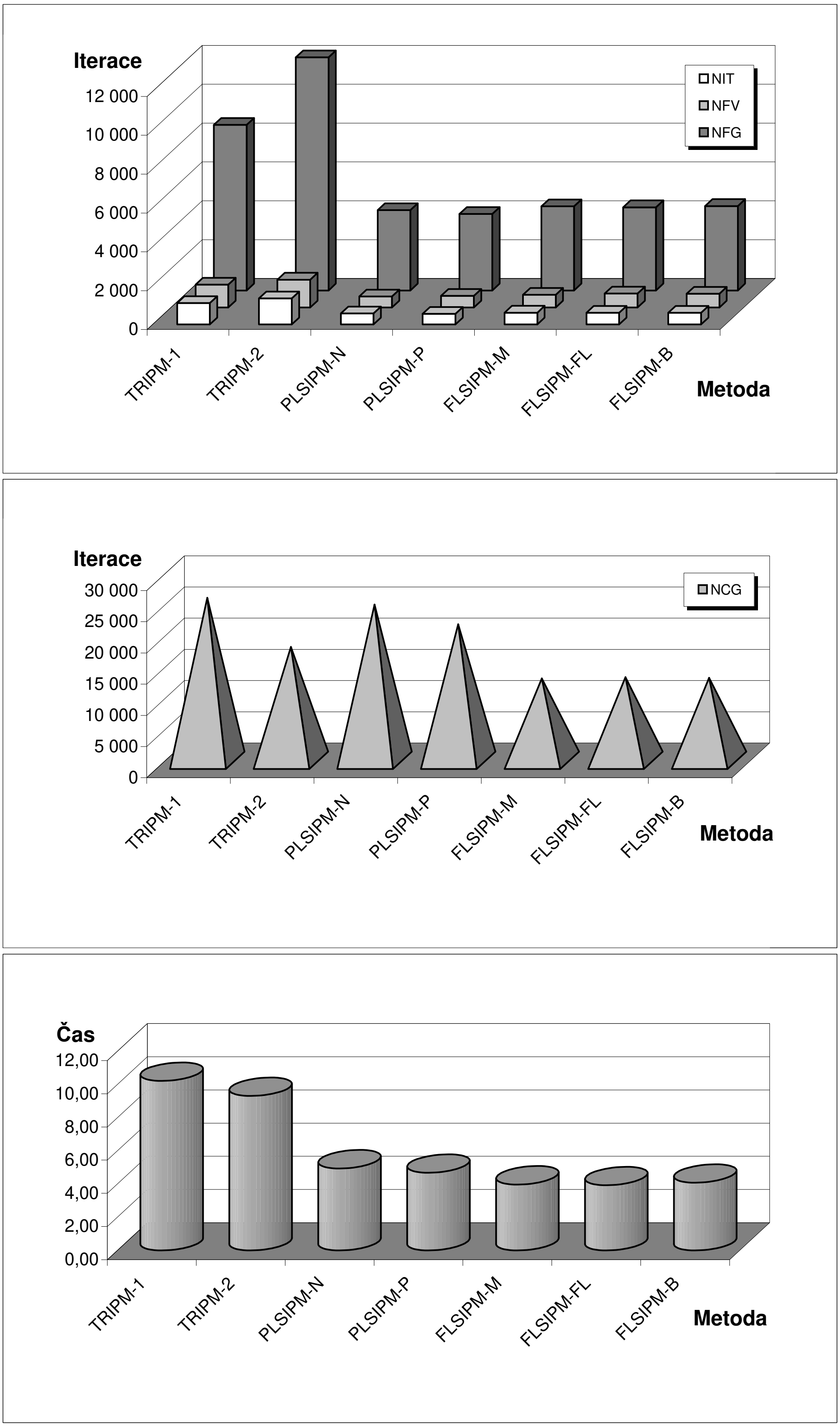
Tabulka 3.3: První množina příkladů pro minimalizaci s omezeními.

Metoda	NIT	NFV	NFG	NCG	NR	T	NF
TRIPM-1	904	998	6 185	10 521	0	6.52	1
TRIPM-2	906	941	6 048	10 448	0	5.82	1
PLSIPM-N	393	393	2 823	10 728	19	2.95	0
PLSIPM-P	476	925	3 403	5 654	73	3.32	1
FLSIPM-M	461	500	3 363	5 924	10	2.61	0
FLSIPM-FL	469	517	3 412	5 979	11	2.63	0
FLSIPM-B	464	508	3 381	5 941	10	2.59	0

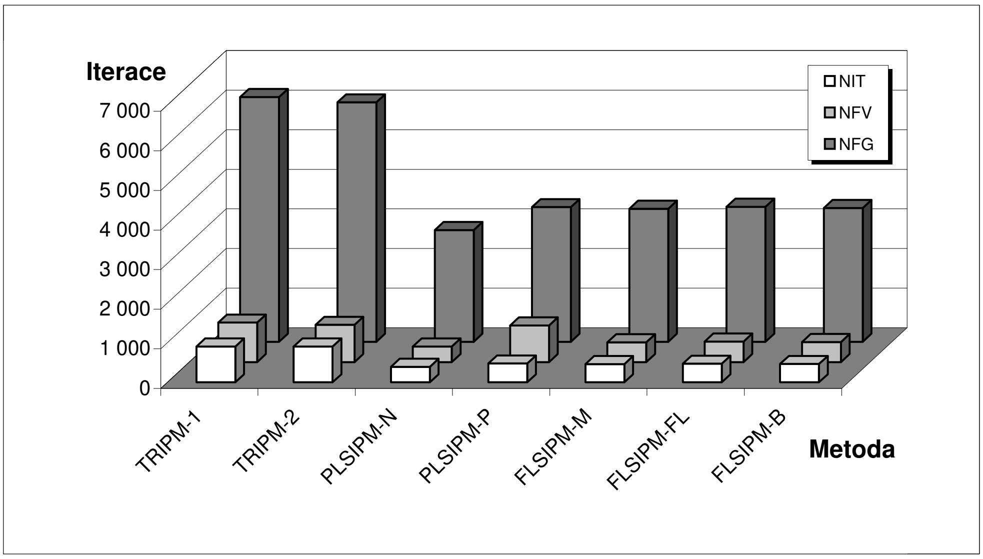
Tabulka 3.4: Druhá množina příkladů pro minimalizaci s omezeními.

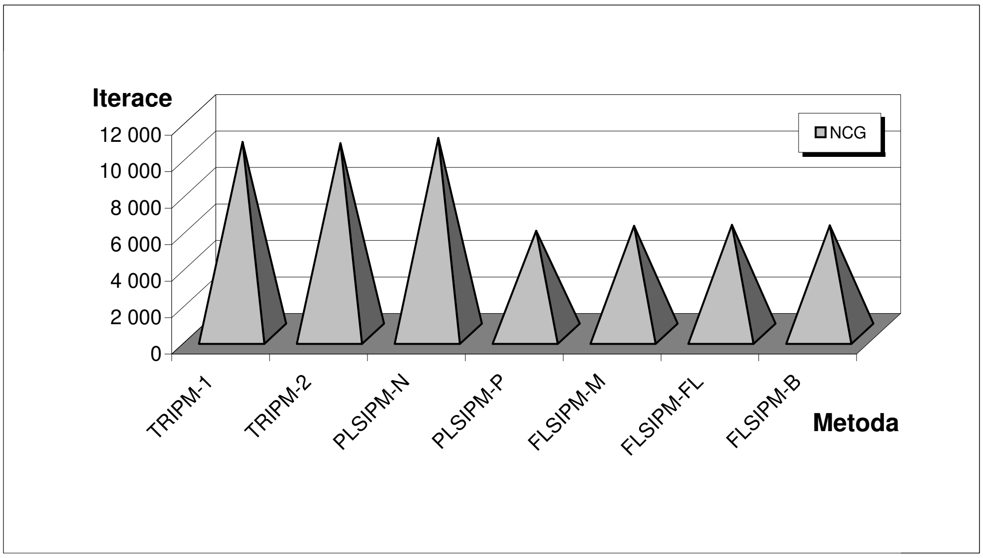
Metoda	NIT	NFV	NFG	NCG	NR	T	NF
TRIPM-1	1 040	1 128	6 878	6 923	0	13.63	1
TRIPM-2	1 459	1 514	9 705	7 485	0	15.98	1
PLSIPM-N	571	571	3 991	2 731	13	6.84	1
PLSIPM-P	572	699	4 294	3 670	25	7.55	0
FLSIPM-M	542	636	3 956	2 630	10	7.11	0
FLSIPM-FL	536	657	3 915	2 636	11	6.77	0
FLSIPM-B	541	632	3 946	2 629	10	6.78	0

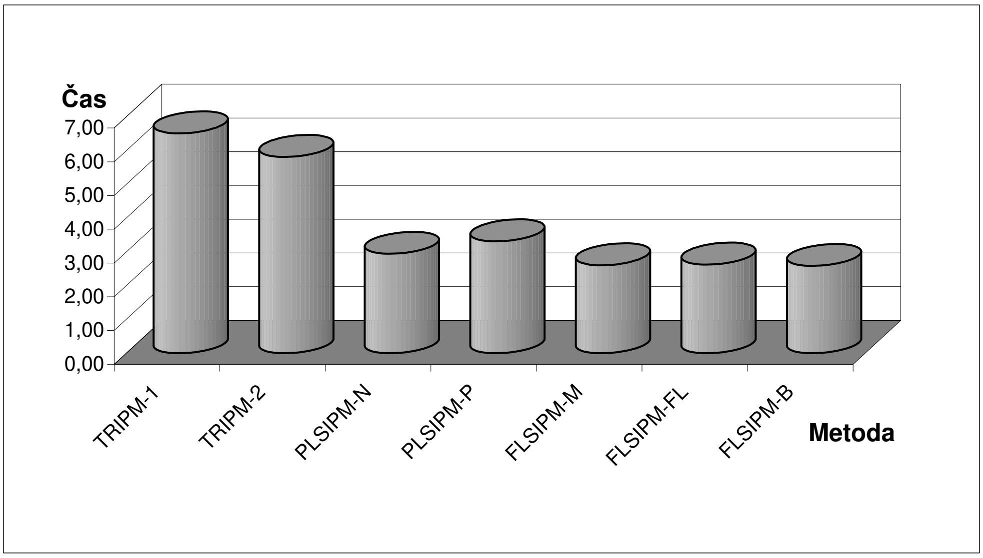
Tabulka 3.5: Třetí množina příkladů pro minimalizaci s omezeními.



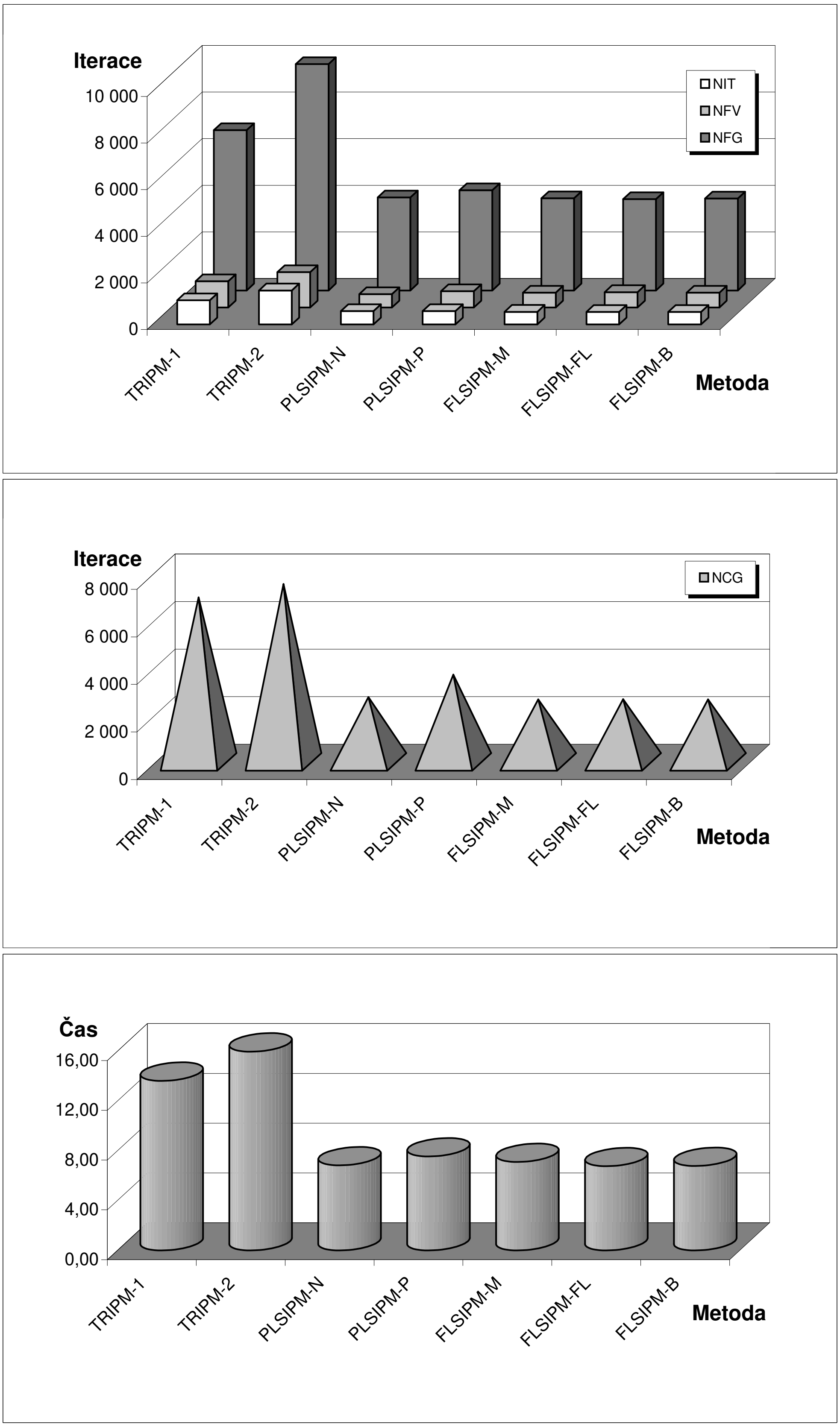
Obrázek 3.4: První množina příkladů pro minimalizaci s omezeními.







Obrázek 3.5: Druhá množina příkladů pro minimalizaci s omezeními.



Obrázek 3.6: Třetí množina příkladů pro minimalizaci s omezeními.

Příloha A

V příloze uvedeme některá pomocná tvrzení a poznámky z algebry, které jsou použity v této práci.

Poznámka A.1 *Označme symbolem* $\mathbf{M}^\dagger$ *Moore-Penroseovu pseudoinverzi matice* $\mathbf{M}$. *Tato pseudoinverze je určena jednoznačně a platí:*

$$\mathbf{M}\mathbf{M}^\dagger\mathbf{M} = \mathbf{M}, \quad \mathbf{M}^\dagger\mathbf{M}\mathbf{M}^\dagger = \mathbf{M}^\dagger, \quad (\mathbf{M}\mathbf{M}^\dagger)^T = \mathbf{M}\mathbf{M}^\dagger, \quad (\mathbf{M}^\dagger\mathbf{M})^T = \mathbf{M}^\dagger\mathbf{M}.$$

Pseudoinverze diagonální matice $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$ *je dána takto:*

$$\mathbf{D}^\dagger = \mathrm{diag}(\tilde{d}_1, \ldots, \tilde{d}_n),$$

kde

$$d_i \neq 0 \quad \Rightarrow \quad \tilde{d}_i = d_i^{-1}; \qquad d_i = 0 \quad \Rightarrow \quad \tilde{d}_i = 0.$$

Poznámka A.2 *Nechť* $\mathbf{A}$ *je symetrická matice řádu* n, λ_j *je libovolné vlastní číslo matice* $\mathbf{A}$ *a* $\mathcal{S}_j = \{q \neq 0 : \mathbf{A}q = \lambda_j q\}$ *je množina vlastních vektorů asociovaných s vlastním číslem* λ_j. *Označme* $\mathbf{P} = (\mathbf{A} - \lambda_j\mathbf{I})(\mathbf{A} - \lambda_j\mathbf{I})^\dagger$.

1. *Platí*

$$\mathbf{P}^2 = (\mathbf{A} - \lambda_j\mathbf{I})(\mathbf{A} - \lambda_j\mathbf{I})^\dagger(\mathbf{A} - \lambda_j\mathbf{I})(\mathbf{A} - \lambda_j\mathbf{I})^\dagger = (\mathbf{A} - \lambda_j\mathbf{I})(\mathbf{A} - \lambda_j\mathbf{I})^\dagger = \mathbf{P}$$

   a tedy $\mathbf{P}$ *je ortogonální projekce.*

2. *Nechť* $y \in \mathcal{R}(\mathbf{A} - \lambda_j\mathbf{I})$, *tzn. existuje* $w \in \mathbb{R}^n$, *že* $(\mathbf{A} - \lambda_j\mathbf{I})w = y$. *Platí*

$$\mathbf{P}y = (\mathbf{A} - \lambda_j\mathbf{I})(\mathbf{A} - \lambda_j\mathbf{I})^\dagger y = (\mathbf{A} - \lambda_j\mathbf{I})(\mathbf{A} - \lambda_j\mathbf{I})^\dagger(\mathbf{A} - \lambda_j\mathbf{I})w = (\mathbf{A} - \lambda_j\mathbf{I})w = y$$

   a tedy $\mathbf{P}$ *je projekce na prostor* $\mathcal{R}(\mathbf{A} - \lambda_j\mathbf{I})$.

3. *Nechť* $g \perp \mathcal{S}_j$. *Vektor* g *lze napsat ve tvaru* $g = \sum_{i=1}^n \vartheta_i q_i$, *kde* $q_1, \ldots, q_n$ *jsou vlastní vektory matice* $\mathbf{A}$ *příslušné vlastním číslům* $\lambda_1, \ldots, \lambda_n$. *Protože* $g \perp \mathcal{S}_j$, *je* $\vartheta_j = 0$. *Platí* $g \in \mathcal{R}(\mathbf{A} - \lambda_j\mathbf{I})$, *kde* $w = \sum_{i=1, i\neq j}^n \frac{\vartheta_i}{\lambda_i - \lambda_j}\, g$, *a tudíž* $\mathbf{P}g = g$.

Totéž platí i pro $\mathbf{P} = (\mathbf{A} - \lambda_j\mathbf{I})^\dagger(\mathbf{A} - \lambda_j\mathbf{I})$.

Uvažujme nyní Lanczosovu metodu popsanou v § 2.6, konkrétně vztah (2.63). Platí následující tvrzení o vztahu vlastních čísel matic $\mathbf{A} \in \mathbb{R}^{n\times n}$ a $\mathbf{T}_k \in \mathbb{R}^{(k+1)\times(k+1)}$, kde $k + 1 \leq n$.

Lemma A.1 *Buď (2.63) maticový tvar Lanczosovy metody (algoritmus 2.11). Nechť* $\mathbf{T}_k$ *je nedegenerovaná matice a* $\gamma_{k+1} = 0$. *Pak platí:*

1. *Nechť* $\{\lambda_i, w_i\}$ *je vlastní pár matice* $\mathbf{T}_k$. *Označme* $\mathbf{Q}_k w_i = z_i$. *Pak* $\{\lambda_i, z_i\}$ *je vlastní pár matice* $\mathbf{A}$ *a* $g^T z_i \neq 0$.

2. *Nechť* $\{\lambda_i, z_i\}$ *je vlastní pár matice* $\mathbf{A}$ *a* $g^T z_i \neq 0$. *Označme* $\mathbf{Q}_k^T z_i = w_i$. *Pak* $\{\lambda_i, w_i\}$ *je vlastní pár matice* $\mathbf{T}_k$.

DŮKAZ: Předně, protože $\gamma_{k+1} = 0$, platí $\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_k \mathbf{T}_k$ podle (2.63). Z toho, že $\mathbf{T}_k$ je nedegenerovaná, plyne $w_i^{(1)} \neq 0$ podle lemmatu 2.11 a $g = \gamma_0 q_0$ podle (2.65). Nyní:

1. Platí
$$\mathbf{T}_k w_i = \lambda_i w_i \quad \Rightarrow \quad \mathbf{Q}_k \mathbf{T}_k w_i = \lambda_i \mathbf{Q}_k w_i \quad \Rightarrow \quad \mathbf{A}\mathbf{Q}_k w_i = \lambda_i \mathbf{Q}_k w_i.$$
Dále
$$z_i = \mathbf{Q}_k w_i = (q_0, \ldots, q_k)(w_i^{(1)}, \ldots, w_i^{(k+1)})^T \neq 0,$$
protože $q_0, \ldots, q_k$ jsou lineárně nezávislé a $w_i^{(1)} \neq 0$. Tedy $\{\lambda_i, z_i\}$ je vlastní pár matice $\mathbf{A}$. Konečně
$$g^T z_i = \gamma_0 q_0^T \mathbf{Q}_k w_i = \gamma_0 w_i^{(1)} \neq 0.$$

2. Platí
$$\mathbf{A} z_i = \lambda_i z_i \quad \Rightarrow \quad \mathbf{Q}_k^T \mathbf{A} z_i = \lambda_i \mathbf{Q}_k^T z_i \quad \Rightarrow \quad \mathbf{T}_k \mathbf{Q}_k^T z_i = \lambda_i \mathbf{Q}_k^T z_i.$$
Dále
$$w_i = \mathbf{Q}_k^T z_i = (q_0, \ldots, q_k)^T z_i = (\frac{1}{\gamma_0} g^T z_i, \ldots) \neq 0,$$
tedy $\{\lambda_i, w_i\}$ je vlastní pár matice $\mathbf{T}_k$. □

Důsledkem je, že matice $\mathbf{T}_k$ má právě ta vlastní čísla λ_i, $i = 1, \ldots, k+1$, (bez uspořádání) matice $\mathbf{A}$, pro která platí $g^T z_i \neq 0$.

Dále uvedeme některé vlastnosti matic.

Lemma A.2 *Nechť jsou dány soustavy* $\mathbf{M}x = b$ *a* $(\mathbf{M} + \alpha \mathbf{I})y = b$, *kde* $\mathbf{M}$ *je symetrická a positivně semidefinitní matice,* $\alpha > 0$ *a* $b \neq 0$. *Pak platí:*
$$\|x\| > \|y\|.$$

DŮKAZ: Pro symetrickou matici $\mathbf{M}$ použijeme vlastní rozklad $\mathbf{Q}\mathbf{D}\mathbf{Q}^T$:
$$\mathbf{M}x = (\mathbf{M} + \alpha \mathbf{I})y \quad \Leftrightarrow \quad \mathbf{Q}\mathbf{D}\mathbf{Q}^T x = \mathbf{Q}\mathbf{D}\mathbf{Q}^T y + \alpha y.$$
Protože $\mathbf{Q}$ je ortonormální matice, platí $\|\bar{x}\| = \|\mathbf{Q}^T x\| = \|x\|$ a $\|\bar{y}\| = \|\mathbf{Q}^T y\| = \|y\|$. Máme tedy dokázat, že $\|\bar{x}\| > \|\bar{y}\|$, kde
$$\mathbf{Q}\mathbf{D}\bar{x} = \mathbf{Q}\mathbf{D}\bar{y} + \alpha \mathbf{Q}\bar{y} \quad \Leftrightarrow \quad \mathbf{D}\bar{x} = \mathbf{D}\bar{y} + \alpha \bar{y} \quad \Leftrightarrow \quad \mu_i \bar{x}_i = \mu_i \bar{y}_i + \alpha \bar{y}_i, \quad i = 1, \ldots, n,$$
kde μ_i jsou vlastní čísla matice $\mathbf{M}$.

1. Protože $\mathbf{M} \neq 0$, existuje alespoň jeden index j takový, že $\mu_j > 0$. Pro takový index j platí
$$\bar{x}_j = \bar{y}_j + \frac{\alpha}{\mu_j} \bar{y}_j > \bar{y}_j.$$

2. Jestliže existuje index k takový, že $\mu_k = 0$, pak platí

$$0 = \mu_k \bar{x}_k = \mu_k \bar{y}_k + \alpha \bar{y}_k = \alpha \bar{y}_k \quad \Rightarrow \quad \bar{y}_k = 0.$$

Spojením obou částí dostáváme požadovanou nerovnost $\|\bar{x}\| > \|\bar{y}\|$. □

Následující věty pojednávají o rozložení vlastních čísel dvou matic. První je Cauchyho věta, druhá je věta o monotonnosti, [28], [47].

Věta A.1 *Nechť* $\mathbf{N} = \begin{pmatrix} \mathbf{M} & \mathbf{U}^T \\ \mathbf{U} & \mathbf{V} \end{pmatrix} \in \mathbb{R}^{n\times n}$ *a* $\mathbf{M} \in \mathbb{R}^{m\times m}$ *jsou symetrické matice,* $m < n$. *Označme vlastní čísla matic* $\mathbf{N}$ *a* $\mathbf{M}$ *jako*

$$\nu_1 \le \nu_2 \le \ldots \le \nu_n \quad a \quad \mu_1 \le \mu_2 \le \ldots \le \mu_m.$$

Pak platí:

$$\nu_i \le \mu_i \le \nu_{i+n-m}, \quad i = 1, 2, \ldots, m$$

nebo ekvivalentně z druhé strany:

$$\mu_{j-n+m} \le \nu_j \le \mu_j, \quad j = 1, 2, \ldots, n, \quad kde \quad \mu_k = \begin{cases} -\infty & pro\ k < 1 \\ \infty & pro\ k > m \end{cases}$$

Specielně pro $m = n - 1$ *platí*

$$\nu_1 \le \mu_1 \le \nu_2 \le \mu_2 \le \ldots \le \mu_{n-1} \le \nu_n.$$

Důkaz: Viz např. [28], [47]. □

Věta A.2 *Nechť* $\mathbf{A} = \mathbf{M} + \mathbf{N}$, *kde* $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{n\times n}$ *jsou symetrické matice. Označme vlastní čísla matic* $\mathbf{A}, \mathbf{M}, \mathbf{N}$ *postupně jako*

$$\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n, \quad \mu_1 \le \mu_2 \le \ldots \le \mu_n, \quad \nu_1 \le \nu_2 \le \ldots \le \nu_n.$$

Pak pro $i, j = 1, \ldots, n$ *platí*

$$\mu_j + \nu_{i-j+1} \le \lambda_i \quad pro \quad i \ge j; \qquad \lambda_i \le \mu_j + \nu_{i-j+n} \quad pro \quad i \le j.$$

Specielně pro $i = j$ *platí*

$$\mu_j + \nu_1 \le \lambda_j \le \mu_j + \nu_n, \quad j = 1, \ldots, n.$$

Důkaz: Viz např. [28], [47]. □

Důsledek A.1 *Nechť* $\mathbf{N} = \mathrm{diag}(\nu, 0, \ldots, 0) \in \mathbb{R}^{n\times n}$, *kde* $\nu \ne 0$. *Pak platí*

$$\nu > 0 \quad \Rightarrow \quad \mu_1 \le \lambda_1 \le \mu_2 \le \lambda_2 \le \ldots \le \mu_n \le \lambda_n,$$

$$\nu < 0 \quad \Rightarrow \quad \lambda_1 \le \mu_1 \le \lambda_2 \le \mu_2 \le \ldots \le \lambda_n \le \mu_n.$$

Na závěr něco o čísle podmíněnosti.

Poznámka A.3 *Nechť* $\mathbf{M} \in \mathbb{R}^{n\times n}$ *je symetrická matice. Potom číslo podmíněnosti* $\kappa(\mathbf{M})$ *matice* $\mathbf{M}$ *je definováno jako*

$$\kappa(\mathbf{M}) = \|\mathbf{M}\|\|\mathbf{M}^{-1}\|,$$

kde $\kappa(\mathbf{M}) = \infty$ *pro singulární matici* $\mathbf{M}$. *Číslo* $\kappa(\mathbf{M})$ *závisí na použité normě, například ve dvojkové (euklidovské) normě platí*

$$\kappa_2(\mathbf{M}) = \|\mathbf{M}\|_2\|\mathbf{M}^{-1}\|_2 = \frac{\lambda_n(\mathbf{M})}{\lambda_1(\mathbf{M})},$$

kde $\lambda_1(\mathbf{M})$, *resp.* $\lambda_n(\mathbf{M})$ *je nejmenší, resp. největší vlastní číslo matice* $\mathbf{M}$. *Pokud je* $\kappa(\mathbf{M})$ *velké, je* $\mathbf{M}$ *špatně podmíněná matice. Dále platí*

$$\kappa(\mathbf{M}) \geq 1$$

pro jakoukoli maticovou normu. Matice, které mají malé číslo podmíněnosti, jsou dobře podmíněné. Například ve dvojkové normě pro ortogonální matici $\mathbf{Q}$ *platí* $\kappa_2(\mathbf{Q}) = 1$. *Je-li* $\mathbf{M}$ *navíc positivně definitní, takže existuje* $\mathbf{M}^{\frac{1}{2}}$, *platí*

$$\kappa(\mathbf{M}^{\frac{1}{2}})^2 = \kappa(\mathbf{M}).$$

Závěr

Metody s lokálně omezeným krokem, které jsou hlavní náplní této práce, jsou efektivním nástrojem pro řešení optimalizačních úloh zejména tehdy, je-li účelová funkce nekonvexní nebo je-li úloha špatně podmíněná. Ve druhé kapitole bylo kromě popisu a zhodnocení všech známých publikovaných metod navrženo několik nových algoritmů. Efektivně jsme využili Lanczosovy metody a nové metody LCG a CGL, zvláště pak první metoda, se ukázaly být v praxi velmi efektivní. Metoda výpočtu optimálního lokálně omezeného kroku s použitím Choleského rozkladu a následnou aktualizací parametru ξ pomocí Newtonovy metody, § 2.1, byla detailně propracována s uvedením možných komplikací a jejich odstraněním.

Nedávno vyvinutá metoda používající parametrizovaný problém vlastních čísel je velmi dobrá z teoretického hlediska, avšak při testování této metody bylo zjištěno, že v případě rozsáhlých strukturovaných úloh je výpočet nejmenších vlastních čísel a jim příslušných vlastních vektorů pomocí programů knihovny ARPACK [29] časově příliš náročný. Otevřenou otázkou do budoucna zůstává, jak tuto metodu vylepšit, aby byla i v tomto případě efektivní.

Při testování jednotlivých metod se ukázalo, že nový algoritmus 2.13, který používá Lanczosův proces způsobem dovolujícím předpodmínění, algoritmus 2.8 používající v metodě psí nohy více kroků metody sdružených gradientů a algoritmus 2.2 realizující výpočet optimálního lokálně omezeného kroku pomocí řídkého Choleského rozkladu jsou nejefektivnější pro typy úloh, které byly testovány.

V oblasti minimalizace s obecnými omezeními jsme se zabývali metodami vnitřních bodů, kde se princip lokálně omezeného kroku používá opět proto, aby byly odstraněny problémy s nekonvexitou a špatnou podmíněností. Ve třetí kapitole je popsán nový algoritmus, ve kterém počítáme směrové vektory pomocí metod s lokálně omezeným krokem s využitím poznatků druhé kapitoly. Vyvinutý algoritmus 3.2 byl úspěšně implementován a aplikován na různé typy problémů. Protože pokutová funkce výrazně ovlivňuje volbu poloměru oblasti přijatelnosti, je tento algoritmus velmi citlivý na výběr této funkce. Ukazuje se, že standardní pokutové funkce vedou na algoritmy, které jsou méně efektivní než metody spádových směrů. Otázkou tedy zůstává, jak snížit tento vliv. Další otázkou zůstává zajištění globální konvergence při použití efektivní pokutové funkce.

Závěr práce patří zcela nové oblasti, použití principu filtru namísto pokutové funkce. Její idea tkví v tom, že nemusíme uvažovat jakou pokutovou funkci použít nebo jak aktualizovat pokutový parametr σ. Protože metody s lokálně omezeným krokem mají výborné konvergenční vlastnosti a použití filtrů je velmi efektivní, jak ukázaly numerické experimenty na novém algoritmu 3.4, je úkolem do budoucna spojit obě teorie dohromady, což by mohlo dát dobré výsledky. Také zůstává otázkou globální konvergence metod používajících princip filtru.

Poděkování

V roce 1998 jsem nastoupil na postgraduální studium na katedře Numerické matematiky MFF UK pod vedením Jana Zítko. Díky němu jsem získal mnoho zkušeností, děkuji mu za jeho trpělivost, přízeň a podporu a hlavně za pečlivé pročtení celého textu. Jeho odborné rady a připomínky k obsahu a úpravě přispěly ke zlepšení celé práce.

Rovněž děkuji všem ostatním spolupracovníkům z katedry, kterými jsem byl ovlivněn v prvních třech letech mého doktorského studia, za jejich přátelství a podporu.

V témže roce jsem současně nastoupil do Ústavu Informatiky AV ČR, kde disertační práce postupně vznikala. Předně děkuji Julovi Štullerovi a Zdeňku Strakošovi, kteří mě povzbuzovali a umožnili mi pracovat v klidu a v pohodě. Dále děkuji kolegům Mirkovi Tůmovi a Mirovi Rozložníkovi za jejich přízeň a podporu při psaní práce, knihovnicím Ludmile Nývltové a Nině Ramešové za ochotu a přátelství a také Hance Bílkové, které vděčím za dobrou náladu na pracovišti. Zvláště pak děkuji Petru Tichému, laskavému člověku a dobrému příteli, jehož odborné diskuse při zpracování textu a obrázků rovněž výrazně pomohly ke zkvalitnění práce.

Můj největší dík potom patří Ladislavu Lukšanovi. Jeho spolupráce od počátku mého doktorského studia byla pro mne velice přínosná. Poskytoval mi své životní zkušenosti, díky jeho bohaté zásobě literatury jsem získal přehled v dané problematice a velmi hodnotná byla též spolupráce při programování jednotlivých metod. Děkuji mu za jeho trpělivost, moudré rady, za to, že podrobně pročetl celý text, poskytl mi hodnotné návrhy a komentáře, věnoval mi přízeň a čas.

Literatura

[01] Benson H.Y., Vanderbei R.J., Shanno D.F.: *Interior-Point Methods for Nonconvex Nonlinear Programming: Filter Methods and Merit Functions;* Computational Optimization and Applications, 23, 257-272, 2000

[02] Bunch J.R., Parlett B.N.: *Direct methods for solving symmetric indefinite systems of linear equations;* SIAM J. Numer. Anal. 8, 639-655, 1971

[03] Byrd R.H., Gilbert J.C., Nocedal J.: *A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming;* 1996

[04] Byrd R.H., Hribar M.E., Nocedal J.: *An Interior Point Algorithm for Large Scale Nonlinear Programming;* 1997

[05] Chin C.M.: *A Global Convergence Theory of a Filter Line Search Method for Nonlinear Programming;* Numerical Optimization Report, 2002

[06] Coleman T.F., Li Y.: *An Interior Trust Region Approach for Nonlinear Minimization Subject to Bounds;* SIAM J. Optimization, Vol.6, No.2, 1996

[07] Conn A.R., Gould N.I.M., Toint P.L.: *Global Convergence of a Class of Trust Region Algorithms for Optimization with Simple Bounds;* SIAM J. Numer. Anal., Vol.25, No.2, 1988

[08] Conn A.R., Gould N.I.M., Toint P.L.: *Trust-region methods;* SIAM, 2000

[09] Das I.: *An interior point algorithm for the general nonlinear programming problem with trust region globalization*

[10] Dembo R.S., Steihaug T.: *Truncated-Newton algorithms for large-scale unconstrained optimization;* Mathematical Programming, 26(1983), 190-212

[11] Dennis J.E., Mei H.H.W.: *An unconstrained optimization algorithm which uses function and gradient values;* TR 75-246, 1975

[12] Dennis J.E., Vicente L.N.: *On the convergence theory of trust-region-based algorithms for equality-constrained optimization;* SIAM J. Optimization, Vol.7, 1997, pp.927-950

[13] Fletcher R.: *Practical methods of optimization;* John Wiley & Sons, 1987, 2nd edition

[14] Fletcher R., Gould N.I.M., Leyffer S., Toint P.L.: *Global Convergence of Trust-Region SQP-Filter Algorithms for General Nonlinear Programming;* Report 99/03, 1999

[15] Fletcher R., Leyffer S.: *Nonlinear Programming without a penalty function;* University of Dundee Numerical Analysis Report, NA/171, 1997

[16] Fletcher R., Leyffer S., Toint P.L.: *On the Global Convergence of an SLP-Filter Algorithm;* Report 98/13, 1998

[17] Gertz E.M., Gill P.E.: *A primal-dual trust region algorithm for nonlinear optimization;* Optimization Technical Report 02-09, 2002

[18] Gill P.E., Murray W.: *Newton type methods for unconstrained and linearly constrained optimization;* Math. Programming 7, 311-350, 1974

[19] Golub G.H., Van Loan C.F.: *Matrix computations;* The Johns Hopkins University Press, 1989, 2nd edition

[20] Gould N.I.M., Hribar M.E., Nocedal J.: *On the Solution of Equality Constrained Quadratic Programming Problems Arising in Optimization;* 2000

[21] Gould N.I.M., Leyffer S.: *An introduction to algorithms for nonlinear optimization;* RAL-TR-2002-031

[22] Gould N.I.M., Leyffer S., Toint P.L.: *A Multidimensional Filter Algorithm for Nonlinear Equations and Nonlinear Least Squares;* RAL-TR-2003-004, 2003

[23] Gould N.I.M., Lucidi S., Roma M., Toint P.L.: *Solving the trust-region subproblem using Lanczos method;* RAL-TR-97-028, 1997

[24] Gould N.I.M., Toint P.L.: *Global Convergence of a Non-monotone Trust-Region Filter Algorithm for Nonlinear Programming;* RAL-TR-2003-003, 2003

[25] Hager W.W.: *Minimizing a Quadratic Over a Sphere;* 1999

[26] He B.: *Solving Trust Region Problem in Large Scale Optimization;* J. of Computational Mathematics, Vol.18, No.1, 2000

[27] Hribar M.E.: *Large-scale constrained optimization;* a dissertation, 1996

[28] Ikebe Y., Inagaki T., Miyamoto S.: *The Monotonicity Theorem, Cauchy's Interlace Theorem, and the Courant-Fischer Theorem;* The American Mathematical Monthly, Vol. 94, No.4, 1987, pp.352-354

[29] Lehoucq R.B., Sorensen D.C., Yang C.: *ARPACK Users' Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods;* SIAM, Philadelphia, 1998

[30] Lescrenier M.: *Convergence of Trust Region Algorithms for Optimization with Bounds when Strict Complementarity Does Not Hold;* SIAM J. Numer. Anal., Vol.28, No.2, 1991

[31] LIANG X., XU C.: *A Trust Region Algorithm for Bound Constrained Optimization;* Optimization, Vol.41, pp.279-289, 1997

[32] LIN C., MORÉ J.J.: *Newton's Method for Large Bound-Constrained Optimization Problems;* SIAM J. Optim., Vol.9, No.4, pp.1100-1127, 1999

[33] LUKŠAN L.: *Hybrid methods for large sparse nonlinear least squares;* J. Optim. Theory Appl., 89, 1996, 575-595

[34] LUKŠAN L.: *Metody s proměnnou metrikou;* Academia Praha, 1990

[35] LUKŠAN L.: *Numerické optimalizační metody pro úlohy bez omezujících podmínek;* Výzkumná zpráva č. V-640, 1995

[36] LUKŠAN L., MATONOHA C., VLČEK J.: *Interior point method for non-linear non-convex optimization;* Numerical Linear Algebra with Applications, 2004

[37] LUKŠAN L., MATONOHA C., VLČEK J.: *Nonsmooth equation method for nonlinear nonconvex optimization;* sborník konference „Conjugate Gradient Algorithms and Finite Element Methods", Springer-Verlag, Berlín, 2004.

[38] LUKŠAN L., SPEDICATO E.: *Variable metric methods for unconstrained optimization and nonlinear least squares;* J. of Computational and Applied Mathematics, vol. 124, 2000

[39] LUKŠAN L., TŮMA M., ŠIŠKA M., VLČEK J., RAMEŠOVÁ N.: *UFO 2002, Interactive System for Universal Functional Optimization;* Technical report No.883, 2002

[40] LUKŠAN L., VLČEK J.: *Indefinitely Preconditioned Inexact Newton Method for Large Sparse Equality Constrained Non-linear Programming Problems;* Numerical Linear Algebra with Applications, Vol.5, 1998, p.219-247

[41] LUKŠAN L., VLČEK J.: *Numerical experience with iterative methods for equality constrained nonlinear programming problems;* Optimization Methods and Software, Vol.16, 2001

[42] LUKŠAN L., VLČEK J.: *Sparse and Partially Separable Test Problems for Unconstrained and Equality Constrained Optimization;* Technical Report No. 767, 1998

[43] MORÉ J.J., SORENSEN D.C.: *Computing a trust region step;* Applied Mathematics Division, ANL-81-83, 1981

[44] NIE P.Y.: *Sequential Penalty Quadratic Programming Filter Methods for Nonlinear Programming*

[45] NOCEDAL J., WRIGHT S.J.: *Numerical Optimization;* Springer Series in Operations Research, 1999

[46] PALAGI L.: *Large scale trust region problems;* Tech. Rep. 07-99

[47] PARLETT B.N.: *The symmetric eigenvalue problem;* Prentice-Hall, Inc., Englewood Cliffs, N.J., 07632, 1980

[48] PLANTENGA T.D.: *Large-scale nonlinear constrained optimization using trust regions;* a dissertation, 1994

[49] POWELL M.J.D.: *A New Algorithm for Unconstrained Optimization;* Nonlinear programming, Academic press, 1970

[50] POWELL M.J.D.: *On the global convergence of trust region algorithms for unconstrained optimization;* Mathematical Programming 29(1984) 297-303

[51] POWELL M.J.D.: *Trust region calculations revisited;* 17th Biennial Conference on Numerical Analysis (Dundee), 1997

[52] RALSTON A.: *Základy numerické matematiky;* Academia Praha, 1973

[53] RENDL F., WOLKOWICZ H.: *A semidefinite framework for trust region subproblems with applications to large scale minimization;* Math. Prog., 77(1997)

[54] ROJAS M.: *A Large-Scale Trust-Region Approach to the Regularization of Discrete Ill-Posed Problems;* a dissertation, 1998

[55] ROJAS M., SANTOS S.A., SORENSEN D.C.: *A New Matrix-Free Algorithm for the Large-Scale Trust-Region Subproblem;* SIAM J. Optim., Vol.11, No.3, 2000

[56] ROJAS M., SORENSEN D.C.: *A Trust-Region Approach to the Regularization of Large-Scale Discrete Ill-Posed Problems;* Technical Report 99-26, Dep. of Computational and Applied Mathematics, Rice Univ., Houston, 2000

[57] SANTOS S.A., SORENSEN D.S.: *A New Matrix-Free Algorithm for the Large-Scale Trust-Region Subproblem;* CRPC-TR95551, 1995

[58] SORENSEN D.C.: *Minimization of a large-scale quadratic function subject to a spherical constraint;* SIAM J. Optim., Vol.7, No.1, 1997

[59] SORENSEN D.C.: *Newton's method with a model trust region modification;* SIAM J. Numer. Anal., Vol.19, No.2, 1982

[60] STEIHAUG T.: *The conjugate gradient method and trust regions in large scale optimization;* SIAM J. Numer. Anal., Vol.20, No.3, 1983

[61] SUN W., YUAN Y.: *A Conic Model Trust Region Method for Nonlinearly Constrained Optimization;* submitted to Mathematics of Computation, 1997

[62] TITS A.L., WÄCHTER A., BAKHTIARI S., URBAN T.J., LAWRENCE C.T.: *A Primal-Dual Interior Point Method for Nonlinear Programming with Strong Global and Local Convergence;* ISR Technical Report TR 2002-29, 2002

[63] ULBRICH M., ULBRICH S., VICENTE L.N.: *A Globally Convergent Primal-Dual Interior-Point Filter Method for Nonconvex Nonlinear Programming;* TR00-12, 2000

[64] ULBRICH S.: *On the superlinear local convergence of a filter-SQP method;* 2002

[65] VANDERBEI R.J., SHANNO D.F.: *An interior point algorithm for nonconvex nonlinear programming;* Computational Optimization and Applications, 13, 231-252, 1999

[66] WRIGHT S.J.: *Primal-dual interior point methods;* SIAM, 1997

[67] YUAN Y.: *A review of trust region algorithms for optimization*

[68] YUAN Y.: *Matrix computation problems in trust region algorithms for optimization*

[69] ZHANG J., XU C.: *A class of indefinite dogleg path methods for unconstrained minimization;* SIAM J. Optim., Vol.9, No.3, pp.646-667