Základy numerické matematiky

Miloslav Feistauer
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 $x_0, \ldots, x_n$ a předepsané hodnoty $f(x_0), \ldots, f(x_n)$. Hledáme funkci p daného typu, která vyhovuje podmínkám $p(x_i) = f(x_i),\ i = 0, \ldots, n$. Je známo, že existuje právě jeden (Lagrangeův) interpolační polynom p_n stupně nejvýše 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 každé $x \in [a, b]$ existuje $\xi \in (a, b)$ tak, že*

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

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

Důsledek 2 *Buď $f \in C^\infty[a, b]$, $|f^{(k)}| \leqq K$ v $[a, b]$. Pak*

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

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

Cvičení 1 Uvažujme funkci $f(x) = \sin x$ na intervalu $[0, 1]$. Jaké největší chyby se můžeme dopustit, aproximujeme-li f pomocí p_9?

Pokaždé však nedosáhneme tak vynikající aproximace. Příkladem může být funkce $f(x) = \frac{1}{1+x^2}$ na intervalu $[-5, 5]$. Dá se dokázat, že 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, že hodnoty interpolačního polynomu jsou silně ovlivněny i hodnotami funkce ve vzdálených uzlech. Řešením je aproximovat f po částech. Při tomto přístupu je naším cílem aproximovat funkci f v intervalu $[x_0, x_n]$ pomocí funkce φ takové, že $\varphi|_{[x_i, x_{i+1}]}$ je polynom. Většinou se navíc požaduje, aby φ byla třídy C^k pro dané k. Takovou funkci φ nazýváme *spline*.

Nejjednodušším případem ($k = 0$) je aproximace pomocí funkce po částech lineární, jejímž grafem je lomená čára. V praxi je většinou třeba lepší aproximace. Přijatelným řešením je tzv. *kubický spline*.

Definice 1 Řekneme, že funkce $\varphi : [x_0, x_n] \to I\!R$ je *kubický spline*, jestliže

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

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

Říkáme, že φ je *kubický interpolační spline* k f v bodech $x_0, \dots x_n$, jestliže 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 však dávají jen $4n - 2$ podmínek. Dá se tedy očekávat, že budeme muset ještě dva parametry zvolit. Nejčastěji se používají 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) (jíž je (γ) speciálním případem). Předpokládejme nejprve, že již 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

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

Integrováním dostaneme φ_i' a φ_i:

$$\begin{aligned}\varphi_i'(x) &= -M_i\frac{(x_{i+1} - x)^2}{2h_i} + M_{i+1}\frac{(x - x_i)^2}{2h_i} + A_i\\ \varphi_i(x) &= M_i\frac{(x_{i+1} - x)^3}{6h_i} + M_{i+1}\frac{(x - x_i)^3}{6h_i} + A_i(x - x_i) + B_i.\end{aligned}$$

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:

$$f(x_i) = \frac{1}{6}M_i h_i^2 + B_i \qquad \Longrightarrow \qquad B_i = f(x_i) - \frac{1}{6}M_i h_i^2,$$

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

$\Longrightarrow$

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

Zbývá určit hodnoty momentů: M_0 a M_n máme zadané, ostatní vypočteme ze spojitosti první derivace (derivací funkce φ_i v bodech x_i, x_{i+1} se rozumí příslušná jednostranná derivace).

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

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

Z rovnosti obou derivací dostaneme po úpravách

$$M_{i-1}\frac{h_{i-1}}{6} + M_i\left(\frac{h_{i-1}}{3}+\frac{h_i}{3}\right) + M_{i+1}\frac{h_i}{6} = \frac{f(x_{i+1}-f(x_i)}{h_i} - \frac{f(x_i)-f(x_{i-1})}{h_{i-1}};$$

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 přepsat ve tvaru

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

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

$$\begin{array}{lllll} 2M_1 & + \mu_1 M_2 & & = g_1 - \lambda_1 f''_0 \\ \lambda_2 M_1 & + \ 2M_2 & + \mu_2 M_3 & = g_2 \\ \lambda_3 M_2 & + \ 2M_3 & + \mu_3 M_4 & = g_3 \\ & & & \vdots \\ \lambda_{n-1}M_{n-2} & + 2M_{n-1} & & = g_{n-1} - \mu_{n-1} f''_n. \end{array}$$

Dokážeme-li nyní existenci a jednoznačnost řešení, budeme hotovi. Všimněme si, že prvky na diagonále matice soustavy jsou vždy 2, zatímco součet všech 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 Dokažte, že každá ostře diagonálně dominantní matice je regulární.

Navíc je matice soustavy tzv. třídiagonální matice, na které se poměrně jednoduše provádí eliminace. Soustava vypadá takto (uvažujeme 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) nám zmizí všechny b_j a na diagonále se nám postupně objeví členy $A_{jj} = \eta_j$, kde

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

a ve vektoru pravých stran analogicky vzniknou ξ_j, kde

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

Dostaneme tedy soustavu

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

z níž už snadno vypočítáme neznámé y_j:

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

Snadno se ukáže, že pro ostře diagonálně dominantní matici vyjdou po eliminaci prvky na diagonále nenulové, takže nemusíme prohazovat řádky a neznámé přímo vypočítáme podle uvedených vztahů.

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: Nechť $K > 0$ je konstanta. Dále uvažujme dělení D intervalu $[a, b]$, které je tvořeno body $a = x_0 < \ldots < x_n = b$ a které splňuje podmínku*

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

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

$$\left| f^{(k)} - \varphi^{(k)} \right| \leqq CK \, h^{4-k}, \qquad (k = 0, \ldots, 3),$$

kde $h = \max h_i$, přičemž v případě $k = 3$ uvažujeme v dělících bodech derivace zleva nebo zprava.

Důsledek 4 *Máme-li posloupnost dělení D' takových, že $h \to 0$, dostáváme*

$$\varphi^{(k)} \rightrightarrows f^{(k)}.$$

Poznámka 5 Pokud používáme interpolační spline pro interpolaci několika naměřených hodnot, nemůžeme použít okrajové podmínky ve tvaru rovnosti druhých derivací; hodila by se tedy obdobná věta pro přirozený interpolační spline. Nahradíme-li okrajové podmínky ve větě podmínkami $\varphi''(x_0) = 0$, $\varphi''(x_n) = 0$, dostaneme (slabší) odhad

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

1.1.3 *Spline s napětím*

Ne vždy představuje kubický interpolační spline ideální řešení problému. Například při aproximaci nespojitých funkcí nebo funkcí s nespojitou derivací dostáváme nepříjemnou oscilaci v okolí nespojitosti. Proto se někdy používá modifikovaná konstrukce tzv. *splinu s napětím.* Při této konstrukci opět požadujeme $\varphi \in C^2[a, b]$, $\varphi(x_i) = f(x_i)$, ale požadavek, aby φ byla po částech kubická se nahrazuje požadavkem, aby $\varphi|_{[x_i,x_{i+1}]}$ byla řešením diferenciální rovnice

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

kde $\tau \geqq 0$ je tzv. *napěťový parametr.* Pokud bychom položili $\tau = 0$, dostali bychom přesně polynomy stupně nejvýše 3. Pro $\tau > 0$ místo funkcí x^2 a x^3 dostaneme fundamentální řešení $\cosh(t\sqrt{\tau}) - 1$ a $\sinh(t\sqrt{\tau})$. Všimněme si, že při pevné funkci f a dělení D pro $\tau \to \infty$ dostaneme $\varphi'' \to 0$. Význam napěťového parametru je tedy takový, že pro velké τ má spline tendenci být téměř lineární v důsledku většího napětí.

1.1.4 *Hermiteův spline*

Další modifikací je *Hermiteův spline.* Při této konstrukci požadujeme pouze $\varphi \in C^1[a, b]$, $\varphi|_{[x_i,x_{i+1}]}$ je opět polynom stupně nejvýše 3 a $\varphi(x_i) = f(x_i)$, $\varphi'(x_i) = f'(x_i)$.

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

$$\frac{f(x_{i+1}) - f(x_{i-1})}{x_{i+1} - x_{i-1}}$$

(pro $i = 1, \ldots, n-1$). Pokud je totiž $f \in C^2[a, b]$, je pro $h \to 0$

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

Cvičení 3 Dokažte platnost vztahu $(*)$.

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 ŘEŠENÍ OBYČEJNÝCH DIFERENCIÁLNÍCH ROVNIC

Protože rovnici vyššího řádu lze vždy jednoduše převést na soustavu rovnic prvního řádu, budeme se v této části zabývat pouze rovnicemi prvního řádu, navíc rozřešenými vzhledem k derivaci, tj. rovnicemi tvaru

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

což může reprezentovat buď jednu rovnici nebo celou soustavu – v tom případě řešení $y = (y^1, \dots, y^s) : I\!R \to I\!R^s$ a funkce $f = (f^1, \dots, f^n) : I\!R \times I\!R^s \to I\!R^s$ jsou vektorové funkce. Numerické metody mají své uplatnění hlavně v situacích, kdy přesné (analytické) řešení nedokážeme najít.

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)

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

což je vlastně soustava tří obyčejných diferenciálních rovnic druhého řádu, kterou lze převést na soustavu šesti obyčejných diferenciálních rovnic prvního řádu. Funkce $\boldsymbol{F}$ zde vyjadřuje závislost působící síly na čase, poloze částice a její rychlosti. Obecně je však tato závislost natolik složitá, že není šance tuto soustavu vyřešit analyticky. Mnohdy nám však stačí najít numerickými metodami řešení přibližné.

Příklad 2 I u velmi jednoduchých rovnic se nám může stát, že přesné řešení nedokážeme najít. Typickým příkladem je rovnice

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

o níž je dokázáno, že žádné její řešení nelze vyjádřit pomocí elementárních funkcí a jejich neurčitých integrálů.

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

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

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 řešení lze najít metodou mocninných řad, ale pro většinu argumentů x řada konverguje pomalu. Nahradíme-li Dx^2 členem Dx^3, nelze metodu mocninných řad použít.

2.1 Příklady diskrétních metod

Uvažujme rovnici $y' = f(x, y)$ na intervalu $[a, b]$ s počáteční podmínkou $y(a) = \eta$. Nechť $y : [a, b] \to \mathbb{R}$ je řešení uvažovaného problému. Uvažujme dělení intervalu $[a, b]$ tak, že $a = x_0 < \ldots < x_N \leq b,\ x_n = a + nh$, kde $h > 0$ nazýváme krokem metody. Budeme se snažit přiřadit bodům x_i přibližné hodnoty y_i. Uvedeme zde tři jednoduché příklady.

2.1.1 *Eulerova metoda*

Je to nejjednodušší metoda numerického řešení ODR. Předpokládejme, že máme řešení y diferenciální rovnice a v intervalu $[a, b]$ máme dva body $x_n < x_{n+1}$ uvažovaného dělení. Předpokládejme navíc, že řešení 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:

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

2.1.2 *Rungeova-Kuttova metoda*

Vyjdeme-li ze vztahu uvedeného v předchozím cvičení a vztahů

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

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.

2.1.3 *Dvoukroková metoda*

Společnou vlastností obou uvedených metod bylo, že hodnota y_{n+1} se vypočítávala pouze z *jedné* předcházející hodnoty y_n (a samozřejmě z x_n a h). U vícekrokových metod používáme rekurentní vyjádření z více předešlých 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
$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{1}{2}h^2 y''(x_n) + O(h^3), \tag{1}$$
$$y(x_{n+2}) = y(x_n) + 2hy'(x_n) + 2h^2 y''(x_n) + O(h^3). \tag{2}$$
Odečtením čtyřnásobku (1) od (2) 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
$$y_{n+2} - 4y_{n+1} + 3y_n = -2hf(x_n, y_n),$$
kde $y_0 = \eta$ a y_1 vypočteme pomocí některé jednokrokové metody. Popsaná metoda je dvoukroková – další hodnotu počítáme pomocí dvou předcházejících.

2.2 Obecné jednokrokové metody

Při těchto metodách je dán krok $h > 0$ a počáteční podmínka $y(x_0) = \eta$. Uzly jsou $x_n = a + nh$, $n \geqq 0$. Hodnoty přibližného řešení počítáme podle rekurentního vztahu
$$y_0 = \eta, \quad y_{n+1} = y_n + h\Phi(x_n, y_n, h).$$
Funkce Φ (která závisí na f) se nazývá *přírůstková funkce.* Například u Eulerovy metody jsme měli $\Phi(x,y,h) = f(x,y)$, 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 $e_n = y_n - y(x_n)$, tzv. *akumulovanou diskretizační chybu.* Naším cílem je

(1) Najít odhad e_n v závislosti na h;
(2) ukázat, že v jistém smyslu je $e_n \to 0$ pro $h \to 0$.

Abychom měli zaručenu existenci a jednoznačnost řešení, budeme předpokládat, že $f : [a,b] \times \mathbb{R} \to \mathbb{R}$ je spojitá a že je lipschitzovská vzhledem k y, tj. že

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

Z Picardovy věty (viz přednášky z matematické analýzy) vyplývá, že za těchto předpokladů má úloha

$$y' = f(x,y), \quad y(a) = \eta \tag{$*$}$$

právě jedno řešení $y : [a,b] \to \mathbb{R}$. Navíc předpokládejme (budeme se zabývat jen rozumnými metodami), že i funkce $\Phi : [a,b] \times \mathbb{R} \times [0,h_0] \to \mathbb{R}$ je spojitá a splňuje Lipschitzovu podmínku vzhledem k y.

Definice 2 Řekneme, že jednokroková metoda s přírůstkovou funkcí Φ je konvergentní, jestliže platí následující tvrzení: kdykoli je $y : [a,b] \to \mathbb{R}$ řešením úlohy (*), je

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

kde y_n je přibližné řešení v uzlu x_n.

Definice 3 *Jiná definice konvergence* Krok $h > 0$ nám určuje uzly x_n, k nimž pomocí jednokrokové metody přiřadíme přibližné hodnoty řešení. Označme $e(h)$ maximální chybu, tj.

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

Řekneme, že metoda je *konvergentní*, jestliže platí následující tvrzení: kdykoli je $y : [a,b] \to \mathbb{R}$ řešením úlohy (*), je

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

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

Před odvozením některých jednokrokových metod odvodíme jedno pomocné lemma, které nám bude často užitečné.

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

$$|\xi_{n+1}| \leq A|\xi_n| + B, \qquad n = 0,\dots,N-1.$$

Potom platí odhad

$$|\xi_n| \leq A^n|\xi_0| + \begin{cases} \dfrac{A^n - 1}{A-1}B & \textit{pro } A \neq 1, \\[2ex] Bn & \textit{pro } A = 1. \end{cases} \tag{$*$}$$

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 $(*)$ plyne

$$|\xi_n| \le e^{n\delta}|\xi_0| + \frac{e^{n\delta}-1}{\delta}B. \qquad (**)$$

Pro $L > 0$ a $x \in I\!R$ označme

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

(E_L je tzv. *Lipschitzova funkce.*)

2.2.1 *Konvergence jednokrokových metod*

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

$$|\Phi(x,y,h) - \Phi(x,y^*,h)| \le L|y-y^*|, \quad x \in [a,b],\ y,y^* \in I\!R,\ h \in [0,h_0]. \quad (+)$$

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

$$\Phi(x,y,0) = f(x,y), \qquad x \in [a,b],\ y \in I\!R.$$

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 metody 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 řešení úlohy $y' = f(x, y)$ v intervalu $[a, b]$, $y(a) = \eta$, y_n pro $x_n \in [a, b]$ jsou hodnoty přibližného řešení vypočtené pomocí jednokrokové metody s přírůstkovou funkcí Φ splňující předpoklad (P_Φ). Nechť navíc existují konstanty N, $p > 0$ takové, že*

$$|\Delta(x, y(x), h) - \Phi(x, y(x), h)| \leq Nh^p, \quad x, x + h \in [a, b],\ h \in (0, h_0).$$

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

$$|e_n| \leq Nh^p E_L(x_n - a), \qquad x_n \in [a, b],\ h \in (0, h_0). \tag{2.2.1}$$

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 (2.2.1), dostaneme ihned nerovnost

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

Aplikace lemmatu 1 dává odhad

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

□

Poznámka 9 Jednokrokové metody mají tu výhodu, že v každém kroku (při přechodu od x_n k x_{n+1}) lze měnit délku kroku. Tzn., že můžeme uvažovat obecné dělení $a = x_0 < x_1 < \ldots \leq b$ intervalu $[a, b]$, položit $h_n = x_{n+1} - x_n$ a definovat

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

2.3 Odvození některých jednokrokových metod

2.3.1 *Metody založené na přímém použití Taylorova vzorce*

V tomto odstavci budeme konstruovat jednokrokovou metodu p-tého řádu za předpokladu, že přesné řešení y je třídy C^{p+1}. Z Taylorova vzorce

$$y(x+h) = y(x) + \sum_{k=1}^{p} \frac{y^{(k)}(x)}{k!} h^p + \frac{y^{(p+1)}(\tilde{x})}{(p+1)!} h^{p+1} \tag{2.3.2}$$

plyne

$$\frac{1}{h}(y(x+h) - y(x)) = \sum_{k=1}^{p} \frac{y^{(k)}(x)}{k!} h^{k-1} + \frac{y^{(p+1)}(\tilde{x})}{(p+1)!} h^{p}, \tag{2.3.3}$$

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

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

Pro zjednodušení zápisu definujme *diferenciální operátor* [1] takto: pro funkci $\varphi \in C^1(G)$, kde $G \subset \mathbb{R}^2$, buďiž

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

Mocninou rozumíme skládání, tj. $D^0\varphi := \varphi$ a pro $k \geqq 0$ klademe $D^{k+1}\varphi := D(D^k\varphi)$. Pomocí operátoru D můžeme 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 (2.3.3) dostaneme

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

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

což nás vede k tomu, abychom použili přírůstkovou funkci

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

Pak je totiž

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

kde

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

Cvičení 7 Vypočtěte D^2f, D^3f a pro $f(x,y)=x^2+y^2$ vypočtěte D^4f.

2.3.2 *Rungeovy-Kuttovy metody*

Získali jsme sice jednokrokovou metodu libovolně vysokého řádu, ale pro praxi je téměř nepoužitelná. Za prvé proto, že počítaní mocnin operátoru D vede brzy k příliš složitým výrazům (viz cvičení 7). Druhý (a hlavní) důvod spočívá v tom, že v definici funkce Φ vystupuje nejen funkce f sama, ale i její parciální derivace. V dalším označme

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

Tuto přírůstkovou funkci použijeme k odvození *Rungeových-Kuttových metod*, u nichž se Φ počítá pouze pomocí hodnot f. Jejich přírůstková funkce bude ve tvaru

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

kde

$$\begin{aligned}
k_1(x,y,h)&=f(x,y),\\
k_2(x,y,h)&=f(x+\alpha_2h,y+\beta_{21}hk_1(x,y,h)),\\
&\;\vdots\\
k_i(x,y,h)&=f\left(x+\alpha_ih,y+h\sum_{j=1}^{i-1}\beta_{ij}k_j(x,y,h)\right),\quad i=2,\dots,P.
\end{aligned}$$

Zde P, w_i, α_i a β_{ij} jsou vhodně zvolené konstanty (tak, aby výsledná metoda byla řádu p). V dalším se budeme snažit určit hodnoty těchto konstant tak, abychom dostali odhad $\Phi(x,y,h)-\tilde{\Phi}(x,y,h)=O(h^p)$.

Příklad 5 Pokusme se odvodit Rungeovu-Kuttovu metodu druhého řádu . Tzn., že $p = 2$. Zkusíme zvolit počet členů také $P = 2$. Napišme si 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 rozepišme funkční hodnotu z prvního vztahu podle Taylorova vzorce:[2]

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

Poslední sčítanec na pravé straně ovšem není nic jiného než $O(h^2)$. Dostáváme tedy

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

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 můžeme zvolit; zvolme $w_2 = \alpha \neq 0$ (jinak by nemohl platit vztah $\alpha_2 w_2 = \frac{1}{2}$). Dostaneme

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

[2]Taylorův vzorec pro funkce n proměnných: buď $f \in C^k(\Omega)$, $\Omega \subseteq I\!R^n$ otevřená, $a, b \in \Omega$ a předpokládejme, že i celá úsečka s krajními body a, b leží 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 a vyjádřit $g^{(j)}$ pomocí f.

Tyto hodnoty nám dávají celou třídu Rungeých-Kuttových metod druhého řádu. Nejpoužívanější 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, že máme více metod stejného řádu, vybíráme metodu buď tak, aby koeficienty byly co nejjednodušší (jednoduchá metoda), nebo tak, aby konstanta N v odhadu

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

byla co nejmenší.

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)

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

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:

$$w_1 = w_4 = \tfrac{1}{8}, \qquad w_2 = w_3 = \tfrac{3}{8}$$

$$\alpha_2 = \tfrac{1}{3}, \qquad \alpha_3 = \tfrac{2}{3}, \qquad \alpha_4 = 1$$

$$\beta_{21} = \tfrac{1}{3}, \qquad \beta_{31} = -\tfrac{1}{3}, \qquad \beta_{32} = 1$$

$$\beta_{41} = 1, \qquad \beta_{42} = -1, \qquad \beta_{43} = 1$$

Gillova formule:

$$w_1 = w_4 = \tfrac{1}{6}, \qquad w_2 = \tfrac{1}{3}\left(1 - \tfrac{1}{\sqrt{2}}\right), \qquad w_3 = \tfrac{1}{3}\left(1 + \tfrac{1}{\sqrt{2}}\right)$$

$$\alpha_2 = \alpha_3 = \tfrac{1}{2}, \qquad \alpha_4 = 1$$

$$\beta_{21} = \tfrac{1}{2}, \qquad \beta_{31} = \tfrac{1}{2}(\sqrt{2} - 1), \qquad \beta_{32} = 1 - \tfrac{1}{\sqrt{2}}$$

$$\beta_{41} = 0, \qquad \beta_{42} = -\tfrac{1}{\sqrt{2}}, \qquad \beta_{43} = 1 + \tfrac{1}{\sqrt{2}}$$

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

2.4 Použitelnost odhadů chyb

Podívejme se nyní na efektivnost odhadu, který jsme získali. Uvažujme např. Eulerovu metodu. Ukázali jsme, že pokud je přesné řešení třídy C^2, je

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

kde N je polovina maxima absolutní hodnoty druhé derivace přesného řešení y na intervalu $[a, b]$. Tento odhad se dá o něco zlepšit, uvědomíme-li si, že pro odhad chyby v bodě x_n stačí řešit rovnici na $[a, x_n]$ (interval $(x_n, b]$ nemá žádný vliv).

Můžeme tedy položit $b = x_n$, $N(x_n) = \frac{1}{2}\max_{[a,x_n]}|y''|$ a dostaneme zlepšený odhad

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

Uvažujme dva příklady:

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

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

$$\frac{1}{2}h(e^5 - 1) \le 10^{-3}, \text{ takže}$$

$$h \le \frac{2\ .\ 10^{-3}}{e^5 - 1} \doteq \frac{1}{73707}.$$

Experimentálně však zjistíme, že při volbě $h = 2^{-6} = \frac{1}{64}$ vyjde $e_n = -0,000261$. Vidíme tedy, že odhad z věty je silně nadsazený; při jeho odvozování se v každém kroku počítá s nejhorší možností. Ve skutečnosti většinou dostáváme výsledky výrazně lepší.

Je ale třeba vzít v úvahu, že konstanta N závisí na přesném řešení, které obvykle neznáme. Proto odhad z věty 8 nám dává pouze představu o chování chyby v závislosti na kroku metody a není prakticky použitelný. Proto hledáme odhady jiného typu, které můžeme získat na základě vypočteného přibližného řešení a dat úlohy. Toto jsou tzv. *aposteriorní odhady* . Jedním takovým odhadem se budeme yabývat dále.

2.4.1 *Odhad chyby metodou polovičního kroku*

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

$$\left|\frac{1}{h}(y(x+h) - y(x)) - \Phi(x, y(y), h)\right| \le Nh^p,$$

potom platí

$$|y_n - y(x_n)| \le NE_L(x_n - a)h^p.$$

Dá se navíc ukázat, že pokud jsou funkce f a Φ dostatečně hladké, existuje funkce $e : [a, b] \to I\!R$ taková, že

$$e_n = h^p e(x_n) + O(h^{p+1}). \qquad (2.4.4)$$

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

Předpokládejme, že na tutéž úlohu použijeme postupně metodu s krokem $2h$ a h. Dostaneme uzly a odpovídající hodnoty přibližného řešení:

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

Vidíme, že $x_n^{(2h)} = x_{2n}^{(h)}$. V těchto bodech můžeme porovnat hodnoty přibližných řešení. Užitím vztahu (2.4.4) 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 plyne, že

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

Aproximaci

$$e_{2n}^{(h)} \approx \frac{y_n^{(2h)} - y_{2n}^{(h)}}{2^p - 1} \tag{2.4.5}$$

nazýváme *odhadem chyby metodou polovičního kroku.* Jedná se o tzv. *aposteriorní odhad chyby* , který získáme na základě přibližného řešení (a případně dat úlohy). Numerické experimenty ukazují, že v řadě praktických aplikacích tento odhad dává spolehlivé výsledky.

Poznámka 10 U jednokrokových metod můžeme v libovolném bodě změnit délku kroku. Pracujeme tak, že počítáme s krokem $2h$ a h. Pokud zjistíme, že odhad chyby v nějakém bodě překračuje povolenou mez, pokračujeme od tohoto bodu s polovičním krokem. Naopak, pokud je chyba výrazně nižší než naše tolerance, můžeme krok opět zvětšit (abychom ušetřili časovou náročnost).

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

2.4.2 *Zaokrouhlovací chyby*

Začněme jednoduchým příkladem. Uvažujme 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 $10^{-6} - 1$, což je dostatečně daleko od nuly. Matice soustavy je tedy regulární. Použijeme-li Gaussovu eliminaci, dostaneme řešení

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

Přibližně máme $x_2 \doteq 0,999999$. Počítejme nyní na pět platných číslic. Pak ovšem vyjde $x_2 = 1$, a odtud $x_1 = 0$. Při zkoušce dostaneme v první rovnici $0 + 1 = 1$, ale ve druhé $0 + 1 = 1 \neq 2$, což je příliš velká nepřesnost. Pokud bychom ale rovnice prohodili, vyšly by přibližné hodnoty x_1 i x_2 rovny jedné, tedy správné (i zkouška by ve zvolené přesnosti vyšla správně). Obecně je vhodné při řešení soustavy lineárních rovnic prohazovat rovnice tak, aby se na hlavní diagonále neobjevovala příliš malá čísla. (Tento proces se nazývá *pivotace*.)

Zaokrouhlovací chyby u jednokrokových metod

Dosud jsme přepokládali, že při výpočtu přibližného řešení všechny operace probíhají přesně. Nyní se zamyslíme nad vlivem zaokrouhlovacích chyb. Zaokrouhlování modelujeme tak, že každému $\alpha \in \mathbb{R}$ přiřadíme jeho zaokrouhlenou hodnotu α^*. Pro jednoduchost předpokládejme, že vstupní data jsou už zaokrouhlená, tedy $a = a^*$, $h = h^*$, $\eta = \eta^*$. Přibližné řešení 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 ještě *akumulovanou zaokrouhlovací chybu* v uzlu x_n jako $r_n = \tilde{y}_n - y_n$.

Věta 11 *Buď $\Phi(x, y, h) : [a, b] \times \mathbb{R} \times [0, h_0] \to \mathbb{R}$ L-lipschitzovská v y. Nechť platí $|\varepsilon_k| \leq \varepsilon$ pro $x_k, x_{k+1} \in [a, b]$. 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,

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

Použijeme-li (stejně jako u věty o odhadu diskretizační chyby) „kumulační lemma“ ze cvičení 5, dostaneme

$$|r_n| \leq \frac{(1 + hL)^n - 1}{hL}\varepsilon \leq \frac{e^{nL} - 1}{L}\frac{\varepsilon}{h} = E_L(x_n - a)\frac{\varepsilon}{h}.$$

□

Poznámka 12 Velikost horního odhadu samozřejmě o ničem nesvědčí. Jedná se o *nejhorší možný scénář* (worst scenario0, který může nastat. Numerické experimenty však ukazují, že v praxi je pro malá h opravdu $r_n \approx \frac{1}{h}$. To ve svém důsledku znamená, že zmenšením kroku sice zmenšíme diskretizační chybu metody, ale zvýšíme vliv zaokrouhlovacích chyb.

Hodilo by se tedy najít způsob, jak zjistit velikost zaokrouhlovací chyby. To je možné, pokud můžeme úlohu současně řešit v jednoduché a dvojnásobné přesnosti. Potom bude zaokrouhlovací chyba při dvojnásobné přesnosti zanedbatelná ve srovnání s jednoduchou přesností a můžeme tak získat aproximaci zaokrouhlovacích chyb. Převažuje-li diskretizační chyba, je třeba krok zjemnit, převažuje-li zaokrouhlovací chyba, je třeba krok zvětšit.

2.5 Soustavy lineárních diferenčních rovnic

Označme $I\!N = \{1, 2, 3, \ldots\}$, $I\!N = \{0, 1, 2, \ldots\}$.

Definice 5 Buď $k \in I\!N$, $F_n : I\!R^{k+1} \to I\!R$ (případně $F_n : \mathbb{C}C^{k+1} \to \mathbb{C}C$), $n \in I\!N_0$. Pak systém vztahů

$$F_n(y_n, \ldots, y_{n+k}) = 0, \qquad n \in I\!N_0 \tag{2.5.6}$$

nazýváme *soustavou diferenčních rovnic.* Řešením soustavy nazveme posloupnost $(y_n)_{n=0}^{\infty}$ splňující (2.5.6).

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

$$y_{n+3}^2 - \left(y_{n+2}^2 - y_{n+1}^2 + \sqrt[3]{y_n^2}\right)^4 = 0, \qquad n \in I\!N_0.$$

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

Cvičení 8 Kolik řešení má tato soustava pro pevně zvolené počáteční podmínky y_0, y_1, y_2?

Definice 6 Jsou-li funkce F_n lineární, mluvíme o *soustavě lineárních diferenčních rovnic.* Můžeme ji psát ve tvaru

$$\sum_{\nu=0}^{k} \alpha_{\nu n} y_{n+\nu} = \gamma_n, \quad n \in I\!N_0. \tag{2.5.7}$$

Čísla γ_n se nazývají *pravé strany.* Jsou-li všechny pravé strany nulové, dostaneme homogenní soustavu. Soustavě (2.5.7) můžeme vždy přiřadit homogenní soustavu

$$\sum_{\nu=0}^{k} \alpha_{\nu n} y_{n+\nu} = 0, \quad n \in I\!N_0. \tag{2.5.8}$$

Lemma 2 *Uvažujme soustavu (2.5.7) a k ní příslušnou homogenní soustavu (2.5.8). Pak platí:*

(i) Množina V všech řešení soustavy (2.5.8) je lineární prostor.

(ii) Jsou-li $\{y_n\}_{n=0}^{\infty}$ a $\{z_n\}_{n=0}^{\infty}$ řešení soustavy (2.5.7), je jejich rozdíl řešením soustavy (2.5.8).

(iii) Je-li $\{w_n\}_{n=0}^{\infty}$ řešením (2.5.7) a $\{y_n\}_n \in V$, pak $\{w_n + y_n\}_{n=0}^{\infty}$ je řešením soustavy (2.5.7).

(iv) je-li $\{w_n\}_{n=0}^{\infty}$ řešením soustavy (2.5.7), pak k libovolnému řešení $\{z_n\}_{n=0}^{\infty}$ soustavy (2.5.7) existuje právě jedno $\{y_n\}_n \in V$ takové, že $z_n = y_n + w_n$; jinak řečeno, označíme-li W množinu řešení soustavy (2.5.7), můžeme psát $W = V + w$.

Důkaz pěnecháme čtenáři.

Definice 7 Řekneme, že soustava (2.5.7) je *řádu* k, jestliže všechny koeficienty α_{kn}, $n \in \mathbb{N}_0$, jsou nenulové.

Definice 8 Posloupnosti $\{y_n^\mu\}_{n\in\mathbb{N}_0} \in V$, $\mu = 1,\dots,m$, nazveme *lineárně nezávislé*, jestliže platí:

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

Je zřejmé, že tato podmínka je splněna, právě když matice

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

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

Definice 9 Systém k prvků z V lineárně nezávislých nazveme *fundamentálním systémem* (FS) soustavy homogenních rovnic.

Věta 13 *Nechť $\{z_n^\mu\}_{n\in\mathbb{N}_0} \in V$, $\mu = 1,\dots,k$, je FS řešení soustavy k-tého řádu a $\{y_n\}_{n\in\mathbb{N}_0} \in V$. Pak posloupnost $\{y_n\}_{n\in\mathbb{N}_0}$ je lineární kombinací posloupností $\{z_n^\mu\}_{n\in\mathbb{N}_0}$, $\mu = 1,\dots,k$.*

Důkaz: Z definice FS plyne, že existují konstanty $a_1,\dots,a_k$ takové, že

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

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 $\alpha_{kn} \neq 0$ pro všechna $n \in \mathbb{N}_0$, nutně $z_n = 0$ pro všechna $n \in \mathbb{N}_0$ 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.*

2.5.1 *Nalezení fundamentálního systému*

Definice 10 Soustavou tvaru

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = 0, \qquad n \in \mathbb{N}_0, \tag{2.5.9}$$

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

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

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

Hledejme řešení této soustavy ve tvaru $y_n = \xi^n$, kde $\xi \in C$. 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 (2.5.9) 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 řešením soustavy (2.5.9), právě když ξ je kořenem charakteristického polynomu ρ. Tento polynom má právě k kořenů, počítáme-li každý tolikrát, kolik činí jeho násobnost.

Rozlišme dva případy

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

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

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

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

Tudíž uvažované posloupnosti tvoří FS.

2) *Obecný případ.*

Věta 15 *Nechť* $\xi_1, \ldots, \xi_m \in C$ $(m \leq k)$ *jsou navzájem různé kořeny charakteristického polynomu* ρ *s násobnostmi* $p_1, \ldots, p_m$ *(takže* $k = \sum_{\mu=1}^{m} p_\mu$*). Pak* k *posloupností s prvky*

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

tvoří FS řešení soustavy (2.5.9) řádu k*. (Je-li* $\xi_\mu = 0$*, pak klademe* $0 \,.\, \xi_\mu^{-\nu} = 0$ *pro* $\nu > 0$*.)*

Důkaz: Dokažme, že uvedené posloupnosti jsou řešení uvažované soustavy. Jestliže $\xi_\mu = 0$, pak má charakteristický polynom tvar

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

(tudíž $\alpha_0, \ldots, \alpha_{p_\mu - 1} = 0$) a soustava (2.5.9) má tvar

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

Dosazením snadno ověříme, že posloupnosti (P) jsou řešeními, 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

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

Tzn., že posloupnosti (P) jsou řešeními soustavy (2.5.9).

Dokažme nyní, že řešení (P) jsou lineárně nezávislá, což je ekvivalentní s tím, že příslušná matice $\mathbb{M}$, jejíž řádky jsou tvořeny prvními k prvky posloupností (P), je regulární. V n-tém sloupci této matice ($n = 0, \dots, k-1$) se nachází 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, $n = 0, \dots, k-1$, že $\sum_{n=0}^{k-1} |a_n| > 0$ a

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

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 $\sum_{\mu=1}^{m} p_\mu = k$.

□

2.5.2 *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, jeho kořeny tedy jsou

$$\xi_1, \dots, \xi_r \in I\!R$$

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

(všechny kořeny opakujeme tolikrát, kolik činí jejich násobnost, tedy $r + 2s = k$). Fundamentální systém získaný z naší 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 můžeme převzít do našeho 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 Re $z = \frac{1}{2}(z+\bar{z})$, Im $z = \frac{1}{2i}(z-\bar{z})$ vidíme snadno, že nové posloupnosti jsou lineárními kombinacemi starých, jsou to tedy řešení. K důkazu lineární nezávislosti si stačí uvědomit, že staré posloupnosti lze podobně jednoduše vyjádřit pomocí nových použitím vztahů $z = \text{Re } z + i \text{ Im } z$, $\bar{z} = \text{Re } z - i \text{ Im } z$.

Poznámka 16 Kniha [H] obsahuje kapitolu věnovanou soustavám diferenčních rovnic. Je vidět, že 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 řešení nehomogenní soustavy pomocí fundamentálního systému řešení příslušné homogenní rovnice (*Duhamelův princip*). Tato metoda se podobá variaci konstant, blíže viz [H].

2.6 Vícekrokové metody

Opět budeme hledat přibližné řešení úlohy

$$y' = f(x, y), \qquad y(a) = \eta$$

v uzlech $x_n = a + nh$. Na rozdíl od jednokrokových metod budeme tentokrát počítat hodnoty přibližného řešení y_n pomocí několika předchozích hodnot.

Příklad 7 Dříve jsme odvodili metodu

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

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

$$\sum_{\nu=0}^{k} a_\nu y_{n+\nu} = h \sum_{\nu=0}^{k} \beta_\nu f_{n+\nu}, \tag{2.6.10}$$

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

Pomocí předpisu (2.6.10) můžeme počítat hodnoty přibližného řešení až od y_k. Hodnoty $y_0 (= \eta), y_1, \ldots, y_{k-1}$ se nazývají počáteční podmínky pro k-krokovou metodu. Hodnoty $y_1, \ldots, y_{k-1}$ vypočteme některou jednokrokovou metodou. Podle způsobu výpočtu je třeba rozlišit dva případy.

V jednodušším případě $\beta_k = 0$ máme

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

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

Protože lze hodnotu y_{n+k} přímo vypočítat, mluvíme o *explicitní metodě*. Hlavní výhodou těchto (např. oproti jednokrokovým metodám vyšších řádů) je to, že každou z hodnot f_j stačí vypočítat jednou (v dalších krocích už si ji pamatujeme). Na jednu hodnotu přibližného řešení tak připadá jedna vypočítaná hodnota funkce f.

Ve složitějším 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.* Naše rovnice (obecně nelineární) pro y_{n+k} je tvaru $y_{n+k} = F(y_{n+k})$, kterou řešíme iterační metodou: zvolíme počáteční hodnotu y^0_{n+k} a pro $s \geqq 0$ 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 menší než 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 použití je však třeba odpovědět na dvě otázky: jak volit počáteční hodnotu y^0_{n+k} a kolik iterací provést. Pro volbu počáteční hodnoty je nejjednodušší 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

$$y^0_{n+k} = \frac{1}{\alpha^*_k} \left(h \sum_{\nu=0}^{k-1} \beta^*_\nu f_{n+\nu} - \sum_{\nu=0}^{k-1} \alpha^*_\nu y_{n+\nu} \right). \tag{2.6.11}$$

Potom vypočteme několik aproximací y^s_{n+k} uvažované implicitní metody a klademe $y_{n+k} := y^m_{n+k}$. Nejjednodušší volba je $m = 1$:

$$y_{n+k} = \left(h\beta_k f(x_{n+k}, y^0_{n+k}) + \sum_{\nu=0}^{k-1} \beta_\nu f_{n+\nu} - \sum_{\nu=0}^{k-1} \alpha_\nu y_{n+\nu} \right) / \alpha_k. \tag{2.6.12}$$

Z rovnice (2.6.11) vypočteme y^0_{n+k}, dosadíme do (2.6.12)) a vypočteme y_{n+k}. Tato metoda se nazývá *metoda prediktor-korektor.* Explicitní metoda (2.6.11) se nazývá *prediktorová formule*, implicitní metoda (2.6.12) je *korektorová formule.* Metody prediktor-korektor jsou nejefektivnější vícekrokové metody. V každém kroku se počítají pouze dvě hodnoty funkce f.

Cvičení 9 Uvažujme metodu danou následujícími formulemi:

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

Zapište tuto metodu ve tvaru jednoho vzorce pro výpočet y_{n+1}. Ukažte, že výsledkem je Rungeova-Kuttova metoda druhého řádu.

2.7 Některé vlastnosti obecných vícekrokových metod

Uvažujme obecnou k-krokovou metodu

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = h \sum_{\nu=0}^{k} \beta_\nu f_{n+\nu}, \tag{2.7.13}$$

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

$$y_0 = \eta, y_1, \ldots, y_{k-1}. \tag{2.7.14}$$

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

Definice 11 Řekneme, že k-kroková metoda (2.7.13) je *konvergentní*, jestliže pro řešení $y : [a, b] \to I\!R$ úlohy $y' = f(x, y)$, $y(a) = \eta$ s libovolnou f, která je spojitá v $[a, b] \times I\!R$ a lipschitzovská v y a pro libovolné η platí následující tvrzení: jestliže y_n je přibližné řešení získané pomocí metody (2.7.13) s počátečními podmínkami (2.7.14) takovými, že $\lim\limits_{h\to 0+} \eta_\mu(h) = \eta$, 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 12 Říkáme, že metoda (2.7.13) je *stabilní* (podle Dahlquista), jestliže všechny kořeny polynomu $\rho(\xi) = \sum \alpha_\nu \xi^\nu$ jsou v absolutní hodnotě nejvýše rovny jedné a všechny kořeny, jejichž absolutní hodnota je rovna jedné, jsou jednoduché.

Definice 13 Řekneme, že metoda (2.7.13) je *řádu* $p \geqq 0$, jestliže 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 (2.7.13) řád $p \geqq 1$, říkáme, že je *konsistentní*.

Věta 17 *Každá konvergentní metoda je stabilní.*

Důkaz: Mějme konvergentní metodu (2.7.13). Přibližné řešení y_n konverguje za podmínek uvedených v definici k přesnému řešení. Uvažujme úlohu $y' = 0$, $y(0) = 0$: její přesné řešení je $y(x) = 0$ (tato úloha určitě podmínky splňuje). Pak máme

$$\sum \alpha_\nu y_{n+\nu} = 0 \qquad n = 0, 1, \dots \tag{2.7.15}$$

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

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

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

Vztah (2.7.15) je soustava lineárních diferenčních rovnic s konstantními koeficienty a charakteristickým polynomem $\rho(\xi) = \sum \alpha_\nu \xi^\nu$. Nechť $\xi = re^{i\varphi}$ $(r \geqq 0)$ je kořen polynomu ρ: pak posloupnost $\{\xi^n\}_n$ je řešením (2.7.15). Protože koeficienty α_ν jsou reálné, víme z teorie, že 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 řešení (2.7.15). Jsou to tedy přibližná řešení vyhovující počátečním podmínkám

$$hr^\mu \cos \mu\varphi \xrightarrow{h \to 0+} 0, \qquad hr^\mu \sin \mu\varphi \xrightarrow{h \to 0+} 0$$

Odtud pro každé $x > 0$

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

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

Předposlední ekvivalence vychází z toho, že ta poněkud záhadně vyhlížející limita není ve skutečnosti nic jiného než limita pro $n \to \infty$ s dosazením $h = \frac{x}{n}$.

Nyní mějme kořen $\xi = re^{i\varphi}$ násobnosti alespoň dvě. Potom je $\{n\xi^{n-1}\}_{n=0}^\infty$ řešením (2.7.15). Stejně jako výše vidíme, že 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 řešení. Jsou to tedy přibližná řešení pro počáteční podmínky

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

Odtud stejným způsobem jako v první části dostaneme

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

□

Věta 18 *Je-li metoda (2.7.13) 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: Uvažujme úlohu $y' = 0$ v $[0, b], b > 0$, tentokrát s počáteční podmínkou $y(0) = 1$. Pro přibližné řešení opět platí (2.7.15), 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 (2.7.13) je konvergentní, tedy

$$\forall x > 0 : \lim_{\substack{h \to 0+ \\ x_n = x}} y_n = 1, \tag{2.7.16}$$

přičemž limitu opět chápeme jako limitu pro $n \to \infty$, kde y_n je hodnota přibližného řešení s krokem $h = \frac{x}{n}$. Počáteční podmínky ani koeficienty rovnice (2.7.15) nezávisí na kroku h, nezávisí na něm tedy ani přibližné řešení. Ze vztahu (2.7.16)

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

Provedeme-li tedy limitní přechod v (2.7.15), dostaneme $\sum \alpha_\nu = 0$.

Pro druhou podmínku uvažujme úlohu $y' = 1$, $y(0) = 0$ (řešení je $y(x) = x$). Metoda (2.7.13) tady vypadá takto:

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = h \sum_{\nu=0}^{k} \beta_\nu \qquad n = 0, 1, \ldots. \tag{2.7.17}$$

Pro počáteční podmínky vyžadujeme, aby $y_\mu = \eta_\mu(h) \to 0$ pro $h \to 0+$, $\mu = 0, 1, \ldots, k-1$. Hledejme řešení (2.7.17) ve tvaru $y_n = Kx_n = Khn$. Po dosazení do (2.7.17) 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, že $\sum_{\nu=0}^{k} \alpha_\nu = 0$, takže $\rho(1) = 0$. Z předcházející věty vyplývá, že 1je jednoduchým kořenem polynomu ρ. Tudíž, $\rho'(1) \neq 0$, což znamená, že $\sum_{\nu=0}^{k} \alpha_\nu \nu \neq 0$. Můžeme 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 $y_n = Kx_n,\ n = 0, 1, \ldots,$ řešením (2.7.17). Zvolíme-li počáteční podmínky metody $\eta_\mu(h) = K\mu h \to 0$ pro $h \to 0+$, bude $y_n = Kx_n$ přibližným řešením, získaným metodou (2.7.13). Protože předpokládáme, že je tato metoda konvergentní, máme pro každé x

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

□

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

Cvičení 10 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.

Význam stability metody

Uvažujme rovnici $y' = -y$, $y(0) = 1$ s řešením $y(x) = e^{-x}$. Aplikujeme-li některou z předchozích metod, zjistíme, že pro rostoucí n hodnoty přibližného řešení oscilují okolo přesného se zvětšujícím se rozptylem.

Význam řádu metody

Definice řádu nijak nenaznačuje, k čemu je tento pojem dobrý. To vyplyne z následující věty. Mějme obecnou k-krokovou metodu (2.7.13), $y : [a, b] \to \mathbb{R}$ buď přesné řešení naší úlohy. Dosaďme $y(x_n)$ místo y_n v (2.7.13):

$$\sum_{\nu=0}^{k} \alpha_\nu y(x_{n+\nu}) = h \sum_{\nu=0}^{k} \beta_\nu f(x_{n+\nu}, y(x_{n+\nu})) + h\delta_n. \tag{2.7.18}$$

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

Věta 19 *Nechť metoda (2.7.13) je řádu p a nechť přesné řešení je třídy $C^{p+1}[a, b]$. Pak existují h_0, $K > 0$ tak, že $|\delta_n| \le Kh^p$ pro všechna n taková, že x_n, $x_{n+k} \in [a, b]$ a pro všechna $h < h_0$.*

Důkaz: Uvažujme intervaly $[x_n, x_{n+\nu}] \subset [a, b]$ pro $\nu = 0, \ldots, k$ a $n = 0, 1, \ldots$. Z Taylorova vzorce plyne:

$$y(x_{n+\nu}) = y(x_n + \nu h) = y(x_n) + \sum_{s=1}^{p} \frac{y^{(s)}(x_n)}{s!} \nu^s h^s + R_{\nu,p+1}, \tag{2.7.19}$$

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

Protože y je řešení rovnice, máme $f(x_{n+\nu}, y(x_{n+\nu})) = y'(x_{n+\nu})$; opět použijme Taylorův vzorec:

$$hf(x_{n+\nu}, y(x_{n+\nu})) = hy'(x_n + \nu h) = \sum_{s=1}^{p} \frac{y^{(s)}(x_n)}{(s-1)!}\nu^{s-1}h^s + \tilde{R}_{\nu,p+1}, \quad (2.7.20)$$

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

Zbytky R, $\tilde{R}$ jsou $O(h^{p+1})$, protože $y^{(p+1)}$ je spojitá, a tedy v $[a,b]$ omezená; označme Y maximum její absolutní hodnoty. Vztahy (2.7.19) a (2.7.20) dosadíme do (2.7.18) a upravíme pomocí definice řádu metody:

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

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

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

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 řešení rovnice dostatečně hladké, platí pro chybu $\delta_n = O(h^p)$. Bez důkazu uveďme následující fundamentální výsledek:

Věta 20 *Vícekroková metoda je konvergentní, právě když je stabilní a konzistentní.*

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

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

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

kde $f \in C([a,b]) \times I\!R)$ splňuje Lipschitzovu podmínku s konstantou $L > 0$ a nechť $\{y_n\}_{x_n\in[a,b]}$ je přibližné řešení vypočtené pomocí vícekrokové stabilní metody řádu $p \geqq 1$. Označme

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

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

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

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

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

Definice 14 Vícekroková metoda je *silně stabilní* jestliže má tvar

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

Tzn., že polynom ρ má pouze kořeny 0 a 1 a kořen 1 je jednonásobný.

Nyní dokážeme předchozí větu o odhadu chyby pro případ silně stabilních metod.

Důkaz: Přibližné řešení 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 (M) řádu p a přesné řešení je $C^{p+1}([a,b])$, 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

$$|\delta_n| \le Nh^p, \quad h \in (0,h_0).$$

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

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

Poněvadž funkce f splňuje Lipschitzovu podmínku, máme

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

a tedy

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

Nechť $h_1 > 0$ je takové, že

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

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

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

Tudíž,

$$|e_{n+k}| \leq |e_{n+k-1}| + h \sum_{\nu=0}^{k-1} \vartheta_\nu |e_{n+\nu}| + \tilde{\vartheta}, \tag{2.7.21}$$

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 použijeme následující výsledek.

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

$$u_{n+k} \leq u_{n+k-1} + h \sum_{\nu=0}^{k-1} \vartheta_\nu u_{n+\nu} + \tilde{\vartheta} \tag{2.7.22}$$

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

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

Důkaz:

1) Nechť

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

$$\xi_0 = \delta.$$

Potom podle lemmatu …

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

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

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

Nechť $u_j \leq \xi_j$ pro $j = 0, \dots, n+k-1$, $n \geqq 0$. Potom z (2.7.22) plyne, že

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

3) Použijeme-li výsledek z 1), dostaneme

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

což jsme chtěli dokázat.

□

Nyní můžeme pokračovat v důkazu věty tak, že budeme aplikovat lemma na nerovnosti (2.7.21). Vzhledem k tomu, že $nh = x_n - a$, $\tilde{\vartheta} = \tilde{N}h^{p+1}, \tilde{N} = 2N$, snadno zjistíme, že

$$\begin{aligned} |e_n| &\leq e^{(x_n - a)\vartheta}\delta + \frac{e^{(x_n-a)\vartheta} - 1}{\vartheta h}\tilde{N}h^{p+1} = \\ &= e^{(x_n-a)\vartheta}\delta + E_\vartheta(x_n - a)\tilde{N}h^p, \quad x_n \in [a, b]. \end{aligned}$$

□

Podobným způsobem lze odvodit odhad pro akumulovanou zaokrouhlovací chybu. Stejně jako u jednokrokových metod budeme symbolem $\tilde{y}_n$ značit přibližné řešení 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 $|\varepsilon_n| \leq \varepsilon$ pro $x_n, x_{n+k} \in [a, b]$ a $\tilde{y}_\mu = y_\mu$, $\mu = 0, \dots, k-1$. Přibližné řešení 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 $\tilde{\vartheta} = \varepsilon$. Dostaneme pak odhad akumulované zaokrouhlovací chyby $r_n = y_n - \tilde{y}_n$ ve tvaru

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

Tento odhad nám dává *nejhorší scénář* pro chování chyby r_n v závislosti na h. Numerické experimenty bohužel ukazují, že pro $h \to 0+$ zaokrouhlovací chyby mohou růst tak, že znehodnotí výpočet.

2.8 Odvození některých vícekrokových metod

2.8.1 *Interpolační polynom a zpětné diference*

Nechť $J \subset I\!R$ je interval, $z : J \to I\!R$ a nechť $x_0, \dots, x_q \in J$ ($q \geqq 1$) jsou navzájem různé body. Pak existuje právě jeden interpolační polynom P takový, že

a) stupeň $P \leq q$,

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

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 $x_i = x_0 + ih$ pro $i = 0, \dots, q$ a $h > 0$, 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 $p, p-1, \ldots, p-q$, lze napsat ve tvaru*

$$P(x) = \sum_{m=0}^{q} (-1)^m \binom{-s}{m} \nabla^m z_p, \tag{2.8.23}$$

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

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

Důkaz: Je zřejmé, že polynom na pravé straně vztahu (2.8.23) je stupně $\leq q$. Je-li $x = x_{p-r}$, potom $s = -r$ a $\binom{r}{m} = 0$ pro $m > r$. Pak podle inverzní formule

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

□

2.8.2 *Metody založené na numerické integraci*

Přesné řešení rovnice $y' = f(x, y)$ splňuje vztah

$$y(x+\tau) - y(x) = \int_x^{x+\tau} f(t, y(t)) dt,$$

pro $x, x+\tau \in [a, b]$. Nyní nahradíme funkci $f(t, y(t))$ 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 $x_p, x_{p-1}, \ldots, x_{p-q}$. Uvažovaný interpolační polynom má tvar

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

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

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

kde

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

α) Takto získaným metodám (pro různá q) se říká *Adams-Bashforthovy metody.* Pro zvolené q dostaneme $(q+1)$-krokovou metodu: k výpočtu y_{p+1} používáme hodnoty f v $x_{p-q}, \dots, x_p$, 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) Položme $x = x_{p-1}$, $\tau = h$. Dosaďme do (2.7.13):

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

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

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

Hodnoty zpětných diferencí se počítají poměrně snadno, ale přesto můžeme naše metody přepsat do tvaru, kdy místo hodnot $\nabla^m f_p$ používáme 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

$$y_{p+1} - y_p = \sum_{m=0}^{q} \gamma_m \nabla^m f_p = h \sum_{m=0}^{q} \gamma_m \sum_{\rho=0}^{m} (-1)^\rho \binom{m}{\rho} f_{p-\rho} =$$
$$= h \sum_{\rho_0}^{q} f_{p-\rho} \left((-1)^\rho \sum_{m=\rho} q \binom{m}{\rho} \gamma_m \right).$$

2.8.3 *Příklady některých metod*

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

	metoda	k	p
1.	$y_{n+1} = y_n + hf_n$	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.	$y_{n+1} = y_n + hf_{n+1}$	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 využít uvedené metody pro použití v metodách prediktor-korektor, je výhodné vzít kombinace 1-7 a 2-8. Z teoretického rozboru totiž vyplývá, že v případě, kdy jako prediktor použijeme metodu řádu o jedna menšího než je řád korektoru, celková metoda bude mít stejný řád jako korektor. Dostaneme tedy jednokrokovou metodu řádu 2 a dvoukrokovou metodu řádu 3.

2.9 Metoda sítí pro řešení parciálních diferenciálních rovnic

Uveďme na závěr jednoduchý příklad numerické aproximace parciálních diferenciálních rovnic. Představme si pružnou blánu, na okraji upevněnou, na niž působí kolmo nějaká síla. Modelem bude oblast (otevřená souvislá množina) $\Omega \subset I\!R^2$, sílu budeme reprezentovat hustotou jejího rozložení, tedy funkcí $f : \Omega \to I\!R$ a výchylku blány hledanou funkcí $u : \bar{\Omega} \to I\!R$. Za předpokladu, že jsme dosáhli stabilního stavu (a že výchylka je malá) bude u splňovat vztah $-\Delta u = f$, kde Δ je takzvaný *Laplaceův operátor* definovaný vztahem

$$\Delta u = \sum_{i=1}^{n} \frac{\partial^2 h}{\partial x_i^2} \qquad (h : I\!R^n \to I\!R).$$

Upevnění na okrajích je vyjádřeno okrajovou podmínkou $u|\partial\Omega = 0$. (V přídě jednodimenzionální analogie místo Ω máme interval (a, b), rovnici je $-u'' = f$ a okrajové podmínky $u(a) = u(b) = 0$.) Řešme úlohu přibližně pro $\Omega = (0, 2) \times (0, 1)$. K tomu potřebujeme situaci diskretizovat: obdélník $\bar{\Omega}$ si rozdělme na $2n \times n$ čtverečků o hraně $h = \frac{1}{n}$. Souřadnice uzlů označme $x_i = y_i = ih$. Hodnoty u, f v uzlu (x_i, y_j) označme u_{ij}, f_{ij}. Okrajová podmínka dává $u_{0,j} = u_{2n,j} = u_{i,0} = u_{i,n} = 0$. Podobně jako derivaci funkce jedné proměnné lze aproximovat podle vztahu $y' \approx \frac{1}{h}(y(x + h) - y(x))$, lze aproximovat nesmíšené parciální derivace druhého řádu:

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

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 $\boldsymbol{u} = (u_{1,1}, \ldots, u_{1,n-1}, u_{2,1}, \ldots, u_{2n-1,n-1})^T$ 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

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

a C je $-E$, kde E je jednotková matice. Matice A se skládá z $(2n-1)\times(2n-1)$ bloků typu $(n-1)\times(n-1)$. Matici s takovouto strukturou se říká *blokově třídiagonální matice*. Struktura této matice samozřejmě závisí na očíslování „políček“. Popsané metodě se nazývá *metoda sítí* nebo též *metoda konečných diferencí* .

3

NUMERICKÉ METODY OPTIMALIZACE

V této kapitole se věnujeme výkladu elementů konvexní analýzy v $I\!R^n$ 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 nejblíže k y_i. Použijeme-li metodu nejmenších 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*

Snažíme se najít co nejvýhodnější tvar průřezu žebra. Tento tvar můžeme reprezentovat uzavřenou křivkou okolo počátku; protože chceme, aby tato křivka byla symetrická, můžeme ji reprezentovat pomocí nezáporné funkce $p : [0, b] \to [0, +\infty)$. Naše křivka pak vznikne doplněním grafu této funkce o části souměrně s ním sdružené podle počátku a obou os. Rozložení 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 množství 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 množinu všech takových funkcí (bez požadavku minimality), hledáme minimum

[3]tento příklad má pouze ilustrovat, jak se složitější 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 použití numerických metod se musíme ještě omezit na nějakou množinu $\tilde{U} \subset U$ funkcí určenou konečně mnoha parametry $u_1, \ldots, u_n$. Každé takové n-tici parametrů přiřadíme funkci $p \in U$, k té spočítáme teplotu T, hodnoty $I(p)$ a $F(p)$ a položíme $\tilde{F}(u_1, \ldots, u_n) := F(p)$. Tím už 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, že tato úloha má řadu praktických aplikací, a proto byly vyvinuty různé algoritmy pro její řešení.

3.1 Základy konvexní analýzy

V této části uvedeme některé základní pojmy a výsledky konvexní analýzy, které se využívají při numerickém hledání extrémů. V dalším bude U podmnožina $\mathbb{R}^n$ a J zobrazení U do $\mathbb{R}$.

Definice 15 Hodnotu $J(\bar{u})$, $\bar{u} \in U$ nazveme (absolutním, globálním) *minimem* funkce J na U, jestliže $\forall u \in U : J(u) \geqq J(\bar{u})$. Bod $\bar{u}$ nazýváme *bod minima*, píšeme $J(\bar{u}) = \min_U J$ nebo $\bar{u} = \arg\min_U J$.

Hodnotu $J(\bar{u})$, $\bar{u} \in U$ nazveme *lokálním minimem* funkce J, jestliže existuje okolí $V(\bar{u})$ takové, že $J(\bar{u})$ 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 množina a $J : U \to \mathbb{R}$ spojitá funkce. Potom J nabývá na U minima.*

Definice 16 Nechť $U \subset \mathbb{R}^n$ je neomezená množina a $J : U \to \mathbb{R}$. Řekneme, že funkce J je *koercivní* , jestliže

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

Věta 23 *Nechť $U \subset \mathbb{R}^n$ je neomezená, uzavřená množina. Buď $J : U \to \mathbb{R}$ spojitá a koercivní. Pak J nabývá na U minima.*

Důkaz: Nechť $a \in U$. Z podmínky koercivity vyplývá, že existuje takové $R > 0$, že

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

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: Nechť $U \subset I\!R^n$ je otevřená, $J \in C^1(U)$. Pak pro každé $u \in U$, $\varphi \in I\!R^n$ 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

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

je *gradient* funkce J.

Definice 17 Buď $U \subset I\!R^n$ konvexní množina. Řekneme, že $J : U \to I\!R$ je *konvexní funkce*, jestliže

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

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

Věta 24 *$U \subset I\!R^n$ buď otevřená konvexní množina, $J \in C^1(U)$. Pak*

(i) J je konvexní právě tehdy, když $J(v) \ge J(u) + J'(u, v-u)$ pro každé $u, v \in U$;

(ii) J je ryze konvexní právě tehdy, když $J(v) > J(u) + J'(u, v-u)$ pro každé $u, v \in U$, $u \neq v$.

Důkaz: Nechť 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 uvažujme $\theta \to 0+$, pak

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

Opačná implikace: nechť je pro každé $u, v \in U$ splněna podmínka

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

Z toho plyne

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

Obdobně

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

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 nechť 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 užili (uvažujeme-li ostré nerovnosti) dostaneme, že f je ryze konvexní.

Předpokládejme naopak, že f je ryze konvexní. Z definice vyplývá, že pro $\theta \in (0,1)$, $u \neq v$

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

a odtud ihned plyne

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

□

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 každé $\varphi \in \mathbb{R}^n$ (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 nejvýše jeden bod minima.

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

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

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

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

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

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

přičemž $J'(u_1, u_2 - u_1) = 0$, protože u_1 je bod minima. To je ovšem spor. $\square$

Nechť $J \in C^2(U)$, $U \subset \mathbb{R}^n$. Pak má J totální diferenciál druhého řádu v libovolném bodě $u \in U$ a směrové derivace druhého řádu $J''(u, \varphi, \psi) = \varphi^T J''(u)\psi$, 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). Taylorův vzorec lze psát ve tvaru

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

pro nějaké $\theta \in [0, 1]$. Použitím tohoto vzorce lze dokázat tuto větu:

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

Důkaz: Nechť J je dvakrát spojitě diferencovatelná. Nejprve dokážeme koercivitu. Z Taylorova vzorce máme pro $\theta \in [0, 1]$

$$\begin{aligned} J(u) &= J(0) + J'(0, u) + \frac{1}{2} J''(\theta u, u, u) \geq \\ &\geq J(0) - M\|u\| + \frac{\alpha}{2}\|u\|^2 \xrightarrow{\|u\| \to \infty} +\infty. \end{aligned}$$

Jest totiž podle Cauchyovy nerovnosti

$$|J'(0, u)| = |\nabla J(0) \cdot u| \leq \underbrace{\|\nabla J(0)\|}_{=M} \|u\|$$

a podle předpokladů je

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

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

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

čímž je důkaz hotov. $\square$

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.

Nechť $J : \mathbb{R}^n \to \mathbb{R}$ a je dán počáteční bod $u^0 \in \mathbb{R}^n$. Nechť jsme již určili $u^k \in \mathbb{R}^n$. Pak položíme

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

kde $k \geq 0$, $\varphi^k \in \mathbb{R}^n$, $\rho_k > 0$. Veličinu φ nazýváme *směr spádu*, ρ_k určuje vzdálenost – *délku kroku* ve směru φ^k. Pro praktické použití 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 dostatečněmalé). 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 *Nechť je $J \in C^2(\mathbb{R}^n)$ a platí (3.2.2). Potom existuje $\tilde{\rho}_0 > 0$ takové, že $J(u^{k+1}) < J(u^k)$, 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 $u^k + \rho_k \varphi^k$. Nechť $K > 0$. Potom existuje $M > 0$ tak, že

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

Máme tedy

$$J(u^{k+1}) = J(u^k) + \rho_k \left(\underbrace{J'(u^k, \varphi_k)}_{<0} + \frac{1}{2}\rho_k \underbrace{J''(\beta, \varphi^k, \varphi^k)}_{|J''| \leq M} \right).$$

Odtud plyne, že existuje $\tilde{\rho}_0 > 0$ tak, že

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

$\square$

Příklad 12 *Směr největšího spádu* definujeme jako

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

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 dalším uvedeme a budeme analyzovat dva typy metody největšího spádu.

Algoritmus I - metoda největšího spádu s konstantním krokem

Nechť je dáno: $u^0 \in I\!R^n$, $M \geq 1$ celé, $\varepsilon > 0$, $\rho > 0$ (krok). 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 největšího 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 dalším 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.

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

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

Zvolme $\rho = \frac{2}{\lambda+\Lambda}$. Pak posloupnost $\{u^k\}_{k=0}^{\infty}$ daná algoritmem 1 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

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

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 1,

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

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

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

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 $\mathbb{R}^n$:

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

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

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

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

Důkaz: Nechť $\mathrm{Sp}(\mathbb{A}) = \{\lambda_1, \ldots, \lambda_n\}$, kde $|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_m|$. 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 $c_i \in \mathbb{R}^n$ takové, že $u = \sum_{i=1}^n c_i \varphi_i$. Dále snadno zjistíme, že

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

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

Lemma 6 *Nechť*

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

Pak

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

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

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

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 $\varphi \in \mathbb{R}^n$ máme

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

Integrací dostaneme

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

a tedy

$$\left| \varphi^T \left(I - \rho \int_0^1 J''(\overline{u} + t(u^k - \overline{u}))\,\mathrm{d}t \right) \varphi \right| \leq q := \max\left(|1-\rho\Lambda|,\ |1-\rho\lambda|\right).$$

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

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

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

Nevýhodou algoritmu 1 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 2.

Věta 30 *Nechť v algoritmu 2 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 2 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

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

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

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

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

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

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

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

V důsledku této nerovnosti platí:

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

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 (3.2.15), (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 přdpoklady 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 řešení diferenciálních rovnic*, 1981.
2. Přikryl P.: *Numerické metody matematické analýzy*, (MVŠT 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.*