Základy numerické matematiky

Miloslav Feistauer, Václav Kučera
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

OBSAH

1

INTERPOLACE FUNKCÍ

V této části se budeme zabývat následujícím problémem: Nechť je dán interval $[a, b]$, v něm« jsou dány body $a = x_0 < x_1 < \ldots < x_n = b$ a jsou předepsané hodnoty $f(x_0), \ldots, f(x_n)$. Hledáme funkci p daného typu, která vyhovuje podmínkám $p(x_i) = f(x_i),\ i = 0, \ldots, n$. Příkladem je *Lagrangeův interpolační polynom* p_n stupně nejvýe n, pro který je $p_n(x_j) = f(x_j), j = 0, ..., n$. Pro přesnost aproximace dostatečně hladké funkce máme následující větu:

Věta 1 *Je-li $f \in C^{n+1}[a, b]$, $x_0, \ldots, x_n \in [a, b]$, pak pro ka«dé $x \in [a, b]$ existuje $\xi \in (a, b)$ tak, «e*

$$f(x) - p_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi) \prod_{j=0}^{n} (x - x_j).$$

Je-li aproximovaná funkce dokonce třídy C^∞ a má stejně omezené derivace, dostáváme následující výsledek.

Důsledek 2 *Buď $f \in C^\infty([a, b])$, $|f^{(k)}| \le CK^k$ v $[a, b]$ pro vechna $k = 0, 1, \ldots$, s konstantami $C, K > 0$. Pak*

$$\max_{x \in [a,b]} |f(x) - p_n(x)| \le C \frac{K^{n+1}(b-a)^{n+1}}{(n+1)!} \xrightarrow[n \to \infty]{} 0.$$

Tudí«, $p_n \rightrightarrows f$ v $[a, b]$ pro $n \to \infty$.

Cvičení 1 Uva«ujme funkci $f(x) = \sin x$ na intervalu $[0, 1]$. Jaké největí chyby se mů«eme dopustit, aproximujeme-li f pomocí p_9?

Poka«dé vak nedosáhneme tak vynikající aproximace. Příkladem mů«e být funkce $f(x) = \frac{1}{1+x^2}$ na intervalu $[-5, 5]$. Dá se dokázat, «e pro interpolační polynomy p_n s uzly v bodech ekvidistantního dělení je posloupnost $\{\|f - p_n\|_{C([a,b])}\}_{n=1}^{\infty}$ neomezená.

1.1 Po částech polynomiální interpolace

Jednou z nevýhod aproximace interpolačním polynomem je skutečnost, «e hodnoty interpolačního polynomu jsou silně ovlivněny hodnotami funkce f i ve vzdálených uzlech. To způsobuje, «e interpolační polynom mů«e silně oscilovat. Tuto potí« mů«eme odstranit tak, «e funkci f budeme aproximovat pomocí polynomů po částech. To znamená, «e funkce f je aproximována v intervalu $[a, b]$ pomocí funkce φ takové, «e $\varphi|_{[x_i, x_{i+1}]}$ je vhodný polynom. Většinou se navíc po«aduje, aby φ byla třídy C^k pro dané k. Takovou funkci φ nazýváme *spline.*

Nejjednoduím případem ($k = 0$) je aproximace pomocí funkce po částech lineární, jejím« grafem je lomená čára. V praxi je většinou třeba lepí aproximace. Přijatelným řeením je tzv. *kubický spline.*

Definice 1 Řekneme, «e funkce $\varphi : [x_0, x_n] \to \mathbb{R}$ je *kubický spline*, jestli«e

(i) $\varphi \in C^2([x_0, x_n])$,

(ii) $\varphi|_{[x_i, x_{i+1}]}$ je polynom stupně nejvýe 3.

Říkáme, «e φ je *kubický interpolační spline* k f v bodech $x_0, \dots x_n$, jestli«e jsou navíc splněny podmínky $\varphi(x_i) = f(x_i), i = 0, \dots, n$.

Restrikci φ na $[x_i, x_{i+1}]$ označíme φ_i. Funkci φ_i lze psát ve tvaru

$$\varphi_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3.$$

Funkce φ je tedy určena celkem $4n$ parametry. Podmínky z definice kubického interpolačního splinu nám vak dávají jen $4n - 2$ podmínek (hodnoty funkce φ v bodech $x_0, \dots, x_n$ a spojitost $\varphi, \varphi', \varphi''$ v bodech $x_1, \dots x_n$). Dá se tedy očekávat, «e budeme muset jetě dva parametry zvolit. Nejčastěji se pou«ívají okrajové podmínky v krajních bodech x_0 a x_n, a to tří typů:

(α) $\varphi'(x_0) = f'_0$, $\varphi'(x_n) = f'_n$;

(β) $\varphi''(x_0) = f''_0$, $\varphi''(x_n) = f''_n$;

(γ) $\varphi''(x_0) = 0$, $\varphi''(x_n) = 0$.

Kubický interpolační spline určený podmínkou (γ) se nazývá *přirozený spline.*

1.1.1 *Konstrukce kubického splinu*

Budeme konstruovat spline určený podmínkou (β) (jí« je (γ) speciálním případem). Předpokládejme nejprve, «e ji« známe čísla $M_i = \varphi''(x_i)$, tzv. *momenty splinu.* Funkce φ'' je spojitá a po částech lineární. Označíme-li tedy $h_i = x_{i+1} - x_i$, potom pro $x \in [x_i, x_{i+1}]$ máme

$$\varphi''_i(x) = M_i + (M_{i+1} - M_i)\frac{x - x_i}{x_{i+1} - x_i} = M_i\frac{x_{i+1} - x}{h_i} + M_{i+1}\frac{x - x_i}{h_i}. \quad (1.1.1)$$

Integrací dostaneme φ'_i a φ_i:

$$\varphi'_i(x) = -M_i\frac{(x_{i+1} - x)^2}{2h_i} + M_{i+1}\frac{(x - x_i)^2}{2h_i} + A_i \quad (1.1.2)$$

$$\varphi_i(x) = M_i\frac{(x_{i+1} - x)^3}{6h_i} + M_{i+1}\frac{(x - x_i)^3}{6h_i} + A_i(x - x_i) + B_i. \quad (1.1.3)$$

Pomocí známých hodnot $\varphi_i(x_i) = f(x_i)$, $\varphi_i(x_{i+1}) = f(x_{i+1})$ určíme konstanty A_i a B_i: $f(x_i) = \frac{1}{6}M_i h_i^2 + B_i$. Tudí«,

$$B_i = f(x_i) - \frac{1}{6}M_i h_i^2. \quad (1.1.4)$$

Dále, $f(x_{i+1}) = \frac{1}{6}M_{i+1}h_i^2 + A_i h_i + B_i$ a tedy

$$A_i = \frac{1}{h_i}\left(f(x_{i+1}) - \frac{1}{6}M_{i+1}h_i^2 - B_i\right) = \frac{f(x_{i+1}) - f(x_i)}{h_i} - \frac{h_i}{6}(M_{i+1} - M_i). \tag{1.1.5}$$

Zbývá určit hodnoty momentů M_i. M_0 a M_n máme zadané, ostatní vypočteme ze spojitosti derivace φ' v bodech $x_i, i = 1, \ldots, n-1$, což anamená, že $\varphi'_i(x_i+) = \varphi'_{i-1}(x_i-)$. Podle (1.1.2) máme

$$\begin{aligned}\varphi'_{i-1}(x_i-) &= \frac{1}{2}M_i h_{i-1} + A_{i-1} = \\ &= \frac{1}{2}M_i h_{i-1} + \frac{f(x_i) - f(x_{i-1})}{h_{i-1}} - \frac{h_{i-1}}{6}(M_i - M_{i-1})\end{aligned}$$

$$\begin{aligned}\varphi'_i(x_i+) &= -\frac{1}{2}M_i h_i + A_i = \\ &= -\frac{1}{2}M_i h_i + \frac{f(x_{i+1}) - f(x_i)}{h_i} - \frac{h_i}{6}(M_{i+1} - M_i).\end{aligned}$$

Z rovnosti obou derivací dostaneme po úpravách

$$\begin{aligned}M_{i-1}\frac{h_{i-1}}{6} + M_i\left(\frac{h_{i-1} + h_i}{3}\right) + M_{i+1}\frac{h_i}{6} \\ = \frac{f(x_{i+1} - f(x_i)}{h_i} - \frac{f(x_i) - f(x_{i-1})}{h_{i-1}}.\end{aligned} \tag{1.1.6}$$

Označíme-li $\lambda_i = \frac{h_{i-1}}{h_{i-1}+h_i}$, $\mu_i = 1 - \lambda_i = \frac{h_i}{h_{i-1}+h_i}$, lze rovnici (1.1.6) přepsat ve tvaru

$$\lambda_i M_{i-1} + 2M_i + \mu_i M_{i+1} = g_i, \tag{1.1.7}$$

kde g_i je pravá strana rovnice (1.1.6) vynásobená výrazem $\frac{6}{h_{i-1}+h_i}$. Dostáváme soustavu

$$\begin{aligned}2M_1 + \mu_1 M_2 &= g_1 - \lambda_1 f_0'' \\ \lambda_2 M_1 + 2M_2 + \mu_2 M_3 &= g_2 \\ \lambda_3 M_2 + 2M_3 + \mu_3 M_4 &= g_3 \\ &\vdots \\ \lambda_{n-1} M_{n-2} + 2M_{n-1} &= g_{n-1} - \mu_{n-1} f_n''.\end{aligned} \tag{1.1.8}$$

Doká«eme-li nyní existenci a jednoznačnost řeení soustavy (1.1.8), budeme hotovi. Vimněme si, «e prvky na diagonále matice soustavy jsou v«dy 2, zatímco součet vech ostatních prvků v libovolném řádku je mezi 0 a 1 (s výjimkou prvního a posledního dokonce právě 1). Matice je tedy ostře diagonálně dominantní a tedy i regulární.

Cvičení 2 Doka«te, «e ka«dá ostře diagonálně dominantní matice je regulární.

Matice soustavy (1.1.8) je navíc třídiagonální, na které se poměrně jednodue provádí eliminace. Soustava vypadá takto (uva«ujeme obecný případ – matici $n \times n$):

$$\begin{pmatrix} a_1 & c_1 & 0 & \dots & 0 & 0 \\ b_1 & a_2 & c_2 & \dots & 0 & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & & & \dots & b_{n-1} & a_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_n. \end{pmatrix}$$

Při přímém chodu Gaussovy eliminace (tvorba horní trojúhelníkové matice) zmizí vechny prvky b_j a na diagonále se postupně objeví členy $A_{jj} = \eta_j$, kde

$$\eta_1 = a_1,\ \eta_2 = a_2 - \frac{b_1}{\eta_1} c_1, \quad \text{obecně } \eta_i = a_i - \frac{b_{i-1}}{\eta_{i-1}} c_{i-1} \quad (i = 2, \dots, n),$$

a ve vektoru pravých stran analogicky vzniknou ξ_j, kde

$$\xi_1 = d_1,\ \xi_2 = d_2 - \frac{b_1}{\eta_1} \xi_1, \quad \text{obecně } \eta_i = d_i - \frac{b_{i-1}}{\eta_{i-1}} \xi_{i-1} \quad (i = 2, \dots, n).$$

Dostaneme tedy soustavu

$$\begin{pmatrix} \eta_1 & c_1 & 0 & \dots & 0 & 0 \\ 0 & \eta_2 & c_2 & \dots & 0 & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & & & \dots & 0 & \eta_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n, \end{pmatrix}$$

z ní« u« snadno vypočítáme neznámé y_j:

$$y_n = \frac{\xi_n}{\eta_n},\ y_{n-1} = \frac{\xi_{n-1} - c_{n-1} y_n}{\eta_{n-1}}, \quad \text{obecně } y_i = \frac{\xi_i - c_i y_{i+1}}{\eta_i} \quad (i = 1, \dots, n-1).$$

Je mo«né dokázat, «e pro ostře diagonálně dominantní matici vyjdou po eliminaci prvky na diagonále nenulové.

1.1.2 *Odhad chyby*

Zabývejme se nyní otázkou, jaké chyby se dopustíme, aproximujeme-li funkci interpolačním kubickým splinem. Zhruba řečeno, je-li f dostatečně hladká v $[a, b]$ a dělení intervalu neomezeně zjemňujeme *vhodným způsobem*, pak interpolační kubické spliny konvergují stejnoměrně k f (případně i s některými derivacemi). Přesnou formulaci tohoto výsledku dává následující věta, kterou uvedeme bez důkazu.

Věta 3 *Buď $f \in C^4[a,b]$. Pak existuje konstanta $C > 0$ taková, že platí: Nech¸ $K > 0$ je konstanta. Uva«ujme dělení D intervalu $[a,b]$, které je tvořeno body $a = x_0 < \ldots < x_n = b$ a které splňuje podmínku*

$$\frac{\max h_i}{\min h_i} \leqq K, \tag{1.1.9}$$

kde $h_i = x_{i+1} - x_i$. Dále uva«ujme okrajové podmínky pro interpolační kubický spline $\varphi''(x_0) = f''(x_0)$, $\varphi''(x_n) = f''(x_n)$. Potom

$$\left| f^{(k)}(x) - \varphi^{(k)}(x) \right| \leqq CK\, h^{4-k}, \quad x \in [a,b], \quad k = 0, \ldots, 3,$$

kde $h = \max h_i$, přičem« v případě $k = 3$ uva«ujeme v dělících bodech derivaci zleva nebo zprava.

Důsledek 4 *Máme-li posloupnost dělení intervalu $[a,b]$ splňujíc[ích podmínku (1.1.9) takových, «e $h \to 0$, dostáváme*

$$\varphi^{(k)} \rightrightarrows f^{(k)}, \quad k = 0, \ldots, 3.$$

Poznámka 5 Pokud pou«íváme interpolační spline pro interpolaci naměřených hodnot funkce f, obvykle neznáme hodnoty $f''(a), f''(b)$. Nahradíme-li okrajové podmínky ve větě 3 podmínkami $\varphi''(x_0) = 0$, $\varphi''(x_n) = 0$, dostaneme (slabí) odhad

$$|f(x) - \varphi(x)| \leqq CK\, h^2, \quad x \in [a,b].$$

1.1.3 *Spline s napětím*

Ne v«dy představuje kubický interpolační spline ideální aproximaci interpolované funkce. Například při aproximaci nespojitých funkcí nebo funkcí s nespojitou derivací dostáváme nepříjemné oscilace v okolí nespojitosti. Proto se někdy pou«ívá modifikovaná konstrukce tzv. *splinu s napětím.* Při této konstrukci opět předpokládáme, že $\varphi \in C^2[a,b]$, $\varphi(x_i) = f(x_i)$, ale po«adavek, aby φ byla funkce po částech kubická, se nahrazuje po«adavkem, aby funkce $\varphi_i = \varphi|_{[x_i, x_{i+1}]}$ byla řeením diferenciální rovnice

$$\varphi_i^{(4)} - \tau \varphi_i'' = 0,$$

kde $\tau \geqq 0$ je tzv. *napě¸ový parametr.* Pokud bychom polo«ili $\tau = 0$, dostali bychom polynomy stupně nejvýe 3 a tedy kubický inrepolační spline. Pro τ blízké k ∞ dostaneme $\varphi'' \approx 0$. Význam napě¸ového parametru je tedy takový, «e pro velké τ má spline tendenci být téměř lineární v důsledku většího napětí.

1.1.4 *Hermiteův spline*

Dalí modifikací je *Hermiteův spline.* Při této konstrukci po«adujeme pouze, že $\varphi \in C^1[a,b]$, $\varphi|_{[x_i, x_{i+1}]}$ je opět polynom stupně nejvýe 3 a $\varphi(x_i) = f(x_i)$, $\varphi'(x_i) = f'(x_i)$, $i = 0, \ldots, n$.

V praxi při interpolaci naměřených hodnot ovem není derivace f' známa. V takových případech se hodnota $f'(x_i)$ aproximujeme výrazem

$$\frac{f(x_{i+1}) - f(x_{i-1})}{x_{i+1} - x_{i-1}}, \quad i = 1, \ldots, n-1.$$

Pokud je toti« $f \in C^2[a,b]$, je pro $h \to 0$

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{x_{i+1} - x_{i-1}} + O(h). \tag{1.1.10}$$

Dále uva«ujeme okrajové podmínky zadávající hodnoty $\varphi'(x_0)$ a $\varphi'(x_n)$.

Cvičení 3 Doka«te platnost vztahu (1.1.10)

Cvičení 4 Za předpokladu $f \in C^3[a,b]$ najděte přesnější aproximaci $f'(x_i)$ pro $i = 1, \ldots, n-1$ s chybou $O(h^2)$.

2

NUMERICKÉ ŘEENÍ OBYČEJNÝCH DIFERENCIÁLNÍCH ROVNIC

V této kapitole se budeme zabývat numerickým řeením obyčejných diferenciálních rovnic. Proto«e rovnici vyího řádu lze v«dy jednodue převést na soustavu rovnic prvního řádu, budeme se v této části zabývat pouze rovnicemi prvního řádu, navíc rozřeenými vzhledem k derivaci. Abychom zajistili jednoznačnost jejich řeení, budeme zadávat tzv. počáteční podmínky. Budeme se zabývat řeením úlohy najít funkci y definovanou v intervalu $[a, b]$ splňující diferenciální rovnici tvaru

$$y' = f(x, y),$$

k ní« přidáme počáteční podmínku $y(a) = \eta$, kde η je předepsaná hodnota. Uva«ovaná rovnice mů«e reprezentovat buď jednu diferenciální rovnici nebo soustavu diferenciálních rovnic. V tomto případě řeení $y = (y^1, \ldots, y^s) : [a, b] \to \mathbb{R}^s$ a pravá strana rovnice $f = (f^1, \ldots, f^n)$ jsou s-rozměrné vektorové funkce.

V řadě případů je mo«né najít tzv. analytické řeení diferenciálních rovnic, které je vyjádřené pomocí elementárních funkcí a jejich neurčitých integrálů. Pro většinu diferenciálních rovnic to ale není mo«né. V těchto případech je třeba pou«ít numerické metody.

2.1 Příklady některých diferenciálních rovnic

V tomto odstavci uvedeme několik příkladů diferenciálních rovnic, které je třeba řeit pomocí vhodných numerických metod.

Příklad 1 *Pohyb částice v silovém poli* Trajektorii pohybu částice lze popsat jednou vektorovou funkcí $\boldsymbol{x} = \boldsymbol{x}(t)$, kde proměnná t má fyzikální význam času a $\boldsymbol{x}(t)$ je poloha částice v čase t. Newtonův pohybový zákon pak lze formulovat ve tvaru (m_p je hmotnost částice)

$$m_p \boldsymbol{x}'' = \boldsymbol{F}(\boldsymbol{x}', \boldsymbol{x}, t),$$

co« je vlastně soustava tří obyčejných diferenciálních rovnic druhého řádu, kterou lze převést na soustavu esti obyčejných diferenciálních rovnic prvního řádu. Funkce $\boldsymbol{F}$ zde vyjadřuje závislost působící síly na čase, poloze částice a její rychlosti. Obecně je vak tato závislost natolik slo«itá, «e není ance tuto soustavu vyřeit analyticky. Mnohdy nám vak stačí najít numerickými metodami řeení přibli«né.

Příklad 2 I u velmi jednoduchých rovnic se nám mů«e stát, «e přesné řeení nedoká«eme najít. Typickým příkladem je rovnice

$$y' = x^2 + y^2,$$

o ní« je dokázáno, «e «ádné její řeení nelze vyjádřit pomocí elementárních funkcí a jejich neurčitých integrálů.

Příklad 3 I v případě, «e umíme přesné řeení najít, se mů«e stát, «e se bez numerických metod neobejdeme. Rovnice $y' = 1 - 2xy$ s počáteční podmínkou $y(0) = 0$ má řeení

$$y(x) = e^{-x^2} \int_0^x e^{t^2} \, \mathrm{d}t.$$

Chceme-li znát jeho hodnotu v nějakém bodě x, potřebujeme numericky vypočítat určitý integrál.

Příklad 4 Rovnice

$$y'' + \frac{A}{x} y' + \left(\frac{B}{x^2} + C + Dx^2 \right) y = 0$$

je příkladem rovnice Fuchsova typu. Její řeení lze najít metodou mocninných řad, ale pro většinu argumentů x řada konverguje pomalu. Nahradíme-li Dx^2 členem Dx^3, nelze metodu mocninných řad pou«ít.

V dalím se budeme zabývat řeením řeením následující úlohy. Hledáme funkci y definovanou v intervalu $[a, b]$ takovou, že

$$y' = f(x, y), \ x \in [a, b], \quad y(a) = \eta. \tag{2.1.1}$$

2.2 Příklady nejjednoduích diskrétních metod

Nech˛ $y : [a, b] \to I\!R$ je řeení uva«ovaného problému (2.1.1). Uva«ujme dělení intervalu $[a, b]$ tak, «e $a = x_0 < \ldots < x_N \leq b$, $x_n = a + nh$, kde $h > 0$ nazýváme krokem metody. Budeme se sna«it přiřadit bodům x_i přibli«né hodnoty y_i aproximující hodnoty přesného řeení $y(x_i)$. Uvedeme zde tři jednoduché příklady.

2.2.1 *Eulerova metoda*

Je to nejjednoduí metoda numerického řeení ODR. Předpokládejme, «e máme řeení y diferenciální rovnice a v intervalu $[a, b]$ máme dva body $x_n < x_{n+1}$ uva«ovaného dělení. Předpokládejme navíc, «e řeení je třídy C^2. Taylorův vzorec nám dává $y(x_{n+1}) = y(x_n) + hy'(x_n) + O(h^2)$. Odtud

$$y'(x_n) = \frac{y(x_{n+1}) - y(x_n)}{h} + O(h).$$

Zanedbáme-li výraz $O(h)$ a nahradíme hodnoty $y(x_i)$ přesného řešení hodnotami y_i přibližného řešení, dostaneme formuli

$$\frac{1}{h}(y_{n+1} - y_n) = f(x_n, y_n).$$

Dostáváme tedy rekurentní vztahy

$$\begin{aligned} y_0 &= \eta = \text{ počáteční podmínka k diferenciální rovnici,} \\ y_{n+1} &= y_n + hf(x_n, y_n), \quad n \geq 0. \end{aligned}$$

Cvičení 5 Je-li dokonce $y \in C^3[a, b]$, lze derivaci aproximovat přesněji:

$$y'\left(x_n + \tfrac{h}{2}\right) = \frac{y(x_{n+1}) - y(x_n)}{h} + O(h^2).$$

2.2.2 *Rungeova-Kuttova metoda*

Vyjdeme-li ze vztahu uvedeného v předchozím cvičení a vztahů

$$\begin{aligned} y'\left(x_n + \tfrac{h}{2}\right) &= f(x_n + \tfrac{h}{2}, y(x_n + \tfrac{h}{2})), \\ y(x_n + \tfrac{h}{2}) &= y(x_n) + \tfrac{h}{2}y'(x_n) + O(h^2), \\ y'(x_n) &= f(x_n, y(x_n)), \end{aligned}$$

a u«ijeme-li aproximace $y_i \approx y(x_i)$, dostaneme

$$\frac{1}{h}(y_{n+1} - y_n) = f\left(x_n + \tfrac{h}{2}, y_n + \tfrac{h}{2}f(x_n, y_n)\right).$$

Tímto postupem jsme dospěli k rekurentním vztahům

$$y_0 = \eta, \quad y_{n+1} = y_n + hf\left(x_n + \tfrac{h}{2}, y_n + \tfrac{h}{2}f(x_n, y_n)\right), \quad n \geq 0.$$

Tato metoda je *Rungeova-Kuttova metoda druhého řádu*. Obecným postupem při odvození Rungeových-Kuttových metod se budeme zabývat později.

2.2.3 *Dvoukroková metoda*

Společnou vlastností obou uvedených metod bylo, «e hodnota y_{n+1} se vypočítávala pouze z *jedné* předcházející hodnoty y_n (a samozřejmě z x_n a h). U vícekrokových metod pou«íváme rekurentní vyjádření z více předelých hodnot.

Mějme nyní opět $y \in C^3[a, b]$, x_n, x_{n+1}, x_{n+2} tři po sobě jdoucí body ekvidistantního dělení $[a, b]$ (tedy $x_n = a + nh$, kde h je krok metody). Pak máme

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{1}{2}h^2y''(x_n) + O(h^3), \tag{2.2.2}$$

$$y(x_{n+2}) = y(x_n) + 2hy'(x_n) + 2h^2y''(x_n) + O(h^3). \tag{2.2.3}$$

Odečtením čtyřnásobku (2.2.2) od (2.2.3) dostaneme

$$y(x_{n+2}) - 4y(x_{n+1}) = -3y(x_n) - 2hy'(x_n) + O(h^3),$$

odkud dosazením rovnice $y'(x_n) = f(x_n, y(x_n))$ dojdeme k rekurentnímu vztahu

$$y_{n+2} - 4y_{n+1} + 3y_n = -2hf(x_n, y_n). \tag{2.2.4}$$

Zde klademe $y_0 = \eta$ a y_1 vypočteme pomocí některé jednokrokové metody. Popsaná metoda je dvoukroková – hodnotu y_{n+2} přibli«ného řeení počítáme pomocí dvou předcházejících hodnot y_{n+1} a y_n.

2.3 Obecné jednokrokové metody

V obecných jadnokrokových metodách je dán krok $h > 0$ a počáteční podmínka $y(x_0) = \eta$. Uzly jsou $x_n = a + nh$, $n \geq 0$. Hodnoty přibli«ného řeení počítáme podle rekurentního vztahu

$$y_0 = \eta, \quad y_{n+1} = y_n + h\Phi(x_n, y_n, h), \quad x_n, x_{n+1} \in [a, b]. \tag{2.3.5}$$

Funkce Φ (která závisí na f) se nazývá *přírůstková funkce.* Například u Eulerovy metody máme $\Phi(x, y, h) = f(x, y)$, u Rungeovy-Kuttovy metody druhého řádu je $\Phi(x, y, h) = f(x + \frac{h}{2}, y + \frac{h}{2}f(x, y))$.

Chybou metody (v bodě x_n) rozumíme rozdíl $e_n = y_n - y(x_n)$. Někdz mluvíme o tzv. *akumulované diskretizační chybě.* Naím cílem je

(1) Ukázat, «e v jistém smyslu je $e_n \to 0$ pro $h \to 0$.
(2) Najít odhad e_n v závislosti na h.

Abychom měli zaručenou existenci a jednoznačnost řeení, budeme předpokládat, «e funkce $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ je spojitá a je lipschitzovská vzhledem k y, tj. «e

$$|f(x, y_1) - f(x, y_2)| \leq L_f|y_1 - y_2| \qquad \forall x \in [a, b], \quad y_1, y_2 \in \mathbb{R}.$$

Z Picardovy věty (viz **doplnit citaci**) vyplývá, «e za těchto předpokladů má úloha (2.1.1) právě jedno řeení $y : [a, b] \to \mathbb{R}$.

2.3.1 *Konvergence jednokrokových metod*

Definice 2 Řekneme, «e jednokroková metoda (2.3.5) je konvergentní, jestli«e platí následující tvrzení: je-li $y : [a, b] \to \mathbb{R}$ řeením úlohy (2.1.1), pak

$$\forall x \in [a, b] : \lim_{\substack{h \to 0+ \\ x_n = x}} y_n = y(x),$$

kde y_n je přibli«né řeení v uzlu x_n.

Definice 3 *Jiná definice konvergence* Označme $e(h)$ maximální chybu metody v intervalu $[a, b]$, tj.

$$e(h) = \max_{\substack{x_n \in [a,b] \\ x_n = a + nh}} |e_n|.$$

Řekneme, «e metoda (2.3.5) je konvergentní, jestli«e platí následující tvrzení: je-li $y : [a, b] \to \mathbb{R}$ řeením úlohy (2.1.1), pak

$$\lim_{h \to 0+} e(h) = 0.$$

Poznámka 6 Snadno nahlédneme, «e konvergence podle druhé definice implikuje konvergenci podle první definice.

Před analýzou jednokrokových metod uvedeme jedno pomocné lemma, které nám bude často u«itečné.

Lemma 1 *Nechť $A, B \geq 0$, $N \geq 1$ celé a nechť je splněna podmínka*

$$|\xi_{n+1}| \leq A|\xi_n| + B, \quad \text{pro } n = 0, \ldots, N-1.$$

Potom pro $n = 0, \ldots, N$ platí odhad

$$|\xi_n| \leq A^n|\xi_0| + \begin{cases} \dfrac{A^n - 1}{A - 1} B & \text{pro } A \neq 1, \\ Bn & \text{pro } A = 1. \end{cases} \tag{2.3.6}$$

Cvičení 6 Dokažte toto lemma.

Při aplikaci lemmatu 1 je často $A = 1 + \delta$, $\delta > 0$. Použijeme-li nerovnost $1 + \delta < e^{\delta}$, pak z (2.3.6) plyne

$$|\xi_n| \leq e^{n\delta}|\xi_0| + \frac{e^{n\delta} - 1}{\delta} B. \tag{2.3.7}$$

Pro $L > 0$ a $x \in \mathbb{R}$ označme

$$E_L(x) = \frac{e^{Lx} - 1}{L}. \tag{2.3.8}$$

(E_L je tzv. *Lipschitzova funkce.*)

Předpoklad (P_Φ): Předpokládejme, že přírůstková funkce $\Phi : [a, b] \times \mathbb{R} \times [0, h_0] \to \mathbb{R}$, $\Phi = \Phi(x, y, h)$, je spojitá a splňuje Lipschitzovu podmínku vzhledem k y s konstantou $L > 0$:

$$|\Phi(x, y, h) - \Phi(x, y^*, h)| \leq L|y - y^*|, \quad x \in [a, b],\ y, y^* \in \mathbb{R},\ h \in [0, h_0]. \quad (+)$$

Definice 4 Řekneme, že jednokroková metoda s přírůstkovou funkcí Φ pro řešení diferenciální rovnice $y' = f(x, y)$ je *konsistentní*, jestliže

$$\Phi(x, y, 0) = f(x, y), \qquad x \in [a, b],\ y \in \mathbb{R}.$$

Věta 7 *Jednokroková metoda s přírůstkovou funkcí Φ splňující předpoklad (P_Φ) je konvergentní, právě když je konsistentní.*

Důkaz vynecháme, protože nás zajímá odhad chyby metody e_n. Odhad chyby provádíme ve dvou krocích.

a) Dosadíme přesné řešení do uvažované metody. Dostaneme vztah

$$y(x_n + h) - y(x_n) - h\Phi(x_n, y(x_n), h) = h\delta_n, \quad x_n, x_n + h \in [a, b],\ h \in (0, h_0).$$

b) Odhadneme δ_n a z tohoto odhadu odvodíme odhad chyby e_n.

Veličinu δ_n nazýváme lokální relativní diskretizační chybou v uzlu x_n. Obecně *lokální relativní diskretizační chybu* (krátce chybu diskretizace) v bodě x definujeme jako výraz

$$\Delta(x, y(x), h) - \Phi(x, y(x), h),$$

kde

$$\Delta(x, y(x), h) = \frac{y(x+h) - y(x)}{h}, \qquad x, x+h \in [a, b],\ h \in (0, h_0),$$

je tzv. *přesný relativní přírůstek.*

Věta 8 *Nechť $y : [a, b] \to \mathbb{R}$ je přesné řeení úlohy (2.1.1) v intervalu $[a, b]$ a y_n pro $x_n \in [a, b]$ jsou hodnoty přibližného řeení vypočtené pomocí jednokrokové metody (2.3.5) s přírůstkovou funkcí Φ splňující předpoklad (P_Φ). Nechť navíc existují konstanty N, $p > 0$ takové, že*

$$|\Delta(x, y(x), h) - \Phi(x, y(x), h)| \le N h^p, \quad x, x+h \in [a, b],\ h \in (0, h_0). \tag{2.3.9}$$

Potom pro chybu metody (tj. akumulovanou diskretizační chybu) $e_n = y_n - y(x_n)$ platí odhad

$$|e_n| \le N h^p E_L(x_n - a), \qquad x_n \in [a, b],\ h \in (0, h_0). \tag{2.3.10}$$

Důkaz: Zřejmě máme $e_0 = y_0 - y(x_0) = 0$. Dále máme vztahy

$$y_{n+1} = y_n + h\Phi(x_n, y_n, h),$$

$$y(x_{n+1}) = y(x_n) + h\Phi(x_n, y(x_n), h) + h\delta_n,$$

kde

$$\delta_n = \Delta(x_n, y(x_n), h) - \Phi(x_n, y(x_n), h)$$

pro $x_n, x_{n+1} \in [a, b]$, $h \in (0, h_0)$. Odtud vyplývá odečtením druhé rovnice od první

$$e_{n+1} = e_n + h(\Phi(x_n, y_n, h) - \Phi(x_n, y(x_n), h)) - h(\Delta(x_n, y(x_n), h) - \Phi(x_n, y(x_n), h)).$$

Použijeme-li předpoklad (P_Φ) a vztah (2.3.9), dostaneme ihned nerovnost

$$|e_{n+1}| \le (1 + hL)|e_n| + N h^{p+1}, \quad x_n, x_{n+1} \in [a, b],\ h \in (0, h_0).$$

Aplikace lemmatu 1 dává odhad

$$\begin{aligned} |e_n| &\le \frac{(1+hL)^n - 1}{hL} N h^{p+1} \le \\ &\le \frac{e^{nhL} - 1}{L} N h^p = N h^p E_L(x_n - a), \quad x_n \in [a, b],\ h \in (0, h_0), \end{aligned}$$

□

Výe uvedené výsledky nás vedou k zavedení řádu jednokrokové metody.

Definice 5 *Řekneme, «e metoda (2.3.5) je řádu p, jestli«e podmínka (2.3.9) je splněna pro ka«dé dostatečně hladké řeení y rovnice (2.1.1).*

Cvičení 7 Uka«te, «e Eulerova metoda odvozená v 2.2.1 je řádu $p = 1$.

Poznámka 9 Jednokrokové metody mají tu výhodu, «e v ka«dém kroku (při přechodu od x_n k x_{n+1}) lze měnit délku kroku. Tzn., «e mů«eme uva«ovat obecné dělení $a = x_0 < x_1 < \ldots \leq b$ intervalu $[a, b]$, polo«it $h_n = x_{n+1} - x_n$ a definovat

$$y_{n+1} = y_n + h_n \Phi(x_n, y_n, h_n).$$

2.4 Odvození některých jednokrokových metod

Naím cílem bude odvodit jednokrokové metody vyího řádu p.

2.4.1 *Metody zalo«ené na přímém pou«ití Taylorova vzorce*

V tomto odstavci budeme konstruovat jednokrokovou metodu p-tého řádu za předpokladu, «e pravá strana f diferenciální rovnice (2.1.1) je třídy C^p. Potom lze ukázat, «e přesné řeení y je třídy C^{p+1}. Nech¸ $x, x + h \in [a, b]$. Z Taylorova vzorce plyne, že

$$y(x+h) = y(x) + \sum_{k=1}^{p} \frac{y^{(k)}(x)}{k!} h^k + \frac{y^{(p+1)}(\tilde{x})}{(p+1)!} h^{p+1}, \quad \tilde{x} \in [x, x+h]. \tag{2.4.11}$$

Tudí«,

$$\Delta(x, y(x), h) = \frac{1}{h}(y(x+h) - y(x)) = \sum_{k=1}^{p} \frac{y^{(k)}(x)}{k!} h^{k-1} + \frac{y^{(p+1)}(\tilde{x})}{(p+1)!} h^p, \tag{2.4.12}$$

kde $\tilde{x} \in [x, x+h]$. Nyní pou«ijeme diferenciální rovnici a pokusíme se vypočítat prvních několik derivací funkce y:

$$\begin{aligned}
y'(x) &= f(x, y(x)), \\
y''(x) &= \frac{\mathrm{d}}{\mathrm{d}x} f(x, y(x)) = \frac{\partial f}{\partial x}(x, y(x)) + y'(x) \frac{\partial f}{\partial y}(x, y(x)) = \\
&= \frac{\partial f}{\partial x}(x, y(x)) + f(x, y(x)) \frac{\partial f}{\partial y}(x, y(x)), \\
y'''(x) &= \frac{\mathrm{d}}{\mathrm{d}x} \left\{ \left(\frac{\partial f}{\partial x} + f \frac{\partial f}{\partial y} \right)(x, y(x)) \right\} = \frac{\partial}{\partial x} \{\ldots\} + f \frac{\partial}{\partial y} \{\ldots\}.
\end{aligned}$$

Pro zjednoduení zápisu definujme *diferenciální operátor* [1] takto: pro funkci $\varphi \in C^1(G)$, kde $G \subset I\!R^2$, položme

$$D\varphi = \frac{\partial \varphi}{\partial x} + f \cdot \frac{\partial \varphi}{\partial y}.$$

[1] Jedná se o zobrazení z jistého prostoru funkcí do nějakého jiného prostoru funkcí.

Mocninou rozumíme skládání, tj. $D^0\varphi := \varphi$ a pro $k \geq 0$ klademe $D^{k+1}\varphi := D(D^k\varphi)$. Pomocí operátoru D mů«eme snadno vyjádřit derivace funkce y:

$$\begin{aligned} y'(x) &= (D^0 f)(x, y(x)), \\ y''(x) &= (D^1 f)(x, y(x)), \\ y'''(x) &= (D^2 f)(x, y(x)), \\ &\vdots \\ y^{(k)}(x) &= (D^{k-1} f)(x, y(x)). \end{aligned}$$

Dosazením do (2.4.12) dostaneme

$$\Delta(x, y(x), h) = \sum_{k=1}^{p} \frac{(D^{k-1} f)(x, y(x))}{k!} h^{k-1} + \frac{y^{(p+1)}(\tilde{x})}{(p+1)!} h^p,$$

co« nás vede k tomu, abychom definovali přírůstkovou funkci ve tvaru

$$\Phi(x, y, h) = \sum_{k=1}^{p} \frac{(D^{k-1} f)(x, y)}{k!} h^{k-1}.$$

Pak je toti«

$$|\delta(x, h)| = |\Delta(x, y(x), h) - \Phi(x, y(x), h)| \leq \left| \frac{y^{(p+1)}(\tilde{x})}{(p+1)!} h^p \right| \leq N h^p,$$

kde

$$N = \max_{x \in [a,b]} \left| \frac{y^{(p+1)}(x)}{(p+1)!} \right|.$$

Cvičení 8 Vypočtěte $D^2 f$, $D^3 f$ a pro $f(x, y) = x^2 + y^2$ vypočtěte $D^4 f$.

2.4.2 *Rungeovy-Kuttovy metody*

Získali jsme sice jednokrokovou metodu libovolně vysokého řádu, ale pro praxi je téměř nepou«itelná. Za prvé proto, «e počítaní mocnin operátoru D vede brzy k příli slo«itým výrazům (viz cvičení 8). Druhý důvod spočívá v tom, «e v definici funkce Φ vystupuje nejen funkce f sama, ale i její parciální derivace. V dalím označme

$$\tilde{\Phi}(x, y, h) = \sum_{k=1}^{p} \frac{(D^{k-1} f)(x, y)}{k!} h^{k-1}.$$

Tuto přírůstkovou funkci pou«ijeme k odvození *Rungeových-Kuttových metod*, u nich« se Φ počítá pouze pomocí hodnot f. Jejich přírůstkovou funkci budeme psát ve tvaru

$$\Phi(x, y, h) = \sum_{k=1}^{P} w_i k_i(x, y, h),$$

kde

$$\begin{aligned} k_1(x,y,h) &= f(x,y), \\ k_2(x,y,h) &= f(x+\alpha_2 h, y+\beta_{21} h k_1(x,y,h)), \\ &\vdots \\ k_i(x,y,h) &= f\left(x+\alpha_i h, y+h\sum_{j=1}^{i-1}\beta_{ij}k_j(x,y,h)\right), \quad i=2,\ldots,P. \end{aligned}$$

Zde P, w_i, α_i a β_{ij} jsou vhodně zvolené konstanty (tak, aby výsledná metoda byla řádu p). V dalím se budeme sna«it určit hodnoty těchto konstant tak, aby platila rovnost $\Phi(x,y,h) - \tilde{\Phi}(x,y,h) = O(h^p)$.

Příklad 5 Pokusme se odvodit Rungeovu-Kuttovu metodu druhého řádu . Tzn., že $p = 2$. Zkusíme zvolit počet členů také $P = 2$. Nech¸ $f \in C^2$. Napime vztahy pro Φ a $\tilde{\Phi}$:

$$\begin{aligned} \Phi(x,y,h) &= w_1 f(x,y) + w_2 f(x+\alpha_2 h, y+\beta_{21} h f(x,y)) \\ \tilde{\Phi}(x,y,h) &= f(x,y) + \frac{h}{2}(Df)(x,y) = \\ &= f(x,y) + \frac{h}{2}\left(\frac{\partial f}{\partial x}(x,y) + f(x,y)\frac{\partial f}{\partial y}(x,y)\right). \end{aligned}$$

Dále rozepime výraz $f(x+\alpha_2 h, y+\beta_2 h f(x,z))$ podle Taylorova vzorce:[2]

$$\begin{aligned} f(x+\alpha_2 h, y+\beta_{21} h f(x,y)) &= f(x,y) + \left(\alpha_2 h\frac{\partial}{\partial x} + \beta_{21} h f \frac{\partial}{\partial y}\right) f(x,y) + \\ &+ \frac{1}{2!}\left(\alpha_2 h\frac{\partial}{\partial x} + \beta_{21} h f \frac{\partial}{\partial y}\right)^2 f(x+\theta h, y+\theta\beta_{21} h f(x,y)) = \\ &= f(x,y) + \alpha_2 h\frac{\partial f}{\partial x}(x,y) + \beta_{21} h f(x,y)\frac{\partial f}{\partial y}(x,y) + \\ &+ \frac{1}{2}\left[\alpha_2^2 h^2 \frac{\partial^2 f}{\partial x^2} + 2\alpha_2\beta_{21} h^2 f(x,y)\frac{\partial^2 f}{\partial x \partial y} + \beta_{21}^2 h^2 f^2(x,y)\frac{\partial^2 f}{\partial y^2}\right] \\ &\quad (x+\theta h, y+\theta\beta_{21} h f(x,y)). \end{aligned}$$

[2]Taylorův vzorec pro funkce n proměnných: Buď $f \in C^k(\Omega)$, $\Omega \subseteq \mathbb{R}^n$ otevřená, $a, b \in \Omega$ a předpokládejme, «e i celá úsečka s krajními body a, b le«í v Ω; potom

$$f(b) = \sum_{j=0}^{k-1}\frac{1}{j!}\left[\sum_{i=1}^{n}(b_i - a_i)\frac{\partial}{\partial x_i}\right]^j f(a) + \frac{1}{k!}\left[\sum_{i=1}^{n}(b_i - a_i)\frac{\partial}{\partial x_i}\right]^k f(a+\theta(b-a))$$

pro nějaké $\theta \in [0,1]$. Umocňováním hranatých závorek se rozumí umocňování (tedy skládání) tohoto diferenciálního operátoru. K důkazu stačí rozepsat $g(b)$, kde $g(t) = f(a+t(b-a))$, podle jednorozměrného Taylorova vzorce se středem v bodě 0 a vyjádřit derivace $g^{(j)}$ pomocí derivací funkce f.

Poslední sčítanec na pravé straně je ovem řádu $O(h^2)$. Dostáváme tedy

$$\Phi(x,y,h) = (w_1+w_2)f(x,y)+w_2\alpha_2 h\frac{\partial f}{\partial x}(x,y)+w_2\beta_{21}hf(x,y)\frac{\partial f}{\partial y}(x,y)+O(h^2).$$

Srovnejme nyní koeficienty u f, $\frac{\partial f}{\partial x}$, $f\frac{\partial f}{\partial y}$ ve vyjádření Φ a $\tilde{\Phi}$. Budou-li stejné, bude i $\Phi - \tilde{\Phi} = O(h^2)$. Získáváme vztahy

$$w_1 + w_2 = 1, \qquad w_2\alpha_2 = \frac{1}{2}, \qquad w_2\beta_{21} = \frac{1}{2},$$

což jsou tři rovnice pro čtyři neznámé. Jednu neznámou si tedy mů«eme zvolit; zvolme $w_2 = \alpha \neq 0$ (jinak by nemohl platit vztah $\alpha_2 w_2 = \frac{1}{2}$). Dostaneme

$$w_1 = 1-\alpha, \qquad w_2 = \alpha, \qquad \alpha_2 = \beta_{21} = \frac{1}{2\alpha}.$$

Tyto hodnoty nám dávají celou třídu Rungeých-Kuttových metod druhého řádu. Nejpou«ívanější hodnoty parametru α jsou 1, $\frac{3}{4}$ a $\frac{1}{2}$.

$$\begin{aligned}
\alpha = 1: &\quad y_{n+1} = y_n + hf(x_n + \frac{h}{2}, y_n + \frac{h}{2}f(x_n,y_n)) \\
\alpha = \frac{3}{4}: &\quad y_{n+1} = y_n + \frac{h}{4}\left(f(x_n,y_n) + 3f(x_n+\frac{2}{3}h, y_n+\frac{2}{3}hf(x_n,y_n))\right) \\
\alpha = \frac{1}{2}: &\quad y_{n+1} = y_n + \frac{h}{2}\left(f(x_n,y_n) + f(x_n+h, y_n+hf(x_n,y_n))\right)
\end{aligned}$$

V případě, «e máme více metod stejného řádu, vybíráme metodu buď tak, aby koeficienty byly co nejjednoduí nebo tak, aby konstanta N v odhadu

$$|\Delta(x,y(x),h) - \Phi(x,y(x),h)| \leq Nh^p$$

byla co nejmení.

Rungeovy-Kuttovy metody 3. řádu

V tomto případě postačuje $P = 3$, tedy $\Phi(x,y,h) = w_1k_1+w_2k_2+w_3k_3$. Uvedeme dva příklady:

(a)

$$\begin{aligned}
y_{n+1} &= y_n + h\left(\frac{2}{9}k_1 + \frac{1}{3}k_2 + \frac{4}{9}k_3\right) \\
k_1 &= f(x,y) \\
k_2 &= f(x+\tfrac{h}{2}, y+\tfrac{h}{2}k_1) \\
k_3 &= f(x+\tfrac{3}{4}h, y+\tfrac{3}{4}hk_2)
\end{aligned}$$

(b)

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{6}\left(k_1 + 4k_2 + k_3\right) \\ k_1 &= f(x, y) \\ k_2 &= f(x + \tfrac{h}{2}, y + \tfrac{h}{2}k_1) \\ k_3 &= f(x + h, y - hk_1 + 2hk_2) \end{aligned}$$

Rungeovy-Kuttovy metody 4. řádu

Nejčastěji jsou používány metody čtvrtého řádu. U těchto metod máme naposledy $P = p$, tedy $\Phi(x, y, h) = w_1k_1 + w_2k_2 + w_3k_3 + w_4k_4$. (U metod vyššího řádu je $P > p$.) Uvedeme tři příklady.

Standardní formule:

$$\begin{aligned} w_1 = w_4 = \tfrac{1}{6}, &\qquad w_2 = w_3 = \tfrac{1}{3} \\ k_1 &= f(x, y) \\ k_2 &= f(x + \tfrac{h}{2}, y + \tfrac{h}{2}k_1) \\ k_3 &= f(x + \tfrac{h}{2}, y + \tfrac{h}{2}k_2) \\ k_4 &= f(x + h, y + hk_3) \end{aligned}$$

"Tříosminová" formule:

$$\begin{gathered} w_1 = w_4 = \tfrac{1}{8}, \qquad w_2 = w_3 = \tfrac{3}{8} \\ \alpha_2 = \tfrac{1}{3}, \qquad \alpha_3 = \tfrac{2}{3}, \qquad \alpha_4 = 1 \\ \beta_{21} = \tfrac{1}{3}, \qquad \beta_{31} = -\tfrac{1}{3}, \qquad \beta_{32} = 1 \\ \beta_{41} = 1, \qquad \beta_{42} = -1, \qquad \beta_{43} = 1 \end{gathered}$$

Gillova formule:

$$\begin{gathered} w_1 = w_4 = \tfrac{1}{6}, \qquad w_2 = \tfrac{1}{3}\left(1 - \tfrac{1}{\sqrt{2}}\right), \qquad w_3 = \tfrac{1}{3}\left(1 + \tfrac{1}{\sqrt{2}}\right) \\ \alpha_2 = \alpha_3 = \tfrac{1}{2}, \qquad \alpha_4 = 1 \\ \beta_{21} = \tfrac{1}{2}, \qquad \beta_{31} = \tfrac{1}{2}(\sqrt{2} - 1), \qquad \beta_{32} = 1 - \tfrac{1}{\sqrt{2}} \\ \beta_{41} = 0, \qquad \beta_{42} = -\tfrac{1}{\sqrt{2}}, \qquad \beta_{43} = 1 + \tfrac{1}{\sqrt{2}} \end{gathered}$$

Gillova formule je sice o něco slo«itější, ale byla u ní provedena optimalizace konstanty N. Odvozením Rungeových-Kuttových metod řádu $p > 4$ se zabýval např. prof. Hu ¸a z Bratislavy. Pro $p > 4$ vyjde $P > p$ (např. pro $p = 6$ musíme vzít osm členů), metody jsou tedy slo«itější, a proto se příliš nepou«ívají.

2.5 Použitelnost odhadů chyb

Podívejme se nyní na efektivnost odhadu, který jsme získali. Uva«ujme např. Eulerovu metodu. Ukázali jsme, «e pokud je přesné řeení třídy C^2, je

$$|e_n| \leq NE_L(x_n - a)h,$$

kde N je polovina maxima absolutní hodnoty druhé derivace přesného řeení y na intervalu $[a, b]$. Tento odhad se dá o něco zlepit, uvědomíme-li si, «e pro odhad chyby v bodě x_n stačí řeit rovnici na $[a, x_n]$ (interval $(x_n, b]$ nemá «ádný vliv). Mů«eme tedy polo«it $b = x_n$, $N(x_n) = \frac{1}{2}\max_{[a,x_n]} |y''|$ a dostaneme zlepený odhad

$$|e_n| \leq N(x_n)E_L(x_n - a)h.$$

Uva«ujme dva příklady:

$$\begin{aligned} y_0 &= 1,\ y' = y \\ y(x) &= e^x \\ L &= 1 \\ N(x) &= \tfrac{1}{2}e^x \\ E_1(x_n) &= e^{x_n} - 1 \\ |e_n| &\leq \tfrac{1}{2}he^{x_n}(e^{x_n} - 1) \end{aligned} \qquad \begin{aligned} y_0 &= 1,\ y' = -y \\ y(x) &= e^{-x} \\ L &= 1 \\ N(x) &= \tfrac{1}{2} \\ E_1(x_n) &= e^{x_n} - 1 \\ |e_n| &\leq \tfrac{1}{2}h(e^{x_n} - 1) \end{aligned}$$

Řekněme, «e pomocí numerického řeení druhé úlohy chceme vypočítat e^{-5} s přesností 10^{-3}. Chceme tedy najít h tak, aby

$$\frac{1}{2}h(e^5 - 1) \leq 10^{-3},\ \text{takže}$$
$$h \leq \frac{2\ .\ 10^{-3}}{e^5 - 1} \doteq \frac{1}{73707}.$$

Experimentálně vak zjistíme, «e při volbě $h = 2^{-6} = \frac{1}{64}$ vyjde $e_n = -0,000261$. Vidíme tedy, «e odhad z věty je silně nadsazený; při jeho odvozování se v ka«dém kroku počítá s nejhorí mo«ností. Ve skutečnosti většinou dostáváme výsledky výrazně lepí.

Je ale třeba vzít v úvahu, «e konstanta N závisí na přesném řeení, které obvykle neznáme. Proto odhad z věty 8 nám dává pouze představu o chování chyby v závislosti na kroku metody a není prakticky pou«itelný. Proto hledáme odhady jiného typu, které mů«eme získat na základě vypočteného přibli«ného řeení a dat úlohy. Toto jsou tzv. *aposteriorní odhady.* Jedním takovým odhadem se budeme zabývat dále.

2.5.1 *Odhad chyby metodou poloviního kroku*

Víme, «e pokud přírůstková funkce Φ splňuje Lipschitzovu podmínku vzhledem k y a

$$|\Delta(x, y(x), h) - \Phi(x, y(x), h)| \leq Nh^p,$$

potom platí

$$|y_n - y(x_n)| \leq NE_L(x_n - a)h^p, \quad x_n \in [a, b].$$

Dá se navíc ukázat, «e pokud jsou funkce f a Φ dostatečně hladké, existuje funkce $e : [a, b] \to \mathbb{R}$ taková, «e

$$e_n = h^p e(x_n) + O(h^{p+1}). \tag{2.5.13}$$

Tento vztah chybu neodhaduje, ale aproximuje. Funkce e samozřejmě opět závisí na vlastnostech přesného řeení y.

Předpokládejme, «e k řeení uva«ované úlohy pou«ijeme postupně metodu s krokem $2h$ a h. Dostaneme uzly a odpovídající hodnoty přibli«ného řešení:

$$\begin{array}{ll} x_n^{(2h)} = a + n(2h) \ldots & y_n^{(2h)} \\ x_n^{(h)} = a + nh \quad \ldots & y_n^{(h)} \end{array}$$

Vidíme, «e $x_n^{(2h)} = x_{2n}^{(h)}$. V těchto bodech mů«eme porovnat hodnoty přibli«ných řeení a odpovídající chyby. Užitím vztahu (2.5.13) dostaneme

$$\begin{aligned} e_n^{(2h)} &= y_n^{(2h)} - y(x_n^{(2h)}) = (2h)^p e(x_n^{(2h)}) + O(h^{p+1}), \\ e_{2n}^{(h)} &= y_{2n}^{(h)} - y(x_{2n}^{(h)}) = h^p e(x_n^{(2h)}) + O(h^{p+1}) \end{aligned}$$

Odtud dostaneme postupně vztahy

$$\begin{aligned} y_n^{(2h)} - y_{2n}^{(h)} &= (2^p - 1)h^p e(x_n^{(2h)}) + O(h^{p+1}), \\ h^p e(x_n^{(2h)}) &= \frac{y_n^{(2h)} - y_{2n}^{(h)}}{2^p - 1} + O(h^{p+1}), \\ e_{2n}^{(h)} &= \frac{y_n^{(2h)} - y_{2n}^{(h)}}{2^p - 1} + O(h^{p+1}). \end{aligned}$$

Zanedbáním výrazu $O(h^{p+1})$ dostaneme *odhad chyby metodou poloviního kroku* v uzlech $x_n^{(2h)}$):

$$e_{2n}^{(h)} \approx \frac{y_n^{(2h)} - y_{2n}^{(h)}}{2^p - 1}. \tag{2.5.14}$$

Tento odhad jsme získali pouze na základě přibli«ného řeení (a případně dat úlohy). Mluvíme o *aposteriorním odhadu chyby.* Numerické experimenty ukazují, «e v řadě praktických aplikacích tento odhad dává spolehlivé výsledky.

Poznámka 10 U jednokrokových metod mů«eme v libovolném bodě změnit délku kroku. Postupujeme tak, «e výpočet provádíme s krokem $2h$ a h. Pokud zjistíme, «e odhad chyby v nějakém bodě překračuje povolenou mez, pokračujeme od tohoto bodu s polovičním krokem. Naopak, pokud je chyba výrazně ni«í ne« nae tolerance, mů«eme krok opět zvětit (abychom uetřili časovou náročnost výpočtu).

Existují dokonce *adaptivní metody,* u nich« se automaticky mění délka kroku a řád metody s cílem minimalizovat chybu a časovou náročnost.

2.5.2 *Zaokrouhlovací chyby*

Začněme jednoduchým příkladem. Uva«ujme soustavu dvou lineárních rovnic

$$\begin{pmatrix} 10^{-6} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Determinant matice soustavy je $10^{-6} - 1$, co« je dostatečně daleko od nuly. Matice soustavy je tedy regulární. Pou«ijeme-li Gaussovu eliminaci, dostaneme řešení

$$x_1 = (1 - x_2)\ 10^6, \qquad x_2 = \frac{2 - 10^6}{1 - 10^6}.$$

Přibli«ně máme $x_2 \doteq 0,999999$. Počítejme nyní na pět platných číslic. Pak ovem vyjde $x_2 = 1$, a odtud $x_1 = 0$. Při zkouce dostaneme v první rovnici $0 + 1 = 1$, ale ve druhé $0 + 1 = 1 \neq 2$, co« je příli velká nepřesnost. Pokud bychom ale rovnice prohodili, vyly by přibli«né hodnoty x_1 i x_2 rovny jedné, tedy správné (i zkouka by ve zvolené přesnosti vyla správně). Obecně je vhodné při řeení soustavy lineárních rovnic prohazovat rovnice tak, aby se na hlavní diagonále neobjevovala příli malá čísla. (Tento proces se nazývá *pivotace.*)

Zaokrouhlovací chyby u jednokrokových metod

Dosud jsme analyzovali metody pro výpočet přibli«ného řeení za předpokladu, «e vechny aritmetické operace probíhají přesně. Nyní se zamyslíme nad vlivem zaokrouhlovacích chyb. Zaokrouhlování modelujeme tak, «e ka«dému $\alpha \in \mathbb{R}$ přiřadíme jeho zaokrouhlenou hodnotu α^*. Pro jednoduchost předpokládejme, «e vstupní data jsou u« zaokrouhlená, tedy $a = a^*$, $h = h^*, x_n = x_n^*$, $\eta = \eta^*$. Přibli«né řeení počítané se zaokrouhlováním označme $\tilde{y}_n$. Můžeme psát

$$\begin{aligned} \tilde{y}_0 &= \eta = y_0, \\ \tilde{y}_{n+1} &= \tilde{y}_n + h\Phi(x_n, \tilde{y}_n, h) + \varepsilon_{n+1}, \end{aligned}$$

kde ε_{n+1} se nazývá *lokální zaokrouhlovací chyba.* Zaveďme jetě *akumulovanou zaokrouhlovací chybu* v uzlu x_n jako $r_n = \tilde{y}_n - y_n$.

Věta 11 *Nech, funkce $\Phi = \Phi(x, y, h) : [a, b] \times \mathbb{R} \times [0, h_0] \to \mathbb{R}$ splňuje Lipschitzovu podmínku vzhledem k y s konstantou $L > 0$ a nech, existuje konstanta $\varepsilon > 0$ taková, «e $|\varepsilon_k| \leq \varepsilon$ pokud $x_k, x_{k+1} \in [a, b]$. Potom*

$$|r_n| \leq \frac{\varepsilon}{h} E_L(x_n - a), \quad x_n \in [a, b],\ h \in (0, h_0).$$

Důkaz: Máme $r_0 = 0$ a

$$\begin{aligned} \tilde{y}_{n+1} &= \tilde{y}_n + h\Phi(x_n, \tilde{y}_n, h) + \varepsilon_{n+1}, \\ y_{n+1} &= y_n + h\Phi(x_n, y_n, h). \end{aligned}$$

Tudíž,

$$|r_{n+1}| \le |r_n + hLr_n| + |\varepsilon_{n+1}| \le |r_n|(1+hL) + \varepsilon.$$

Pou«ijeme-li nyní lemma 1, dostaneme

$$|r_n| \le \frac{(1+hL)^n - 1}{hL}\varepsilon \le \frac{e^{nhL}-1}{L}\frac{\varepsilon}{h} = E_L(x_n - a)\frac{\varepsilon}{h}, \quad x_n \in [a,b].$$

□

Poznámka 12 Velikost horního odhadu samozřejmě o ničem nesvědčí. Jedná se o *nejhorší možný scénář* (worst scenario), který mů«e nastat. Numerické experimenty však ukazují, «e v praxi je pro malá h opravdu $r_n \approx \frac{1}{h}$. To ve svém důsledku znamená, «e zmenením kroku sice zmeníme diskretizační chybu metody, ale zvýíme vliv zaokrouhlovacích chyb.

Hodilo by se tedy najít způsob, jak zjistit velikost zaokrouhlovací chyby. To je mo«né, pokud mů«eme úlohu současně řeit v jednoduché a vícenásobné přesnosti. Potom bude zaokrouhlovací chyba při vícenásobné přesnosti zanedbatelná ve srovnání s jednoduchou přesností a mů«eme tak získat aproximaci zaokrouhlovacích chyb. Převa«uje-li diskretizační chyba, je třeba krok zjemnit, převa«uje-li zaokrouhlovací chyba, je třeba krok zvětit.

2.6 Soustavy lineárních diferenčních rovnic

Označme $\mathbb{N} = \{1,2,3,\ldots\}$, $\mathbb{N}_0 = \{0,1,2,\ldots\}$.

Definice 6 Buď $k \in \mathbb{N}$, $F_n : \mathbb{R}^{k+1} \to \mathbb{R}$ (případně $F_n : \mathbb{C}^{k+1} \to \mathbb{C}$), $n \in \mathbb{N}_0$. Pak systém vztahů

$$F_n(y_n, \ldots, y_{n+k}) = 0, \qquad n \in \mathbb{N}_0 \tag{2.6.15}$$

nazýváme *soustavou diferenčních rovnic.* Řeením soustavy nazveme posloupnost $(y_n)_{n=0}^{\infty}$ splňující (2.6.15).

Příklad 6 Mějme soustavu rovnic ($k = 3$)

$$y_{n+3}^2 - \left(y_{n+2}^2 - y_{n+1}^2 + \sqrt[3]{y_n^2}\right)^4 = 0, \qquad n \in \mathbb{N}_0.$$

Vidíme, «e y_{n+3} lze vypočítat z předchozích tří členů. Vyjdeme-li z hodnot y_0, y_1, y_2, mů«eme postupně vypočítat y_3, y_4, ... Čísla y_0, y_1, y_2 se nazývají *počáteční podmínky.*

Cvičení 9 Kolik řeení má tato soustava pro pevně zvolené počáteční podmínky y_0, y_1, y_2?

Definice 7 Jsou-li funkce F_n lineární, mluvíme o *soustavě lineárních diferenčních rovnic.* Mů«eme ji psát ve tvaru

$$\sum_{\nu=0}^{k} \alpha_{n\nu} y_{n+\nu} = \gamma_n, \quad n \in \mathbb{N}_0. \tag{2.6.16}$$

Čísla γ_n se nazývají *pravé strany*. Jsou-li vechny pravé strany nulové, dostaneme homogenní soustavu. Soustavě (2.6.16) mů«eme v«dy přiřadit homogenní soustavu

$$\sum_{\nu=0}^{k} \alpha_{n\nu} y_{n+\nu} = 0, \quad n \in I\!N_0. \tag{2.6.17}$$

Lemma 2 *Uva«ujme soustavu (2.6.16) a k ní příslunou homogenní soustavu (2.6.17). Pak platí:*

(i) Mno«ina V vech řeení soustavy (2.6.17) je lineární prostor.

(ii) Jsou-li $\{y_n\}_{n=0}^{\infty}$ a $\{z_n\}_{n=0}^{\infty}$ řeení soustavy (2.6.16), je jejich rozdíl řeením soustavy (2.6.17).

(iii) Je-li $\{w_n\}_{n=0}^{\infty}$ řeením (2.6.16) a $\{y_n\}_n \in V$, pak $\{w_n + y_n\}_{n=0}^{\infty}$ je řeením soustavy (2.6.16).

(iv) je-li $\{w_n\}_{n=0}^{\infty}$ řeením soustavy (2.6.16), pak k libovolnému řeení $\{z_n\}_{n=0}^{\infty}$ soustavy (2.6.16) existuje právě jedno $\{y_n\}_n \in V$ takové, «e $z_n = y_n + w_n$; jinak řečeno, označíme-li W mno«inu řeení soustavy (2.6.16), můžeme psát $W = V + w$.

Důkaz pěnecháme čtenáři.

Definice 8 Řekneme, «e soustava (2.6.16) je *řádu k*, jestli«e všechny koeficienty α_{nk}, $n \in I\!N_0$, jsou nenulové.

Definice 9 Posloupnosti $\{y_n^{\mu}\}_{n \in I\!N_0} \in V$, $\mu = 1, \ldots, m$, nazveme *lineárně nezávislé*, jestli«e platí:

$$\sum_{\mu=1}^{m} a_\mu y_n^\mu = 0, \quad \forall n = 0, \ldots, k-1 \Rightarrow a_1 = a_2 = \ldots = a_m = 0.$$

Je zřejmé, «e tato podmínka je splněna, právě kdy« matice

$$\mathbb{M} = (y_n^\mu)_{\substack{n=0,\ldots,k-1 \\ \mu=1,\ldots,m}}$$

má hodnost m. Maximální počet lineárně nezávislých prvků z V je roven k.

Definice 10 Systém k prvků z V lineárně nezávislých nazveme *fundamentálním systémem* (FS) soustavy homogenních rovnic.

Věta 13 *Nech, $\{z_n^\mu\}_{n \in I\!N_0} \in V$, $\mu = 1, \ldots, k$, je FS řeení soustavy k-tého řádu a $\{y_n\}_{n \in I\!N_0} \in V$. Pak posloupnost $\{y_n\}_{n \in I\!N_0}$ je lineární kombinací posloupností $\{z_n^\mu\}_{n \in I\!N_0}$, $\mu = 1, \ldots, k$.*

Důkaz: Z definice FS plyne, «e existují konstanty $a_1, \ldots, a_k$ takové, «e

$$y_n = \sum_{\mu=1}^{k} a_\mu z_n^\mu \qquad \text{pro } n = 0, \ldots, k-1.$$

Označme

$$z_n = y_n - \sum_{\mu=1}^{k} a_\mu z_n^\mu, \qquad n \in \mathbb{N}_0.$$

Pak $\{z_n\}_{n\in\mathbb{N}_0} \in V$ a $z_0 = \ldots = z_{k-1} = 0$. Poněvadž $\alpha_{nk} \neq 0$ pro vechna $n \in \mathbb{N}_0$, nutně $z_n = 0$ pro vechna $n \in \mathbb{N}_0$ a tudíž,

$$y_n = \sum_{\mu=1}^{k} a_\mu z_n^\mu, \qquad \forall n \in \mathbb{N}_0.$$

□

Důsledek 14 *Prvky FS tvoří bazi v prostoru V.*

2.6.1 *Nalezení fundamentálního systému*

Definice 11 Soustavou tvaru

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = 0, \qquad n \in \mathbb{N}_0, \tag{2.6.18}$$

nazveme soustavou lineárních diferenčních rovnic s *konstantními koeficienty*. Tudíž,

$$\alpha_{n\nu} = \alpha_\nu, \qquad \forall \nu = 0, \ldots, k, \quad \forall n \in \mathbb{N}_0.$$

Tato soustava je řádu k, jestliže $\alpha_k \neq 0$.

Hledejme řeení této soustavy ve tvaru $y_n = \xi^n$, kde $\xi \in \mathbb{C}$. Dosazením do soustavy dostaneme

$$0 = \sum_{\nu=0}^{k} \alpha_\nu \xi^{n+\nu} = \xi^n \sum_{\nu=0}^{k} \alpha_\nu \xi^\nu, \quad n \in \mathbb{N}_0,$$

což je ekvivalentní s podmínkou

$$0 = \rho(\xi) = \sum_{\nu=0}^{k} \alpha_\nu \xi^\nu,$$

kterou nazýváme *charakteristickou rovnicí* a polynom ρ je tzv. *charakteristický polynom*.

Nechť soustava (2.6.18) je řádu k, tj. $\alpha_k \neq 0$. Pak ρ má stupeň k. Je zřejmé, že $\{y_n\}_{n\in\mathbb{N}_0} = \{\xi^n\}_{n\in\mathbb{N}_0}$ je řeením soustavy (2.6.18), právě když ξ je kořenem charakteristického polynomu ρ. Tento polynom má právě k kořenů, počítáme-li každý tolikrát, kolik činí jeho násobnost.

Rozlime dva případy

1) *Polynom ρ má právě k různých kořenů $\xi_1, \dots, \xi_k$.*
Pak lze sestrojit k řeení soustavy $\{y_n^\mu\}_{n\in\mathbb{N}_0}$, $\mu = 1, \dots, k$, kde $y_n^\mu = \xi_\mu^n$. Podle definice tvoří tato řeení FS, právě kdy« matice

$$\mathbb{M} = \begin{pmatrix} 1 & \xi_1 & \dots & \xi_1^{k-1} \\ 1 & \xi_2 & \dots & \xi_2^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \xi_k & \dots & \xi_k^{k-1} \end{pmatrix}$$

je regulární. Její determinant (známý jako Vandermondeův) má hodnotu

$$\det \mathbb{M} = \prod_{\substack{i<j \\ j=1,\dots,k}} (\xi_j - \xi_i) \neq 0.$$

Tudí« uva«ované posloupnosti tvoří FS.

2) *Obecný případ.*

Věta 15 *Nech, $\xi_1, \dots, \xi_m \in \mathbb{C}$ $(m \le k)$ jsou navzájem různé kořeny charakteristického polynomu ρ s násobnostmi $p_1, \dots, p_m$ (tak«e $k = \sum_{\mu=1}^m p_\mu$). Pak k posloupností s prvky*

$$(P) \qquad \begin{aligned} y_n^{\mu,1} &= \xi_\mu^n, \ n \in \mathbb{N}_0, \\ y_n^{\mu,2} &= n\xi_\mu^{n-1}, \ n \in \mathbb{N}_0, \\ &\vdots \\ y_n^{\mu,p_\mu} &= n(n-1)\dots(n-p_\mu+2)\xi_\mu^{n-p_\mu+1}, \ n \in \mathbb{N}_0, \end{aligned} \qquad \mu = 1, \dots, m$$

tvoří FS řeení soustavy (2.6.18) řádu k. (Je-li $\xi_\mu = 0$, pak klademe $0 \,.\, \xi_\mu^{-\nu} = 0$ pro $\nu > 0$.)

Důkaz: Doka«me, «e uvedené posloupnosti jsou řeení uva«ované soustavy. Jestli«e $\xi_\mu = 0$, pak má charakteristický polynom tvar

$$\rho(\xi) = \sum_{\nu=p_\mu}^{k} \alpha_\nu \xi^\nu$$

(tudí« $\alpha_0, \dots, \alpha_{p_\mu - 1} = 0$) a soustava (2.6.18) má tvar

$$\sum_{\nu=p_\mu}^{k} \alpha_\nu y_{n+\nu} = 0, \qquad n \in \mathbb{N}_0.$$

Dosazením snadno ověříme, «e posloupnosti (P) jsou řeeními, poněvadž mají nenulové prvky pouze na pozicích $n < p_\mu$.

Nechť $\xi_\mu \neq 0$ je kořen polynomu ρ s násobností p_μ. Pak ξ_μ je kořenem polynomu $f_n(\xi) = \xi^n \rho(\xi)$ násobnosti p_μ pro každé $n \in \mathbb{N}_0$. Tudíž, $p_\mu - 1$ derivací f_n je rovno nule v bodě $\xi = \xi_\mu$, takže

$$\begin{aligned} \textstyle\sum_{\nu=0}^{k} \alpha_\nu \xi_\mu^{n+\nu} &= 0, \ n \in \mathbb{N}_0, \\ \textstyle\sum_{\nu=0}^{k} \alpha_\nu (n+\nu) \xi_\mu^{n+\nu-1} &= 0, \ n \in \mathbb{N}_0, \\ &\vdots \\ \textstyle\sum_{\nu=0}^{k} \alpha_\nu (n+\nu)(n+\nu-1)\dots(n+\nu-p_\mu+2) \xi_\mu^{n+\nu-p_\mu+1} &= 0, \ n \in \mathbb{N}_0, \end{aligned}$$

Tzn., že posloupnosti (P) jsou řeeními soustavy (2.6.18).
Dokažme nyní, že řeení (P) jsou lineárně nezávislá, což je ekvivalentní s tím, že přísluná matice $\mathbb{M}$, její řádky jsou tvořeny prvními k prvky posloupností (P), je regulární. V n-tém sloupci této matice ($n = 0, \dots, k-1$) se nacházejí prvky

$$\xi_1^n,\ n\xi_1^{n-1}, \dots,\ n(n-1)\dots(n-p_1+2)\xi_1^{n-p_1+1}, \dots,$$

$$\xi_m^n,\ n\xi_m^{n-1}, \dots,\ n(n-1)\dots(n-p_m+2)\xi_m^{n-p_m+1}.$$

Matice $\mathbb{M}$ je regulární, právě když její sloupce jsou lineárně nezávislé k-rozměrné vektory. Nechť tomu tak není. Pak existují takové koeficienty a_n, $n = 0, \dots, k-1$ takové, že $\sum_{n=0}^{k-1} |a_n| > 0$ a

$$\begin{aligned} \textstyle\sum_{n=0}^{k-1} a_n \xi_\mu^n &= 0, \\ \textstyle\sum_{n=0}^{k-1} a_n\ n\xi_\mu^{n-1} &= 0, \\ &\vdots \\ \textstyle\sum_{n=0}^{k-1} a_n\ n(n-1)\dots(n-p_\mu+2)\xi_\mu^{n-p_\mu+1} &= 0, \end{aligned} \qquad \mu = 1, \dots, m,$$

což znamená, že polynom $\sum_{n=0}^{k-1} a_n \xi^n$ stupně $\leq k-1$ má aspoň m různých kořenů $\xi_1, \dots, \xi_m$ s násobnostmi $p_1, \dots, p_m$, což je spor, protože platí, že $\sum_{\mu=1}^{m} p_\mu = k$.

□

2.6.2 *Nalezení reálného fundamentálního systému*

Uvažujme rovnici

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = 0$$

s reálnými koeficienty α_j. Charakteristický polynom má pak také reálné koeficienty a jeho kořeny jsou

$$\xi_1, \dots, \xi_r \in \mathbb{R}$$

$$\gamma_1, \overline{\gamma_1}, \dots, \gamma_s, \overline{\gamma_s}, \in \mathbb{C} \setminus \mathbb{R}$$

(vechny kořeny opakujeme tolikrát, kolik činí jejich násobnost, tedy $r + 2s = k$). Fundamentální systém získaný z naí věty je tvořen posloupnostmi (vyplývá z chování aritmetických operací na $\mathbb{C}$)

$$\begin{array}{lll} \{y_n^\mu\}_{n=0}^\infty & \mu = 1, \dots, r & y_n^\mu \in I\!R \\ \{z_n^\mu\}_{n=0}^\infty & \mu = 1, \dots, s & \\ \{\overline{z_n^\mu}\}_{n=0}^\infty & \mu = 1, \dots, s & z_n^\mu \in \mathbb{C} \end{array}$$

Reálné posloupnosti z prvního řádku mů«eme převzít do naeho nového systému. Jako zbylých $2s$ posloupností vezmeme

$$\{\mathrm{Re}\ z_n^\mu\}_{n=0}^\infty, \qquad \{\mathrm{Im}\ z_n^\mu\}_{n=0}^\infty, \qquad \mu = 1, \dots, s.$$

Ze vztahů $\mathrm{Re}\ z = \frac{1}{2}(z+\bar{z})$, $\mathrm{Im}\ z = \frac{1}{2i}(z-\bar{z})$ vidíme snadno, «e nové posloupnosti jsou lineárními kombinacemi starých, jsou to tedy řeení. K důkazu lineární nezávislosti si stačí uvědomit, «e staré posloupnosti lze podobně jednodue vyjádřit pomocí nových pou«itím vztahů $z = \mathrm{Re}\ z + i\ \mathrm{Im}\ z$, $\bar{z} = \mathrm{Re}\ z - i\ \mathrm{Im}\ z$.

Poznámka 16 Kniha [H] obsahuje kapitolu věnovanou soustavám diferenčních rovnic. Je vidět, «e teorie soustav lineárních diferenčních rovnic s konstantními koeficienty se podobá teorii lineárních diferenciálních rovnic s konstantními koeficienty. Stejně tak existuje i metoda pro nalezení partikulárního řeení nehomogenní soustavy pomocí fundamentálního systému řeení přísluné soustavy homogenních rovnic. Tato metoda se podobá variaci konstant se nazývá *Duhamelův princip*, blí«e viz [H].

2.7 Vícekrokové metody

Opět budeme hledat přibli«né řeení úlohy

$$y' = f(x, y), \qquad y(a) = \eta \tag{2.7.19}$$

v uzlech $x_n = a + nh$. Na rozdíl od jednokrokových metod budeme tentokrát počítat hodnoty přibli«ného řeení y_n pomocí několika předchozích hodnot.

Příklad 7 V odstavci 2.2.3 jsme odvodili dvoukrokovou formuli

$$-y_{n+2} + 4y_{n+1} - 3y_n = 2hf(x_n, y_n).$$

To nás vede k zobecnění: obecnou k-krokovou metodou rozumíme rekurentní předpis typu

$$\sum_{\nu=0}^{k} a_\nu y_{n+\nu} = h \sum_{\nu=0}^{k} \beta_\nu f_{n+\nu}, \tag{2.7.20}$$

kde $f_j = f(x_j, y_j)$, za podmínek $\alpha_k \neq 0$ a $|\alpha_0| + |\beta_0| > 0$.

Pomocí předpisu (2.7.20) mů«eme počítat hodnoty přibli«ného řeení a« od y_k. Hodnoty $y_0(= \eta), y_1, \dots, y_{k-1}$ se nazývají počáteční podmínky pro k-krokovou

metodu. Hodnoty $y_1, \ldots, y_{k-1}$ vypočteme některou jednokrokovou metodou. Podle způsobu výpočtu je třeba rozliit dva případy.

V jednoduím případě $\beta_k = 0$ máme

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = h \sum_{\nu=0}^{k-1} \beta_\nu f_{n+\nu} = h \sum_{\nu=0}^{k-1} \beta_\nu f(x_{n+\nu}, y_{n+\nu})$$

$$y_{n+k} = \frac{1}{\alpha_k} \sum_{\nu=0}^{k-1} (h\beta_\nu f_{n+\nu} - \alpha_\nu y_{n+\nu}).$$

Proto«e lze hodnotu y_{n+k} přímo vypočítat, mluvíme o *explicitní metodě*. Hlavní výhodou těchto (např. oproti jednokrokovým metodám vyích řádů) je to, «e ka«dou z hodnot f_j stačí vypočítat jednou (v dalích krocích u« si ji pamatujeme). Na jednu hodnotu přibli«ného řeení tak připadá jedna vypočtená hodnota funkce f.

Ve slo«itějím případě $\beta_k \neq 0$ máme

$$y_{n+k} = \frac{\beta_k}{\alpha_k} h f(x_{n+k}, y_{n+k}) + \sum_{\nu=0}^{k-1} \left(\frac{\beta_\nu}{\alpha_k} h f(x_{n+\nu}, y_{n+\nu}) - \frac{\alpha_\nu}{\alpha_k} y_{n+\nu} \right).$$

Hodnotu y_{n+k} tedy přímo nevypočteme, ale dostaneme pro ni pouze rovnici. Mluvíme o *implicitní metodě*. Nae rovnice (obecně nelineární) pro y_{n+k} je tvaru $y_{n+k} = F(y_{n+k})$, kterou řeíme iterační metodou: zvolíme počáteční hodnotu y^0_{n+k} a pro $s \geq 0$ počítáme $y^{s+1}_{n+k} = F(y^s_{n+k})$. Pokud f je L-lipschitzovská v y, pak pro dostatečně malé h je F kontrakce (lipschitzovské zobrazení s konstantou mení ne« 1) a podle Banachovy věty o kontrakci je

$$y_{n+k} = \lim_{s \to \infty} y^s_{n+k}$$

pro libovolnou počáteční hodnotu y^0_{n+k}.

Pro praktické pou«ití je vak třeba odpovědět na dvě otázky: jak volit počáteční hodnotu y^0_{n+k} a kolik iterací provést. Pro volbu počáteční hodnoty je nejjednoduí vzít buď přímo předchozí hodnotu ($y^0_{n+k} = y_{n+k-1}$) nebo extrapolovat z předchozích dvou ($y^0_{n+k} = 2y_{n+k-1} - y_{n+k-2}$). Iterační proces ukončíme, pokud $|y^s_{n+k} - y^{s+1}_{n+k}| < \varepsilon$, kde ε je předepsaná přesnost.

Počáteční aproximaci y^0_{n+k} lze také získat pomocí explicitní metody

$$\sum_{\nu=0}^{k} \alpha^*_\nu y_{n+\nu} = h \sum_{\nu=0}^{k-1} \beta^*_\nu f_{n+\nu}.$$

Klademe tedy

$$y^0_{n+k} = \frac{1}{\alpha^*_k} \left(h \sum_{\nu=0}^{k-1} \beta^*_\nu f_{n+\nu} - \sum_{\nu=0}^{k-1} \alpha^*_\nu y_{n+\nu} \right). \tag{2.7.21}$$

Potom vypočteme několik aproximací y_{n+k}^s uva«ované implicitní metody a klademe $y_{n+k} := y_{n+k}^m$. Nejjednodušší volba je $m = 1$:

$$y_{n+k} = \left(h\beta_k f(x_{n+k}, y_{n+k}^0) + \sum_{\nu=0}^{k-1} \beta_\nu f_{n+\nu} - \sum_{\nu=0}^{k-1} \alpha_\nu y_{n+\nu} \right) / \alpha_k. \tag{2.7.22}$$

Z rovnice (2.7.21) vypočteme y_{n+k}^0, dosadíme do (2.7.22)) a vypočteme y_{n+k}. Tato metoda se nazývá *metoda prediktor-korektor*. Explicitní metoda (2.7.21) se nazývá *prediktorová formule*, implicitní metoda (2.7.22) je *korektorová formule*. Metody prediktor-korektor jsou nejefektivnější vícekrokové metody. V ka«dém kroku se počítají pouze dvě hodnoty funkce f.

Cvičení 10 Uva«ujme metodu danou následujícími formulemi:

$$\begin{aligned} y_{n+1} &= y_n + hf_n && \text{(prediktor)} \\ y_{n+1} &= y_n + \tfrac{1}{2}h(f_n + f_{n+1}) && \text{(korektor).} \end{aligned}$$

Zapite tuto metodu ve tvaru jednoho vzorce pro výpočet y_{n+1}. Uka«te, «e výsledkem je Rungeova-Kuttova metoda druhého řádu.

2.8 Některé vlastnosti obecných vícekrokových metod

Uva«ujme obecnou k-krokovou metodu

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = h \sum_{\nu=0}^{k} \beta_\nu f_{n+\nu}, \tag{2.8.23}$$

kde $\alpha_k \neq 0$, $|\alpha_0| + |\beta_0| > 0$, $n = 0, 1, \ldots$ (pokud $x_{n+k} \in [a, b]$). Na začátku výpočtu je třeba určit počáteční podmínky

$$y_0 = \eta, y_1, \ldots, y_{k-1}. \tag{2.8.24}$$

Ty se určí pomocí vhodné jednokrokové metody, takže lze psát $y_\mu = \eta_\mu(h)$, $\mu = 0, \ldots, k-1$.

Definice 12 Řekneme, «e k-kroková metoda (2.8.23) je *konvergentní*, jestli«e pro řeení $y : [a, b] \to I\!R$ úlohy $y' = f(x, y)$, $y(a) = \eta$ s libovolnou pravou stranou f, která je spojitá v $[a, b] \times I\!R$ a lipschitzovská vzhledem k y, a pro libovolné η platí následující tvrzení: jestli«e y_n je přibli«né řeení získané pomocí metody (2.8.23) s počátečními podmínkami (2.8.24) takovými, «e $\lim\limits_{h \to 0+} \eta_\mu(h) = \eta$, pro $\mu = 0, 1, \ldots, k-1$, pak platí

$$\forall x \in [a, b] : \quad \lim_{\substack{h \to 0+ \\ x_n = x}} y_n = y(x).$$

Definice 13 Říkáme, «e metoda (2.8.23) je *stabilní* (podle Dahlquista), jestli«e vechny kořeny polynomu $\rho(\xi) = \sum \alpha_\nu \xi^\nu$ jsou v absolutní hodnotě nejvýe rovny jedné a vechny kořeny, jejich« absolutní hodnota je rovna jedné, jsou jednoduché.

Definice 14 Řekneme, «e metoda (2.8.23) je *řádu* $p \geq 0$, jestli«e platí

$$\sum_{\nu=0}^{k} \alpha_\nu = 0$$

a pro $p > 0$ navíc

$$\sum_{\nu=0}^{k} \frac{\alpha_\nu \nu^s}{s!} = \sum_{\nu=0}^{k} \frac{\beta_\nu \nu^{s-1}}{(s-1)!} \qquad (s = 1, \ldots, p).$$

Má-li metoda (2.8.23) řád $p \geq 1$, říkáme, «e je *konsistentní*.

Věta 17 *Ka«dá konvergentní metoda je stabilní.*

Důkaz: Mějme konvergentní metodu (2.8.23). Přibli«né řeení y_n konverguje za podmínek uvedených v definici k přesnému řeení. Uva«ujme úlohu $y' = 0$, $y(0) = 0$: její přesné řeení je $y(x) = 0$ (tato úloha určitě podmínky splňuje). Pak máme

$$\sum \alpha_\nu y_{n+\nu} = 0 \qquad n = 0, 1, \ldots \tag{2.8.25}$$

$$y_\mu = \eta_\mu(h) \qquad \mu = 0, 1, \ldots, k-1$$

a za předpokladu $\eta_\mu(h) \to 0$ pro $h \to 0+$ víme, «e

$$\forall x > 0 : \quad \lim_{\substack{h \to 0+ \\ x_n = x}} y_n = 0.$$

Vztah (2.8.25) je soustava lineárních diferenčních rovnic s konstantními koeficienty a charakteristickým polynomem $\rho(\xi) = \sum \alpha_\nu \xi^\nu$. Nech¸ $\xi = re^{i\varphi}$ $(r \geq 0)$ je kořen polynomu ρ: pak posloupnost $\{\xi^n\}_{n=0}^{\infty}$ je řeením (2.8.25). Proto«e koeficienty α_ν jsou reálné, víme z teorie, «e i $\{w_n\}_{n=0}^{\infty}$ a $\{z_n\}_{n=0}^{\infty}$, kde $w_n = hr^n \cos n\varphi$ a $z_n = hr^n \sin n\varphi$, jsou řeení (2.8.25). Jsou to tedy přibli«ná řeení vyhovující počátečním podmínkám

$$hr^\mu \cos \mu\varphi \xrightarrow{h \to 0+} 0, \qquad hr^\mu \sin \mu\varphi \xrightarrow{h \to 0+} 0, \quad \mu = 0, \ldots, k-1.$$

Odtud plyne, «e pro ka«dé $x > 0$ platí:

$$\lim_{\substack{h \to 0+ \\ x_n = x}} z_n = \lim_{\substack{h \to 0+ \\ x_n = x}} w_n = 0 \iff \lim_{\substack{h \to 0+ \\ x_n = x}} |z_n + iw_n| = 0 \iff \lim_{\substack{h \to 0+ \\ x_n = x}} hr^n = 0 \iff$$

$$\iff x \lim_{n \to \infty} \frac{1}{n} r^n = 0 \iff |\xi| = r \leq 1.$$

Nyní mějme kořen $\xi = re^{i\varphi}$ násobnosti alespoň dvě. Potom je $\{n\xi^{n-1}\}_{n=0}^{\infty}$ řeením (2.8.25). Stejně jako výe vidíme, «e i z_n a w_n, kde $z_n = n\sqrt{h} r^n \cos n\varphi$

a $w_n = n\sqrt{h}r^n \sin n\varphi$ jsou řeení. Jsou to tedy přibli«ná řeení pro počáteční podmínky

$$\mu\sqrt{h}r^\mu \cos\mu\varphi \to 0, \qquad \mu\sqrt{h}r^\mu \sin\mu\varphi \to 0 \text{ for } h \to 0+, \quad \mu = 0, \ldots, k-1.$$

Odtud stejným způsobem jako v první části dostaneme

$$\sqrt{x} \lim_{n\to\infty} r^n\sqrt{n} = 0 \Longleftrightarrow |\xi| = r < 1.$$

□

Věta 18 *Je-li metoda (2.8.23) konvergentní, pak je také konsistentní, tj.*

$$\sum_{\nu=0}^{k} \alpha_\nu = 0, \qquad \sum_{\nu=0}^{k} \alpha_\nu\nu = \sum_{\nu=0}^{k} \beta_\nu.$$

Důkaz: Uva«ujme úlohu $y' = 0$ v $[0, b], b > 0$, tentokrát s počáteční podmínkou $y(0) = 1$. Pro přibli«né řeení opět platí (2.8.25), tj.

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+v} = 0 \qquad n = 0, 1, \ldots.$$

Pro libovolné $h > 0$ volme počáteční podmínky $y_0 = y_1 = \ldots = y_{k-1} = 1$. Zřejmě je splněna podmínka $y_\mu = \eta_\mu(h) = 1 \to 1$ pro $h \to 0+$. Metoda (2.8.23) je konvergentní, tedy

$$\forall x > 0 : \lim_{\substack{h\to 0+ \\ x_n = x}} y_n = 1, \tag{2.8.26}$$

přičem« limitu opět chápeme jako limitu pro $n \to \infty$, kde y_n je hodnota přibli«ného řeení s krokem $h = \frac{x}{n}$. Počáteční podmínky ani koeficienty rovnice (2.8.25) nezávisí na kroku h, nezávisí na něm tedy ani přibli«né řeení. Ze vztahu (2.8.26)

$$\lim_{n\to\infty} y_n = 1 \Longrightarrow \lim_{n\to\infty} y_{n+\nu} = 1 \qquad \nu = 0, \ldots, k.$$

Provedeme-li tedy limitní přechod v (2.8.25), dostaneme $\sum \alpha_\nu = 0$.

Pro pro důkaz druhé podmínky uva«ujme úlohu $y' = 1$, $y(0) = 0$, jejím« jednoznačným řeením je $y(x) = x$. Metoda (2.8.23) má nyní tvar

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = h\sum_{\nu=0}^{k} \beta_\nu \qquad n = 0, 1, \ldots. \tag{2.8.27}$$

Pro počáteční podmínky vyžadujeme, aby $y_\mu = \eta_\mu(h) \to 0$ pro $h \to 0+$, $\mu = 0, 1, \ldots, k-1$. Hledejme řeení (2.8.27) ve tvaru $y_n = Kx_n = Khn$. Po dosazení do (2.8.27) postupně dostaneme

$$Kh\sum_{\nu=0}^{k}\alpha_\nu(n+\nu)=h\sum_{\nu=0}^{k}\beta_\nu,$$
$$Kh\sum_{\nu=0}^{k}\alpha_\nu n+Kh\sum_{\nu=0}^{k}\alpha_\nu\nu=h\sum_{\nu=0}^{k}\beta_\nu,$$
$$K\sum_{\nu=0}^{k}\alpha_\nu\nu=\sum\beta_\nu.$$

Víme, «e $\sum_{\nu=0}^{k}\alpha_\nu=0$, takže $\rho(1)=0$. Z předcházející věty vyplývá, «e $\xi=1$ je jednoduchým kořenem polynomu ρ. Tudíž, $\rho'(1)\neq 0$, co« znamená, že $\sum_{\nu=0}^{k}\alpha_\nu\nu\neq 0$. Mů«eme tedy poslední rovnost touto sumou vydělit, načež dostáváme

$$K=\frac{\sum_{\nu=0}^{k}\beta_\nu}{\sum_{\nu=0}^{k}\alpha_\nu\nu}.$$

Z výše uvedených úvah plyne, že pro tuto hodnotu K je $y_n=Kx_n,\ n=0,1,\dots,$ řeením (2.8.27). Zvolíme-li počáteční podmínky metody $\eta_\mu(h)=K\mu h\to 0$ pro $h\to 0+$, bude $y_n=Kx_n$ přibli«ným řeením, získaným metodou (2.8.23). Protože předpokládáme, že je tato metoda konvergentní, máme pro ka«dé x

$$\lim_{\substack{h\to 0+\\ x_n=x}} y_n=y(x)=x\Longleftrightarrow K\lim_{n\to\infty}n\frac{x}{n}=Kx\Longleftrightarrow K=1\Longleftrightarrow\sum_{\nu=0}^{k}\alpha_\nu\nu=\sum_{\nu=0}^{k}\beta_\nu.$$

□

Příklad 8 Metoda $-y_{n+2}+4y_{n+1}-3y_n=2hf_n$ není stabilní, nebo¸ polynom $\rho(\xi)=-\xi^2+4\xi-3$ má kořeny 1 a 3.

Cvičení 11 Určete řád této metody.

Příklad 9 Metoda $y_{n+2}+4y_{n+1}-5y_n=h(4f_{n+1}+2f_n)$ má sice řád 3, ale není stabilní.

Jak vidíme, metody z příkladů 8 a 9 nejsou pou«itelné.

Význam stability metody

Aplikujeme-li k řeení obyčejných diferenciálních rovnic nestabilní metodu, zjistíme, «e pro rostoucí n hodnoty přibli«ného řeení oscilují okolo přesného řeení se zvětujícím se rozptylem.

Význam řádu metody

Definice řádu metody nijak nenaznačuje, k čemu je tento pojem dobrý. To vyplyne z následující věty. Mějme obecnou k-krokovou metodu (2.8.23), $y:[a,b]\to$

$I\!R$ buď přesné řeení úlohy (2.1.1). Dosaďme $y(x_n)$ místo y_n v (2.8.23). Dostaneme vztah

$$\sum_{\nu=0}^{k} \alpha_\nu y(x_{n+\nu}) = h \sum_{\nu=0}^{k} \beta_\nu f(x_{n+\nu}, y(x_{n+\nu})) + h\delta_n. \tag{2.8.28}$$

Veličinu δ_n nazveme *lokální relativní diskretizační chybou* v bodě x_{n+k}.

Věta 19 *Nech¸ metoda (2.8.23) je řádu p a nech¸ přesné řeení rovnice $y' = f(x,y)$ je třídy $C^{p+1}[a,b]$. Pak existují h_0, $K > 0$ tak, «e $|\delta_n| \le Kh^p$ pro vechna n taková, «e x_n, $x_{n+k} \in [a,b]$ a pro vechna $h < h_0$.*

Důkaz: Zvolme $h_0 > 0$ tak, že $kh_0 < b-a$. Uvažujme intervaly $[x_n, x_{n+\nu}] \subset [a,b]$ pro $\nu = 0, \dots, k$ a $n = 0, 1, \dots$. Z Taylorova vzorce plyne:

$$y(x_{n+\nu}) = y(x_n + \nu h) = y(x_n) + \sum_{s=1}^{p} \frac{y^{(s)}(x_n)}{s!} \nu^s h^s + R_{\nu,p+1}, \tag{2.8.29}$$

$$R_{\nu,p+1} = \frac{y^{(p+1)}(x_{n,\nu})}{(p+1)!} \nu^{p+1} h^{p+1} = O(h^{p+1}), \quad x_{n,\nu} \in [x_n, x_{n+\nu}].$$

Proto«e y je řeení rovnice, máme $f(x_{n+\nu}, y(x_{n+\nu})) = y'(x_{n+\nu})$; opět pou«ijme Taylorův vzorec:

$$hf(x_{n+\nu}, y(x_{n+\nu})) = hy'(x_n + \nu h) = \sum_{s=1}^{p} \frac{y^{(s)}(x_n)}{(s-1)!} \nu^{s-1} h^s + \tilde{R}_{\nu,p+1}, \tag{2.8.30}$$

$$\tilde{R}_{\nu,p+1} = \frac{y^{(p+1)}(\tilde{x}_{n,\nu})}{p!} \nu^p h^{p+1} = O(h^{p+1}), \quad \tilde{x}_{n,\nu} \in [x_n, x_{n+\nu}].$$

Zbytky R, $\tilde{R}$ jsou $O(h^{p+1})$, proto«e $y^{(p+1)}$ je spojitá, a tedy v $[a,b]$ omezená; označme Y maximum její absolutní hodnoty. Vztahy (2.8.29) a (2.8.30) dosadíme do (2.8.28) a upravíme:

$$\begin{aligned} h\delta_n = {} & y(x_n) \sum_{\nu=0}^{k} \alpha_\nu + \sum_{s=1}^{p} h^s y^{(s)}(x_n) \left(\sum_{\nu=0}^{k} \alpha_\nu \frac{\nu^s}{s!} - \sum_{\nu=0}^{k} \beta_\nu \frac{\nu^{s-1}}{(s-1)!} \right) \\ & + \sum_{\nu=0}^{k} \alpha_\nu R_{\nu,p+1} - \sum_{\nu=0}^{k} \beta_\nu \tilde{R}_{\nu,p+1}. \end{aligned}$$

Nyní z definice řádu metody a vyjádření pro $R_{\nu,p+1}$ a $\tilde{R}_{\nu,p+1}$ plyne, že

$$h\delta_n = h^{p+1} \left(\sum_{\nu=0}^{k} \alpha_\nu \nu^{p+1} \frac{y^{(p+1)}(x_{n,\nu})}{(p+1)!} - \sum_{\nu=0}^{k} \beta_\nu \nu^p \frac{y^{(p+1)}(\tilde{x}_{n,\nu})}{p!} \right)$$

Proto«e je $\left|y^{(p+1)}\right|$ na $[a,b]$ odhadnuta konstantou Y, dostáváme

$$|\delta_n| \le Yh^p \left(\sum \frac{|\alpha_\nu| \nu^{p+1}}{(p+1)!} + \sum \frac{|\beta_\nu| \nu^p}{p!} \right) = Kh^p$$

pro x_n, $x_{n+\nu} \in [a,b]$, $h \in (0, h_0)$. □

Shrnutí: Význam této věty je následující: je-li metoda řádu p a je-li přesné řeení dostatečně hladké, lokální relativní diskterizační chyba $\delta_n = O(h^p)$. Bez důkazu uveďme následující fundamentální výsledek:

Věta 20 *Vícekroková metoda je konvergentní, právě kdy« je stabilní a konsistentní.*

Definujme nyní opět chybu metody (akumulovanou diskretizační chybu) v uzlu x_n jako $e_n = y_n - y(x_n)$. Následující věta dává odhad této chyby.

Věta 21 *Nech¸ $y \in C^{p+1}([a,b])$ je přesným řeením úlohy*

$$y' = f(x,y) \ v\ [a,b], \quad y(a) = \eta,$$

kde $f \in C([a,b]) \times I\!R)$ splňuje Lipschitzovu podmínku vzhledem k proměnné y konstantou $L > 0$ a nech¸ $\{y_n\}_{x_n \in [a,b]}$ je přibli«né řeení vypočtené pomocí vícekrokové stabilní metody řádu $p \geq 1$. Označme

$$\delta = \max_{\mu=0,\ldots,k-1} |y(x_\mu) - y_\mu| .$$

Pak existují konstanty $h_1 > 0$, $\tilde{N} > 0$, $\vartheta > 0$ takové, «e pro chybu metody $e_n = y_n - y(x_n)$ platí odhad

$$|e_n| \leq e^{(x_n - a)\vartheta}\delta + E_\vartheta(x_n - a)\tilde{N}h^p$$

$$\forall x_n \in [a,b], \quad \forall h \in (0, h_1).$$

Úplný důkaz je uveden ve skriptech [1] věta 1.12.4. Zde se omezíme na jednoduí případ tzv. *silně stabilních metod.*

Definice 15 Vícekroková metoda je *silně stabilní*, jestli«e má tvar

$$y_{n+k} - y_{n+k-1} = h\sum_{\nu=0}^{k} \beta_\nu f_{n+\nu}. \tag{2.8.31}$$

To znamená, «e polynom ρ má pouze kořeny 0 a 1 a kořen 1 je jednonásobný.

Nyní doká«eme předchozí větu o odhadu chyby pro případ silně stabilních metod.

Důkaz: Přibli«né řeení splňuje vztahy

$$y_{n+k} - y_{n+k-1} = h\sum_{\nu=0}^{k} \beta_\nu f(x_{n+\nu}, y_{n+\nu}), \quad x_n, x_{n+k} \in [a,b].$$

Pokud je metoda (2.8.31) řádu p a přesné řeení je $y \in C^{p+1}([a,b])$, pak

$$y(x_{n+k}) - y(x_{n+k-1}) = h\sum_{\nu=0}^{k} \beta_\nu f(x_{n+\nu}, y(x_{n+\nu})) + h\delta_n, \quad x_n, x_{n+k} \in [a,b].$$

a

$$|\delta_n| \leq Nh^p, \quad rmpokud\ x_n, x_{n+k} \in [a,b], \quad h \in (0,h_0).$$

Označíme-li $e_{n+\nu} = y_{n+\nu} - y(x_{n+\nu})$ chybu metody, z předchozích vztahů dostaneme

$$e_{n+k} - e_{n+k-1} = h\sum_{\nu=0}^{k} \beta_\nu \left\{f(x_{n+\nu}, y_{n+\nu}) - f(x_{n+\nu}, y(x_{n+\nu}))\right\} - h\delta_n.$$

Poněvad« funkce f splňuje Lipschitzovu podmínku, máme

$$|e_{n+k}| \leq |e_{n+k-1}| + hL\sum_{\nu=0}^{k} |\beta_\nu||e_{n+\nu}| + h|\delta_n|$$

a tedy

$$|e_{n+k}|(1 - hL|\beta_k|) \leq |e_{n+k-1}|(1 + hL|\beta_{k-1}|) + hL\sum_{\nu=0}^{k-2} |\beta_\nu||e_{n+\nu}| + Nh^{p+1}.$$

Nech¸ $h_1 > 0$ je takové, «e

$$1 - h_1 L|\beta_k| \geq \frac{1}{2}.$$

Potom pro $h \in (0,h_1)$ platí $1 - hL|\beta_k| \geq 1 - h_1 L|\beta_k| \geq \frac{1}{2}$. Odtud plyne, že

$$\frac{1}{1 - hL|\beta_k|} \leq 2$$

a

$$\frac{1 + hL|\beta_{k-1}|}{1 - hL|\beta_k|} = \frac{1 - hL|\beta_k| + hL(|\beta_k| + |\beta_{k-1}|}{1 - hL|\beta_k|} \leq 1 + 2hL(|\beta_k| + |\beta_{k-1}|).$$

Tudí«,

$$|e_{n+k}| \leq |e_{n+k-1}| + h\sum_{\nu=0}^{k-1} \vartheta_\nu |e_{n+\nu}| + \tilde{\vartheta}, \tag{2.8.32}$$

kde

$$\vartheta_{k-1} = 2L(|\beta_k| + |\beta_{k-1}|), \quad \vartheta_\nu = 2L|\beta_\nu|,\ \nu = 0,\ldots,k-2, \tilde{\vartheta} = 2Nh^{p+1}.$$

Nyní pou«ijeme následující výsledek.

Lemma 3 *Nechť* $u_n \geq 0$ *pro* $n = 0, \ldots, K$, $\vartheta_\nu \geq 0$ *pro* $\nu = 0, \ldots, k-1$, $\tilde{\vartheta} \geq 0$, $\delta \geq 0$, $h > 0$, $u_0, \ldots, u_{k-1} \leq \delta$,

$$u_{n+k} \leq u_{n+k-1} + h \sum_{\nu=0}^{k-1} \vartheta_\nu u_{n+\nu} + \tilde{\vartheta} \tag{2.8.33}$$

pro $n = 0, \ldots, K-k$ *a nechť* $\vartheta = \sum_{\nu=0}^{k-1} \vartheta_\nu$. *Pak*

$$u_n \leq e^{nh\vartheta}\delta + \frac{e^{nh\vartheta} - 1}{\vartheta h}\tilde{\vartheta}, \quad n = 0, \ldots, K.$$

Důkaz:

1) Nechť

$$\xi_{n+1} = (1 + h\vartheta)\xi_n + \tilde{\vartheta}, \quad n = 0, \ldots, K-1,$$
$$\xi_0 = \delta.$$

Potom podle lemmatu 1

$$\xi_n \leq e^{nh\vartheta}\delta + \frac{e^{nh\vartheta} - 1}{\vartheta h}\tilde{\vartheta}, \quad n = 0, \ldots, K.$$

2) Dokažme indukcí, že $u_n \leq \xi_n$, $n = 0, \ldots, K$. Máme

$$u_0, \ldots, u_{k-1} \leq \delta = \xi_0 \leq \xi_1 \leq \ldots \leq \xi_{k-1} \leq \xi_k \leq \ldots$$

Nechť $u_j \leq \xi_j$ pro $j = 0, \ldots, n+k-1$, $n \geq 0$. Potom z (2.8.33) plyne, že

$$\begin{aligned}
u_{n+k} &\leq u_{n+k-1} + h\vartheta_{k-1}u_{n+k-1} + \ldots + h\vartheta_0 u_n + \tilde{\vartheta} \\
&= u_{n+k-1} + h\sum_{\nu=0}^{k-1} \vartheta_\nu u_{n+\nu} + \tilde{\vartheta} \\
&\leq \xi_{n+k-1} + h\vartheta_{k-1}\xi_{n+k-1} + \ldots + h\vartheta_0\xi_n + \tilde{\vartheta} \\
&= \xi_{n+k-1} + h\sum_{\nu=0}^{k-1} \vartheta_\nu \xi_{n+\nu} + \tilde{\vartheta} \\
&\leq \xi_{n+k-1}\left(1 + h\sum_{\nu=0}^{k-1} \vartheta_\nu\right) + \tilde{\vartheta} \\
&= \xi_{n+k-1}(1 + h\vartheta) + \tilde{\vartheta} = \xi_{n+k}.
\end{aligned}$$

3) Použijeme-li výsledek z 1), dostaneme

$$u_n \leq \xi_n \leq e^{nh\vartheta}\delta + \frac{e^{nh\vartheta} - 1}{\vartheta h}\tilde{\vartheta}, \quad n = 0, \ldots, K,$$

což jsme chtěli dokázat.

□

Nyní můžeme pokračovat v důkazu věty tak, že budeme aplikovat lemma na nerovnosti (2.8.32). Vzhledem k tomu, «e $nh = x_n - a$ a $\tilde{\vartheta} = \tilde{N}h^{p+1}$, kde $\tilde{N} = 2N$, snadno zjistíme, «e

$$|e_n| \le e^{(x_n-a)\vartheta}\delta + \frac{e^{(x_n-a)\vartheta}-1}{\vartheta h}\tilde{N}h^{p+1} =$$

$$= e^{(x_n-a)\vartheta}\delta + E_\vartheta(x_n-a)\tilde{N}h^p, \quad x_n \in [a,b].$$

□

Podobným způsobem lze odvodit odhad pro akumulovanou zaokrouhlovací chybu. Stejně jako u jednokrokových metod budeme symbolem $\tilde{y}_n$ značit přibli«né řeení počítané se zaokrouhlováním. Pak lze psát

$$\tilde{y}_{n+k} - \tilde{y}_{n+k-1} = h\sum_{\nu=0}^{k}\beta_\nu f(x_{n+\nu}, \tilde{y}_{n+\nu}) + \varepsilon_n, \quad x_n, x_{n+k} \in [a,b],$$

kde $|\varepsilon_n| \le \varepsilon$ pro $x_n, x_{n+k} \in [a,b]$. Předpokládejme, «e $\tilde{y}_\mu = y_\mu$, $\mu = 0,\dots,k-1$. Přibli«né řeení počítané přesně bez zaokrouhlování je dáno předpisem

$$y_{n+k} - y_{n+k-1} = h\sum_{\nu=0}^{k}\beta_\nu f(x_{n+\nu}, y_{n+\nu}), \quad x_n, x_{n+k} \in [a,b].$$

Nyní lze aplikovat postup jako v důkazu předchozí věty, kde $\delta = 0$ a $\tilde{\vartheta} = 2\varepsilon$. Dostaneme pak odhad akumulované zaokrouhlovací chyby $r_n = \tilde{y}_n - y_n$ ve tvaru

$$|r_n| \le 2E_\vartheta(x_n-a)\frac{\varepsilon}{h}, \quad x_n \in [a,b].$$

Tento odhad nám dává *nejhorí scénář* pro chování chyby r_n v závislosti na h. Numerické experimenty bohu«el ukazují, «e pro $h \to 0+$ zaokrouhlovací chyby mohou růst tak, «e znehodnotí výpočet.

2.9 Odvození některých vícekrokových metod

2.9.1 *Interpolační polynom a zpětné diference*

Nech˛ $J \subset \mathbb{R}$ je interval, $z : J \to \mathbb{R}$ a nech˛ $x_0,\dots,x_q \in J$ $(q \ge 1)$ jsou navzájem různé body. Pak existuje právě jeden interpolační polynom P takový, «e

a) stupeň $P \le q$,

b) $P(x_i) = z_i := z(x_i), \quad i = 0,\dots,q.$

Polynom P lze napsat v Lagrangeově vyjádření

$$P(x) = \sum_{i=0}^{q} \frac{(x-x_0)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_q)}{(x_i-x_0)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_q)} z(x_i).$$

Pokud $h > 0$ a $x_i = x_0 + ih$ pro $i = p, p-1, \dots, p-q$, pak lze vyjádřit P pomocí *zpětných diferencí* funkce z, definovaných následujícím způsobem:

$$\nabla^0 z_i = z_i \qquad \text{(diference nultého řádu)},$$

$$\nabla z_i = z_i - z_{i-1} \qquad \text{(diference prvního řádu)},$$

dále indukcí

$$\nabla^r z_i = \nabla(\nabla^{r-1} z_i) \qquad \text{(diference } r\text{-tého řádu)},$$

pokud mají výrazy vpravo smysl. Platí:

$$\nabla^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) = z_i - 2z_{i-1} + z_{i-2}.$$

Obecně lze dokázat vztah

$$\nabla^r z_i = \sum_{m=0}^{r} (-1)^m \binom{r}{m} z_{i-m}, \qquad r = 0, 1, \dots$$

Existuje i inverzní formule

$$z_{i-r} = \sum_{m=0}^{r} (-1)^m \binom{r}{m} \nabla^m z_i, \qquad r = 0, 1, \dots$$

Lemma 4 *Interpolační polynom P k funkci z v bodech $x_i = x_0 + ih$, kde i nabývá hodnot $p, p-1, \dots, p-q$, lze napsat ve tvaru*

$$P(x) = \sum_{m=0}^{q} (-1)^m \binom{-s}{m} \nabla^m z_p, \tag{2.9.34}$$

kde $s = s(x) = (x - x_p)/h$ a

$$\binom{-s}{m} = \frac{(-s)(-s-1)\dots(-s-m+1)}{m!}.$$

Důkaz: Označme symbolem $\tilde{P}$ výraz na pravé straně vztahu (2.9.34). Je zřejmé, že $\tilde{P}$ je polynom stupně $\le q$. Je-li $x = x_{p-r}$, potom $s = -r$ a $\binom{r}{m} = 0$ pro $m > r$. Pak podle inverzní formule

$$P(x_{p-r}) = \sum_{m=0}^{q} (-1)^m \binom{r}{m} \nabla^m z_p = \sum_{m=0}^{r} (-1)^m \binom{r}{m} \nabla^m z_p = z_{p-r}.$$

Tudíž, $\tilde{P} = P$. □

2.9.2 *Metody zalo«ené na numerické integraci*

Nech ¸ $\tau > 0$. Je-li funkce $y = y(x)$ přesné řeení rovnice $y' = f(x, y)$, potom

$$y(\overline{x} + \tau) - y(\overline{x}) = \int_{\overline{x}}^{\overline{x}+\tau} f(x, y(x))dx,$$

pro $\overline{x}, \overline{x}+\tau \in [a, b]$. Nyní nahradíme funkci $f(x, y(x))$ interpolačním polynomem majícím v uzlech x_n hodnoty $f_n = f(x_n, y_n)$, kde y_n bylo vypočteno nebo má být právě vypočteno. Za interpolační body vezměme body $x_p, x_{p-1}, \ldots, x_{p-q}$. Uva«ovaný interpolační polynom má tvar

$$P(x) = \sum_{m=0}^{q} (-1)^m \binom{-s}{m} \nabla^m f_p, \qquad s = \frac{x - x_p}{h}.$$

α) Zvolme $\overline{x} = x_p$, $\tau = h$ a aproximujme $y(\overline{x}) \approx y_p$, $y(\overline{x} + \tau) \approx y_{p+1}$, $f(x, y(x)) \approx P(x)$. Dostaneme formuli

$$y_{p+1} - y_p = \int_{x_p}^{x_{p+1}} P(x)\ dx = h \sum_{m=0}^{q} \gamma_m \nabla^m f_p,$$

kde

$$\gamma_m = (-1)^m \frac{1}{h} \int_{x_p}^{x_{p+1}} \binom{-s(x)}{m}\ dx = (-1)^m \int_0^1 \binom{-s}{m}\ ds.$$

Takto získané metody se nazývají *Adamsovy-Bashforthovy metody.* Pro zvolené q dostaneme $(q+1)$-krokovou metodu: k výpočtu y_{p+1} pou«íváme hodnoty f v $x_{p-q}, \ldots, x_p$, jedná se tedy o metodu explicitní. Charakteristický polynom $\rho(\xi) = \xi^{q+1} - \xi^q = \xi^q(\xi - 1)$ má q-násobný kořen 0 a jednoduchý kořen 1, metody jsou tedy silně stabilní. Většinou mají tyto metody řád $q + 1$.

β) Polo«me $\overline{x} = x_{p-1}$, $\tau = h$. Podobným postupem jako výe dostaneme schéma

$$y_p - y_{p-1} = \int_{x_{p-1}}^{x_p} P(x)\,\mathrm{d}x = \int_{x_{p-1}}^{x_p} \sum_{m=0}^{p} (-1)^m \binom{-s}{m} \nabla^m f_p\,\mathrm{d}x =$$
$$= h \sum_{m=0}^{q} \gamma_m^* \nabla^m f_p,$$

$$\text{kde } \gamma_m^* = \frac{(-1)^m}{h} \int_{x_{p-1}}^{x_p} x_p \binom{-s(x)}{m} \mathrm{d}x = (-1)^m \int_{-1}^{0} \binom{-s}{m} \mathrm{d}s.$$

V tomto případě dostáváme implicitní q-krokovou metodu. Tyto metody se nazývají *Adamsovy-Moultonovy.*

Hodnoty zpětných diferencí se počítají poměrně snadno, ale přesto mů«eme nae metody přepsat do tvaru, kdy místo hodnot $\nabla^m f_p$ pou«íváme přímo hodnoty f_{p-r} podle vztahu

$$\nabla^m f_p = \sum_{\rho=0}^{m} (-1)^\rho \binom{m}{\rho} f_{p-\rho}.$$

Dostaneme pak např. Adams-Bashforthovu metodu ve tvaru

$$y_{p+1} - y_p = \sum_{m=0}^{q} \gamma_m \nabla^m f_p = h \sum_{m=0}^{q} \gamma_m \sum_{\rho=0}^{m} (-1)^\rho \binom{m}{\rho} f_{p-\rho} =$$
$$= h \sum_{\rho=0}^{q} f_{p-\rho} \left((-1)^\rho \sum_{m=\rho}^{q} \binom{m}{\rho} \gamma_m \right).$$

2.9.3 *Příklady některých metod*

V následující tabulce jsou uvedeny příklady některých metod, které takto dostaneme.

	metoda	k	p
1.	$y_{n+1} = y_n + h f_n$	1	1
2.	$y_{n+2} = y_{n+1} + \frac{1}{2}h(3f_{n+1} - f_n)$	2	2
3.	$y_{n+3} = y_{n+2} + \frac{h}{12}(23f_{n+2} - 16f_{n+1} + 5f_n)$	3	3
6.	$y_{n+1} = y_n + h f_{n+1}$	1	1
7.	$y_{n+1} = y_n + \frac{h}{2}(f_{n+1} + f_n)$	1	2
8.	$y_{n+2} = y_{n+1} + \frac{h}{12}(5f_{n+2} + 8f_{n+1} - f_n)$	2	3

Metody prediktor-korektor

Budeme-li chtít vyu«ít uvedené metody pro pou«ití v metodách prediktor-korektor, je výhodné vzít kombinace 1-7 a 2-8. Z teoretického rozboru toti« vyplývá, «e v případě, kdy jako prediktor pou«ijeme metodu řádu o jedna meního ne« je řád korektoru, celková metoda bude mít stejný řád jako korektor. Dostaneme tedy jednokrokovou metodu řádu 2 a dvoukrokovou metodu řádu 3.

2.10 Metoda sítí pro řeení parciálních diferenciálních rovnic

Uveďme na závěr jednoduchý příklad numerické aproximace parciálních diferenciálních rovnic. Představme si pru«nou membránu, na okraji upevněnou, na ni« působí kolmo nějaká síla. Modelem bude oblast (tj. otevřená souvislá mno«ina) $\Omega \subset \mathbb{R}^2$, sílu budeme reprezentovat hustotou jejího rozlo«ení, tedy funkcí $f : \Omega \to \mathbb{R}$ a výchylku membrány hledanou funkcí $u : \bar{\Omega} \to \mathbb{R}$. Za předpokladu, «e jsme dosáhli stabilního stavu (a výchylka je malá) bude u splňovat rovnici

$$-\Delta u = f$$

, kde Δ je takzvaný *Laplaceův operátor* definovaný vztahem

$$\Delta u = \sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2}.$$

Upevnění na okrajích je vyjádřeno okrajovou podmínkou

$$u|\partial\Omega = 0$$

.

Řeme uvažovanou úlohu přibli«ně v případě $\Omega = (0,1)\times(0,1)$. Za tím účelem provedeme diskretizaci této úlohy. V mno«ině $\bar{\Omega}$ sestrojíme sí¸ tvořenou uzly (x_i, y_j) kde $x_i = y_i = ih$ a . Hodnoty u, f v uzlu (x_i, y_j) označme u_{ij} a f_{ij}. Okrajová podmínka dává $u_{0,j} = u_{n,j} = u_{i,0} = u_{i,n} = 0$. Parciální derivace druhého řádu funkce u aproximujeme pomocí konečných diferencí:

$$\frac{\partial^2 u}{\partial x^2}(x_i, y_j) \approx \frac{1}{h^2}\left(u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j)\right)$$
$$\frac{\partial^2 u}{\partial y^2}(x_i, y_j) \approx \frac{1}{h^2}\left(u(x_i, y_{j-1}) - 2u(x_i, y_j) + u(x_i, y_{j+1})\right)$$

Dosadíme-li tuto diskretizaci do rovnice, dostaneme soustavu lineárních rovnic tvaru

$$-u_{i,j-1} - u_{i,j+1} + 4u_{i,j} - u_{i-1,j} - u_{i+1,j} = h^2 f_{i,j}.$$

Uspořádáme-li neznámé do jednoho sloupcového vektoru $\boldsymbol{u} = (u_{1,1}, \ldots, u_{1,n-1}, u_{2,1}, \ldots, u_{n-1,n-1})^T$ a stejným způsobem pravé strany do vektoru $\boldsymbol{f}$, dostaneme soustavu $A\boldsymbol{u} = \boldsymbol{f}$, kde matici A lze psát rozdělenou na pole ve tvaru

$$A = \begin{pmatrix} B & C & 0 & 0 & \ldots & 0 \\ C & B & C & 0 & \ldots & 0 \\ 0 & C & B & C & \ldots & 0 \\ \vdots & & & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & B \end{pmatrix}, \qquad B = \begin{pmatrix} 4 & -1 & 0 & 0 & \ldots & 0 \\ -1 & 4 & -1 & 0 & \ldots & 0 \\ 0 & -1 & 4 & -1 & \ldots & 0 \\ \vdots & & & & \ddots & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 \end{pmatrix}$$

a C je $-E$, kde E je jednotková matice. Matice A se skládá z $(n-1)\times(n-1)$ bloků typu $(n-1)\times(n-1)$.?????? Matici s takovouto strukturou se říká *blokově třídiagonální matice*. Struktura této matice samozřejmě závisí na očíslování uzlů. Popsaná technika se nazývá *metoda sítí* nebo též *metoda konečných diferencí* .

3

NUMERICKÉ METODY OPTIMALIZACE

V této kapitole se věnujeme výkladu elementů konvexní analýzy v $\mathbb{R}^n$ a základním numerickým metodám pro hledání minima (lokálního minima) funkcí více proměnných. Na začátku uveďme dva příklady.

Příklad 10 *Prokládání křivek naměřenými daty:* Máme zadány hodnoty x_i, y_i $(i = 1, \ldots, r, x_1 < \ldots < x_r)$. Hledáme funkci $f = f(\alpha_1, \ldots, \alpha_n, x)$ závislou na konečně mnoha parametrech $\alpha_1, \ldots, \alpha_n$ tak, aby její hodnoty v x_i byly co nejblí«e k y_i. Pou«ijeme-li metodu nejmeních čtverců, je třeba minimalizovat funkci

$$J(\alpha_1, \ldots, \alpha_n) = \sum_{i=1}^{r} (f(\alpha_1, \ldots, \alpha_n, x_i) - y_i)^2.$$

Příklad 11 *Optimalizace tvaru radiátoru ústředního topení*

Sna«íme se najít co nejvýhodnější tvar průřezu «ebra. Tento tvar mů«eme reprezentovat uzavřenou křivkou okolo počátku; proto«e chceme, aby tato křivka byla symetrická, mů«eme ji reprezentovat pomocí nezáporné funkce $p : [0, b] \to [0, +\infty)$. Nae křivka pak vznikne doplněním grafu této funkce o části souměrně s ním sdru«ené podle počátku a obou os. Rozlo«ení teploty T bude splňovat rovnici pro vedení tepla ve tvaru[3]

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}T}{\mathrm{d}x}\right) = a\sqrt{1 + \left(\frac{\mathrm{d}p}{\mathrm{d}x}\right)^2} \cdot T$$

s okrajovými podmínkami $T(0) = T_0$, $T(b) = T_b$, kde a je konstanta závisející na vlastnostech materiálu. Celkové mno«ství tepla uvolněného do okolí je

$$I = 2\int_0^b kT\mathrm{d}x,$$

kde k je opět jakýsi koeficient. Úkolem je najít funkci p takovou, aby hodnota funkcionálu $I(p)$ byla maximální, tedy aby hodnota $F(p) = -I(p)$ byla minimální. Přesněji řečeno, hledáme funkci $p^* \in C^1[0, p]$ tak, aby

$$F(p^*) = \min_{p \in C^1[0,p]} F(p)$$

a zároveň $p(b) = 0$, $p(0) = K > 0$ (K je zadáno) a $p > 0$ na $[0, b)$. Označíme-li U mno«inu vech takových funkcí (bez po«adavku minimality), hledáme minimum

[3]tento příklad má pouze ilustrovat, jak se slo«itější praktický problém převádí na standardní situaci hledání minima funkce několika proměnných

F na U. Takto formulovaná úloha spadá do odvětví matematiky zvaného *optimální řízení*. Pro pou«ití numerických metod se musíme jetě omezit na nějakou mno«inu $\tilde{U} \subset U$ funkcí určenou konečně mnoha parametry $u_1, \ldots, u_n$. Ka«dé takové n-tici parametrů přiřadíme funkci $p \in U$, k té spočítáme teplotu T, hodnoty $I(p)$ a $F(p)$ a polo«íme $\tilde{F}(u_1, \ldots, u_n) := F(p)$. Tím u« jsme úlohu převedli na problém minimalizace funkce $\tilde{F} : \mathbb{R}^n \to \mathbb{R}$.

V obou případech jsme dospěli k úloze minimalizovat reálnou funkci n proměnných. Je vidět, «e tato úloha má řadu praktických aplikací, a proto byly vyvinuty různé algoritmy pro její řeení.

3.1 Základy konvexní analýzy

V této části uvedeme některé základní pojmy a výsledky konvexní analýzy, které se vyu«ívají při numerickém hledání extrémů. V dalím bude U podmno«ina $\mathbb{R}^n$ a J zobrazení U do $\mathbb{R}$.

Definice 16 Hodnotu $J(\bar{u})$, $\bar{u} \in U$ nazveme (absolutním, globálním) *minimem* funkce J na U, jestli«e $J(u) \geqq J(\bar{u})$ platí pro vechna $u \in U$. Bod $\bar{u}$ nazýváme *bod minima*, píeme $J(\bar{u}) = \min_U J$ nebo $\bar{u} = \arg\min_U J$.

Hodnotu $J(\bar{u})$, $\bar{u} \in U$ nazveme *lokálním minimem* funkce J, jestli«e existuje okolí $V(\bar{u})$ takové, «e $J(\bar{u})$ je (globálním) minimem na J na $U \cap V(\bar{u})$. Bod $\bar{u}$ se pak nazývá *bodem lokálního minima.*

Věta 22 *Buď $U \subset \mathbb{R}^n$ omezená, uzavřená mno«ina a $J : U \to \mathbb{R}$ spojitá funkce. Potom J nabývá na U minima.*

Definice 17 Nech¸ $U \subset \mathbb{R}^n$ je neomezená mno«ina a $J : U \to \mathbb{R}$. Řekneme, «e funkce J je *koercivní*, jestli«e

$$\lim_{\substack{\|u\| \to \infty \\ u \in U}} J(u) = +\infty.$$

Věta 23 *Nech¸ $U \subset \mathbb{R}^n$ je neomezená, uzavřená mno«ina. Buď $J : U \to \mathbb{R}$ spojitá a koercivní. Pak J nabývá na U minima.*
Důkaz: Nech¸ $a \in U$. Z podmínky koercivity vyplývá, «e existuje takové $R > 0$, «e

$$a \in B(0,R) \cap U \neq \emptyset, \quad \forall u \in U \setminus B(0,R) : J(u) \geq J(a) + \frac{\pi}{17},$$

kde $B(0,R)$ označuje uzavřenou kouli se středem v počátku a poloměrem R. Označíme-li $K = U \cap B(0,R)$, nabývá J na K podle minulé věty v nějakém bodě $\bar{u}$ minima. Pak je ale pro $u \in U \setminus B(0,R)$

$$J(u) \geq J(a) + \frac{\pi}{17} > J(a) \geq J(\bar{u})$$

a hodnota $J(\bar{u})$ je tedy minimem na celé množině U. □

Označení: Nechˌ $U \subset \mathbb{R}^n$ je otevřená, $J \in C^1(U)$. Pak pro ka«dé $u \in U$, $\varphi \in \mathbb{R}^n$ existuje směrová derivace

$$J'(u,\varphi) = \lim_{\theta \to 0} \frac{(J(u+\theta\varphi) - J(u))}{\theta}.$$

Její hodnota je rovna $\nabla J(u)\,.\,\varphi$, kde

$$\nabla J(u) = \operatorname{grad} J(u) = \left(\frac{\partial J}{\partial u_1}(u), \ldots, \frac{\partial J}{\partial u_n}(u) \right)^T$$

je *gradient* funkce J.

Definice 18 Buď $U \subset \mathbb{R}^n$ konvexní mno«ina. Řekneme, «e $J : U \to \mathbb{R}$ je *konvexní funkce*, jestli«e

$$J(u + \theta(v-u)) \le J(u) + \theta(J(v) - J(u))$$

pro libovolné $u, v \in U$, $\theta \in [0,1]$. Řekneme, «e J je *ryze konvexní*, jestli«e pro $u \neq v$, $\theta \in (0,1)$ platí ostrá nerovnost.

Věta 24 *$U \subset \mathbb{R}^n$ buď otevřená konvexní mno«ina, $J \in C^1(U)$. Pak*

(i) J je konvexní právě tehdy, kdy« $J(v) \ge J(u) + J'(u, v-u)$ pro ka«dé $u, v \in U$;

(ii) J je ryze konvexní právě tehdy, kdy« $J(v) > J(u) + J'(u, v-u)$ pro ka«dé $u, v \in U$, $u \neq v$.

Důkaz: Nechˌ je nejprve J konvexní na U. Tzn., že je-li $u, v \in U$, potom

$$J(u + \theta(v-u)) \le J(u) + \theta(J(v) - J(u)), \qquad \forall \theta \in (0,1).$$

Z toho plyne

$$J(v) - J(u) \ge \frac{J(u + \theta(v-u)) - J(u)}{\theta}, \qquad \forall \theta \in (0,1).$$

Nyní uva«ujme $\theta \to 0+$. Pak

$$J(v) - J(u) \ge \lim_{\theta \to 0+} \frac{J(u + \theta(v-u)) - J(u)}{\theta} = J'(u, v-u).$$

Dokažme opačnou implikaci. Nechˌ je pro ka«dé $u, v \in U$ splněna podmínka

$$J(v) \ge J(u) + J'(u, v-u).$$

Z toho plyne, «e pro ka«dé $\theta \in [0,1]$ platí

$$\begin{aligned} J(u) &\ge J(u + \theta(v-u)) - J'(u + \theta(v-u), \theta(v-u)) = \\ &= J(u + \theta(v-u)) - \theta J'(u + \theta(v-u), v-u) \end{aligned}$$

a

$$J(v) \geq J(u+\theta(v-u)) + J'(u+\theta(v-u),(1-\theta)(v-u)) = \\ = J(u+\theta(v-u)) + (1-\theta)J'(u+\theta(v-u),v-u).$$

Přenásobením prvního vztahu členem $(1-\theta)$ a druhého θ a sečtením dostaneme

$$(1-\theta)J(u) + \theta J(v) \geq J(u+\theta(v-u)), \qquad \forall u,v \in U, \forall \theta \in [0,1].$$

Nyní budeme dokazovat druhé tvrzení. Nejprve nech¸ je

$$J(v) > J(u) + J'(u, v-u), \qquad \forall u,v \in U, u \neq v, \theta \in (0,1).$$

Stejným postupem jako jsme právě u«ili (uva«ujeme-li ostré nerovnosti) dostaneme, «e f je ryze konvexní.

Předpokládejme naopak, «e f je ryze konvexní. Tudí«, funkce f je rovně« konvexní. Z definice ryzí konvexity funkce f vyplývá, že pro $\theta \in (0,1)$ a $u, v \in U,\ u \neq v$ máme

$$J(u+\theta(v-u)) < J(u) + \theta(J(v) - J(u)).$$

Odtud a z tvrzení (i) ihned plyne, «e

$$J(v) - J(u) > \frac{J(u+\theta(v-u)) - J(u)}{\theta} \geq \\ \geq \frac{J'(u,\theta(v-u))}{\theta} = J'(u, v-u),$$

co« jsme chtěli dokázat. □

Věta 25 *$U \subset \mathbb{R}^n$ buď otevřená, $J \in C^1(U)$.*

(i) Je-li $\bar{u} \in U$ bodem lokálního minima, je $J'(u,\varphi) = 0$ pro ka«dé $\varphi \in \mathbb{R}^n$ (neboli $\nabla J(\bar{u}) = 0$).

(ii) Je-li navíc U konvexní, J konvexní, potom jsou následující tvrzení ekvivalentní:

1) $\bar{u}$ je bodem lokálního minima,

2) $\bar{u}$ je bodem minima,

3) $\nabla J(\bar{u}) = 0$.

(iii) Je-li navíc J ryze konvexní, má J na U nejvýe jeden bod minima.

Důkaz: Tvrzení (i) je známé z matematické analýzy. Pro důkaz (ii) předpokládejme, že U je konvexní, $J \in C^1(U)$ a J je konvexní. Z toho vyplývá, že

$$J(v) \geq J(\bar{u}) + J'(\bar{u}, v-\bar{u}) \quad \forall v \in U. \tag{3.1.1}$$

Jestli«e $\bar{u}$ je bodem lokálního minima, potom $\nabla J(\bar{u}) = 0$ a $J'(\bar{u},\varphi) = 0$ pro ka«dé $\varphi \in \mathbb{R}^n$. Odtud a z (3.1.1) plyne, že $J(v) \geq J(\bar{u})$ pro všechna $v \in U$. Tudíž,

$$\bar{u} = \arg\min_U J.$$

Naopak, pokud $\nabla J(\bar{u}) = 0$, pak podle předchozího je $\bar{u}$ bodem minima.

Pro důkaz tvrzení (iii) předpokládejme, «e J je ryze konvexní a má v U dva body minima $u_1 \neq u_2$. Potom ale z ryzí konvexity plyne, «e

$$J(u_2) > J(u_1) + J'(u_1, u_2 - u_1),$$

přičem« $J'(u_1, u_2 - u_1) = 0$, proto«e u_1 je bod minima. To je ovem spor. □

Definice 19 *Pokud $\nabla J(u) = 0$, potom bod u nazýváme stacionárním bodem funkce J.*

Nech¸ $U \subset \mathbb{R}^n$ je otevřená množina a nech¸ $J \in C^2(U)$. Pak má J totální diferenciál druhého řádu v libovolném bodě $u \in U$ a směrové derivace druhého řádu $J''(u, \varphi, \psi) = \varphi^T J''(u)\psi$ ve směrech $\varphi, \psi \in \mathbb{R}^n$, kde

$$J''(u) = \left(\frac{\partial^2 J}{\partial u_i \partial u_j}(u) \right)_{i,j=1}^{n}$$

je *Hessova matice* (která je symetrická). Pokud $u, v \in U$ a úsečka s krajními body u, v le«í celá v mno«ině U, potom platí Taylorův vzorec ve tvaru

$$J(v) = J(u) + J'(u, v - u) + \frac{1}{2} J''(u + \theta(v - u), v - u, v - u)$$

pro nějaké $\theta \in [0, 1]$. Pou«itím tohoto vzorce lze dokázat následující větu:

Věta 26 *Nech¸ $J : \mathbb{R}^n \to \mathbb{R}$, $J \in C^2$ a nech¸ existuje $\alpha > 0$ tak, «e $J''(u, \varphi, \varphi) \geq \alpha \|\varphi\|^2$ pro všechna $u, \varphi \in \mathbb{R}^n$. Potom J je koercivní a ryze konvexní na $\mathbb{R}^n$.*

Důkaz: Nejprve doká«eme koercivitu. Z Taylorova vzorce plyne, «e pro ka«dé $u \in \mathbb{R}^n$ existuje $\theta \in [0, 1]$ takové, že

$$J(u) = J(0) + J'(0, u) + \frac{1}{2} J''(\theta u, u, u).$$

Vezmeme-li v úvahu, že

$$|J'(0, u)| = |\nabla J(0) \cdot u| \leq \|\nabla J(0)\| \, \|u\|$$

a že podle předpokladů věty je

$$\frac{1}{2} J''(\theta u, u, u) \geq \frac{\alpha}{2} \|u\|^2,$$

ihned zjistíme, že

$$J(u) \geq J(0) - \|\nabla J(0)\| \, \|u\| + \frac{\alpha}{2} \|u\|^2 \to \infty \text{ if } \|u\| \to \infty.$$

Tím jsme dokázali koercivitu.

Zbývá dokázat ryzí konvexitu. Nechť $u, v \in \mathbb{R}^n,\ u \neq v$. Opět u«ijeme Taylorův vzorec. Existuje $\theta \in [0,1]$ takové, že

$$\begin{aligned} J(v) &= J(u) + J'(u, v-u) + \frac{1}{2}J''(u+\theta(v-u), v-u, v-u) \geq \\ &\geq J(u) + J'(u, v-u) + \frac{1}{2}\alpha\|v-u\|^2 > \\ &> J(u) + J'(u, v-u), \end{aligned}$$

čím« je důkaz hotov. □

Důsledek 27 *Funkce J splňující předpoklady věty 26 má právě jedno minimum v $\mathbb{R}^n$.*

3.2 Numerické metody hledání minima

V této části se věnujeme iteračním procesům, které představují numerické minimalizační algoritmy.

Nech¸ $J : \mathbb{R}^n \to \mathbb{R}$ a je dán počáteční bod $u^0 \in \mathbb{R}^n$. Předpokládejme, «e jsme ji« určili $u^k \in \mathbb{R}^n$ pro nějaké $k \geq 0$. Pak položíme

$$u^{k+1} = u^k + \rho_k \varphi^k,$$

kde $k \geq 0$, $\varphi^k \in \mathbb{R}^n$, $\rho_k > 0$. Vektor φ^k nazýváme *směr spádu*, ρ_k určuje vzdálenost – *délku kroku* ve směru φ^k. Pro praktické pou«ití je potřeba zodpovědět dvě základní otázky:

1) jak volit směr spádu φ_k,
2) jak volit velikosti kroku.

Směr spádu φ^k volíme tak, aby $J(u^{k+1}) < J(u^k)$ pro vhodné $\rho_k > 0$ (např. dostatečně malé). Směr spádu φ^k můžeme zvolit např. tak, že

$$J'(u^k, \varphi^k) = \nabla J(u^k) \cdot \varphi^k < 0. \tag{3.2.2}$$

Věta 28 *Nech¸ je $J \in C^2(\mathbb{R}^n)$ a platí (3.2.2). Potom existuje $\tilde{\rho}_0 > 0$ takové, «e $J(u^{k+1}) < J(u^k)$, pokud $\rho_k \in (0, \tilde{\rho}_0)$.*

Důkaz: Z Taylorova vzorce plyne, že

$$J(u^{k+1}) = J(u^k) + \rho_k J'(u^k, \varphi^k) + \frac{1}{2}\rho_k^2 J''(\beta, \varphi^k, \varphi^k),$$

kde β leží na úsečce mezi body u^k a $u^k + \rho_k\varphi^k$. Nech¸ $R > 0$. Potom existuje $M > 0$ tak, že

$$|J''(\beta, \varphi^k, \varphi^k)| \leq M, \quad \forall\ \beta = u^k + \rho\varphi^k, \quad \rho \in [0, R].$$

Odtud plyne, «e existuje dostatečně malé $\tilde{\rho}_0 > 0$ takové, «e

$$J'(u^k, \varphi_k) + \frac{1}{2}\rho_k J''(\beta, \varphi^k, \varphi^k) \leq J'(u^k, \varphi^k) + \frac{1}{2}\rho M$$
$$\leq J'(u^k, \varphi^k) + \frac{1}{2}\tilde{\rho}_0 M < 0 \quad \forall \rho \in (0, \tilde{\rho}_0).$$

Tudíž pro $\rho_k \in (0, \tilde{\rho}_0)$ máme

$$J(u^{k+1}) = J(u^k) + \rho_k \left(J'(u^k, \varphi^k) + \frac{1}{2}\rho J''(\beta, \varphi^k, \varphi^k) \right) < J(u^k).$$

□

Příklad 12 *Metoda největšího spádu.* Tuto metodu dostaneme, jestliže φ^k definujeme jako směr největšího spádu, tj.

$$\varphi^k = -\nabla J(u^k).$$

Pokud $\nabla J(u^k) \neq 0$, potom

$$J'(u^k, \varphi^k) = -\nabla J(u^k) \cdot \nabla J(u^k) = -\|\nabla J(u^k)\|^2 < 0.$$

V dalím uvedeme a budeme analyzovat dva typy metody největího spádu.

Algoritmus I - metoda největího spádu s konstantním krokem

Nechť je dáno: $u^0 \in I\!R^n$, $M \geq 1$ celé, $\varepsilon > 0$, $\rho > 0$. Pro $k = 0, 1, \ldots, M$ proveďte:

1) Položte $u^{k+1} = u^k - \rho \nabla J(u^k)$
2) Jestliže $\|\nabla J(u^k)\| < \varepsilon$, jděte na 3.
3) Stop.

Algoritmus II - metoda největího spádu s optimálním krokem

Nechť je dáno: $u_0 \in I\!R^n$, $M \geq 1$ celé, $\varepsilon > 0$. Pro $k = 0, 1, \ldots, M$ proveďte:

1) Položte $\varphi^k = -\nabla J(u^k)$.
2) Vypočtěte $\rho_k = \arg\min_{\rho > 0} J(u^k + \rho \varphi^k)$
3) Položte $u^{k+1} = u^k + \rho_k \varphi^k$
4) Je-li $\|\nabla J(u^k)\| < \varepsilon$, pak jděte na 5.
5) Stop.

Pokud v předchozích algoritmech zvolíme $M = \infty$, $\varepsilon = 0$, dostaneme nekonečné iterační procesy. V dalím se budeme zabývat analýzou konvergence těchto iteračních procesů, tj. konvergence posloupnosti $\{u^k\}_{k=0}^{\infty}$ pro $k \to \infty$ k bodu minima nebo lokálního minima nebo ke stacionárnímu bodu.

V praktických výpočtech nelze samozřejmě realizovat tento nekonečný iterační proces. Výpočet ukončíme, pokud je počet vypočtených iterací určený číslem M dostatečně velký nebo je splněna podmínka $\|\nabla J(u^k)\| < \varepsilon$, kde dostatečně malé ε zaručuje, že je přibližně splněna podmínka stacionárního bodu funkce J.

Věta 29 *Nechť $J \in C^2(\mathbb{R}^n)$ a nechť existují konstanty λ, $\Lambda > 0$ takové, že*

$$\lambda\|\varphi\|^2 \le \varphi^T J''(u)\,\varphi \le \Lambda\|\varphi\|^2 \quad \forall\, u,\, \varphi \in \mathbb{R}^n. \tag{3.2.3}$$

Zvolme $\rho = \frac{2}{\lambda+\Lambda}$. Pak posloupnost $\{u^k\}_{k=0}^{\infty}$ daná algoritmem I s $M = \infty$ a $\varepsilon = 0$ konverguje k jednoznačnému řešení úlohy

$$J(\overline{u}) = \min_{u\in\mathbb{R}^n} J(u) \tag{3.2.4}$$

a

$$\|u^k - \overline{u}\| \le \|u^0 - \overline{u}\| \left(\frac{\Lambda-\lambda}{\Lambda+\lambda}\right)^k, \quad k = 1, 2, \dots \tag{3.2.5}$$

Důkaz: Z předpokladů, že $J \in C^2(\mathbb{R}^n)$ a (3.2.3) plyne, že funkce J je spojitá, ryze konvexní a koercivní. Tudíž, existuje právě jeden bod $\overline{u} = \arg\min_{\mathbb{R}^n} J$ a $\nabla J(\overline{u}) = 0$. Podle algoritmu I,

$$\|u^{k+1} - \overline{u}\| = \|u^k - \rho\nabla J(u^k) + \rho\nabla J(\overline{u}) - \overline{u}\|. \tag{3.2.6}$$

Zabývejme se výrazem $\nabla J(u^k) - \nabla J(\overline{u})$. Definujme vektorovou funkci $f(t) = \nabla J(\overline{u}+t(u^k-\overline{u}))$, $t \in [0,1]$. Její derivace má tvar $f'(t) = J''(\overline{u}+t(u^k-\overline{u}))\,(u^k-\overline{u})$. Můžeme psát

$$\begin{aligned}\nabla J(u^k) - \nabla J(\overline{u}) = f(1) - f(0) \;&=\; \int_0^1 f'(t)\,\mathrm{d}t \\ &= \int_0^1 J''(\overline{u} + t(u^k - \overline{u}))\,(u^k - \overline{u})\,\mathrm{d}t.\end{aligned} \tag{3.2.7}$$

Odtud a z (3.2.6) plyne, že

$$\|u^{k+1} - \overline{u}\| \le \left\| I - \rho \int_0^1 J''(\overline{u} + t(u^k - \overline{u}))\,\mathrm{d}t \right\| \|u^k - \overline{u}\|, \tag{3.2.8}$$

kde I je jednotková matice. Je zřejmé, že $I - \rho\int_0^1 J''(\overline{u}+t(u^k-\overline{u}))\,\mathrm{d}t$ je symetrická matice typu $n \times n$.

V dalším budeme potřebovat dva důležité výsledky.

Lemma 5 *Pro symetrickou reálnou matici $\mathbb{A}$ platí:*

$$\|\mathbb{A}\| = \sigma(\mathbb{A}), \tag{3.2.9}$$

kde $\|\mathbb{A}\|$ je norma matice $\mathbb{A}$ indukovaná eukleidovskou normou $\|\cdot\|$ v $\mathbb{R}^n$:

$$\|\mathbb{A}\| = \sup_{0\ne u\in\mathbb{R}^n} \frac{\|\mathbb{A}u\|}{\|u\|},$$

a $\sigma(\mathbb{A})$ je spektrální poloměr matice $\mathbb{A}$:

$$\sigma(\mathbb{A}) = \max_{\lambda\in \mathrm{Sp}(\mathbb{A})} |\lambda|.$$

Zde $\mathrm{Sp}(\mathbb{A}) = \{\lambda;\ \lambda =$ *vlastní číslo matice* $\mathbb{A}\}$ *značí spektrum matice* $\mathbb{A}$.

Důkaz: Nechť $\mathrm{Sp}(\mathbb{A}) = \{\lambda_1, \ldots, \lambda_n\}$, kde $|\lambda_1| \le |\lambda_2| \le \cdots \le |\lambda_m|$. Poněvadž $\mathbb{A}$ je symetrická matice, je $\mathrm{Sp}\,\mathbb{A} \subset \mathbb{R}$. Je známo z algebry, že v $\mathbb{R}^n$ existuje ortonormální báze $\varphi_1, \ldots, \varphi_n$ tvořená vlastními vektory matice $\mathbb{A}$ k vlastním číslům $\lambda_1, \ldots, \lambda_n$: $\mathbb{A}\varphi_i = \lambda_i\varphi_i$, $i = 1, \ldots, n$.

Jestliže $u \in \mathbb{R}^n$, $u \neq 0$, pak existují $c_i \in \mathbb{R}$ takové, že $u = \sum_{i=1}^n c_i\varphi_i$. Dále snadno zjistíme, že

$$\mathbb{A}\,u = \sum_{i=1}^n \lambda_i c_i \varphi_i \quad \text{a} \quad \frac{\|\mathbb{A}u\|^2}{\|u\|^2} = \frac{\sum_{i=1}^n \lambda_i^2 c_i^2}{\sum_{i=1}^n c_i^2} \le \lambda_n^2.$$

Odtud plyne, že $\|\mathbb{A}\| \le |\lambda_n| = \sigma(\mathbb{A})$. Zvolíme-li $u = \varphi_n$, pak dostaneme $\sigma(\mathbb{A}) = |\lambda_n| = \|\mathbb{A}\varphi_n\| \le \|A\|$. Tudíž $\sigma(\mathbb{A}) = \|\mathbb{A}\|$. □

Lemma 6 *Nechť*

$$|\varphi^T \mathbb{A}\varphi| \le \alpha\|\varphi\|^2 \quad \forall\, \varphi \in \mathbb{R}^n. \tag{3.2.10}$$

Pak

$$|\lambda| \le \alpha \quad \forall\, \lambda \in \mathrm{Sp}(\mathbb{A}). \tag{3.2.11}$$

Důkaz: Je-li $\lambda \in \mathrm{Sp}(\mathbb{A})$, pak existuje $\varphi \in \mathbb{R}^n$, $\varphi \neq 0$ takové, že $\mathbb{A}\varphi = \lambda\varphi$. Tudíž, $\varphi^T\mathbb{A}\varphi = \lambda\|\varphi\|^2$ a tedy

$$|\lambda|\,\|\varphi\|^2 = |\varphi^T\mathbb{A}\varphi| \le \alpha\|\varphi\|^2.$$

Odtud plyne (3.2.11). □

Nyní budeme pokračovat v důkazu věty 29. Z lemmatu 5 plyne, že

$$\left\| I - \rho\int_0^1 J''(\overline{u} + t(u^k - \overline{u}))\,\mathrm{d}t \right\| = \sigma\left(I - \rho\int_0^1 J''(\overline{u} + t(u^{k-\overline{u}}))\,\mathrm{d}t \right).$$

Z (3.2.3) plyne, že pro každé $\varphi \in \mathbb{R}^n$ a každé $t \in [0,1]$ máme

$$(1 - \rho\Lambda)\,\|\varphi\|^2 \le \varphi^T\left(I - \rho J''(\overline{u} + t(u^k - \overline{u}))\right)\varphi \le (1 - \rho\lambda)\,\|\varphi\|^2.$$

Integrací dostaneme

$$(1 - \rho\Lambda)\,\|\varphi\|^2 \le \varphi^T\left(I - \rho\int_0^1 J''(\overline{u} + t(u^k - \overline{u}))\,\mathrm{d}t \right)\varphi \le (1 - \rho\lambda)\,\|\varphi\|^2$$

a tedy

$$\left| \varphi^T\left(I - \rho\int_0^1 J''(\overline{u} + t(u^k - \overline{u}))\,\mathrm{d}t \right)\varphi \right| \le q\|\varphi\|^2,$$

kde

$$q := \max\left(|1 - \rho\Lambda|,\, |1 - \rho\lambda|\right).$$

Položíme-li $\rho = \frac{2}{\lambda+\Lambda}$, dostaneme $q = (\Lambda - \lambda)/(\Lambda + \lambda) < 1$. Z lemmat 5 a 6 plyne, že

$$\left\| I - \rho\int_0^1 J''(\overline{u} + t(u^k - \overline{u}))\,\mathrm{d}t \right\| = \sigma\left(I - \rho\int_0^1 J''(\overline{u} + t(u^{k-\overline{u}}))\,\mathrm{d}t \right) \le q.$$

Odtud a z (3.2.8) ihned dostaneme (3.2.5). □

Nevýhodou algoritmu I je, že požadavek (3.2.3), který zaručuje konvergenci, vyžaduje znalost čísel λ, Λ. Ověření podmínky (3.2.3) je v praxi ale obvykle značně obtížné, ne-li nemožné. Proto je často používán algoritmus II.

Věta 30 *Nechť v algoritmu II je $M = \infty$ a $\varepsilon = 0$. Dále předpokládejme, že $J \in C^1(\mathbb{R}^n)$ a ∇J splňuje lokálně Lipschitzovu podmínku v $\mathbb{R}^n$. Pak každý hromadný bod $\hat{u} \in \mathbb{R}^n$ posloupnosti $\{u^m\}_{m=0}^{\infty}$ vypočtené pomocí algoritmu II splňuje podmínku*

$$\nabla J(\hat{u}) = 0. \tag{3.2.12}$$

Důkaz: Nechť existuje podposloupnost $\{u^{m_i}\}_{i=0}^{\infty}$ taková, že

$$u^{m_i} \to \hat{u} \in \mathbb{R}^n \quad \text{pro } i \to \infty. \tag{3.2.13}$$

Z konstrukce posloupnosti $\{u^m\}_{m=0}^{\infty}$ plyne, že posloupnost $\{J(u^m)\}_{m=0}^{\infty}$ je nerostoucí. Tudíž, protože $m_{i+1} \geq m_i + 1$, máme

$$J\left(u^{m_{i+1}}\right) - J\left(u^{m_i}\right) \leq J\left(u^{m_i+1}\right) - J\left(u^{m_i}\right). \tag{3.2.14}$$

Prvek u^{m_i+1} je vypočten z u^{m_i} podle algoritmu 2, podle kterého platí

$$J\left(u^{m_i+1}\right) - J\left(u^{m_i}\right) \leq J\left(u^{m_i} + \rho\varphi^{m_i}\right) - J\left(u^{m_i}\right) \quad \forall\, \rho > 0. \tag{3.2.15}$$

Z věty o střední hodnotě plyne existence $\theta_\rho \in [0,1]$ takového, že

$$\begin{aligned} & J\left(u^{m_i} + \rho\varphi^{m_i}\right) - J\left(u^{m_i}\right) \\ &= \rho\nabla J\left(u^{m_i} + \theta_\rho\rho\varphi^{m_i}\right) \cdot \varphi^{m_i} \\ &= -\rho\nabla J\left(u^{m_i} - \theta_\rho\rho\nabla J(u^{m_i})\right) \cdot \nabla J\left(u^{m_i}\right). \end{aligned} \tag{3.2.16}$$

Ze spojitosti ∇J a předpokladu, že $u^{m_i} \to \hat{u}$ pro $i \to \infty$ plyne, že

$$\varphi^{m_i} = -\nabla J\left(u^{m_i}\right) \to -\nabla J(\hat{u}) \quad \text{pro } i \to \infty. \tag{3.2.17}$$

Dále dokážeme, že existují $i^{**} \geq 0$ celé a $\rho > 0$ tak, že

$$\begin{aligned} & \rho\nabla J\left(u^{m_i} - \theta\rho\nabla J\left(u^{m_i}\right)\right) \cdot \nabla J\left(u^{m_i}\right) \\ & \geq \frac{\rho}{4}\|\nabla J(\hat{u})\|^2 \quad \forall\, i > i^{**},\ \forall\, \theta \in [0,1]. \end{aligned} \tag{3.2.18}$$

Za tím účelem zvolme okolí $\mathcal{O}_\varepsilon(\hat{u})$ o poloměru $\varepsilon > 0$. V důsledku (3.2.13) a (3.2.17) existují $i^* \geq 0$ celé a $\hat{\rho} > 0$ takové, že

$$u^{m_i},\ u^{m_i} - \theta\rho\nabla J\left(u^{m_i}\right) \in \mathcal{O}_\varepsilon(\hat{u}) \quad \forall\, i > i^*,\ \forall\, \rho \in (0,\hat{\rho}),\ \forall\, \theta \in [0,1]. \tag{3.2.19}$$

Dále je zřejmé, že existuje konstanta $K > 0$ tak, že

$$\|\nabla J(x) - \nabla J(y)\| \leq K\|x - y\| \quad \forall\, x,\ y \in \overline{\mathcal{O}_\varepsilon(\hat{u})}. \tag{3.2.20}$$

Odtud a z (3.2.19) vidíme, že

$$\begin{aligned} &\left\|\nabla J\left(u^{m_i} - \theta\rho\nabla J\left(u^{m_i}\right)\right) - \nabla J\left(u^{m_i}\right)\right\| \\ &\leq \rho K \left\|\nabla J\left(u^{m_i}\right)\right\| \quad \forall\, i > i^*,\ \forall\, \rho \in (0, \hat{\rho}),\ \forall\, \theta \in [0,1]. \end{aligned} \tag{3.2.21}$$

V důsledku této nerovnosti platí:

$$\begin{aligned} &\rho\nabla J\left(u^{m_i} - \theta\rho\nabla J\left(u^{m_i}\right)\right) \cdot \nabla J\left(u^{m_i}\right) \\ &= \rho \left\|\nabla J\left(u^{m_i}\right)\right\|^2 + \rho\nabla J\left(u^{m_i}\right) \cdot \left(\nabla J\left(u^{m_i} - \theta\rho\nabla J\left(u^{m_i}\right)\right) - \nabla J\left(u^{m_i}\right)\right) \\ &\geq \rho \left\|\nabla J\left(u^{m_i}\right)\right\|^2 (1 - \rho K) \quad \forall\, i > i^*,\ \forall\, \rho \in (0, \hat{\rho}),\ \forall\, \theta \in [0,1]. \end{aligned} \tag{3.2.22}$$

Z (3.2.17) plyne, že existuje $i^{**} \geq i^*$ takové, že

$$\left\|\nabla J\left(u^{m_i}\right)\right\|^2 \geq \frac{1}{2} \left\|\nabla J(\hat{u})\right\|^2 \quad \forall\, i > i^{**}. \tag{3.2.23}$$

Nyní zvolme $\rho \in (0, \hat{\rho})$ tak malé, že $\frac{1}{2}(1 - \rho K) \geq \frac{1}{4}$. Pak

$$\frac{1}{2}\rho \left\|\nabla J(\hat{u})\right\|^2 (1 - \rho K) \geq \frac{1}{4}\rho \left\|\nabla J(\hat{u})\right\|^2.$$

Odtud, z (3.2.22) a (3.2.23) dostaneme ihned (3.2.18).

Nyní již můžeme dokončit důkaz věty. Shrneme-li (3.2.14), (3.2.15), (3.2.16) a (3.2.18), vidíme, že existují $i^{**} \geq 0$ a $\rho > 0$ takové, že pro všechna $i > i^{**}$ platí odhad

$$J\left(u^{m_{i+1}}\right) - J\left(u^{m_i}\right) \leq -\frac{1}{4}\rho \|\nabla J(\hat{u})\|^2 \leq 0. \tag{3.2.24}$$

Z (3.2.13) a spojitosti funkce J plyne, že

$$J\left(u^{m_{i+1}}\right) - J\left(u^{m_i}\right) \to 0 \quad \text{pro } i \to \infty.$$

takže $\|\nabla J(\hat{u})\|^2 = 0$, což jsme chtěli dokázat. □

Důsledek. Splňuje-li J předpoklady věty 30 a J je koercivní, pak z posloupnosti $\{u^m\}_{m=0}^{\infty}$ lze vybrat podposloupnost konvergentní k nějakému $\hat{u} \in \mathbb{R}^n$ a tudíž $\nabla J(\hat{u}) = 0$. Potom $\hat{u}$ je bodem lokálního minima nebo tzv. sedlovým bodem. Je ale nepravděpodobné, že bychom při praktické realizaci dostali konvergenci k sedlovému bodu.

Jestliže J splňuje předpoklady věty 30 a J je koercivní a ryze konvexní, pak $u^m \to \overline{u} \in \mathbb{R}^n$ pro $m \to \infty$ a $\overline{u}$ je jediným bodem minima funkce J.

REJSTŘÍK

LITERATURA

1. Feistauer M.: *Diskrétní metody řeení diferenciálních rovnic*, 1981.
2. Přikryl P.: *Numerické metody matematické analýzy*, (MVT 24), SNTL.
3. Vitásek E., Prager E. : *Numerická matematika I., II.*
4. Ralston, A.: *Základy numerické matematiky.*

[H]. Henrici: *Discrete Variable Methods for ODE's.*