Úvod do numerické matematiky.
Přednáška pro posluchače informatiky.
Zimní resp. Letní semestr 2/2

Ivo Marek, Petr Mayer a Bohuslav Sekerka

1 Úvodní poznámky.

Vymezení problematiky vystihuje následující charakteristika.

Numerická (výpočtová) matematika je realizace matematických modelů na výpočetní technice.

Numerická matematika je tedy součástí matematiky. Je to však součást mající oproti teoretické matematice jeden významný nedostatek. Ten spočívá v tom, že čistě teoretické matematické modely podstatným způsobem operují s realnými čísly zatímco numerická matematika je závislá na výpočetní technice a tudíž se musí bez tohoto prostředku obejít. Typickým projevem této skutečnosti jsou zaokrouhlovací chyby, což vede ve svých důsledcích k *numerické nestabilitě*.

Zdroj numerické nestability však nemusí být výhradně chyby ze zaokrouhlení. Numerická nestabilita může být zapřičiněna též vlastnostmi matematických modelů samých, na př. realizace matematického špatně podmíněného modelu na počítači je přirozeným zdrojem numerické nestability. Zde použitý pojem *špatně podmíněný model* je charakterizován nespojitou závislostí výstupní informace na informacích vstupních.

Příklad 1.1 Buď

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

kde a_{jk} jsou reálná čísla.

Symbolem $h(A)$ označme hodnost matice A. Snadno si uvědomíme, že h není spojitou funkcí prvků a_{jk}.

Abychom si tuto okolnost uvědomili, stačí položiti

$$A_0 = \begin{bmatrix} a_{11} & 0 \\ 0 & 0 \end{bmatrix},\ a_{11} \neq 0,$$

takže na jedné straně

$$h(A_0) = 1,$$

avšak, na straně druhé, pro libovolné reálné $b \neq 0$

$$h(A(b)) = 2,$$

kde

$$A(b) = \begin{bmatrix} a_{11} & 0 \\ 0 & b \end{bmatrix}.$$

Předchozí příklad ukazuje jednak nespojitost hodnosti matice jakožto funkce jejích prvků jednak další pozoruhodnou vlastnost hodnosti. Ta je obsahem následující věty.

Věta 1.1 *Hodnost matice je zdola polospojitou funkcí jejích prvků.*

Definice 1.1 Skalární reálná funkce f se nazývá *zdola polospojitou v bodě* s_0, jestliže pro každé $\varepsilon > 0$ množina

$$\{s : f(s) > f(s_0)\}$$

je otevřená, t. j. existuje $\delta > 0$ takové, že

$$f(s) - f(s_0) \geq \varepsilon$$

pro $s \in (s_0 - \delta, s_0 + \delta)$.

Dalším příkladem uvedeme numerickou nestabilitu poplatnou chybám ze zaokrouhlení.

Příklad 1.2 Určeme hodnoty integrálů

$$I_n = \frac{1}{e} \int_0^1 t^n e^t dt \tag{1.1}$$

pro *n = 1, 2, ...*

Užitím metody "per partes" zjistíme, že

$$I_n = 1 - nI_{n-1} \tag{1.2}$$

při čemž

$$I_0 = \frac{e-1}{e}. \tag{1.3}$$

Zřejmě platí

$$0 \leq I_n \leq I_{n-1} \leq 1 \tag{1.4}$$

jakož i

$$\lim_{n\to\infty} I_n = 0. \tag{1.5}$$

Výpočet na počítači se realizuje nikoliv v termínech zavedených veličin I_n avšak v termínech veličin počítaných $\tilde{I}_n$.

Vzhledem k zaokrouhlovacím chybám zjistíme, že existují indexy n_1 a n_2 tak, že

$$\tilde{I}_{n_1} < 0 \tag{1.6}$$

a

$$\tilde{I}_{n_2} > 1. \tag{1.7}$$

Vidíme tedy, že vztahy (1.4) neplatí pro veličiny $\tilde{I}_n$.

Důvodem pro platnost vztahů (1.6) a (1.7) je skutečnost vyjádřená rovnostmi

$$\tilde{I}_n = I_n + \delta_n, \quad n = 1, 2, ...$$

z nichž plyne, že

$$\delta_n = n\delta_{n-1},$$

odkud je patrné, že chyba se na každém kroku výpočtu násobí indexem kroku. Toto zesílení chyby je velikosti faktoriálu, což značí, že "destrukce" se explicitně projeví po p krocích, kde p značí počet cifer používaných v průběhu výpočtu.

2 Diferenční rovnice

Připomeňme si pojem *obyčejné diferenciální rovnice.*

Buď $f = f(s, y, z)$ skalární reálná funkce, při čemž

$$s \in I = (a, b), \quad y \in S_1, \quad z \in S_2, \tag{2.1}$$

kde

$$S_1 \subset R, \quad S_2 \subset R,$$

při standartním značení množiny reálných čísel symbolem R.

Úlohu nalézti funkci $x = x(s), s \in I$ takovou,že jednak

$$x(s) \in S_1, x'(s) \in S_2 \tag{2.2}$$

a jednak

$$f(s, x(s), x'(s)) = 0, \tag{2.3}$$

se nazývá *obyčejnou diferenciální rovnicí 1. řádu.*

Příklad 2.1 Položme

$$f(s, y, z) = z - \alpha y.$$

Snadno se přesvědčíme, že

$$x(s) = Cexp\{\alpha s\}$$

je pro $s \in [0, T)$, kde $T > 0$, řešením diferenciální rovnice

$$x' = \alpha x$$

splňující počáteční podmínku

$$x(0) = C.$$

Příklad 2.2 Nechť $I = (-\infty, +\infty), S_0 = S_1 = [-1, 1]$ a

$$f(s, y, z) = z^2 - (1 - y^2).$$

Pak máme co do činění s diferenciální rovnicí

$$x'^2 = 1 - x^2,$$

jejíž řešení jsou dána formulí

$$x(s) = \sin(s - a),$$

kde $a \in R$ je libovolné.

Dále si připomeňme pojem diferenciální rovnice N-tého řádu.

Buď

$$f = f(s, y_0, y_1, ..., y_N)$$

skalární reálná funkce definovaná pro $s \in I \subset R$ a $y_0, ..., y_N$ ležící v podmnožinách reálných čísel $S_0, ...S_N$. Problém nalézt funkci $x = x(s)$ definovanou pro $s \in I$, mající N derivací na I a splňující relace

$$x^{(k)}(s) \in S_k \quad k = 0, 1, ..., N \tag{2.4}$$

a

$$f(s, x(s), x'(s), ..., x^{(N)}(s)) = 0 \tag{2.5}$$

pro všechna $s \in I$, se nazývá *obyčejnou diferenciální rovnicí řádu* N; symbolicky píšeme

$$f(s, x(s), x'(s), ..., x^{(N)}) = 0.$$

Příklad 2.3 Nechť $N = 2$ a $I = S_0 = S_2 = R$. Nechť dále

$$f(s, y_0, y_1, y_2) = y_0 + y_2.$$

Potom každá funkce x tvaru

$$x(s) = A\cos s + B\sin s,$$

kde A a B jsou konstanty, je řešením odpovídající diferenciální rovnice

$$x'' + x = 0.$$

Nyní přistoupíme k úlohám podobným diferenciálním rovnicím. Analogie bude zřejmá z definice. Budeme se zabývati diferenčními rovnicemi.

Uveďme napřed pojem *diferenční rovnice 1. řádu.*

Budiž dána soustava

$$f = \{f_n(y, z)\},$$

kde $f_n = f_n(y, z), n = 1, 2, ...,$ jsou skalární reálné funkce definované na množině $I \subset Z_\infty$, při čemž Z_∞ značí množinu celých čísel a $y, z \in S \subset R$. Úloha nalézt posloupnost $\{x_n\} \subset R$, $n \in I$, splňující

$$x_n \in S, \quad x_{n-1} \in S \tag{2.6}$$

a

$$f_n(x_n, x_{n-1}) = 0 \tag{2.7}$$

se nazývá *diferenční rovnice 1. řádu.* Každá posloupnost $\{x_n\}$ splňující (2.6) a (2.7) se nazývá *řešením diferenční rovnice* (2.7).

Příklad 2.4 Buď I množina celých čísel. Potom

$$f_n(y,z) = y - z - 1$$

vede k diferenční rovnici

$$x_n - x_{n-1} = 1,$$

Její řešení mají tvar

$$x_n = n + c,$$

kde $c \in R$ je libovolné.

Podobně

Příklad 2.5 Nechť I je množina nezáporných celých čísel a

$$f_n(y,z) = y - z - n,$$

takže naše diferenční rovnice má tvar

$$x_n - x_{n-1} = n.$$

Jejím řešením je posloupnost $\{x_n\}$, kde

$$x_n = \frac{n(n-1)}{2} \quad .$$

Příklad 2.6 Buď I množina všech celých čísel a

$$f_n(y,z) = y - qz,$$

Není obtížné ukázat, že řešení splňující $x_0 = 1$ odpovídající rovnice

$$x_n = qx_{n-1}$$

má tvar

$$x_n = q^n.$$

Nyní přistoupíme k vyšetřování *diferenčních rovnic řádu* $N \geq 1$.

Buď

$$f = \{f_n(y_0, y_1, ..., y_N)\}$$

posloupnost funkcí definovaných na množině $I \subset Z_\infty$, př čemž $y_j \in S_j \subset R$, $j = 0, 1, ..., N$.

Úlohu nalézt posloupnost $\{x_n\}, n \in I$ a splňující následují požadavky

$$x_n \in S_n,\ x_{n-1} \in S_{n-1},\ \ ...x_{n-N} \in S_{n-N} \tag{2.8}$$

a

$$f_n(x_n, x_{n-1}, ..., X_{n-N}) = 0. \tag{2.9}$$

Posloupnost $\{x_n\}$ splňující (2.8) a (2.9) se nazývá *řešení diferenční rovnice* (2.9).

Příklad 2.7 Nechť $I = Z_\infty$ *a*

$$f_n(y_0, y_1, y_2) = y_0 - 2\cos\phi y_1 + y_2,$$

kde $\phi \in R$. Řešení takto vzniklé diferenční rovnice

$$x_n - 2\cos\phi\, x_{n-1} + x_{n-2} = 0$$

je dáno formulí

$$x_n = \cos n\phi.$$

Podobně jako v teorii diferenciálních rovnic se pro některé třídy úloh jednoznačné řešitelnosti dociluje vhodnou volbou počátečních podmínek.

Důležitým případem diferenčních rovnic jsou *diferenční rovnice lineární*. V tom případě

$$f_n(y_0, y_1, ..., y_N) = a_{0n}y_0 + a_{1n}y_1 + ... + A_{Nn}y_N + b_n, \tag{2.10}$$

kde $a_{jn}, b_n \in R,\ j = 0, 1, ..., N,\ n = 0, 1, ...$

Příklad 2.8 Všechny diferenční rovnice uvedené v příkladech $2.3 - 2.7$ jsou lineární.

Podobně diferenční rovnice

$$x_n + 5nx_{n-1} + n^2x_{n-2} = 2$$

je lineární, zatímco

$$x_{n-1} - 2x_{n-1}^2 = 0$$

je nelineární.

Analogicky jako v teorii diferenciálních rovnic, obecné řešení diferenčních rovnic umíme plně charakterizovat pro případ diferenčních rovnic lineárních.

Zabývejme se úlohou (analytického) sestrojení řešení diferenčních rovnic 1. řádu. Tedy, řešme rovnici

$$x_n = a_n x_{n-1} + b_n,\ n \in I \subset Z_\infty^+,\ a_n \neq 0, \tag{2.11}$$

kde Z_∞^+ značí množinu všech nezáporných celých čísel.

Vyšetřujme nejprve *homogenní* diferenční rovnici

$$x_n = a_n x_{n-1}. \tag{2.12}$$

Snadno zsjistíme, že řešení splňující podmínku

$$x_0 = c \tag{2.13}$$

je dáno výrazem

$$x_n = c\pi_n, \tag{2.14}$$

kde

$$\pi_0 = 1,\ \pi_n = \prod_{k=1}^{n} a_k,\ n = 1, 2, ... \tag{2.15}$$

K určení řešení *nehomogenní* rovnice (2.12) splňující $x_0 = c$, použijeme metody známé z teorie lineárních diferenciálních rovnic pod názvem *metoda variace konstanty.*

Položme

$$x_n = c_n \pi_n, \tag{2.16}$$

kde $\{c_n\}$ podléhá určení.

Z rovností

$$x_0 = c_0 \pi_0 = c_0,$$

vidíme, že

$$c_0 = c.$$

Po dosazení (2.16) do (2.11) zjistíme, že

$$c_n \pi_n = a_n c_{n-1} \pi_{n-1} + b_n = c_{n-1} \pi_n + b_n.$$

Z předpokladu $a_n \neq 0$ plyne, že též $\pi_n \neq 0$. Předchozí vztahy implikují rovnost

$$c_n = c_{n-1} + \frac{b_n}{\pi_n}$$

a dále pak rovnosti

$$c_n = c_0 + \sum_{k=1}^{n} (c_k - c_{k-1}) = c + \sum_{k=1}^{n} \frac{b_k}{\pi_k}.$$

Výsledek shrneme ve tvaru věty.

Věta 2.1 *Nechť $a_n \neq 0,\ n = 1, 2, ...$ Řešení diferenční rovnice*

$$x_n = a_n x_{n-1} + b_n,$$

splňující $x_0 = c$, je dáno výrazem

$$x_n = \pi_n \left(c + \sum_{k=1}^{n} \frac{b_k}{\pi_k} \right),\ n = 0.1, ... \tag{2.17}$$

při čemž

$$\pi_0 = 1,\ \pi_n = \prod_{k=1}^{n} a_k.$$

Příklad 2.9 Buď I množina nezáporných čísel,

$$f_n(y_0, y_1, y_2) = y_0 - y_1 - y_2.$$

Odpovídající diferenční rovnice

$$x_n - x_{n-1} - x_{n-2} = 0 \tag{2.18}$$

definuje *Fibonacciova čísla* $(x_0 = 0,\ x_1 = 1)$.

Podobně jako v problematice diferenciálních rovnic je jedním z možných způsobů řešení lineárních rovnic řádu N jejich převod na soustavy lineárních rovnic 1. řádu.

Ukážeme si zmíněný postup na příkladě rovnice (2.18).

Nechť

$$X_n = \begin{pmatrix} x_n \\ x_{n-1} \end{pmatrix},\ X_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix},\ n = 1, 2, ...$$

a

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

Snadno nahlédneme, že rovnici (2.18) lze vyjádřiti ve tvaru

$$X_n = AX_{n-1},\ X_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{2.19}$$

Podobně jako v případě lineární diferenční rovnice 1. řádu, řešení rovnice (2.19) má tvar

$$X_n = A^n X_0.$$

Odtud plyne platnost zajímavého vztahu

$$\lim_{n\to\infty} \frac{x_n}{x_{n-1}} = \frac{1+\sqrt{5}}{2}.$$

Vidíme tedy, že limitní poměr veličin x_n a x_{n-1} je dán číslem charakterizujícím t. zv. *zlatý řez*. Protože rovnice (2.18) má vskutku zajímavou biologickou interpretaci, můžeme ve výše uvedeném výsledku spatřovati úkaz estetiky projevující se v některých oblastech života na naší planetě. Model popisující jistý způsob rozmnožování králíků pochází ze 13. století; jeho autorem je právě Fibonacci (r. 1228).

Zcela jinou aplikací lineárních diferenčních rovnic je následující úloha.
Buď

$$p_n(x) = \sum_{k=0}^{n} b_k x^{n-k},\ b_0 \neq 0,\ n = 0, 1, ..., N$$

a položme si úlohu stanovit hodnotu

$$p_n(z),$$

kde z je daný bod na reálné ose.

Naším cílem je pokud možno minimalitovat při tom počet aritmetických operací ve snaze snížit riziko numerické nestability.

Algoritmus 2.1 *Počítejme veličiny $x_0,\ x_1,\ ...,\ x_N$ rekurentně pomocí relací*

$$x_0 = b_0,\ x_n = z x_{n-1} + b_n,\ n = 1,\ ...,\ N. \tag{2.20}$$

Posloupnost $\{x_n\}$ daná v (2.20) je řešení diferenční rovnice (2.11) splňující podmínku $x_0 = b_0$, kde $a_n = z$ pro $n = 1,\ 2,\ ...,$. V tomto případě

$$\pi_n = z^n$$

a tudíž, na základě věty 2.1,

$$x_n = \sum_{k=0}^{n} b_k z^k$$

a, speciálně pro $n = N$,

$$x_N = \sum_{k=0}^{N} b_k z^k = p_N(z).$$

Platí tedy

Věta 2.2 *Veličiny x_n tvořené pomocí algoritmu 2.1 jsou hodnoty polynomů p_n definovaných pomocí*

$$p_n(x) = \sum_{k=0}^{n} b_k x^k,\ n = 0,\ 1,\ ...,\ N$$

v bodě $x = z$.

Algoritmus 2.1 se nazývá *Hornerovým schématem.*

Nyní si ukážeme jak lze uvedený algoritmus zobecnit pro výpočet hodnot derivací polynomu p_n v bodě $x = z$.

Opět je nutné zdůraznit, že přímá aplikace standardních pravidel derivování člen po členu není vhodná a algoritmy typu Hornerova jsou daleko stabilnější a jednodušší z hlediska teorie složitosti.

Vyjdeme ze známých vztahů

$$p(x) = p(z + h) = \sum_{k=0}^{N} c_k h^{N-k},$$

kde

$$c_{N-k} = \frac{1}{k!} p^{(k)}(z),\ k = 0,\ 1,\ ...,\ N. \tag{2.21}$$

$$
\begin{array}{ccccc}
x_0^{(0)} & x_1^{(0)} & x_2^{(0)} & x_3^{(0)} & \underline{x_4^{(0)}} \\
 & & \downarrow & & \\
x_0^{(1)} & x_1^{(1)} & x_2^{(1)} & \underline{x_3^{(1)}} & \\
x_0^{(2)} & x_1^{(2)} & \underline{x_2^{(2)}} & & \\
x_0^{(3)} & \underline{x_1^{(3)}} & & & \\
\underline{x_0^{(4)}} & & & &
\end{array}
$$

Figure 1: Obrázek č. 2.1

Algoritmus 2.2 *Označme prvky tvořené pomocí algoritmu 2.1 symboly* $\{x_n^{(0)}\}$. *Pro* $k = 1,\ 2,\ ...,\ N$ *tvořme posloupnosti* $\{x_n^{(k)}\}$ *rekurzivně pomocí diferenčních schémat*

$$x_0^{(k)} = x_0^{(k-1)},\ x_n^{(k)} = z x_{n-1}^{(k)} + x_n^{(k-1)}, n = 0,\ 1,\ ...,\ N-k. \tag{2.22}$$

Věta 2.3 *Označuje-li* p_n *polynom*

$$p_n(x) = \sum_{k=0}^{n} b_k x^{n-k},\ n = 0,\ 1,\ ...,\ N,$$

pak veličiny určované algoritmem 2.2 splňují

$$x_n^{(k)} = \frac{1}{k!} p_{n+k}^{(k)}(z),\ n = 1,\ ,...,\ N,\ k = 0,\ ...,\ N-n. \tag{2.23}$$

Tudíž koeficienty c_{N-k} požadované ve formuli (2.21) jsou právě ty, jež se nalézají v dolní diagonále na obrázku 2.1 pro $N = 4$ v posloupnosti veličin tvořených algoritmem 2.2.

3 Iterační metody řešení nelineárních rovnic a jejich soustav

3.1 Postupné aproximace

Začněme příkladem úlohy najít kořen funkce g, t. j. nalézt takový bod $\hat{x} \in I = [a, b],\ -\infty < a \leq b < +\infty$, tak, aby

$$g(\hat{x}) = 0$$

za předpokladu, že g je daná spojitá funkce zobrazující $I = I_1$ do $I_2 \subset R$. Speciálně, g může být polynom. A již v tomto speciálním případě si můžeme uvědomit několik důležitých skutečností. Tak především, obecně nemusí v I existovat kořen funkce g. Ale i když takový kořen existuje, nemusí existovat analytické vyjádření tohoto kořene pomocí parametrů charakterizujících g (na př. koeficienty polynomu); takové formule jsou známy pro případ polynomů stupňů nepřevyšujících 4. A jak víme, cena takových formulí je sporná, protože takové formule jsou mnohdy prakticky nepoužitelné (t. zv. *casus irreducibilis*). Důsledky těchto skutečností jsou zřejmé: Naději, že se podaří nalézt kořen g mají spíše numerické než analytické metody.

Mezi numerickými metodami se z důvodů snadné algoritmické realizovatelnosti prosadily *metody iterační* a to navzdory jejich neuniversálnosti a pohříchu poměrně pomalé konvergenci. Právě uvedený nedostatek dal podnět k rozvoji metod urychlování konvergence obecných posloupností, jmenovitě pak takových generovaných iteračními procesy.

Základním iteračním postupem jsou *postupné aproximace*. Jsou určeny k sestrojování aproximací řešení rovnic typu

$$s = f(s), \tag{3.1}$$

kde f je daná funkce, tedy, $g(s) = s - f(s)$.

Algoritmus 3.1 *Zvolme x_0 libovolně a tvořme posloupnost $\{x_n\}$ rekurzívně pomocí relace*

$$x_n = f(x_{n-1}). \tag{3.2}$$

Všímněme si, že k tomu, abychom mohli prováděti jednotlivé kroky dle (3.2) musí f kromě spojitosti míti ještě další vhodné vlastnosti. Budeme předpokládat, že obor hodnot funkce f patří do I_1, tedy $f(I_1) \subset I_1$. V takovém případě spojitost f zaručuje neomezenou proveditelnost procesu (3.2).

Abychom to ozřejmili, položme

$$g(s) = s - f(s),\ a \leq s \leq b,$$

takže

$$g(a) \leq 0, \text{ a } g(b) \geq 0.$$

Ze spojitosti f plyne spojitost g a odtud pak skutečnost, že g nabývá všech hodnot z intervalu $[g(a),\ g(b)] = I_3 \subset I_1$. Existuje tedy alespoň jeden bod $\hat{s}$ takový,

$$g(\hat{s}) = 0.$$

To značí, že

$$\hat{s} = f(\hat{s})$$

a $\hat{s}$ je námi hledaný kořen.

Obecně, množina M všech samodružných bodů funkce f, t. j.

$$M = \{s \in I_1 : s = f(s)\}$$

má mohutnost moh $M \geq 1$.

Jednoznačnosti se dociluje pomocí dalších předpokladů.

Definice 3.1 Předpokládejme, že f zobrazuje $I \subset R$ do R a že existuje konstanta L taková, že nerovnost

$$|f(s) - f(t)| \leq L|s - t| \tag{3.3}$$

platí pro libovolná $s, t \in I$. V takovém případě se f nazývá *lipschitzovskou* a konstanta L *Lipschitzovou konstantou* funkce f.

Všimněme si, že funkce lipschitzovská na I je nutně spojitá na I. Je-li f diferencovatelná na I a platí

$$\sup\{|f'(s)| : s \in I\} \leq L,$$

kde $L > 0$ je konstanta, pak je f na I lipschitzovská a L je Lipschitzovou konstantou funkce f.

Předpokládejme, že Lipschitzova konstanta L splňuje nerovnost

$$0 < L < 1. \tag{3.4}$$

Snadno nahlédneme, že pro $s_1,\ s_2 \in M$ platí

$$|s_1 - s_2| = |f(s_1) - f(s_2)| \leq L|s_1 - s_2|$$

a tudíž

$$s_1 = s_2.$$

Můžeme přistoupit k formulaci základního výsledku o konvergenci postupných aproximací.

Věta 3.1 *Buď $I = [a, b]$ uzavřený konečný interval a nechť funkce f splňuje následující podmínky (i) – (iii).*

(i) $f(s) \in I$ pro $s \in I$.

(ii) f je lipschitzovská s Lipschitzovou konstantou $L < 1$.

Potom pro libovolné nulové přiblížení $x_0 \in I$ posloupnost $\{x_n\}$ definovaná pomocí algoritmu 3.1 konverguje k (jedinému) řešení rovnice $s = f(s)$.

Důkaz. Z předchozích úvah již víme, že existuje alespoň jedno řešení rovnice $s = f(s)$ v I a že, díky předpokladu o L, toto řešení je určeno jednoznačně. Stačí tedy dokázati konvergenci posloupnosti $\{x_n\}$.

Nechť $\hat{x} = f(\hat{x})$. Zřejmě

$$x_n - \hat{x} = f(x_{n-1}) - f(\hat{x}),$$

takže

$$|x_n - \hat{x}| \leq L|x_{n-1} - \hat{x}|$$

a tudíž

$$|x_n - \hat{x}| \leq L^n|x_0 - \hat{x}|. \tag{3.5}$$

Protože

$$\lim_{n\to\infty} L^n = 0,$$

obdržíme žádaný výsledek

$$\lim_{n\to\infty} |x_n - \hat{x}| = 0.$$

|||

Příklad 3.1 Pomocí postupných aproximací hledejme řešení rovnice

$$s = e^{-s}.$$

Protože $e^{-s} > 0$ pro $s \in (-\infty, +\infty)$ a $s > 1 \geq e^{-s}$ *pro* $s > 1$, stačí omezit se na interval $[0, 1]$.

Snadno si uvědomíme, že obor hodnot funkce $f(s) = e^{-s}$ pro $s \in [0, 1]$ leží v intervalu $[e^{-1}, 1] \subset [0, 1]$ a že pro $s_1\ s_2 \in [0, 1]$ platí

$$|f(s_1) - f(s_2)| = |f'(\tilde{s})||s_1 - s_2|,$$

při čemž

$$f'(\tilde{s}) = -e^{-\tilde{s}},\ \tilde{s} \in [s_1, s_2].$$

Protože

$$max\,\{|f'(s)| : s \in [0, 1]\} = 1,$$

Lipschitzova konstanta pro f na $[0, 1]$ je rovna 1, což k platnosti věty 3.1 nestačí.

Avšak zvolíme-li za I interval $[\frac{1}{2}, log\,2]$, snadno zjistíme, že pro $\frac{1}{2} \leq s \leq log\,2$,

$$\frac{1}{2} = e^{-log 2} \leq e^{-s} \leq e^{-\frac{1}{2}} < log\,2$$

a tudíž $[\frac{1}{2}, log\ 2] = I \subset [0, 1]$. Navíc první derivace funkce f, kde $f(s) = e^{-s}$, je klesající, takže

$$max\{|f'(s)| : s \in I\} = f'(\frac{1}{2}) \approx 0,606531 < 1.$$

Iterační proces (3.2) pro funkci $f(s) = e^{-s}$ je tedy konvergentní v I.

Odhad (3.5) nemá bezprostřední praktický význam, závisí totiž na znalosti hledaného řešení. V praxi jsme totiž vždy nuceni omezit se na určitý konečný počet iterací x_n v (3.2). Rádi bychom proto znali nějaký realistický odhad chyby po n - tém kroku.

Všimněme si, že pro libovolné $k \geq 1$ platí vztahy

$$\begin{aligned} |x_{k+1} - x_k| &= |f(x_k) - f(x_{k-1})| \\ &\leq L|x_k - x_{k-1}| \\ &\leq L^k|x_1 - x_0|. \end{aligned}$$

Budˇ n zafixováno pevně a nechť $m = n + p > n$. Potom

$$x_m - x_n = \sum_{k=n}^{n+p-1} (x_{k+1} - x_k).$$

Po aplikaci trojúhelníkové nerovnosti obdržíme vztahy

$$\begin{aligned} |x_m - x_n| &\leq \textstyle\sum_{k=n}^{n+p-1} L^k|x_1 - x_0| \\ &= L^n \textstyle\sum_{k=0}^{p-1} L^k|x_1 - x_0| \\ &\leq L^n \textstyle\sum_{k=0}^{\infty} L^k|x_1 - x_0| \\ &= L^n(1 - L)^{-1}|x_1 - x_0|. \end{aligned}$$

Protože pro $m \to \infty$ plyne, že $x_m \to \hat{x}$, tedy též $p \to \infty$, a platí

$$|x_n - \hat{x}| \leq \frac{L^n}{1-L}|x_1 - x_0|. \tag{3.6}$$

Důsledek 3.1 *Při splnění předpokladů věty 3.1 je odhad chyby po n krocích postupných aproximací definovaných pomocí algoritmu 3.1 dán výrazem (3.6).*

Zkoumejme blíže chování posloupnosti $\{x_n\}$ za předpokladů věty 3.1 doplněných o předpoklad následující.

Nechť f je spojitě diferencovatelná a nechť v celém intervalu I je tato derivace nenulová.

To znamená, že f bud˘ klesá nebo roste v I.

Je-li $x_0 \neq \hat{x} = f(\hat{x})$, pak $f(x_n) \neq \hat{x}$ pro libovolné $n = 1, .2, ...$ Abychom to nahlédli, stačí si uvědomit, že kdyby

$$x_n = f(x_{n-1}) = f(x_n),$$

při čemž $x_n \neq x_{n-1}$, pak

$$0 = f(x_{n-1}) - f(x_n) = f'(\xi)\ (x_{n-1} - x_n),$$

kde ξ leží mezi x_{n-1} a x_n, t.j.

$$(x_{n-1} - \xi)\ (\xi - x_n) < 0.$$

Protože $x_{n-1} - x_n \neq 0$, musí být $f'(\xi) = 0$, ale to je ve sporu s předpokladem. Odtud vyplývá, že chyba

$$d_n = x_n - \hat{x}$$

není pro žádné $n \geq 0$ nulová.

Existuje limita

$$\lim_{n\to\infty} \frac{d_{n+1}}{d_n}?$$

A když ano, pak čemu je rovna?

Snadno zjistíme, že

$$d_{n+1} = x_{n+1} - \hat{x} = f(x_n) - f(\hat{x})$$
$$= f(\hat{x} + d_n) - f(\hat{x}) = f'(\hat{x} + \theta_n\ d_n)\ d_n,$$

kde

$$0 < |\theta_n| < 1.$$

Definujme ε_n pomocí relace

$$f'(\hat{x} + \theta_n\ d_n) = f'(\hat{x}) + \varepsilon_n.$$

Potom

$$d_{n+1} = (f'(\hat{x}) + \varepsilon_n)d_n \tag{3.7}$$

a tudíž, ježto $\varepsilon_n \to 0$ pro $n \to \infty$ plyne na základě spojitosti f' rovnost

$$\lim_{n\to\infty} \frac{d_{n+1}}{d_n} = f'(\hat{x}). \tag{3.8}$$

Lze této skutečnosti využít pro praktické účely? Veličinu $\hat{x}$ a tudíž i $f'(\hat{x})$ neznáme.

Ukážeme, že explicitní znalost uvedených veličin není nutná k tomu, abychom rovnosti (3.8) efektivně využili k sestrojení algoritmu, který poskytne aproximace konvergující k hledanému řešení $\hat{x}$ rychleji než posloupnost $\{x_n\}$.

Předpokládejme na okamžik, že ve formuli (3.7) $\varepsilon_n = 0$ pro nějaké pevné $n \geq 1$. Nechť

$$f'(\hat{x}) = A,$$

takže

$$x_{n+1} - \hat{x} = A\,(x_n - \hat{x}),$$
$$x_{n+2} - \hat{x} = A\,(x_{n+1} - \hat{x}).$$

Snadno zjistíme, že

$$A = \frac{x_{n+2} - x_{n+1}}{x_{n+1} - x_n},$$

takže

$$\hat{x} = \frac{1}{1-A}\,(x_{n+1} - Ax_n)$$
$$= x_n + \frac{1}{1-A}\,(x_{n+1} - x_n) = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}$$

Vidíme tedy, že za předpokladu, že $\varepsilon_n = 0$, pro nějaké $n \geq 1$, lze přesné řešení $\hat{x}$ obdržeti použitím tří po sobě jdoucích iterací. To však je dost akademická situace.

Ačkoliv veličiny $\varepsilon_n \neq 0$, lze očekávat, že ve srovnání s modulem $f'(\hat{x})$ jsou malé.

Pro velká n tak posloupnost

$$y_n = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n} \tag{3.9}$$

skýtá pro $\hat{x}$ lepší aproximaci než x_n.

Má tedy smysl vyšetřovat

Algoritmus 3.2 *Buď $\{x_n\}$ libovolná posloupnost a nechť y_n je tvořeno pomocí (3.9).*

Zaveďme označení

$$\Delta x_n = x_{n+1} - x_n,\ n = 0, 1, ...$$

a

$$\Delta^{k+1} x_n = \Delta(\Delta^k x_n),\ k \geq 1.$$

Na př

$$\Delta^2 x_n = \Delta(x_{n+1} - x_n)$$
$$= x_{n+2} - 2x_{n+1} + x_n$$

Formuli (3.9) lze pak psát ve tvaru

$$y_n = x_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n}. \tag{3.10}$$

Posloupnost $\{y_n\}$ tvořená pomocí algoritmu 3.2 definuje t. zv. *Aitkenův urychlovací* Δ^2 *- proces.*

Jest na místě zdůraznit, že Aitkenův proces je definován pro posloupnosti ne nutně generované pomocí algoritmu 3.1. Pro takto obecný proces platí poměrně silné tvrzení.

Věta 3.2 *Budˇ* $\{x_n\}$ *daná posloupnost taková, že*

$$\hat{x} = \lim_{n\to\infty} x_n.$$

Nechť

$$d_n = x_n - \hat{x} \neq 0 \; n = 0, 1, ... \tag{3.11}$$

a

$$d_{n+1} = (A + \varepsilon_n)\, d_n, \tag{3.12}$$

kde A je konstanta taková, že

$$|A| < 1$$

a

$$\varepsilon_n \to 0 \;\; pro \;\; n \to \infty.$$

Potom posloupnost $\{y_n\}$ *tvořená pomocí (3.9) v algoritmu 3.2 je dobře definovaná pro dostatečně velká n a navíc*

$$\lim_{n\to\infty} \frac{y_n - \hat{x}}{x_n - \hat{x}} = 0, \tag{3.13}$$

což značí, že posloupnost $\{y_n\}$ *konverguje k* $\hat{x}$ *rychleji než posloupnost* $\{x_n\}$.

Důkaz. Použitím formule (3.12) obdržíme

$$d_{n+2} = (A + \varepsilon_{n+1})\,(A + \varepsilon_n)\, d_n,$$

takže

$$\begin{aligned} \Delta^2 x_n &= x_{n+2} - 2x_{n+1} + x_n \\ &= d_{n+2} - 2d_{n+1} + d_n \\ &= [(A-1)^2 + \varepsilon'_n]\;\; d_n, \end{aligned}$$

kde

$$\varepsilon'_n = A[\varepsilon_n + \varepsilon_{n+1}] - 2\varepsilon_n + \varepsilon_n\varepsilon_{n+1}.$$

Z podmínky $\varepsilon_n \to 0$ pro $n \to \infty$ plyne, že

$$\varepsilon'_n \to 0 \text{ pro } n \to \infty. \tag{3.14}$$

Odtud dostáváme,

$$(A-1)^2 + \varepsilon'_n \neq 0$$

pro dostatečně velká n, řekněme pro $n > n_0$. Tudíž

$$\Delta^2 x_n \neq 0 \text{ pro } n > n_0.$$

Dále pak

$$\Delta x_n = \Delta d_n = (A - 1 + \varepsilon_n)\, d_n$$

a tudíž

$$\begin{aligned} y_n - \hat{x} &= d_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n} \\ &= d_n - \frac{(A - 1 + \varepsilon_n)^2\, d_n}{(A-1)^2 + \varepsilon'_n}. \end{aligned}$$

Z $\varepsilon_n \to 0$ pro $n \to \infty$ a (3.14) plyne, že

$$\frac{y_n - \hat{x}}{d_n} = \frac{\varepsilon'_n - 2(A-1)\,\varepsilon_n - \varepsilon_n^2}{(A-1)^2 + \varepsilon'_n} \to 0,$$

což jsme měli dokázati. |||

Důsledek 3.2 *Předpokládejme, že f splňuje podmínky věty 3.1 a navíc, že f má spojitou 1. derivaci na I a $f'(s) \neq 0$ pro $s \in I$. Pro $x_0 \neq \hat{x}$ posloupnost $\{x_n\}$ vyhovuje předpokladům věty 3.2, takže algoritmus 3.2 poskytuje urychlení konvergence.*

Důkaz. Zřejmě stačí ověřit platnost vztahu (3.12). To však je vztah (3.7) odvozený za obecnějších předpokladů. Zbývá tudíž ověřit požadavek $|A| < 1$. Platnost tohoto vztahu však plyne odtud, že proces postupných aproximací je konvergentní. Tímto konstatováním můžeme důkaz ukončit. |||

Znovu připomeňme širší aplikabilitu urychlovacího Aitkenova Δ^2- procesu i mimo oblast metody postupných aproximací.

3.2 Kvadratická konvergence a Newtonova metoda

V předchozím odstavci jsme předpokládali, že $f'(s) \neq 0$ pro $s \in I$, což zaručuje platnost předpokladu (3.12) a obdrželi jsme tak charakteristiku rychlosti konvergence danou pomocí (3.12), t. zv. *lineární konvergenci*. Dá se říci, že v takovém případě počet platných cifer je lineární funkcí indexu iterací.

Učiňme nyní předpoklad

$$f'(\hat{x}) = 0. \tag{3.15}$$

Tento předpoklad je poměrně velmi silný a zabezpečuje splnění některých předpo-kladů věty 3.1.

Věta 3.3 *Buď $I \subset (-\infty, +\infty)$, (nikoliv nutně ohraničený interval) a nechť f je definovaná na I a přitom platí*

(i) f a f' jsou spojité na I.

(ii) Rovnice $s = f(s)$ má řešení $\hat{x} \in I$ takové, že platí (3.15).

Potom existuje $\delta > 0$ takové, že algoritmus 3.1 poskytuje posloupnost $\{x_n\}$ konvergující k $\hat{x}$ pro všechna $|x_0 - \hat{x}| \leq \delta$.

Důkaz. Označme

$$I_\delta = [\hat{x} - \delta, \hat{x} + \delta].$$

Pro dostatečně malá $\delta_1 > 0$, $I_{\delta_1} \subset I$. Potom existuje $\delta_0 > 0$ a $L > 0$ takové, že

$$|f'(s)| \leq L \text{ pro } s \in I_{\delta_0}.$$

Nechť $\delta = \min\{\delta_0, \delta_1\}$.

Podle věty o střední hodnotě, protože $f'(\hat{x}) = 0$ a tedy $0 \leq L < 1$,

$$|f(s) - \hat{x}| = |f(s) - f(\hat{x})| \leq |f'(\xi)||s - \hat{x}| \leq L|s - \hat{x}| < \delta.$$

Tudíž obor hodnot funkce f je obsažen v I_δ a algoritmus 3.1 vede ke konvergenci podle věty 3.1. |||

Zesilme naše požadavky na hladkost funkce f předpokladem existence spojité druhé derivace f'' na I_δ a její nenulovosti.

Podobně jako v předchozích úvahách předpoklad $x_0 \neq \hat{x}$ implikuje platnost vztahů $x_n \neq \hat{x}$, $n = 1, 2, ...$ Postupné aproximace nemohou tedy poskytnout přesné řešení v konečném počtu kroků.

Užitím Taylorova rozvoje se zbytkem, obdržíme vztahy

$$\begin{aligned} d_{n+1} &= x_{n+1} - \hat{x} \\ &= f(x_n) - f(\hat{x}) \\ &= f'(\hat{x})d_n + \frac{1}{2}f''(\hat{x} + \theta_n d_n)d_n^2, \end{aligned}$$

kde $0 < |\theta_n| < 1$.

Z našich předpokladů plyne, že

$$d_{n+1} = \frac{1}{2}f''(\hat{x} + \theta_n d_n)d_n^2. \tag{3.16}$$

Protože $d_n \neq 0$, $n = 1, 2, ...$ a $d_n \to 0$ pro $n \to \infty$, odvodíme, že

$$\lim_{n\to\infty} \frac{d_{n+1}}{d_n^2} = \frac{1}{2}f''(\hat{x}). \tag{3.17}$$

To je pozoruhodný vztah, podle něhož chyba po $(1 + n)$- tém kroku je úměrná *čtverci* chyby po n - tém kroku. Tato skutečnost se označuje jako *kvadratická konvergenvce*. Počet platných cifer se zdvojuje po každém iteračním kroku.

Příklad 3.2 Buď $a > 0$ a nechť $f(s) = (1/2)[s + (a/s)]$ pro $s > 0$. Rovnice $s = f(s)$ má řešení $\sqrt{a}$. Zřejmě $f'(\sqrt{a}) = 0$ a $f''(s) > 0$ pro $s > 0$ *a* $f''(\sqrt{a}) = \frac{1}{\sqrt{a)}}$. Tudíž

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) \tag{3.18}$$

konverguje k $\sqrt{a}$ kvadraticky pro x_0 dostatečně blízká k $\sqrt{a}$. Později ukážeme, že tento proces konverguje pro libovolné výchozí přiblížení $x_0 > 0$.

Nyní přistoupíme k snad nejpoužívanější metodě přibližného řešení nelineárních rovnic a lze říci, že nejen rovnic ale i nelineárních úloh - k *metodě Newtonově*.

V předchozích úvahách jsme dospěli ke kvadratické konvergenci za výrazně akademického předpokladu nulovosti první derivace v přesném řešení. Ani se nechce věřit, že tento akademismus má tak nádherné uplatnění, k jehož výkladu teď přistupujeme.

Buď F definovaná a dvakrát spojitě diferencovatelná na intervalu $I = [a, b] \subset R$ a nechť $F'(s) \neq 0$ pro $s \in I$. Dále nechť rovnice

$$F(s) = 0 \tag{3.19}$$

má řešení $\hat{x} \in (a, b)$ (nikoliv nutně jediné). Dá se ukázat, že toto řešení lze nalézt pomocí postupných aproximací s funkcí

$$f(s) = s + MF(s),$$

kde M je konstanta splňující určité podmínky, na př. $M = (\alpha+\beta)/2$, kde $I = [0, 1]$, $F(0) < 0 < F(1)$ a $0 < \alpha \leq F'(s) \leq \beta$. Víme též, že genericky obdržíme lineární konvergenci. Jak to tedy zařídit, abychom sestrojili algoritmus poskytující kvadratickou konvergenci?

Což připustit místo konstantního faktoru M vhodnou funkci, tedy snažit se sestrojit f ve tvaru

$$f(s) = s + h(s)F(s),$$

kde hodnoty funkce h se "snadno" počítají.

Snažme se tedy splnit tyto požadavky:

$$\hat{x} = f(\hat{x}) \text{ a } f'(\hat{x}) = 0,$$

To nás vede k relacím

$$f'(s) = 1 + h'(s)F(s) + h(s)F'(s)$$

a tudíž

$$h'(\hat{x})F(\hat{x}) + h(\hat{x})F'(\hat{x}) = -1$$

a ježto $F(\hat{x}) = 0$,

$$h(s) = -\frac{1}{F'(s)}.$$

Dospěli jsme tedy k následujícímu algoritmu.

Algoritmus 3.3 *Volme x_0 a určeme posloupnost $\{x_n\}$ pomocí vztahů*

$$x_{n+1} = x_n - \frac{1}{F'(x_n)}F(x_n), n = 0, 1, ... \tag{3.20}$$

Algoritmus 3.3 nese název *Newtonova metoda*, či *metoda Newtonova - Raphsonova*.

Z předchozího výkladu již víme, že, ježto $f(s) = (1/F'(s))F(s)$,

$$f'(\hat{x}) = 0,$$

takže případná konvergence posloupnosti $\{x_n\}$ je kvadratická.

Jednotlivé kroky Newtonovy metody mají velice názornou interpretaci.

V bodě x_n aproximuje graf funkce F pomocí tečny ke grafu F v tomto bodě. Průsečík této tečny s osou x určuje další aproximaci x_{n+1}.

Umíme si snadno představit grafy funkce F, pro něž Newtonův proces diverguje. (To je velice hezké cvičení.)

Podobně jako pro všechny iterační metody typu postupných aproximací, slabina Newtonovy metody spočívá mimo jiné též v silné závislosti na volbě počátečního přiblížení. Ukážeme si jak lze zabezpečit globální konvergenci; požadujeme vyšší hladkost F - totiž znaménkovou stálost druhé derivace F''.

Věta 3.4 *Předpokládejme, že funkce F je definovaná a dvakrát spojitě diferencovatelná v $I = [a, b]$ a nechť splňuje tyto požadavky*

(i) $F(a)F(b) < 0$;

(ii) $F'(s) \neq 0$ *pro* $s \in I$;

(iii) Buď $F''(s) \geq 0$ $s \in I$ *nebo* $F''(s) \leq 0,\ s \in I$;

(iv) Nechť (α) $c = a$ *jestliže*

$$|F'(a)| \leq |F'(b)| \ ,$$

(β) $c = b$, *jestliže*

$$|F'(b)| \leq |F'(a)| \, .$$

Dále nechť

$$\left| \frac{F(c)}{F'(c)} \right| \leq b - a.$$

Potom Newtonova metoda konverguje pro libovolné výchozí přiblížení x_0 k jedinému řešení rovnice $F(s) = 0$.

Poznámky. Podmínka (i) zřejmě zaručuje existenci kořene díky spojitosti F. Z (ii) plyne jednoznačnost kořene v I. Podmínka (iii) říká, že F je buď *konkávní* nebo *konvexní* v I. Posléze (iv) zajišťuje, aby tečna ke křivce $y = F(x)$ v koncovém bodě, kde $|F'(x)|$ je menší, protínala osu x v intervalu I.

Důkaz věty 3.4. Věta 3.4 zahrnuje tyto čtyři odlišné situace

(a) $F(a) < 0,\ F(b) > 0,\ F''(s) \leq 0$ $(c = b)$,

(b) $F(a) > 0,\ F(b) < 0,\ F''(s) \geq 0$ $(c = b)$,

(c) $F(a) < 0,\ F(b) > 0,\ F''(s) \geq 0$ $(c = a)$,

(d) $F(a) > 0,\ F(b) < 0,\ F''(s) \leq 0$ $(c = a)$.

Případy (b) a (d) se snadno redukují na (a) a (c) volbou $-F$ na místě F. (Tato úprava nemění dokonce ani prvky posloupnosti $\{x_n\}$.) Případ (c) se převede na (a) volbou $-x \to x$. (V tomto případě obdržíme místo $\{x_n\} \to \{-x_n\}$. Stačí se tedy zabývati případem (a). Máme pak co do činění s grafem rostoucí konkávní funkce F takové,že $F(a) < 0 < F(b)$.

Buď $\hat{x}$ jediný kořen rovnice $F(s) = 0$. Napřed předpokládejme, že $a \leq x_0 \leq \hat{x}$. Potom, na základě předpokladu, že $F'(x_0) \geq 0$,

$$x_1 = x_0 - \frac{F(x_0)}{F'(x_0)} \geq x_0.$$

Indukcí lze ukázati, že

$$x_0 \leq x_1 \leq ... \leq x_n \leq ... \leq \hat{x}. \tag{3.21}$$

Indukční krok je realizován takto. Z věty o střední hodnotě,

$$-F(x_n) = F(\hat{x}) - F(x_n) = F'(\xi)(\hat{x} - x_n)$$

a přitom $x_n \leq \xi_n \leq \hat{x}$. Z podmínky $F''(s) \leq 0$ plyne, že F' je nerostoucí a tudíž $F'(\xi_n) \leq F'(x_n)$, takže

$$-F(x_n) \leq F'(x_n)(\hat{x} - x_n)$$

a tedy

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)} \leq x_n + (\hat{x} - x_n) = \hat{x}.$$

Aplikací F k oběma stranám nerovnosti obdržíme (F je neklesající), že $F(x_{n+1}) \leq 0$ a tudíž

$$x_{n+2} = x_{n+1} - \frac{F(x_{n+1})}{F'(x_{n+1})} \geq x_{n+1}.$$

Tedy platí (3.21).

Na tomto místě aplikujeme jednu z fundamentálních vět matematické analýzy: *Každá ohraničená monotonní posloupnost reálných čísel je konvergentní.*

Označme tedy

$$\tilde{x} = \lim_{n\to\infty} x_n.$$

Z (3.21) plyne, že $\tilde{x} \leq \hat{x}$. Na druhé straně z definice posloupnosti $\{x_n\}$ obdržíme při $n = \infty$ rovnost

$$\tilde{x} = \tilde{x} - \frac{F(\tilde{x})}{F'(\tilde{x})}$$

a tedy $F(\tilde{x}) = 0$. Z jednoznačnosti pak $\tilde{x} = \hat{x}$. Tím je dokázáno tvrzení věty 3.4 pro případ $x_0 \leq \hat{x}$.

Dále vyšetřujme případ $x_0 \geq \hat{x}$. Opět použitím věty o střední hodnotě

$$F(x_0) = F(x_0) - F(\hat{x}) = F'(\xi_0)(x_0 - \hat{x}),$$

kde $\hat{x} \leq \xi_0 \leq x_0$ a díky monotonii F', $F(x_0) \geq F'(x_0)(x_0 - \hat{x})$. Snadno zjistíme, že

$$x_1 = x_0 - \frac{F(x_0)}{F'(x_0)} \leq x_0 - (x_0 - \hat{x}) = \hat{x}$$

a tudíž

$$F(x_0) \leq F(b) - (b - x_0)F'(b).$$

Na základě podmínky (iv) obdržíme, že platí

$$x_1 = x_0 - \frac{F(x_0)}{F'(x_0)} \geq x_0 - \frac{F(x_0)}{F'(b)}$$

$$\geq x_0 - \frac{F(b)}{F'(b)} + b - x_0$$

$$= x_0 - (b-a) + b - x_0 = a.$$

Tudíž $a \leq x_1 \leq \hat{x}$. Tím jsme se dostali do situace, již jsme vyšetřovali v předchozí části důkazu s tím rozdílem, že na místě x_0 vystupuje x_1. Podle již dokázaného, $\{x_n\}$ konverguje k $\hat{x}$. Tím je důkaz věty 3.4 proveden. |||

Vraťme se ještě k příkladu 3.2. O procesu v něm vyšetřovaném již víme, že je konvergentní pro počáteční přiblížení dostatečně blízká k přesnému řešení, t. j. k $\sqrt{c}$, kde $c > 0$ je dané kladné číslo, jehož odmocninu hledáme.

Položme tedy

$$F(s) = s^2 - c$$

a

$$f(s) = s - \frac{F(s)}{F'(s)} \ .$$

Snadno zjistíme, že

$$F'(s) = 2s > 0,$$

a

$$F''(s) = 2 > 0$$

a tudíž nastává v tomto případě situace (c) popsaná v důkazu věty 3.4. V každém intervalu (a, b), kde $0 < a < \sqrt{c} < b$, platí, že

$$\left|\frac{F(a)}{F'(a)}\right| = \frac{c - a^2}{2a} \leq b - a,$$

pro každé b splňující nerovnost

$$b \geq \frac{1}{2}(a + c/a).$$

Z věty 3.4 plyne konvergence posloupnosti

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{c}{x_n}\right) \to \sqrt{c}$$

pro libovolné $x_0 > 0$.

Zvlášť výhodné je použití metody Newtonovy k hledání nulových bodů polynomů.

V tom případě využijeme té skutečnosti, že umíme efektivně stanoviti jak hodnoty $p(z)$ tak $p'(z)$, kde p je daný mnohočlen a z je daná veličina.

Buď tedy

$$p(z) = \sum_{k=0}^{N} a_k z^{N-k} \ ,$$

pak pomocí vztahů

(3.22) $$b_0 = a_0, \ b_n = z b_{n-1} + a_n, \ n = 1, \ 2, \ ..., \ N;$$

a

(3.23) $$c_0 = b_0, \ c_n = z c_{n-1} + b_n, \ n = 1, \ 2, \ ..., \ N-1,$$

obdržíme rovnosti

$$b_N = p(z) \text{ a } c_{n-1} = p'(z).$$

Schématicky postupujeme jak znázorněno níže

$$\begin{array}{ccccccc} a_0 & & a_1 & & a_2 & ... & a_{N-1} & & a_N \\ \downarrow & & \downarrow & & \downarrow & ... & \downarrow & & \downarrow \\ b_0 & \Rightarrow & b_1 & \Rightarrow & b_2 & \Rightarrow ... & b_{N-1} & \Rightarrow & \underline{b_N} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \\ c_0 & \rightarrow & c_1 & \rightarrow & c_2 & \rightarrow ... & \underline{c_{N-1}} & & \end{array}$$

při čemž $\rightarrow$ znamená sčítání a $\Rightarrow$ násobení faktorem z a sčítání.

Budˇ nyní z kořen polynomu p. Pak

$$p(x) = (x - z)q(x),$$

kde q je polynom stupně $N - 1$,

$$q(x) = \sum_{k=0}^{N-1} b_k x^{N-k-1},\ b_N = 0.$$

Snadno zjistíme, že

$$a_0 = b_0$$

a

$$a_{n-1} = -b_n z + b_n,\ n = 1,\ ...,\ N - 1.$$

Tudíž

$$\sum_{k=0}^{N} a_k x^{N-k} = \sum_{k=0}^{N-1} b_k x^{N-k} - z \sum_{k=0}^{N-1} b_k x^{N-k-1},$$

při čemž

$$a_n = b_n - z b_{n-1},$$

neboli,

(3.24)
$$b_n = a_n + z b_{n-1}.$$

Vidíme, že jsme obdrželi vztahy (2.20) algoritmu 2.1.
Platí tedy

Věta 3.5 *Je-li z kořen polynomu p, kde*

$$p(x) = \sum_{k=0}^{N} a_k x^{N-k},\ a_0 \neq 0,$$

a koeficienty b_0, ..., b_{N-1} se počítají podle formule (3.24), pak

$$q(x) = (x - z)^{-1} p(x) = \sum_{k=0}^{N-1} b_k x^{N-k},\ b_0 = a_0.$$

Mohlo by se zdát, že je výhodné hledat další kořeny polynomu p jakožto nulové body polynomu nižšího stupně q. Ukazuje se však, že taková procedura je numericky nestabilní a je tudíž vhodnější počítat všechny kořeny polynomu pomocí původního polynomu p a nikoliv modifikovaného polynomu q.

Cvičení 3.1 *Analyzujte způsobem obdobným analýze Newtonovy metody metodu následující,* metodu regula falsi*, definovanou předpisem*

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1})F(x_n)}{F(x_n) - F(x_{n-1})},$$

kde F je funkce, jejíž kořeny hledáme, t. j. $F(x) = 0$.

Cvičení 3.2 *Jaké vlastnosti má Newtonova metoda v případě, kdy $F'(x^*) = 0$, kde $F(x^*) = 0$.*

Cvičení 3.3 *Analyzujte* modifikovanou *Newtonovu (Whittakerovu - Robinsonovu) metodu danou předpisem*

$$x_{n+1} = x_n - \frac{1}{m}F(x_n),$$

kde m je pevně zvolená konstanta, na př. $m = F'(x_0)$.

4 Iterační metody řešení soustav nelineárních rovnic

4.1 Věta o kontrakci

Předmětem našeho studia bude vyšetřování následující úlohy. Nalézt prvek $x^\star \in R^N$ takový, že pro $x = x^*$ platí

$$\left\{\begin{array}{ll} x_1 & = f_1(x_1, ..., x_N), \\ x_2 & = f_2(x_1, ..., x_N), \\ & . \\ & . \\ & . \\ x_N & = f_N(x_1, ..., x_N), \end{array}\right. \tag{4.1}$$

neboli, psáno vektorově,

$$\mathbf{x} = \mathbf{f}(\mathbf{x}), \tag{4.2}$$

kde $\mathbf{x} = (x_1, ..., x_N)^T$ a $\mathbf{f} = (f_1, ..., f_N)^T$.

Příklad 4.1 Nechť $N = 2$ a $f_1(x_1, x_2) = x_1^2 + x_2^2$, $f_2(x_1, x_2) = x_1^2 - x_2^2$.Potom má (4.1) tvar

$$x_1 = x_1^2 + x_2^2,$$

$$x_2 = x_1^2 - x_2^2.$$

Výraz $x_1^2 + x_2^2 - x_1 = 0$ je rovnicí kružnice se středem v bodě $(1/2, 0)$ a $x_1^2 - x_2^2 - x_2 = 0$ je rovnicí hyperboly se středem v bodě $(0, -1/2)$. Obě tyto křivky procházejí počátkem, což značí, že jedním z nulových bodů je $x_1 = 0, x_2 = 0$. Pomocí grafů pro $(x_1 - 1/2)^2 + x_2^2 = 1/4$ a $(x_2 + 1/2)^2 - x_1^2 = 1/4$ snadno nahlédneme, že v okolí bodu $\tilde{x}_1 = 0,8, \tilde{x}_2 = 0,4$ leží další (netriviální) nulový bod soustavy (4.1).

Při analýze důkazů konvergence metod studovaných v předchozí kapitole zjistíme, že významným důkazovým prostředkem byly vlastnosti absolutní hodnoty reálných čísel:

$$|x| = 0 \Leftrightarrow x = 0;$$

$$|cx| = |c||x|;$$

a

$$|x + y| \leq |x| + |y|$$

pro c, x a $y \in R$.

V souvislosti s prostory R^N dimenze $N \geq 2$, to vede k nutnosti zobecnit pojem absolutní hodnoty. Matematika na přelomu 19. a 20. století dospěla při tomto zobecňování k pojmu *normy* a *normovaného prostoru*.

Definice 4.1 *Buď E lineární prostor nad tělesem reálných čísel. Funkce $\nu : E \to R^1_+$ se nazývá* normou *na E, jestliže platí následující relace:*

(i) $\nu(x) = 0 \Leftrightarrow x = 0$;
(ii) $\nu(cx) = |c|\nu(x)\ x \in E, c \in R^1$;
(iii) $\nu(x + y) \leq \nu(x) + \nu(y),\ x, y \in E$.

Existuje-li na E norma, tak se prostor E se nazývá normovaným.

Poznámka. Obvykle se norma na E označuje symbolem $||\ .\ ||$. Je-li zapotřebí zvýraznit, že norma je brána na prostoru E, pak se značí následovně $||\ .\ ||_E$.

Definice 4.2 *Posloupnost $\{x_n\},\ x_n \in \mathcal{E}$ se nazývá* konvergentní *k $x \in \mathcal{E}$ a x její limitou, jestliže platí*

$$\lim_{n\to\infty} ||x_n - x|| = 0.$$

Definice 4.3 *Posloupnost $\{x_n\},\ x_n \in \mathcal{E}$, se nazývá* cauchyovskou, *jestliže pro každé $\varepsilon > 0$ existuje N_0 takové, že $||x_{n+p} - x_n|| < \varepsilon$ pro $n \geq N_0$ a $p \geq 1$.*

Normovaný prostor E se nazývá Banachovým prostorem *je-li v něm každá cauchyovská posloupnost konvergentní.*

Příklad 4.2 Nechť $\mathcal{E} = \mathcal{R}^{\mathcal{N}}$ a

(a)

$$||\mathbf{x}||_p = \left(\sum_{k=1}^{N} |x_k|^p\right)^{1/p},\ 1 \leq p < +\infty,$$

(b)

$$||\mathbf{x}||_\infty = \max\{|x_k| :\ k = 1, ...,\ N\}.$$

Všímněme si, že pro $N = 1$

$$\|\mathbf{x}\|_\infty = \|x\|_p = |x|,$$

takže absolutní hodnota v R^1 je norma.

Dále nechť $C = (c_{jk})$ je regulární matice typu $N \times N$. Potom výraz

$$||x||_C = ||Cx||_{R^N}$$

je norma na R^N.

Speciálně nechť $\mathcal{E} = (R^N,\ ||.||_2)$ *a* $C = (c_{jk})$, $detC \neq 0$.

Dále nechť

$$(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{N} x_k y_k.$$

Položme $B = C^T C$ *a*

$$||\mathbf{x}||_{2,B}^2 = (B\mathbf{x}, \mathbf{x}).$$

Snadno se ukáže, že $||.||_{2,B}$ je norma na R^N.

Vraťme se však k vyšetřování soustav nelineárních rovnic v R^N.

Metoda postupných aproximací je definována podobně jako pro případ jedné rovnice.

Algoritmus 4.1 *Zvolme prvek* $\mathbf{x} \in \mathcal{E}$ *a počítejme posloupnost vektorů* $\{\mathbf{x}_n\}$ *rekurzívně pomocí formule*

$$\mathbf{x}_{n+1} = \mathbf{f}(\mathbf{x}_n),\ n = 1,\ ... , \tag{4.3}$$

Opět si klademe otázku, zda je proces (4.3) konvergentní a jaké vlastnosti má případná limita; zřejmě se nabízí řešení rovnice (4.2).

Odpovědi na položené otázky jsou obsahem následujícího tvrzení.

Věta 4.1 *Nechť* $\mathcal{I} = \prod_{j=1}^{N} \mathcal{I}_j$, *kde* $\mathcal{I}_j = [a_j,\ b_j],\ j = 1,\ ...,\ N$, *a nechť funkce* $f_1,\ ...\ f_N$ *vyhovují následujícím podmínkám:*

(i) $f_1,\ ...,\ f_N$ *jsou definovány a jsou spojité na* $\mathcal{I}$*;*

(ii) Pro každý prvek $\mathbf{x} \in \mathcal{I}$ *též vektor* $\mathbf{f}(\mathbf{x}) \in \mathcal{I}$*;*

(iii) Existuje konstanta L, $0 < L < 1$, *taková, že pro libovolná* $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{I}$ *platí*

$$||\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)|| \leq L||\mathbf{x}_1 - \mathbf{x}_2||. \tag{4.4}$$

Potom platí následující výroky

(a) Rovnice (4.2) má právě jedno řevsení $\mathbf{x}^* \in \mathcal{I}$.

(b) Pro libovolný výchozí prvek $\mathbf{x}_0 \in \mathcal{I}$ *posloupnost* $\{\mathbf{x}_n\}$ *určená v algoritmu 4.1 je definována pro každé* n *a konverguje k* $\mathbf{x}^*$.

(c) Pro libovolné $n \geq 1$ *platí odhad*

$$||\mathbf{x}_n - \mathbf{x}^*|| \leq \frac{L^n}{1 - L}||\mathbf{x}_1 - \mathbf{x}_0||. \tag{4.5}$$

Definice 4.4 *Konstanta* L, *požadovaná v (4.4) se nazývá* Lipschitzovou konstantou *(vzhledem k normě* $||.||$*) vektorové funkce* $\mathbf{f}$. *Vztah (4.4) se nazývá* Lipschitzovou podmínkou *pro* $\mathbf{f}$ *vzhledem k normě* $||.||$.

Poznámky. Podmínky kladené ve větě 4.1 na Lipschitzovu konstantu L napovídají interpretaci metody. Vzdálenost obrazů je při takovém zobrazení $\mathbf{f}$ ostře majorizována vzdáleností vzorů, což dává takovému zobrazení jméno: *kontraktivní zobrazení*, čili *kontrakce*.

Je na místě podotknout, že tvrzení (a) je v případě $N \geq 2$ daleko hlubším výrokem než je tomu pro $N = 1$, kdy existenci řešení lze nahlédnout na př. vyšetřením grafu zkoumané funkce. Tvrzení (b) vypovídá o konvergenci algoritmu, zatímco v (c) je dán horní odhad chyby.

Důkaz věty 4.1. Z požadavku (ii) plyne, že posloupnost $\{\mathbf{x}_n\}$ je definována pro každé $n \geq 1$ a $\mathbf{x}_0 \in \mathcal{I}$.

Na základě (iii) odvodíme snadno, že

$$||\mathbf{x}_{n+1} - \mathbf{x}_n|| \leq L^n ||\mathbf{x}_1 - \mathbf{x}_0||, \ n = 0, \ 1, ... \tag{4.6}$$

Fixujme n pevně a vezměme $m > n, \ m = n + p, \ p \geq 1$. Vidíme, že

$$\mathbf{x}_m - \mathbf{x}_n = \sum_{k=1}^{p} (\mathbf{x}_{n+k} - \mathbf{x}_{n+k-1}),$$

a tedy,

$$||\mathbf{x}_m - \mathbf{x}_n|| \leq \sum_{k=1}^{p} ||\mathbf{x}_{n+k} - \mathbf{x}_{n+k-1}| \, |.$$

Použitím (4.6) obdržíme vztahy

$$\begin{aligned} ||\mathbf{x}_m - \mathbf{x}_n|| \ &\leq \textstyle\sum_{k=1}^{p} L^{n+k-1} \, ||\mathbf{x}_1 - \mathbf{x}_0| \, | \\ &\leq \tfrac{L^n}{1-L} ||\mathbf{x}_1 - \mathbf{x}_0||. \end{aligned} \tag{4.7}$$

Protože $\mathcal{E} = R^N$ je úplný (Banachův) prostor, plyne odtud konvergence posloupnosti $\{\mathbf{x}_n\}$.

Nechť

$$\hat{\mathbf{x}} = \lim_{n \to \infty} \mathbf{x}_n,$$

což je totéž co

$$\lim_{n \to \infty} ||\mathbf{x}_n - \hat{\mathbf{x}}|| = 0.$$

Vzhledem ke spojitosti funkcí $f_1, \ ..., \ f_N$ obdržíme z (4.7)

$$\lim_{n \to \infty} \mathbf{f}(\mathbf{x}_n) = \mathbf{f}(\hat{\mathbf{x}})$$

a tudíž

$$\hat{\mathbf{x}} = \lim_{n \to \infty} \mathbf{x}_{n+1} = \lim_{n \to \infty} \mathbf{f}(\mathbf{x}_n) = \mathbf{f}(\hat{\mathbf{x}}),$$

a tak $\hat{\mathbf{x}} = \mathbf{x}^*$ je řešením soustavy rovnic (4.2).

Jednoznačnost se dokazuje zcela analogicky jako v případě $N = 1$.

Nechť $\mathbf{x}_1^*$ a $\mathbf{x}_2^*$ jsou dvě řešení soustavy (4.2). Potom, na základě (4.4)

$$||\mathbf{x}_1^* - \mathbf{x}_2^*|| = ||\mathbf{f}(\mathbf{x}_1^*) - \mathbf{f}(\mathbf{x}_2^*)|| \leq L||\mathbf{x}_1^* - \mathbf{x}_2^*||,$$

což je možné pouze když $\mathbf{x}_1^* = \mathbf{x}_2^*$.

Platnost (4.5) plyne z (4.7) pro $p \to \infty$:

$$||\mathbf{x}_n - \mathbf{x}^*|| = \lim_{n \to \infty} ||\mathbf{x}_n - \mathbf{x}_{n+p}|| \leq \frac{L^n}{1 - L} ||\mathbf{x}_1 - \mathbf{x}_0||.$$

Tím je důkaz věty 4.1 proveden. |||

4.2 Kvadratická konvergence a Newtonova metoda pro soustavy

Zabývejme se otázkou explicitního vyjádření vektoru chyby

$$\mathbf{d}_n = \mathbf{x}_n - \mathbf{x}^*$$

při $n \to \infty$.

Předpokládejme, že $\mathbf{f}$ splňuje předpoklady věty 4.1 a navíc, že funkce $f_1, ..., f_N$ mají v $\mathcal{I}$ spojité parcální derivace až do řádu 2 včetně. Používáme značení

$$(f_j)_{x_k}(x_1, ..., x_N) \equiv \frac{\partial f_j}{\partial x_k}(x_1, ..., x_N),\ j, k = 1, ..., N$$

a

$$\begin{aligned}(f_j)_{x_k,x_l}(x_1, ..., x_N) &\equiv \tfrac{\partial^2 f_j}{\partial x_k \partial x_l}(x_1, ..., x_N) \\ &= \tfrac{\partial (f_j)_{x_k}}{\partial x_l}(x_1, ..., x_N),\ j, k, l = 1, ..., N\end{aligned}$$

Aplikací Taylorovy věty pro funkce N proměnných zjistíme, že

$$\text{(4.8)}\ \mathbf{d}_{n+1} = \mathbf{x}_{n+1} - \mathbf{x}^* = \mathbf{f}(\mathbf{x}_n) - \mathbf{f}(\mathbf{x}^*) = \mathbf{f}(\mathbf{x}^* + \mathbf{d}_n) - \mathbf{f}(\mathbf{x}_n) = \mathcal{J}(\mathbf{x}^*)\mathbf{d}_n + O(||\mathbf{d}_n||^2),$$

kde

$$\text{(4.9)}\qquad \mathcal{J}(\mathbf{x}) = \begin{pmatrix} (f_1)_{x_1}(x_1, ..., x_N) & ... & (f_1)_{x_N}(x_1, ..., x_N) \\ ... & ... & ... \\ (f_N)_{x_1}(x_1, ..., x_N) & ... & (f_N)_{x_N}(x_1, ..., x_N) \end{pmatrix}$$

a $O(||\mathbf{d}_n^2||)$ značí výrazy mající majorantu tvaru $c\ ||\mathbf{d}_n||^2$, při čemž c je konstanta nezávislá jak na n tak na $\mathbf{x} \in \mathcal{I}$.

Definice 4.5 *Matice definovaná v (4.9) se nazývá* Jacobiovou maticí *soustavy (4.1) resp. (4.2).*

Všimněme si toho, že vztah (4.8) je N-dimenzionálním analogem vztahu (3.12).

Je-li $\mathcal{J}(\mathbf{x}^*)$ nenulová matice, pak (4.8) říká, že vektor chyby se v každém iteračním kroku násobí Jacobiovou maticí (4.9), což odpovídá *lineární konvergenci.* Je-li $\mathcal{J}(\mathbf{x}^*) = 0$, je z (4.8) patrné, že chyba po 1+n- tém kroku je úměrná čtverci chyby po n-tém kroku - v tom případě máme co do činění s *kvadratickou konvergencí.*

Opět můžeme formulovat analog věty 3.3 pro $N > 1$.

Věta 4.2 *Nechť funkce f_1, ..., f_N definované na $\mathcal{I}$ splňují na $\mathcal{I}$ tyto požadavky:*
(i) Existují spojité derivace

$$\frac{\partial f_j}{\partial x_l},\ 1 \le j, l \le N;$$

(ii) Soustava (4.1) má uvnitř $\mathcal{I}$ řešení $\mathbf{x}^*$ takové, že $\mathcal{J}(\mathbf{x}^*) = 0$.

Potom existuje číslo $\delta > 0$ takové, že posloupnost definovaná pomocí algoritmu 4.1 konverguje k $\mathbf{x}^*$ pro libovolný startovací vektor $\mathbf{x}_0$, pro nějž

$$||\mathbf{x}_0 - \mathbf{x}^*|| < \delta.$$

Tedy opět, jako pro $N = 1$, je-li výchozí přiblížení dostatečně blízké přesnému řešení, dochází ke konvergenci.

Globální konvergence je pro vícerozměrný případ problém principiálně komplikovanější než pro $N = 1$ (konvexita ev. konkavita řešení se jeví pro $N = 1$ dost akademicky). Podmínky zaručující globální konvergenci jsou dost složité a jejich analýza patří mezi speciální partie numerických metod.

Hledejme řešení soustavy

$$\mathbf{F}(\mathbf{x}) = 0, \tag{4.10}$$

kde $\mathbf{F} = (F_1, ..., F_N)^T$.

Algoritmus 4.2 *(Newtonova metoda pro soustavy) Volme $\mathbf{x}_0 \in \mathcal{I}$ a sestrojme posloupnost $\{\mathbf{x}_n\}$ podle schematu*

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mathbf{h}_n, \tag{4.11}$$

kde $\mathbf{h}_n = (h_1, ..., h_N)^T$ je řešením soustavy

$$\mathcal{J}(\mathbf{x}_n)\mathbf{h}_n = \mathbf{F}(\mathbf{x}_n) \tag{4.12}$$

a

$$\mathcal{J} = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_N} \\ . & \cdots & . \\ \frac{\partial F_N}{\partial x_1} & \cdots & \frac{\partial F_N}{\partial x_N}. \end{pmatrix}$$

Jak je patrné z (4.11) při provádění algoritmu 4.2 musíme při každém iteračním kroku řešiti soustavu lineárních algebraických rovnic (4.12). Jest proto přirozené požadovati

$$det\mathcal{J}(\mathbf{x}_n) \neq 0,\ n = 0,\ 1, ... \tag{4.13}$$

Cvičení 4.1 *Nalezněte matici $H = H(\mathbf{x})$ tak, aby užití postupných aproximací na rovnici*

$$\mathbf{x} = \mathbf{x} - H(\mathbf{x})F(\mathbf{x})$$

poskytlo kvadratickou konvergenci. Rozhodněte, zda se vždy obdrží Newtonova metoda.

5 Řešení soustav lineárních algebraických rovnic

5.1 Obecné soustavy a řešení ve smyslu nejmenších čtverců

Budeme se zabývat řešením soustav typu

$$Ax = b, \tag{5.1}$$

kde

$$x \in R^N,\ b \in R^M,$$

$$A = (a_{jk}),\ 1 \le j \le M,\ 1 \le k \le N.$$

Poznámka. Je zřejmé, že lze vždy docílit situace, kdy $M = N$ a to rozšířením matice A o vhodný počet nulových sloupců pro případ $N < M$ či řádků pro případ $N > M$.

Řešitelnost soustavy v klasickém smyslu, t. j. kdy vztah (5.1) je chápán jako soustava rovností

$$\sum_{k=1}^{N} a_{jk}x_k = b_j,\ j = 1, ..., M,$$

nelze vyžadovat.

Jako příklad může sloužit častý případ, kdy soustava (5.1) representuje nějaký model v němž některé z prvků a_{jk} a b_j se obdrží měřením. Obecně nelze vyloučit, aby odpovídající soustava pro dvě různá měření byla v jednom případě řešitelná a v druhém nikoliv.

V praktických aplikacích se ukazuje potřeba následujícího tvrzení.

Každá soustava typu (5.1) má právě jedno "řešení".

Takové tvrzení nemůže platit pro klasické řešení (podmínka o hodnostech). Aby požadované tvrzení platilo, musíme vhodným způsobem zobecnit pojem řešení. Proto zavedeme *zobecněné řešení*.

Buďte $(.\ ,.)_M$ a $(.\ ,.)_N$ skalární součiny na R^M a R^N.

Definujme

$$F(x,\ x) = (Ax - b,\ Ax - b)_M \tag{5.2}$$

a hledejme $\overline{x} \in R^N$ tak, aby

$$F(\overline{x}) = min\{F(x) :\ x \in R^N\}. \tag{5.3}$$

Vektor $\overline{x}$ splňující relaci (5.3) se nazývá *zobecněné řešení soustavy* (5.1).

Věta 5.1 *Existuje alespoň jeden vektor $\overline{x} \in R^N$ takový, že platí (5.3).*

Důkaz. Definujme *obor hodnot matice A*

$$range(A) = \left\{u \in \mathcal{R}^M : \exists x \in \mathcal{R}^N \text{ for which } u = Ax\right\}$$

a *nulovou množinu matice* neboli *jádro A*

$$ker(A) = \left\{x \in \mathcal{R}^N : x \in \mathcal{R} \text{ for which } Ax = 0\right\}.$$

Snadno ověříme, že platí relace

$$R^M = rangeA \oplus kerA^* \tag{5.4}$$

$$R^N = rangeA^* \oplus kerA, \tag{5.5}$$

kde A^* je definována pomocí vztahu $A^*v = y$, při čemž

$$(Ax, v)_M = (x, y)_N$$

pro libovolná $x \in R^N$ a $v \in R^M$.

Existence i jednoznačnost y je důsledkem Rieszovy věty o reprezentaci spojitého lineárního funkcionálu na Hilbertově prostoru $(R^N, (.\ ,\ .)_N)$.

Na základě (5.4) a (5.5) obdržíme, že platí

$$x = A^*u + v = w + v,\ Av = 0,$$

a

$$b = Ac + d,\ A^*d = 0.$$

Odtud pak

$$F(x) = (Aw - Ac,\ Aw - Ac)_M + (d,\ d)_M.$$

Vzhledem k tomu, že d je pevný vektor, stačí tedy položiti

$$\bar{x} = c,$$

takže

(5.6) $$F(\bar{x}) = (d,\ d)_M = min\{F(x) : x \in R^N\}.$$

Tím je věta 5.1 dokázána.

Poznámka. Protože (5.6) je hodnota globálního minima kvadratické funkce F, je veličina (5.6) táž pro všechna eventuální zobecněná řešení x soustavy $Ax = b$. Tato hodnota se nazývá *cenou* uvedené úlohy najít zobecněné řešení, neboli cenou zobecněného řešení. Všimněme si té skutečnosti, že cena zobecněného řešení x je nulová právě když x je řešení klasické.

Nechť

$$\mathcal{M} = \{x \in R^N :\ F(x) = (d,\ d)_M\}.$$

Z věty 5.1 plyne, že $\mathcal{M} \neq \emptyset$ a ze spojitosti F pak, že $\mathcal{M}$ je uzavřená. Díky subaditivitě normy snadno nahlédneme, že $\mathcal{M}$ je konvexní. Existuje tedy jediný prvek $x^* \in \mathcal{M}$ takový, že

$$||x^*||_N = min\{||x||_N :\ x \in \mathcal{M}\}.$$

Zobecněné řešení x^* se nazývá *normálním řešením* soustavy (5.1).

Věta 5.2 *Existuje právě jedno normální řešení libovolné soustavy $Ax = b$.*

Vidíme tedy, že námi požadované tvrzení uvedené na počátku této kapitoly, platí pro námi zavedené zobecněné řešení.

Poznámka. Zobecněná řešení a tedy též normální řešení jsou podstatně závislá na skalárních součinech $(.\ ,\ .)_M$ a $(.\ ,\ .)_N$.

Volíme-li speciálně

$$(u,v)_M = \sum_{j=1}^{M} (u)_j (v)_j, \tag{5.7}$$

kde $(u)_j$ značí j-tou komponentu vektoru $u \in R^M$, obdržíme funkci F ve tvaru

$$F(x) = \sum_{j=1}^{M} \left[\sum_{k=1}^{N} a_{jk}(x)_k - (b)_j \right]^2 .$$

Odtud získalo zobecněné řešení takto zavedené název *řešení ve smyslu nejmenších čtverců*. Řešení ve smyslu nejmenších čtverců patří mezi nejrozšířenější. Toto zobecněné řešení bylo zavedeno a velmi zevrubně aplikováno již Gaussem.

Zobecněné inverzní operátory

Buď A matice typu $M \times N$.
Vyšetřujme soustavu relací

(i)$AXA = A$, (ii)$XAX = X$,

(iii) $(AX)^* = AX$, (iv) $(XA)^* = XA$.

Příklad.

Nechť $M = N$ a $\det A \neq 0$.
Položme $X = A^{-1}$.
Splnění vztahů (i) - (iv) s $X = A^{-1}$ je očividné.

Napřed si ukážeme, že matice X typu $N \times N$ splňující vztahy (i) - (iv) existuje nejvýše jedna.

Buďte tedy X_1 a X_2 dvě takové matice. Jest potom

$$X_1 = X_1 A X_1 = X_1 (A X_1)^* = X_1 (A X_2 A X_1)^* =$$

$$X_1 (A X_1)^* (A X_2)^* = X_1 A X_1 A X_2 = X_1 A X_2 =$$

$$(X_1 A)^* X_2 A X_2 = (X_1 A)^* (X_2 A)^* X_2 .$$

Na druhé straně,

$$(X_2 A X_1 A)^* X_2 = X_1 A X_2 A X_2 = (X_1 A)^* (X_2 A)^* X_2 ,$$

takže

$$X_1 = (X_1 A X_2 A)^* X_2 = X_2 A X_2 = X_2 .$$

Již víme, že v případě $M = N$ a $\det A \neq 0$ jediným řešením soustavy vztahů (i) - (iv) je inverzní matice A^{-1}.

Existuje řešení v obecném případě?
Ano!

Položme

$$T = A^* A.$$

Snadno nahlédneme, že matice T je typu $N \times N$, je symetrická a též je *pozitivně semi-definitní*, t. j.

$$(Tx, x)_N \geq 0.$$

Existují tedy projekce $P_j,\ j = 1, ...s,\ s \leq N$ takové, že platí vztahy

$$T = \sum_{k=1}^{s} \lambda_k P_k,\ TP_0 = 0, \tag{5.8}$$

při čemž

$$P_j P_k = P_k P_j = \delta_{jk} P_k,\ P_j^* = P_j,\ \sum_{k=1}^{s} P_k = I - P_0,\ j, k = 0, 1, ..., N \tag{5.9}$$

a dále pak

$$(T - \lambda_j) P_j = 0,\ \lambda_j > 0,\ j = j, ..., N,\ \lambda_0 = 0. \tag{5.10}$$

Položme

$$Z = \sum_{j=1}^{N} \lambda_j^{-1} P_j A^*. \tag{5.11}$$

Věta 5.3 *Ke každé matici A typu $M \times N$ existuje právě jedna matice typu $N \times M$ taková, že pro ni platí vztahy (i) - (iv).*

Definice 5.1 *Matice, jejíž existenci a jednoznačnost zaručuje věta 5.3, se nazývá* (Mooreovou - Penroseovou) pseudoinverzní maticí *a označuje se symbolem A^+.*

Důkaz. Důkaz jednoznačnosti je elementární a lze jej přenechat čtenáři.
Na základě (5.11) a (5.8) snadno zjistíme, že

$$AZA = A \sum_{j=1}^{N} \lambda_j^{-1} P_j A^* A = A \sum_{j=1}^{N} \sum_{k=1}^{N} \lambda_j^{-1} \lambda_k P_j P_k.$$

Z tohoto vyjádření pomocí (5.9) odvodíme, že

$$AZA = A(I - P_0). \tag{5.12}$$

Buď $x \in \mathcal{R}^N$. Podle předpokladu (5.8) platí, že

$$0 = (TP_0 x,\ P_0 x)_N = (AP_0 x,\ AP_0 x)_N,$$

takže

$$AP_0 x = 0.$$

Protože x je libovolný, plyne odtud, že

$$AP_0 = 0.$$

Dokázali jsme tak, že, díky (5.12), matice Z splňuje vztah (i).

Platnost zbývajících vztahů (ii) - (iv) lze dokázat analogicky, což přenecháme čtenáři a pokládáme tím větu 5.3 za dokázanou s tím, že

$$A^+ = Z,$$

při čemž Z je definována v (5.11). □

Položme

(5.13) $$x^+ = A^+ b.$$

Dokážeme, že platí

Věta 5.4 *Buďte A matice typu $M \times N$ a $b \in \mathcal{R}^M$, $b = b_1 + b_2$, $b_1 \in \mathrm{range}(A)$, $b_2 \in \ker(A)$.*
Potom platí rovnost

(5.14) $$x^+ = x^*.$$

Důkaz. Zřejmě

$$AA^+ b = b_1$$

a proto

$$F(x^+) = \left(Ax^+ - b,\ Ax^+ - b\right)_M = \left(AA^+ b - b,\ AA^+ b - b\right)_M = (b_2,\ b_2)_M.$$

Stačí tedy ukázati, že $x^+ \in R(A^*)$. To však plyne odtud, že

$$P_0 A^+ = \sum_{j=1}^{s} \frac{1}{\lambda_j} P_0 P_j A^* = 0$$

a tedy též $P_0 x^+ = 0$ a tudíž, $x^+ = (I - P_0)x^+$. Protože $A^+A = I - P_0$ a $(I - P_0)range(A^*) = range(A^*)$, je tvrzení dokázáno. □

Vyšetřujme opět soustavu $Ax = b$, kde A je $M \times N$ matice a $b \in \mathcal{R}^N$ a $b = b_1 + b_2$, při čemž $b_1 \in range(A)$ t.j. $b_1 = Ac_1$, dále pak a $A^* b_2 = 0$.

Snadno zjistíme, že platí

$$A^*Ax = A^*b = A^*b_1,$$

kterážto soustava se nazývá *normální soustavou.* Tato soustava má vždy klasické řešení. To vyplývá odtud, že $(A^*A)^* = A^*A$ a pro každé řešení y homogenní soustavy $A^*Ay = 0$ platí vztahy

$$(y, A^*b)_N = (y, A^*Ac_1)_N = (A^*Ay, c_1)_N = 0.$$

Hledat řešení normální soustavy rovnic je numericky velmi náročné, protože číslo podmíněnosti

$$\kappa(A^*A) = \frac{s_{\max}}{s_{\min}},$$

kde

$$s_{\max} = \|A^* A\|_2^{1/2}$$

a $s_{\min}$ je odmocnina nejmenší kladné vlastní hodnoty matice A^*A, je zpravidla velmi veliké; v případě čtvercové matice A je rovno čtverci čísla podmíněnosti matice A. Připomeňme, že číslo podmíněnosti čtvercové matice A je definováno jakožto výraz

$$\|A\|\|A^{-1}\|.$$

Normální řešení je tedy (klasické) řešení normální soustavy ležící v $range(A^+)$.
Dále platí

Lemma 5.1 *Každé zobecněné řešení $\overline{x} \in \mathcal{M}$ má tvar*

$$\overline{x} = x^+ + y, \tag{5.15}$$

kde

$$Ay = 0.$$

Důkaz. Buď $\overline{\overline{x}}$ další řešení soustavy $Ax = b$ a nechť

$$\overline{\overline{x}} = a_1 + a_2,\ a_1 = A^* c_1,\ Aa_2 = 0.$$

Potom platí, že

$$\overline{x} - \overline{\overline{x}} = x^+ - a_1 + y - a_2.$$

Protože $Aa_1 = b_1$, odvodíme, že platí vztahy

$$A(x^+ - a_1) = AA^+ b_1 - b_1 = 0.$$

Jest tudíž

$$\|A^+ b_1 - A^* c_1\|^2 = (A^* \tilde{b}_1 - A^+ c_1)_N = (\tilde{b}_1 - c_1, A[(x^+ - a_1)_M = 0,$$

kde $A^+ b_1 = A^* \tilde{b}_1$ a tedy

$$\overline{\overline{x}} = \overline{x} = x^+.$$

Protože $\overline{\overline{x}} - \overline{x} \in ker(\mathrm{A})$, tvrzení lemmatu je dokázáno. □

5.2 Pomocné prostředky ke konstrukci speciálních reprezentací matic

Nejprve si zopakujme některé pojmy, jež se ukáží jako vhodné pro konstrukci a vyšetřování určitých algoritmů.

Předpokládáme, že všechny matice v tomto článku jsou obecně komplexní typu $N \times N$.
Matice A je *hermiteovsky sdružená*, jestliže

$$A^H = A,\ A^H = (a^*_{jk})$$

při čemž

$$a^*_{jk} = \bar{a}_{kj},\ j, k = 1, ..., N,$$

kde pro $\alpha = a + ib$, $a, b \in \mathcal{R}^1, i^2 = -1$, $\overline{\alpha} = a - ib$. Je-li matice A reálná, t. j. je-li

$$\bar{a}_{jk} = a_{jk}$$

nazýváme matici hermiteovskou *symetrickou*. Jest totiž

$$A^T = A,\ A^T = (a^T_{jk}),\ a^T_{jk} = a_{kj},\ j, k = 1, ..., N.$$

Permutační matice $P = (p_{jk})$ splňuje vztahy

$$p_{jk} = \left\{ \begin{array}{l} 0 \\ 1 \end{array} \right. ,\ \sum_{k=1}^{N} p_{jk} = \sum_{k=1}^{N} p_{kj} = 1,$$

a

$$\det P \neq 0.$$

Zřejmě

$$P^H = P^T = P^{-1}.$$

Projekční maticí či *projekcí* se nazývá matice splňující

$$P^2 = P.$$

Speciálně tedy, $I^2 = I$ a $0^2 = 0$.

Permutační matice, která realizuje záměnu l - té a k - té složky vektorůz $\mathcal{R}^N$, se nazývá *elementární permutační* maticí neboli *transpoziční* maticí.

Taková elementární permutační matice má tvar

$$\begin{array}{cc} & \begin{array}{ccccccccccc} & & & & \mathrm{k} & & & & \mathrm{l} & & \end{array} \\ & \begin{array}{ccccccccccc} & & & & \downarrow & & & & \downarrow & & \end{array} \\ \begin{array}{c} \\ \\ \\ \\ k \ \rightarrow \\ \\ l \ \rightarrow \\ \\ \\ \\ \end{array} & \left(\begin{array}{ccccccccccc} 1 & . & . & . & 0 & . & . & . & 0 & . & . & . & 0 \\ . & . & & & & & & & & & & & . \\ . & & . & & & & & & & & & & . \\ . & & & . & & & & & & & & & . \\ 0 & & & & 0 & . & . & . & 1 & & & & 0 \\ . & & & & . & . & . & . & . & & & & . \\ 0 & & & & 1 & . & . & . & 0 & & & & 0 \\ . & & & & & & & & & . & & & . \\ . & & & & & & & & & & . & & . \\ . & & & & & & & & & & & . & . \\ 0 & . & . & . & 0 & . & . & . & 0 & . & . & . & 1 \end{array} \right) \end{array}$$

a při její aplikaci platí vztahy $y = Px$, při čemž

$$y_j = x_j,\ j \neq k, l,$$

zatímco

$$y_k = x_l \text{ a } y_l = x_k.$$

Unitární matice U je definová na vztahy

$$UU^H = U^H U = I,$$

což pro reálnou matici V značí, že

$$VV^T = V^T V = I.$$

V takovém pří padě hovoříme o matici *ortogonální*.

Pro *involutorní* matici B platí

$$B^2 = I.$$

Tedy speciálně $I^2 = I$.

Buďte $y \in \mathcal{R}^N$ $v \in \mathcal{R}^M$. Definujeme matici

$$y \otimes v$$

kladouce

$$[y \otimes v]\, x = (x, v)_N y.$$

Jinými slovy,

$$y \otimes v = y\, v^T.$$

Matice E_k se nazývá *elementární dolní trojúhelníkovou indexu* k, jestliže

$$E_k = I_N - v \otimes e_k$$

a přitom

$$e_j^T v = 0,\ j = 1, ..., k.$$

To značí, že

$$E_k = \begin{pmatrix} 1 & & & & & & & \\ & \cdot & & & & & & \\ & & \cdot & & & & & \\ & & & \cdot & & & & \\ & & & & 1 & & & \\ & & & & -v_{k+1} & 1 & & \\ & & & & \cdot & & \cdot & \\ & & & & \cdot & & & \cdot \\ & & & & \cdot & & & & \cdot \\ & & & & -v_N & & & & 1 \end{pmatrix}.$$

Snadno prověříme, že $y \otimes v$ splňuje následující relaci

$$[y \otimes v]^2\, x = (y, v)_N\, [y \otimes v]\, x.$$

takže, je - li $(y, v)_N = 0$,

$$[y \otimes v]^2 = 0$$

a tedy,

$$\sigma(y \otimes v) = \{0\}$$

neboli, $y \otimes v$ je *nilpotentní matice*.

V případě dolní elementární matice E_k vidíme, že $(e_k, v)_N = 0$, takže platí

Věta 5.5 *Buď* $E_k = I_N - v_k \otimes e_k$, *při čemž*

$$(v_k, e_j)_N = 0,\ j = 1, ..., k.$$

Potom existuje inverzní matice E_k^{-1} *a platí*

$$E_k^{-1} = I_N + v_k \otimes e_k.$$

Věta 5.6 *Nechť* $x^T = (\xi_1, ..., \xi_N)$ *a nechť* $0 \neq \xi_k = e_k^T x$.

Potom existuje jednoznačně určená elementární dolní trojúhelníková matice indexu k E_k *taková, že*

$$E_k x = (\xi_1, ..., \xi_k, 0, ..., 0)^T.$$

Důkaz. Hledejme E_k ve tvaru

$$E_k = I_N - v_k \otimes e_k.$$

Z požadavku elementarity E_k plyne, že

$$\nu_1 = ... = \nu_k = 0, \tag{5.16}$$

kde

$$v_k^T = (\nu_1, ..., \nu_N).$$

Dále pak

$$E_k x = x - (x, e_k)_N v = x - \xi_k v$$

a protože žádáme, aby

$$\xi_j - \xi_k \nu_j = 0,\ j = k+1, ..., N, \tag{5.17}$$

z předpokladu $\xi_k \neq 0$ odvodíme vztahy

$$\nu_j = \frac{\xi_j}{\xi_k},\ j = k+1, ..., N. \tag{5.18}$$

Tedy, E_k existuje a je vztahy (5.16), (5.17), (5.18) určena jednoznačně. |||

Householderovu matici zavedeme pomocí vztahů

$$H = I - 2ww^H,\ w^H w = 1,$$

kde w je (obecně komplexní) sloupcový vektor a w^H odpovídající vektor řádkový.

Zřejmě Householderova matice je hermiteovsky sdružená $H^H = H$ a též unitární , neboť

$$H^H H = H^2 = (I - 2ww^H)^2 = I.$$

Všimněme si, že pro $x \in C^N$ a $y = Hx$ platí, že

$$y = x - 2w^H x w,$$

takže

$$y^H y = (y, y) = (Hx, Hx) = (H^H Hx, x) = (x, x)$$

a

$$(x, y) = \overline{(y, x)}.$$

Lemma 5.2 *K danému vektoru $v \neq 0$ existuje ortogonální matice Q taková, že*

$$Qv = -\sigma||v||e_1, \tag{5.19}$$

kde

$$\sigma = \begin{cases} +1 & pro \quad v_1 = (v, e_1) \geq 0 \\ -1 & pro \quad v_1 < 0. \end{cases}$$

Důkaz. Nechť

$$u = v + \sigma\sqrt{(v,v)}e_1$$

a

$$Q = I - 2\frac{uu^H}{u^H u}.$$

Zřejmě

$$Q^H = I - 2\frac{uu^H}{u^H u} = Q$$

a

$$Q^H Q = Q^2 = I.$$

Dále pak

$$Qv = Qu - \sigma\|v\|Qe_1.$$

Avšak

$$Qu = -u$$

a

$$Qe_1 = e_1 - 2\sigma\|v\|u(u)_1\frac{1}{(u,u)}.$$

Odtud plyne, že

$$Qv = -u - \sigma\|v\|e_1 + 2\sigma^2\|v\|^2\frac{(u)_1}{(u,u)}u,$$

a protože

$$(u)_1 = (v)_1 + \sigma\|v\|,$$

zjistíme, že

$$-u + 2\sigma\|v\|\frac{(v)_1 + \sigma\|v\|}{(u,u)} = \frac{1}{(u,u)}\left\{-2\sigma^2(v,v) - 2\sigma\|v\|(v)_1 + 2\sigma^2\|v\|^2 + 2\sigma\|v\|(v)_1\right\} = 0.$$

Tedy

$$Qv = -\sigma\|v\|e_1.$$

□

Matice $H = (h_{jk})$ taková, že

$$h_{jk} = 0 \text{ pro } j - k > 1 \text{ a } j, k = 1, ..., N-2,$$

se nazývá *horní Hessenbergova*.

Názorně je patrno rozmístění prvků na následujícím schématu

$$\begin{pmatrix} \times & \times & \times & . & . & . & \times & \times \\ \times & \times & \times & . & . & . & \times & \times \\ 0 & \times & \times & . & . & . & \times & \times \\ 0 & 0 & \times & . & . & . & \times & \times \\ . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & \times & \times \end{pmatrix}.$$

Lemma 5.3 *Ke každé matici A existuje unitární matice T, taková, že*

$$A = THT^H,$$

kde matice H je horní Hessenbergova.

Čtvercová matice $A = (a_{jk})$ $a_{jk} \in \mathcal{R}$ $j, k = 1, ..., N$, se nazývá *slabě diagonálně dominantní*, jestliže platí

$$\sum_{k=1, k\neq j}^{N} |a_{jk}| \leq |a_{jj}|, \ j = 1, ..., N, \tag{5.20}$$

při čemž existuje alespoň jeden index j_0, $1 \leq j_0 \leq N$, tak, že

$$\sum_{k=1, k\neq j_0}^{N} |a_{j_0 k}| < |a_{j_0 j_0}|. \tag{5.21}$$

Nastane-li v (5.20) ostrá nerovnost pro všechny indexy $j = 1, ..., N$, pak se matice A nazývá *diagonálně dominantní.*

Věta 5.7 *(Geršgorinova) Buď $A = (a_{jk})$ matice $N \times N$, při čemž a_{jk} jsou komlexní čísla. Dále buď λ vlastní hodnota matice A. Potom existuje index j_0, $1 \leq j_0 \leq N$ takový, že platí vztahy*

$$|\lambda - a_{j_0 j_0}| \leq \sum_{k=1, k\neq j_0}^{N} |a_{j_0 k}|. \tag{5.22}$$

Důkaz. Buď $x \in \mathcal{R}^N \oplus \mathrm{i}\,\mathcal{R}^N$ vlastní vektor odpovídající vlastní hodnotě λ. Tedy,

$$Ax = \lambda x, \ x \neq 0.$$

Položme $|x_{j_0}| = \max\{|x_j| \ : j = 1, ..., N\}$. Jest potom

$$\begin{aligned} |\lambda - a_{j_0}| \ &\leq \textstyle\sum_{k=1, \neq j_0}^{N} |a_{j_0 k}| \frac{|x_k|}{|x_{j_0}|} \\ &\leq \textstyle\sum_{k=1, \neq j_0}^{N} |a_{j_0 k}|. \end{aligned}$$

Je tak dokázána platnost relace (5.22) a tím i věta 5.7. |||

Důsledek 5.1 *Každá diagonálně dominantní matice A je regulární.*

Důkaz. Stačı si uvědomit, že čtvercová matice je regulární právě když 0 není její vlastní hodnotou. To, že 0 není vlastní hodnotou dané čtvercové matice však je bezprostředním důsledkem diagonální dominance. Kdyby totiž 0 byla vlastní hodnotou matice A, platily by vztahy (5.22) pro $\lambda = 0$,

$$|a_{j_0 j_0}| = |\lambda - a_{j_0 j_0}| \leq \sum_{k=1, k\neq j_0}^{N} |a_{j_0 k}| < |a_{j_0 j_0}|.$$

Tento spor dokazuje platnost tvrzení. |||

Věta 5.8 *Buď $0 \neq A = (a_{jk})$, diagonálně dominantní a buď $E_1 = I_N - v_1 \otimes e_1$ elementární dolní trojúhelníková matice indexu 1, kde*

$$v_1 = \begin{pmatrix} 0 \\ a_{21} \,/\, a_{11} \\ . \\ . \\ . \\ a_{N1} \,/\, a_{11} \end{pmatrix}.$$

Potom

$$A^{(1)} = E_1 A$$

je též diagonálně dominantní.

Důkaz. Vyšetřujme výraz a^1_{jk} rovný dle definice,

$$a^1_{jk} = a_{jk} - \frac{a_{j1}}{a_{11}} a_{1k},\ 1, ..., N.$$

Jest tedy,

$$\begin{aligned}
&|a^1_{jj}| - \textstyle\sum_{k=2,\ k\neq j}^{N} |a^1_{jk}| \\
&\geq |a_{jj}| - \tfrac{1}{|a_{11}|}|a_{j1}a_{1j}| - \textstyle\sum_{k=2,\ k\neq j}^{N} |a_{jk}| - \tfrac{1}{|a_{11}|} \textstyle\sum_{k=2,\ k\neq j}^{N} |a_{j1}a_{1k}| \\
&= |a_{jj}| - \textstyle\sum_{k=2,\ k\neq j}^{N} |a_{jk}| - \tfrac{1}{|a_{11}|} \textstyle\sum_{k=1,\ k\neq j}^{N} |a_{j1}a_{1k}| \\
&\geq |a_{jj}| - \textstyle\sum_{k=1,\ k\neq j}^{N} |a_{jk}| \\
&\geq 0,\ j = 2, ..., N.
\end{aligned}$$

|||

Čtvercová matice $A = (a_{jk})$ nad tělesem komplexních čísel se nazývá *pozitivně definitní*, jestliže existuje konstanta $\kappa > 0$ taková, že

$$(Ax, x)_N \geq \kappa (x, x)_N \ \forall x \in \mathcal{C}.$$

Jsou - li $a_{jk} \in \mathcal{R}$, pak navíc požadujeme, aby $A = A^T$.

5.3 Přímé metody řešení soustav lineárních algebraických rovnic

Pod přímými metodami řešení soustav lineárních algebraických rovnic rozumíme takové metody, které jsou založeny na algoritmech, jež poskytují přesné řešení po provedení jistého konečného počtu aritmetických operací.

Je zřejmé, že pro soustavy, jejichž matice jsou trojúhelníkové - horní či dolní - se nabízejí přirozené přímé metody. Jsou to t. zv. *zpětná* a *přímé dosazení*.

Vyšetřujme soustavy

$$Ly = c, \tag{5.23}$$

a

$$Ux = b, \tag{5.24}$$

kde

$$L = \begin{pmatrix} l_{11} & . & . & . & 0 \\ l_{21} & . & . & . & 0 \\ . & . & . & . & . \\ l_{N1}. & . & . & . & l_{NN} \end{pmatrix}$$

a

$$U = \begin{pmatrix} u_{11} & u_{12} & . & . & . & u_{1N} \\ 0 & u_{22} & . & . & . & u_{2N} \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & u_{NN} \end{pmatrix}.$$

Nechť

$$l_{jj} \neq 0,\ j = 1, ..., N$$

a

$$u_{jj} \neq 0,\ j = 1, ..., N.$$

Potom

$$y_1 = \frac{c_1}{l_{11}} \tag{5.25}$$

a

$$y_k = \frac{c_k}{l_{kk}} - \sum_{j=1}^{k-1} \frac{l_{kj} c_j}{l_{jj} l_{kk}},\ k = 2, ..., N, \tag{5.26}$$

a podobně

$$x_N = \frac{b_N}{u_{NN}}, \tag{5.27}$$

a

$$x_{N-k} = \frac{b_{N-k}}{u_{N-k,N-k}} - \sum_{j=k+1}^{N} \frac{u_{N-k,j} b_j}{u_{N-k,N-k} u_{jj}},\ k = 1, ..., N-1. \tag{5.28}$$

Nejznámější a přitom nejdůležitější přímou metodou řešení soustav linárních algebraických rovnic je *Gaussova eliminační metoda*. Její základní myšlenka spočívá v transformaci původní soustavy s maticí A na soustavu s trojúhelníkovou maticí $\tilde{A}$. Matici $\tilde{A}$ lze získat postupným

eliminováním, jež lze reprezentovat elementárními eliminačními maticemi popsanými v odstavci 5.2.

Buď tedy $A = (a_{jk})$ daná regulární matice s $a_{11} \neq 0$ a

$$E_1 = I - v_1 \otimes e_1$$

elementární dolní trojúhelníková matice indexu 1, kde

$$v_1 = \frac{1}{a_{11}} \begin{pmatrix} 0 \\ a_{12} \\ . \\ . \\ . \\ a_{N1} \end{pmatrix}.$$

Jest tedy

$$A^{(0)} = A$$

a

$$A^{(1)} = \begin{pmatrix} a_{11} & a_{12} & . & . & . & a_{1N} \\ 0 & a_{22} - \frac{a_{12}}{a_{11}} a_{21} & . & . & . & a_{2N} - \frac{a_{1N}}{a_{11}} a_{21} \\ . & . & . & . & . & . \\ 0 & a_{N2} - \frac{a_{12}}{a_{11}} a_{N1} & . & . & . & a_{NN} - \frac{a_{1N}}{a_{11}} a_{N1} \end{pmatrix}.$$

Zvláštní pozornosti si zasluhuje skutečnost, že první řádek transformované matice A^1 je totožný s prvním řádkem původní matice A.

Obecně, buď $1 < k \leq N-1$, $a_{kk}^{k-1} \neq 0$, kde

$$A^{(k-1)} = \begin{pmatrix} a_{11}^{(k-1)} & . & . & . & . & . & . & a_{1N}^{(k-1)} \\ & . & & & & & & \\ & & . & & & & & \\ & & & . & & & & \\ 0 & . & . & . & a_{kk}^{(k-1)} & . & . & a_{(kN)} \\ & . & & & & . & & \\ & & . & & & & . & \\ & & & . & & & & \\ 0 & . & . & . & a_{Nk}^{k-1} & . & . & a_{NN}^{(k-1)} \end{pmatrix},$$

při čemž řádky s indexy $1 \leq j \leq k$, jsou totožné pro všechny matice A^l, $l = j, ..., k-1$.

Nechť

$$E_k = I_N - v_k \otimes e_k,$$

kde

$$v_k = \begin{pmatrix} a_{k+1,k}^{(k-1)} / a_{kk}^{(k-1)} \\ . \\ . \\ . \\ a_{Nk}^{(k-1)} / a_{kk}^{(k-1)} \end{pmatrix}.$$

Potom

$$A^{(k)} = E_k A^{(k-1)} = \begin{pmatrix} a_{11}^{(k)} & . & . & . & a_{1k}^{(k)} & a_{1k+1}^{(k)} & . & . & . & a_{1N}^{(k)} \\ & . & & & & & & & & \\ & & . & & & & & & & \\ & & & . & & & & & & \\ 0 & & & & a_{kk}^{(k)} & a_{kk+1}^{(k)} & . & . & . & a_{kN}^{(k)} \\ & . & & & & & & & & \\ & & . & & & & & & & \\ & & & . & & & & & & \\ 0 & . & . & . & 0 & a_{Nk+1}^{(k)} & . & . & . & a_{NN}^{(k)} \end{pmatrix},$$

při čemž řádky s indexy $1 \leq j \leq k+1$ jso totožné pro matice $A^{(l)}$, kde $l = j, ..., k$.

Právě provedený indukční krok ukazuje, že ve "fyziologickém" případě, t. j. když

$$a_{kk}^{(k-1)} \neq 0,$$

výše uvedený proces vede po konečném počtu kroků, $k \leq N-1$, k situaci, kdy $A^{(N-1)}$ je horní trojúhelníková matice.

Hledané řešení původní soustavy je evidentně totožné s řešením soustavy

$$A^{(N-1)}x = b^{(N-1)},$$

kde

$$b^{(N-1)} = E_{N-1}...E_1 b.$$

Schematicky lze tuto skutečnost postihnout tak, že sestrojíme obdélníkovou matici typu $N \times N+1$

$$B = (A, b)$$

a provádíme postupně transformace s touto maticí, tedy konstruujeme postupně matice

$$B^{(k)} = E_k B^{(k-1)},\ k = 1, ...,,$$

při čemž

$$B^{(0)} = B.$$

Po $N-1$ krocích obdržíme matici B^{N-1} mající tvar

$$B^{(N-1)} = \begin{pmatrix} \times & \times & . & . & . & \times & . & . & . & \times \\ 0 & \times & . & . & . & \times & . & . & . & \times \\ . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & . & . & . & \times . & . & . & . & \times \end{pmatrix}.$$

Je vcelku zřejmé, jak uvedený postup zobecnit na případ současného řešení soustav s danou maticí soustavy a s více pravými stranami $b_1, ..., b_p,\ p \geq 1$. V takovém případě

$$B_{(p)} = (A, b_1, ..., b_p)$$

s

$$B_{(p)}^{(k)} = E_k B_{(p)}^{(k-1)},$$

kde

$$B_{(p)}^{(0)} B_{(p)}.$$

Již víme, že řešení soustav

$$Ax = B_j,\ j = 1, ..., p,$$

se převádí na řešení soustav

$$A^{(N-1)}x = b_j^{(N-1)},\ j = 1, ...p,$$

kde b_j je $N+j$ - tý sloupec matice $B_{(p)}^{(N-1)}$. Matice těchto soustav jsou však horní trojúhelníkové a nalezení hledaných řešení je záležitostí rutinní. To vše za předpokladu, že

$$a_{kk}^{(k-1)} \neq 0.$$

Tento předpoklad je rozhodující. Numerická analýza spojená s tímto předpoladem si vynutí posléze úpravu vlastního algoritmu řešení soustav lineárních algebraických rovnic - t. zv. *pivotní strategie*.

Ještě než se budeme numerickými aspekty eliminačních metod zabývat, uveďme složitost výpočtu dle výše uvedeného schematu za předpokladu, že proces je algoritmicky proveditelný, t.j.

$$a_{kk}^{(k-1)} \neq 0,\ k = 1, ..., N-1.$$

Transformace matice $B_{(p)}$ $(= B_{(p)}^{(0)})$ do lichoběžníkového tvaru *stojí* asymptoticky

$$\sum_{j=1}^{N}(N - j + p)(N - 1) \sim \frac{1}{3}N^3 + \frac{1}{2}pN^2.$$

Je - li na daném počítači časová spotřeba na jednu aritmetickou operaci ν sekund, pak stanovení řešení dané soustavy s maticí typu $N \times N$ pro p pravých stran se realizuje v následujících časových mezích:

N	$\frac{1}{3}N^3$ (eliminace)	$\frac{1}{2}N^2$ (zpětná substituce)
10	$3,3.10^{-4}\nu$ sek	$5.10^{-5}\nu$ sek
10^2	$3,3.10^{-1}\nu$ sek	$5.10^{-2}\nu$ sek
10^3	$3,3.10^{5}\nu$ sek	$5.10^{-1}\nu$ sek
10^6	$3,3.10^{12}\nu$ sek	$5.10^{5}\nu$ sek

Poznámka 5.1 *Jistě stojí za povšimnutí, že spotřeba času uvedená v posledním řádku činí řádově $\nu.10^9$ hodin! To ukazuje, že pro velké dimenze jsou metody eliminační na soudobých seriálních počítačích prakticky nepoužitelné.*

Uvedená poznámka osvětluje nejrůznější snahy obejít komplikace spojené s časovou náročností eliminačních metod. Tendence jsou zřejmé - využít buď speciální struktury počítače (víceprocesorové) a nebo speciální struktury dané třídy matic řešených soustav (na př. pozitivně definitní a pásové).

Pivotní strategie

Snadno nahlédneme, že předpoklad nenulovosti diagonálních prvků $a_{kk}^{(k-1)}$, $k = 1, ..., N-2$, je podstatný a že algoritmus eliminace se zhroutí při jeho nesplnění. Nepříjemné je též, že totéž lze říci o numerické realizaci, t. j. eliminační algoritmy se zhroutí, pakliže uvedené prvky $a_{kk}^{(k-1)}$ nejsou dostatečně velké v absolutní hodnotě.

Demonstrujme takovou situaci na příkladě následující soustavy.

Nechť tedy

$$\left\{\begin{array}{rcrcrcl} x_1 & + & x_2 & + & x_3 & = & 1 \\ x_1 & + & x_2 & + & 2x_3 & = & 2 \\ x_1 & + & 2x_2 & + & 2x_3 & = & 1 \end{array}\right. \tag{5.29}$$

Matice této soustavy je regulární a existuje tudíž jediné řešení

$$x_1 = -x_2 = x_3 = 1.$$

Avšak po provedení prvního kroku eliminace obdržíme

$$\left\{\begin{array}{rcrcl} & & x_3 & = & 1 \\ x_2 & + & x_3 & = & 0, \end{array}\right. \tag{5.30}$$

a dále eliminovat nelze

$$a_{22}^{(1)} = 0. \tag{5.31}$$

V obecném případě kdy $detA \neq 0$ a přitom

$$a_{kk}^{k-1} = 0,$$

musí existovat pro nějaké $j > k$ prvek

$$a_{kj}^{(k-1)} \neq 0,$$

jinak by k - tý řádek byl nulový, což vylučuje regularitu matice soustavy. Vhodnou záměnou řádků či sloupců soustavy lze posléze docílit toho, aby po takové úpravě diagonáln í prvek byl roven vhodně vybranému prvku k - tého řádku.

Tato úvaha má oprávnění i v situaci, kdy posuzujeme nenulovost z pohledu daného počítače. Jinými slovy, permutaci proměnných či řádků je nutné provádět kdykoliv absolutní hodnota prvku $a_{kk}^{(k-1)}$ není dostatečně velká.

Dokumentujme takový případ soustavou

$$\left\{\begin{array}{rcrcrcl} x_1 & + & x_2 & + & x_3 & = & 1 \\ & & 0,0001x_2 & + & x_3 & = & 1 \\ & & & & 9,999x_3 & = & 10 \end{array}\right. \tag{5.32}$$

Matice této soustavy je regulární horní trojúhelníková. Provedeme - li zpětnou substituci užívajíce tří platných cifer v pohyblivé desetinné čárce, zjistíme, že

$$\tilde{x}_1 = 0,\ \tilde{x}_2 = 0,\ \tilde{x}_3 = 1$$

splňují požadované vztahy. Avšak přesné řešení (vzaté se třemi polatnými ciframi) je dáno výrazy

$$x_1 = 1,\ -x_2 = x_3 = 1.$$

Jak vysvětlit tento rozpor tedy numerickou nestabilitu? Odpověď je smutná, je to v podstatě věci: *Gaussova eliminační metoda je numericky nestabilní.*

V souvislosti s předchozím příkladem si všímněme toho, že současnou záměnou 2. a 3. řádku a 2. a 3. sloupce, t. j.

$$y_1 = x_1,\ y_2 = x_3\ y_3 = x_2,$$

se obdrží soustava

$$\begin{array}{rcrcrcl} y_1 & + & y_2 & + & y_3 & = & 1 \\ & & 9,999y_2 & & & = & 10 \\ & & y_2 & + & 0,0001y_3 & = & 1. \end{array}$$

Zde již

$$y_1 = y_2 = -y_3 = 1,$$

takže po zpětné substituci $y \to x$ obdržíme přesné řešení v rámci tří platných cifer.

Částečná pivotace

Hledejme přirozené číslo t tak, aby

$$t = \min\{j : k \le j \le N\}$$

a přitom

$$\left|a_{tk}^{(k-1)}\right| = \max\left\{|a_{jk}^{(k-1)}| :\ j = 1, ..., N\right\}.$$

Úplná pivotace

Hledejme přirozená čísla t a s taková, aby

$$\begin{array}{c} t = \min\left\{j :\ k \le j \le N\right\} \\ s = \min\left\{j :\ k \le l \le N\right\} \end{array}$$

a přitom

$$\left|a_{ts}^{(k-1)}\right| = \max\left\{\left|a_{jl}^{(k-1)}\right| :\ k \le j \le N,\ k \le l \le N\right\}.$$

Důležité doporučení plynoucí z praktických zkušeností praví, že částečná pivotáž je postačující; úplná pivotáž jednak výpočet zpomaluje, jednak numerickou stabilitu podstatně nezvyšuje.

Formální výraz předchozích úvah je obsahem následujících vět.

Věta 5.9 *(o úplné pivotaci) Buď A matice typu $M \times N$ a nechť $A \neq 0$.*

Potom existují celá kladná čísla $t \geq 1$ a $s \leq min\ (M-1,\ N)$, elementární permutační matice $P_j,\ Q_l,\ j = 1, ..., M-1,\ l = 1, ...N$ a elementární dolní trojúhelníkové matice E_j indexu $N-j,\ j = 1, ..., s$, tak, že platí vyjádření

$$\hat{A} = E_s P_s ... E_1 P_1\ A\ Q_1 ... Q_t$$

$$= \begin{pmatrix} \hat{A}_{11} & \hat{A}_{12} \\ 0 & 0 \end{pmatrix},$$

při čemž $\hat{A}_{11}$ je horní trojúhelníková matice typu $r \times r$ s

$$\det \hat{A} \neq 0.$$

Dále pak r je hodnost matice A.

Věta 5.10 *(o částečné pivotaci) Buď A matice typu $M \times N$ a nechť $s = min\ (M-1,\ N)$.*

Potom existují elementární permutační matice P_j a elementární dolní trojúhelníkové matice E_j indexu $M-j,\ j = 1, ..., s$ takové, že platí vyjádření

$$\tilde{A} = E_s\ P_s ... E_1\ P_1\ A,$$

při čemž $\tilde{A}$ je dolní lichoběžníková, t. j.

$$\tilde{a}_{jk} = 0,\ \ j > k.$$

Již víme, že při řešení soustav typu

$$Ax = b,$$

kde A je matice typu $N \times N$, Gaussovou eliminační metodou je nutné provádět pivotaci.

Tedy, buď podle schématu úplné pivotace

$$A = E_{N-1}\ P_{N-1} ... E_1\ P_1\ A\ Q_1\ Q_{N-1}$$

nebo užitím pivotace částečné

$$A = E_{N-1}\ P_{N-1} ... E_1\ P_1\ A.$$

Obě tyto strategie lze bez obtíží přenést na případ simultanního řešení p soustav typu

$$Ax = b_j,\ j = 1, ..., p.$$

V takovém případě se potřebné transformace provádějí s maticí

$$B_p = (A,\ b_1, ..., b_p)\,.$$

Zde je zřejmě podstatný předpoklad o nezávislosti vektoru b_j na řešení soustav s indexy $l < j$. Nejsou však řídké případy, kdy

$$b_{j+1} = b_{j+1}(x_1, ..., x_p)$$

a v tom případě naznačený postup selže.

Věta 5.11 *Buď A matice typu $N \times N$ a A_k nechť označují matici typu $k \times k$ tvořenou prvky ležícími v průniku řádků a sloupců s indexy $1, ..., k$.*

Je - li

$$\det A_k \neq 0,\ \forall k = 1, ..., N-1, \tag{5.33}$$

pak existuje právě jedna dolní trojúhelníková matice $L = (l_{jk})$ typu $N \times N$ taková, že

$$l_{jj} = 1,\ j = 1, ..., N,$$

a právě jedna horní trojúhelníková matice $U = (u_{jk})$ typu$N \times N$ taková, že

$$A = L\,U.$$

Důkaz. Nechť $N = 1$. Potom

$$l_{11} = 1,\ u_{11} = a_{11}$$

a tvrzení platí. Nechť tvrzení platí pro všechny matice řádu $l \leq k-1$. Vyšetřujme

$$A_k = \begin{pmatrix} A_{k-1} & a \\ c^T & a_{kk} \end{pmatrix},$$

za předpokladu, že

$$A_{k-1} = L_{k-1}\,U_{k-1}.$$

Položme

$$L_k = \begin{pmatrix} L_{k-1} & 0 \\ l^T & 1 \end{pmatrix},$$

a

$$U_k = \begin{pmatrix} U_{k-1} & u \\ 0^T & u_{kk} \end{pmatrix},$$

kde a a c jsou známé vektory a l je vektor typu $(k-1)$, jenž je zapotřebí určit podobně jako vektor u a číslo u_{kk}.

Abychom určili součin $L_k\,U_k$, vycházíme z relací

$$L_{k-1}\,U_{k-1} = A_{k-1}, \tag{5.34}$$

$$L_{k-1}u = a, \tag{5.35}$$

(5.36) $$l^T U_{k-1} = c^T$$

a

(5.37) $$l^T u + u_{kk} = a_{kk}.$$

Dle indukčního předpoklagu jsou L_{k-1} a U_{k-1} určeny jednoznačně a navíc

$$\begin{aligned} \det U_{k-1} &= \det L_{k-1} \det U_{k-1} \\ &= \det A_{k-1} \neq 0, \end{aligned}$$

takže (5.35) je jednoznačně řešitelná vzhledem k u. Podobně je tomu s vektorem l v (5.36). Díky jednoznačnosti l a u je posléze jednoznačně určen i prvek u_{kk} pomocí (5.37). |||

Není - li předpoklad (5.33) splněn, věta 5.11 neplatí; na př.

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

Předpokládejme, že naopak,

$$A = L\,U,$$

kde

$$L = \begin{pmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{pmatrix},\ U = \begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix},$$

takže

$$L\,U = \begin{pmatrix} l_{11}u_{11} & l_{11}u_{12} \\ l_{21}u_{11} & l_{21}u_{12} + l_{22}u_{22} \end{pmatrix}.$$

odkud plyne, že buď

$$l_{11} = 0$$

nebo

$$u_{11} = 0,$$

ale potom buď

$$L = \begin{pmatrix} 0 & 0 \\ l_{21} & l_{22} \end{pmatrix}$$

a nebo

$$U = \begin{pmatrix} 0 & u_{12} \\ 0 & 0 \end{pmatrix},$$

tedy spor.

Ukážeme nyní, že existence LU rozkladu pro A je zaručena vždy, pakliže lze Gaussův algoritmus realizovat bez pivotace.

Předpokládejme tedy, že A je taková, že Gaussova eliminace projde bez pivotace, to značí, že

$$A^{(N-1)} = E_{N-1} \ldots E_1\, A$$

a tudíž

$$A = E_1^{-1}...E_{N-1}^{-1}\ A^{(N-1)},$$

takže

$$A = L\,U,$$

kde

$$L = E_1^{-1}...E_{N-1}^{-1}$$

a

$$U = A^{(N-1)} = \begin{pmatrix} a_{11}^{(N-1)} & . & . & . & a_{1N}^{(N-1)} \\ & . & & & \\ & & . & & \\ & & & . & \\ 0 & & . & . & a_{NN}^{N-1} \end{pmatrix}.$$

Protože zřejmě $l_{jj} = 1$,$j = 1, ..., N$, výše uvedené tvrzení platí.

Obecně však se Gaussova eliminace, byť pro symetrické matice bez permutace řádků či sloupců provádět nedá. To ukazuje příklad

$$\begin{pmatrix} 0 & 1 \\ 1 & \epsilon \end{pmatrix}.$$

Pro pozitivně definitní symetrické matice však lze Gaussovu eliminaci provádět bez pivotace.

Pozitivně definitní matice jsou totiž charakterizovány skutečností, že platí

Věta 5.12 *(*Sylvesterovo kriterium*) Matice $A = (a_{jk})$ typu $N \times N$ je pozitivně definitní právě když*

(5.38) $$0 < \det A_k,\ k = 1, ...N,$$

kde A_k jsou tvořeny prvky ležícími v průseku prvních k řádků a sloupců.

Důkaz. Nutnost je triviální. Je-li A pozitivně definitní, potom, podle předpokladu platí nerovnosti

$$\exists \kappa > 0 :\ \kappa(x,\ x)_N \leq (Ax,\ x)_N,$$

takže volbou

$$x = \begin{pmatrix} \tilde{x} \\ 0 \end{pmatrix},\ \tilde{x} \in \mathcal{R}^k,$$

obdržíme, že platí

$$(A_k\tilde{x},\ \tilde{x})_k \geq \kappa(\tilde{x},\ \tilde{x})_k.$$

K důkazu postačitelnosti nechť naopak platí (5.38). Předpokládejme, že existuje vektor $y \in \mathcal{R}^N$ takový, že

(5.39) $$(Ay,\ y)_N \leq 0,$$

Je tedy zaručena existence vlastní hodnoty $\lambda \leq 0$ a tudíž,

$$0 = \det(\lambda I - A) = \sum_{k=0}^{N} a_{N-k}\lambda^k, \tag{5.40}$$

kde

$$\text{sign}\,(a_{N-k}) = (-1)^k,\ k = 1, ..., N.$$

Vidíme tedy, že rovnost (5.40) je s $\lambda \leq 0$ vyloučena.

Tím je důkaz věty 5.12 proveden. |||

Pozitivně definitní matice

Buď $A = (a_{jk})$ symetrická pozitivně definitní matice typu $N \times N$.

Ukážeme si, že provedení Gaussovy eliminační transformace zachovává symetrii těch hlavních submatic, které ještě nejsou horní trojúhelníkové. Tedy, je - li

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & . & . & . & a_{1k}^{(k)} & a_{1k+1}^{(k)} & . & . & . & a_{1N}^{(k)} \\ & . & & & & & & & & \\ & & . & & & & & & & \\ & & & . & & & & & & \\ 0 & . & . & . & a_{kk}^{k(}) & a_{kk+1}^{(k)} & . & . & . & a_{kN}^{(k)} \\ 0 & . & . & . & 0 & a_{k+1k+1}^{(k)} & . & . & . & a_{k+1N}^{(k)} \\ . & . & . & . & . & . & . & . & . & . \\ 0 & . & . & . & 0 & a_{Nk+1}^{(k)} & . & . & . & a_{NN}^{(k)} \end{pmatrix},$$

potom tvrdíme, že

$$a_{jl}^{(k)} = a_{lj}^{(k)},\ j, l = k+1, ..., N. \tag{5.41}$$

Nechť tedy

$$a_{jl}^{(k-1)} = a_{lj}^{(k-1)},\ j, l = k, ..., N. \tag{5.42}$$

Protože

$$a_{jl}^{(k)} = a_{jl}^{(k-1)} - \frac{a_{kl}^{(k-1)}}{a_{kk}^{(k-1)}} a_{jk}^{(k-1)},$$

plyne z (5.42), že

$$a_{jl}^{k} = a_{lj}^{(k-1)} - \frac{a_{lk}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)} = a_{lj}^{k}.$$

Platí tedy (5.41). Indukce dává obecný výsledek.

Poznámka 5.2 *Tato skutečnost zřejmě snižuje počet operací nutných k provedení eliminační transformace zhruba na polovinu.*

Ze Sylvesterova kriteria (věta 5.12) plyne, že

$$a_{kk} > 0,\ k = 1, ..., N$$

a dále

Věta 5.13 *Je - li A symetrická pozitivně definitní matice, potom*

$$|a_{jk}|^2 \leq a_{jj}\, a_{kk},\ j,k = 1,...,N. \tag{5.43}$$

Tudíž maximální prvek matice A leží na její diagonále.

Důkaz. Buď P permutační matice. Potom

$$(PAP^T)^T = PA^TP^T = PAP^T$$

je symetrická.

Pro pevnou dvojici indexů j a k existuje elementární permutační matice P tak, že PAP^T je symetrická a pozitivně definitní matice. Na základě Sylvesterova kriteria potom platí relace

$$0 < \det \begin{pmatrix} a_{jj} & a_{jk} \\ a_{kj} & a_{kk} \end{pmatrix} = a_{jj}a_{kk} - a_{jk}^2$$

a tudíž (5.43).

Kdyby

$$a_{jk} = \mathrm{Max}\left\{a_{lt} :\ l,\ t = 1,...,N\right\},\ j \neq k,$$

pak by na jedné straně, dle předpokladu,

$$a_{jk} > a_{kk},\ a_{jk} > a_{jj},$$

tedy

$$a_{jk}^2 > A_{jj}a_{kk}$$

a současně

$$a_{jk} \leq \sqrt{a_{jj}}\ \sqrt{a_{kk}},$$

spor. □

Vidíme tedy, že Gaussovu eliminaci lze pro pozitivně definitní matice provádět vždy a to bez pivotace. Hrozí jen, že se při provádění transformací vytvoří dělením malými čísly prvky velké. Již víme, že ty však musí ležet na diagonále. Dále pak z vyjádření

$$a_{jj}^{(k)} = a_{jj}^{(k-1)} - \frac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}} a_{jk}^{(k-1)}$$

odvodíme, že

$$a_{jj}^{(k)} \leq a_{jj}^{(k-1)} + \frac{|a_{kj}^{(k-1)}|}{a_{kk}^{(k-1)}} \left|a_{jk}^{(k-1)}\right| \leq 2a_{jj}^{(k-1)}.$$

Vidíme, tedy, že růst prvků transformovaných matic je nezávislý na velikosti prvků transformovaných matic. To nás vede k závěru, že Gaussova eliminace bez pivotace je pro pozitivně definitní matice bezpečná.

Analog věty o LU-rozkladu (Věta 5.11, str. 51) lze pro pozitivně definitní matice formulovat takto:

Věta 5.14 *Buď A symetrická pozitivně definitní matice.*
Potom existuje právě jedna horní trojúhelníková matice R taková, že platí

$$A = R^T R.$$

Důkaz. Nechť

$$A = L\,U,$$

kde $l_{jj} = 1,\ j = 1, ..., N$. Tento rozklad existuje na základě věty 5.11 a věty 5.12.
Na základě tohoto rozkladu odvodíme, že

$$a_{11} = u_{11} > 0$$

a

$$u_{jj} = \frac{\det A_j}{\det A_{j-1}},\ j = 1, ..., N.$$

Položme

$$D = \mathrm{diag}\{u_{11}, ..., u_{NN}\}.$$

Potom

$$A = LDD^{-1}U = LD\tilde{U},\ \tilde{U} = D^{-1}U,$$

při čemž

$$\mathrm{diag}L = I = \mathrm{diag}\tilde{U}.$$

Ze symetrie A plyne, že

$$A^T = A = \tilde{U}^T D L^T,$$

neboli, na základě jednoznačnosti LU-rozkladu,

$$\tilde{U}^T = L,\ DL^T = U. \tag{5.44}$$

Položme dále

$$R = D^{-\frac{1}{2}}U,$$

kde

$$D^{-\frac{1}{2}} = \mathrm{diag}\left\{d_{11}^{-\frac{1}{2}}, ..., d_{NN}^{-\frac{1}{2}}\right\}.$$

Potom

$$R^T R = U^T D^{-\frac{1}{2}} D^{-\frac{1}{2}} U = \tilde{U}^T U$$

a dle (5.44) tedy

$$R^T R = L\,U = A.$$

□

Třídiagonální matice

Matice $A = (a_{jk})$ se nazývá *pásová*, jestliže jestliže existují indexy $p \geq 1$ a $q \geq 1$ takové, že

$$a_{jk} = 0$$

pro

$$j > k + p \text{ a } j - q < k.$$

Speciálně pro $p = q = 1$ se takové pásové matice nazývají *třídiagonáln í.*

Je zřejmé, že Gaussova eliminace pásovost zachovává.

Obecně platí i zde, že pivotace je nezbytná i v případě symetrických matic. Výjimkami jsou matice diagonálně dominantní a matice pozitivně definitní .

Vyšetřujme případ třídiagonálních matic. Matici

$$A = \begin{pmatrix} a_1 & c_1 & 0 & . & . & . & 0 & 0 & 0 \\ b_2 & a_2 & c_2 & . & . & . & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & b_{N-1} & a_{N-1} & c_{N-1} \\ 0 & 0 & 0 & . & . & . & 0 & b_N & a_N \end{pmatrix}$$

se snažme vyjádřiti ve tvaru

$$A = \begin{pmatrix} 1 & 0 & 0 & . & . & . & 0 & 0 & 0 \\ \beta_2 & 1 & 0 & . & . & . & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & \beta_{N-1} & 1 & 0 \\ 0 & 0 & 0 & . & . & . & 0 & \beta_N & 1 \end{pmatrix} \times$$

$$\begin{pmatrix} \alpha_1 & \gamma_1 & 0 & . & . & . & 0 & 0 & 0 \\ 0 & \alpha_2 & \gamma_2 & . & . & . & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & 0 & \alpha_{N-1} & \gamma_{N-1} \\ 0 & 0 & 0 & . & . & . & 0 & 0 & a\alpha_N \end{pmatrix}$$

Odtud, za předpokladu, že

$$a_k \neq 0, \quad \forall k = 1, ..., N,$$

$$b_k \neq 0, \quad \forall k = 1, ..., N-1,$$

$$c_k \neq 0, \quad \forall k = 2, ..., N,$$

vyplývají vztahy

$$\gamma_k = c_k \quad \forall k = 1, ..., N-1,$$

$$\alpha_1 = a_1$$

$$\beta_k = \frac{1}{a_{k-1}} \quad \forall k = 2, ..., N,$$

$$\alpha_k = a_k - \beta_k c_{k-1}, \quad \forall k = 2, ..., N.$$

a z nich posléze algoritmus řešení. Řešení se poté obdrží obvyklým postupem, dopřednou eliminací a zpětnou substitucí.

Cvičení 5.1 *Buď A třídiagonální pozitivně definitní matice. Dále buďte $b_j,\ j = 1, ...; p,\ p \geq 2$. Takto formulovanou úlohu lze řešit na př. těmito dvěma odlišnými postupy:*

1. *Napřed určíme inverzní matici A^{-1} a počítáme postupně $A^{-1}b_j,\ j = 1, ..., p$.*
2. *Řešíme postupně soustavy*

$$Ax = b_j,\ j = 1, ..., p.$$

Rozhodněte který z těchto postupů je výhodnější z hlediska složitosti algoritmů.

5.4 Metody řešení soustav lineárních algebraických rovnic založené na principu optimalizace

Metody uvedené v záhlaví tohoto odstavce jsou typické pro soustavy s pozitivně definitní maticí.

Buď tedy H pozitivně definitní matice typu $N \times N$. Vyšetřujme soustavu

$$H\mathbf{x} = \mathbf{b},\ \mathbf{b} \in \mathcal{R}^N. \tag{5.45}$$

Vyšetřujme funkcionál $f : \mathcal{R}^N \to \mathcal{R}^1$

$$f(\mathbf{x}) = (H\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x}),\ \mathbf{x} \in \mathcal{R}^N \tag{5.46}$$

a zkoumejme jeho vlastnosti.

Snadno se přesvědčíme, že

$$f(\mathbf{x}) = (H(\mathbf{x} - \hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}}) - (\mathbf{b}, \hat{\mathbf{x}}),\ \mathbf{x} \in \mathcal{R}^N, \tag{5.47}$$

kde

$$\hat{\mathbf{x}} = H^{-1}\mathbf{b}, \tag{5.48}$$

takže díky nezápornosti výrazu $(H\mathbf{y}, \mathbf{y}),\ \mathbf{y} \in \mathcal{R}^N$, existuje

$$\min\left\{f(\mathbf{x}) : \mathbf{x} \in \mathcal{R}^N\right\} = f(\hat{\mathbf{x}}) = -(\mathbf{b}, \hat{\mathbf{x}}), \tag{5.49}$$

kde $\hat{\mathbf{x}}$ splňuje (5.48).

Snadno nahlédneme, že úloha nalézt řešení soustavy (5.45) s pozitivně definitní maticí je ekvivalentní s úlohou nalézt (lokální) minimum funkcionálu (5.46). S podobnou situací jsme se již setkali při vyšetřování zobecněných řešení v v odstaci 5.1.

Zabývejme se tedy otázkou sestrojování extrémů funkcionálů na $\mathcal{R}^N$.

Definice 5.2 *Funkcionál $f : \mathcal{R}^N \to \mathcal{R}^1$ se nazývá* kvadratický*, jestliže je polynomem druhého stupně jakožto funkce N proměnných $x_1, ..., x_N$.*

Snadno zjistíme, že f je kvadratický funkcionál na $\mathcal{R}^N$ právě když existuje symetrická matice H typu $N \times N$, vektor $\mathbf{x} \in \mathcal{R}^N$ a číslo $c \in \mathcal{R}^1$, takové, že platí

$$f(\mathbf{x}) = \frac{1}{2}(H\mathbf{x}, \mathbf{x}) + (\mathbf{b}, \mathbf{x}) + \mathbf{c},\ \mathbf{x} \in \mathcal{R}^{\mathbf{N}}. \tag{5.50}$$

Poznámka 5.3 *Z definice (5.46) je patrné, že tímto vztahem definovaný funkcionál je kvadratický.*

Symbolem $\mathcal{C}^{(m)}(\mathcal{S})$, $m \geq 0$, označujeme množinu funkcionálů, které mají spojité parciální derivace až do řádu m na množině $\mathcal{S}$), kde $\mathcal{S} \supset \{\mathbf{x} \in \mathcal{R}^N : \|\mathbf{x} - \mathbf{x}_0\| < \delta\}$, pro nějaké $\delta > 0$ a $\mathbf{x}_0 \in \mathcal{R}^N$.

Definice 5.3 *Nechť $f \in \mathcal{C}^{(1)}(\mathcal{S})$. Vektor*

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial f}{\partial x_N}(\mathbf{x}) \end{pmatrix}$$

se nazývá gradientem funkcionálu f.

Jestliže $f \in \mathcal{C}^{(2)}(\mathcal{S})$, matice

$$H(\mathbf{x}) = \left(\frac{\partial^2 f}{\partial x_j \partial x_k}(\mathbf{x}) \right)$$

se nazývá Hessiánem funkcionálu f.

Je-li $f \in \mathcal{C}^{(1)}(\mathcal{S})$, pak

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + (\mathbf{h}, \mathbf{g}(\mathbf{x})) + o(\|\mathbf{h}\|), \tag{5.51}$$

zatímco pro $f \in \mathcal{C}^{(2)}(\mathcal{S})$

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + (\mathbf{h}, \mathbf{g}(\mathbf{x})) + \frac{1}{2}(H(\mathbf{x})\mathbf{h}, \mathbf{h}) + o(\|\mathbf{h}\|^2), \tag{5.52}$$

kde

$$o(\alpha) = F(\mathbf{x}; \alpha),\ F : \mathcal{R}^N \times \mathcal{R}^1_+ \to \mathcal{R}^1,$$

při čemž

$$\lim_{\alpha \to 0} \frac{|F(\mathbf{x}; \alpha)|}{\alpha} = 0.$$

Poznámka 5.4 *Všimněme si toho, že na základě (5.52) každý nelineární funkcionál, mající gradient a Hessián, se lokálně chová jako funkcionál kvadratický.*

Definice 5.4 *Prvek $\tilde{\mathbf{x}} \in \mathcal{S}$ se nazývá* stacionárním prvkem *funkcionálu $f \in \mathcal{C}^{(1)}(\mathcal{S})$, jestliže platí*

$$\mathbf{g}(\tilde{\mathbf{x}}) = 0,$$

kde $\mathbf{g}(\mathbf{x})$ značí gradient f v $\tilde{\mathbf{x}}$.

Věta 5.15 *Nechť $f \in \mathcal{C}^{(1)}(\mathcal{S})$. Je-li $\hat{\mathbf{x}}$ prvkem lokálního minima funkcionálu f, pak f je stacionární v prvku $\mathbf{x}$.*

Důkaz. V (5.51) položme $\mathbf{x} = \hat{\mathbf{x}}$ a $\mathbf{h} = -hg(\hat{\mathbf{x}})$, kde $h \in \mathcal{R}^1$ je reálná proměnná veličina z nějakého intervalu $[0, h_0]$.

Platí tedy

$$f(\hat{\mathbf{x}} + \mathbf{h}) = f(\hat{\mathbf{x}}) - h\|\mathbf{g}(\hat{\mathbf{x}})\|^2 + o(h).$$

Nechť f není stacionární v $\hat{\mathbf{x}}$. Potom $g(\hat{\mathbf{x}}) \neq 0$ a pro dostatečně malá h zjistíme, že platí

$$-h\|\mathbf{g}(\hat{\mathbf{x}})\|^2 + o(h) < 0,$$

neboli

$$f(\hat{\mathbf{x}} + \mathbf{h}) < f(\hat{\mathbf{x}})$$

a tudíž $\hat{\mathbf{x}}$ nemůže být prvkem lokálního minima. □

Stacionarita funkcionálu v $\hat{\mathbf{x}}$ není však postačující podmínkou pro extrém v $\hat{\mathbf{x}}$.

Věta 5.16 *Nechť $f \in \mathcal{C}^{(2)}(\mathcal{S})$ a nechť f je stacionární v $\hat{\mathbf{x}} \in \mathcal{S}$. Je-li $H(\hat{\mathbf{x}})$ pozitivně definitní pak je $\hat{\mathbf{x}}$ prvkem ostrého lokálního minima funkcionálu f.*

Důkaz. Z (5.52) a z podmínky $\mathbf{g}(\hat{\mathbf{x}}) = 0$ odvodíme, že

$$f(\hat{\mathbf{x}} + \mathbf{h}) = f(\hat{\mathbf{x}}) + \frac{1}{2}\left(H(\hat{\mathbf{x}})\mathbf{h}, \mathbf{h}\right) + o(\|\mathbf{h}\|^2).$$

Pozitivní definitnost $H(\hat{\mathbf{x}})$ poskytuje existenci $\gamma > 0$ takového, že

$$\left(H(\hat{\mathbf{x}})\mathbf{h}, \mathbf{h}\right) \geq \gamma(\mathbf{h}, \mathbf{h}).$$

Tudíž

$$f(\hat{\mathbf{x}} + \mathbf{h}) - f(\hat{\mathbf{x}}) \geq \frac{1}{2}\gamma(\mathbf{h}, \mathbf{h}) + o(\|\mathbf{h}\|^2) \tag{5.53}$$

a protože výraz v pravé části vztahu (5.53) je kladný pro dostatečně malá $\|\mathbf{h}\|^2$, $\hat{\mathbf{x}}$ je prvkem ostrého minima funkcionálu f. □

Situace je podobná vyšetřujeme-li f z hlediska lokálních maxim. V tom případě totiž $-f$ má v odpovídajících prvcích lokální minima, takže je-li $\hat{\mathbf{x}}$ prvkem lokálního maxima, pak nutně f je stacionární v $\hat{\mathbf{x}}$ a je-li Hessián $H(\hat{\mathbf{x}})$ negativně definitní , $\hat{\mathbf{x}}$ je potom prvkem ostrého lokálního maxima.

Pro indefinitní Hessián nelze žádné závěry stran extrémů obecně zaručit. Není-li $\hat{\mathbf{x}}$ prvkem lokálního extrému, pak Hessián nemůže být definitní , tedy $H(\hat{\mathbf{x}})$ má ve svém spektru jak kladné tak záporné vlastní hodnoty. Dokonce, je-li Hessián $H(\hat{\mathbf{x}})$ semidefinitní, f může ale nemusí mít lokální extrém v $\hat{\mathbf{x}}$.

Dále si připomeňme klasickou definici z elemntární matematické analýzy.

Definice 5.5 *Nechť $f \in \mathcal{C}^{(2)}(\mathcal{S})$, $\mathbf{x} \in \mathcal{S} \subset \mathcal{R}^N$ s $\mathbf{y} \in \mathcal{R}^N$, $\|\mathbf{y}\| = 1$.*

Výraz

$$f^{(m)}(\mathbf{x};\mathbf{y}) = \left.\frac{d^m f(\mathbf{x}+\tau\mathbf{y})}{d\tau^m}\right|_{\tau=0} \tag{5.54}$$

se nazývá směrovou derivací funkcionálu f řádu m v prvku $\mathbf{x}$ ve směru $\mathbf{y}$.

Výpočet směrových derivací se provádí použitím obvyklých pravidel derivování.
Tak pro $f \in \mathcal{C}^{(1)}(\mathcal{S})$, $\mathbf{x} \in \mathcal{R}^N$ a $\mathbf{y} \in \mathcal{R}^N$ s $\|\mathbf{y}\| = 1$,

$$f^{(1)}(\mathbf{x};\mathbf{y}) = \sum_{j=1}^{N} \frac{\partial f}{\partial x_j}(x_1, ..., x_N) y_j = (g(\mathbf{x}), \mathbf{y})\,. \tag{5.55}$$

Dále pak,

$$f^{(2)}(\mathbf{x};\mathbf{y}) = (H(\mathbf{x})\mathbf{y}, \mathbf{y})\,. \tag{5.56}$$

Ze Schwarzovy nerovnosti odvodíme pro $f \in \mathcal{C}^{(1)}(\mathcal{S})$, že

$$\max\left\{\left|f^{(1)}(\mathbf{x};\mathbf{y})\right| : \mathbf{y} \in \mathcal{R}^N,\ \|\mathbf{y}\| = 1\right\} = \left|f^{(1)}(\mathbf{x};\hat{\mathbf{y}})\right| = (g(\mathbf{x}), \hat{\mathbf{y}})\,,$$

kde

$$\hat{\mathbf{y}} = \frac{g(\mathbf{x})}{\|g(\mathbf{x})\|}.$$

Z (5.55) je patrné, že $f^{(1)}(\mathbf{x};\mathbf{y}) = 0$ pro všechny směry $\mathbf{y}$ nastane právě když $g(\mathbf{x}) = 0$, tedy, je-li $\mathbf{x}$ stacionárním prvkem funkcionálu f.

Z (5.56) odvodíme zase, že $f^{(2)}(\mathbf{x};\mathbf{y}) > 0$ pro všechny směry $\mathbf{y} \neq 0$ právě když $H(\mathbf{x})$ je pozitivně definitní .

Předchozí věty lze posléze přeformulovat takto.

Věta 5.17 *Nechť $f \in \mathcal{C}^{(1)}(\mathcal{S})$, a $\hat{\mathbf{x}} \in \mathcal{S}$ je prvkem lokálního minima pro f.*

Potom $f^{(1)}(\hat{\mathbf{x}};\mathbf{y}) = 0$ pro všechny směry $\mathbf{y} \in \mathcal{R}^N$.

Věta 5.18 *Nechť $f \in \mathcal{C}^{(2)}(\mathcal{S})$ a $\hat{\mathbf{x}} \in \mathcal{S}$ je takový, že $f^{(1)}(\hat{\mathbf{x}};\mathbf{y}) = 0$ pro všechny směry $\mathbf{y} \in \mathcal{R}^N$.*

Potom $\hat{\mathbf{x}}$ je prvkem ostrého lokálního minima pro f v $\mathcal{S}$, jestliže $f^{(2)}(\hat{\mathbf{x}};\mathbf{y}) > 0$ pro všechny směry $\mathbf{y} \in \mathcal{R}^N$ s $\|\mathbf{y}\| = 1$.

Poznámka 5.5 *Věty 5.17 a 5.18 umožňují snadná zobecnění na případy, kdy vyšetřované funkcionály operují na obecnějších prostorech než jsou aritmetické vektorové prostory.*

Definice 5.6 *Předpokládejme, že funkcionál f je definován na $\mathcal{S} \subset \mathcal{R}^N$ a nechť $\eta \in$ range $f = \{\alpha \in \mathcal{R}^1 : \exists \mathbf{x} \in \mathcal{S} \Rightarrow f(\mathbf{x}) = \alpha\}$.*

Množina

$$\mathcal{L}_\alpha = \{\mathbf{x} \in \mathcal{S} :\ f(\mathbf{x}) = \alpha\}$$

se nazývá úrovňovou plochou funkcionálu f pro hodnotu α.

Situace s kvadratickými funkcionály je natolik jednoduchá, že lze používat názoru, který napovídá , že v okolí ostrého lokálního minima $\hat{\mathbf{x}}$ s $f(\hat{\mathbf{x}}) = \hat{\alpha}$ se úrovňové plochy $\mathcal{L}_\alpha$ s α blízkými k k α chovají jako nadelipsoidy (t. j. obecně jako elipsoidy ve vícerozměrných prostorech).

Obecně řečeno, numerické metody hledání ostrých lokálních extrémů se "chovají lépe", když úrovňové plochy v okolí extrémů jsou blízké plochám kulovým a "hůře, jsou-li tyto plochy výrazně deformovány od tvaru plochy kulové. Veličina, která výstihuje míru odchylky úrovňových ploch od tvaru sféry, je dána výrazem

$$\delta_\alpha = \inf\left\{\frac{\sup\{\|\mathbf{x}-\mathbf{y}\| : \mathbf{x} \in \mathcal{L}_\alpha\}}{\inf\{\|\mathbf{x}-\mathbf{y}\| : \mathbf{x} \in \mathcal{L}_\alpha\}} : \mathbf{y} \in \mathcal{S}\right\}, \tag{5.57}$$

při čemž

$$\mathcal{S}_\alpha = \mathrm{Int}\mathcal{L}_\alpha = \left\{\mathbf{x} \in \mathcal{L}_\alpha : \ \exists \rho > 0 \text{ tak, že } \forall \mathbf{h} \in \mathcal{R}^N \ \|\mathbf{h}\| < \rho \Rightarrow \mathbf{x}+\mathbf{h} \in \mathcal{L}_\alpha\right\}.$$

Zřejmě, $\delta_\alpha = 1, \ \alpha \to 0$, implikuje, že $\mathcal{L}_\alpha$ jsou kulové plochy.

Cvičení 5.2 Jednoduše určíme odchylku úrovňové plochy $\mathcal{L}_\alpha$ od plochy kulové pro případ kvadratického funkcionálu charakterizovaného pozitivně definitním Hessiánem H. Tehdy je totiž

$$\delta_\alpha = \frac{\lambda_{\max}}{\lambda_{\min}},$$

konstantní a přitom

$$\lambda_{\min} = \min\{\lambda \in \sigma(H)\}, \quad \lambda_{\max} = \max\{\lambda \in \sigma(H)\},$$

takže δ_α je rovno *spektrálnímu číslu podmíněnosti matice* H. V další části tohoto článku ještě uvidíme, jaký význam má tato skutečnost v analýze konvergence některých numerických metod. Dá se očekávat, že pro Hessiány H s velkými spektrálními čísly podmíněnosti budou odpovídající výpočtové procesy numericky nestabilní (odstrašující je případ, kdy $\mathcal{L}_\alpha$ je "doutník").

Iterační metody sestrojování prvků extrémů

Velké množství numerických metod hledání lokálního minima funkcionálů má iterační charakter a je zpravidla tvaru

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \tau_k \mathbf{d}^k, \tag{5.58}$$

v němž $\mathbf{d}^k$ je *zkusmý směr* zatímco τ_k je parametr zajišťující lokální minimalizaci či alespoň redukci velikosti funkcionálu f na dané množině.

Iterační proces (5.58) vyžaduje tedy strategii pro určení dvou typů objektů: výběr $\mathbf{d}^k$ a určení τ_k s ohledem na chování f podél přímky $\mathbf{x}^k + \tau\mathbf{d}^k, \ -\infty < \tau < +\infty$.

Definice 5.7 *Předpokládejme, že pro funkcionál f a vektory $\mathbf{x}$ a $\mathbf{d}$ existuje $\tau_0 > 0$ takové, že*

$$f(\mathbf{x}+\tau\mathbf{d}) < f(\mathbf{x}), \ 0 < \tau < \tau_0.$$

Potom nazýváme směr $\mathbf{d}$ směrem klesání *funkcionálu* f.

Věta 5.19 *Nechť $f \in \mathcal{C}^{(1)}(\mathcal{R}^N)$ a nechť* [$\mathbf{g}(\mathbf{x})$ *označuje (jako obvykle v tomto odstavci) gradient funkcionálu f v $\mathbf{x}$. Jestliže vektor $\mathbf{d}$ splňuje relaci*

$$(\mathbf{d}, \mathbf{g}(\mathbf{x})) < 0,$$

pak $\mathbf{d}$ je směrem klesání funkcionálu f v $\mathbf{x}$.

Důkaz. Z (5.51) obdržíme, že

$$f(\mathbf{x} + \tau\mathbf{d}) = f(\mathbf{x}) + \tau\,(\mathbf{d}, \mathbf{g}(\mathbf{x})) + o(\tau),$$

a protože podle předpokladu $(\mathbf{d}, g(\mathbf{x})) < 0$, musí platit nerovnost

$$\tau\,(\mathbf{d}, \mathbf{g}(\mathbf{x})) + o(\tau) < 0$$

pro dostatečně malá τ.

Poznámka 5.6 *Je-li $\mathbf{g}(\mathbf{x}) = 0$,, pak nelze stanovit, zda $\mathbf{d}$ je či není směrem klesání f v $\mathbf{x}$ bez další dodatečné informace.*

Věta 5.20 *Nechť $f \in \mathcal{C}^{(2)}(\mathcal{R}^N)$, nechť $(\mathbf{d}, g(\mathbf{x})) = 0$ a $(H(\mathbf{x})\mathbf{d}, g(\mathbf{x})) < 0$ pro nějaká $\mathbf{x}$ a $\mathbf{d}$ ($H(\mathbf{x})$ značí jako obvykle Heessián f).*

Potom $\mathbf{d}$ je směrem klesání f v $\mathbf{x}$.

Důkaz. Z (5.52) plyne

$$f(\mathbf{x} + \tau\mathbf{d}) = f(\mathbf{x}) + (1/2)\tau^2\,(H(\mathbf{x})\mathbf{d}, \mathbf{d})) + o(\tau^2), \tag{5.59}$$

odkud tvrzení plyne. □

Věta 5.21 *Nechť $f \in \mathcal{C}^{(1)}(\mathcal{R}^N)$. Potom mezi všemi zkusmými směry $\mathbf{d}$ funkcionálu f v $\mathbf{x}$ tím směrem, v němž f klesá v okolí $\mathbf{x}$ nejrychleji je $\mathbf{d} = -g(\mathbf{x})$.*

Důkaz. Naším úkolem je minimalizovat směrovou derivaci f v $\mathbf{x}$ vzhledem ke všem zkusmým směrům $\mathbf{d}$. Z (5.55) plyne, že to znamená minimalizovat výraz $(\mathbf{d}, g(\mathbf{x}))$ pro všechny směry $\mathbf{d}$, pro něž $\|\mathbf{d}\| = 1$. Protože $|(\mathbf{d}, g(\mathbf{x}))| \leq |g((\mathbf{x})|$, minimum se nabývá pro

$$\mathbf{d} = -\frac{\mathbf{g}(\mathbf{x})}{\|\mathbf{g}(\mathbf{x})\|}.$$

□

Dále se zabývejme problémem určení τ jsou-li zadány $\mathbf{d}$ a $\mathbf{x}$. Snažme se minimalizovat f podél přímky $\mathbf{y} = \mathbf{x} + \tau\mathbf{d}$, $-\infty < \tau < +\infty$. Pro obecné funkcionály stanovení parametru τ_k je poměrně složitá úloha a její řešení
vyžaduje další iterační proces. Pro případ kvadratického funkcionálu to však není žádný problém.

Nechť tedy

$$f(\mathbf{x}) = (1/2)(H\mathbf{x}, \mathbf{x}) + (\mathbf{b}, \mathbf{x}) + c.$$

Okamžitě lze ověřit, že

$$f(\mathbf{x}+\tau\mathbf{d}) - f(\mathbf{x}) = (1/2)\tau^2(H\mathbf{d},\mathbf{d}) + \tau(\mathbf{d},\mathbf{g}(\mathbf{x})).$$

Je-li H pozitivně definitní a $\mathbf{d} \neq 0$, potom $(H\mathbf{d},\mathbf{d}) > 0$ a pro

$$\tau = \frac{(\mathbf{d}, g(\mathbf{x}))}{(H\mathbf{d},\mathbf{d})}$$

$f(\mathbf{x}+\tau\mathbf{d})$ nabývá minima vzhledem k $\tau \in (-\infty, +\infty)$.

Tedy, nezávisle na volbě $\mathbf{d}^k$ v (5.58) je pro kvadratický funkcionál f

$$\tau = -\frac{(\mathbf{d}, g(\mathbf{x}))}{(H\mathbf{d},\mathbf{d})} \tag{5.60}$$

přii pustná hodnota lokálního optimalizačního parametru.

Metoda největšího spádu

Na základě výsledku věty 5.21 je přirozené klást $\mathbf{d}^k = -g(\mathbf{x}^k)$ v (5.58), takže

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \tau_k \mathbf{g}(\mathbf{x}^k). \tag{5.61}$$

Tento vztah spolu se strategií určování τ_k definuje *metodu největšího spádu* minimalizace obecného funkcionálu f.

Pro případ kvadratického funkcionálu ve tvaru (5.59) poskytuje (5.60) evidentní způsob určení τ_k. Položme

$$\mathbf{g}^k = \mathbf{g}(\mathbf{x}^k),\ \forall k = 0, 1, ...$$

Pro (5.59) je metoda největšího spádu určena formulemi

$$\mathbf{g}^k = H\mathbf{x}^k - b, \tag{5.62}$$

$$\tau_k = \frac{(\mathbf{g}^k, \mathbf{g}^k)}{(H\mathbf{g}^k, \mathbf{g}^k)}, \tag{5.63}$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \tau_k \mathbf{g}^k, \tag{5.64}$$

kde $k = 0, 1, ...$ a $\mathbf{g}^k = \mathbf{g}(\mathbf{x}^k)$.

Zřejmě (5.62) lze psát ve tvaru

$$\mathbf{g}^{k+1} = \mathbf{g}^k - \tau_k H\mathbf{g}^k, \tag{5.65}$$

a to umožňuje výpočet gradientů v (5.62) - (5.64) rekurentně.

Příslušný algoritmus se řídí formulemi

$$\tau_k = \frac{(\mathbf{g}^k, \mathbf{g}^k)}{(H\mathbf{g}^k, \mathbf{g}^k)}, \tag{5.66}$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \tau_k \mathbf{g}^k, \tag{5.67}$$

$$\mathbf{g}^{k+1} = \mathbf{g}^k - \tau_k H\mathbf{g}^k, \tag{5.68}$$

kde $k = 0, 1, ...$ a $\mathbf{g}^0 = H\mathbf{x}^0 - b$.

Poznamenejme, že algoritmus určený formulemi (5.66) - (5.68) lze implementovat za použití jediného násobení maticí H na jednu iteraci na rozdíl od standardsního algoritmu (5.62) - (5.64), kdy násobení maticí H jsou zapotřebí dvě.

Analýza konvergence metody největšího spádu dává podnět k obecnější úvaze. Vzhledem k existenci obecně nekonečně mnoha norem jest možno klást si otázku, kterou normu zvolit k odhadu chyby metody tak, abychom obdrželi výsledek co možno nejpřirozenější, nejrealističtější, nejvhodnější, nejjednodušší, nej- atd.

Na první pohled by se mohlo zdát, že výběr normy není nijak důležitý; vždyť všechny normy na $\mathcal{R}^N$ jsou ekvivalentní. Při bližším zkoumání zjistíme, že pro některé normy je taková analýza dost obtížná. Rozhodujícím kriteriem pro volbu normy bude tedy co nejmenší analytická obtížnost vyšetřování konvergence. Tu se ukazuje, jako velice výhodná a vskutku přirozená t.zv. energetická norma.

Odvoďme tedy větu o rychlosti konvergence metody největšího spádu v energetické normě.

Definice 5.8 *Buď H pozitivně definitní matice typu $N \times N$. Norma $\|.\|_H$ definovaná pomocí skalárního součinu*

$$\phi_H(x, y) = (Hx, y)\,,$$

kde $(x, y) = \sum_{k=1}^{N}(x)_k(y)_k$, se nazývá H-energetickou normou*; tedy, $\|x\|_H^2 = (Hx, x)$.*

Věta 5.22 *Nechť $\{x^k\}$ je posloupnost získaná pomocí metody největšího spádu pro kvadratický funkcionál f. Nechť $\hat{x}$ značí prvek globálního minima funkcionálu f.*

Potom platí odhad

$$\|x^k - \hat{x}\|_H \leq \frac{(\kappa(H) - 1)^k}{(\kappa(H) + 1)^k}\|x^0 - \hat{x}\|_H. \tag{5.69}$$

Dále nechť pro libovolné $\epsilon > 0$ $p(\epsilon)$ značí nejmenší index takový, že $\|x^k - \hat{x}\|_H \leq \epsilon\|x^0 - \hat{x}\|_H$, pro každé $x^0 \in \mathcal{R}^N$, potom platí nerovnost

$$p(\epsilon) \leq \frac{1}{2}\kappa(H)log\frac{1}{\epsilon}. \tag{5.70}$$

Důkaz. Nechť

$$E(\mathbf{x}^k - \hat{\mathbf{x}}) = (H(\mathbf{x} - \hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}}) = \left(H^{-1}\mathbf{g}(\mathbf{x}), \mathbf{g}(\mathbf{x})\right).$$

Dále si uvědomme, že platí

$$E(\mathbf{x} - \hat{\mathbf{x}}) = \left(H^{-1}\mathbf{g}^k, \mathbf{g}^k\right). \tag{5.71}$$

Platnost (5.71) plyne odtud, že

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\hat{\mathbf{x}}) = \frac{1}{2}\left(H(\mathbf{x} - \hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}}\right),$$

$$E(\mathbf{x}-\hat{\mathbf{x}}) = 2\{\mathbf{f}(\mathbf{x}) - \mathbf{f}(\hat{\mathbf{x}})\}$$

a

$$\mathbf{g}(\mathbf{x}) = H(\mathbf{x}-\hat{\mathbf{x}}).$$

Z vyjádření

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\hat{\mathbf{x}}) = \frac{1}{2}\,(H(\mathbf{x}-\hat{\mathbf{x}}), \mathbf{x}-\hat{\mathbf{x}})$$

totiž plyne, že

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} \frac{\partial \mathbf{f}}{\partial x_1}(\mathbf{x}) \\ . \\ . \\ . \\ \frac{\partial \mathbf{f}}{\partial x_N}(\mathbf{x}) \end{pmatrix} = \frac{1}{2} H(\mathbf{x}-\hat{\mathbf{x}})$$

a tudíž

$$\mathbf{x}-\hat{\mathbf{x}} = H^{-1}\mathbf{g}(\mathbf{x}).$$

Položíme-li tedy $\mathbf{x} = \mathbf{x}^{k+1}$ a použijeme-li rekurentní formule pro prvky $\mathbf{g}^k$, obdržíme, že

$$\begin{aligned} E(\mathbf{x}^{k+1}-\hat{\mathbf{x}}) &= \left(H^{-1}\mathbf{g}^{k+1}, \mathbf{g}^{k+1}\right) \\ &= \left(H^{-1}(\mathbf{g}^k - \tau_k H\mathbf{g}^k), \mathbf{g}^k - \tau_k H\mathbf{g}^k\right) \\ &= \left(H^{-1}(I-\tau_k H)\mathbf{g}^k, (I-\tau_k H)\mathbf{g}^k\right) \\ &= \left(H^{-1}(I-\tau_k H)^2\mathbf{g}^k, \mathbf{g}^k\right) \\ &= \left(H^{-1}\mathbf{g}^k, \mathbf{g}^k\right) - 2\tau_k(\mathbf{g}^k, \mathbf{g}^k) + \tau_k^2(H\mathbf{g}^k, \mathbf{g}^k) \\ &= \left(H^{-1}\mathbf{g}^k, \mathbf{g}^k\right) - \tfrac{(\mathbf{g}^k, \mathbf{g}^k)^2}{(H\mathbf{g}^k, \mathbf{g}^k)^2}(H\mathbf{g}^k, \mathbf{g}^k). \end{aligned}$$

Výsledkem je vztah

$$E(\mathbf{x}^{k+1}-\hat{\mathbf{x}}) - E(\mathbf{x}^k-\hat{\mathbf{x}}) = \frac{(\mathbf{g}^k, \mathbf{g}^k)^2}{(H\mathbf{g}^k, \mathbf{g}^k)(H^{-1}\mathbf{g}^k, \mathbf{g}^k)} E(\mathbf{x}^k-\hat{\mathbf{x}}),$$

odkud pak

$$E(\mathbf{x}^{k+1}-\hat{\mathbf{x}}) = \left\{1 - \frac{(\mathbf{g}^k, \mathbf{g}^k)^2}{(H\mathbf{g}^k, \mathbf{g}^k)(H^{-1}\mathbf{g}^k, \mathbf{g}^k)}\right\} E(\mathbf{x}^k-\hat{\mathbf{x}}).$$

Aplikujeme-li nyní Bergstromovo lemma na pozitivně definitní matici H, mající vlastní čísla $\lambda_1, ..., \lambda_N$, podle něhož

$$\begin{aligned} 1 \leq\ & \tfrac{(Hx,x)}{(x,x)} \tfrac{(x,H^{-1}x)}{(x,x)} \\ & \leq \tfrac{(\lambda_1+\lambda_N)^2}{4\lambda_1\lambda_N},\ \mathbf{x} \in \mathcal{R}^N, \end{aligned}$$

obdržíme odhad

$$E(\mathbf{x}^{k+1} - \hat{\mathbf{x}}) \leq \left[1 - \frac{4\lambda_1\lambda_N}{(\lambda_1 + \lambda_N)^2}\right] E(\mathbf{x}^k - \hat{\mathbf{x}}),$$

což vyjádřeno pomocí spektrálního čísla podmíněnosti $\kappa(H) = (\lambda_N/\lambda_1)$ poskytuje odhad

$$E(\mathbf{x}^{k+1} - \hat{\mathbf{x}}) \leq \left(\frac{\kappa(H) - 1}{\kappa(H) + 1}\right)^2 E(\mathbf{x}^k - \hat{\mathbf{x}}).$$

Vidíme tedy, že konvergenci metody největšího spádu lze popsat pomocí energetické normy vztahem

$$\|\mathbf{x}^{k+1} - \hat{\mathbf{x}}\|_H \leq \left(\frac{\kappa(H) - 1}{\kappa(H) + 1}\right) \|\mathbf{x}^k - \hat{\mathbf{x}}\|_H,$$

kterým je dokázána platnost prvního tvrzení dokazované věty.

Abychom odvodili platnost nerovnosti (5.70) uvědomme si, že platnost nerovnosti $\|\mathbf{x}^k - \hat{\mathbf{x}}\|_H \leq \epsilon\|\mathbf{x}^0 - \hat{\mathbf{x}}\|_H$ pro $k \geq \hat{k}$ je zaručena platností nerovnosti

$$\left(\frac{\kappa(H) - 1}{\kappa(H) + 1}\right)^k \leq \epsilon$$

pro $k \geq \hat{k}$.

Hledejme tedy nejmenší k takové, aby

$$\frac{1}{\epsilon} \leq \left(\frac{\kappa(H) + 1}{\kappa(H) - 1}\right)^k.$$

Jest totiž

$$\log\frac{1}{\epsilon} \leq k \log(1 + \sigma),$$

kde

$$\sigma = \frac{2}{\kappa - 1}.$$

Protože

$$\frac{1}{\sigma}\log\frac{1}{\epsilon} \leq \frac{1}{2}(\kappa - 1)\log\frac{1}{\epsilon},$$

odvodíme snadno, že hledaný index $p(\epsilon)$ splňuje nerovnost

$$p(\epsilon) \leq \frac{1}{2}\kappa(H)\log\frac{1}{\epsilon} + 1.$$

Tím je dokázána platnost vztahu (5.70) a tedy též věty 5.22. □

Předpodmínění.

Předpokládejme, že C je pozitivně definitní $N \times N$ matice faktorizovaná na tvar $C = EE^T$ a že

$$\mathbf{f}(\mathbf{x}) = \frac{1}{2}(H\mathbf{x}, \mathbf{x}) + (\mathbf{x}, \mathbf{b}) + \mathbf{c}, \tag{5.72}$$

kde H je pozitivně definitní $N \times N$ matice. Zaveďme další funkcionál $\tilde{\mathbf{f}}$ předpisem

$$\tilde{\mathbf{f}}(\mathbf{y}) = \mathbf{f}(E^T\mathbf{y}) = \frac{1}{2}(\tilde{H}\mathbf{y}, \mathbf{y}) + (\mathbf{y}, \tilde{\mathbf{b}}) + \tilde{\mathbf{c}}, \tag{5.73}$$

při čemž

$$\tilde{H} = E^{-1}H(E^T)^{-1},\ \tilde{\mathbf{b}} = E^{-1}\mathbf{b},\ \tilde{\mathbf{c}} = \mathbf{c}. \tag{5.74}$$

Zřejmě je $\tilde{H}$ symetrická a navíc ze vztahu $(\tilde{H}\mathbf{y}, \mathbf{y}) = (H\mathbf{x}, \mathbf{x}) > 0$ pro všechna $\mathbf{x} = (E^T)^{-1}\mathbf{y}$, plyne, že $\tilde{H}$ je pozitivně definitní.

Podobnostní transformace

$$(E^T)^{-1}\tilde{H}E^T = (E^T)^{-1}E^{-1}H = C^{-1}H$$

ukazuje, že matice $\tilde{H}$ a $C^{-1}H$ mají totožná spektra. Označme vlastní hodnoty matice $\tilde{H}$ symboly $\tilde{\lambda_1}, ..., \tilde{\lambda_N}$. Číslo podmíněnosti

$$\kappa(\tilde{H}) = \frac{\tilde{\lambda_N}}{\tilde{\lambda_1}} \tag{5.75}$$

je určeno maticemi C a H přesto, že $\tilde{H}$ závisí ještě na rozkladu $C = EE^T$.

Aplikujme nyní metodu největšího spádu na funkcionál (5.73). Obdržíme tak tento postup

$$\tilde{\mathbf{g}}^k = \tilde{H}\mathbf{y}^k - \tilde{\mathbf{b}}, \tag{5.76}$$

$$\tilde{\tau}_k = \frac{(\tilde{\mathbf{g}}^k, \tilde{\mathbf{g}}^k)}{(\tilde{H}\tilde{\mathbf{g}}^k, \tilde{\mathbf{g}}^k)}, \tag{5.77}$$

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \tilde{\tau}_k\tilde{\mathbf{g}}^k,\ k = 0, 1, ... \tag{5.78}$$

kde $\mathbf{y}^0$ lze volit libovolně.

Již víme, že platí

$$\lim_{k\to\infty} \mathbf{y}^k = \hat{\mathbf{y}} = \tilde{H}^{-1}b, \tag{5.79}$$

a ž rychlost konvergence je určena číslem podmíněnosti $\kappa(\tilde{H})$.

Položme $\mathbf{x}^k = (E^T)^{-1}\mathbf{y}^k$ a $\mathbf{g}^k = H\mathbf{x}^k - \mathbf{b}$, $k = 0, 1, ...$ Snadno se přesvědčíme, že platí vztahy

$$\tilde{\mathbf{g}}^k = (E^T)^{-1}\mathbf{g}^k,\ \tilde{\tau}_k = \frac{(\mathbf{g}^k, \mathbf{h}^k)}{(H\mathbf{g}^k, \mathbf{g}^k)},\ \mathbf{y}^{k+1} = E^T(\mathbf{x}^k - \tilde{\tau}_k\mathbf{h}^k),$$

kde $\mathbf{h}^k = C^{-1}\mathbf{g}^k$.

Jelikož $\mathbf{y}^{k+1} = E^T\mathbf{x}^{k+1}$, výše uvedené vztahy lze realizovat též pomocí schématu

$$\mathbf{g}^k = H\mathbf{x}^k - \mathbf{b}, \tag{5.80}$$

(5.81) $$\mathbf{h}^k = C^{-1}\mathbf{g}^k,$$

(5.82) $$\tau_k = \frac{(\mathbf{h}^k, \mathbf{g}^k)}{(H\mathbf{h}^k, \mathbf{g}^k)},$$

(5.83) $$\mathbf{x}^{k+1} = \mathbf{x}^k - \tau_k \mathbf{h}^k,\ k = 0, 1, ...$$

Extrémní vektory pro $\mathbf{f}$ a $\tilde{\mathbf{f}}$ jsou dány výrazy

$$\hat{\mathbf{x}} = H^{-1}\mathbf{b},\ \text{a}\ \hat{\mathbf{y}} = \tilde{H}^{-1}\tilde{\mathbf{b}},$$

při čemž

$$\hat{\mathbf{y}} = E^T\hat{\mathbf{x}}.$$

Dále pak

$$\mathbf{y}^k - \hat{\mathbf{y}} = E^T(\mathbf{x}^k - \hat{\mathbf{x}})$$

a

$$\|\mathbf{y}^k - \hat{\mathbf{y}}\|_{\tilde{H}} = \|\mathbf{x}^k - \hat{\mathbf{x}}\|_{\tilde{H}}.$$

Na základě věty 5.22 však platí

$$\|\mathbf{y}^k - \hat{\mathbf{y}}\|_{\tilde{H}} \leq \left(\frac{\kappa(\tilde{H})-1}{\kappa(\tilde{H})+1}\right)^k \|\mathbf{y}^k - \hat{\mathbf{y}}\|_{\tilde{H}}$$

$$= \left(\frac{\kappa(\tilde{H})-1}{\kappa(\tilde{H})+1}\right)^k \|\mathbf{x}^k - \hat{\mathbf{x}}\|_H.$$

Posloupnost $\{\mathbf{x}^k\}$, kde $\mathbf{x}^0$ lze volit libovolně, konstruovaná pomocí vztahů (5.80) - (5.83), konverguje k $\hat{\mathbf{x}}$ a její rychlost konvergence je určena číslem $\kappa(\tilde{H})$. Lze-li najít matici C tak, aby $\kappa(\tilde{H}) < \kappa(H)$, pak odhad (5.70) napovídá, že rychlost konvergence (5.80) - (5.83) je lepší než (5.66) - (5.68).

Matice C se nazývá *matice předpodmínění metody největšího spádu*. Metoda určená vzorci (5.76) - (5.79) se nazývá *transformovanou předpodmíněnou metodou největšího spádu*, zatímco metoda určená vzorci (5.80) - (5.83) se nazývá *netransformovanou předpodmíněnou metodou největšího spádu*. Názvy jsou odvozeny odtud, že extrém $\hat{\mathbf{y}} = E^T\hat{\mathbf{x}}$, pro metodu danou vzorci (5.76) - (5.79) je transformovaným extrémem pro (5.80) - (5.83), tedy pro $\hat{\mathbf{x}}$. Platí též, že $\mathbf{x}^k = (E^T)^{-1}\mathbf{y}^k$, a tuto transformaci lze ale není nutné provádět pro každé $k = 0, 1, ...$

Výpočtový proces prováděný na základě výše uvedených algoritmů se budeme snažit ocenit z pozic jejich složitosti.

Potřebné gradienty lze stanovit rekurzivně. Jmenovitě pak (5.76) a (5.80) lze nahradit formulemi

$$\tilde{\mathbf{g}}^k = \tilde{\mathbf{g}}^{k-1} - \tilde{\tau}_{k-1}\tilde{H}\tilde{\mathbf{g}}^{k-1}$$

a

$$\mathbf{g}^k = \tau_{k-1}H\mathbf{h}^{k-1}.$$

Vektory $\tilde{H}\tilde{\mathbf{g}}^{k-1}$ a $H\mathbf{h}^{k-1}$ jsou k dispozici z předchozího iteračního kroku.

Nás zajímají případy, v nichž rozměr matice H je značný a matice H sama je řídká. Matice E bývá zpravidla řídká trojúhelníková matice. Při realizaci transformované metody narážíme na skutečnost, že $\tilde{H}$ nebývá řídká a tudíž se explicitnímu výpočtu snažíme vyhnout.

Snadno nahlédneme, že v rámci jednoho iteračního kroku je výpočet vektoru $\tilde{H}\tilde{\mathbf{g}}^k$ v (5.77) nejpracnější. Je tudíž záhodno provádět jeho výpočet nepřímo. Řešíme proto napřed soustavu

$$E^T\mathbf{z} = \mathbf{g}^k$$

vzhledem k $\mathbf{z}$, poté položíme

$$\mathbf{z}^* = H\mathbf{z}$$

a řešíme

$$E\mathbf{z}^{**} = \mathbf{z}^*$$

vzhledem k $\mathbf{z}^{**} = E^{-1}H(E^T)^{-1}\mathbf{g}^k = \tilde{H}^{-1}\mathbf{g}^k$. Je patrné, že abychom tento vektor určili, musíme řešit dvě soustavy s trojúhelníkovými maticemi a provézt jedno násobení maticí H. Dále pak se musí provézt transformace $\mathbf{x}^k = (E^T)^{-1}\mathbf{y}^k$ a to opět vyžaduje řešení jedné soustavy s trojúhelníkovou maticí.

Nejpodstatnější složkou výpočtu podle netransformované metody (5.80) - (5.83) je krok (5.81), kde se počítá $C^{-1}\mathbf{g}^k$ a krok (5.82) kde se stanovuje $H\mathbf{h}^k$. Vektor $C^{-1}\mathbf{g}^k$ se vypočítává tak, že se řeší

$$E\mathbf{z} = \mathbf{g}^k$$

vzhledem k $\mathbf{z}$ a

$$E^T\mathbf{h}^k = \mathbf{z}$$

se řeší vzhledem k $\mathbf{h}^k$. Potom zřejmě $\mathbf{h}^k = C^{-1}\mathbf{g}^k$.

I v netransformované metodě se řeší dvě soustavy s trojúhelníkovými maticemi. Mohlo by se zdát, že netransformovaná metoda je výhodnější, protože poskytuje přímo hledaný extrémní vektor $\hat{\mathbf{x}}$. Existují však vhodné transformační matice, pro něž výpočtová náročnost není větší než pro metodu netransformovanou. Toto tvrzení se pokusíme ozřejmit pro případ předpodmíněné metody sdružených gradientů, tedy metody značně efektivnější než je metoda největšího spádu. Tvrzení samo je však pravdivé i pro metodu největšího spádu.

5.5 Metoda sdružených gradientů

V tomto článku ukážeme jak sestrojovat účinné aproximace řešení soustavy lineárních algebraických rovnic

$$Ax = b,\ x, b \in \mathcal{E} = \mathcal{R}^N, \tag{5.84}$$

kde A je pozitivně definitní $N \times N$ matice, pomocí extremalizace odpovídajícího funkcionálu

$$F(x) = (Ax, x) - (b, x) - (x, b). \tag{5.85}$$

Odvození provedeme simultánně jak pro základní verzi metody sdružených gradientů tak pro verzi předpodmíněnou.

Libovolně zvolme základní vektor aproximace x_0 a položme $r_0 = b - Ax_0$. Dále definujme *Krylovovy prostory* (vzhledem k r_0 a A) kladouce

$$\mathcal{V}_k = Lin\left\{r_0, Ar_0, ..., A^{k-1}r_0\right\},$$

při čemž symbol *Lin* znamená lineární obal vektorů $\{r_0, ..., A^{k-1}r_0\}$.

Jakožto k-tou iteraci metody sdružených gradientů definujeme Galerkinovu aproximaci $\tilde{x}_k$ vektoru $x_k - x$ v prostoru $\mathcal{V}_k$, tedy

$$(A\tilde{x}_k, \phi) = (r_0, \phi),\ \forall\phi \in \mathcal{V}_k,\ k = 1, 2, ... \tag{5.86}$$

Položme $x_k = \tilde{x}_k + x_0$. Jest potom (5.86) totéž co

$$(Ax_k, \phi) = (b, \phi)\ \forall\phi \in \mathcal{V}_k \tag{5.87}$$

a $x_k - x_0$ patří do $\mathcal{V}_k$.

Dále odvodíme metodu sdružených gradientů s předpodmíněním. Buď B další symetrická pozitivně definitní matice typu $N \times N$. Jest tudíž soustava (5.84) ekvivalentní se soustavou

$$BAx = Bb. \tag{5.88}$$

Definujme dále nový skalární součin na $\mathcal{E}$ formulí $[.,.] = (B^{-1}.,.)$. Snadno zjistíme, že operátor $\mathcal{A} = BA$ je symetrický a pozitivně definitní vzhledem k součinu $[.,.]$, t.j. $[\mathcal{A}x, y] = x, \mathcal{A}y]$ a existuje $\gamma > 0$ takové, že $[\mathcal{A}x, x] \geq \gamma[x, x], x \in \mathcal{E}$.

Buď $y_0 \in \mathcal{E}$ libovolný startovní nástřel a $z_0 = Bb - \mathcal{A}y_0$. Definujme

$$\tilde{\mathcal{V}} = Lin\left\{z_0, \mathcal{A}z_0, ..., \mathcal{A}^{k-1}z_0\right\}, k = 1, 2, ...$$

k-tý člen metody sdružených gradientů pro soustavu s maticí A a předpodmiňovačem B (PCG) je definována jako Galerkinova aproximace $\tilde{y}_k$ vektoru $x - y_0$ t.j. pro $k = 1, 2, ...$

$$[\mathcal{A}\tilde{y}_k, \phi] = [z_0, \phi],\ \forall\phi \in \tilde{\mathcal{V}}. \tag{5.89}$$

Pro $y_k = \tilde{y}_k + y_0$ je posléze (5.89) totéž co

$$(Ay_k, \phi) = (b, \phi) \tag{5.90}$$

$\forall\phi \in \tilde{\mathcal{V}}$ a $y_k - y_0$.

Porovnáme-li (5.86) a (5.89) zjistíme, že PCG je formálně podobné s CG dosadíme-li $\mathcal{A}$ namísto A a $[.,.]$ namísto $(.,.)$. Za zmínku stojí též skutečnost, že PCG aproximaci $y_k - y_0$ lze interpretovat jako standardní Galerkinovu aproximaci vektoru $x - x_0$ v prostoru $\tilde{\mathcal{V}}_k$. Volíme-li $B = I$ pak $\tilde{\mathcal{V}}_k = \mathcal{V}_k$ a CG je tak speciální případ PCG.

Poznámka 5.1 Podle Hamiltonovy-Cayleyovy věty (blíže se o ní čtenář dozví v monografii I. Marek, K. Žitný: *Analytická teorie matic pro aplikované vědy*, Teubner Verlag, Díl I. z 1984 a díl II. z 1986) lze inverzní matici A^{-1} vyjádřiti jakožto polynom stupně nejvýše $N - 1$, kde N je dimenze prostoru $\mathcal{E}$. Protože $x - x_0 = A^{-1}b \in \mathcal{V}_N$, při exaktním počítání metoda CG docílí hledané řešení po nejvýše N krocích. Tato okolnost však ustupuje svou důležitostí do pozadí skutečnosti jiné, že totiž potřebnou přesnost aproximace lze docílit za pomoci daleko menšího počtu kroků než je N. V praxi lze totiž docíliti toho, že při vhodném předpodmiňovači počet kroků potřebný k zaručení žádané přesnosti je typická situace kdy, řekněme $k = 10$ či $k = 20$ zatímco dimenze $N = 10^6$ a ta může být i větší.

Analýza konvergence

Metody CG a PCG, vlastně však jen PCG, budeme analyzovat za použití energetické normy $[\mathcal{A}.,.]^{1/2} = (A.,.)^{1/2}$.

Z formule (5.89) plyne platnost vztahů

$$[\mathcal{A}(\tilde{y}_k - (x - y_0)), \phi] = 0, \ \forall \phi \in \tilde{\mathcal{V}}_k. \tag{5.91}$$

Položíme-li tedy $e_k = \tilde{y}_k - (x - y_0) = y_k - x$ vyplynou nám následující vztahy

$$[\mathcal{A}e_k, e_k] = [\mathcal{A}e_k, y - (x - y_0)], \tag{5.92}$$

platné pro všechny prvky $y \in \tilde{\mathcal{V}}_k$. Nyní použijeme Schwarzovy nerovnosti pro skalární součin (5.92). Postupně obdržíme

$$\begin{aligned}
[\mathcal{A}e_k, y - (x - y_0)] &= (Ae_k, y - (x - y_0)) \\
&= (A^{1/2}e_k, A^{1/2}(y - (x - y_0)) \\
&\leq (A^{1/2}e_k, A^{1/2}e_k)^{1/2}(A^{1/2}(y - (x - y_0), A^{1/2}(y - (x - y_0))^{1/2} \\
&= (Ae_k, e_k)^{1/2}(A(y - (x - y_0), (y - (x - y_0))^{1/2}
\end{aligned}$$

takže též,

$$[\mathcal{A}e_k, e_k] \leq [\mathcal{A}e_k, e_k]^{1/2}[\mathcal{A}(y - (x - y_0), (y - (x - y_0)]^{1/2}$$

a posléze

$$[\mathcal{A}e_k, e_k] \leq [\mathcal{A}(y - (x - y_0), y - (x - y_0)] \tag{5.93}$$

pro jakýkoliv vektor $y \in \tilde{\mathcal{V}}_k$.

Protože $y \in \tilde{\mathcal{V}}_k$, lze zvolit libovolný polynom P_{k-1} stupně nejvýše $k - 1$ a položit

$$y = P_{k-1}(\mathcal{A})z_0 \equiv P_{k-1}(\mathcal{A})\mathcal{A}(x - y_0).$$

S touto volbou nabude formule (5.93) tvaru

$$[\mathcal{A}e_k, e_k] \leq [\mathcal{A}(P_{k-1}(\mathcal{A})\mathcal{A} - I)(x - y_0), (P_{k-1}(\mathcal{A})\mathcal{A} - I)(x - y_0)]. \tag{5.94}$$

Nechť

$$Q_k(\mathcal{A}) = I - P_{k-1}(\mathcal{A})\mathcal{A}.$$

Potom

$$\begin{cases} [\mathcal{A}Q_k(\mathcal{A})(x - y_0), Q_k(\mathcal{A})(x - y_0)] \\ \\ \leq [[Q_k(\mathcal{A})]]^2[\mathcal{A}(x - y_0), (x - y_0)], \end{cases} \tag{5.95}$$

kde $[[.]]$ značí normu indukovanou skalárním součinem $[.,.]$. Tudíž (5.94) a (5.95) implikují platnost relace

$$(Ae_k, e_k) \leq [[Q_k(\mathcal{A})]]^2(\mathcal{A}(x - y_0), x - y_0) \tag{5.96}$$

pro každý polynom Q_k stupně k splňující podmínku $Q_k(0) = 1$. Protože $\mathcal{A}$ je pozitivně definitní operátor, jeho spektrum je reálné a kladné. Buďte $\underline{\lambda}$ a $\overline{\lambda}$ hranice intervalu, v němž leží spektrum $\sigma(\mathcal{A})$. Položme ještě

$$\mathcal{M} = I - \frac{2}{\underline{\lambda} + \overline{\lambda}}\,\mathcal{A}$$

a nechť pro každý polynom $\tilde{Q}_k$ stupně k splňující $\tilde{Q}_k(1) = 1$ položme

$$Q_k(t) = \tilde{Q}_k(1 - \frac{2}{\underline{\lambda} + \overline{\lambda}}\,t).$$

Potom platí,že Q_k je polynom stupně k, splňuje $Q_k(0) = 1$ a

$$Q_k(\mathcal{A}) = \tilde{Q}_k(\mathcal{M}).$$

Tudíž,

$$[[Q_k(\mathcal{A})]] = [[\tilde{Q}_k(\mathcal{M})]]. \tag{5.97}$$

Nyní pro takový polynom $\tilde{Q}_k$

$$[[\tilde{Q}_k(\mathcal{M})]] = \mathrm{Max}\left\{|\tilde{Q}_k(\lambda)| : \lambda \in \sigma(\mathcal{M})\right\} = \mathrm{Max}\left\{|\tilde{Q}_k(\lambda)| : \lambda \in [-\rho, \rho]\right\}, \tag{5.98}$$

kde

$$\rho = \frac{\overline{\lambda} - \underline{\lambda}}{\underline{\lambda} + \overline{\lambda}} < 1.$$

Z (5.96), (5.97) a (5.98) plyne platnost formule

$$(Ae_k, e_k) \leq \mathrm{Max}\left\{|\tilde{Q}_k(\lambda)| : \lambda \in [-\rho, \rho]\right\} (A(x - y_0), x - y_0)^{1/2} \tag{5.99}$$

pro jakýkoliv polynom $\tilde{Q}_k$ splňující $\tilde{Q}_k(1) = 1$. Se znalostí jediné veličiny ρ je optimální volbou polynom určený pomocí Čebyshevových polynomů T_k, t.j.

$$\tilde{Q}_k(\lambda) = \frac{T_k(\lambda/\rho)}{T_k(1/\rho)}$$

Čebyševovy polynomy jsou definovány pomocí formulí

$$T_k(t) = \begin{cases} \cos(k\cos^{-1}t), & \text{je-li } |t| \leq 1 \\ \\ \cosh(k\cosh^{-1}t), & \text{je-li } t \geq 1. \end{cases}$$

Platí tedy,

$$\mathrm{Max}\left\{|\tilde{Q}_k(\lambda)| : \lambda \in [-\rho, \rho]\right\} \leq \frac{1}{T_k(1/\rho)}. \tag{5.100}$$

Protože $\rho < 1$, můžeme klást $\sigma = \cosh^{-1}(1/\rho)$, takže

$$T_k(1/\rho) = \frac{e^{k\sigma} + e^{-k\sigma}}{2} = e^{k\sigma}\left[\frac{1 + e^{-2k\sigma}}{2}\right].$$

Dále pak,

$$\frac{1}{T_k(\frac{1}{\rho})} = e^{-k\sigma}\left[\frac{2}{1+e^{-2k\sigma}}\right] \leq 2e^{-k\sigma}. \tag{5.101}$$

Avšak cosh $\sigma = \frac{1}{\rho}$ a sinh $\sigma = \sqrt{\frac{1}{\rho^2}-1}$, takže cosh$\sigma$+sinh $\sigma = e^{\sigma}$ a tudíž, log $\left[\frac{1}{\rho}+\sqrt{\frac{1}{\rho^2}-1}\right]$ a tedy,

$$e^{-k\sigma} = \left[\frac{1}{\rho}+\sqrt{\frac{1}{\rho^2}-1}\,\right]^{-k} = \left(\frac{\rho}{1+\sqrt{1-\rho^2}}\right)^k. \tag{5.102}$$

Do hry se dostává číslo podmíněnosti $\kappa(\mathcal{A}) = \frac{\overline{\lambda}}{\underline{\lambda}}$ takže jest potom

$$\rho = \frac{\kappa(\mathcal{A})-1}{\kappa(\mathcal{A})+1}.$$

Kombinujíce (5.101) a (5.102) zjistíme, že

$$\frac{1}{T_k(1/\rho)} \leq 2\left(\frac{\kappa(\mathcal{A})-1}{\kappa(\mathcal{A})+1}\right)^k \tag{5.103}$$

a žádaný odhad má konečný tvar

$$(Ae_k, e_k)^{1/2} \leq 2\left(\frac{\kappa(\mathcal{A})-1}{\kappa(\mathcal{A})+1}\right)^k (A(x-y_0), x-y_0)^{1/2}.$$

Za zmínku jistě stojí, že asymptoticky dává metoda sdružených gradientů lepší odhad než stacionární iterační proces s iterační maticí $\mathcal{A} = BA$. Jest totiž

$$\frac{\kappa(BA)-1}{\kappa(BA)+1} = \frac{\kappa(BA)-1}{\kappa(BA)+1+2\kappa^{1/2}(BA)} \leq \frac{\kappa(BA)-1}{\kappa(BA)+1}.$$

5.6 Zobecněné Bergstromovo lemma pro pozitivně definitní matice

Při vyšetřování konvergence některých metod řešení soustav lineárních algebraických rovnic bude užitečné následující lemma.

Lemma 5.4 *Buď H pozitivně definitní matice typu $N \times N$, $N \geq 1$. Nechť pro její spektrum $\sigma(H) = \{\lambda_j : j = 1, ..., s\}$, $s \leq N$ platí vztahy*

$$\lambda_{\min} = \lambda_1 < \lambda_j < ... < \lambda_k < ... < \lambda_s = \lambda_{\max},\ j, k = 1, ..., s,\ j \neq k.$$

Potom pro libovolný vektor $x \in \mathcal{R}^N$ platí nerovnosti

$$(x,x)^2 \leq (Hx,x)\,(H^{-1}x,x) \leq \frac{1}{\lambda_{\min}\lambda_{\max}}\left(\frac{\lambda_{\min}+\lambda_{\max}}{2}\right)^2 (x,x)^2. \tag{5.104}$$

Poznámka 5.7 *Lemma 5.4 podává elegantní odhad dolní hranice čísla podmíněno- sti matice H dané veličinou $V(x) = (Hx, x)\ (H^{-1}x, x)$ pomocí podílu čtverců aritmetického a geometrického průměru jejích extrémních vlastních hodnot.*

Důkaz provedeme za použití následujících dvou lemmat.

Lemma 5.5 *Nechť $0 < \mu \leq \nu$,*

$$f(x) = \mu\frac{1}{x} + \nu x,\ 0 < a \leq x \leq 1.$$

Potom

$$f(x) \leq \max\left\{\mu\frac{1}{a} + \nu a, \mu + \nu\right\}. \tag{5.105}$$

Důkaz. Snadno nahlédneme, že ze vztahů

$$f'(x) = -\frac{\mu}{x^2} + \nu = 0,\ f''(x) = \frac{2\mu}{x^2} > 0,$$

plyne, že f je konvexní takže

$$\max\{f(x) : x \in [a, 1]\} = \max\{f(a), f(1)\}$$

a odtud tvrzení snadno plyne. □

Nechť $\zeta \in \mathcal{R}^N,\ \zeta^T = (\zeta_1, ..., \zeta_N),\ \zeta_j \geq 0, j = 1, ..., s$ a

$$\sum_{j=1}^{s} \zeta_j = 1 \tag{5.106}$$

Definujme dvě veličiny

$$\phi(\zeta) = \sum_{j=1}^{s} \zeta_j\lambda_j,\ s = \mathrm{card}\sigma(T), \tag{5.107}$$

$$\psi(\zeta) = \sum_{j=1}^{s} \zeta_j\lambda_j^{-1}. \tag{5.108}$$

Lemma 5.6 *Pro veličiny definované v (5.107) a (5.108) platí následující tvrzení (a), (b), (c):*

(a) $\lambda_{\min}\lambda_{\max}\psi(\zeta) \leq (\lambda_{\min} + \lambda_{\max}) - \phi(\zeta)$,

(b) Existuje $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ *takové, že* $\phi(\zeta) = \lambda$.

(c)

$$\phi(\zeta)\psi(\zeta) \leq \left(\frac{\frac{1}{2}\left(\lambda_{\min} + \lambda_{\max}\right)}{\left(\sqrt{\lambda_{\min}\lambda_{\max}}\right)}\right)^2.$$

Proof. Abychom ukázali platnost tvrzení (a), vyšetřujme výrazy

$$\begin{aligned}\phi(\zeta)+\lambda_{\min}\lambda_{\max}\psi(\zeta) &= \lambda_{\min}\textstyle\sum_{j=1}^{s}\zeta_j\frac{\lambda_j}{\lambda_{\min}}+\lambda_{\max}\sum_{j=1}^{s}\zeta_j\frac{\lambda_{\min}}{\lambda_j}\\ &= \textstyle\sum_{j=1}^{s}\zeta_j\left[\lambda_{\min}\frac{\lambda_j}{\lambda_{\min}}+\lambda_{\max}\frac{\lambda_{\min}}{\lambda_j}\right].\end{aligned}$$

Aplikujeme-li Lemma 5.5 s

$$a=\frac{\lambda_{\min}}{\lambda_{\max}},\ \mu=\lambda_{\min},\ \nu=\lambda_{\max},$$

obdržíme, že

$$\phi(\zeta)+\lambda_{\min}\lambda_{\max}\psi(\zeta)\le\sum_{j=1}^{s}\zeta_j\left(\lambda_{\min}+\lambda_{\max}\right)=\lambda_{\min}+\lambda_{\max}.$$

Tudíž, (a) platí.

Protože

$$\lambda_{\min}\le\phi(\zeta)\le\lambda_{\max},$$

platí i (b).

Konečně, na základě platnosti vztahů (a) a (b), a vědomi si toho, že pro libovolné reálné $\alpha>0$ platí vztahy

$$\max\left\{(\alpha-\lambda)\lambda:0<\lambda_{\min}\le\lambda\le\lambda_{\max}\right\}=(1/4)\alpha^2$$

odvodíme, že

$$\begin{aligned}\phi(\zeta)\psi(\zeta) &\le \tfrac{\lambda}{\lambda_{\min}\lambda_{\max}}\left(\lambda_{\min}+\lambda_{\max}-\lambda\right)\\ &\le \tfrac{1}{\lambda_{\min}\lambda_{\max}}\left(\tfrac{1}{2}\left(\lambda_{\min}+\lambda_{\max}\right)\right)^2,\end{aligned}$$

Tím je důkaz lemmatu 5.5 proveden. □

Důkaz lemmatu 5.4. Je li $s = 1$, pak $H = \lambda I$, a tvrzení platí.
Nechť tedy $s > 1$. Vyjděme ze spektrálního vyjádření

$$H = \sum_{j=1}^{s} \lambda_j P_j,$$

v němž

$$P_j^T = P_j = P_j^2,$$

$$P_j P_k = P_k P_j = \delta_{jk} P_j,\ j,k = 1,...,s,\ \sum_{j=1}^{s} P_j = I.$$

Pro $0 \neq x \in \mathcal{R}^N$ jest tedy,

$$V(x) = (Hx,x)\ (H^{-1}x,x) = \sum_j \lambda_j (P_j x,x) \sum_k \lambda_k^{-1} (P_k x,x)$$

a dále pak

$$V(x) = \sum_j (P_j x,x)^2 + \sum_j \sum_{k>j} \left(\frac{\lambda_j}{\lambda_k} + \frac{\lambda_k}{\lambda_j} \right) (P_j x,x)\ (P_k x,x), \tag{5.109}$$

takže

$$V(x) \geq \sum_j (P_j x,x)^2 + 2 \sum_j \sum_{k>j} (P_j x,x)\ (P_k x,x) =$$

$$\left(\sum_j (P_j x,x) \right)^2 = (x,x)^2.$$

Na druhé straně, položme

$$\zeta_j = \frac{(P_j x,x)}{(x,x)},\ j = 1,...,s$$

a aplikujme lemma 5.6. Tím však obdržíme platnost tvrzení dokazovaného v lemmatu 5.4.
□

Poznámka 5.8 *Je zřejmé, že pro*

$$H = \lambda I,$$

se nabude extrémní dolní (i horní) hranice

$$(Hx,x)\ (H^{-1}x,x) = (x,x)^2.$$

Podobně, pro

$$H = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix},$$

se nabývá extrémní horní hranice

$$(Hx,x)\ (H^{-1}x,x) = x_1^4 + x_2^4 + \left(\frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1} \right)\ x_1^2 x_2^2 =$$

$$(x_1^2 + x_2^2)^2 \left(\frac{1}{2}(\lambda_1 + \lambda_2) \right)^2 \frac{1}{\lambda_1 \lambda_2}.$$

Cvičení 5.3 *Nalezněte všechny případy pozitivně definitních matic H, pro něž nastane některá z extrémních situací ve formuli (5.104).*

Zobecněné Bergstromovo lemma pro operátory na nekonečně dimenzionálním Hilbertově prostoru

Předpokládejm,e, že $\mathcal{E}$ je Hilbertův prostor a že $\mathbf{B}(\mathcal{E})$ značí prostor ohraničených lineárních operátorů zobrazujících $\mathcal{E}$ do sebe.

Lemma 5.7 *Buď $T \in \mathbf{B}(\mathcal{E})$ samoadjungovaný pozitivně definitní operátor.*
Potom pro libovolný prvek $x \in \mathcal{E}$ platí vztahy

$$(x,x)^2 \leq (Hx,x)\,(H^{-1}x,x) \leq \frac{1}{\lambda_{\min}\lambda_{\max}}\,\left(\frac{\lambda_{\min}+\lambda_{\max}}{2}\right)^2\,(x,x)^2, \tag{5.110}$$

při čemž $\lambda_{\min}$ a $\lambda_{\max}$ značí hranice spektra operátoru T, t. j. $\sigma(T) \subset [\lambda_{\min}, \lambda_{\max}]$ a $\lambda_{\min}, \lambda_{\max} \in \sigma(T)$.

Důkaz. Podobně jako formulace tak i důkaz lemmatu 5.7 je v podstatě totožný s důkazem lemmatu 5.4.

Buď $E = E(\lambda)$ spektrální míra operátoru T.

Vzhledem k předpokladu o nekonečnosti dimenze $\mathcal{E}$ je nutné upravit důkaz čáasti (a) v lemmatu 5.6, při čemž klademe

$$\phi(x) = \int_{\lambda_{\min}}^{\lambda_{\max}} \lambda(\mathrm{d}E(\lambda)x,x),\ \psi(x) = \int_{\lambda_{\min}}^{\lambda_{\max}} \lambda^{-1}(\mathrm{d}E(\lambda)x,x).$$

Platí tedy,

$$\phi(x) + \lambda_{\min}\lambda_{\max}\psi(x) = \int_{\lambda_{\min}}^{\lambda_{\max}} \left[\lambda_{\min}\frac{\lambda}{\lambda_{\min}} + \lambda_{\max}\frac{\lambda_{\min}}{\lambda}\right](\mathrm{d}E(\lambda)x,x)$$

a tedy též

$$\begin{cases} \phi(x) + \lambda_{\min}\lambda_{\max}\psi(x) \leq (\lambda_{\min}+\lambda_{\max})\int_{\lambda_{\min}}^{\lambda_{\max}}(\mathrm{d}E(\lambda)x,x) \\ \\ \qquad\qquad = (\lambda_{\min}+\lambda_{\max})(x,x). \end{cases} \tag{5.111}$$

Nerovnost ve vztahu (5.111) jsme odvodili pomocí lemmatu 5.5, kde jsme položili

$$a = \frac{\lambda_{\min}}{\lambda_{\max}},\ \mu = \lambda_{\min},\ \nu = \lambda_{\max}.$$

Dokázali jsme tak platnost tvrzení analogické k (a) v lemmatu 5.6. Ostatní části důkaz lemmatu 5.7 jsou shodné s odpovídajícími částmi důkazu lemmatu 5.6. Tímto konstatováním lze ukončit důkaz lemmatu 5.7. □

6 Problémy vlastních hodnot

Jest na místě si uvědomit, že úloha nalézt vlastní číslo a příslušný vlastní vektor dané čtvercové matice je úloha *nelineární*. Zkušenost nám napovídá, že na rozdíl od lineárních úloh, pro které zpravidla jsme schopni nalézt finitní metody k nalezení přesných řešení, pro nelineární úlohy takové metody známy nejsou. Tak je tomu i v případě hledání vlastních čísel a vlastních vektorů. Dá se dokonce říci, že iterační metody jsou v tomto případě jedinými spolehlivými prostředky jak sestrojovat patřičné aproximace hledaných vlastních prvků.

Základní metodou v problematice vlastních čísel je *mocninná metoda*.

6.1 Mocninná metoda

Konvergence mocninné metody je založena na jednoduchém pozorování, že totiž posloupnost $\{\lambda^k\}$, kde λ je komplexní číslo, je buď *divergentní* a to když $|\lambda| \geq 1$, $\lambda \neq 1$, *stacionární*, když $\lambda = 1$ anebo *konvergentní k nule*, když $|\lambda| < 1$.

Tuto skutečnost reflektuje posloupnost $\{T^k\}$, kde T je matice typu $N \times N$ s komplexními prvky t_{jk}, $j, k = 1, 2..., N$.

Protože chování posloupnosti mocnin $\{T^k\}$ je pro vyšetřování algoritmů k nalezení vlastních hodnot matice T rozhodující, bude užitečné připomenout si některé důležité vlastnosti funkcí, jejichž argumentem je matice. K tomu nám poslouží známé tvrzení z lineární algebry

Tvrzení 6.1 *Buď $T = (t_{jk})$ matice typu $N \times N$ nad tělesem komplexních čísel. Buď*

$$\sigma(T) = \{\lambda_1, \ ..., \lambda_N\}, \ s \leq N,$$

spektrum matice T a nechť

$$r_1, \ ..., \ r_s, \ \sum_{j=1}^{s} r_j = N,$$

jsou násobnosti vlastních hodnot $\lambda_1, ..., \lambda_s$ v charakteristickém mnohočenu

$$\chi_T(\lambda) = det(\lambda I - T) = \prod_{j=1}^{s} (\lambda - \lambda_j)^{r_j}.$$

Dále nechť ϕ označuje minimální polynom *matice T, t.j.*

$$\phi(\lambda) = \prod_{j=1}^{s} (\lambda - \lambda_j)^{q_j},$$

při čemž $1 \leq q_j \leq r_j$.

Potom existuje $N \times N$ matice V, $\det V \neq 0$, *taková, že*

$$VTV^{-1} = \begin{pmatrix} \mathcal{J}_1 & & & \\ & . & & \\ & & . & \\ & & & . \\ & & & & \mathcal{J}_s \end{pmatrix}. \tag{6.1}$$

V tomto vyjádření je každý z bloků $\mathcal{J}_{jr_j}$ sám o sobě blokově diagonální matice. Pro případ $q_j = 1$ jsou její diagonální bloky skalární matice *t.j. $\mathcal{J}_{jl} = \lambda_j I$, kde I je jednotková matice příslušného rozměru. Pro $q_j > 1$ mohou některé diagonální bloky být skalární matice avšak alespoň jeden z bloků má tvar*

$$J_{jl} = \begin{pmatrix} \lambda_j & 1 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 1 \\ & & & & \lambda_j \end{pmatrix},$$

$$\dim J_{jl} = t_{j_l}$$

při čemž $t_1, ..., t_s$ jsou přirozená čísla, (tedy nikoliv nula) a platí rovnosti

$$\sum_{l=1}^{t_j} \dim J_{jl} = r_j,$$

$$\max\{\dim J_{jl} : \ j = 1, ...s; \ l = 1, .., t_j\} = q_j.$$

Posléze diagonální blok $\mathcal{J}_j$ ve vyjádření (6.1) tvoří všechny bloky odpovídající vlastní hodnotě $\lambda_j, j = 1, ..., s$.

Příklad 6.1 *Nechť spektrum matice T tvoří vesměs navzájem různá vlastní čísla, tedy $s = N$. Potom $q_j = r_j = 1$ a všechny bloky $\mathcal{J}_{jl}$, $j = 1, ..., N$, jsou nutně diagonální, takže*

$$VTV^{-1} = \mathrm{diag}\{\lambda_1, ..., \lambda_N\}.$$

Příklad 6.2 *Nechť $T = \lambda I$, $N > 1$. Zřejmě $\sigma(T) = \{\lambda\}$. Tudíž $s = 1 < N$ a $r_1 = N$, zatímco $q_1 = 1$.*

Příklad 6.3 *Nechť $N > 1$ a*

$$T = \begin{pmatrix} \lambda & 1 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 1 \\ & & & & \lambda \end{pmatrix}.$$

Zřejmě, $\sigma(T) = \{\lambda\}$, $s = 1$ a $r_1 = N$. Na rozdíl od příkladu 6.1, $q_1 = N$.

Definice 6.1 *Matice T typu $N \times N$ má* jednoduchou strukturu, *jestliže $q_j = 1$ pro $j = 1, ..., s$, kde s značí počet různých vlastních hodnot matice T.*

Poznámka 6.1 *Matice v příkladech 6.1 a 6.2 mají jednoduchou strukturu; matice z příkladu 6.3 jednoduchou strukturu nemá.*

Zaveďme následující označení.

$$B_{j1} = V^{-1} \begin{pmatrix} 0 & & & & & \\ & . & & & & \\ & & . & & & \\ & & & I_{r_j} & & \\ & & & & . & \\ & & & & & . \\ & & & & & & 0 \end{pmatrix} V,$$

kde I_{r_j} je jednotková matice rozměru $r_j = \sum_{l=1}^{t_j} \dim J_{jl}$. Vidíme, že

$$\sum_{j=1}^{s} B_{j1} = I,$$

kde $I = I_N$.

Dále nechť

$$B_{jk} = V^{-1} \begin{pmatrix} 0 & & & & & \\ & . & & & & \\ & & . & & & \\ & & & \mathcal{J}_j - \lambda_j I_{r_j} & & \\ & & & & . & \\ & & & & & . \\ & & & & & & 0 \end{pmatrix} V, j = 1, ...s,\ k = 1, 2...$$

Snadno se přesvědčíme, že platí vztahy

$$B_{jk+1} = (T - \lambda_j I) B_{jk} = (T - \lambda_j I)^k B_{j1},$$

tedy,

$$B_{jk} = 0 \text{ pro } k > q_j.$$

Dále pak platí formule

$$B_{j1} B_{t1} = \delta_{jt} B_{j1},\ j, t = 1, ..., s$$

a

$$B_{jk} B_{tl} = \delta_{jt} B_{jk+l-1};$$

speciálně

$$B_{j1}^2 = B_{j1},\ j = 1, ..., s,$$

Podobně se lze přesvědčit, že

$$T = \sum_{j=1}^{s} (\lambda_j B_{j1} + B_{j2})$$

a tedy,

$$T^k = \sum_{j=1}^{s} (\lambda_j B_{j1} + B_{j2})^k$$

$$= \sum_{j=1}^{s} \sum_{t=1}^{q_j} \frac{f_k^{(t-1)}(\lambda_j)}{(t-1)!} B_{jt},$$

kde

$$f_k(\lambda) = \lambda^k,$$

takže

$$f_k'(\lambda) = k\lambda^{k-1},$$
$$f_k''(\lambda) = k(k-1)\lambda^{k-2},$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$f_k^{(t)}(\lambda) = k(k-1)...(k-t+1)\lambda^{k-t}.$$

Typickou pro konvergenci mocninné metody je následující situace.

Nechť spektrum $\sigma(T) = \{\lambda_1, ..., \lambda_s\}$ má strukturu danou vztahy

$$|\lambda_1| > |\lambda_j|,\ j > 1,\ q_1 = 1. \tag{6.2}$$

Potom

$$\left(\frac{1}{\lambda_1} T\right)^k = B_{11} + \sum_{j=2}^{s} \sum_{t=1}^{q_j} \frac{g_k^{(t-1)}(\lambda_j)}{(q_j-1)!} B_{jt},$$

kde

$$g_k(\lambda) = \left(\frac{\lambda}{\lambda_1}\right)^k.$$

Je okamžitě patrné, že

$$\lim_{k\to\infty} \left(\frac{1}{\lambda_1} T\right)^k = B_{11},$$

neboť

$$\lim_{k\to\infty} g_k^{(t)}(\lambda_j) = 0$$

pro $j = 2, ..., s$ a $t = 0, 1, ..., q_j - 1,\ j > 1$.

Věta 6.1 *Je - li $q_1 = 1$, pak posloupnost*

$$\left\{\left(\frac{1}{\lambda_1} T\right)^k\right\}$$

konverguje k B_{11} s rychlostí s jakou skalární posloupnost

$$\max\left\{\left|\frac{\lambda_j}{\lambda_1}\right|^k k^{q_j-1} : j = 2, ..., s\right\}$$

konverguje k nule.

Je - li $q_1 > 1$, pak platí

$$(q_1-1)! \lim_{k\to\infty} \frac{1}{\lambda_1^{k-q_1+1}} k^{-q_1+1} T^k = B_{1q_1}. \tag{6.3}$$

Předchozí teorie nabízí následující obecný algoritmus mocninné metody.

Algoritmus 6.1 *Buď $\{x'_k\}$ posloupnost lineárních funkcionálů na $\mathcal{C}^N$ takových, že*

$$x'_k(B_{11}x_0) \neq 0 \text{ a } \lim_{k\to\infty} x'_k = x'_\infty, \tag{6.4}$$

$$(B_{11})^T x'_\infty \neq 0. \tag{6.5}$$

Buď $\varepsilon > 0$ zadaná tolerance.

1. *Zvolme x_0.*
2. *Položme $k = 0$.*
3. *Sestrojme vektor x_k kladouce*
$$x_k = Tx_k.$$
4. *Určeme $\lambda^{(k)}$ pomocí předpisu*
$$\lambda^{(k)} = \frac{x'_k(x_k)}{x'_k(x_{k-1})}$$
a
$$x_{k+1} = \frac{1}{\lambda^{(k)}} x_k.$$
5. *Prověřme, zda*
$$\left|\lambda^{(k)} - \lambda^{(k-1)}\right| \begin{cases} \geq \varepsilon & \text{(NE)}, \\ < \varepsilon & \text{(ANO)} \end{cases} \tag{6.6}$$
5. *Nastane - li v (6.6) případ* NE, *položme $k+1 \to k$ a* GOTO *3.*
6. *Nastane - li v (6.6) případ* ANO , GOTO *7.*
7. *Položme $\lambda_1 = \lambda^{(k)}$ a* STOP.

Potřebné vlastnosti algoritmu 6.1 jsou obsahem následující věty, již formulujeme ve zjednodušené verzi

Věta 6.2 *Buď T matice typu $N \times N$, taková, že platí vztahy (6.2), (6.5) a $x'_k = x'$ pro $k = 0, 1, ...$ takže (6.5) platí automaticky.*

Potom platí rovnosti

$$\lim_{k\to\infty} x_k = x_\infty,\ Tx_\infty = \lambda_1 x_\infty, \tag{6.7}$$

a

$$\lim_{k\to\infty} \lambda^{(k)} = \lambda_1. \tag{6.8}$$

Důkaz. Snadno ověříme, že

$$x_{k+1} = \prod_{l=0}^{k} \frac{1}{\lambda^{(l)}} T^{k+1} x_0$$

a

$$\prod_{l=0}^{k} \lambda^{(l)} = \prod_{l=0}^{k} \frac{x'(Tx_l)}{x'_l(x)}$$

$$= \frac{x'(T^{k+1}x_0)}{x'(x_0)}$$

a tudíž

$$x_{k+1} = \frac{x'(x_0)}{x'(T^{k+1}x_0)} T^{k+1}x_0.$$

Na základě věty 6.1 odvodíme, že

$$\lim_{k\to\infty} x_k = x'(x_0) \lim_{k\to\infty} \frac{B_{1q_1}x_0 + Z_k x_0}{x'(B_{1q_1}x_0) + x'(Z_k)},$$

kde, vzhledem k (6.3)

$$\lim_{k\to\infty} Z_k x_0 = 0.$$

Posléze tedy,

$$\lim_{k\to\infty} x_k = \frac{1}{x'(B_{1q_1}x_0)} B_{1q_1}x_0.$$

Jako důsledek obdržíme vztahy

$$\lim_{k\to\infty} \lambda^{(k)} =$$

$$= \lim_{k\to\infty} \frac{x'(T^k x_0)}{x'(T^{k-1}x_0)}$$

$$= \lambda_1 \lim_{k\to\infty} \frac{x'(k^{-q_1+1}(\frac{1}{\lambda_1}T)^k x_0)}{x'(k^{-q_1+1}(\frac{1}{\lambda_1}T)^{k-1}x_0)} = \lambda_1.$$

Tím je důkaz věty 6.2 proveden. |||

6.2 Algoritmy LR a QR

Definice 6.2 *Buď A matice typu $N \times N$ nad tělesem reálných čísel.* **LR** rozkladem *matice A nazveme dvojici matic typu $N \times N$ takovou, že platí:*

(a) $L = (l_{jk})$ je dolní trojúhelníková s jednotkovou diagonálou, t.j.

$$l_{jk} = \begin{cases} 1 & \text{pro } j = k \\ 0 & \text{pro } j < k \end{cases} \quad j,k = 1,...N,$$

(b) $R = (r_{jk})$ je horní trojúhelníková, t.j.

$$r_{jk} = 0 \text{ pro } \; j > k, \; j,k = 1,...,N.$$

(c) $A = LR$.

Poznámka 6.2 *Postačující podmínkou pro existenci LR rozkladu $N \times N$ matice A je* plná regularita *matice A, t.j. všechny hlavní minory matice A jsou regulární.*

Algoritmus 6.2 *Buď A matice typu $N \times N$ a $\varepsilon > 0$ daná tolerance.*

(1) Položme $A = A_1$ a $k = 1$.

(2) Nechť

$$A_k = L_k R_k$$

je LR rozklad matice A_k.

(3) Dále položme

$$R_k L_k = A_{k+1}, \; R_k = (r_{jl}^{(k)}).$$

(4) Posuďme, zda

$$\max \left\{ \left| r_{jl}^{(k)} \right| : \; j,l = 1,...N, \; j > l \right\} \begin{cases} \geq \varepsilon & \text{NE} \\ < \varepsilon & \text{ANO.} \end{cases}$$

(5) Je -li v (4) NE, *položme $k+1 \to k$ a* GOTO *(2).*

(6) Je - li v (4) ANO, *položme*

$$R_\infty = R_{k+1}$$

a GOTO *(7).*

(7) STOP.

Tvrzení 6.2 *Nechť LR rozklady*

$$A_k = L_k R_k \tag{6.9}$$

existují pro $k = 1, 2, ...$

Potom platí

A_{k+1} *a* A_k *jsou podobné matice; jmenovitě*

$$A_{k+1} = L_k^{-1} A_k L_k, \tag{6.10}$$

$$A_{k+1} = (L_1 ... L_k)^{-1} A_1 (L_1 ... L_k). \tag{6.11}$$

Matice T_k*, kde*

$$T_k = L_1 ... L_k$$

je dolní trojúhelníková a matice U_k*,*

$$U_k = R_k ... R_1$$

je horní trojúhelníková, při čemž

$$A^k = A_1^k = T_k U_k. \tag{6.12}$$

Důkaz. Vztahy (6.10) a (6.11) jsou zřejmé. Pomocí nich pak odvodíme, že

$$L_1 ... L_k A_{k+1} = A_1 L_1 ... L_k.$$

Dále pak,

$$\begin{aligned} T_k U_k &= (L_1 ... \, L_{k-1}) \, (L_k R_k) \, (R_k ... R_1) \\ &= (L_1 ... L_{k-1}) \, A_k \, (R_{k-1} ... R_1) \\ &= A_1 \, (L_1 ... L_{k-1}) \, (R_{k-1} ... R_1) \\ &= A_1 \, T_{k-1} \, U_{k-1} \\ &= ..... \\ &= A_1^k = A^k. \end{aligned}$$

$$|||$$

Matice T_k a U_k tvoří LR rozklady pro A^k, neboť zřejmě

$$\operatorname{diag} T_k = I.$$

Naším cílem je ukázat, že

$$\lim_{k \to \infty} L_k = I \tag{6.13}$$

a

$$\lim_{k \to \infty} A_k = \lim_{k \to \infty} R_k = \begin{pmatrix} \lambda_1 & \times & . & . & . & \times \\ 0 & \lambda_2 & . & . & . & \times \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & \lambda_N \end{pmatrix}, \tag{6.14}$$

kde $\lambda_1, ..., \lambda_N$ jsou vlastní čísla.

Věta 6.3 *Nechť $A = A_1$ splňuje*

1^0. *LR rozklady matic R_k L_k existují pro $k = 1, 2, ...$*

2^0. *Nechť*

$$|\lambda_1| > |\lambda_2| > \ ... \ |\lambda_N| > 0. \tag{6.15}$$

3^0. *Předpokládejme, že pro matice X a $Y = X^{-1}$, kde $A = XDY$ je Jordanova forma matice A, existují LR rozklady*

$$X = L_X R_X, \ \ Y = L_Y R_Y, \ \mathrm{diag} L_X = \mathrm{diag} L_Y = I. \tag{6.16}$$

Potom platí (6.13) a (6.14).

Důkaz. Podle předpokladu existuje D^{-1} takže

$$A^k = XD^kY = L_X R_X D^k L_Y R_Y = L_X R_X \ (D^k L_Y D^{-k}) \ D^k R_Y.$$

Avšak

$$D^k \ L_Y \ D^{-k} = \left(l_{jt}^{(k)}\right)$$

je dolní trojúhelníková matice, při čemž

$$l_{jt}^{(k)} = \left(\frac{\lambda_j}{\lambda_t}\right)^k l_{jt}, \text{ kde } L_Y = (l_{jt} \text{ pro } j > t)$$

a

$$l_{jt}^{(k)} = \begin{cases} 1 & \text{pro} \quad j = t \\ 0 & \text{pro} \quad j < t \end{cases}$$

Protože

$$|\lambda_t| > |\lambda_j| \text{ pro } j > t,$$

odvodíme, že

$$\lim_{k\to\infty} l_{jt}^{(k)} = \lim_{k\to\infty} \left(\frac{\lambda_j}{\lambda_t}\right)^k = 0 \text{ pro } j > t.$$

Tudíž

$$D^{-k} \ L_Y \ D^k = I + E_k,$$

při čemž

$$\lim_{k\to\infty} E_k = 0.$$

Z předchozího plyne, že

$$\begin{aligned} A^k &= L_X R_X \ (I + E_k) \ D^k R_Y \\ &= L_X \ (I + R_X E_k R_X^{-1}) \ R_X D^k R_Y \\ &= L_X \ (I + F_k) \ R_X D^k R_Y, \end{aligned}$$

kde

$$F_k = R_X E_k R_X^{-1}.$$

Protože $E_k \to 0$, pro všechna dostatečně velká k existují LR rozklady

$$I + F_k = \tilde{L}_k \tilde{R}_k, \ \mathrm{diag} \tilde{L}_k = I, \ \tilde{l}_{jt} = 0 \text{ pro } j < t.$$

Díky tomu, že $E_k \to 0$, též $F_k \to 0$, odvodíme snadno, že

$$\lim_{k\to\infty} \tilde{L}_k = I = \lim_{k\to\infty} \tilde{R}_k.$$

Posléze,

$$A^k = (L_X \tilde{L}_k)\,(\tilde{R}_k R_X D^k R_Y).$$

Z jednoznačnosti trojúhelníkových rozkladů plyne, že

$$T_k = L_1 ... L_k = L_X \tilde{L}_k$$

a

$$U_k = R_k ... R_1 = \tilde{R}_k R_X D^k R_Y.$$

Protože $\tilde{L}_k \to I,\ \tilde{R}_k \to I$, obdržíme, že platí vztahy

$$\lim_{k\to\infty} T_k = L_X,$$

$$\lim_{k\to\infty} L_k = \lim_{k\to\infty} T_{k-1}^{-1} T_k = I,$$

$$\begin{aligned}
\lim_{k\to\infty} R_k &= \lim_{k\to\infty} U_k U_{k-1}^{-1} \\
&= \lim_{k\to\infty} \tilde{R}_k R_X D^k R_Y R_Y^{-1} D^{(-k+1)} R_X^{-1} \tilde{R}_k^{-1} \\
&= R_X D R_X^{-1}.
\end{aligned}$$

Avšak

$$R_X D R_X^{-1} = \begin{pmatrix} \lambda_1 & \times & & \times & \times \\ 0 & \lambda_2 & & \times & \times \\ & & \cdot & & \\ & & & \cdot & \\ & & & \lambda_{N-1} & \times \\ 0 & 0 & & 0 & \lambda_N \end{pmatrix}$$

a tím je důkaz věty 6.3 proveden. □

7 Počáteční a okrajové problémy pro obyčejné diferenciální rovnice

7.1 Základní poznatky z teorie obyčejných diferenciálních rovnic

V této kapitole budeme vyšetřovati metod přibližného řešení rovnic typu

$$y' = f(x, y),\ x \in [a, b], \tag{7.1}$$

$$y(a) = y_0, \tag{7.2}$$

kde $-\infty < a < b < +\infty$ a soustav takových rovnic.

$$\left\{ \begin{array}{l} y_1' = f_1(x, y_1, ..., y_N) \\ ..... \\ y_N' = f_N(x, y_1, ..., y_N) \end{array} \right. , \ x \in [a, b] \tag{7.3}$$

a

$$y_j(0) = y_{j0}, \ j = 1, ..., N. \tag{7.4}$$

Poznamenejme, že problém (7.3) - (7.4) zahrnuje jako speciální úlohu

$$z^N = f(x, z, z', ..., z^{N-1}), \ x \in [a, b], \tag{7.5}$$

s podmínkami

$$z^{j-1}(a) = z_{j0}, \ j = 1, ..., N-1. \tag{7.6}$$

Stačí položit

$$z = y_1, \ z' = y_2, \ ..., \ z^{N-1} = y_N,$$

takže

$$y_1' = y_2, \ y_2' = y_3, \ ..., \ y_{N-2}' = y_{N-1}$$

a

$$y_N' = z^{(N)} = f(x, z, z', ..., z^{N-1}) = f(x, y_1, ..., y_N).$$

Kromě úloh typu (7.1) - (7.2) a (7.3) - (7.4) budeme vyšetřovati též úlohy typu

$$y' = f(x, y),$$

avšak místo počáteční podmínky (7.2) budeme požadovati, aby platila rovnost

$$r(y(a), y(b)) = 0, \tag{7.7}$$

kde r je vhodná funkce dvou proměnných.

Obecněji budeme vyšetřovat úlohy

$$y_j' = f_j(x, y_1, ..., y_N), \ j = 1, ..., N,$$

s podmínkami

$$\mathbf{0} = \mathbf{r}(\mathbf{y}(a), \mathbf{y}(b)), \tag{7.8}$$

kde

$$\mathbf{r} = \mathbf{r}(\mathbf{u}, \mathbf{v}) = \left[\begin{array}{c} r_1(u_1, ..., u_N; v_1, ..., v_N) \\ ..... \\ r_N(u_1, ..., u_N; v_1, ..., v_N) \end{array} \right]. \tag{7.9}$$

Úlohy typu (7.1) a (7.7), respektive (7.3) a (7.8) se nazývají *úlohami okrajovými* na rozdíl od úloh *počátečního* typu (7.1) a (7.2) respektive (7.3) a (7.4).

Numerické metody přibližného řešení výše uvedených úloh je založeno na vlastnostech řešení těchto úloh. Zformulujme některá tvrzení potřebná pro další vyšetřování.

Věta 7.1 *Předpokládejme, že komponenty $f_1, ..., f_N$ vektoru $\mathbf{f}$ jsou definovány a jsou spojité na oblasti*

$$\mathcal{S} = \left\{(x, y) : \ a \leq x \leq b, \ \mathbf{y} \in R^N\right\}.$$

Dále nechť existuje lipschitzovská konstanta L tak; že

$$||\mathbf{f}(x, \bar{\mathbf{y}}) - \mathbf{f}(x, \tilde{\mathbf{y}})|| \ \leq L||\bar{\mathbf{y}} - \tilde{\mathbf{y}}||, \tag{7.10}$$

*platí pro všechna $x \in [a, b]$ a $\mathbf{y} \in R^N$ (*Lipschitzova podmínka*).*

Potom ke každému $x_0 \in [a, b]$ a $\mathbf{y}^0 \in R^N$ existuje právě jedna vektor - funkce $\mathbf{y}(x)$ tak, že

1) $\mathbf{y}$ je spojitá a spojitě diferencovatelná vektor - funkce v $[a, b]$;

2) $\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \ x \in [a, b]$;

3) $\mathbf{y}(x_0) = \mathbf{y}^0$.

Poznámka. Již víme, vze Lipschitzova podmínka je splněna, pakliže parciální derivace $\frac{\partial f_j}{\partial x}, \ j = 1, ..., N$ existují a jsou spojité na $\mathcal{S}$.

V praxi se můžeme setkat převážně s případy, kdy $f_j, \ j = 1, ..., N$, jsou spojité na $\mathcal{S}$, avšak derivace $\frac{\partial f_j}{\partial x}$ mohou být na $\mathcal{S}$ neohraničené. V takovém případě má počáteční úloha (7.3) - (7.4) řešení, ale to nemusí být definováno na celé oblasti.

Příklad 7.1 *Nechť $N = 1$ a*

$$y' = y^2, \ y(0) = 1, \ 0 \leq x < +\infty.$$

Vidíme, že

$$y(x) = \frac{1}{1 - x}$$

a to je ohraničená funkce jen pro $x < 1$.

Pro numerické počítání je důležité, že řešení závisejí na počátečních datech spojitě. Tvrzení na toto téma je obsahem následující věty.

Věta 7.2 *Za předpokladů věty 7.1 platí odhady*

$$||\mathbf{y}(x, \mathbf{a}^1) - \mathbf{y}(x, \mathbf{a}^2)|| \ \leq e^{L(x-a)}||\mathbf{a}^1 - \mathbf{a}^2||, \tag{7.11}$$

kde

$$\mathbf{y}'(x, \mathbf{a}^j) = \mathbf{f}(x, \mathbf{y}), \ \mathbf{y}(a) = \mathbf{a}^j, \ j = 1, 2. \tag{7.12}$$

Významné pro teorii i praxi jsou lineární úlohy typu

$$\mathbf{Y}' = T(x)\mathbf{Y}, \ \mathbf{Y}(a) = \mathbf{Y}^a, \tag{7.13}$$

přičemž $T = T(x)$ je matice typu $N \times N$. *Růst* řešení je charakterizován v následující větě.

Věta 7.3 *Nechť prvky matice $T(x)$ jsou spojité funkce na $[a,b]$ a nechť*

$$\kappa(x) = ||T(x)||. \tag{7.14}$$

Potom platí odhad

$$||\mathbf{Y}(x) - I|| \ \leq \ exp\left\{\int_a^x \kappa(t)\,dt\right\} - 1 \ pro\ x \geq a. \tag{7.15}$$

7.2 Racionální aproximace exponenciální funkce

Jednou z řady ekvivalentních definic exponenciální funkce může posloužit formule daná Taylorovou řadou. Pro všechna komplexní čísla je platný rozvoj

$$\exp\{z\} \equiv e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!},\ z \in \mathcal{C}^1. \tag{7.16}$$

Snadno si uvědomíme, že platí

$$|e^z| \leq 1 \text{ pro } z \in \mathcal{C}^1, \Re z \leq 0 \tag{7.17}$$

a dále, že

$$\lim_{z\to\infty} e^{|z|} = 0. \tag{7.18}$$

Dále nechť $P = P(z)$ a $Q = Q(z)$ označují polynomy s komplexními koeficienty. Bude-li to vhodné, budeme indexy vyznačovat odpovídající stupeň. Tedy, na př.

$$P_0(z) = 1,\ P_1(z) = 2z + 1/2,\ Q_1(z) = 1 - z,$$

a tak podobně.

Buď $R = R(z)$ racionální funkce daná polynomy P a Q:

$$R(z) = \frac{P(zP}{Q(z)},\ z \in \mathcal{C}^1. \tag{7.19}$$

V následujícím textu budeme předpokládat, že polynomy v (7.19) jsou nesoudělné, tedy, že neexistují polynomy $\tilde{P}, \tilde{Q}$, při čemž stupeň polynomu $\tilde{Q}$ je nižší než stupeň polynomu Q, takové, že

$$R(z) = \frac{\tilde{P(z)}}{\tilde{Q(z)}},\ z \in \mathcal{C}^1.$$

Definice 7.1 *Řekneme, že racionální funkce R je* Padéovou aproximací exponenciální funkce řádu *p, pakliže platí vztahy*

$$e^z - R(z) = g(z)z^{p+1} = 0(z^{p+1}),\ z \in \Omega_0, \tag{7.20}$$

kde Ω_0 je nějaké okolí bodu $0 \in \mathcal{C}^1$ a přitom

$$0 \neq \kappa = \sup\{|g(z)| : z \in \Omega_0\} < +\infty.$$

Příklady.

$$R_{1,0}(z) = \frac{P_1(z)}{Q_0(z)} = 1 + z,\ P_1(z) = z,\ Q_0(z) = 1, \tag{7.21}$$

$$R_{0,1}(z) = \frac{P_0(z)}{Q_1(z)} = \frac{P_0(z)}{Q_1(z)} = \frac{1}{1-z},\ P_0(z) = 1, Q_1(z) = 1 - z, \tag{7.22}$$

$$R_{1,1}(z) = \frac{P_1(z)}{Q_1(z)} = \frac{2+z}{2-z},\ P_1(z) = 2 + z,\ Q_1(z) = 2 - z. \tag{7.23}$$

Snadno zjistíme, že

$$R_{0,1}(z) = 1 + z + z^2 + ...,$$

a

$$R_{1,1}(z) = 1 + z + \frac{1}{2}z^2 + ...$$

a tedy,

$$|e^z - R_{1,0}(z)| = 0(z^2),\ p = 1, \tag{7.24}$$

$$|e^z - R_{0,1}(z)| = 0(z^2),\ p = 1, \tag{7.25}$$

$$|e^z - R_{1,1}(z)| = 0(z^3),\ p = 2. \tag{7.26}$$

Definice 7.2 *Racionální aproximace R exponenciální funkce se nazývá* A−přijatelnou*, platí-li vztahy*

$$|R(z)| \leq 1 \text{ pro } \Re z \leq 0. \tag{7.27}$$

Příklady.

Snadno ověříme, že $R_{1,0}$ v (7.21) není A-přijatelná. Na druhé straně však, $R_{0,1}$ v (7.22) a $R_{1,1}$ v (7.23) A-přijatelné jsou a navíc pro ně platí

$$\lim_{z\to\infty} |R_{0,1}(z)| = 0 \tag{7.28}$$

a

$$\lim_{z\to\infty} |R_{1,1}(z)| = 1. \tag{7.29}$$

Definice 7.3 *Racionální aproximace $R_{1,0}, R_{0,1}$ a $R_{1,1}$ v (7.21)-(7.23) nesou názvy svých původních autorů. Aproximace $R_{1,0}$ se tedy nazývá* Eulerovou přímou aproximací*, $R_{0,1}$* Eulerovou zpětnou aproximací *a $R_{1,1}$* aproximací Cranka-Nicholsonové*.*

Test 7.1 *Buď λ takové, že $\Re\lambda < 0$ je v absolutní hodnotě dostatečně velké číslo. Jest tedy číslo $|e^\lambda|$ velmi malé. Ptáme se, jak tuto skutečnost zprostředkovávají jednotlivé racionální aproximace exponenciální funkce.*

Protože v tomto případě,

$$\lim_{\Re\lambda\to-\infty} |e^\lambda| = \lim_{\Re\lambda\to-\infty} e^{\Re\lambda} = 0,$$

1^0 *Eulerova dopředná aproximace dává*

$$\begin{aligned}\left|e^\lambda - R_{1,0}(\lambda)\right| &= \left|e^\lambda - 1 - \lambda\right| \\ &\geq |\lambda| - 1 - e^{\Re\lambda}.\end{aligned}$$

Vidíme tedy, že tato aproximace poskytuje libovolně velkou chybu.

Naproti tomu,

2^0 *pro zpětnou Eulerovu metodu zjistíme, že platí*

$$\begin{aligned}\left|e^\lambda - R_{0,1}(\lambda)\right| &= \left|e^\lambda - \tfrac{1}{1-\lambda}\right| \\ &\leq \textstyle\sum_{k=2}^{\infty} \frac{z^k}{k!} - \sum_{k=2}^{\infty} z^k \\ &= O(|z|^2)\end{aligned}$$

takže chyba aproximace je pro $\Re\lambda < 0$ malá.

3^0 *Ještě uspokojivější se jeví situace pro aproximaci Crankovu-Nicholsonové,*

$$\left|e^\lambda - R_{1,1}(\lambda)\right| = \left|e^\lambda - \frac{2+\lambda}{2-\lambda}\right| = O(|z|^3).$$

7.3 Počáteční úlohy

Jako motivaci vyšetřujme skalární úlohu

$$y' = f(x,y),\ y(x_0) = y_0,\ x_0,\ x \in [0,1],\ y \in R^1. \tag{7.30}$$

Buď $h > 0$. Nahraďme v (7.30) derivaci diferencí čímž obdržíme

$$\frac{y(x+h) - y(x)}{h} = f(x, y(x)),$$

neboli

$$y(x+h) = y(x) + hf(x, y(x)). \tag{7.31}$$

Obdrželi jsme tak algoritmus pro řešení.

Algoritmus 7.1 *Buď M celé kladné. Zvolme*

$$h = \frac{x - x_0}{M}$$

a počítejme rekurzivně pomocí formulí

$$\eta_0 = y_0,$$

$$\eta_{k+1} = \eta_k + hf(x_k, \eta_k),$$

kde

$$x_{k+1} = x_k + h,\ k = 0,\ 1, ...,\ M - 1.$$

Algoritmus 7.1 popisuje t. zv. *Eulerovu polygonální metodu.* Všímněme si, že v Algoritmu 7.1 platí vztahy

$$\eta(x_0; h) = y_0,$$

a

$$\eta(x_0 + h; h) = \eta(x; h) + hf(x, \eta(x; h).$$

To navádí na myšlenku definovati algortimus, v němž roli $f(x; \eta(x; h))$ převezme vhodná funkce $\Phi = \Phi(x, y; h, f)$.

Algoritmus 7.2 *Volme $h > 0$ a nechť*

$$\begin{cases} \eta_0 = y_0 \\ \eta_{k+1} = \eta_k + h\Phi(x_k, \eta_k; h; f) \text{ pro } k = 1, 2, ... \end{cases} \tag{7.32}$$

kde

$$x_{k+1} = x_k + h.$$

Již víme, že pro Eulerovu polygonální metodu jest

$$\Phi(x, y, h, f) = f(x, y),$$

Φ tedy nezávisí na h.

Pro další účely označme pro libovolné $x \in [a, b]$ a $\mathbf{y} \in R^N$ symbolem $\mathbf{z} = \mathbf{z}(\mathrm{x})$ přesné řešení úlohy

$$\mathbf{z}'(t) = \mathbf{f}(x, \mathbf{z}(t)),\ \mathbf{z}(x) = \mathbf{y}(x). \tag{7.33}$$

Definujme funkci Δ předpisem

$$\Delta(x, y; h; \mathbf{f}) = \begin{cases} \frac{\mathbf{z}(x+h)-\mathbf{y}}{h} & \text{pro } h > 0 \\ \mathbf{f}(x, \mathbf{y}) & \text{pro } h = 0. \end{cases} \tag{7.34}$$

V dalším textu nebudeme vyznačovati f v argumentu funkce Δ a Φ.

Pomocí funkce Δ definujeme t. zv. *lokální diskretizační chybu*, tedy veličinu

$$\tau(x, \mathbf{y}, h) = \Delta(x, \mathbf{y}; h) - \Phi(x, \mathbf{y}; h). \tag{7.35}$$

V algoritmu 7.2 zavedené metody se nazývají *jednokrokové*. To evidentně proto, že k získání hodnoty přiblížení v bodě x_k se vyžaduje znalost hodnoty v příslušném bodě předchozím x_{k-1}.

Jednokroková metoda se ukazuje jako vhodná, jestliže

$$\lim_{h\to 0} \tau(x, \mathbf{y}; h) = 0, \tag{7.36}$$

což, díky tomu, že

$$\lim_{h\to 0} \mathbf{\Delta}(x, \mathbf{y}, h) = \mathbf{f}(x, \mathbf{y}), \tag{7.37}$$

vede k podmínce

$$\mathbf{\Phi}(x, \mathbf{y}, 0) = \mathbf{f}(x, \mathbf{y}). \tag{7.38}$$

Říkáme, že jednokroková metoda je *konzistentní*, jestliže platí relace (7.38).

Poznámka. Eulerova polygonální metoda je zřejmě konzistentní.

Příklad 7.2 *Předpokládejme, že $N = 1$ a f má vyšší hladkost než je pouhá spojitost, řekněme, že $\frac{\partial y}{\partial x}$ jsou spojité na $[a, b]$. Navíc nechť*

$$z(x+h) = z(x) + hz'(x) + \frac{1}{2}h^2 z''(x) + ... + \frac{h^p}{p!} z^{(p)}(x+\theta h),\ 0 < \theta < 1.$$

Protože

$$\begin{aligned} z''(x) &= \frac{d}{dt} f(t, z(t))\,|_{t=x} \\ &= f_x(t, z(t))\,|_{t=x} + f_y(t, z(t))\,|_{t=x} \\ &= f_x(x, y) + f_y(x, y) f(x, y), \end{aligned}$$

a

$$z'''(x) = f_{xx}(x, y) + 2f_{xy}(x, y)[f(x, y)]^2 + f_y(x, y) z''(x),$$

zjistíme, že

$$\left\{ \begin{array}{ll} \Delta(x, y; h) & = z'(x) + \frac{h}{2} z''(x) + ... + \frac{h^p}{p!} z^{(p)}(x+\theta h) \\ & = f(x, y) + \frac{h}{2} \left[f_x(, y) + f_y(x, y) f(x, y) \right] \\ & \quad + ... \end{array} \right. \tag{7.39}$$

Na př. pro Eulerovu metodu

$$\begin{aligned} \tau(x, y; h) &= \Delta(x, y; h) - \Phi(x, y; h) \\ &= \frac{h}{2} \left[f_x(x, y) + f_y(x, y) f(x, y) \right] + O(h). \end{aligned}$$

Obecně, říkáme, že *jednokroková metoda je řádu p*, jestliže platí

$$\tau(x, \mathbf{y}; h) = O(h^p) \tag{7.40}$$

pro $x \in [a,b]$, $\mathbf{y} \in R^N$ a pro všechny $\mathbf{f} \in [C^{(p)}([a,b])]^N$.

Z předchozího příkladu odvodíme poměrně obecný způsob konstrukce metod vyšších řádů.

Na př.

$$\Phi(x,y;h) = f(x,y) + \frac{h}{2}\left[f_x(x,y) + f_y(x,y)f(x,y)\right] \tag{7.41}$$

definuje metodu 2. řádu.

Na tomto místě je nutné uvézt na pravou míru skutečnost, že metody vysokých řádů (> 2) jsou v praxi víceméně bezcenné, neúčinné, složité. To je způsobeno tím, že výpočet vyšších derivací je problém složitý v tom smyslu, že je *špatně podmíněný* (*non well posed*). Připomeňme, že to značí, že hodnota výsledku výpočet není spojitě závislá na datech.

Uveďme některé příklady metod vyšších řádů.

Položme

$$\Phi(x,y;h) = a_1 f(x,y) + a_2 f(x + p_1 h, y + p_2 hy)), \tag{7.42}$$

kde $a_1,\ a_2,\ p_1,\ p_2$ jsou vhodné konstanty, které jest určit tak, aby platily vztahy určující řád metody.

Pro Φ z (7.42) platí

$$\Phi(x,y;h) = (a_1 + a_2)f(x,y) + a_1 h\left[p_1 f_x(x,y) + p_2 f_y(x,y)f(x,y)\right] + O(h^p),$$

odkud vyplývá, splnění podmínek

$$a_1 + a_2 = 1,\ a_2 p_1 = 1/2,\ a_2 p_2 = 1/2, \tag{7.43}$$

je postačující aby metoda byla 2. řádu.

Speciální volbou

$$a_1 = a_2 = \frac{1}{2},\ p_1 = p_2 = 1$$

obdržíme *metodu Heunovu*

$$\Phi(x,y;h) = \frac{1}{2}\left[f(x,y) + f(x+h, y + hf(x,y))\right]. \tag{7.44}$$

Je patrné, že tato metoda vyžaduje výpočet dvou funkčních hodnot na jeden krok metody a to je cenou, již je nutno zaplatit za vyšší řád.

Jinou metodu obdržíme volbou

$$a_1 = 0,\ a_2 = 1,\ p_1 = p_2 = 1/2.$$

Ta je dána *schématem Collatzovým*

$$\Phi(x,y;h) = f\left(x+\frac{h}{2}, y+\frac{h}{2}f(x,y)\right). \tag{7.45}$$

Toto schéma je opět řádu 2. a vyžaduje výpočet dvou funkčních hodnot na jeden krok metody.

Další metody vyšší řádů jsou spojeny se jmény *Runge* a *Kutta*.

Hledejme jenokrokovou metodu pomocí schématu

$$\Phi(x,y;h) = \frac{1}{6}\left[k_1+2k_2+2k_3+k_4\right], \tag{7.46}$$

kde

$$\begin{aligned} k_1 &= f(x,y), \\ k_2 &= f\left(x+\tfrac{h}{2},\ y+\tfrac{h}{2}k_1\right), \\ k_3 &= f\left(x+\tfrac{h}{2},\ y+\tfrac{1}{2}hk_2\right), \\ k_4 &= f\left(x+h,\ y+hk_3\right). \end{aligned}$$

Dá se ukázati, že

$$\Phi(x,y;h) - \Delta(x,y;h) = O(h^4).$$

Znamená to, že tato, dá se říci, základní metoda *Runge - Kuttovy* třídy, je řádu 4.

Aplikujme některé jednokrokové metody na případ úlohy

$$y'(x) = f(x),\ y(x_0) = y_0.$$

Přesné řešení této úlohy je dáno výrazem

$$y(x) = \int_{x_0}^{x} f(t)\mathrm{dt} + y_0.$$

Snadno nahlédneme, že v problematice numerické kvadratury, což je vlastně námi vyšetřovaná úloha, Eulerova metoda vede na známou metodu obdélníkovou zatímco Heunova metoda na metodu lichoběžníkovou.

Runge Kuttova metoda poskytuje metodu aproximace integrálu, jež je známá pod názvem metoda Simpsonova, kterážto metoda integruje polynomy až do řádu 3. stupně včetně.

Otázku konvergence jednokrokových metod zodpovídá následující věta. Soustředíme se na skalární případ s tím, že odpovídající věty mají takřka shodné formulace i důkazy.

Věta 7.4 *Nechť $f \in C([a,b])$, $x_0 \in [a,b]$ a $y_0 \in R^1$. Buď $y = y(x)$ řešení počáteční úlohy*

$$y' = f(x,y),\ y(x_0) = y_0.$$

Buď Φ funkce spojitá na

$$\mathcal{G} = \{(x,y,h):\ a \le x \le b,\ |y-y(x)| \le \gamma, 0 \le |h| \le h_0, , h_0 > 0,\ \gamma > 0\}$$

a nechť existují konstanty $\kappa_1,\ \kappa_2$ *tak, že*

$$|\Phi(x, y_1; h) - \Phi(x, y_2; h)| \leq \kappa_1 |y_1 - y_2|$$

a

$$|\tau(x, y(x); h)| = |\Delta(x, y(x); h) - \Phi(x, y(x); h)| \leq \kappa_2 |h|^p,\ p > 0,$$

platí pro všechna $x \in [a, b],\ |h| \leq h_0$.

Potom existuje $\hat{h},\ 0 < \hat{h} \leq h_0$ *tak, že pro* globální diskretizační chybu e, *kde*

$$e = e(x) = \eta(x; h) - y(x)$$

platí odhad

$$|e(x; h_k)| \leq |h_k|^p\, \kappa_2 \frac{\exp\{\kappa_1 |x - x_0|\} - 1}{\kappa_1}$$

pro všechna $x \in [a, b]$ *a všechna*

$$h_k = \frac{x - x_0}{k},\ k = 1, 2, ...,$$

pro něž $|h_k| \leq \hat{h}$. *Pro* $\gamma = \infty$ *je* $\hat{h} = h_0$.

Obecnější metody pro řešení počátečních úloh lze shrnout pod obecné schéma

$$\eta_{k+r} + a_{r-1}\eta_{k+r-1} + ... + a_0\eta_k = hF(x_k, \eta_{k+r}, \eta_{k+r-1}, ..., \eta_k; h; f). \tag{7.47}$$

Zde $F = F(x, u_1, ...u_{r+1}; h; f)$ je vhodná funkce a $a_0, ...a_{r-1}$ jsou parametry metody. Je zřejmé, že k nastartování metod typu (7.47) je nutné znát r počátečních hodnot $\eta_0, ..., \eta_{r-1}$.

Teorie *vícekrokových* a obecněji, metod typu (7.47), byť přirozeně složitější než ta, již jsme naznačili pro metody jednokrokové, je do šíře i do hloubky propracována a dovedena do realizace v podobě softwarových balíků.

7.4 Okrajové úlohy

Elementy matematické analýzy slabých řešení.

V tomto článku se omezíme na lineární okrajové úlohy s poznámkou, že nelineární úlohy lze v principu vyšetřovat a řešit podobným způsobem jako v problematice lineární; rozdíl spočívá pouze v tom, že výsledné diskrétní soustavy jsou nelineární.

Významným nástrojem vyšetřování okrajových úloh je pojem *slabé (variační) formulace* okrajových úloh.

Výklad budeme provádět na příkladě typické, byť akademické, úlohy

$$-y'' = f \text{ v } \Omega = [a, b], \tag{7.48}$$

$$\mu_1 y'(a) + \mu_0 y(a) = \alpha \text{ a } \nu_1 y'(b) + \nu_0 y(b) = \beta. \tag{7.49}$$

O funkci f předpokládáme, že patří do $L^2(a,b)$; prakticky však vždy je f po částech spojitá, t. j. f má v $[a,b]$ konečný počet bodů nespojitiosti.

Formálně vyžaduje t. zv. *variační přístup* pojem zobecněné derivace.

Funkce $g \in L^2(a,b)$, t. j. g, pro níž

$$\int_a^b |g(t)|^2 dt < +\infty,$$

má v $[a,b]$ zobecněnou derivaci, pakliže platí, že existuje $h \in L^2(a,b)$ taková, že

$$\int_a^b g(t)v'(t)dt = -\int_a^b h(t)v(t)dt + g(b)v(b) - g(a)v(a). \tag{7.50}$$

Pak říkáme, že h je *zobecněnou derivací (1. řádu)* funkce g.

Dá se ukázati, že v našem případě, kdy vyšetřujeme oblast $\Omega \subset \mathcal{R}^1$, funkce g má v $[a,b]$ zobecněnou derivaci 1. řádu právě když (*klasická*) derivace g' existuje v $[a,b]$ s vyjímkou množiny $E \subset [a,b]$ mající Lebesgueovu míru $m(E) = 0$. Je známo, že spojitá funkce g má v $[a,b]$ zobecněnou derivaci h právě když g je *absolutně spojitá* v $[a,b]$. Množina funkcí mající zobecněné derivace 1. řádu integrovatelné se čtvercem na Ω se značí symbolem $W_2^{(1)}(\Omega)$ a nazývá se *Sobolevovým prostorem*. V tomto prostoru lze zavézti skalární součin

$$(u,v)_{W_2^1(\Omega)} = (u',v')_2 + (u,v)_2, \tag{7.51}$$

kde pro $u,\ v\ \in L^2(\Omega)$,

$$(u,v)_2 = \int_\Omega u(t)v(t)dt$$

a u' a v' značí zobecněné derivace funkcí u a v. Předpokládá se, že $u',\ v' \in L^2(\Omega)$.

Předpokládejme, že $w \in L^2(\Omega)$ je funkce splňující okrajové podmínky

$$\mu_0 w(a) + \mu_1 w'(a) = \alpha \text{ a } \nu_0 w(b) + \nu_1 w'(b) = \beta. \tag{7.52}$$

Dá se ukázati, že existuje hustá ve smyslu normy prostoru $L^2(\Omega)$ množina tvořená dostatečně hladkými funkcemi splňujícími (7.52); na př. množina všech polynomů splňujících (7.52).

Položme v (7.48)

$$y(s) = u(s) - w(s).$$

Jest tedy,

$$-y'' = w'' + f \text{ v } \Omega$$

a

$$\mu_1 y'(a) + \mu_0 y(a) = \mu_0 \left(u(a) - w(a)\right) + \mu_1 \left(u'(a) - w'(a)\right) = 0,$$

$$\nu_1 y'(b) + \nu_0 y(b) = \nu_1 \left(u'(b) - w'(b)\right) + \nu_0 \left(u(b) - w(b)\right) = 0.$$

Známe-li y, známe automaticky i hlednaé u.

K určení y máme tedy vztahy

$$-y'' = f + w'' \text{ v } \Omega \tag{7.53}$$

$$\mu_0 y(a) + \mu_1 y'(a) = \nu_0 y(b) + \nu_1 y'(b) = 0. \tag{7.54}$$

Protožee w'' je známá funkce, lze (7.53) psát ve tvaru

$$-y'' = F \text{ v } \Omega, \tag{7.55}$$

při čemž F je daná funkce.

Všímněme si, že ve formulaci (7.48) - (7.49), neboli (7.54) - (7.55), vyžadujeme splnění těchto vztahů identicky v $[a, b]$. To však vyžaduje znalost F ve *všech bodech* oblasti Ω a obecněji, znalost všech dat v Ω. Toto je jednak ne příliš praktické a jednak dokonce ne vždy splnitelné. Ve vícerozměrných oblastech zadání okrajových podmínek na některých částech hranice může být nedefinovatelné. Na př. nechť $\Omega = [0, 1] \times [0, 1]$ a nechť $u(s, 0) = 0$ pro $s \in (0, 1)$, zatímco $u(0, y) = 1$ pro $y \in (0, 1)$. Otázkou je, jakou hodnotu může nabývat u v bodě $(1, 0)$, t. j. , čemu je rovno $u(1, 0)$. Uvědomme si, že u má být spojitá na Ω.

V takových problémech se ukáže jako velice vhodná *slabá* (*variační*) formulace problému.

Označme symbolem V podprostor Sobolevova prostoru $W_2^1(\Omega)$ splňujících okrajové podmínky (7.54).

Vynásobme rovnici (7.55) funkcí $v \in V$ a integrujme podél Ω. Obdržíme vztahy

$$-\int_a^b y''(t)v(t)\mathrm{dt} = \int_a^b F(t)v(t)\mathrm{dt},\ v \in V,$$

takže požadované řešení lze hledat pomocí vztahů

$$\int_a^b y'(t)v'(t)\mathrm{dt} = \int_a^b F(t)v(t)\mathrm{dt},\ v \in V, \tag{7.56}$$

což plyne z (7.56) protože v splňuje tytéž okrajové podmínky (7.54) jako hledané řešení y (viz Cvičení 7.3).

Příklad 7.3 Buď $\mathcal{L}$ množina všech funkcí y majících spojité derivace 2. řádu na $[a, b]$ a splňujících okrajové podmínky (7.54). Ukažte, že rovnosti

$$\int_a^b y''(t)v(t)\mathrm{dt} = \int_a^b u(t)v''(t)\mathrm{dt}$$

platí pro všechny prvky $y \in \mathcal{L}$, přičemž v je funkce mající na $[a, b]$ spojitou derivaci 2. řádu, právě když $v \in \mathcal{L}$.

Pro jednoduchost položme

$$a = 0 \text{ a }, b = 1$$

a

$$\mu_1 = \nu_1 = 0.$$

V tomto případě se okrajové podmínky (7.54) redukují na

$$y(0) = y(1) = 0 \tag{7.57}$$

a tedy též

$$v(0) = v(1) = 0. \tag{7.58}$$

Formulujme nyní naši okrajovou úlohu *variačně* neboli *slabě*, či přesněji řečeno, *ve slabém smyslu.*

Úloha (U). *Nalézt* $y \in \tilde{W}_2^{(1)}(0,1)$ *tak, aby vztahy*

$$\int_0^1 y'(t)v'(t)\mathrm{dt} = \int_0^1 F(t)v(t)\mathrm{dt} \tag{7.59}$$

platily pro všechny prvky $v \in \tilde{W}_2^{(1)}(0,1)$, *kde* $\tilde{W}_2^{(1)}(0,1)$ *značí uzávěr prostoru* $\mathcal{C}_0^\infty([0,1])$ *funkcí s kompaktním nosičem na* $[0,1]$ *majících na* $[0,1]$ *spojité derivace všech řádů. Uzávěr se míní ve smyslu metriky prostoru* $W_2^{(1)}(0,1)$.

Dá se ukázat, že úloha **U** má právě jedno řešení. Má-li okrajová úloha

$$y'' = F \text{ v } [0,1]$$

s podmínkami (7.58) klasické řešení, pak toto klasické řešení je totožné s řešením úlohy **U**. Tato tvrzení jsou založena na platnosti slavného *Lax - Milgramova lemmatu*

Lemma 7.1 *(Lax − Milgram) Buď* V *Hilbertův prostor. Buď* $B = B(u,v)$ *bilineární forma, jež je spojitá a* $V-$*eliptická, t.j. existují konstanty* $c > 0$ *a* $\tilde{c} > 0$ *tak, že*

$$|B(u,v)| \le c||u||\,||v||, \quad u,v \in V \tag{7.60}$$

a

$$|B(u,u)| \ge \tilde{c}||u||^2, \; u \in V. \tag{7.61}$$

Potom pro každý prvek $f \in V$ *existuje právě jeden prvek* $u^* \in V$ *takový, že*

$$B(u^*,v) = (F,v), \quad \forall v \in V. \tag{7.62}$$

Platí přitom

$$||u^*|| \le \kappa||F||, \tag{7.63}$$

kde κ *nezávisí na* F.

V našem případě $V = \tilde{W}_2^{(1)}(0,1)$ a

$$B(u,v) = \int_0^1 u'(t)v'(t)\mathrm{dt}.$$

Protože

$$||u||^2 = \int_0^1 \left[||u||^2 + ||u'||^2\right]\mathrm{dt},$$

platnost požadavku (7.61) je důsledkem *Friedrichsovy nerovnosti*

$$\int_\Omega u'(t)^2\mathrm{dt} \ge \tau \int_\Omega u^2\mathrm{dt}, \; \forall u \in \tilde{W}_2^{(1)}(0,1),$$

kde τ je kladná konstanta nezávislá na u ani u'.

Požadavek (7.60) lemmatu 7.1 je automaticky splněn díky tomu, že skalární součin v $\tilde{W}_2^1(0,1)$ je dán výrazem (7.51).

Na základě lemmatu 7.1 tedy platí

Důsledek 7.1 *Existuje právě jedno slabé řevsení úlohy (U).*

Diskretizace.

Buď $V_h \subset V = \tilde{W}_2^1(0,1)$, kde

$$\dim V_h = N < +\infty.$$

Nechť

$$\{\phi_1, ..., \phi_N\}$$

tvoří basi prostoru V_h, tedy

$$V_h = \mathrm{Span}\{\phi_1, ..., \phi_N\}.$$

Tudíž,

$$u_h = \sum_{k=1}^{N} \nu_k \phi_k, \tag{7.64}$$

pro každý prvek $u_h \in V_h$.

Diskretizací úlohy (U) se rozumí úloha

($\mathbf{U_h}$) Nalézt $u_h^\star \in V_h$ takové, že platí

$$B(u_h^\star, v_h) = (f, v_h), \ \forall v_h \in V_h. \tag{7.65}$$

Věta 7.5 *Splňuje-li forma $B = B(u,v)$ požadavky Laxova - Milgramova lemmatu, pak lze nalézt $h_0 > 0$ tak, že pro každé $0 < h \leq h_0$, existuje právě jedno řešení $u_h^\star$ úlohy (U_h).*

Navíc platí odhad

$$\|u^\star - u_h^\star\|_V \leq \kappa d_h(u_h^\star)\|f\|,$$

kde κ nezávisí ani na h ani na f a

$$d_h(v) = \inf\{\|v - v_h\|_V : \ v_h \in V_h\}.$$

Příklad 7.4 *Pro případ, kdy v problému (U_h)*

$$h = \frac{1}{N}, \text{ a } x_k = kh, \ k = 0, 1, ..., N,$$

a basi aproximačního podprostoru tvoří po částech lineární funkce definované formulemi

$$\phi_k(x) = \begin{cases} 0 & \text{pro} \quad 0 \leq x \leq x_{k-1}, \\ \frac{1}{h}(x - (k-1)h) & \text{pro} \quad x_{k-1} \leq x_k, \\ 1 - \frac{1}{h}(x - x_k) & \text{pro} \quad x_k \leq x \leq x_{k+1}, \\ 0 & \text{pro} \quad x_{k+1} \leq x \leq 1, \end{cases}$$

lze odvodit, že

$$d_h(u) = h\|u\|_V.$$

Důsledkem je posléze odhad chyby v metodě konečných prvků *pro aproximaci úlohy (U) ve tvaru*

$$\|u^\star - u_h^\star|_{\tilde{W}_2^1(0,1)} \leq \kappa h \|f\|_{L^2(0,1)}.$$

Vyplývá odtud speciálně, že

$$\|u^\star - u_h^\star\|_{L^2(0,1)} \leq \hat{\kappa} h^2 \|f\|_{L^2(0,1)}.$$

Po dosazení vyjádření (7.64) do (7.65) obdržíme vztahy

$$\sum_{k=1}^{N} \nu_k B(\phi_k\ , \phi_j) = (f,\ \phi_j),\ j = 1, ..., N. \tag{7.66}$$

Vyplnění podmínek Laxova - Milgramova lemmatu zaručuje, že

$$\det\left(B(\phi_k,\ \phi_j)\right) \neq 0$$

a soustava (7.66) má tudíž právě jedno řešení $\nu^\star = (\nu_1^\star,\ ...\nu_N^\star)^T$.

Matice

$$(B(\phi_k,\ \phi_j))$$

se díky její interpretaci v mechanice kontinua nazývá *maticí tuhosti* úlohy (U_h).

Je-li forma $B = B(u,\ v)$ symetrická a jsou-li splněny podmínky Laxova - Milgramova lemmatu, pak matice tuhosti úlohy (U_h) je pozitivně definitní, t.j.

$$(B(u_h,\ u_h)) \geq \tau(u_h,\ u_h)\ \forall u_h,\ u_h \in V_h,\ \tau > 0.$$

Výpočet prvků matice tuhosti spočívá ve stanovení veličin

$$\int_{a+(p-1)h}^{a+ph} \phi_k'(x)\phi_j'(x)\mathrm{dx},\ j, k = 1, ..., N,\ p = 1, ..., N-1,$$

při čemž $\Omega = (a,\ b),\ -\infty < a < b < +\infty$.

Pro výše uvedený případ jednorozměrné oblasti s Ω a pro částech lineárních funkcí ϕ_k lze tyto integrály stanovit přesně. Obecně je nutné prvky matice tuhosti určovat přibližně pomocí vhodných kvadraturních vzorců.