Akademie věd České republiky

Základy teorie numerických metod pro řešení diferenciálních rovnic

OBSAH

PŘEDMLUVA

Kniha, kterou dostává čtenář do rukou, si klade za cíl podat elementární a ne příliš rozsáhlý úvod do problematiky numerického řešení diferenciálních rovnic. Vznikla na základě mého dlouholetého působení jako učitele na matematicko-fyzikální fakultě Karlovy University, kdy moje přednáška tvořila (a dosud tvoří) tu část základního čtyřsemestrového kursu numerických metod pro posluchače specializace numerická matematika, která se týká diferenciálních rovnic. Obsah knihy je tedy v podstatě dán studijním plánem zmíněné specializace a zahrnuje základní metody pro řešení obyčejných diferenciálních rovnic (úloh s počátečními podmínkami i okrajových úloh) a parciálních diferenciálních rovnic eliptického a parabolického typu. Historicky vzniklo, že diferenciální rovnice hyperbolického typu nejsou zatím pojaty do základního kursu numerické matematiky, což tvoří jeden z důvodů, proč je čtenář nenalezne ani v této knize. Druhý důvod je ten, že chování rovnic hyperbolického typu je natolik odlišné od chování eliptických a parabolických rovnic, že by se při jejich vyšetřování dalo užít jen velmi málo z ostatního obsahu knihy. Její rozsah by tedy vzrostl neúnosně. Určitou mezeru, která tak vznikne, bude třeba zaplnit teprve v budoucnosti.

Kniha je rozdělena do čtyř kapitol. V první kapitole věnované problematice numerického řešení úloh s počátečními podmínkami pro obyčejné diferenciální rovnice se studují běžné lineární k-krokové metody a obecné jednokrokové metody čítaje v to důležitý speciální případ metod Runga a Kutty. Kromě toho jsou zde pojednány také problémy spojené s řešením diferenciálních rovnic se silným tlumením. Obsah druhé kapitoly je tvořen základními metodami pro řešení okrajových úloh pro obyčejné diferenciální rovnice. Omezujeme se zde převážně na lineární problémy a jmenovitě jde o metody převodu na úlohy s počátečními podmínkami (metoda střelby, metoda přesunu okrajové podmínky), o metodu sítí a o základy variačních metod zejména pak o metodu konečných prvků. Kapitola třetí pojednává o metodě sítí a o základech metody konečných prvků pro řešení parciálních diferenciálních rovnic eliptického typu. Konečně v kapitole čtvrté si všímáme problematiky přibližného řešení parciálních diferenciálních rovnic parabolického typu. Zabýváme se zde metodou sítí a klasickými i variačními semidiskrétními metodami. Kromě toho je

ke každé kapitole připojen komentovaný seznam doporučené literatury čítaje v to samozřejmě ty publikace, které jsou v textu bezprostředně použity. Nesnažil jsem se přitom podat vyčerpávající výčet literatury, ale spíše ukázat studentu základní prameny, z nichž může rozšířit své znalosti.

Už na začátku jsem uvedl, že kniha představuje elementární úvod do výše popsané problematiky. Rád bych k tomu dodal, že mně při jejím psaní šlo o to dát čtenáři základní orientaci zejména v teoretických otázkách spjatých se studovanými metodami. Měl jsem přitom na mysli ani ne tak podat co nejobecnější výsledky a co nejúplnější seznam stávajících metod, jako zdůraznit základní myšlenku příslušného postupu a užívat přitom co nejméně předběžných matematických znalostí. Skutečně k tomu, aby čtenář plně porozuměl celé knize, vystačí pouze s těmi nejelementárnějšími pojmy a poznatky z lineární algebry a reálné analýzy (pojem vektoru, matice a lineárního prostoru, základní vlastnosti derivace a integrálu). Jen na několika málo místech se užívají některá speciálnější tvrzení z teorie funkcí komplexní proměnné.

Kniha je zaměřena tedy spíše teoreticky, proto i cvičení připojená k jednotlivým kapitolám jsou teoretická a slouží k hlubšímu porozumění vykládané látky. S tím souvisí i charakter konkrétních numerických příkladů, které v ní čtenář nalezne. Tyto příklady nejsou myšleny jako přímé vzory k řešení konkrétních praktických problémů — k tomu slouží, nebo by měla sloužit praktická cvičení přímo na počítači — ale jako ilustrace někdy i nečekaných jevů, s nimiž se lze v praxi za většinou podstatně složitějších okolností setkat. Mimoto slouží některé z těchto příkladů jako motivace k prováděným teoretickým úvahám.

Závěrem je mou milou povinností poděkovat všem, kteří v té či oné míře přispěli k vytvoření této knížky. Jsou to zejména všichni moji kolegové z oddělení konstruktivních metod matematické analýzy Matematického ústavu AV ČR a řada kolegů z katedry numerické matematiky matematicko-fyzikální fakulty Karlovy univerzity, kteří byli kdykoliv ochotni vést se mnou dlouhé diskuse o obsahu a zaměření knihy. Jmenovitě bych rád uvedl prof. dr. Ivo Marka, člena korespondenta AV ČR, který jako vědecký redaktor věnoval publikaci péči a pozornost daleko přesahující běžné zvyklosti. Dále pak vděčím svým přátelům RNDr. Petru Přikrylovi, CSc., a RNDr. Michalu Křížkovi, DrSc., kteří pečlivě přečetli celý rukopis a řadou připomínek přispěli ke zkvalitnění textu. Co nejvřeleji také děkuji Radě Fondu pro vydávání vědecké literatury AV ČR, jejíž subvence byla pro vydání knihy rozhodující. Manželům Šidákovým pak patří můj dík za nesmírně pečlivou přípravu matrice pro tisk. Konečně co nejupřímněji děkuji své ženě za oporu, kterou mně byla při psaní této knihy, a za trpělivost, se kterou znovu a znovu přepisovala náročný rukopis.

Praha, leden 1989 *Autor*

Kapitola I.

Obyčejné diferenciální rovnice – úlohy s počátečními podmínkami

1 Úvod

V této kapitole se budeme zabývat problematikou numerického řešení úloh s počátečními podmínkami pro obyčejné diferenciální rovnice. *Úlohou s počátečními podmínkami* pro soustavu m diferenciálních rovnic prvního řádu

$$
\begin{aligned}
{}^1y' &= {}^1f(x, {}^1y, \dots, {}^my), \\
{}^2y' &= {}^2f(x, {}^1y, \dots, {}^my), \\
&\vdots \\
{}^my' &= {}^mf(x, {}^1y, \dots, {}^my)
\end{aligned} \tag{1.1}
$$

rozumíme úlohu nalézt m funkcí ${}^1y, \dots, {}^my$ proměnné x definovaných, spojitých a spojitě diferencovatelných v intervalu $\langle a, b\rangle$, které vyhovují rovnicím (1.1) a pro něž platí

$$
{}^1y(\xi) = \eta_1, {}^2y(\xi) = \eta_2, \dots, {}^my(\xi) = \eta_m, \tag{1.2}
$$

kde $\boldsymbol{\eta} = (\eta_1, \dots, \eta_m)^T$ je daný m-dimenzionální vektor a ξ pevně zvolený bod v intervalu $\langle a, b\rangle$. Často to bývá levý krajní bod daného intervalu, což také v dalším výkladu budeme předpokládat. Doplňující podmínky (1.2) kladené na řešení soustavy (1.1) se nazývají *počáteční podmínky*.

Protože každou diferenciální rovnici m-tého řádu

$$
y^{(m)} = f(x, y, y', \dots, y^{(m-1)}) \tag{1.3}
$$

lze psát po zavedení nových neznámých funkcí

$$
\begin{aligned}
{}^1y &= y, \\
{}^2y &= {}^1y' \ (\equiv y'), \\
{}^3y &= {}^2y' \ (\equiv y''), \\
&\vdots \\
{}^my &= {}^{(m-1)}y' \ (\equiv y^{(m-1)}),
\end{aligned} \tag{1.4}
$$

jako soustavu diferenciálních rovnic prvního řádu

$$
\begin{aligned}
{}^1y' &= {}^2y,\\
{}^2y' &= {}^3y,\\
&\vdots\\
{}^my' &= f(x, {}^1y, \dots, {}^my),
\end{aligned}
\tag{1.5}
$$

je i tento typ rovnic zahrnut v případě (1.1).

Při vyšetřování konkrétních metod se skoro všude omezíme na jednu diferenciální rovnici

$$y' = f(x, y), \quad x \in \langle a, b\rangle \tag{1.6}$$

s počáteční podmínkou

$$y(a) = \eta. \tag{1.7}$$

Toto omezení není většinou nikterak podstatné. Přejdeme-li totiž k vektorovému zápisu a položíme-li

$$
\begin{aligned}
&y = ({}^1y, \dots, {}^my)^T, \quad f = ({}^1f, \dots, {}^mf)^T,\\
&\eta = (\eta_1, \dots, \eta_m)^T,
\end{aligned}
\tag{1.8}
$$

můžeme soustavu (1.1) s počátečními podmínkami (1.2) psát ve tvaru

$$y' = f(x, y), \quad y(\xi) = \eta. \tag{1.9}$$

K přenesení výsledků, které zformulujeme pro případ rovnice (1.6) s počáteční podmínkou (1.7) na případ soustavy (1.1) s počátečními podmínkami (1.2) tedy stačí obvykle pouze předpokládat, že y v rovnici (1.6) je vektor a f vektorová funkce.

Dále se omezíme na *metody diskrétní*, které spočívají v tom, že přibližné hodnoty hledané funkce hledáme pouze v nějaké diskrétní množině bodů $x_n \in \langle a, b\rangle$, $n = 0, 1, \dots, N$. Tyto body budeme volit pro jednoduchost převážně ekvidistantní, tj. položíme

$$x_n = a + nh, \quad n = 0, 1, \dots, \tag{1.10}$$

kde h je konstanta, kterou nazveme *integračním krokem*. Rovněž toto omezení se většinou snadno odstraní.

Naproti tomu metody typu *postupných* (nebo *Picardových*) *aproximací*, které spočívají v tom, že se diferenciální rovnice (1.6) s počáteční podmínkou (1.7) nahradí ekvivalentní integrální rovnicí

$$y(x) = \eta + \int_a^x f(t, y(t))\,\mathrm{d}t \tag{1.11}$$

a pro tuto rovnici se konstruují postupné aproximace předpisem

$$y_{n+1}(x) = \eta + \int_a^x f(t, y_n(t))\,\mathrm{d}t, \quad n = 1, 2, \dots, \tag{1.12}$$

a které hrají významnou roli v teorii diferenciálních rovnic, ponecháváme stranou našeho zájmu. Důvod je ten, v že efektivitě nemohou konkurovat výše zmíněným diskrétním metodám.

O pravé straně f diferenciální rovnice (1.6) budeme předpokládat všude v této kapitole, i když to nebudeme vždy znovu opakovat, že

(i) je definovaná a spojitá jako funkce dvou proměnných v pásu $a \leqq x \leqq b$, $-\infty < y < \infty$ (a, b jsou konečná čísla) a že

(ii) splňuje v tomto pásu Lipschitzovu podmínku vzhledem k y s konstantou nezávislou na x, tj. že existuje konstanta L taková, že platí

$$|f(x, y) - f(x, z)| \leqq L|y - z| \tag{1.13}$$

pro každé $x \in \langle a, b\rangle$ a pro libovolná y a z.

K splnění této podmínky stačí např., aby ve zmíněném pásu měla pravá strana dané diferenciální rovnice omezenou první parciální derivaci podle y, jak plyne ihned z věty o střední hodnotě. To je splněno zejména v případě lineární diferenciální rovnice.

Globální předpoklady o pravé straně dané diferenciální rovnice, které jsme zformulovali, zaručují, jak hned ukážeme, existenci a jednoznačnost řešení úlohy (1.6), (1.7) v celém intervalu $\langle a, b\rangle$, zatímco obvyklé lokální předpoklady, které se uvádějí ve většině příruček o teorii obyčejných diferenciálních rovnic, zaručují existenci řešení pouze lokálně. Uvedené předpoklady nás tedy ušetří úvah, zda v té části intervalu, do níž jsme dospěli s přibližným řešením, ještě řešení existuje apod. Na druhé straně je však předpoklad (1.13) značně silný a často se tedy v praktické situaci stane, že není splněn. To však není tak tragické, jak se může na první pohled zdát, neboť numericky většinou řešíme takové diferenciální rovnice, o nichž už předem máme k dispozici nějakou informaci umožňující odhadnout polohu hledaného řešení. Pak už obvykle snadno změnou definice dané pravé strany v těch částech pásu $a \leqq x \leqq b$, $-\infty < y < \infty$, kde řešení zaručeně neleží, dosáhneme splnění předpokladu (ii).

Další důležitá přednost předpokladů (i), (ii) souvisí s tím, že u všech metod, které uvedeme, bude v příslušném algoritmu nutno vypočítávat hodnoty $f(x, y)$, kde x bude ležet v intervalu $\langle a, b\rangle$, avšak hodnotu y nebudeme mít možnost předem odhadnout. Předpoklad, že funkce f je definována pro libovolné y, teoretické zkoumání takové metody značně zjednoduší.

Jako nejjednodušší příklad metody, kterou máme na mysli, uveďme tzv. *Eulerovu metodu*, ve které přibližné hodnoty y_n řešení $y(x)$ diferenciální rovnice (1.6)

v bodech $x = x_n$ daných rovnicí (1.10) počítáme z rekurence

$$y_0 = \eta,$$
$$y_{n+1} = y_n + hf(x_n, y_n), \quad n = 0, 1, \ldots, x_n \in \langle a, b\rangle . \tag{1.14}$$

Zavedení této metody je motivováno geometricky. Geometricky řečeno udává totiž diferenciální rovnice (1.6) v pásu $a \leqq x \leqq b$, $-\infty < y < \infty$ směrové pole. Nalézt její řešení splňující danou počáteční podmínku (1.7) značí tedy nalézt křivku, jejíž graf prochází bodem (a, η) a má v každém svém bodě směrnici určenou tímto směrovým polem. Eulerova metoda v této interpretaci tedy značí aproximaci hledané křivky lomenou čarou procházející body (x_i, y_i), $i = 0, 1, \ldots$, přičemž směrnice úseček, které ji vytvářejí, souhlasí se směrovým polem vždy v levém krajním bodě.

Z praktického hlediska je jistě rozumné žádat, aby přibližné řešení y_n získané danou metodou konvergovalo v nějaké vhodném smyslu při zmenšování integračního kroku k přesnému řešení y dané úlohy. Tento požadavek totiž aspoň teoreticky zaručí, že řešení úlohy (1.6), (1.7) lze získat s libovolnou přesností volbou dostatečně malého h. Zavedeme-li tedy tzv. *celkovou diskretizační chybu* e_n rovnicí $e_n = y_n - y(x_n)$, bude její chování jako funkce integračního kroku h důležitou charakteristikou dané metody. Zdůrazněme, že chování celkové diskretizační chyby při $h \to 0$ je třeba vyšetřovat při pevném $x_n = x \in \langle a, b\rangle$, neboť jen tak udává tato veličina velikost chyby, které se dopouštíme při aproximaci hodnoty přesného řešení v bodě x. Vzhledem k rovnici $x_n = a + nh$ to má ovšem za následek, že při zvoleném x probíhá integrační krok nikoliv spojitou, ale diskrétní množinu, a to těch hodnot h, pro něž je $(x - a)/h$ celé číslo; při $h \to 0$ musí tedy platit $n \to \infty$.

Konvergenční vlastnosti celkové diskretizační chyby budeme vyšetřovat na různé úrovni. Podaří-li se dokázat, že pro $h \to 0$ a $x_n = x$ platí $e_n \to 0$, mluvíme o *prosté konvergenci*. Nalezneme-li funkci $\varphi(h)$, pro niž je $\lim \varphi(h) = 0$ pro $h \to 0$ a platí-li $e_n = O\big(\varphi(h)\big)$ pro $h \to 0$ a $x_n = x$, řekneme, že jsme zjistili *rychlost konvergence*. Zápis $e_n = O\big(\varphi(h)\big)$ totiž znamená, že existuje konstanta M taková, že pro malá h a pro n definovaná rovnicí $x_n = x$ platí $|e_n| \leqq M\varphi(h)$. Chyba tedy konverguje k nule nejméně tak rychle jako známá funkce $\varphi(h)$ (v praxi je funkce φ nejčastěji tvaru h^p). Ještě silnější informací je *odhad chyby* $|e_n| \leqq \psi(h)$, kde ψ je známá funkce, pro niž platí $\psi(h) \to 0$ pro $h \to 0$. Konečně, podaří-li se nalézt funkci η takovou, že platí $e_n/\eta(h) \to 0$ pro $h \to 0$ a $x_n = x$, mluvíme o funkci η jako o *asymptotickém odhadu chyby*.

Ze všech právě uvedených informací o chování chyby se zdá být na první pohled nejužitečnější odhad chyby, neboť jen on umožňuje volit předem velikost integračního kroku tak, aby chyba, které se dopustíme, nepřesáhla předepsanou hodnotu. V dalším však uvidíme, že tato úvaha vede skoro vždy k zbytečně malým integračním krokům, takže ji z praktických důvodů nelze většinou použít. Proto znalost rychlosti konvergence nebo asymptotického chování chyby poskytuje často podstatně více prakticky použitelné informace o dané metodě než odhad chyby.

Pro získání představy o celkové diskretizační chybě bývá mnohdy užitečné znát

tzv. *lokální diskretizační chybu* dané metody. Rozumíme jí chybu, které se dopustíme tím, že provedeme jeden krok dané metody a předpokládáme, že všechny hodnoty, které k jeho realizaci potřebujeme, jsou přesné. V konkrétním případě Eulerovy metody je tedy lokální diskretizační chyba dána výrazem

$$L\big(y(x); h\big) \equiv y(x + h) - y(x) - hf\big(x, y(x)\big), \tag{1.15}$$

kde y je přesné řešení dané diferenciální rovnice. I lokální diskretizační chyba se většinou vyjadřuje jako funkce integračního kroku. Při numerickém řešení diferenciální rovnice se dopouštíme lokální diskretizační chyby v každém kroku, a celková diskretizační chyba je tedy výsledkem nakupení lokálních diskretizačních chyb. Přitom je třeba brát navíc v úvahu, že každý krok vychází z hodnot, které už jsou zatíženy chybou z předešlého výpočtu. Je tedy žádoucí, aby daná metoda měla tu vlastnost, že při jejím užití nedochází ke katastrofální akumulaci lokálních diskretizačních chyb. O metodách, u nichž tomu tak je, se obvykle mluví jako o stabilních metodách. Pojmů stability je přitom řada; některé z nich v dalším rovněž vyšetříme.

Při všech předchozích úvahách jsme mlčky předpokládali, že přibližné řešení splňuje rovnici Eulerovy metody (nebo rovnice složitější metody) přesně. Tak tomu v důsledku toho, že aritmetické operace jsme schopni provádět pouze s konečným počtem čísel o konečném počtu cifer, samozřejmě není a místo veličiny y_n vypočteme veličinu $\tilde{y}_n$, která splňuje rovnici (1.14) pouze přibližně. Celková chyba přibližného řešení, které dostaneme jako výsledek konkrétního výpočtu, se tedy skládá ze dvou částí: celkové diskretizační chyby e_n a z veličiny $r_n = \tilde{y}_n - y_n$, kterou nazveme *celkovou zaokrouhlovací chybou*. Mají-li být naše znalosti o dané metodě uspokojivé, musíme mít představu i o chování této zaokrouhlovací chyby.

Všechny právě zmíněné aspekty budeme mít na mysli při vyšetřování konkrétních metod v dalších článcích této kapitoly. Začneme přitom podrobným vyšetřením Eulerovy metody, ne ani proto, že bychom ji nějak zvlášť doporučovali pro praktické počítání — v efektivitě nemůže většinou soutěžit s důmyslnějšími metodami, jak uvidíme — ale hlavně proto, že mnohé charakteristické postupy použitelné i u složitějších metod, jsou u ní snadněji patrné, protože nejsou tolik zastřeny komplikovaným zápisem. Protože však Eulerovu metodu použijeme v důkazu slíbené existenční věty, zformulujeme jednu její jednoduchou vlastnost už zde. Začneme pomocným tvrzením, které bude užitečné i v dalších článcích.

Lemma 1.1. *Buďte A a B nezáporné konstanty a nechť pro posloupnost čísel $\varphi_0, \ldots, \varphi_N$ platí*

$$|\varphi_{n+1}| \leqq A|\varphi_n| + B, \quad n = 0, 1, \ldots, N - 1. \tag{1.16}$$

Pak platí

$$|\varphi_n| \leqq A^n|\varphi_0| + \begin{cases} \frac{A^n - 1}{A - 1}B, & A \neq 1 \\ nB, & A = 1 \end{cases} \tag{1.17}$$

pro $n = 0, \ldots, N$.

D ů k a z se provede velice snadno např. úplnou indukcí, a proto jej přenecháme čtenáři.

Obsahem následujícího lemmatu je apriorní odhad přibližného řešení získaného Eulerovou metodou.

Lemma 1.2. *Nechť f splňuje předpoklady* (i) a (ii) a *nechť h je libovolné kladné číslo. Nechť N je největší takové číslo, že je $x_N \in \langle a, b\rangle$ a nechť $y_0, \dots, y_N$ je posloupnost získaná Eulerovou metodou* (1.14). *Pak platí*

$$|y_n| \leqq Y, \quad n = 0, \dots, N, \tag{1.18}$$

kde

$$Y = |\eta| e^{L(b-a)} + cE_L(b-a), \tag{1.19}$$

$$E_L(x) = \frac{\mathrm{e}^{Lx} - 1}{L} \quad \text{pro } L > 0, \qquad = x \quad \text{pro } L = 0 \tag{1.20}$$

a

$$c = \max_{x \in \langle a, b\rangle} |f(x, 0)|. \tag{1.21}$$

D ů k a z . Předně si všimněme, že číslo c definované rovnicí (1.21) je konečné, neboť z předpokladu (i) plyne, že funkce $|f(\cdot, 0)|$ je spojitá na kompaktní množině. Z (1.14) plyne ihned, že platí

$$|y_{n+1}| \leqq |y_n| + h|f(x_n, y_n)|. \tag{1.22}$$

Z předpokladu (ii) a trojúhelníkové nerovnosti dostáváme

$$|f(x_n, y_n)| - |f(x_n, 0)| \leqq |f(x_n, y_n) - f(x_n, 0)| \leqq L|y_n|, \tag{1.23}$$

neboli

$$|f(x_n, y_n) \leqq L|y_n| + |f(x_n, 0)| \leqq L|y_n| + c, \quad n = 0, \dots, N, \tag{1.24}$$

neboť je $x_n \in \langle a, b\rangle$. Dosadíme-li tento výsledek do nerovnosti (1.22), máme

$$|y_{n+1}| \leqq (1 + hL)|y_n| + ch \tag{1.25}$$

a z lemmatu 1.1 plyne, že platí

$$|y_n| \leqq (1 + hL)^n |\eta| + \frac{(1 + hL)^n - 1}{hL} hc. \tag{1.26}$$

pro $L > 0$ a

$$|y_n| \leqq |\eta| + nhc \tag{1.27}$$

pro $L = 0$. Použijeme-li nyní zřejmé nerovnosti $1 + hL \leqq \mathrm{e}^{hL}$ a faktu, že $nh = x_n - a \leqq b - a$, dostáváme požadovaný výsledek ihned z nerovností (1.26) a (1.27).

Nyní už se dostáváme k formulaci slíbené globální existenční věty.

Věta 1.1. *Nechť pravá strana f diferenciální rovnice* (1.6) *splňuje předpoklady* (i) a (ii) *a nechť η je libovolné reálné číslo. Pak existuje jediná funkce y, pro niž platí*

(i) *y je definována a spojitá v intervalu $\langle a, b\rangle$,*

(ii) *y je spojitě diferencovatelná v $\langle a, b\rangle$ a platí $y'(x) = f(x, y(x))$ pro každé $x \in \langle a, b\rangle$,*

(iii) *$y(a) = \eta$.*

D ů k a z . K důkazu existence řešení použijeme, jak už jsme se zmínili v předchozím textu, Eulerovu metodu. Označme k tomu cíli symbolem H libovolnou posloupnost kladných čísel h takových, že $h \to 0$ a že $(b-a)/h$ je celé číslo a uvažujme posloupnost $y_0, \dots, y_N$ ($N = (b-a)/h$) získanou Eulerovou metodou (1.14) s integračním krokem $h \in H$. Přiřaďme dále této posloupnosti (a tedy vlastně integračnímu kroku $h \in H$) spojitou funkci y_h tak, že její graf je lomená čára složená z úseček spojujících body (x_n, y_n), $n = 0, \dots, N$. Je tedy

$$y_h(x) = y_n + f(x_n, y_n)(x - x_n) \tag{1.28}$$

pro $n = 0, \dots, N-1$ a pro $x_n \leqq x \leqq x_{n+1}$. Protože podle lemmatu 1.2 platí, že je $|y_n| \leqq Y$ pro $n = 0, \dots, N$ a číslo Y nezávisí na h, je právě zavedená posloupnost funkcí $\{y_h;\, h \in H\}$ stejnoměrně ohraničená. Dále, každá funkce y_h je po částech diferencovatelná a platí

$$y_h'(x) = f(x_n, y_n), \quad x_n < x < x_{n+1}, \quad n = 0, \dots, N-1. \tag{1.29}$$

Položíme-li nyní

$$R = \{(x, y);\, a \leqq x \leqq b, |y| \leqq Y\}, \tag{1.30}$$

je R kompaktní a funkce f je na ní vzhledem k předpokladu (i) ohraničená. Pro každé $h \in H$ a pro každé $x \in x_n$, $n = 0, \dots, N$ tedy platí

$$|y_h'(x)| \leqq M, \tag{1.31}$$

kde

$$M = \max_{(x,y) \in R} |f(x, y)|. \tag{1.32}$$

Odtud ihned dostáváme, že pro libovolná x, $x^* \in \langle a, b\rangle$ platí

$$|y_h(x) - y_h(x^*)| = \left| \int_x^{x^*} y_h'(t)\, \mathrm{d}\, t \right| \leqq M|x^* - x|. \tag{1.33}$$

Posloupnost $\{y_h;\, h \in H\}$ je tedy na intervalu $\langle a, b\rangle$ nejen stejnoměrně ohraničená, ale i stejně spojitá. Podle Arzelovy-Ascoliovy věty tedy existuje množina $H^* \subset H$ taková, že posloupnost $\{y_h;\, h \in H^*\}$ je stejnoměrně konvergentní. Označme její limitu y. Je to zřejmě spojitá funkce, pro niž platí $y(x) = \eta$. Dokažme ještě, že

splňuje danou diferenciální rovnici. Abychom toho dosáhli, uvědomme si nejprve, že platí

$$y_n = \eta + h\sum_{\nu=0}^{n-1} f(x_\nu, y_\nu), \quad n = 0, 1, \ldots, \; x_n \in \langle a, b\rangle, \tag{1.34}$$

jak plyne ihned z rovnic (1.14). Zvolme dále pevně $x \in \langle a, b\rangle$ a při fixovaném $h \in H^*$ určeme přirozené číslo $n = n(h, x)$ tak, aby platilo $x_{n-1} < x \leqq x_n$. Z definice funkce y_h plyne, že platí

$$\begin{aligned} &\left|h\sum_{\nu=0}^{n-1} f(x_\nu, y_\nu) - h\sum_{\nu=0}^{n-1} f(x_\nu, y(x_\nu))\right| = \\ &= \left|h\sum_{\nu=0}^{n-1} f(x_\nu, y_h(x_\nu)) - h\sum_{\nu=0}^{n-1} f(x_\nu, y(x_\nu))\right| \leqq \\ &\leqq hL\sum_{\nu=0}^{n-1} |y_h(x_\nu) - y(x_\nu)| \leqq (b-a)L \max_{x\in\langle a,b\rangle} |y_h(x) - y(x)| \to 0 \end{aligned} \tag{1.35}$$

pro $h \to 0$, $h \in H^*$. Dále zřejmě platí

$$h\sum_{\nu=0}^{n-1} f(x_\nu, y(x_\nu)) \to \int_a^x f(t, y(t))\,\mathrm{d}\,t \tag{1.36}$$

pro $h \to 0$, $h \in H^*$, neboť pro $h \to 0$ bod x_n konverguje k bodu x. Přejdeme-li tedy v (1.34) k limitě pro $h \to 0$, $h \in H^*$, dostáváme

$$y(x) = \eta + \int_a^x f(t, y(t))\,\mathrm{d}\,t. \tag{1.37}$$

Z rovnice (1.37) však už bezprostředně plyne, že funkce y je diferencovatelná a že splňuje danou diferenciální rovnici.

Abychom dokázali jednoznačnost, předpokládejme, že existují dvě řešení y a z. Z podmínky (ii) bezprostředně plyne, že je

$$|y(x) - z(x)| \leqq L\int_a^x |y(t) - z(t)|\,\mathrm{d}\,t. \tag{1.38}$$

Rozdíl $y - z$ je na intervalu $\langle a, b\rangle$ zřejmě spojitý, a proto existuje konstanta M_0 tak, že platí

$$\max_{x\in\langle a,b\rangle} |y(x) - z(x)| = M_0. \tag{1.39}$$

Dosadíme-li do nerovnosti (1.38) z (1.39), dostáváme

$$|y(x) - z(x)| \leqq M_0 L(x - a). \tag{1.40}$$

Dosadíme-li právě získaný odhad znovu do (1.38), máme

$$|y(x) - z(x)| \leqq M_0 \frac{L^2(x-a)^2}{2!}. \tag{1.41}$$

Pokračujeme-li v tomto postupu dále, dostaneme

$$|y(x) - z(x)| \leqq M_0 \frac{L^k(x-a)^k}{k!} \tag{1.42}$$

pro libovolné k a libovolné $x \in \langle a, b\rangle$. Odtud ale plyne, že platí $y(x) = z(x)$ pro každé $x \in \langle a, b\rangle$. Věta je dokázána.

Poznamenejme, že z jednoznačnosti plyne, že k řešení sestrojenému v důkazu věty 1.1 konverguje nejen vybraná posloupnost $\{y_h;\, h \in H^*\}$, ale dokonce celá posloupnost $\{y_h;\, h \in H\}$. Skutečně, kdyby tomu tak nebylo, existovala by posloupnost $\tilde{H} \subset H$ a $\varepsilon_0 > 0$ takové, že by platilo

$$\max_{x\in\langle a,b\rangle} |y_h(x) - y(x)| \geqq \varepsilon_0, \quad h \in \tilde{H}. \tag{1.43}$$

Stejně jako v důkazu věty 1.1 se nyní dá ukázat, že musí existovat $\tilde{H}^* \subset \tilde{H}$ tak, že posloupnost $\{y_h;\, h \in \tilde{H}^*\}$ je stejnoměrně konvergentní. Limitní funkce je však řešením dané diferenciální rovnice, jak plyne opět z důkazu věty 1.1. Vzhledem k nerovnosti (1.43) je toto řešení nutně různé od řešení y, což odporuje jednoznačnosti.

Tato poznámka tedy dává vlastně odpověď na otázku konvergence Eulerovy metody. V následujícím odstavci dokážeme konvergenci Eulerovy metody ještě jednou, a to poněkud jiným způsobem. Tento postup bude vhodnější k bezprostřednímu přenesení na obecnější případy, k jeho realizaci však bude třeba existenci řešení úlohy (1.6), (1.7) předpokládat.

2 Eulerova metoda

Jak už jsme se dohodli v úvodu, budeme všude v této kapitole předpokládat, že pravá strana dané diferenciální rovnice splňuje předpoklady (i) a (ii) z prvního odstavce, takže řešení problému (1.6), (1.7) existuje a je jediné. Vyšetřování Eulerovy metody začneme přesnou definicí pojmu konvergence, i když je tento pojem z předešlého textu intuitivně jasný.

Definice 2.1. Řekněme, že Eulerova metoda je *konvergentní*, jestliže pro každé $x \in \langle a, b\rangle$ platí

$$\lim_{\substack{h\to 0\\ x_n = x}} y_n = y(x), \tag{2.1}$$

kde y_n je přibližné řešení získané z rovnic (1.14) při integračním kroku h a y je přesné řešení. Limitní přechod ve vzorci (2.1) je přitom nutno chápat tak, že současně $h \to 0$ a $n \to \infty$ tak, že platí $x_n = a + nh = x$.

Jak už jsme upozornili dříve, je limitní přechod v rovnici (2.1) diskrétní, neboť při pevně zvoleném x bereme v úvahu jen ta h, pro něž platí, že $(x-a)/h$ je celé číslo. Tohoto omezení je možné se zbavit tak, že při zmíněném limitním přechodu požadujeme platnost podmínky $x_n \to x$. Za n v (2.1) je pak třeba vzít, např. celistvou část podílu $(x-a)/h$.

Než zformulujeme konvergenční větu pro Eulerovu metodu, připomeneme pojem modulu spojitosti spojité funkce, který se zavádí v matematické analýze jako kvantitativní charakteristika stejnoměrně spojitých funkcí.

Definice 2.2. Buď φ funkce spojitá na uzavřeném intervalu $\langle a, b\rangle$. Pak funkci ω_φ definovanou pro nezáporná δ předpisem

$$\omega_\varphi(\delta) = \sup_{\substack{|x-x^*| \leqq \delta \\ x, x^* \in \langle a,b\rangle}} |\varphi(x) - \varphi(x^*)| \tag{2.2}$$

nazveme *modulem spojitosti funkce* φ.

Funkce ω_φ je zřejmě spojitá zprava v bodě $\delta = 0$, tj. platí

$$\lim_{\delta \to 0+} \omega_\varphi(\delta) = 0. \tag{2.3}$$

Věta 2.1. *Nechť jsou splněny předpoklady* (i) a (ii) *z odst. 1, nechť* y_n *je přibližné řešení úlohy* (1.6), (1.7) *získané Eulerovou metodou* (1.14) *s integračním krokem* h *a nechť* y *je přesné řešení. Pak platí*

$$|y_n - y(x_n)| \leqq \omega(h) E_L(x_n - a), \quad n = 0, \dots, N, \tag{2.4}$$

kde N *je největší přirozené číslo, pro něž je* $x_N \in \langle a, b\rangle$, *funkce* E_L *je definována rovnicemi* (1.20) *a symbol* ω *značí modul spojitosti první derivace přesného řešení, tj. je*

$$\omega(h) = \sup_{\substack{|x-x^*| \leqq h \\ x, x^* \in \langle a,b\rangle}} |y'(x) - y'(x^*)|. \tag{2.5}$$

D ů k a z . Přibližné řešení y_n splňuje rovnici

$$y_{n+1} = y_n + h f(x_n, y_n), \quad n = 0, \dots, N-1. \tag{2.6}$$

Použijeme-li definice lokální diskretizační chyby Eulerovy metody (srv. vzorec (1.15)), dostáváme pro přesné řešení analogickou rovnici

$$y(x_{n+1}) = y(x_n) + h f\big(x_n, y(x_n)\big) + L\big(y(x_n); h\big), \quad n = 0, \dots, N-1. \tag{2.7}$$

Odečteme-li tuto rovnici od rovnice (2.6) a položíme-li $e_n = y_n - y(x_n)$, máme

$$e_{n+1} = e_n + h\big[f(x_n, y_n) - f\big(x_n, y(x_n)\big)\big] - L\big((x_n); h\big), \quad n = 0, \dots, N-1. \tag{2.8}$$

Odtud užitím toho, že funkce f splňuje Lipschitzovu podmínku vzhledem k proměnné y, dostáváme

$$|e_{n+1}| \leqq (1 + hL)|e_n| + |L\big(y(x_n); h\big)|, \quad n = 0, \dots, N-1. \tag{2.9}$$

Abychom tedy byly schopni odhadnout celkovou diskretizační chybu Eulerovy metody, je třeba mít k dispozici odhad její lokální diskretizační chyby. Podle věty o střední hodnotě však existuje číslo θ_n, $0 < \theta_n < 1$, tak, že platí

$$y(x_{n+1}) - y(x_n) = h y'(x_n + \theta_n h), \quad n = 0, \dots, N-1, \tag{2.10}$$

a tedy z rovnic (2.7) $\big(f\big(x_n, y(x_n)\big) = y'(x_n)\big)$, (2.10) a z definice modulu spojitosti dostáváme

$$|L\big(y(x_n); h\big)| \leqq h\omega(h). \tag{2.11}$$

Dosadíme-li tento odhad do (2.9) a na vzniklé nerovnosti užijeme lemma 1.1, máme

$$\begin{aligned} |e_n| &\leqq \frac{(1+hL)^n - 1}{hL} h\omega(h), \quad L > 0 \\ &\leqq nh\omega(h), \quad L = 0, \end{aligned} \tag{2.12}$$

neboť $e_0 = 0$. Z nerovností (2.12) a ze zřejmé nerovnosti $1 + hL < \mathrm{e}^{hL}$ (kterou jsme v podstatě ve stejné souvislosti ostatně užili v důkazu lemmatu 1.2) už tvrzení věty plyne bezprostředně. Věta je dokázána.

Vzhledem k tomu, že funkce E_L je na konečném intervalu omezená, plyne z této věty konvergence Eulerovy metody. Je třeba poznamenat, že věta 2.1 nedokazuje v té podobě, jak je formulována, skutečně nic víc než pouhou konvergenci, i když nerovnost (2.4) má formálně tvar odhadu chyby. O funkci ω, která v této nerovnosti vystupuje, totiž víme pouze, že existuje a že pro $h \to 0$ konverguje k nule, avšak zmíněná věta nic neříká o jejím konkrétním tvaru.

Aby nerovnost (2.4) byla skutečným odhadem chyby, museli bychom být schopni popsat (nebo odhadnout) funkci ω na základě vstupních dat dané úlohy. Pokusme se tedy o to.

V lemmatu 2.1 jsme zjistili, že pro každé přibližné řešení y_n získané Eulerovou metodou platí $|y_n| \leqq Y$, kde konstanta Y je dána vztahem (1.19). Z důkazu věty 1.1 však plyne, že tentýž odhad platí i pro přesné řešení rovnice (1.6) s počáteční podmínkou (1.7). Definujme pro nezáporná δ funkci Ω vztahem

$$\Omega(\delta) = \sup_{\substack{|x-x^*| \leqq \delta \\ (x,y),(x^*,y) \in R}} |f(x,y) - f(x^*,y)|, \tag{2.13}$$

kde R je obdélník definovaný rovností (1.30). Protože f je spojitá v R a R je kompaktní množina, je f stejnoměrně spojitá na R, což implikuje, že platí

$$\lim_{\delta \to 0+} \Omega(\delta) = 0. \tag{2.14}$$

Buďte nyní x, x^* libovolné dva body z intervalu $\langle a, b\rangle$, pro něž platí $|x - x^*| \leqq h$. Protože je vzhledem k tomu, co jsme řekli výše, $|y(x)| \leqq Y$, $|y(x^*)| \leqq Y$ je $(x, y(x)) \in R$, $(x^*, y(x^*)) \in R$. Platí tedy

$$\begin{aligned} |y'(x) - y'(x^*)| &= |f(x, y(x)) - f(x^*, y(x^*))| \leqq \\ &\leqq |f(x, y(x)) - f(x^*, y(x))| + \\ &\quad + |f(x^*, y(x)) - f(x^*, y(x^*))| \leqq \\ &\leqq \Omega(h) + L|y(x) - y(x^*)|. \end{aligned} \tag{2.15}$$

Pro rozdíl $y(x) - y(x^*)$ však platí podle věty o střední hodnotě odhad

$$\begin{aligned} |y(x) - y(x^*)| &= |y'(\xi)||x - x^*| = \\ &= |f(\xi, y(\xi))||x - x^*| \leqq Mh, \end{aligned} \tag{2.16}$$

kde bod ξ leží mezi body x a x^* a M je konstanta definovaná vztahem (1.32). Dosadíme-li tedy (2.16) do (2.15), dostáváme pro funkci ω z věty 2.1 odhad

$$\omega(h) \leqq \Omega(h) + LMh \tag{2.17}$$

a z nerovnosti (2.4) nerovnost

$$|y_n - y(x_n)| \leqq [\Omega(h) + LMh] E_L(x_n - a), \tag{2.18}$$

která už je odhadem chyby. Počítat funkci Ω je však velice nepohodlné a často prakticky vůbec neproveditelné. Praktická cena odhadu (2.18) je tedy značně problematická. Pokusme se proto nalézt použitelnější odhad než (2.18) za cenu, že zesílíme předpoklady na danou úlohu.

Věta 2.2. *Nechť jsou splněny předpoklady* (i) a (ii) a *nechť přesné řešení rovnice* (1.6) *určené počáteční podmínkou* (1.7) *má v intervalu* $\langle a, b\rangle$ *dvě spojité derivace. Buď*

$$M(x) = \tfrac{1}{2} \max_{t \in \langle a, x\rangle} |y''(t)|. \tag{2.19}$$

Pak platí

$$|y_n - y(x_n)| \leqq hM(x_n) E_L(x_n - a), \quad n = 0, \ldots, N, \tag{2.20}$$

přičemž význam symbolů y_n, N a E_L *je stejný jako ve větě 2.1.*

D ů k a z . Zvolme $h > 0$ pevně a buď n libovolné pevně zvolené přirozené číslo takové, že je $x_n \in \langle a, b\rangle$. Ze vzorce (2.9) je zřejmé, že pro $m = 0, \ldots, n-1$ platí

$$|e_{m+1}| \leqq (1 + hL)|e_m| + \left|L(y(x_m); h)\right|. \tag{2.21}$$

Z Taylorova vzorce plyne, že pro tatáž m existují čísla θ_m, $0 < \theta_m < 1$, tak, že platí

$$L(y(x_m); h) = \tfrac{1}{2} h^2 y''(x_m + \theta_m h). \tag{2.22}$$

Pro $m = 0, \ldots, n-1$ však platí podle definice funkce M

$$\left|\tfrac{1}{2} h^2 y''(x_m + \theta_m h)\right| \leqq h^2 M(x_n). \tag{2.23}$$

Tedy celkem

$$|e_{m+1}| \leqq (1 + hL)|e_m| + M(x_n)h^2, \quad m = 0, \ldots, n-1. \tag{2.24}$$

Odtud dostáváme podle lemmatu 1.1, že pro $m = 0, \ldots, n$ platí

$$\begin{aligned} |e_m| &\leqq \frac{(1 + hL)^m - 1}{hL} h^2 M(x_n), \quad L > 0, \\ &\leqq mh^2 M(x_n), \quad L = 0. \end{aligned} \tag{2.25}$$

Položíme-li zde speciálně $m = n$ a užijeme už dříve několikrát užitou nerovnost $1 + hL < \mathrm{e}^{hL}$, dostáváme nerovnost (2.20). Věta je dokázána.

Věta 2.2 odpovídá vlastně na dvě otázky týkající se konvergence Eulerovy metody. Známe-li funkci $M(x)$, představuje vzorec (2.20) odhad chyby. Víme-li jen, že tato funkce existuje, říká vzorec (2.20), že rychlost konvergence Eulerovy metody je h.

Poznamenejme, že žádným dalším zvyšováním požadavků na hladkost řešení dané diferenciální rovnice nelze obecně docílit vyšší rychlosti konvergence. Skutečně, řešíme-li Eulerovou metodou např. diferenciální rovnici $y' = 2x$ s počáteční podmínkou $y(0) = 0$, jejíž přesné řešení $y\ (= x^2)$ je dokonce analytické, snadno vypočteme, že platí $y_n - y(x_n) = -x_n h$.

Uveďme ještě, jak lze jednoduše odhadnout funkci M z věty 2.2. Za předpokladu, že pravá strana dané diferenciální rovnice má v obdélníku R daném vztahem (1.30) spojité parciální derivace podle x a y, platí zřejmě

$$2M(x) \leqq \max_{(x,y) \in R} |f_x + f_y f|. \tag{2.26}$$

(Symboly f_x a f_y zde značí parciální derivace podle x a y.)

Před uvedením jednoduchého příkladu ilustrujícího výsledky, ke kterým jsme do této chvíle dospěli, uveďme ještě jednu větu, která bude v dalším užitečná.

Věta 2.3. *Nechť platí tatáž označení jako ve větě 2.2 a nechť jsou rovněž splněny předpoklady této věty. Buď dále posloupnost* $\tilde{y}_n$, $n = 0, \ldots, N$ *definována rekurencí*

$$\begin{aligned} \tilde{y}_0 &= \eta \\ \tilde{y}_{n+1} &= \tilde{y}_n + hf(x_n, \tilde{y}_n) + \theta_n C h^2, \quad n = 0, \ldots, N-1, \end{aligned} \tag{2.27}$$

kde C *je konstanta a* θ_n *jsou libovolná reálná čísla, pro něž platí* $|\theta_n| \leqq 1$ *pro* $n = 0, \ldots, N-1$. *Pak platí*

$$|\tilde{y}_n - y(x_n)| \leqq h[M(x_n) + C] E_L(x_n - a), \quad n = 0, \ldots, N. \tag{2.28}$$

D ů k a z této věty je zřejmý a je v podstatě opakováním důkazu věty 2.2. Jediný rozdíl je v tom, že na pravé straně nerovnosti (2.21) vystupuje navíc člen $\theta_n Ch^2$, který se snadno majorizuje výrazem Ch^2.

Věta 2.3 vlastně říká, že k přesnému řešení dané diferenciální rovnice konverguje nejen přibližné řešení získané Eulerovou metodou, ale i řešení, které vznikne tak, že se v každém kroku Eulerovy metody dopustíme navíc chyby velikosti h^2. Je to důsledkem skutečnosti, které si pozorný čtenář už jistě všiml, a která spočívá v tom, že akumulací lokálních chyb se ztratí jeden řád (měřeno mocninou integračního kroku h).

Příklad 2.1. Řešme Eulerovou metodou v intervalu $\langle 0, 5\rangle$ diferenciální rovnice $y' = y$, resp. $y' = -y$, obě s počáteční podmínkou $y(0) = 1$. Přesná řešení jsou tedy funkce e^x, resp. e^{-x}. Lipschitzova konstanta L je v obou případech rovna jedné a $2M(x) = e^x$, resp. $2M(x) = 1$, takže výrazy $he^{x_n}(e^{x_n} - 1)/2$, resp. $h(e^{x_n} - 1)/2$ představují odhad chyby ve smyslu věty 2.2. Výsledky jsou pro integrační krok $h = 1/64$, uspořádány v tab. 2.1 a 2.2 a nejsou příliš uspokojivé, neboť získané odhady jsou velmi pesimistické. Tak např., chceme-li vypočítat e^{-5} s chybou, která nepřevýší 10^{-3}, je podle odhadu nutno vzít h tak, aby platilo $h(e^5 - 1)/2 \leqq 10^{-3}$, tj. $h \leqq 1/73707$; ve skutečnosti stačilo $h = 1/64$.

Tabulka 2.1

Přibližné řešení diferenciální rovnice $y' = y$

x_n	1	2	3	4	5
y_n	2,69735	7,27567	19,62499	52,93537	142,7850
e_n	−0,02093	−0,11339	−0,46055	−1,66278	−5,6282
odhad e_n	0,03649	0,36882	2,99487	22,86218	170,9223

Tabulka 2.2

Přibližné řešení diferenciální rovnice $y' = -y$

x_n	1	2	3	4	5
y_n	0,364987	0,133215	0,048622	0,017746	0,006477
e_n	−0,002892	−0,002120	−0,001165	−0,000570	−0,000261
odhad e_n	0,013424	0,049914	0,149106	0,418735	1,151666

Tento jev, i když jsme jej ukázali pouze na zcela triviálních příkladech, je pro odhady typu (2.20) typický a představuje silně limitující faktor pro jejich praktické použití. Je tedy velmi žádoucí mít k dispozici jinou možnost, pomocí níž by bylo

Tabulka 2.3

Závislost chyby na integračním kroku

$-\log_2 h$	y_n	e_n	$h^{-1}e_n$
1	0,250000	−0,117879	−0,235758
2	0,316406	−0,051473	−0,205893
3	0,343609	−0,024270	−0,194164
⋮			
6	0,364987	−0,002892	−0,185147
7	0,366438	−0,001441	−0,184540
8	0,367160	−0,000719	−0,184239

možné učinit si o chybě realističtější představu. Abychom jeden takový postup získali, začneme opět jednoduchým příkladem. V tab. 2.3 je ukázána závislost výrazu e_n/h na integračním kroku v případě diferenciální rovnice $y' = -y$ s počáteční podmínkou $y(0) = 1$. Z předešlého výkladu víme, že tento výraz je při $h \to 0$ a při $x_n = x$ ohraničený. Z tabulky se však zdá, že tento výraz je nejen ohraničený, ale že má dokonce při $h \to 0$ limitu. Kdyby tomu tak bylo a kdybychom tuto limitu znali, dávala by nám pro malá h asi daleko lepší představu o skutečné chybě než uvažovaný maximalistický odhad. Následující věta ukazuje, že za vhodných předpokladů zmíněná limita skutečně existuje.

Věta 2.4. *(Asymptotický vzorec pro chybu.) Nechť jsou splněny předpoklady* (i) a (ii) *a nechť navíc má pravá strana f dané diferenciální rovnice spojité první a druhé parciální derivace podle obou proměnných v množině R definované rovnicí* (1.30). *Pak celková diskretizační chyba e_n přibližného řešení vypočteného Eulerovou metodou se dá psát ve tvaru*

$$e_n = e(x_n)h + O(h^2), \tag{2.29}$$

kde funkce e je řešením diferenciální rovnice

$$e' = f_y\big(x, y(x)\big)e - \tfrac{1}{2}y''(x) \tag{2.30}$$

s počáteční podmínkou $e(a) = 0$.

D ů k a z . Z předpokladů o hladkosti funkce f plyne, že přesné řešení dané diferenciální rovnice má v intervalu $\langle a, b\rangle$ tři derivace spojité. Z Taylorova vzorce tedy plyne

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \tfrac{1}{2}h^2y''(x_n) + \tfrac{1}{6}h^3y''(\xi_n), \tag{2.31}$$

kde bod ξ_n leží mezi body x_n a x_{n+1}. Odečtením této rovnice od (1.14) dostáváme

$$e_{n+1} = e_n + h\big[f(x_n, y_n) - f(x_n, y(x_n))\big] - \tfrac{1}{2}h^2y''(x_n) - \tfrac{1}{6}h^3y'''(\xi_n). \tag{2.32}$$

Protože je $e_n = y_n - y(x_n)$, můžeme v rozdílu v hranaté závorce na pravé straně poslední rovnice psát $y(x_n) + e_n$ místo y_n a opět podle Taylorova vzorce máme, že

$$f(x_n, y_n) - f\big(x_n, y(x_n)\big) = f_y\big(x_n, y(x_n)\big)e_n + \tfrac{1}{2}f_{yy}(x_n, y^*)e_n^2, \tag{2.33}$$

kde bod y^* leží mezi body y_n a $y(x_n)$. Zaveďme nyní veličinu $\bar{e}_n = e_n/h$. Podle věty 2.2 je $|\bar{e}_n| \leqq M(x_n)E_L(x_n - a)$, a protože jak funkce M, tak funkce E_L jsou na intervalu $\langle a, b\rangle$ omezené, existuje konstanta C_1 taková, že platí

$$|\bar{e}_n| \leqq C_1. \tag{2.34}$$

Dělíme-li rovnici (2.32) integrační krokem a použijeme-li rovnici (2.33), dostáváme

$$\bar{e}_{n+1} = \bar{e}_n + hf_y\big(x_n, y(x_n)\big)\bar{e}_n + \tfrac{1}{2}h^2 f_{yy}(x_n, y^*)\bar{e}_n^2 - \\ - \tfrac{1}{2}hy''(x_n) - \tfrac{1}{6}h^2 y'''(\xi_n) \tag{2.35}$$

neboli

$$\bar{e}_{n+1} = \bar{e}_n + h\big[f_y\big(x_n, y(x_n)\big)\bar{e}_n - \tfrac{1}{2}y''(x_n)\big] + h^2 r_n, \tag{2.36}$$

kde jsme položili

$$r_n = \tfrac{1}{2}f_{yy}(x_n, y^*)\bar{e}_n^2 - \tfrac{1}{6}y'''(\xi_n) \tag{2.37}$$

Přitom existuje konstanta C_2 taková, že platí

$$|r_n| \leqq C_2. \tag{2.38}$$

Na rekurenci (2.36) se tedy můžeme dívat jako na Eulerovu metodu pro řešení diferenciální rovnice (2.30) s počáteční podmínkou $e(a) = 0$ (neboť je $\bar{e}_0 = 0$), kde se v každém kroku dopouštíme navíc chyby, která nepřevýší C_2h^2. Vzhledem k předpokladům o hladkosti funkce f má pravá strana rovnice (2.30) spojité parciální derivace podle x a e; její řešení má tedy dvě spojité derivace. Podle věty 2.3 tedy platí

$$|h^{-1}e_n - e(x)| \leqq hC_3E_{L_1}(x_n - a), \tag{2.39}$$

kde

$$C_3 = \tfrac{1}{2}\max_{x\in\langle a,b\rangle}|e''(x)| + C_2, \quad L_1 = \max_{x\in\langle a,b\rangle}\big|f_y\big(x, y(x)\big)\big|. \tag{2.40}$$

Věta je dokázána.

Ilustrujeme právě dokázanou větu opět jednoduchým příkladem.

Příklad 2.2. Řešme diferenciální rovnici $y' = -y$ s počáteční podmínkou $y(0) = 1$. V tomto případě je $y(x) = \mathrm{e}^{-x}$, $f_y\big(x, y(x)\big) = -1$, $y'' = \mathrm{e}^{-x}$, a tedy $e(x) = -x\mathrm{e}^{-x}/2$. V tabulce 2.4 je uveden průběh funkce $he(x_n)$ při $h = 2^{-6}$. Je z ní vidět velmi dobrá shoda skutečné chyby s hodnotou $he(x_n)$, takže znalost funkce e (tj. znalost asymptotického chování celkové diskretizační chyby) může být velice cenná.

Tabulka 2.4

Asymptotický odhad chyby

x_n	1	2	3	4	5
e_n	−0,002892	−0,002120	−0,001165	−0,000570	−0,000261
$he(x_n)$	−0,002874	−0,002114	−0,001167	−0,000572	−0,000263
odhad met. pol. kroku	−0,002902	−0,002123	−0,001165	−0,000568	−0,000260

Určení funkce e podle věty 2.4 však vyžaduje nutnost řešení další diferenciální rovnice, v níž navíc vystupuje hledané řešení, tj. neznámá funkce. Tato obtíž se sice dá za určitých okolností obejít; my však nyní ukážeme jiný postup ocenění celkové diskretizační chyby, který se sice také opírá o větu 2.4, zužitkovává však přímo pouze její existenční část.

Mějme dán pevně bod $x \in \langle a, b\rangle$ a značme pro tuto chvíli symbolem $y(x; h)$ přibližné řešení v bodě x získané Eulerovou metodou s krokem h (který musí samozřejmě být takový, že $(x-a)/h$ je celé číslo). Jsou-li splněny předpoklady věty 2.4, platí

$$y(x; h) = y(x) + e(x)h + O(h^2). \tag{2.41}$$

Píšeme-li zde $h/2$ místo h, dostáváme

$$y(x; h/2) = y(x) + \tfrac{1}{2}e(x)h + O(h^2). \tag{2.42}$$

Odečteme-li rovnici (2.42) od rovnici (2.41), máme

$$y(x; h) - y(x; h/2) = \tfrac{1}{2}e(x)h + O(h^2), \tag{2.43}$$

a tedy pro chybu přibližného řešení platí

$$y(x; h) - y(x) = 2[y(x; h) - y(x; h/2)] + O(h^2). \tag{2.44}$$

Až na veličiny vyššího řádu lze tedy získat velikost celkové diskretizační chyby tak, že přibližné řešení vypočteme dvakrát se dvěma různými integračními kroky. I když není zřejmě nutné, aby v právě popsaném postupu byl jeden z užitých integračních kroků polovina druhého, je tento případ nejčastější, a proto se tomuto způsobu ocenění chyby často také říká *metoda polovičního kroku*. V tab. 2.4 jsou uvedeny hodnoty chyby získané touto metodou. Opět vidíme velmi dobrou shodu se skutečností. Je však třeba důrazně upozornit, že vzorec (2.44) platí pouze asymptoticky, tj. až na veličiny vyššího řádu, a proto je třeba při jeho užití zachovávat jistou opatrnost.

Poznamenejme ještě, že výraz $2[y(x; h) - y(x; h/2)]$ (a také výraz $e(x)h$), který se rovná až na veličiny vyššího řádu celkové diskretizační chybě, se alternativně nazývá *hlavní část celkové diskretizační chyby*. Uveďme také, že vzorec (2.44), který je

možné sestavit až na základě provedeného výpočtu, je příkladem tzv. *aposteriorního odhadu* na rozdíl od *apriorního odhadu*, který je dán už pouhými vstupními daty dané úlohy (srv. např. odhad v větě 2.2).

Poslední otázka spojená s Eulerovou metodou, kterou zde stručně objasníme, je otázka ovlivnění přibližného řešení zaokrouhlováním. Jak už jsme uvedli v úvodu, v důsledku zaokrouhlování vypočteme místo veličiny y_n veličinu $\tilde{y}_n$ a naším úkolem je učinit si představu o velikosti rozdílu $r_n = \tilde{y}_n - y_n$, který nazveme *celkovou zaokrouhlovací chybou.*

Předpokládejme k tomu cíli, že veličiny $\tilde{y}_n$ splňují rovnice

$$\begin{aligned} &\tilde{y}_0 = \eta, \\ &\tilde{y}_{n+1} = \tilde{y}_n + hf(x_n, \tilde{y}_n) + \varepsilon_n, \quad 0, \dots, N-1. \end{aligned} \tag{2.45}$$

Veličina ε_n vyskytující se v těchto rovnicích je mírou nepřesnosti, které se dopustíme při výpočtu veličiny $\tilde{y}_{n+1}$ za předpokladu, že $\tilde{y}_n$ je zadáno přesně a je ovlivněna zejména tím, s jakou přesností jsme schopni vypočítat součet v (2.45). Je tomu tak proto, že integrační krok h je většinou malé číslo, takže chyba, se kterou vypočteme funkční hodnotu pravé strany dané diferenciální rovnice, se tolik neprojeví. Z těchto důvodů je rozumné nazvat veličinu ε_n *lokální zaokrouhlovací chybou* a předpokládat o ní, že je v průběhu výpočtu omezená. Všimněme si ještě, že rovnice (2.45) obsahují předpoklad, že počáteční podmínka je zadána přesně. Tento předpoklad není na újmu obecnosti, jak bude dále na první pohled vidět, a je činěn v podstatě pouze z pohodlí, abychom si v dalším ušetřili psaní. I o dalším předpokladu, který je v (2.45) obsažen implicitně, je třeba se zmínit. Je to předpoklad, že jak bod a, tak body x_n jsou zadány bez zaokrouhlovacích chyb. V praxi je splnění takovéhoto předpokladu zřejmě snadno dosažitelné.

Odečteme-li od rovnic (2.45) rovnice (1.14), zjistíme, že pro celkovou zaokrouhlovací chybu platí

$$r_{n+1} = r_n + h[f(x_n, \tilde{y}_n) - f(x_n, y_n)] + \varepsilon_n, \quad n = 0, \dots, N-1. \tag{2.46}$$

Je-li nyní $|\varepsilon_n| \leqq \varepsilon$ pro $n = 0, \dots, n$, dostáváme z rovnice (2.46) nerovnost

$$|r_{n+1}| \leqq (1 + hL)|r_n| + \varepsilon, \quad n = 0, \dots, N-1. \tag{2.47}$$

Z nerovnosti (2.47) a z lemmatu 1.1 však ihned plyne následující věta.

Věta 2.5. *Nechť funkce f splňuje předpoklady* (i) a (ii) *a nechť lokální zaokrouhlovací chyba ε_n je omezená, tj. nechť existuje konstanta ε taková, že platí $|\varepsilon_n| \leqq \varepsilon$ pro $n = 0, \dots, N-1$. Pak platí*

$$|r_n| \leqq \frac{\varepsilon}{h} E_L(x_n - a), \quad n = 0, \dots, N. \tag{2.48}$$

Z této věty plyne několik prakticky velmi důležitých důsledků. Nerovnost (2.48), kterou nelze obecně zlepšit, jak se snadno nahlédne (stačí k tomu např. uvažovat

diferenciální rovnici $y' = Ay$, kde A je konstanta a $\varepsilon_n = \varepsilon$), tak říká, že při omezených lokálních zaokrouhlovacích chybách celková zaokrouhlovací chyba roste při $h \to 0$ jako $1/h$. Skutečně vypočtené řešení je zatíženo dvěma chybami: celkovou diskretizační chybou, která při $h \to 0$ konverguje k nule, a celkovou zaokrouhlovací chybou, která při $h \to 0$ roste. Pokud tedy nejsme schopni zaručit při výpočtu více než jen to, že lokální zaokrouhlovací chyba je ohraničená, vypadá situace tak, že pro velká h převládne celková diskretizační chyba a celková chyba se bude při zmenšování h zmenšovat. Při dalším zmenšování integračního kroku h dospějeme však k takovému h_0, při němž převládne celková zaokrouhlovací chyba, která při dalším zmenšování h roste. Další zmenšování integračního kroku je tedy zbytečné a nejen to, dokonce škodlivé, protože celkovou chybu jen zvětšuje. V této situaci tedy existuje pro danou metodu a danou velikost lokální zaokrouhlovací chyby určitá *mezní přesnost*, kterou nelze překročit. Je-li tedy nezbytné z jakéhokoliv důvodu vypočíst řešení přesněji, než činí tato mezní přesnost, nemáme na základě toho, co dosud víme, jinou možnost, než zmenšit lokální zaokrouhlovací chybu a posunout tak mezní přesnost k menším integračním krokům. To však značí užít dvojnásobnou nebo případně i vícenásobnou aritmetiku, což jistě není příliš pohodlné řešení.

V dalších článcích uvidíme, že naštěstí existuje jiné východisko, které spočívá v tom, že se užije jiná metoda než Eulerova, a to taková, že její celková diskretizační chyba je při daném h menší než u Eulerovy metody. Chování celkové zaokrouhlovací chyby je totiž do značné míry nezávislé, jak uvidíme, na konkrétní metodě a tato skutečnost zřejmě umožňuje zmenšit velikost kritické chyby.

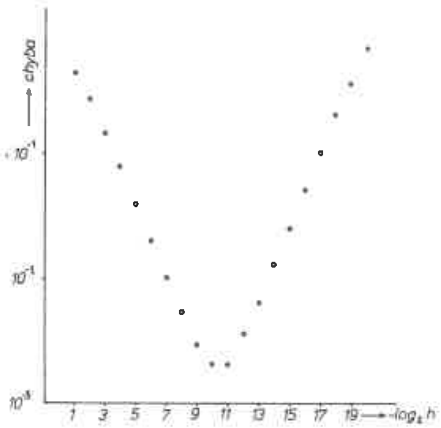
V dalších článcích si všimneme řady možností konstrukce těchto obecnějších metod. Než k nim však přejdeme, ilustrujme problematiku mezní přesnosti jednoduchým příkladem.

Příklad 2.3. Vypočtěme Eulerovou metodou hodnotu přibližného řešení diferenciální rovnice $y' = y$ s počáteční podmínkou $y(0) = 1$ v bodě $x = 1$ a užijme přitom integrační kroky $h = 1/2^s$, $s = 1, \dots, 20$.

Na obr. 2.1 a 2.2 je znázorněn průběh celkové chyby (tj. celkové diskretizační chyby a celkové zaokrouhlovací chyby) v závislosti na velikosti integračního kroku h. Z obou obrázků je vidět velmi dobrá shoda s předpověděným chováním. Celková chyba se nejprve zmenšuje, a to lineárně s h, neboť Eulerova metoda je řádu jedna, a po dosažení minima se v důsledku toho, že převládne zaokrouhlovací chyba, lineárně zvětšuje. Velikost kritické chyby je přitom na druhém obrázku menší než na prvním. Je to způsobeno tím, že příslušné výpočty byly v případě obr. 2.1 provedeny na počítači, u něhož je reálné číslo uloženo v 32 bitech, zatímco v případě obr. 2.2 byl užit počítač s přesnější aritmetikou o délce reálného čísla 40 bitů.

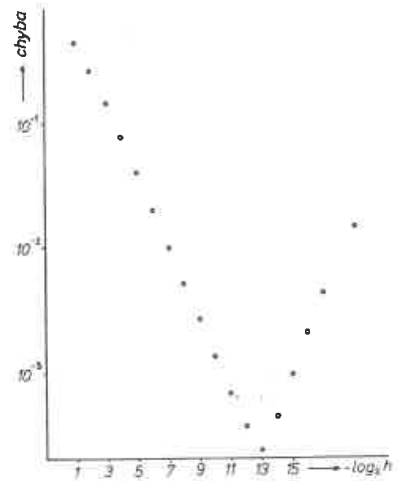
Obr. 2.1

Závislost celkové chyby na integračním kroku



Obr. 2.2

Závislost celkové chyby na integračním kroku



3 Obecná jednokroková metoda

Jedním z charakteristických rysů Eulerovy metody je, že přibližné řešení y_{n+1} v bodě x_{n+1} se při jejím užití počítá pouze ze znalosti přibližného řešení y_n v bodě x_n. Můžeme ji tedy pokládat za speciální případ *obecné jednokrokové metody*, když obecnou jednokrokovou metodou rozumíme jakýkoliv algoritmus pro řešení problému (1.6), (1.7), v němž se přibližné řešení y_{n+1} v bodě x_{n+1} počítá pouze na základě znalosti veličin x_n, y_n a h. Funkcionální závislost veličiny y_{n+1} na veličinách x_n, y_n a h je užitečné zapsat ve tvaru

$$y_{n+1} = y_n + h\Phi(x_n, y_n, h), \tag{3.1}$$

kde funkce Φ tří reálných proměnných závisí na dané diferenciální rovnici (resp. na její pravé straně). Vypočítat přibližné řešení y_n problému (1.6), (1.7) obecnou jednokrokovou metodou pak znamená generovat posloupnost aproximací $y_0, \ldots, y_N$ opakovaným užitím vzorce (3.1), přičemž se klade $y_0 = \eta$.

Intuitivně se dá očekávat, že i v případě obecné jednokrokové metody bude hrát lokální chyba stejně důležitou roli, jako tomu bylo u Eulerovy metody. Začneme proto vyšetřování obecné jednokrokové metody přesnou definicí její lokální chyby.

Definice 3.1. Výraz $L(y(x); h)$ definovaný rovnicí

$$L(y(x); h) = y(x + h) - y(x) - h\Phi(x, y(x), h), \tag{3.2}$$

kde y je přesné řešení problému (1.6), (1.7), nazveme *lokální chybou obecné jednokrokové metody.*

Veličina $h\Phi(x; y, h)$ udává zřejmě, oč vzroste na intervalu $\langle x, x + h\rangle$ přibližné řešení, které vychází z bodu (x, y); veličina $\Phi(x; y, h)$ je tedy relativní přírůstek přibližného řešení (tj. přírůstek vztažený na jednotku nezávisle proměnné). Lokální chyba dané jednokrokové metody bude tedy malá, bude-li se přírůstek přibližného řešení málo lišit od přírůstku přesného řešení, které prochází tímtéž bodem. Abychom tento požadavek mohli snadno zapisovat, zavedeme následující označení:

Nechť jsou splněny předpoklady (i) a (ii) z čl. 1 a buďte $x \in \langle a, b\rangle$ a y libovolná čísla. Podle věty 1.1 existuje v intervalu $\langle a, b\rangle$ právě jedna funkce z, která v tomto intervalu řeší diferenciální rovnici

$$\frac{\mathrm{d}\,z}{\mathrm{d}\,t} = f(t, z) \tag{3.3}$$

a pro kterou platí

$$z(x) = y. \tag{3.4}$$

Pro každé h takové, že je $x + h \in \langle a, b\rangle$ definujme funkci $\Delta(x, y, h)$ předpisem

$$\begin{aligned} \Delta(x, y, h) &= \frac{z(x + h) - y}{h}, \quad h \neq 0, \\ &= f(x, y), \quad h = 0. \end{aligned} \tag{3.5}$$

Funkce $h\Delta(x,y,h)$ tedy představuje přírůstek přesného řešení dané diferenciální rovnice procházejícího bodem (x,y) a funkce $\Delta(x,y,h)$ relativní přírůstek přesného řešení.

Podle definice 3.1 zřejmě platí

$$L(y(x);h) = h[\Delta(x,y(x),h) - \Phi(x,y(x),h)] \tag{3.6}$$

Funkci Φ, jíž je definována obecná jednokroková metoda, se tedy budeme snažit volit tak, aby rozdíl $\Phi(x,y,h) - \Delta(x,y,h)$ byl pro malá h co nejmenší. Dříve než si všimneme podrobněji problematiky spjaté s obecnou metodou typu (3.1), uvedeme některé důležité speciální případy. Jak už jsme uvedli výše, budeme tyto speciální jednokrokové metody konstruovat tak, aby jejich lokální chyba byla pro malé hodnoty integračního kroku malá. Chování lokální chyby se obvykle charakterizuje veličinou, která se nazývá řád obecné jednokrokové metody. Tento pojem bude hrát důležitou roli při konstrukci konkrétních jednokrokových metod, a to je také důvod, proč jej zavedeme už nyní.

Definice 3.2. Symbolem $\mathfrak{M}_0$ označme množinu diferenciálních rovnic, jejichž pravé strany splňují požadavky (i) a (ii) z čl. 1, a buď $\mathfrak{M} \subset \mathfrak{M}_0$. Maximální přirozené číslo p takové, že platí

$$L(y(x);h) = O(h^{p+1}) \tag{3.7}$$

pro každou funkci y, která je řešením diferenciální rovnice z množiny $\mathfrak{M}$, nazveme *řádem dané obecné jednokrokové metody na množině* $\mathfrak{M}$.

V předešlém článku jsme viděli, že některá tvrzení, která jsme tam dokázali, jako např. věta 2.1, platí pro celou množinu $\mathfrak{M}_0$, jiná (příkladem může být věta 2.2) platí pouze pro vhodnou vlastní podmnožinu množiny $\mathfrak{M}_0$. V důkaze zmíněných vět jsme také viděli, že o lokální chybě Eulerovy metody můžeme pro celou třídu $\mathfrak{M}_0$ říci jen, že se chová jako $h\omega(h)$, kde ω je modul spojitosti první derivace řešení, kdežto pro třídu $\mathfrak{M}_1$ diferenciálních rovnic, jejichž řešení mají spojitou nejen první, ale i druhou derivaci, je lokální chyba řádu h^2. Eulerova metoda má tedy na množině $\mathfrak{M}_1$ řád 1, kdežto na množině $\mathfrak{M}_0$ se o jejím řádu ve smyslu definice 3.2 nedá dost dobře mluvit. V dalším uvidíme, že i v případě obecné jednokrokové metody je tomu podobně; zavedení množiny $\mathfrak{M}$ v definici 3.2 je tedy nezbytné. Upozorněme také, že v smyslu definice 3.2 lze mluvit o řádu dané jednokrokové metody pro dané řešení.

3.1 Speciální případy

Předpokládejme v tomto odstavci, že pravá strana f dané diferenciální rovnice má pro $x \in \langle a,b\rangle$ a pro libovolné y parciální derivace až do řádu p, kde p je nějaké pevné přirozené číslo. Pak řešení y má derivace až do řádu $p+1$ a tyto derivace jsou úplně určeny funkcí f. Tak např. platí

$$y'' = f_x(x,y) + f_y(x,y)f(x,y) \tag{3.8}$$

a analogické vzorce lze psát pro y''', y'''' atd. Abychom zjednodušili zápisy, pišme $y'' = f'(x,y)$, $y''' = f''(x,y)$ atd. Čárka tedy značí totální derivaci funkce f podle x (tj. derivaci funkce f podle x, v níž y je též funkce proměnné x). Použijeme-li tohoto označení, platí pro přírůstek $h\Delta(x,y,h)$ přesného řešení podle Taylorova vzorce

$$h\Delta(x,y,h) = z(x+h) - y = \tag{3.9}$$
$$= hf(x,y) + \tfrac{1}{2}h^2 f'(x,y) + \ldots + \frac{1}{p!}h^p f^{(p-1)}(x,y) + O(h^{p+1}),$$

a tedy pro funkci Δ

$$\Delta(x,y,h) = f(x,y) + \tfrac{1}{2}hf'(x,y) + \ldots + \frac{1}{p!}h^{p-1}f^{(p-1)}(x,y) + O(h^p). \tag{3.10}$$

Vzhledem k tomu, co jsme řekli výše, se zdá přirozené volit funkci Φ tak, aby souhlasila s Δ tak dobře, jak je to jen možné. Vzorec (3.5) nabízí okamžité řešení spočívající v tom, že položíme

$$\Phi(x,y,h) = f(x,y) + \tfrac{1}{2}hf'(x,y) + \ldots + \frac{1}{p!}h^{p-1}f^{(p-1)}(x,y). \tag{3.11}$$

Metodu danou vzorcem (3.11) nazveme *metodou Taylorova rozvoje řádu p*. Její početní nevýhody jsou očividné: K tomu, abychom vypočetli hodnotu funkce Φ, je třeba mít k dispozici hodnoty funkcí $f, f', \ldots, f^{(p-1)}$. Určit tvar těchto funkcí však nebude v konkrétních případech většinou vůbec jednoduché. Čtenář se o tom může přesvědčit na příkladě velice nevinně vyhlížející diferenciální rovnice $y' = x^2 + y^2$. Proto je snaha sestrojit funkci Φ tak, aby se lišila od funkce Δ až teprve ve členech řádu h^p a aby zároveň nebylo třeba k výpočtu jejích hodnot počítat žádné derivace funkce f, zcela přirozená. Vzniká jen otázka, je-li možné. Klíčem ke kladné odpovědi na tuto otázku je známá skutečnost, že derivaci libovolného řádu každé funkce lze aproximovat lineární kombinací jejích hodnot v nějakých vhodně zvolených bodech. Základní myšlenkou příslušného postupu, který dá vzniknout tzv. *Rungovým-Kuttovým metodám*, si ukážeme pro jednoduchost na případě $p = 2$.

Hledejme funkce Φ ve tvaru

$$\Phi(x,y,h) = w_1k_1 + w_2k_2, \tag{3.12}$$

kde

$$k_1 = f(x,y), \quad k_2 = f(x+\alpha h, y+\beta hk_1) \tag{3.13}$$

a w_1, w_2, α a β jsou konstanty, které určíme tak, aby rozvoj funkce Φ jako funkce proměnné h souhlasil co možno nejdále s rozvojem funkce Δ. Rozviňme tedy funkci

Φ podle h. Dostaneme

$$(3.14)\quad \Phi(x,y,h) = w_1k_1 + w_2k_2 = w_1f(x,y) + w_2f\big(x+\alpha h, y+\beta h f(x,y)\big) = \\ = (w_1+w_2)f(x,y) + w_2[\alpha f_x(x,y) + \beta f_y(x,y)f(x,y)]h + O(h^2).$$

Porovnáním koeficientů u stejných mocnin h ve vzorci (3.14) a (3.10) dostáváme

$$(3.15)\qquad w_1 + w_2 = 1, \quad w_2\alpha = \tfrac{1}{2}, \quad w_2\beta = \tfrac{1}{2}.$$

Tato soustava má nekonečně mnoho řešení a každé řešení dostaneme zřejmě pomocí vzorců

$$(3.16)\qquad w_1 = 1-t, \quad w_2 = t, \quad \alpha = \frac{1}{2t}, \quad \beta = \frac{1}{2t},$$

kde t je libovolné reálné číslo různé od nuly. Dále se dá také snadno přímým výpočtem ukázat, že žádnou volbou parametrů w_1, w_2, α a β nelze docílit, aby v rozvojích (3.14) a (3.10) souhlasily pro obecnou diferenciální rovnici i členy s h^2. Zvolíme-li tedy ve vzorci (3.12) a (3.13) parametry w_1, w_2, α a β podle (3.16), dostáváme obecnou jednokrokovou metodu, pro niž platí $\Phi(x,y,h) - \Delta(x,y,h) = O(h^2)$, a která je tedy řádu 2 na množině $\mathfrak{M}_2$ diferenciálních rovnic, jejichž řešení mají 3 spojité derivace. Lokální chyba této metody má, jak vidíme, stejnou vlastnost jako lokální chyba metody Taylorova rozvoje řádu 2. Nutnost počítat derivace pravé strany dané diferenciální rovnice je však nahrazena tím, že se počítají v každém kroku dvakrát hodnoty funkce f.

V obecné Rungově-Kuttově metodě je funkce Φ dána předpisem

$$(3.17)\qquad \Phi(x,y,h) = w_1k_1 + \ldots + w_sk_s,$$

kde

$$(3.18)\qquad k_1 = f(x,y),$$
$$k_i = f\Big(x+\alpha_i h, y + h\sum_{j=1}^{i-1}\beta_{ij}k_j\Big), \quad i = 2,\ldots,s,$$

a k výpočtu její hodnoty je tedy třeba s-krát vypočítat hodnotu pravé strany dané diferenciální rovnice. Postupem, který jsme ukázali výše, o mnoho však pracnějším a komplikovanějším, se dá ukázat, že parametry v každé Rungově-Kuttově metodě lze volit tak, že pro její řád $p = p(s)$ platí $p(s) \to \infty$ pro $s \to \infty$. Řádu $p(s)$ lze přitom dosáhnout nekonečně mnoha volbami příslušných koeficientů. Do dalších podrobností nebudeme zacházet, spokojíme se pouze s konstatováním, že pro $s \leqq 4$ je na množině diferenciálních rovnic s dostatečně hladkými řešeními $p(s) = s$, zatímco pro $s > 4$ roste funkce p už pomaleji než s. To je také v podstatě důvod, proč se v praxi jen zřídka užívá Rungových-Kuttových metod řádu většího než 4. Druhý důvod je ten, že koeficienty každé Rungovy-Kuttovy metody se určují ze značně složité soustavy nelineárních algebraických rovnic, při jejímž řešení vznikají

pro velká s prakticky nezvládnutelné problémy. Aby si čtenář mohl udělat představu o jejich charakteru, uveďme tuto soustavu pro $s = 4$:

$$(3.19)\qquad \begin{aligned} &\alpha_2 = \beta_{21},\\ &\alpha_3 = \beta_{31} + \beta_{32},\\ &\alpha_4 = \beta_{41} + \beta_{42} + \beta_{43},\\ &w_1 + w_2 + w_3 + w_4 = 1,\\ &w_2\alpha_2 + w_3\alpha_3 + w_4\alpha_4 = \tfrac{1}{2},\\ &w_2\alpha_2^2 + w_3\alpha_3^2 + w_4\alpha_4^2 = \tfrac{1}{3},\\ &w_2\alpha_2^3 + w_3\alpha_3^3 + w_4\alpha_4^3 = \tfrac{1}{4},\\ &w_3\alpha_2\beta_{32} + w_4(\alpha_2\beta_{42} + \alpha_3\beta_{43}) = \tfrac{1}{6},\\ &w_3\alpha_2^2\beta_{32} + w_4(\alpha_2^2\beta_{42} + \alpha_3^2\beta_{43}) = \tfrac{1}{12},\\ &w_3\alpha_2\alpha_3\beta_{32} + w_4(\alpha_2\beta_{42} + \alpha_3\beta_{43})\alpha_4 = \tfrac{1}{8},\\ &w_4\alpha_2\beta_{32}\beta_{43} = \tfrac{1}{24}. \end{aligned}$$

V následujících řádcích uvedeme některé důležité speciální Rungovy-Kuttovy metody. Dvě metody druhého řádu

$$(3.20)\qquad y_{n+1} = y_n + hf\big(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hf(x_n,y_n)\big)$$

a

$$(3.21)\qquad y_{n+1} = y_n + \tfrac{1}{2}h[f(x_n,y_n) + f(x_n+h, y_n + hf(x_n,y_n)],$$

které vzniknou ze vzorce (3.12) a (3.13) dosazeními podle (3.16) s $t = 1$, resp. $t = 1/2$, se vyskytují v literatuře pod názvem *modifikovaná Eulerova metoda* a *Heunova metoda*. Geometricky lze obě tyto metody interpretovat podobně jako Eulerovu metodu. Řešení se opět aproximuje lomenou čarou procházející body (x_n, y_n); její směrnice se však volí o něco důmyslněji. Je to právě metoda (3.20), kterou navrhl už koncem minulého století Runge a dal tak vlastně podnět ke vzniku všech ostatních metod tohoto typu.

Z metod třetího řádu uveďme metodu

$$(3.22)\qquad \begin{aligned} y_{n+1} &= y_n + \tfrac{1}{4}h(k_1 + 3k_3),\\ k_1 &= f(x_n,y_n),\\ k_2 &= f(x_n + \tfrac{1}{3}h, y_n + \tfrac{1}{3}hk_1),\\ k_3 &= f(x_n + \tfrac{2}{3}h, y_n + \tfrac{2}{3}hk_2) \end{aligned}$$

známou rovněž jako *Heunova metoda.*

Dvě velmi známé metody čtvrtého řádu jsou

$$
\begin{aligned}
(3.23)\qquad y_{n+1} &= y_n + \tfrac{1}{6}h(k_1 + 2k_2 + 2k_3 + k_4),\\
k_1 &= f(x_n, y_n),\\
k_2 &= f(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hk_1),\\
k_3 &= f(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hk_2),\\
k_4 &= f(x_n + h, y_n + hk_3)
\end{aligned}
$$

a

$$
\begin{aligned}
(3.24)\qquad y_{n+1} &= y_n + \tfrac{1}{8}h(k_1 + 3k_2 + 3k_3 + k_4),\\
k_1 &= f(x_n, y_n),\\
k_2 &= f(x_n + \tfrac{1}{3}h, y_n + \tfrac{1}{3}hk_1),\\
k_3 &= f(x_n + \tfrac{2}{3}h, y_n - \tfrac{1}{3}hk_1 + hk_2),\\
k_4 &= f(x_n + h, y_n + hk_1 - hk_2 + hk_3).
\end{aligned}
$$

Daleko nejužívanější Rungovou-Kuttovou metodou je metoda (3.23); proto také, mluví-li se o Rungově-Kuttově metodě, myslí se tím často právě tato konkrétní metoda. Budeme o ní proto v dalším textu mluvit jako o *standardní Rungově-Kuttově metodě*. Metodě (3.24) se často říká *třiosminové pravidlo*.

Z metod pátého řádu je dosti populární tzv. *Fehlbergova metoda*, která vyžaduje šest funkčních hodnot a je dána vzorci

$$
\begin{aligned}
(3.25)\qquad y_{n+1} &= y_n + h\Big(\tfrac{16}{135}k_1 + \tfrac{6\,656}{12\,825}k_3 + \tfrac{28\,561}{56\,430}k_4 - \tfrac{9}{50}k_5 + \tfrac{2}{55}k_6\Big),\\
k_1 &= f(x_n, y_n),\\
k_2 &= f\Big(x_n + \tfrac{1}{4}h, y_n + \tfrac{1}{4}hk_2\Big),\\
k_3 &= f\Big(x_n + \tfrac{3}{8}h, y_n + \tfrac{1}{32}h(3k_1 + 9k_2)\Big),\\
k_4 &= f\Big(x_n + \tfrac{12}{13}h, y_n + \tfrac{1}{2\,197}h(1\,932k_1 - 7\,200k_2 + 7\,296k_3)\Big),\\
k_5 &= f\Big(x_n + h, y_n + h\Big(\tfrac{439}{216}k_1 - 8k_2 + \tfrac{3\,680}{513}k_3 - \tfrac{845}{4\,104}k_4\Big)\Big),\\
k_6 &= f\Big(x_n + \tfrac{1}{2}h, y_n + \\
&\qquad + h\Big(-\tfrac{8}{27}k_1 + 2k_2 - \tfrac{3\,544}{2\,565}k_3 + \tfrac{1\,859}{4\,104}k_4 - \tfrac{11}{40}k_5\Big)\Big).
\end{aligned}
$$

Tato metoda má tu vlastnost, že šest hodnot k v ní užitých se dá zkombinovat tak, že se dostane nová aproximace y^*_{n+1},

$$
(3.26)\qquad y^*_{n+1} = y_n + h\Big(\tfrac{25}{216}k_1 + \tfrac{1\,408}{2\,565}k_3 + \tfrac{2\,197}{4\,104}k_4 - \tfrac{1}{5}k_5\Big),
$$

která je čtvrtého řádu. Rozdíl $y^*_{n+1} - y_{n+1}$ pak lze pokládat za přibližný odhad velikosti lokální chyby. Je tomu tak proto, že přibližné řešení y_{n+1} vypočtené metodou pátého řádu je většinou podstatně přesnější než přibližné řešení y^*_{n+1}, takže je lze pro účel odhadnutí chyby ztotožnit s přesným řešením.

Huťova metoda

$$
\begin{aligned}
(3.27)\qquad y_{n+1} &= y_n + \tfrac{1}{840}h(41k_1 + 216k_3 + 27k_4 + 272k_5 + 27k_6 + \\
&\qquad + 216k_7 + 41k_8),\\
k_1 &= f(x_n, y_n),\\
k_2 &= f(x_n + \tfrac{1}{9}h, y_n + \tfrac{1}{9}hk_1),\\
k_3 &= f\big(x_n + \tfrac{1}{6}h, y_n + \tfrac{1}{24}h(k_1 + 3k_2)\big),\\
k_4 &= f\big(x_n + \tfrac{1}{3}h, y_n + \tfrac{1}{6}h(k_1 - 3k_2 + 4k_3)\big),\\
k_5 &= f\big(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{8}h(-5k_1 + 27k_2 - 24k_3 + 6k_4)\big),\\
k_6 &= f\big(x_n + \tfrac{2}{3}h, y_n + \tfrac{1}{9}h(221k_1 - 981k_2 + 867k_3 - 102k_4 + k_5)\big),\\
k_7 &= f\big(x_n + \tfrac{5}{6}h, y_n + \tfrac{1}{48}h(-183k_1 + 678k_2 - 472k_3 - 66k_4 + \\
&\qquad + 80k_5 + 3k_6)\big),\\
k_8 &= f\big(x_n + h, y_n + \tfrac{1}{82}h(716k_1 - 2\,079k_2 + 1\,002k_3 + 834k_4 - \\
&\qquad - 454k_5 - 9k_6 + 72k_7)\big),
\end{aligned}
$$

je šestého řádu a vyžaduje v každém kroku osmkrát vypočítat hodnotu pravé strany dané diferenciální rovnice.

Huťova metoda i další zde explicite uvedené Rungovy-Kuttovy metody vysokých řádů se užívají velmi zřídka, ačkoliv k tomu není vlastně žádný vážný důvod. Většinou se argumentuje v jejich neprospěch tím, že k provedení jednoho kroku je třeba mnohokrát vypočíst pravou stranu dané diferenciální rovnice (což činí beze sporu největší podíl v práci vynaložené na provedení jednoho kroku metody). Tyto metody jsou však vysokého řádu, takže je lze užít s větším integračním krokem, čímž může být zmíněná nevýhoda do značné míry kompenzována.

3.2 Konvergence obecné jednokrokové metody

V tomto odstavci zformulujeme požadavky, které je nutné klást na obecnou jednokrokovou metodu, abychom zaručili její konvergenci, resp. abychom byli schopni odhadnout její chybu.

Definice 3.3. Řekněme, že obecná jednokroková metoda je *regulární*, platí-li:
(i) funkce Φ je definovaná a spojitá v oboru $a \leqq x \leqq b$, $-\infty < y < \infty$, $0 \leqq h \leqq h_0$ (h_0 je kladná konstanta);
(ii) existuje konstanta L taková, že platí

$$
(3.28)\qquad |\Phi(x, y, h) - \Phi(x, z, h)| \leqq L|y - z|
$$

pro každé $x \in \langle a, b\rangle$, $h \in \langle 0, h_0\rangle$ a pro libovolné y a z.

Definice 3.4. . Řekněme, že obecná jednokroková metoda daná funkcí Φ je *konzistentní* s danou diferenciální rovnicí (1.6), je-li regulární a platí-li

$$\Phi(x, y, 0) = f(x, y) \tag{3.29}$$

pro $a \leqq x \leqq b$ a $-\infty < y < \infty$.

Věta 3.1. *Nechť jsou splněny předpoklady* (i) a (ii) *z odst. 1 a nechť y je přesné řešení diferenciální rovnice* (1.6) *s počáteční podmínkou* (1.7). *Buď dále y_n přibližné řešení vypočtené obecnou jednokrokovou metodou* (3.1), *která je konzistentní s danou diferenciální rovnicí. Pak platí*

$$|y_n - y(x_n)| \leqq [\omega(h) + \varphi(h)]E_L(x_n - a) \tag{3.30}$$

pro $n = 0, \ldots, N$ a pro $h \leqq h_0$, kde N je celá část podílu $(b-a)/h$, ω je modul spojitosti první derivace přesného řešení (tj. funkce ω je dána vzorcem (2.5)), *funkce φ je definována rovnicí*

$$\varphi(h) = \sup_{x \in \langle a,b \rangle} \left|\Phi\big(x, y(x), h\big) - \Phi\big(x, y(x), 0\big)\right| \tag{3.31}$$

a E_L *je funkce definovaná v* (1.20).

D ů k a z . Podle definice lokální chyby obecné jednokrokové metody je $y(x_{n+1}) = y(x_n) + h\Phi\big(x_n, y(x_n), h\big) + L\big(y(x_n); h\big)$. Odečteme-li tuto rovnici od rovnice (3.1), dostaneme pro celkovou diskretizační chybu $e_n = y_n - y(x_n)$ rovnici

$$e_{n+1} = e_n + \big[\Phi(x_n, y_n, h) - \Phi\big(x_n, y(x_n), h\big)\big] - L\big(y(x_n); h\big) \tag{3.32}$$

a odtud

$$|e_{n+1}| \leqq (1 + hL)|e_n| + \left|L\big(y(x_n); h\big)\right|, \quad n = 0, \ldots, N-1. \tag{3.33}$$

Tato nerovnost je úplně stejná jako nerovnost (2.9) v důkazu konvergenční věty pro Eulerovu metodu. K získání nerovnosti (3.30) je tedy stejně jako tam třeba odhadnout vhodně lokální diskretizační chybu. Podle věty o střední hodnotě existuje číslo θ_n, $0 < \theta_n < 1$, takové, že platí

$$y(x_{n+1}) = y(x_n) + hy'(x_n + \theta_n h), \quad n = 0 \ldots, N-1. \tag{3.34}$$

Píšeme-li ještě

$$\begin{aligned} h\Phi\big(x_n, y(x_n), h\big) = {} & h\Phi\big(x_n, y(x_n), 0\big) + \\ & + h\big[\Phi\big(x_n, y(x_n), h\big) - \Phi\big(x_n, y(x_n), 0\big)\big]. \end{aligned} \tag{3.35}$$

a užijeme-li podmínku konzistence, dostáváme

$$\begin{aligned} L\big(y(x_n); h\big) = {} & h\big[y'(x_n + \theta_n h) - f\big(x_n, y(x_n)\big)\big] - \\ & - h\big[\Phi\big(x_n, y(x_n), h\big) - \Phi\big(x_n, y(x_n), 0\big)\big]. \end{aligned} \tag{3.36}$$

První sčítanec na pravé straně poslední rovnice lze odhadnout výrazem $h\omega(h)$, neboť je $f\big(x_n, y(x_n)\big) = y'(x_n)$ a druhý výrazem $h\varphi(h)$, jak plyne ihned z definice funkce φ. Odtud už úplně stejným postupem jako v důkazu věty 2.1 plyne platnost nerovnosti (3.30). Věta je dokázána.

Z věty 3.1 plyne ihned konvergence obecné jednokrokové metody, tj. platnost vztahu

$$\lim_{\substack{h \to 0 \\ x_n = x}} y_n = y(x), \tag{3.37}$$

neboť jak funkce ω, tak funkce φ konvergují pro $h \to 0$ k nule. O funkci ω to už víme, o funkci φ to plyne ze stejnoměrné spojitosti funkce $\Phi\big(x, y(x), h\big)$ na množině $a \leqq x \leqq b$, $0 \leqq h \leqq h_0$.

Z nerovnosti (3.30) můžeme dostat odhad chyby analogicky jako u Eulerovy metody. Zesílením požadavků na danou metodu můžeme rovněž tak sestrojit odhad analogický odhadu z věty 2.2.

Věta 3.2. *Buď dána množina diferenciálních rovnic $\mathfrak{M} \subset \mathfrak{M}_0$, buď y přesné řešení diferenciální rovnice z $\mathfrak{M}$ a buď y_n jeho aproximace získaná obecnou jednokrokovou metodou, která je regulární a je řádu $p \geqq 1$ na $\mathfrak{M}$. Pak existuje konstanta M taková, že platí*

$$|y_n - y(x_n)| \leqq Mh^p E_L(x_n - a) \tag{3.38}$$

pro $n = 0, \ldots, N$ a pro $h \leqq h_0$.

D ů k a z je v podstatě opakováním důkazu věty 3.1. Jediný rozdíl je v tom, že za předpokladů uvedených ve znění věty platí pro lokální chybu vyšetřované metody odhad

$$\left|L\big(y(x_n); h\big)\right| \leqq Mh^{p+1}. \tag{3.39}$$

Užitím tohoto odhadu v nerovnosti (3.33) dostaneme ihned požadovaný výsledek.

Na první pohled by se mohlo zdát, že požadavek konzistence se ve větě 3.2 oproti větě 3.1 ztratil. Ve skutečnosti tomu tak není, neboť z toho, že daná metoda je řádu aspoň jedna na vhodné množině, plyne, že je pro tuto množinu diferenciálních rovnic i konzistentní. Skutečně, z nerovnosti $|L(y(x); h)| \leqq Mh^2$ plyne nerovnost

$$\left|\frac{y(x+h) - y(x)}{h} - \Phi\big(x, y(x), h\big)\right| \leqq Mh. \tag{3.40}$$

Protože pravá strana této nerovnosti konverguje pro $h \to 0$ k nule, platí totéž i pro levou stranu. První sčítanec na levé straně má však pro $h \to 0$ limitu, a to rovnou číslu $y'(x) = f\big(x, y(x)\big)$, tutéž limitu má tedy i druhý sčítanec.

Zcela analogicky jako větu 3.2 lze dokázat i následující větu, která tvoří paralelu k větě 2.3.

Věta 3.3. *Nechť jsou splněny předpoklady věty 3.2 a nechť $\tilde{y}_n$ je posloupnost čísel definovaná rekurencí*

$$\tilde{y}_0 = \eta, \tag{3.41}$$
$$\tilde{y}_{n+1} = \tilde{y}_n + h[\Phi(x_n, \tilde{y}_n, h) + h^q \theta_n K], \quad n = 0, \dots, N,$$

kde K je konstanta a θ_n jsou čísla, pro něž platí $|\theta_n| \leqq 1$. Pak pro $n = 0, \dots, N$ a pro $h \leqq h_0$ platí

$$|\tilde{y}_n - y(x_n)| \leqq M_1 h^r E_L(x_n - a), \tag{3.42}$$

kde $r = \min(p, q)$ a $M_1 = M h_0^{p-r} + K h_0^{q-r}$.

Větu 3.2 můžeme interpretovat podobně jako analogickou větu v případě Eulerovy metody dvojím způsobem. Pokud o konstantě ve vzorci (3.38) víme pouze, že existuje, ale neznáme její konkrétní hodnotu, je věta 3.2 jen výrokem o rychlosti konvergence dané metody. Pokud tuto konstantu známe nebo jsme schopni ji odhadnout, představuje vzorec (3.38) odhad chyby. Všimněme si proto problému odhadu konstanty M v nerovnosti (3.39) podrobněji.

Buď dána diferenciální rovnice (1.6) s počáteční podmínkou (1.7) a buď y její přesné řešení. Nechť dále funkce Φ definuje obecnou jednokrokovou metodu, která je pro toto řešení řádu p. Konečně předpokládejme, že funkce f, resp. Φ má p spojitých parciálních derivací podle obou, resp. všech tří proměnných v množině R, resp. $R \times \langle 0, h_0 \rangle$, kde

$$R = \{(x, y);\ a \leqq x \leqq b,\ -Y \leqq y \leqq Y\} \tag{3.43}$$

a Y je definováno vztahem (1.19). Protože platí (3.6), plyne z Taylorova vzorce a z předpokladu, že daná metoda je řádu p

$$\frac{\partial^k}{\partial h^k}[\Delta(x, y(x), h) - \Phi(x, y(x), h)]\Big|_{h=0} = 0 \tag{3.44}$$

pro $k = 0, \dots, p-1$ a

$$\Delta(x, y(x), h) - \Phi(x, y(x), h) = \frac{1}{p!} h^p \frac{\partial^p}{\partial h^p}[\Delta(x, y(x), h) - \Phi(x, y(x), h)]\Big|_{h=h_1} = 0, \tag{3.45}$$

kde h_1 je vhodné číslo, pro něž platí $0 < h_1 < h$. Vypočtěme tedy p-tou derivaci podle h rozdílu $\Delta - \Phi$. Pokud není dán konkrétní tvar funkce Φ, nelze s výrazem $\partial^p \Phi(x, y(x), h)/\partial h^p$ nic dalšího dělat. Pro výraz $\partial^p \Delta(x, y(x), h)/\partial h^p$ však platí

podle Leipnitzova pravidla pro derivování součinu

$$\frac{\partial^p}{\partial h^p}\Delta(x, y(x), h) = \frac{\partial^p}{\partial h^p}\frac{y(x+h) - y(x)}{h} = \tag{3.46}$$
$$= \sum_{\nu=0}^{p} (-1)^{p-\nu} \binom{p}{\nu} (p-\nu)! \frac{1}{h^{p-\nu+1}} \frac{\partial^\nu}{\partial h^\nu}[y(x+h) - y(x)] =$$
$$= (-1)^{p+1} \frac{p!}{h^{p+1}} \left[y(x) - \sum_{\nu=0}^{p} \frac{1}{\nu!} y^{(\nu)}(x+h)(-h)^\nu \right].$$

Z Taylorova vzorce užitého na funkci y a na interval $\langle x+h, x+h+(-h) \rangle$ nyní plyne, že existuje h_2, $0 < h_2 < h$ tak, že výraz v hranaté závorce na pravé straně rovnosti (3.46) je roven výrazu $y^{(p+1)}(x+h_2)(-h)^{p+1}/(p+1)!$. Celkem tedy dostáváme

$$\Delta(x, y(x), h) - \Phi(x, y(x), h) = \tag{3.47}$$
$$= \left[\frac{1}{(p+1)!} f^{(p)}(x + h_2, y(x+h_2)) - \frac{1}{p!} \frac{\partial^p}{\partial h^p} \Phi(x, y(x), h_1) \right] h^p$$

Přesné řešení dané diferenciální rovnice, které se vyskytuje v tomto vzorci, sice neznáme, víme však, že dvojice $(x, y(x))$ leží pro každé $x \in \langle a, b \rangle$ v množině R dané rovnici (3.43). Za výše uvedených předpokladů lze tedy položit

$$M = \max_{\substack{(x,y) \in R \\ 0 \leqq h \leqq h_0}} \left| \frac{1}{(p+1)!} f^{(p)}(x+h, z(x+h)) - \frac{1}{p!} \frac{\partial^p \Phi(x, y, h)}{\partial h^p} \right|, \tag{3.48}$$

kde z je řešení diferenciální rovnice $z' = f(t, z)$ s počáteční podmínkou $z(x) = y$.

Aplikujme nyní uvedenou teorii velmi stručně na speciální jednokrokové metody zavedené v předešlém odstavci. Všimněme si z tohoto hlediska nejprve Rungovy-Kuttovy metody (3.12), (3.13). Příslušná funkce Φ je zřejmě spojitá, je-li pravá strana f dané diferenciální rovnice spojitá. Splňuje-li pravá strana dané diferenciální rovnice Lipschitzovu podmínku vzhledem k proměnné y — příslušnou konstantu označíme nyní L_0 — platí (položíme-li $k_i^* = k(x, y^*, h)$) $|k_1^* - k_1| \leqq L_0 |y^* - y|$, $|k_2^* - k_2| \leqq L_0 |y^* + h k_1^*/(2t) - y - h k_1|(2t)| \leqq L_0 [|y^* - y| + h|k_1^* - k_1|/(2|t|)] \leqq$ $\leqq L_0 (1 + h L_0/(2|t|)) |y^* - y|$. Celkem tedy je

$$|\Phi(x, y^*, h) - \Phi(x, y, h)| \leqq \left[|1-t| L_0 + |t| \left(1 + \frac{h L_0}{2|t|} \right) L_0 \right] |y^* - y| \tag{3.49}$$

a pro konstantu L v nerovnosti (3.28) platí odhad

$$L \leqq |1-t| L_0 + |t| \left(1 + \frac{h_0 L_0}{2|t|} \right) L_0. \tag{3.50}$$

Jsou-li tedy splněny předpoklady (i) a (ii) z čl. 1, je uvažovaná Rungova-Kuttova metoda regulární. Je zřejmé, že totéž je pravda i pro libovolnou jinou Rungovu-Kuttovu metodu, jen odhad (3.50) je třeba nahradit jiným odhadem, který se od-

vodí stejně snadno. Uveďme jeho konkrétní tvar ještě pro standardní Rungovu-Kuttovu metodu:

$$L \leqq \Big(1 + \frac{h_0 L_0}{2} + \frac{h_0^2 L_0^2}{6} + \frac{h_0^3 L_0^3}{24}\Big) L_0 \leqq \frac{e^{h_0 L} - 1}{h_0}. \tag{3.51}$$

Všimněme si dále u metody (3.12), (3.13) problému odhadu konstanty M v nerovnosti (3.39). Uvažovaná metoda je řádu 2, neboť tak bylo odvozena, takže chceme-li k odhadu této konstanty užít rovnici (3.48), musíme vypočítat (a odhadnout) derivace $\partial^2\Phi/\partial h^2$ a f''. Pišme k tomu cíli funkci Φ ve tvaru

$$\Phi(x, y, h) = (1 - t) f(x, y) + t f(\tilde{x}, \tilde{y}), \tag{3.52}$$

kde jsme položili

$$\tilde{x} = x + \frac{1}{2t} h, \quad \tilde{y} = y + \frac{1}{2t} h f(x, y). \tag{3.53}$$

Odtud však již snadno vypočteme, že platí

$$\frac{\partial^2 \Phi}{\partial h^2} = \frac{1}{4t} \{ f_{xx}(\tilde{x}, \tilde{y}) + 2 f_{xy}(\tilde{x}, \tilde{y}) f(x, y) + f_{yy}(\tilde{x}, \tilde{y}) [f(x, y)]^2 \} \tag{3.54}$$

Určíme-li nyní konstanty M_0 a K tak, aby platilo $M_0 \geqq 1$ a

$$|f(x, y)| \leqq M_0, \quad \Big|\frac{\partial^{i+k} f}{\partial x^i \partial y^k}\Big| \leqq \frac{K}{M_0^{i+k-1}}, \quad i + k \leqq 2 \tag{3.55}$$

v R, lze výraz (3.54) odhadnout takto:

$$\Big|\frac{\partial^2 \Phi(x, y, h)}{\partial h^2}\Big| \leqq \frac{1}{|t|} K M_0. \tag{3.56}$$

Protože je

$$\begin{aligned} f''(x, y) = {} & f_{xx}(x, y) + f_{xy}(x, y) f(x, y) + f_{yy}(x, y) f^2(x, y) + \\ & + f_y(x, y) [f_x(x, y) + f_y(x, y) f(x, y)], \end{aligned} \tag{3.57}$$

dostáváme snadno nerovnost

$$|f''(x, y)| \leqq 4 K M_0 + 2 K^2 M_0. \tag{3.58}$$

Pro konstantu M máme tedy odhad

$$M \leqq K M_0 \Big(\frac{1}{2|t|} + \frac{2}{3} + \frac{1}{3} K\Big). \tag{3.59}$$

Podobným způsobem, jen podstatně komplikovanějším a pracnějším, odvodil L. Bieberbach pro standardní Rungovu-Kuttovu metodu dnes už klasický odhad

$$M \leqq 6 M_0 K (1 + K + K^2 + K^3 + K^4). \tag{3.60}$$

Je přitom samozřejmě třeba předpokládat, že druhý odhad (3.55) platí pro $i + k \leqq 4$, neboť uvažovaná metoda je 4. řádu.

Obraťme se k metodě Taylorova rozvoje řádu p. Odhad konstanty M podle vzorce (3.48) je triviální, neboť v tomto případě je zřejmě $\partial^p \Phi / \partial h^p = 0$. Vyšetření regularity je zde však o něco komplikovanější než u Rungových-Kuttových metod a abychom ji zaručili, je třeba klást na pravou stranu dané diferenciální rovnice podstatně silnější požadavky. Silnější předpoklady o funkci f jsou ostatně nutné už z toho důvodu, aby uvažovaná metoda měla vůbec smysl. Předpokládáme-li např., že funkce f má spojité derivace dostatečně vysokého řádu v pásu $a \leqq x \leqq b$, $-\infty < y < \infty$, položíme-li

$$L_k = \sup_{\substack{a \leqq x \leqq b \\ -\infty < y < \infty}} \Big|\frac{\partial}{\partial y} f^{(k)}(x, y)\Big|, \quad k = 0, 1, \ldots \tag{3.61}$$

a předpokládáme-li dále, že tato čísla jsou konečná, je metoda Taylorova rozvoje řádu p regulární a pro příslušnou konstantu L platí odhad

$$L \leqq L_0 + \frac{h_0}{2} L_1 + \ldots + \frac{h_0^{p-1}}{p!} L_{p-1}, \tag{3.62}$$

jak plyne z věty o střední hodnotě.

Zakončeme tento odstavec příkladem ilustrujícím dosažené výsledky.

Příklad 3.1. Řešme standardní Rungovou-Kuttovou metodou v intervalu $\langle 0, 5 \rangle$ diferenciální rovnici $y' = y$ s počáteční podmínkou $y(0) = 1$. Výsledky výpočtu pro $h = 1/8$ spolu s odhadem chyby získaným z věty 3.2 za použití odhadů (3.51) a (3.60) jsou uvedeny v tab. 3.1. Opět vidíme tentýž jev, na který jsme upozornili už u Eulerovy metody, totiž, že apriorní odhad je nesmírně pesimistický, a tedy prakticky nepoužitelný.

Tabulka 3.1

Řešení diferenciální rovnice $y' = y$ Rungovou-Kuttovou metodou

x_n	1	2	3	4	5
y_n	2,718277	7,389029	20,08543	54,59775	148,4118
e_n	−0,000005	−0,000027	−0,00011	−0,00040	−0,0014
odhad e_n	0,05	0,40	2,95	21,8	161,33

3.3. Asymptotický vzorec pro chybu

V tomto odstavci vyšetříme asymptotické chování chyby obecné jednokrokové metody. Nechť tedy funkce Φ definuje obecnou jednokrokovou metodu řádu $p \geqq 1$. Už u Eulerovy metody jsme viděli, že k získání asymptotického vzorce pro chybu bylo třeba klást na regularitu dané diferenciální rovnice přísnější požadavky, než když jsme chtěli získat pouze údaje o maximální rychlosti konvergence. Budeme

tedy i zde klást silnější hladkostní požadavky, než které nám umožnily odhadnout konstantu ve větě 3.2 pomocí rovnice (3.48). Konkrétně budeme předpokládat, že funkce Φ a f mají $p+1$ spojitých derivací podle všech svých proměnných v oboru $R \times \langle 0, h_0\rangle$, resp. R, kde množina R je definována rovnicí (3.43). Na základě úplně stejných úvah, které nás vedly v předešlém odstavci k odhadu konstanty M, můžeme tvrdit, že za výše uvedených předpokladů existuje funkce φ a konstanta K taková, že platí

$$\Delta(x,y,h) - \Phi(x,y,h) = \varphi(x,y)h^p + \theta K h^{p+1}, \tag{3.63}$$

kde číslo θ závisí na x, y a h, vždy však pro ně platí $|\theta| \leqq 1$. Funkce φ je přitom dána vzorcem

$$\begin{aligned}\varphi(x,y) &= \frac{1}{p!}\frac{\partial^p}{\partial h^p}[\Delta(x,y,h) - \Phi(x,y,h)]\Big|_{h=0} = \\ &= \frac{1}{(p+1)!}f^{(p)}(x,y) - \frac{1}{p!}\frac{\partial^p \Phi(x,y,h)}{\partial h^p}\Big|_{h=0}\end{aligned} \tag{3.64}$$

Výraz $\varphi\big(x,y(x)\big)h^{p+1}$ je *hlavní část lokální diskretizační chyby*, neboť se jí podle (3.6) a (3.63) až na veličiny řádu h^{p+2} rovná.

Pro Eulerovu metodu — abychom uvedli triviální příklad — je $\Phi(x,y,h) = f(x,y)$ a $p=1$; tedy $\varphi(x,y) = f'(x,y)/2 = (f_x + f_y f)/2$. Pro klasickou Rungovu-Kuttovu metodu lze, ovšem už zdaleka ne tak triviálně, vypočítat, že je

$$\begin{aligned}\varphi(x,y) = &-\tfrac{1}{2\,880}f^{(4)} + \tfrac{1}{576}f_y f^{(3)} - \tfrac{1}{288}(f_y^2 - f_{xy} - f_{yy}f)f'' - \\ &- \tfrac{1}{192}(2f_{xy}f_y + 3f_{yy}f_y f - 2f_y^3 + f_{yy}f_x)f'\end{aligned} \tag{3.65}$$

Pro chybu $e_n = y_n - y(x_n)$ platí rovnice (3.32). Použijeme-li vztah (3.6), lze tuto rovnici psát ve tvaru

$$\begin{aligned}e_{n+1} = e_n + h\big[&\Phi(x_n, y_n, h) - \Phi\big(x_n, y(x_n), h\big) + \\ &+ \Phi\big(x_n, y(x_n), h\big) - \Delta\big(x_n, y(x_n), h\big)\big].\end{aligned} \tag{3.66}$$

Za námi učiněných předpokladů jsou splněny předpoklady věty 3.2, existuje tedy konstanta K_1 (nezávislá na h) taková že pro $n = 0, \ldots, N$ platí

$$|e_n| \leqq K_1 h^p. \tag{3.67}$$

Užijme této skutečnosti k získání odhadu rozdílu $\Phi(x_n, y_n, h) - \Phi\big(x_n, y(x_n), h\big)$, který bude přesnější než ten, který se dostane tak, že se užije pouze lipschitzovskosti funkce Φ.

Předně podle Taylorova vzorce je

$$\begin{aligned}&\Phi(x_n, y_n, h) - \Phi\big(x_n, y(x_n), h\big) = \\ &= \Phi\big(x_n, y(x_n) + e_n, h\big) - \Phi\big(x_n, y(x_n), h\big) = \\ &= \Phi_y\big(x_n, y(x_n), h\big)e_n + \tfrac{1}{2}\Phi_{yy}(x_n, y^*, h)e_n^2 = \\ &= \big[\Phi_y\big(x_n, y(x_n), 0\big) + \Phi_{yh}\big(x_n, y(x_n), h^*\big)h\big]e_n + \tfrac{1}{2}\Phi_{yy}(x_n, y^*, h)e_n^2,\end{aligned} \tag{3.68}$$

kde y^* leží mezi body $y(x_n)$, y_n a pro h^* platí $0 < h^* < h$, neboť $p \geqq 1$ a funkce Φ má $p+1$ spojitých derivací podle všech proměnných, tedy alespoň dvě. Protože uvažovaná metoda splňuje podmínku konzistence, jak plyne opět z toho, že je $p \geqq 1$, máme

$$\Phi_y\big(x_n, y(x_n), 0\big) = f_y\big(x_n, y(x_n)\big). \tag{3.69}$$

Vezmeme-li v úvahu, že druhé derivace funkce Φ jsou v $R \times \langle 0, h_0\rangle$ ohraničené, dostáváme z (3.67), (3.68) a (3.69), že existuje konstanta K_2 (která ovšem závisí mimo jiné na konstantě K_1, nezávisí však na h) taková, že platí

$$\big|\Phi(x_n, y_n, h) - \Phi\big(x_n, y(x_n), h\big) - f_y\big(x_n, y(x_n)\big)e_n\big| \leqq K_2 h^{p+1}, \tag{3.70}$$

a tedy existuje θ' takové, že $|\theta'| \leqq 1$ a že

$$\Phi(x_n, y_n, h) - \Phi\big(x_n, y(x_n), h\big) = f_y\big(x_n, y(x_n)\big)e_n + \theta' K_2 h^{p+1}. \tag{3.71}$$

Dosadíme-li do rovnice (3.66) podle (3.63) a (3.71), máme

$$e_{n+1} = e_n + h\big[f_y\big(x_n, y(x_n)\big)e_n + \theta' K_2 h^{p+1} - \varphi\big(x_n, y(x_n)\big)h^p + \theta K h^{p+1}\big]. \tag{3.72}$$

Položíme-li $\bar{e}_n = h^{-p}e_n$, dostáváme pro tuto veličinu rekurenci

$$\bar{e}_{n+1} = \bar{e}_n + h\big[f_y\big(x_n, y(x_n)\big)\bar{e}_n - \varphi\big(x_n, y(x_n)\big)\big] + \theta'' K_3 h^2, \tag{3.73}$$

kde $|\theta''| \leqq 1$ a $K_3 = K + K_2$. Protože je zřejmě $\bar{e}_0 = 0$, představuje rovnici (3.73) Eulerovou metodou pro řešení diferenciální rovnice

$$e' = f_y\big(x, y(x)\big)e - \varphi\big(x, y(x)\big) \tag{3.74}$$

s počáteční podmínkou $e(a) = 0$, kde se v každém kroku dopouštíme navíc chyby velikosti h^2. Podle věty 2.3, jejíž předpoklady jsou v důsledku předpokladů o hladkosti funkcí Φ a f zřejmě splněny, dostáváme tedy výsledek, který zformulujeme v následující větě.

Věta 3.4. *Nechť jsou splněny předpoklady* (i) a (ii) *z čl.* 1 *a buď y_n přibližné řešení vypočtené obecnou jednokrokovou metodou řádu $p \geqq 1$ a buď y příslušné přesné řešení. Nechť dále funkce Φ a f mají v $R \times \langle 0, h_0\rangle$, resp. v R $p+1$ spojitých parciálních derivací podle všech proměnných. Pak platí*

$$y_n - y(x_n) = e(x_n)h^p + O(h^{p+1}), \tag{3.75}$$

kde e je řešením diferenciální rovnice (3.74) *s počáteční podmínkou $e(a) = 0$.*

O vzorci (3.75) platí úplně totéž, co bylo řešeno o analogickém vzorci pro Eulerovu metodu. Funkce e dává ve většině případů velice dobrou představu o chování skutečné chyby, získat ji však explicite je, jak je na první pohled vidět, velice obtížné. Tu skutečnost však, že tato funkce existuje, i když neznáme její konkrétní tvar,

lze zužitkovat úplně analogicky, jako jsme to učinili u Eulerovy metody. Použijeme-li symbol $y(x,h)$ opět k označení přibližného řešení v bodě x získaného obecnou jednokrokovou metodou s krokem h, platí

$$y(x,h) - y(x) = \frac{2^p}{2^p - 1}[y(x,h) - y(x,h/2)] + O(h^{p+1}), \tag{3.76}$$

jak se snadno zjistí úplně stejným postupem jako v čl. 2. Aposteriorní odhad (3.76) získaný metodou poloviČního kroku, je ve většině praktických případů velmi účinný a bývá také v podstatě podkladem mnoha algoritmů pro řešení úloh s počátečními podmínkami s automatickou volbou integračního kroku.

Uveďme závěrem opět jednoduchý příklad.

Příklad 3.2. Vypočtěme pro diferenciální rovnici $y' = y$ s počáteční podmínkou $y(0) = 1$ z příkladu 3.1 a pro standardní Rungovu-Kuttovu metodu asymptotický odhad chyby a odhad metodou poloviČního kroku. Snadno se zjistí, že zde je $\varphi(x,y) = y/120$, takže $e(x) = -xe^x/120$. Z tabulky 3.2 vidíme podobně jako u Eulerovy metody opět velmi dobrou shodu se skutečností.

Tabulka 3.2

Asymptotický odhad chyby

x_n	1	2	3	4	5
e_n	−0,000005	−0,000027	−0,00011	−0,00040	−0,0014
$e(x_n)h^4$	−0,000006	−0,000030	−0,00012	−0,00044	−0,0015
odhad met. pol. kroku	−0,000005	−0,000027	−0,00011	−0,00039	−0,0013

3.4. Problematika zaokrouhlovacích chyb

V tomto odstavci ukážeme, že obecná jednokroková metoda se co do závislosti na zaokrouhlovacích chybách chová úplně stejně jako Eulerova metoda. Postup je v principu úplně stejný jako u Eulerovy metody, a proto budeme postupovat s maximální stručností. Zavedeme-li i zde lokální zaokrouhlovací chybu ε_n rovnicí

$$\tilde{y}_{n+1} = \tilde{y}_n + h\Phi(x_n, \tilde{y}_n, h) + \varepsilon_n, \tag{3.77}$$

dostáváme ihned následující větu.

Věta 3.5. *Buď dána regulární obecná jednokroková metoda. Nechť dále platí, že je* $|\varepsilon_n| \leqq \varepsilon$ *pro* $n = 0, \ldots, N$. *Pak pro celkovou zaokrouhlovací chybu* $\tilde{y} - y_n$ *platí odhad*

$$|\tilde{y} - y_n| \leqq \frac{\varepsilon}{h} E_L(x_n - a), \quad n = 0, \ldots, N. \tag{3.78}$$

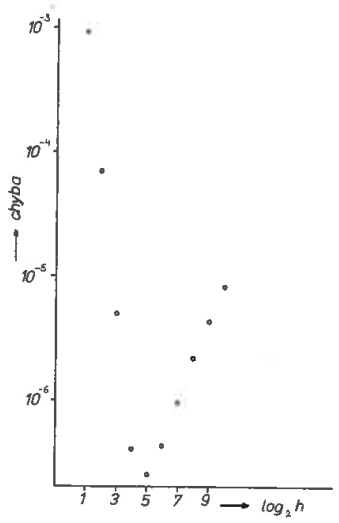
Argumentace, že předpoklad o omezenosti lokální zaokrouhlovací chyby je i v případě jednokrokové metody z praktického hlediska rozumný, je zřejmě i zde stejná jako u Eulerovy metody. Na základě předešlého tedy vidíme, že užití obecné jednokrokové metody řádu p s $p > 1$ má oproti Eulerově metodě z praktického hlediska dvojí efekt: Výpočet je ve většině případů ekonomičtější, neboť vzhledem k tomu, že pro malá h je h^p podstatně menší než h, většinou stačí k dosažení požadované přesnosti brát u obecné jednokrokové metody s $p > 1$ podstatně větší hodnotu h než u Eulerovy metody, a provádět tedy méně kroků. Provádí-li se však méně kroků, je výpočet méně ovlivněn zaokrouhlováním. To má většinou za následek, že při užití metody s $p > 1$ se zmenší kritická chyba, o níž jsme mluvili v souvislosti s Eulerovou metodou, aniž je nutno použít vícenásobnou aritmetiku.

Ilustrujme to na jednoduchém příkladě.

Příklad 3.3. Řešme diferenciální rovnici $y' = y$ a s počáteční podmínkou $y(0) = 1$ standardní Rungovou-Kuttovou metodou s integračním krokem $h = 2^{-s}$, $s = 1, 2, \ldots, 10$.

Obr. 3.1

Závislost celkové chyby na integračním kroku



Na obr. 3.1 je znázorněn průběh celkové chyby v závislosti na velikosti integračního kroku. K výpočtu byla použita 40ti bitová aritmetika. Porovnáním s obr. 2.2,

který znázorňuje tutéž závislost získanou v téže aritmetice pro Eulerovu metodu, vidíme, že kritická chyba skutečně velmi podstatně poklesla.

4 Mnohokrokové metody

V předchozím článku jsme viděli, že výpočet přibližného řešení dané diferenciální rovnice jednokrokovou metodou probíhá tak, že po provedení jednoho kroku se vlastně řeší nový problém s počáteční podmínkou zadanou v tom bodě, do něhož jsme s řešením v předešlém kroku dospěli. To má své výhody: metoda je teoreticky jednoduchá, jednoduše se programuje, změna velikosti integračního kroku je natolik jednoduchá, že jsme o této otázce vlastně dosud nepotřebovali ani explicite mluvit, apod. Na druhé straně však mají tyto metody některé podstatné nevýhody. S jednou z nich spočívající v tom, že odhad lokální diskretizační chyby je (alespoň v případě metody Rungova-Kuttova typu) značně složitý, jsme se už setkali, také to, že v tom okamžiku, kdy vypočítáme přibližné řešení v jednom bodě, zapomeneme všechny informace o řešení v předchozích bodech, se zdá neprozíravé. Naopak intuitivně se zdá zřejmé, že užijeme-li k získání přibližného řešení v bodě $x = x_{n+1}$ nejen to, že známe přibližné řešení v bodě $x = x_n$, ale i to, že je známe také např. v bodech $x = x_{n-1}$ a $x = x_{n-2}$, dostaneme nutně přesnější výsledek. Praxe ukazuje, že je tomu skutečně tak a že metody, které ke konstrukci přibližného řešení užívají nejen informace z bezprostředně předchozího bodu, ale i informace z několika předchozích bodů, a které jsou tedy v tomto smyslu *mnohokrokové*, jsou většinou — hlavně v případech, kdy se vyžaduje vysoká přesnost — efektivnější než metody jednokrokové. Cena, která se za to zaplatí, spočívá, jak uvidíme, ve větší složitosti z programovacího hlediska (je třeba speciální startovní procedury, změna velikosti integračního kroku představuje dosti komplikovaný problém).

Pro mnohokrokovou metodu je tedy charakteristické, že k výpočtu přibližného řešení y_{n+k} v bodě $x = x_{n+k}$ (x_s je zde stejně jako u jednokrokové metody tvaru $x_s = a + sh$) použijeme k předchozích hodnot přibližného řešení. Intuitivně se nabízejí hned dvě možnosti, jak generovat metody tohoto typu. V první z nich vycházíme z identity

$$y(x_r) - y(x_u) = \int_{x_u}^{x_r} f\big(t, y(t)\big)\,\mathrm{d}\,t, \tag{4.1}$$

kterou splňuje přesné řešení dané diferenciální rovnice (1.6) pro libovolné dva body x_r a x_u z intervalu $\langle a, b\rangle$, v níž nahradíme neznámou funkci za integračním znamením vhodným interpolačním polynomem. Metody, které vzniknou na základě tohoto obecného principu, se souhrnně nazývají *metody numerické integrace* a jsou v současné době dosti populární. Nejčastěji se zmíněný interpolační polynom sestrojuje tak, aby v bodech x_ν, $\nu = n, \ldots, n+k-1$, resp. $\nu = n \ldots, n+k$ nabýval hodnot $f_\nu = f(x_\nu, y_\nu)$ a za dvojici x_u, x_r se volí x_{n+k-1}, x_{n+k} nebo x_{n+k-2}, x_{n+k}.

Druhá možnost konstrukce metod, které máme na mysli, vychází přímo z dané diferenciální rovnice. Postupuje se tak, že se sestrojí polynom, který v bodech x_ν, $\nu = n, \ldots, n+k$ nabývá hodnot y_ν, vypočte se derivace tohoto polynomu ve vhodném bodě a položí se rovna hodnotě pravé strany dané diferenciální rovnice v odpovídajícím bodě. Přirozený název pro metody tohoto typu je *metody numerického derivování*.

Obě zmíněné skupiny speciálních mnohokrokových metody popíšeme podrobně v následujících odstavcích.

4.1 Speciální případy

Dříve než začneme popisovat jednotlivé konkrétní metody, uvedeme některé speciální výsledky z teorie interpolace, které nám budou při tomto popisu užitečné.

4.1.1 Interpolace při ekvidistantních argumentech

V tomto odstavci se budeme zabývat úlohou sestrojit polynom stupně nejvýše q, který v bodech $x_p, x_{p-1}, \ldots, x_{p-q}$ ($x_p = a + ph$) nabývá zadaných hodnot $z_p, z_{p-1}, \ldots, z_{p-q}$. Řešení tohoto problému je čtenáři jistě dobře známo: Hledaný polynom existuje, je jediný a je možné jej napsat ve tvaru Lagrangeova interpolačního polynomu. Zde odvodíme jednodušší vzorec, který zužitkovává to, že uzly interpolace jsou ekvidistantní. V této situaci se k zápisu interpolačních polynomů užívá většinou různých typů diferencí. V literatuře byla popsána řada nejrůznějších diferencí a interpolačních vzorců založených na nich. Pro naše účely je vhodné pracovat s pojmem diference zpět. *První diferenci zpět* budeme značit symbolem ∇ a definovat rovností

$$\nabla z_p \equiv z_p - z_{p-1}; \tag{4.2}$$

m-tou diferenci zpět pak označíme symbolem ∇^m a definujeme rekurentně rovností

$$\nabla^m z_p \equiv \nabla(\nabla^{m-1} z_p), \quad m = 2, 3, \ldots, \tag{4.3}$$

kde klademe $\nabla^1 = \nabla$. Z důvodu zjednodušení dalších zápisů položíme ještě $\nabla^0 z_p = z_p$. Úplnou indukcí snadno zjistíme, že platí

$$\nabla^q z_p = \sum_{m=0}^{q} (-1)^m \binom{q}{m} z_{p-m}, \quad q = 0, 1, \ldots, \tag{4.4}$$

kde je

$$\binom{q}{0} = 1, \quad \binom{q}{m} = \frac{q(q-1)\ldots(q-m+1)}{m(m-1)\ldots 1}, \quad m = 1, 2, \ldots \tag{4.5}$$

a naopak, že pro každé $q = 0, 1, \ldots$ platí

$$z_{p-q} = \sum_{m=0}^{q} (-1)^m \binom{q}{m} \nabla^m z_p; \tag{4.6}$$

q-tou diferenci zpět lze tedy vyjádřit jako lineární kombinaci hodnot z_p, z_{p-1}, ..., z_{p-q} a naopak hodnotu z_{p-q} jako lineární kombinaci nulté, první atd. až q-té diference zpět.

Pomocí pojmu diference zpět už snadno zapíšeme hledaný polynom.

Věta 4.1. *Položme*

$$P(x) = \sum_{m=0}^{q} (-1)^m \binom{-s}{m} \nabla^m z_p, \tag{4.7}$$

kde

$$s = \frac{x - x_p}{h}. \tag{4.8}$$

Pak P je polynom stupně nejvýše q v x a platí

$$P(x_{p-r}) = z_{p-r}, \quad r = 0, \ldots, q. \tag{4.9}$$

D ů k a z . Předně výraz $\binom{-s}{m} = \binom{(x_p - x)/h}{m}$ je podle vzorců (4.5) zřejmě polynom stupně m v x, a tedy P je polynom stupně nejvýše q. Protože je $\binom{r}{m} = 0$ pro celé kladné r a $m > r$ (viz vzorec (4.5), který platí pro libovolné reálné q), platí

$$P(x_{p-r}) = \sum_{m=0}^{q} (-1)^m \binom{r}{m} \nabla^m z_p = \sum_{m=0}^{r} (-1)^m \binom{r}{m} \nabla^m z_p. \tag{4.10}$$

Poslední součet v rovnici (4.10) je však podle (4.6) roven číslu z_{p-r}. Věta je dokázána.

Věta 4.2. *Buď I interval obsahující body $x_p, x_{p-1}, \ldots, x_{p-q}$ a buď z funkce definovaná na I, která má v I $q+1$ derivací a buď P polynom (4.7) s $z_{p-r} = z(x_{p-r})$, $r = 0, \ldots, q$. Pak ke každému $x \in I$ existuje bod $\xi \in I$ takový, že platí*

$$z(x) - P(x) = (-1)^{q+1} h^{q+1} \binom{-s}{q+1} z^{(q+1)}(\xi). \tag{4.11}$$

D ů k a z . Je-li $x = x_{p-r}$ pro některé r, je rovnice (4.11) splněna a to dokonce pro libovolné ξ. Můžeme tedy předpokládat, že je $x \neq x_{p-r}$ pro $r = 0, \ldots, q$. Definujme v tomto případě funkci φ proměnné $t \in I$ předpisem

$$\varphi(t) = z(t) - P(t) - \frac{z(x) - P(x)}{\Omega(x)} \Omega(t), \tag{4.12}$$

kde

$$\Omega(t) = (t - x_p)(t - x_{p-1}) \ldots (t - x_{p-q}). \tag{4.13}$$

Funkce φ má v intervalu I zřejmě $q+1$ derivací a platí $\varphi(x_{p-r}) = 0$ pro $r = 0, \ldots, q$ a $\varphi(x) = 0$. V intervalu I tedy existuje $q+2$ navzájem různých bodů,

ve kterých je hladká funkce φ rovna nule. Opakovaným užitím Rolleovy věty ihned plyne existence takového $\xi \in I$, že platí

$$\varphi^{(q+1)}(\xi) = 0. \tag{4.14}$$

Je však $P^{(q+1)}(t) \equiv 0$, neboť stupeň polynomu P je nejvýše q; dále je $\Omega^{(q+1)}(t) = (q+1)!$. Dosazením do rovnice (4.14) dostáváme

$$z(x) - P(x) = \frac{1}{(q+1)!} \Omega(x) z^{(q+1)}(\xi). \tag{4.15}$$

Použijeme-li nyní ve vzorci (4.13) rovnosti $x - x_{p-r} = x - x_p + rh = sh + rh = (-h)(-s - r)$, dostaneme z (4.15) ihned vzorec (4.11). Věta je dokázána.

4.1.2 Adamsova-Bashforthova metoda

Tato metoda je speciálním případem metody numerické integrace. Klademe v ní $x_r = x_{n+k}$, $x_u = x_{n+k-1}$ ($k \geq 1$) a interpolační polynom P prokládáme body (x_ν, f_ν), $\nu = n+k-1, \ldots, n$. Položíme-li tedy v odst. 4.1.1 $p = n+k-1$ a $q = k-1$, je polynomem P tvaru

$$P(x) = \sum_{m=0}^{k-1} (-1)^m \binom{-s}{m} \nabla^m f_{n+k-1}, \tag{4.16}$$

kde $s = (x - x_{n+k-1})/h$. Odtud integrací v mezích od x_{n+k-1} do x_{n+k} už snadno odvodíme, že Adamsova-Bashforthova metoda je dána vzorcem

$$y_{n+k} - y_{n+k-1} = h \sum_{m=0}^{k-1} \gamma_m \nabla^m f_{n+k-1}, \tag{4.17}$$

kde

$$\gamma_m = \frac{1}{h} \int_{x_{n+k-1}}^{x_{n+k}} (-1)^m \binom{-s}{m} \mathrm{d}x = (-1)^m \int_0^1 \binom{-s}{m} \mathrm{d}s \tag{4.18}$$

nebo vzorcem

$$y_{n+k} - y_{n+k-1} = h \sum_{\nu=0}^{k-1} \beta_\nu^{(k)} f_{n+\nu}, \tag{4.19}$$

kde

$$\beta_\nu^{(k)} = (-1)^{k-1-\nu} \sum_{m=k-1-\nu}^{k-1} \binom{m}{k-1-\nu} \gamma_m, \tag{4.20}$$

chceme-li, aby na pravé straně vzorce vystupovaly přímo hodnoty funkce f. Vzorec (4.19) se snadno dostane za vzorce (4.17) užitím rovnice (4.4). Všimněme si, že koeficienty ve vyjádření (4.17) na rozdíl od vyjádření (4.19) nezávisí na k. To je také výhoda tvaru (4.17), který byl populární zejména v předpočítačové éře.

Při výpočtu přibližného řešení se vzorec (4.17) nebo (4.19) užívá postupně pro $n = 0, 1, \ldots$. Na začátku výpočtu musíme mít tedy k dispozici nejen hodnotu y_0, která je dána počáteční podmínkou, jako tomu bylo u obecné jednokrokové metody, ale ještě hodnoty $y_1, \ldots, y_{k-1}$, které musíme získat nějakým jiným způsobem (např. jednokrokovou metodou). Tento jev je pro mnohokrokové metody typický.

Počítat koeficienty Adamsova-Bashforthova vzorce podle (4.18) je nepohodlné. Odvoďme proto jiný způsob jejich výpočtu, který nevyžaduje počítání žádných integrálů.

Definujme k tomu cíli funkci G komplexní proměnné t předpisem

$$G(t) = \int_0^1 (1-t)^{-s}\,\mathrm{d}\,s. \tag{4.21}$$

Protože zřejmě platí

$$G(t) = -\frac{t}{(1-t)\ln(1-t)}, \tag{4.22}$$

je funkce G holomorfní v kruhu $|t| < 1$. Podle binomické věty však platí

$$(1-t)^{-s} = \sum_{m=0}^{\infty} \binom{-s}{m} (-t)^m \tag{4.23}$$

a tato řada konverguje absolutně a stejnoměrně uvnitř jednotkového kruhu. Lze ji tedy integrovat člen po členu, čímž dostaneme

$$G(t) = \sum_{m=0}^{\infty} (-1)^m \left[\int_0^1 \binom{-s}{m} \mathrm{d}\,s\right] t^m = \sum_{m=0}^{\infty} \gamma_m t^m. \tag{4.24}$$

Dosadíme-li tento výsledek do (4.22), dostaneme identitu

$$-\frac{\ln(1-t)}{t} \sum_{m=0}^{\infty} \gamma_m t^m = \frac{1}{1-t} \tag{4.25}$$

nebo

$$\sum_{\nu=0}^{\infty} \frac{1}{\nu+1} t^\nu \sum_{m=0}^{\infty} \gamma_m t^m = \sum_{m=0}^{\infty} t^m \tag{4.26}$$

platnou pro každé $|t| < 1$. Je však

$$\sum_{\nu=0}^{\infty} \frac{1}{\nu+1} \sum_{m=0}^{\infty} \gamma_m t^{m+\nu} = \sum_{\nu=0}^{\infty} \frac{1}{\nu+1} \sum_{m=\nu}^{\infty} \gamma_{m-\nu} t^m = \tag{4.27}$$

$$= \sum_{m=0}^{\infty} \left(\sum_{\nu=0}^{m} \frac{1}{\nu+1} \gamma_{m-\nu} \right) t^m,$$

neboť upravovaná řada konverguje absolutně a lze tedy libovolně přerovnat pořádek jejích členů. Dosadíme-li do identity (4.26) podle (4.27) a porovnáme-li koeficienty

u stejných mocnin proměnné t na levé a pravé straně, dostáváme

$$\sum_{\nu=0}^{m} \frac{1}{\nu+1} \gamma_{m-\nu} = 1 \tag{4.28}$$

pro $m = 0, 1, \ldots$. Koeficienty γ_m můžeme tedy počítat z (4.28) rekurentně. V tab. 4.1 jsou uvedeny tyto koeficienty a v tab. 4.2 koeficienty $\beta_\nu^{(k)}$ pro několik prvních hodnot indexů.

Tabulka 4.1

Koeficienty γ Adamsovy-Bashforthovy metody

m	0	1	2	3	4	5	6
γ_m	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$	$\frac{19\,087}{60\,480}$

Tabulka 4.2

Koeficienty β Adamsovy-Bashforthovy metody

ν	0	1	2	3	4	5
$\beta_\nu^{(1)}$	1					
$2\beta_\nu^{(2)}$	−1	3				
$12\beta_\nu^{(3)}$	5	−16	23			
$24\beta_\nu^{(4)}$	−9	37	−59	55		
$720\beta_\nu^{(5)}$	251	−1 274	2 616	−2 774	1 901	
$1\,440\beta_\nu^{(6)}$	−425	2 627	−6 798	9 482	−7 673	4 227

Všimněme si ještě vztahu Adamsovy-Bashforthovy metody k původní diferenciální rovnici. U obecné jednokrokové metody jsme její kvalitu měřili velikostí lokální chyby, tj. velikostí výrazu $y(x+h) - y(x) - h\Phi(x, y(x), h)$ jako funkce h. U Adamsovy-Bashforthovy metody tomuto výrazu odpovídá výraz

$$R_k^{AB} = y(x_{n+k}) - y(x_{n+k-1}) - h \sum_{m=0}^{k-1} \gamma_m \nabla^m f\big(x_{n+k-1}, y(x_{n+k-1})\big) = \tag{4.29}$$

$$= y(x_{n+k}) - y(x_{n+k-1}) - h \sum_{m=0}^{k-1} \gamma_m \nabla^m y'(x_{n+k-1}),$$

který tak hraje roli lokální chyby. Za předpokladu, že přesné řešení uvažované diferenciální rovnice má $k+1$ spojitých derivací, platí podle věty 4.2 ($p = n+k-1$,

$q = k - 1$, $z(x) = y'(x)$)

$$y'(x) = \sum_{m=0}^{k-1} (-1)^m \binom{-s}{m} \nabla^m y'(x_{n+k-1}) + (-1)^k h^k \binom{-s}{k} y^{(k+1)}(\xi), \tag{4.30}$$

kde $s = (x - x_{n+k-1})/h$. Integrací v mezích od x_{n+k-1} do x_{n+k} odtud plyne, že je

$$R_k^{AB} = (-1)^k h^k \int_{x_{n+k-1}}^{x_{n+k}} \binom{-s}{k} y^{(k+1)}(\xi)\, \mathrm{d}\, x. \tag{4.31}$$

Protože funkce $\binom{-s}{k}$ nemění v intervalu $\langle x_{n+k-1}, x_{n+k}\rangle$ znaménko, dává první věta o střední hodnotě integrálního počtu jednoduchý výsledek

$$R_k^{AB} = (-1)^k h^k y^{(k+1)}(\eta) \int_{x_{n+k-1}}^{x_{n+k}} \binom{-s}{k} \mathrm{d}\, x = \gamma_k y^{(k+1)}(\eta) h^{k+1}. \tag{4.32}$$

4.1.3 Adamsova-Moultonova metoda

Tato metoda je velmi podobná Adamsově-Bashforthově metodě. Jediný rozdíl je ten, že interpolační polynom prokládáme body (x_ν, f_ν) pro $\nu = n, \ldots, n + k$, takže je $p = n + k$ a $q = k$. Metoda je tedy dána vzorcem

$$y_{n+k} - y_{n+k-1} = h \sum_{m=0}^{k} \gamma_m^* \nabla^m f_{n+k}, \tag{4.33}$$

kde

$$\gamma_m^* = (-1)^m \int_{-1}^{0} \binom{-s}{m} \mathrm{d}\, s, \quad m = 0, 1, \ldots, \tag{4.34}$$

nebo vzorcem

$$y_{n+k} - y_{n+k-1} = h \sum_{\nu=0}^{k} \beta_\nu^{*(k)} f_{n+\nu}, \tag{4.35}$$

kde

$$\beta_\nu^{*(k)} = (-1)^{k-\nu} \sum_{m=k-\nu}^{k} \binom{m}{k-\nu} \gamma_m^*, \quad \nu = 0, \ldots, k. \tag{4.36}$$

Podobně jako u Adamsovy-Bashforthovy metody je třeba se na vzorec (4.33) nebo (4.35) dívat tak, že pomocí něj máme vypočítat y_{n+k} za předpokladu, že hodnoty $y_{n+k-1}, \ldots, y_n$ jsou už známé. Hodnota y_{n+k} se zde však na rozdíl od Adamsovy-Bashforthovy metody vyskytuje i na pravé straně vzorce (4.33) (nebo (4.35)). Rovnice (4.33) tedy představuje obecně nelineární rovnici pro určení y_{n+k}. V tomto smyslu mluvíme o Adamsově-Moultonově metodě jako o metodě *implicitní* na rozdíl od *explicitní* Adamsovy-Bashforthovy metody. V daný okamžik tedy vlastně nevíme, je-li Adamsova-Moultonova metoda vzorcem (4.33) vůbec dobře definována, neboť zatím nic nevíme o řešitelnosti zmíněné nelineární rovnice. Později

uvidíme, že pro dostatečně malá h lze tuto rovnici řešit postupnými aproximacemi. Tím sice zatím existující mezeru v definici Adamsovy-Moultonovy metody vyplníme, vyvstane ovšem ihned přirozená otázka, zda má vůbec smysl zavádět implicitní metody, které jsou vzhledem k nutnosti řešení nelineární rovnice z hlediska výpočetní práce patrně méně efektivní než explicitní metody. Později uvidíme, že jiné kladné vlastnosti tuto nevýhodu implicitních metod bohatě vyváží.

Koeficienty γ_m^* lze počítat pomocí integrálů (4.34) nebo, patrně ekonomičtěji, z rekurence

$$\gamma_0^* = 1, \quad \sum_{\nu=0}^{m} \frac{1}{\nu + 1} \gamma_{m-\nu}^* = 0, \quad m = 1, 2, \ldots, \tag{4.37}$$

která se odvodí úplně stejně jako v případě Adamsovy-Bashforthovy metody. Některé konkrétní hodnoty koeficientů γ_m^* a $\beta_\nu^{*(k)}$ jsou uvedeny v tab. 4.3. a 4.4.

Tabulka 4.3

Koeficienty γ^* Adamsovy-Moultonovy metody

m	0	1	2	3	4	5	6
γ_m^*	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$	$-\frac{863}{60\,480}$

Tabulka 4.4

Koeficienty β^* Adamsovy-Moultonovy metody

ν	0	1	2	3	4	5
$2\beta_\nu^{*1}$	1	1				
$12\beta_\nu^{*2}$	−1	8	5			
$24\beta_\nu^{*3}$	1	−5	19	9		
$720\beta_\nu^{*4}$	−19	106	−264	646	251	
$1\,440\beta_\nu^{*5}$	27	−173	482	−798	1 427	475

Stejně snadno jako u Adamsovy-Bashforthovy metody se odvodí, že pro lokální chybu R_k^{AM} Adamsovy-Moultonovy metody platí (samozřejmě za předpokladu dostatečné hladkosti řešení)

$$R_k^{AM} \equiv y(x_{n+k}) - y(x_{n+k-1}) - h \sum_{m=0}^{k} \gamma_m^* \nabla^m y'(x_{n+k}) = \\ = \gamma_{k+1}^* y^{(k+2)}(\eta) h^{k+2}. \tag{4.38}$$

4.1.4 Nyströmova metoda

Jde o explicitní metodu, která vznikne integrací polynomu P, jehož graf prochází body (x_ν, f_ν) pro $\nu = n+k-1, \dots, n$, v mezích x_{n+k-2} do x_{n+k}. Pro $k = 2, 3, \dots$ tedy platí

$$y_{n+k} - y_{n+k-2} = h \sum_{m=1}^{k-1} \kappa_m \nabla^m f_{n+k-1}, \tag{4.39}$$

kde koeficienty κ_m jsou dány integrály

$$\kappa_m = (-1)^m \int_{-1}^{1} \binom{-s}{m} \mathrm{d}\, s, \tag{4.40}$$

nebo rekurencemi

$$\kappa_0 = 2, \quad \sum_{\nu=0}^{m} \frac{1}{\nu+1} \kappa_{m-\nu} = 1, \quad m = 1, 2, \dots \tag{4.41}$$

a jejichž hodnoty jsou pro $m = 0, \dots, 6$ uvedeny v tab. 4.5.

Tabulka 4.5

Koeficienty κ Nyströmovy metody

m	0	1	2	3	4	5	6
κ_m	2	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{29}{90}$	$\frac{14}{45}$	$\frac{1139}{3780}$

Vzorec (4.39) lze samozřejmě přepsat do tvaru, který obsahuje pouze hodnoty funkce f. Pro příslušné koeficienty se odvodí podobné vzorce jako v případě Adamsovy-Bashforthovy nebo Adamsovy-Moultonovy metody. Naproti tomu pro lokální chybu Nyströmovy metody už obecně neplatí jednoduchý vzorec typu (4.32) nebo (4.38).

4.1.5 Zobecněná Milnova-Simpsonova metoda

Tato implicitní metoda je pro $k = 2, 3, \dots$ dána vzorcem

$$y_{n+k} - y_{n+k-2} = h \sum_{m=0}^{k} \kappa_m^* \nabla^m f_{n+k}. \tag{4.42}$$

Pro koeficienty κ_m^* platí

$$\kappa_m^* = (-1)^m \int_{-2}^{0} \binom{-s}{m} \mathrm{d}\, s, \tag{4.43}$$

nebo

$$\kappa_0^* = 2, \quad \kappa_1^* = -2, \quad \kappa_m^* = \sum_{\nu=0}^{m} \frac{1}{\nu+1} \kappa_{m-\nu}^*, \quad m = 2, 3, \dots, \tag{4.44}$$

a jsou pro $m = 0, \dots, 6$ uvedeny v tab. 4.6.

Tabulka 4.6

Koeficienty κ^* zobecněné Milnovy-Simpsonovy metody

m	0	1	2	3	4	5	6
κ^*	2	−2	$\frac{1}{3}$	0	$-\frac{1}{90}$	$-\frac{1}{90}$	$-\frac{37}{3780}$

Čtenář si jistě už snadno sám sestaví polynom, jehož integrací tato metoda vznikla. Poznamenejme jen ještě, že ani u Milnovy-Simpsonovy metody nelze odvodit tak jednoduchý vzorec pro lokální chybu, jako u Adamsových metod.

4.1.6 Metody založené na numerickém derivování

Základní myšlenka konstrukce těchto metod spočívá, jak už bylo řečeno, v tom, že se sestrojí polynom

$$P(x) = \sum_{m=0}^{k} (-1)^m \binom{-s}{m} \nabla^m y_{n+k}, \quad s = \frac{x - x_{n+k}}{h}, \tag{4.45}$$

který nabývá v bodech $x_{n+k}, \dots, x_n$ hodnot $y_{n+k}, \dots, y_n$, zvolí se bod x_{n+k-r} ($0 \leqq r \leqq k$) a za derivaci v diferenciální rovnici (1.6) psané pro bod $x = x_{n+k-r}$ se dosadí derivace tohoto polynomu v tomto bodě. Vzniklá metoda se tedy dá zapsat ve tvaru

$$\sum_{m=1}^{k} \delta_{rm} \nabla^m y_{n+k} = h f_{n+k-r}, \tag{4.46}$$

kde

$$\begin{aligned} \delta_{rm} &= (-1)^m h \frac{\mathrm{d}}{\mathrm{d}\, x} \binom{-s}{m}\Big|_{x = x_{n+k-r}} = \\ &= (-1)^m \frac{\mathrm{d}}{\mathrm{d}\, s} \binom{-s}{m}\Big|_{s=-r}, \quad m = 1, 2, \dots \end{aligned} \tag{4.47}$$

Pro $r = 0$ jde o metodu implicitní, jinak jsou to metody explicitní. Položíme-li

$$D_r(t) = \sum_{m=1}^{\infty} \delta_{rm} t^m, \tag{4.48}$$

je

$$D_r(t) = \sum_{m=1}^{\infty}(-t)^m \frac{\mathrm{d}}{\mathrm{d}s}\binom{-s}{m}\Big|_{s=-r} = \frac{\mathrm{d}}{\mathrm{d}s}\left[\sum_{m=0}^{\infty}\binom{-s}{m}(-t)^m\right]\Big|_{s=-r} = \tag{4.49}$$

$$= \frac{\mathrm{d}}{\mathrm{d}s}(1-t)^{-s}\Big|_{s=-r} = -(1-t)^r \ln(1-t).$$

Odtud snadno plyne, že platí

$$D_{r+1}(t) = (1-t)D_r(t). \tag{4.50}$$

Dosadíme-li (4.48) do (4.50), porovnáme-li koeficienty u stejných mocnin na levé a pravé straně a položíme-li ještě $\delta_{r0} = 0$ pro $r = 0, 1, \ldots$, dostáváme rovnici

$$\delta_{r+1,m} = \delta_{rm} - \delta_{r,m-1}, \tag{4.51}$$

která platí pro $r = 0, 1, \ldots$ a $m = 1, 2, \ldots$. Položíme-li v (4.49) $r = 0$ a vyjádříme-li funkci na pravé straně ve tvaru mocninné řady, máme

$$\delta_{0m} = \frac{1}{m}, \quad m = 1, 2, \ldots. \tag{4.52}$$

Z rovnic (4.51) a (4.52) však už snadno generujeme koeficienty δ_{rm} rekurentně. Prvních několik jejich hodnot je uvedeno v tab. 4.7.

Tabulka 4.7

Koeficienty δ metody založené na numerickém derivování

$r\backslash m$	1	2	3	4	5	6
0	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$
1	1	$-\frac{1}{2}$	$-\frac{1}{6}$	$-\frac{1}{12}$	$-\frac{1}{20}$	$-\frac{1}{30}$
2	1	$-\frac{3}{2}$	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{30}$	$\frac{1}{60}$
3	1	$-\frac{5}{2}$	$\frac{7}{6}$	$-\frac{17}{12}$	$-\frac{1}{20}$	$-\frac{1}{60}$

Vzorec (4.46) lze psát také ve tvaru

$$\sum_{\nu=0}^{k}\alpha_\nu^{(k)} y_{n+\nu} = h f_{n+k-r}, \tag{4.53}$$

který neobsahuje žádné diference. V tomto případě se určí koeficienty $\alpha_\nu^{(k)}$ ze vzorce

$$\alpha_\nu^{(k)} = (-1)^{k-\nu}\sum_{m=k-\nu}^{k}\delta_{rm}\binom{m}{k-\nu}, \quad \nu = 0, \ldots, k, \tag{4.54}$$

jak se snadno zjistí užitím rovnice (4.4).

Pro lokální chybu R_k^D popsané metody, která je definována v tomto případě vzorcem

$$R_k^D \equiv \sum_{m=1}^{k}\delta_{rm}\nabla^m y(x_{n+k-r}) - hy'(x_{m+k-r}), \tag{4.55}$$

platí

$$R_k^D = \delta_{r,k+1} y^{(k+1)}(\eta) h^{k+1}. \tag{4.56}$$

Tento vzorec se odvodí stejně snadno jako obdobný vzorec pro Adamsovy metody.

4.2 Obecná lineární mnohokroková metoda

Všechny metody vyšetřované v předešlém odstavci měly společné, že svazovaly přibližné hodnoty hledaného řešení a jeho derivací lineárně. Je proto přirozené omezit se v dalším na tento druh závislosti a za *obecnou lineární k-krokovou metodu* považovat metodu danou vzorcem

$$\sum_{\nu=0}^{k}\alpha_\nu y_{n+\nu} = h\sum_{\nu=0}^{k}\beta_\nu f_{n+\nu}, \tag{4.57}$$

kde $\alpha_0, \ldots, \alpha_k$, $\beta_0, \ldots, \beta_k$ jsou reálné konstanty, h je integrační krok a f_s je zkrácený zápis pro $f(x_s, y_s)$. Všude v dalším budeme předpokládat, že je $\alpha_k \neq 0$ a že koeficienty α_0 a β_0 nejsou současně rovny nule. O k-krokové metodě budeme také všude tam, kde není konkrétní hodnota parametru k podstatná, mluvit jako o *mnohokrokové metodě*.

Na rovnici (4.57) je třeba se dívat tak, že z ní máme určit přibližnou hodnotu y_{n+k} řešení v bodě $x = x_{n+k}$ za předpokladu, že přibližné hodnoty $y_{n+k-1}, \ldots, y_n$ jsou známé a tento postup opakovat postupně pro $n = 0, 1, \ldots, N-k$, kde $N = [(b-a)h]$. Je tedy především nutno předpokládat, že přibližná řešení $y_0, \ldots, y_{k-1}$ jsou dána. Vzorec (4.57) tak vyžaduje na rozdíl od jednokrokové metody k počátečních hodnot. Na tuto skutečnost jsme už ostatně upozornili při popisu speciálních mnohokrokových metod.

Aby metoda daná rovnicí (4.57) měla vůbec rozumný smysl, je třeba se přesvědčit, že veličina y_{n+k} je jí (za předpokladu, že veličiny $y_n, y_{n+1}, \ldots, y_{n+k-1}$ jsou známé) skutečně jednoznačně určena. To je zřejmé na první pohled v případě, že je $\beta_k = 0$. V tomto případě budeme mluvit o *explicitní* metodě. V případě, že je $\beta_k \neq 0$, nazveme příslušnou metodu *implicitní*; u ní je třeba vypočítat y_{n+k} jako řešení obecně nelineární rovnice

$$y = F(y), \tag{4.58}$$

kde

$$F(y) = h\frac{\beta_k}{\alpha_k} f(x_{n+k}, y) + c \tag{4.59}$$

a

$$c = \frac{1}{\alpha_k}\left(h\sum_{\nu=0}^{k-1}\beta_\nu f_{n+\nu} - \sum_{\nu=0}^{k-1}\alpha_\nu y_{n+\nu}\right), \tag{4.60}$$

a je to tedy známé číslo. Předpokládáme-li, že jsou splněny předpoklady (i) a (ii) z čl. 1, je funkce F definovaná pro libovolná y a platí pro ni

$$|F(y) - F(y^*)| \leqq hL\frac{|\beta_k|}{|\alpha_k|}\,|y - y^*|. \tag{4.61}$$

Splňuje-li h podmínku

$$h < \frac{|\alpha_k|}{L|\beta_k|}, \tag{4.62}$$

je zobrazení F kontrakce a podle Banachovy věty o pevném bodu má rovnice (4.58) právě jedno řešení a toto řešení lze vypočítat metodou postupných aproximací.

Ať už je tedy daná metoda explicitní nebo implicitní, má — alespoň pro malá h — smysl.

Pojem lokální chyby a s ním související pojem řádu lineární mnohokrokové metody bude hrát stejně důležitou roli, jako tomu bylo u jednokrokové metody. Uveďme proto přesné definice těchto pojmů.

Definice 4.1. Buď dána lineární mnohokroková metoda a buď y libovolná funkce, která je v $\langle a, b\rangle$ definovaná, spojitá a spojitě diferencovatelná. Výraz L definovaný rovnicí

$$L(y(x); h) = \sum_{\nu=0}^{k}\alpha_\nu y(x+\nu h) - h\sum_{\nu=0}^{k}\beta_\nu y'(x+\nu h) \tag{4.63}$$

nazveme *lokální chybou* dané *mnohokrokové metody.*

Všimněme si, že v důsledku linearity dané metody není nutné v definici její lokální chyby vůbec mluvit o řešení dané diferenciální rovnice.

Definice 4.2. Řekneme, že lineární mnohokroková metoda je *řádu* p, platí-li $C_0 = C_1 = \ldots = C_p = 0$, $C_{p+1} \neq 0$, kde C_p jsou konstanty, definované rovnicemi

$$\begin{aligned} C_0 &= \alpha_0 + \ldots + \alpha_k, \\ C_1 &= \alpha_1 + 2\alpha_2 + \ldots k\alpha_k - (\beta_0 + \ldots + \beta_k), \\ C_q &= \frac{1}{q!}(\alpha_1 + 2^q\alpha_2 + \ldots + k^q\alpha_k) - \\ &\quad - \frac{1}{(q-1)!}(\beta_1 + 2^{q-1}\beta_2 + \ldots + k^{q-1}\beta_k), \quad q = 2, 3 \ldots \end{aligned} \tag{4.64}$$

Abychom objasnili důvody, které nás vedly k této definici, předpokládejme na okamžik, že funkce y je nekonečně diferencovatelná a že ji lze v okolí každého bodu

rozvinout v Tayolorovu řadu. Pak platí

$$\begin{aligned} L(y(x), h) &= \sum_{\nu=0}^{k}\alpha_\nu\left[y(x) + \nu h y'(x) + \frac{1}{2!}\nu^2h^2y''(x) + \ldots\right] - \\ &\quad - \sum_{\nu=0}^{k}\beta_\nu\left[hy'(x) + \nu h^2y''(x) + \frac{1}{2!}\nu^2h^3y'''(x) + \ldots\right] = \\ &= C_0y(x) + C_1hy'(x) + C_2h^2y''(x) + \ldots + C_qh^qy^{(q)}(x) + \ldots \end{aligned} \tag{4.65}$$

Rovnice (4.64) tak skutečně vyjadřují intuitivně očekávanou skutečnost, totiž, že pro každou dostatečně hladkou funkci y je

$$L(y(x); h) = O(h^{p+1}). \tag{4.66}$$

Vidíme také, že linearita dané metody, která už zjednodušila definici lokální chyby, umožnila definovat její řád čistě algebraicky, tj. pouze pomocí algebraických vztahů pro její koeficienty.

Vezeme-li v úvahu výsledky, k nimž jsme dospěli v případě jednokrokové metody, zdá se rozumné konstruovat lineární mnohokrokové metody tak, aby měly pokud možno vysoký řád. Předtím, než budeme tyto otázky zkoumat teoreticky, uveďme jednoduchý příklad.

Příklad 4.1. Řešme diferenciální rovnici $y' = -y$ s počáteční podmínkou $y(0) = 1$ na intervalu $\langle 0, 1\rangle$ metodou

$$y_{n+2} + 4y_{n+1} - 5y_n = h(4f_{n+1} + 2f_n), \tag{4.67}$$

která je řádu 3, jak se snadno vypočte z rovnic (4.64). Výsledky jsou pro $h = 1/10$ a $1/20$ uvedeny v tab. 4.8. Protože užitá metoda je dvoukroková, je třeba kromě hodnoty y_0 $(= 1)$ dodat ještě hodnotu y_1; za tu jsme zvolili hodnotu přesného řešení v bodě $0,1$, resp. $0,05$. Z tabulky je na první pohled vidět, že jsme dostali zcela nesmyslné výsledky. U metody (4.67) lze tedy jen stěží očekávat konvergenci.

Uvedený příklad ukazuje, že skutečnost, že lokální chyba je pro malá h malá, nemusí ještě u mnohokrokové metody stačit k tomu, aby byla malá i globální chyba. Abychom uhádli, jaké podmínky kromě malosti lokální chyby je třeba klást na danou metodu, aby byla zaručena konvergence, začneme teoretické zkoumání obecné mnohokrokové metody vyšetřováním nutných podmínek konvergence.

4.2.1 Nutné podmínky konvergence

Zatímco u jednokrokové metody záviselo přibližné řešení kromě na počáteční podmínce už jen na integračním kroku, u mnohokrokové metody závisí ještě na dalších počátečních podmínkách. Přibližné řešení získané k-krokovou metodou tedy závisí na integračním kroku nejen proto, že na něm závisí daná metoda, ale také proto, že

Tabulka 4.8

Numerické řešení diferenciální rovnice $y' = -y$ metodou (4.67)

x	$h = 0,1$		$h = 0,05$	
	přibližné řešení	chyba	přibližné řešení	chyba
0,0	1,000000	0	1,000000	0
0,1	0,904837	0	0,904836	−0,000001
0,2	0,818715	−0,000015	0,818711	−0,000019
0,3	0,740872	0,000054	0,740315	−0,000503
0,4	0,669997	−0,000323	0,656994	−0,013326
0,5	0,608200	0,001669	0,252935	−0,353596
0,6	0,539907	−0,008905	−8,833877	−9,382689
0,7	0,543768	0,047183	−248,4746	−248,9711
0,8	0,198971	−0,250358	−6606,041	−6606,490
0,9	1,734617	1,328048	−175303,9	−175304,3
1,0	−6,677257	−7,045136	−4651727	−4651728

na něm závisí i další počáteční podmínky. Veškeré úvahy o konvergenci k-krokové metody musí proto brát zřetel na tuto závislost. Studium konvergence proto začneme přesnou definicí tohoto pojmu.

Definice 4.3. Řekneme, že lineární k-kroková metoda (4.57) je *konvergentní*, platí-li

$$\lim_{\substack{h\to 0\\ x_n = x}} y_n = y(x) \tag{4.68}$$

pro každé $x \in \langle a, b\rangle$, kde y je řešení libovolné diferenciální rovnice (1.6) s počáteční podmínkou (1.7), jejíž pravá strana splňuje požadavky (i) a (ii) z čl. 1 a y_n je libovolné přibližné řešení této rovnice vypočtené metodou (4.57), určené integračním krokem h a počátečními podmínkami $y_s = y_s(h)$, $s = 0, \dots, k-1$, pro něž platí

$$\lim_{h\to 0} y_s(h) = \eta, \quad s = 0, \dots, k-1. \tag{4.69}$$

Abychom mohli formulovat první důležité tvrzení týkající se konvergence obecné lineární k-krokové metody, je třeba zavést ještě další pojem.

Definice 4.4. Řekneme, že metoda (4.57) je *stabilní ve smyslu Dahlquista*[1] nebo stručně *D-stabilní*, platí-li pro všechny kořeny ξ_ν polynomu ϱ definovaného

[1] Podle švédského matematika G. Dahlquista, který je tvůrcem celé teorie, která bude následovat.

rovnicí

$$\varrho(\xi) = \sum_{\nu=0}^{k} \alpha_\nu \xi^\nu \tag{4.70}$$

$|\xi_\nu| \leqq 1$ a jsou-li ty z nich, pro něž je $|\xi_\nu| = 1$, jednoduché.

Poznámka 4.1. Polynom ϱ se nazývá *první charakteristický polynom* dané metody. Zavedeme-li ještě tzv. *druhý charakteristický polynom* σ rovnicí

$$\sigma(\xi) = \sum_{\nu=0}^{k} \beta_\nu \xi^\nu, \tag{4.71}$$

je zřejmé, že zadat metodu (4.57) je ekvivalentní se zadáním polynomů ϱ a σ.

Věta 4.3. *Konvergentní metoda* (4.57) *je D-stabilní.*

D ů k a z . Daná metoda je konvergentní pro každou diferenciální rovnici, jejíž pravá strana splňuje podmínky (i) a (ii), tedy speciálně je konvergentní pro diferenciální rovnici $y' = 0$ s počáteční podmínkou $y(a) = 0$. Je-li tedy y_n libovolné řešení rovnice

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = 0, \tag{4.72}$$

pro něž platí

$$\lim_{h\to 0} y_s = 0, \quad s = 0, \dots, k-1, \tag{4.73}$$

musí platit

$$\lim_{\substack{h\to 0\\ x_n = x}} y_n = 0, \tag{4.74}$$

neboť přesné řešení uvažované diferenciální rovnice je funkce identicky rovná nule. Protože je $x_n = a + nh = x$, můžeme rovnici (4.74) psát ve tvaru

$$\lim_{\substack{n\to\infty\\ h=\frac{x-a}{n}}} y_n = 0. \tag{4.75}$$

Buď nyní $\xi_\nu = re^{i\varphi_\nu}$ kořen polynomu ϱ a předpokládejme, že je $r > 1$. Posloupnost $\{r^n e^{in\varphi_\nu};\ n = 0, 1 \dots\}$ řeší zřejmě rovnici (4.72) a protože tato rovnice nezávisí explicitně na h, řeší ji rovněž posloupnost $\{hr^n e^{in\varphi_\nu}\}$, a tedy také reálné posloupnosti $\{hr^n \cos(n\varphi_\nu)\}$ a $\{hr^n \sin(n\varphi_\nu)\}$, neboť koeficienty dané mnohokrokové metody jsou reálné. Protože pro obě tyto posloupnosti platí (4.73), platí pro ně také (4.75). Je však $|hr^n e^{in\varphi_\nu}| = hr^n$. Protože je $r > 1$ a $h = (x-a)/n$, platí $|hr^n e^{i\varphi_\nu}| \to \infty$ pro $n \to \infty$. Totéž tedy platí aspoň pro jednu z posloupností $\{\mathrm{Re}(hr^n e^{in\varphi_\nu})\}$ a $\{\mathrm{Im}(hr^n e^{in\varphi_\nu})\}$. To je však ve sporu s (4.75). Je tedy $r \leqq 1$.

Buď dále $\xi_\nu = re^{i\varphi_\nu}$ kořen polynomu ϱ o násobnosti větší než jedna. V tomto případě řeší rovnici (4.72) posloupnost $\{nr^n e^{in\varphi_\nu}\}$. Posloupnosti $\{\mathrm{Re}(h^{1/2}nr^n e^{in\varphi_\nu})\}$ a $\{\mathrm{Im}(h^{1/2}nr^n e^{in\varphi_\nu})\}$ jsou tedy také jejími řešeními a platí pro ně (4.73), což implikuje platnost rovnice (4.75). Je však $|h^{1/2}nr^n e^{in\varphi_\nu}| = h^{1/2}nr^n$. Předpokládáme-li tedy, že je $r \geqq 1$, platí $|h^{1/2}nr^n e^{in\varphi_\nu}| \to \infty$ pro $n \to \infty$ a $h = (x-a)/n$. To je však spor dokazující, že platí $r < 1$. Věta je dokázána.

Věta 4.4. *Řád konvergentní metody* (4.57) *je nejméně 1 (tj. konstanty* C_0 *a* C_1 *z definice 4.2. jsou nulové).*

D ů k a z . Dokažme nejprve, že je $C_0 = 0$. Daná metoda je konvergentní speciálně pro diferenciální rovnici $y' = 0$ s počáteční podmínkou $y(a) = 1$, jejíž přesné řešení je funkce $y(x) \equiv 1$ v $\langle a, b\rangle$. Tedy každé řešení rovnice (4.72), pro jehož počáteční podmínky platí $\lim_{h\to 0} y_s = 1$ pro $s = 0, \ldots, k-1$, musí konvergovat k jedničce. Zvolíme-li tedy speciálně $y_s = 1$ pro $s = 0, \ldots, k-1$, musí pro řešení určené těmito počátečními podmínkami platit

$$\lim_{\substack{n\to\infty \\ h=(x-a)/n}} y_n = 1. \tag{4.76}$$

Toto speciální řešení však nezávisí explicitně na h, tedy pro ně platí

$$\lim_{n\to\infty} y_n = 1. \tag{4.77}$$

Přejdeme-li tedy v rovnici (4.72) k limitě pro $n \to \infty$, dostaneme

$$\sum_{\nu=0}^{k} \alpha_\nu = 1, \tag{4.78}$$

což není nic jiného než, že $C_0 = 0$.

Dokažme nyní, že je $C_1 = 0$. K tomu cíli uvažme diferenciální rovnici $y' = 1$ s počáteční podmínkou $y(a) = 0$, jejíž přesné řešení je funkce $y(x) = x - a$. Pro každé řešení y_n rovnice

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = h \sum_{\nu=0}^{k} \beta_\nu, \tag{4.79}$$

pro něž platí (4.73), musí tedy platit

$$\lim_{\substack{h\to 0 \\ x_n = x}} y_n = x - a, \quad x \in \langle a, b\rangle. \tag{4.80}$$

Položme

$$y_n = \frac{\sigma(1)}{\varrho'(1)} nh, \quad n = 0, \ldots, N\ (= [(b-a)/h]), \tag{4.81}$$

kde ϱ a σ jsou první a druhý charakteristický polynom dané metody. (Jde tedy o polynomy definované rovnicí (4.70) a (4.71).) Vzorec (4.81) má smysl, neboť

podle první části důkazu je $C_0 = \varrho(1) = 0$; jednička je tedy kořenem polynomu ϱ a podle věty 4.3 je tento kořen jednoduchý, takže je $\varrho'(1) \neq 0$. Funkce y_n definovaná rovnicí (4.81) splňuje zřejmě požadavky (4.73). Rovnice (4.79) je rovněž splněna, neboť je

$$\begin{aligned} \sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} - h\sum_{\nu=0}^{k} \beta_\nu &= h\frac{\sigma(1)}{\varrho'(1)} \sum_{\nu=0}^{k} \alpha_\nu (n+\nu) - h\sigma(1) = \\ &= h\frac{\sigma(1)}{\varrho'(1)} [n\varrho(1) + \varrho'(1)] - h\sigma(1) = 0. \end{aligned} \tag{4.82}$$

Podle (4.80) tedy dostáváme, že pro každé $x \in \langle a, b\rangle$ platí

$$x - a = \lim_{\substack{h\to 0 \\ x_n = x}} y_n = \frac{\sigma(1)}{\varrho'(1)} \lim_{\substack{h\to 0 \\ x_n = x}} nh = \frac{\sigma(1)}{\varrho'(1)}(x-a). \tag{4.83}$$

Odtud však plyne, že je $\sigma(1) = \varrho'(1)$, což je jen jiný zápis podmínky $C_1 = 0$. Věta je dokázána.

Poznámka 4.2. Podmínky z věty 4.4 můžeme pomocí polynomů ϱ a σ psát alternativně také ve tvaru

$$\varrho(1) = 0, \quad \varrho'(1) = \sigma(1). \tag{4.84}$$

Protože tyto podmínky vyjadřují tu skutečnost, že lokální chyba příslušné metody se chová jako $O(h^2)$, říkáme jim analogicky jako v případě jednokrokové metody podmínky *konzistence.*

Z důkazů vět 4.3 a 4.4 je vidět, že jde o velmi elementární tvrzení, která vlastně neříkají nic víc než jen to, že přibližná řešení konvergují k přesným řešením v případě diferenciálních rovnic, jejichž řešení jsou postupně nula (D-stabilita), konstanta ($C_0 = 0$) a polynom prvního stupně ($C_1 = 0$). Proto čtenáře možná překvapí, že jednoduché podmínky, které jsou obsahem těchto vět, jsou nejen nutné, ale i postačující podmínky pro konvergenci.

4.2.2 Postačující podmínky konvergence

Hlavním cílem tohoto odstavce je dokázat, že D-stabilita a konzistence jsou postačujícími podmínkami konvergence. Dále pak zde sestrojíme apriorní odhad chyby.

Začneme několika pomocnými tvrzeními.

Lemma 4.1. *Nechť* φ_n *a* ψ_n *jsou posloupnosti reálných čísel definované pro* $n = 0, \ldots, N$ *a nechť* m *je nezáporná konstanta. Nechť dále platí*

$$\varphi_n \leqq \psi_n + m\sum_{\nu=0}^{n-1} \varphi_\nu, \quad n = 0, \ldots, N.^{1)} \tag{4.85}$$

[1] Součet, jehož horní mez sčítání je menší než dolní, pokládáme za rovný nule.

Pak platí

$$\varphi_n \leqq \psi_n + m\sum_{\nu=0}^{n-1}\psi_\nu(1+m)^{n-1-\nu}, \quad n=0,\dots,N. \tag{4.86}$$

D ů k a z . Položme

$$R_n = m\sum_{\nu=0}^{n-1}\varphi_\nu, \quad n=0,\dots,N. \tag{4.87}$$

Podle (4.85) platí $R_{n+1}-R_n = m\varphi_n \leqq m\psi_n + mR_n$, neboli

$$R_{n+1} \leqq (1+m)R_n + m\psi_n, \quad m=0,\dots,N-1. \tag{4.88}$$

Definujeme-li nyní funkci Z_n, $n=0,\dots,N$, rekurencí

$$\begin{aligned} Z_0 &= 0, \\ Z_{n+1} &= (1+m)Z_n + m\psi_n, \quad n=0,\dots,N-1, \end{aligned} \tag{4.89}$$

je

$$Z_n = m\sum_{\nu=0}^{n-1}\psi_\nu(1+m)^{n-1-\nu}, \quad n=0,\dots,N, \tag{4.90}$$

jak se snadno přesvědčíme např. úplnou indukcí. Na druhé straně, položíme-li $V_n = Z_n - R_n$, je zřejmě $V_0 = 0$ a $V_{n+1} \geqq (1+m)V_n$ pro $n=0,\dots,N-1$. Odtud však plyne, že je $V_n \geqq 0$ pro $n=0,\dots,N$, tj. $Z_n \geqq R_n$. Z této nerovnosti, z rovnic (4.90), (4.87) a z nerovnosti (4.85) už tvrzení snadno plyne.

Lemma 4.2. *Buď dána D-stabilní mnohokroková metoda* (4.57) *a buď A matice řádu k definovaná rovnicí*

$$A = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ & & & & 0 \\ 0 & \dots & & 0 & 1 \\ -\frac{\alpha_0}{\alpha_k} & & \dots & & -\frac{\alpha_{k-1}}{\alpha_k} \end{bmatrix} \tag{4.91}$$

Pak existuje konstanta Γ *taková, že platí*

$$\|A^n\| \leqq \Gamma, \quad n=0,1,\dots. \tag{4.92}$$

Symbolem $\|.\|$ přitom rozumíme $\mathscr{L}_\infty$-(Čebyševovu) normu matice, tj. maticovou normu indukovanou $\mathscr{L}_\infty$-normou vektoru (tj. normou vektoru, která je dána jeho složkou, která má maximální absolutní hodnotu).

D ů k a z . Charakteristický polynom matice A je zřejmě polynom ϱ. Buď T (regulární) matice taková, že matice $J = T^{-1}AT$ je v Jordanově kanonickém tvaru.

Protože je zřejmě $A^n = TJ^nT^{-1}$, stačí dokázat omezenost n-tých mocnin matice J. Protože však J je blokově diagonální matice, stačí se dále omezit pouze na vyšetřování mocnin jednotlivých bloků. Vzhledem k podmínce stability jsou všechna vlastní čísla matice A v absolutní hodnotě menší nebo rovna jedné a ta z nich, která jsou v absolutní hodnotě rovna jedné, jsou jednoduchá. Jordanovy bloky příslušné vlastním číslům rovným v absolutní hodnotě jedné jsou tedy matice řádu jedna, tj. přímo tato vlastní čísla. Jejich mocniny jsou tedy zřejmě omezené. Bloky J_i řádu r_i většího než jedna nutně přísluší vlastním číslům λ, pro něž je $|\lambda| < 1$. Kromě toho, úplnou indukcí se snadno zjistí, že pro ně platí

$$J_i^n = \begin{bmatrix} \lambda^n & \binom{n}{1}\lambda^{n-1} & \dots & \dots & \binom{n}{r_i-1}\lambda^{n-r_i+1} \\ 0 & \ddots & & & \vdots \\ \vdots & \ddots & & & \\ & & & & \binom{n}{1}\lambda^{n-1} \\ 0 & \dots & \dots & 0 & \lambda^n \end{bmatrix} \tag{4.93}$$

Nyní zbývá už pouze uvědomit si, že $\mathscr{L}_\infty$-norma matice je maximum ze součtů absolutních hodnot prvků v jednotlivých řádcích. Utvoříme-li totiž tyto řádkové součty, jsou všechny sčítance typu polynom v n krát $|\lambda|^{n-\alpha}$, kde α je omezené. V důsledku toho, že je $|\lambda| < 1$, konverguje každý sčítanec pro $n \to \infty$ k nule. Odtud však už lemma bezprostředně plyne.

Poznámka 4.3. Lemma 4.2 platí zřejmě nejen pro $\mathscr{L}_\infty$-normu, ale pro libovolnou maticovou normu.

Lemma 4.3. *Nechť* z_n *splňuje rovnici*

$$\sum_{\nu=0}^{k}\alpha_\nu z_{n+\nu} = h\sum_{\nu=0}^{k}\beta_{n\nu}z_{n+\nu} + \lambda_n \tag{4.94}$$

pro $n=0,1,\dots$, *kde* α_ν *jsou koeficienty D-stabilní k-krokové metody a* $\beta_{n\nu}$, $\nu = 0,\dots,k$, $n=0,1,\dots$ *jsou libovolná čísla. Nechť dále* Z, Λ, B *a* β *jsou takové konstanty, že platí*

$$|z_\mu| \leqq Z, \quad \mu=0,\dots,k-1, \tag{4.95}$$

$$|\lambda_n| \leqq \Lambda, \quad n=0,1,\dots, \tag{4.96}$$

$$\sum_{\nu=0}^{k-1}\left|\beta_{n\nu} - \frac{\beta_{nk}}{\alpha_k}\alpha_\nu\right| \leqq B, \quad n=0,1,\dots \tag{4.97}$$

a

$$|\beta_{nk}| \leqq \beta, \quad n=0,1,\dots. \tag{4.98}$$

Nechť je konečně h_0 číslo, pro něž je

$$h_0 < \frac{|\alpha_k|}{\beta}. \tag{4.99}$$

Pak pro libovolné $h \leqq h_0$ a libovolné $n = 0, 1, \ldots$ platí

$$|z_n| \leqq \Gamma Z \mathrm{e}^{L^* n h} + \Gamma^* \Lambda n \mathrm{e}^{L^* n h}, \tag{4.100}$$

kde

$$\Gamma^* = \frac{\Gamma}{|\alpha_k|\left(1 - \frac{\beta}{|\alpha_k|} h_0\right)}, \tag{4.101}$$

$$L^* = \Gamma^* B \tag{4.102}$$

a Γ *je konstanta z lemmatu 4.2.*

Důkaz. Přičtěme k levé i pravé straně rovnice (4.94) číslo $-\sum_{\nu=0}^{k} h \frac{\beta_{nk}}{\alpha_k} \alpha_\nu z_{n+\nu}$. Po jednoduchých úpravách dostaneme

$$\sum_{\nu=0}^{k} \alpha_\nu \left(1 - \frac{h\beta_{nk}}{\alpha_k}\right) z_{n+\nu} = h \sum_{\nu=0}^{k-1} \left(\beta_{n\nu} - \frac{\beta_{nk}}{\alpha_k} \alpha_\nu\right) z_{n+\nu} + \lambda_n. \tag{4.103}$$

Odtud plyne, že pro $h \leqq h_0$ je

$$\sum_{\nu=0}^{k} \alpha_\nu z_{n+\nu} = q_n, \tag{4.104}$$

kde

$$q_n = \frac{h}{\left(1 - h\frac{\beta_{nk}}{\alpha_k}\right)} \sum_{\nu=0}^{k-1} \left(\beta_{n\nu} - \frac{\beta_{nk}}{\alpha_k} \alpha_\nu\right) z_{n+\nu} + \frac{1}{\left(1 - h\frac{\beta_{nk}}{\alpha_k}\right)} \lambda_n, \tag{4.105}$$

neboť pro $h \leqq h_0$ je výraz $1 - h\beta_{nk}/\alpha_k$ vzhledem k nerovnostem (4.98) a (4.99) kladný, a můžeme jím tedy dělit. Buď nyní z_n k-dimenzionální vektor o složkách $z_n, z_{n+1}, \ldots, z_{n+k-1}$ a q_n k-dimenzionální vektor o složkách $0, 0, \ldots, q_n/\alpha_k$. Pomocí těchto vektorů a matice A z lemmatu 4.2 můžeme rovnici (4.104) zapsat maticově takto:

$$z_{n+1} = A z_n + q_n, \tag{4.106}$$

neboť prvních $k-1$ rovnic v (4.106) jsou identity a poslední rovnice je právě rovnice (4.104). Rovnice (4.106) je však ekvivalentní s rovnicí

$$z_n = A^n z_0 + \sum_{\nu=0}^{n-1} A^{n-1-\nu} q_\nu, \tag{4.107}$$

jak se snadno zjistí úplnou indukcí. Podle lemmatu 4.2 je tedy

$$\|z_n\| \leqq \Gamma \|z_0\| + \Gamma \sum_{\nu=0}^{n-1} \|q_\nu\|. \tag{4.108}$$

Podle (4.105) je

$$\|q_\nu\| = \frac{1}{|\alpha_k|} |q_\nu| = \tag{4.109}$$

$$= \frac{1}{|\alpha_k|\left(1 - h\frac{\beta_{nk}}{\alpha_k}\right)} \left| h \sum_{s=0}^{k-1} \left(\beta_{\nu s} - \frac{\beta_{\nu k}}{\alpha_k} \alpha_s\right) z_{\nu+s} + \lambda_\nu \right|.$$

Protože je $|z_{\nu+s}| \leqq \|z_\nu\|$ pro $s = 0, \ldots, k-1$, dostáváme odtud, z (4.108) a z (4.95) až (4.98), že pro $h \leqq h_0$ platí

$$\|z_n\| \leqq \Gamma Z + \sum_{\nu=0}^{n-1} h L^* \|z_\nu\| + \Gamma^* \Lambda n. \tag{4.110}$$

Položíme-li v lemmatu 4.1 $\varphi_n = \|z_n\|$, $\psi_n = \Gamma Z + \Gamma^* \Lambda n$ a $m = L^* h$, dostaneme z nerovnosti (4.110) nerovnost

$$\|z_n\| \leqq \Gamma Z + \Gamma^* \Lambda n + h L^* \sum_{\nu=0}^{n-1} (\Gamma Z + \Gamma^* \Lambda \nu)(1 + h L^*)^{n-1-\nu} \tag{4.111}$$

Upravme vzniklý výraz tak, že provedeme naznačené součty. Zřejmě platí

$$\sum_{\nu=0}^{n-1} (1 + h L^*)^{n-1-\nu} = \frac{(1 + h L^*)^n - 1}{h L^*} \tag{4.112}$$

a

$$\sum_{\nu=0}^{n-1} \nu (1 + h L^*)^{n-1-\nu} = \frac{(1 + h L^*)^n - 1}{(h L^*)^2} - \frac{n}{h L^*}, \tag{4.113}$$

jak se snadno zjistí např. úplnou indukcí. Dosadíme-li do nerovnosti (4.111) ze vzorců (4.112) a (4.113), dostaneme po jednoduchých úpravách

$$\|z_n\| \leqq \Gamma Z (1 + h L^*)^n + \Gamma^* \Lambda \frac{(1 + h L^*)^n - 1}{h L^*}. \tag{4.114}$$

Požadované tvrzení pak už plyne odtud, z nerovnosti $|z_n| \leqq \|z_n\|$ a z takřka zřejmých nerovností

$$(1 + h L^*)^n < \mathrm{e}^{n h L^*}, \quad \frac{(1 + h L^*)^n - 1}{h L^*} < n \mathrm{e}^{n h L^*} \tag{4.115}$$

Právě zformulované a dokázané lemma hraje v teorii mnohokrokových metod stejnou roli jako hrálo lemma 1.1 u jednokrokových metod. Konvergence i další důležitá tvrzení týkající se mnohokrokových metod se pomocí něj už snadno dokáží.

Věta 4.5. *D-stabilní a konzistentní lineární k-kroková metoda je konvergentní.*

D ů k a z . Buď y řešení diferenciální rovnice (1.6), jejíž pravá strana splňuje předpoklady (i) a (ii) z čl. 1, s počáteční podmínkou (1.7) a buď y_n řešení rovnice (4.57) s počátečními podmínkami $y_s = y_s(h)$, $s = 0, \dots, k-1$. Položme

$$\delta(h) = \max_{s=0,\dots,k-1} |y_s(h) - y(x_s)| \tag{4.116}$$

a předpokládejme, že platí

$$\lim_{h\to 0} \delta(h) = 0. \tag{4.117}$$

Vzhledem k definici 4.1 máme dokázat, že platí (4.68). Položme jako dříve $e_n = y_n - y(x_n)$ a odhadujme tuto veličinu. Přesné řešení dané diferenciální rovnice má v intervalu $\langle a, b\rangle$ spojitou derivaci, a je tedy (viz definice 4.1)

$$\sum_{\nu=0}^{k} \alpha_\nu y(x_{n+\nu}) - h\sum_{\nu=0}^{k} \beta_\nu f\big(x_{n+\nu}, y(x_{n+\nu})\big) = L\big(y(x_n); h\big). \tag{4.118}$$

Odečteme-li tuto rovnici do rovnice (4.57), dostaneme

$$\sum_{\nu=0}^{k} \alpha_\nu e_{n+\nu} = \tag{4.119}$$
$$= h\sum_{\nu=0}^{k} \beta_\nu \big[f(x_{n+\nu}, y_{n+\nu}) - f\big(x_{n+\nu}, y(x_{n+\nu})\big)\big] - L\big(y(x_n); h\big).$$

Položíme-li ještě

$$g_n = \frac{f(x_n, y_n) - f\big(x_n, y(x_n)\big)}{e_n}, \quad e_n \neq 0, \tag{4.120}$$
$$g_n = 0, \quad e_n = 0,$$

můžeme rovnici (4.119) psát ve tvaru

$$\sum_{\nu=0}^{k} \alpha_\nu e_{n+\nu} = h\sum_{\nu=0}^{k} \beta_\nu g_{n+\nu} e_{n+\nu} - L\big(y(x_n); h\big), \tag{4.121}$$

což je rovnice typu (4.94). K odhadu jejího řešení použijeme tedy lemmatu 4.3.

Protože je zřejmě $|g_n| \leqq L$, jsou předpoklady (4.95), (4.97) a (4.98) splněny, klademe-li

$$Z = \delta(h), \quad B = L\sum_{\nu=0}^{k-1} \Big(|\beta_\nu| + \frac{|\beta_k|}{|\alpha_k|}|\alpha_\nu|\Big), \quad \beta = L|\beta_k|. \tag{4.122}$$

(L je lipschitzovská konstanta pravé strany dané diferenciální rovnice.) Protože podmínky stability jsou také splněny, zbývá odhadnout lokální chybu $L\big(y(x_n); h\big)$. K tomu použijeme stejně jako v případě jednokrokové metody modulu spojitosti ω

derivace přesného řešení. Podle definice modulu spojitosti je $|y'(x_{n+\nu}) - y'(x_n)| \leqq \leqq \omega(\nu h)$, a tedy existuje číslo $\theta_{n\nu}$ takové, že je $|\theta_{n\nu}| \leqq 1$ a že platí

$$y'(x_{n+\nu}) = y'(x_n) + \theta_{n\nu}\omega(\nu h). \tag{4.123}$$

Dále, podle věty o střední hodnotě existuje číslo $\xi_{n\nu}$ takové, že je $x_n < \xi_{n\nu} < x_{n+\nu}$ a že platí $y(x_{n+\nu}) = y(x_n) + \nu h y'(\xi_{n\nu})$. Existuje tedy číslo $\theta'_{n\nu}$ takové, že je $|\theta'_{n\nu}| \leqq 1$ a že platí

$$y(x_{n+\nu}) = y(x_n) + \nu h[y'(x_n) + \theta'_{n\nu}\omega(\nu h)]. \tag{4.124}$$

Dosadíme-li do výrazu (4.118) podle (4.123) a (4.124), máme

$$L\big(y(x_n); h\big) = y(x_n)\sum_{\nu=0}^{k} \alpha_\nu + hy'(x_n)\Big(\sum_{\nu=1}^{k} \nu\alpha_\nu - \sum_{\nu=0}^{k} \beta_\nu\Big) + \tag{4.125}$$
$$+ h\sum_{\nu=1}^{k} \nu\alpha_\nu\theta'_{n\nu}\omega(\nu h) - h\sum_{\nu=0}^{k} \beta_\nu\theta_{n\nu}\omega(\nu h).$$

První dva sčítance na pravé straně rovnice (4.125) se rovnají nule v důsledku konzistence; platí tedy (je $0 \leqq \omega(\nu h) \leqq \omega(kh)$, $\nu = 1, 2, \dots, k$)

$$\big|L\big(y(x_n); h\big)\big| \leqq Kh\omega(kh), \tag{4.126}$$

kde

$$K = \sum_{\nu=0}^{k} \big(\nu|\alpha_\nu| + |\beta_\nu|\big). \tag{4.127}$$

Podle lemmatu 4.3 platí pro $h < |\alpha_k|/(L|\beta_k|)$ a pro n takové, že je $x_n \in \langle a, b\rangle$

$$|e_n| \leqq [\Gamma\delta(h) + \Gamma^* K(x_n - a)\omega(kh)]\mathrm{e}^{(x_n - a)L^*}, \tag{4.128}$$

kde

$$\Gamma^* = \frac{\Gamma}{|\alpha_k|\big(1 - h\big|\frac{\beta_k}{\alpha_k}\big|L\big)} \tag{4.129}$$

a $L^* = \Gamma^* B$. Z nerovnosti (4.128) však už tvrzení věty ihned plyne. Věta 4.5 je dokázána.

Ta část vzorce (4.128), která obsahuje funkci ω, je zcela analogická obdobnému vzorci, který jsme odvodili pro Eulerovu nebo pro obecnou jednokrokovou metodu. Kromě toho však tento vzorec obsahuje navíc sčítanec, který charakterizuje chybu, které se dopouštíme nepřesným zadáním počátečních podmínek. Přítomnost tohoto členu je pro mnohokrokové metody typická, a rychlost konvergence je tedy u těchto metod dána nejen tím, jak rychle konverguje k nule lokální chyba, ale také tím, jak dobře jsou aproximovány počáteční podmínky.

Z důkazu věty 4.5 je zřejmé, že lepší odhad, než je vzorec (4.128), dostaneme, odhadneme-li lépe lokální chybu. Za předpokladu, že hledané řešení je dostatečně hladké, se to provede např. pro Adamsovu-Bashforthovu nebo pro Adamsovu-Moultonovu metodu snadno. Pro obě tyto metody totiž platí

$$L(y(x);h) = C_{p+1}y^{(p+1)}(\xi)h^{p+1} \tag{4.130}$$

pro vhodné $\xi \in \langle a,b\rangle$, jak plyne ihned ze vzorců (4.32) a (4.38). Má-li tedy hledané řešení spojitou $(p+1)$-ní derivaci, je lokální chyba Adamsovy-Bashforthovy a Adamsovy-Moultonovy metody řádu h^{p+1}. Vztah analogický k (4.130) sice obecně pro lineární k-krokovou metodu neplatí, platí však následující tvrzení.

Lemma 4.4. *Buď dána lineární k-kroková metoda řádu $p \geqq 1$ a buď y funkce mající $p+1$ spojitých derivací v intervalu $\langle a,b\rangle$. Pak platí*

$$|L(y(x);h)| \leqq GYh^{p+1}, \tag{4.131}$$

kde

$$Y = \max_{x\in\langle a,b\rangle} |y^{(p+1)}(x)|, \tag{4.132}$$

$$G = \int_0^k |g(s)|\,\mathrm{d}\,s, \tag{4.133}$$

funkce g je dána rovnicí

$$g(s) = \sum_{\nu=0}^{k}\left[\frac{\alpha\nu}{p!}(\nu-s)_+^p - \frac{\beta\nu}{(p-1)!}(\nu-s)_+^{p-1}\right] \tag{4.134}$$

a symbol z_+ je definován jako z, je-li $z \geqq 0$, a jako 0, je-li $z < 0$.

D ů k a z . Spokojíme-li se pouze s existencí konstanty G (nezávislé na h) takové, že platí (4.131), stačí k důkazu této nerovnosti elementární užití Taylorova vzorce. Chceme-li dostat vyjádření (4.133) pro G, je vhodnější užít Taylorův vzorec s integrálním tvarem zbytku. Proto jej nejprve připomeňme. Má-li funkce y spojité derivace až do řádu $q+1$ v $\langle a,b\rangle$ a jsou-li x a $\tilde{x}$ dva body z tohoto intervalu, platí

$$y(\tilde{x}) = y(x) + (\tilde{x}-x)y'(x) + \ldots + \frac{1}{q!}(\tilde{x}-x)^q y^{(q)}(x) + \tag{4.135}$$
$$+ \frac{1}{q!}\int_x^{\tilde{x}} (\tilde{x}-t)^q y^{(q+1)}(t)\,\mathrm{d}\,t.$$

Užijeme-li tento vzorec pro funkce y a y' a pro $\tilde{x} = x+\nu h$, máme

$$y(x+\nu h) = y(x) + \nu h y'(x) + \ldots + \frac{1}{p!}\nu^p h^p y^{(p)}(x) + \tag{4.136}$$
$$+ \frac{h^{p+1}}{p!}\int_0^{\nu} (\nu-s)^p y^{(p+1)}(x+hs)\,\mathrm{d}\,s$$

a

$$y'(x+\nu h) = y'(x) + \nu h y''(x) + \ldots + \frac{1}{(p-1)!}\nu^{p-1}h^{p-1}y^{(p)}(x) + \tag{4.137}$$
$$+ \frac{h^p}{(p-1)!}\int_0^{\nu} (\nu-s)^{p-1}y^{(p+1)}(x+hs)\,\mathrm{d}\,s.$$

Vynásobíme-li rovnici (4.136) číslem α_ν, rovnici (4.137) číslem $-h\beta_\nu$ a sečteme pro $\nu = 0$ až k, dostaneme (srov. vzorce (4.64))

$$L(y(x);h) = C_0y(x) + C_1y'(x)h + \ldots + C_py^{(p)}(x)h^p + \tag{4.138}$$
$$+ h^{p+1}\sum_{\nu=0}^{k}\int_0^{\nu}\left[\frac{1}{p!}\alpha_\nu(\nu-s)^p - \frac{1}{(p-1)!}\beta_\nu(\nu-s)^{p-1}\right]y^{(p+1)}(x+sh)\,\mathrm{d}\,s.$$

Metoda je řádu p; proto je $C_0 = C_1 = \ldots = C_p = 0$. Dále je

$$\int_0^{\nu}\left[\frac{1}{p!}\alpha_\nu(\nu-s)^p - \frac{1}{(p-1)!}\beta_\nu(\nu-s)^{p-1}\right]y^{(p+1)}(x+sh)\,\mathrm{d}\,s = \tag{4.139}$$
$$= \int_0^{\nu}\left[\frac{1}{p!}\alpha_\nu(\nu-s)_+^p - \frac{1}{(p-1)!}\beta_\nu(\nu-s)^{p-1}\right]y^{(p+1)}(x+sh)\,\mathrm{d}\,s =$$
$$= \int_0^{k}\left[\frac{1}{p!}\alpha_\nu(\nu-s)_+^p - \frac{1}{(p-1)!}\beta_\nu(\nu-s)_+^{p-1}\right]y^{(p+1)}(x+sh)\,\mathrm{d}\,s,$$

neboť je $\nu - s \geqq 0$ pro $0 \leqq s \leqq \nu$ a $\nu - s < 0$ pro $s > \nu$. Dosazením do (4.138) je tedy

$$L(y(x);h) = h^{p+1}\int_0^k g(s)y^{(p+1)}(x+sh)\,\mathrm{d}\,s. \tag{4.140}$$

Odtud však už lemma ihned plyne.

Poznámka 4.4. Nemění-li funkce g v intervalu $\langle 0,k\rangle$ znaménko, platí pro lokální chybu dokonce

$$L(y(x);h) = C_{p+1}h^{p+1}y^{(p+1)}(\xi) \tag{4.141}$$

a

$$C_{p+1} = \int_0^k g(s)\,\mathrm{d}\,s, \tag{4.142}$$

jak plyne ihned z první věty o střední hodnotě integrálního počtu.

Z lemmatu 4.4 a z důkazu věty 4.5 už snadno plyne, že platí následující věta.

Věta 4.6. *Nechť je dána D-stabilní mnohokroková metoda řádu $p \geqq 1$. Nechť jsou splněny předpoklady* (i) a (ii) *z čl. 1 a nechť řešení dané diferenciální rovnice má $p+1$ spojitých derivací v intervalu $\langle a,b\rangle$. Pak pro $h < |\alpha_k|/(L|\beta_k|)$ platí*

$$|y_n - y(x_n)| \leqq [\Gamma\delta(h) + \Gamma^*(x_n-a)GYh^p]\mathrm{e}^{(x_n-a)L^*} \tag{4.143}$$

pro každé n takové, že je $x_n \in \langle a, b\rangle$. Funkce δ je přitom definována vzorcem (4.116) *a význam konstant* Γ, Γ^*, L^*, G a Y *je stejný jako v lemmatech 4.3 a 4.4.*

Znovu připomeňme, že k tomu, aby chyba, která vznikne řešením diferenciální rovnice (1.6) mnohokrokovou metodou řádu p, byla řádu h^p, nestačí jen dostatečná hladkost řešení, jako tomu bylo u obecné jednokrokové metody. Navíc musí být i k počátečních podmínek, jimiž je přibližné řešení určeno, zadáno s přesností h^p.

4.2.3 Asymptotický odhad chyby

Vyšetřování asymptotického chování chyby je v případě mnohokrokové metody podstatně komplikovanější než v případě jednokrokové metody. Dalo se to ostatně očekávat, neboť u mnohokrokové metody závisí přibližné řešení nejen na integračním kroku, ale i na počátečních podmínkách. Abychom získali alespoň orientační výsledky, omezíme se na jednoduchý případ diferenciální rovnice

$$y' = Ay \tag{4.144}$$

s počáteční podmínkou $y(0) = 1$, kde A je nenulová konstanta. Přesné řešení tohoto problému je tedy funkce $y(x) = e^{Ax}$. Buď dále dána D-stabilní k-kroková metoda (4.57) řádu $p \geqq 1$ a předpokládejme navíc, že příslušné charakteristické polynomy ϱ a σ nemají netriviální společné dělitele. Tento předpoklad není příliš omezující a ve všech uvedených konkrétních příkladech mnohokrokových metod je splněn. Aplikujeme-li danou metodu na rovnici (4.144), dostaneme pro přibližné řešení y_n rovnici

$$\sum_{\nu=0}^{k} (\alpha_\nu - hA\beta_\nu) y_{n+\nu} = 0. \tag{4.145}$$

Rovnice (4.145) je příkladem lineární diferenční rovnice k-tého řádu s konstantními koeficienty. Z teorie těchto rovnic je známo, že položíme-li

$$\tilde{\varrho}(\xi) = \sum_{\nu=0}^{k} (\alpha_\nu - hA\beta_\nu)\xi^\nu = \varrho(\xi) - hA\sigma(\xi), \tag{4.146}$$

označíme-li $\tilde{\xi}_1, \ldots, \tilde{\xi}_k$ kořeny polynomu $\tilde{\varrho}$ a předpokládáme-li, že jsou navzájem různé, dá se každé řešení rovnice (4.145) psát ve tvaru

$$y_n = \sum_{s=1}^{k} A_s \tilde{\xi}_s^n, \tag{4.147}$$

kde A_s jsou vhodné konstanty.

Pokud čtenář není obeznámen s teorií diferenčních rovnic, dokáže si platnost uvedeného tvrzení snadno takto: Předně je hned vidět, že každá funkce tvaru (4.147) řeší rovnici (4.145). Dále, protože každé řešení rovnice (4.145) je jednoznačně určeno

svými počátečními podmínkami, stačí k důkazu požadovaného tvrzení dokázat, že soustava lineárních algebraických rovnic

$$\sum_{s=1}^{k} A_s \tilde{\xi}_s^r = y_r, \quad r = 0, \ldots, k-1 \tag{4.148}$$

pro neznámé $A_1, \ldots, A_k$ má při libovolných $y_0, \ldots, y_{k-1}$ řešení. Determinant této soustavy je však Vandermondův determinant pro čísla $\tilde{\xi}_1, \ldots, \tilde{\xi}_k$ a protože tato čísla jsou navzájem různá, je tento determinant různý od nuly.

Abychom tedy vyšetřili chování řešení rovnice (4.145), musíme začít zkoumáním vlastností kořenů polynomu $\tilde{\varrho}$. To provedeme v následujícím lemmatu.

Lemma 4.5. *Nechť polynomy ϱ a σ jsou nesoudělné a nechť ξ_s je kořen polynomu ϱ násobnosti r. Pak existuje funkce Φ komplexní proměnné t, která je holomorfní v okolí bodu $t = 0$ a taková, že položíme-li $\tilde{\xi}_s = \Phi(h^{1/r})$, kde $h^{1/r}$ je jedna z r (navzájem různých) hodnot r-té odmocniny z h, je toto $\tilde{\xi}_s$ pro dostatečně malá h kořen polynomu $\tilde{\varrho}$.*

D ů k a z . Předpokládejme nejprve, že ξ_s je jednoduchý kořen polynomu ϱ. Protože polynomy ϱ a σ nemají netriviálního společného dělitele, má funkce

$$h = h(\xi) = \frac{\varrho(\xi)}{A\sigma(\xi)} \tag{4.149}$$

v nějakém okolí bodu $\xi = \xi_s$ smysl a pro dané ξ definuje takové h, že toto ξ je kořenem polynomu $\tilde{\varrho}(\xi) = \varrho(\xi) - hA\sigma(\xi)$. Funkce h je zřejmě holomorfní v okolí bodu $\xi = \xi_s$ a platí pro ni $h(\xi_s) = 0$, $h'(\xi_s) = \varrho'(\xi_s)/(A\sigma(\xi_s)) \neq 0$, neboť ξ_s je jednoduchý kořen polynomu ϱ, a je tedy $\varrho'(\xi_s) \neq 0$. Podle věty o inverzi holomorfní funkce existuje funkce $\xi = \Phi(h)$, která je holomorfní v okolí počátku a která je inverzní k funkci $h = h(\xi)$. Tato funkce udává tedy při daném h takové ξ, které řeší rovnici $\tilde{\varrho}(\xi) = 0$.

Buď nyní ξ_s kořen polynomu ϱ násobnosti $r > 1$. Za tohoto předpokladu platí, že je $h(\xi) = (\xi - \xi_s)^r \varphi(\xi)$ a φ je holomorfní a různá od nuly v bodě $\xi = \xi_s$. Existuje tedy funkce ψ, která je holomorfní v okolí bodu $\xi = \xi_s$ a pro niž platí $\psi^r(\xi) = \varphi(\xi)$ (holomorfní větev r-té odmocniny z funkce φ). Definujme funkci $t = t(\xi)$ předpisem $t(\xi) = (\xi - \xi_s)\psi(\xi)$. Tato funkce je holomorfní v okolí bodu $\xi = \xi_s$ a platí pro ni $t(\xi_s) = 0$, $t'(\xi_s) = \psi(\xi_s) \neq 0$. Existuje tedy funkce $\xi = \Phi(t)$, která je holomorfní v okolí počátku a která je inverzní k funkci $t = t(\xi)$ (viz např. Černý (1967)). K dokončení důkazu už stačí uvědomit si, že funkce $\xi = \Phi(t)$ udává tu hodnotu ξ, která je při daném t kořenem rovnice $\varrho(\xi) = t^r A\sigma(\xi) = 0$. Důkaz je hotov.

Z uvedeného lemmatu plyne, že pro dostatečně malé h leží v okolí každého jednoduchého kořene polynomu ϱ právě jeden kořen polynomu $\tilde{\varrho}$ a v okolí každého kořene polynomu ϱ násobnosti r právě r kořenů polynomu $\tilde{\varrho}$. Tyto kořeny polynomu $\tilde{\varrho}$, na něž se rozštěpí r-násobný kořen polynomu ϱ, jsou navzájem různé, neboť vzniknou

dosazením r-různých hodnot r-té odmocniny z h do holomorfní funkce, která je prostá. Kořeny polynomu $\tilde{\varrho}$ jsou tedy navzájem různé a řešení rovnice (4.145) se dá psát pro dostatečně malé h ve tvaru (4.147).

Buďte $\xi_1, \dots, \xi_k$ kořeny polynomu ϱ (každý z nich píšeme tolikrát, kolik činí jeho násobnost) a očíslujme je speciálně. Protože daná metoda je stabilní, platí pro každé s, že $|\xi_s| \leqq 1$. Má-li polynom ϱ m kořenů v absolutní hodnotě rovných jedné, přidělme jim indexy $1, \dots, m$ a speciálně položme $\xi_1 = 1$ (takový kořen v důsledku toho, že je $p \geqq 1$ zaručeně existuje). Tyto kořeny nazveme *podstatné*. V důsledku D-stability o nich víme, že jsou všechny jednoduché. Zbývající kořeny (pokud existují) označíme $\xi_{m+1}, \dots, \xi_k$ a nazveme je *nepodstatné*. Kořeny $\tilde{\xi}_s$ polynomu $\tilde{\varrho}$ očíslujeme tak, že pro dostatečně malé h leží kořen $\tilde{\xi}_s$ v okolí kořene ξ_s, $s = 1, \dots, k$.

Je-li $\tilde{\xi}_s$ kořen polynomu $\tilde{\varrho}$, který se pro $h \to 0$ blíží k nepodstatnému kořenu polynomu ϱ (tj. je-li $s = m+1, \dots, k$), existuje konstanta $t < 1$ taková, že platí

$$|\tilde{\xi}_s| \leqq t \tag{4.150}$$

pro $s = m+1, \dots, k$ a pro dostatečně malé h. To je zřejmé, neboť nepodstatné kořeny polynomu ϱ leží uvnitř jednotkového kruhu a za číslo t lze tedy vzít např.

$$t = \tfrac{1}{2}(1 + \max_{s=m+1,\dots,k} |\xi_s|). \tag{4.151}$$

Kořeny $\tilde{\xi}_s$ závisí pro $s = 1, \dots, m$ na integračním kroku h podle lemmatu 4.5 analyticky, neboť podstatné kořeny jsou jednoduché. Platí tedy

$$\tilde{\xi}_s = \alpha + \beta h + O(h^2) \tag{4.152}$$

pro $h \to 0$ a pro $s = 1, \dots, m$. Určeme koeficienty α a β. Zřejmě je $\alpha = \xi_s$, neboť $\tilde{\xi}_s \to \xi_s$ pro $h \to 0$. Dosazením do rovnice $\tilde{\varrho}(\xi) = 0$ a užitím Taylorova vzorce dostaneme, že existují čísla ξ_s^* a ξ_s^{**} taková, že platí

$$\begin{aligned} \varrho(\xi_s) + [\beta h + O(h^2)]\varrho'(\xi_s) + \tfrac{1}{2}[\beta h + O(h^2)]^2 \varrho''(\xi_s^*) - \\ - hA\{\sigma(\xi_s) + [\beta h + O(h^2)]\sigma'(\xi_s^{**})\} = 0, \end{aligned} \tag{4.153}$$

neboli

$$[\beta\varrho'(\xi_s) - A\sigma(\xi_s)]h + O(h^2) = 0, \tag{4.154}$$

protože je $\varrho(\xi_s) = 0$. Porovnáním koeficientů u stejných mocnin h na levé a pravé straně rovnice (4.154) dostaneme, že pro $s = 1, \dots, m$ platí

$$\tilde{\xi}_s = \xi_s[1 + \lambda_s A h + O(h^2)], \tag{4.155}$$

kde číslo

$$\lambda_s = \frac{\sigma(\xi_s)}{\xi_s \varrho'(\xi_s)} \tag{4.156}$$

nazveme *růstovým parametrem* příslušným k podstatnému kořenu ξ_s. Všimněme si, že v důsledku D-stability má definice růstových parametrů smysl a že v důsledku konzistence je $\lambda_1 = 1$.

Vyšetřeme nyní chování mocnin kořenů $\tilde{\xi}_s$ při $h \to 0$ a $x_n = nh = \text{konst}$. Platí

$$\begin{aligned} [1 + \lambda_s A h + O(h^2)]^n &= \mathrm{e}^{n \ln[1 + \lambda_s A h + O(h^2)]} = \mathrm{e}^{\frac{x_n}{h} \ln[1 + \lambda_s A h + O(h^2)]} = \\ &= \mathrm{e}^{\frac{x_n}{h}[\lambda_s A h + O(h^2)]} = \mathrm{e}^{\lambda_s A x_n} + O(h), \end{aligned} \tag{4.157}$$

a tedy, položíme-li

$$\xi_s = \mathrm{e}^{\mathrm{i}\varphi_s} \tag{4.158}$$

pro $s = 1, \dots, m$, máme

$$\tilde{\xi}_s^n = \mathrm{e}^{\mathrm{i}n\varphi_s}[\mathrm{e}^{\lambda_s A x_n} + O(h)]. \tag{4.159}$$

Speciálně pro $s = 1$ platí $\tilde{\xi}_1^n = \mathrm{e}^{A x_n} + O(h)$. Určeme funkci $O(h)$ v této rovnici přesněji. Podle rovnice (4.155) je

$$\tilde{\xi}_1 = \mathrm{e}^{Ah} + \tau, \tag{4.160}$$

kde τ je analytická funkce h, která se pro $h \to 0$ chová jako $O(h^2)$. Abychom určili tuto funkci přesněji, dosaďme do rovnice $\tilde{\varrho}(\xi) = 0$. Užitím Taylorova vzorce dostaneme

$$\varrho(\mathrm{e}^{Ah}) + \tau\varrho'(\mathrm{e}^{Ah}) + O(\tau^2) - hA\sigma(\mathrm{e}^{Ah}) + O(h\tau) = 0. \tag{4.161}$$

Předně platí $O(\tau^2) + O(h\tau) = O(h\tau)$ pro $h \to 0$, neboť je $\tau = O(h^2)$. Dále, pro každou funkci y, která má $p+2$ spojitých derivací v $\langle a, b\rangle$, platí

$$\begin{aligned} \sum_{\nu=0}^{k} \alpha_\nu y(x + \nu h) - h \sum_{\nu=0}^{k} \beta_\nu y'(x + \nu h) = \\ = C_{p+1} y^{(p+1)}(x) h^{p+1} + O(h^{p+2}), \end{aligned} \tag{4.162}$$

neboť daná metoda je řádu p. Platnost vztahu (4.162) se snadno ověří analogickým postupem jako v důkazu lemmatu 4.4. Speciálně, dosadíme-li do (4.162) za y funkci e^{Ax}, dostáváme

$$\begin{aligned} \sum_{\nu=0}^{k} \alpha_\nu \mathrm{e}^{Ax}\mathrm{e}^{Ah\nu} - hA \sum_{\nu=0}^{k} \beta_\nu \mathrm{e}^{Ax}\mathrm{e}^{Ah\nu} = \\ = C_{p+1} A^{p+1} h^{p+1} \mathrm{e}^{Ax} + O(h^{p+2}), \end{aligned} \tag{4.163}$$

neboli, zkrátíme-li funkcí e^{Ax},

$$\varrho(\mathrm{e}^{Ah}) - hA\sigma(\mathrm{e}^{Ah}) = C_{p+1}(Ah)^{p+1} + O(h^{p+2}). \tag{4.164}$$

Dosadíme-li do rovnice (4.161) z rovnice (4.164), máme

$$\tau\varrho'(\mathrm{e}^{Ah}) = -C_{p+1}(Ah)^{p+1} + O(h\tau) + O(h^{p+2}). \tag{4.165}$$

Vezmeme-li v úvahu, že pro $h \to 0$ platí $O(h\tau) = \tau O(h)$, $\varrho'(\mathrm{e}^{Ah}) = \varrho'(1) + O(h) = \varrho'(1)[1 + O(h)]$ a $1/[1 + O(h)] = 1 + O(h)$, plyne z (4.165), že pro $h \to 0$ platí

$$\tau = -C(Ah)^{p+1} + O(h^{p+2}), \tag{4.166}$$

kde jsme položili

$$C = \frac{C_{p+1}}{\varrho'(1)}. \tag{4.167}$$

Poznamenejme, že konstanta C patří k jedné z charakteristik dané metody a nazývá se *konstanta chyby*.

Dosadíme-li do (4.160) z (4.166), máme

$$\begin{aligned} \tilde{\xi}_1 &= \mathrm{e}^{Ah} - C(Ah)^{p+1} + O(h^{p+2}) = \\ &= \mathrm{e}^{Ah}\{1 - [C(Ah)^{p+1} + O(h^{p+2})]\mathrm{e}^{-Ah}\} = \\ &= \mathrm{e}^{Ah}\{1 - [C(Ah)^{p+1} + O(h^{p+2})][1 + O(h)]\} = \\ &= \mathrm{e}^{Ah}[1 - C(Ah)^{p+1} + O(h^{p+2})]. \end{aligned} \tag{4.168}$$

Odtud ihned plyne, že pro $h \to 0$ a $x_n = nh = \text{konst.}$ platí

$$\tilde{\xi}_1^n = \mathrm{e}^{Ax_n}[1 - x_n A^{p+1} C h^p + O(h^{p+1})], \tag{4.169}$$

neboť je

$$\begin{aligned} &[1 - C(Ah)^{p+1} + O(h^{p+2})]^n = \mathrm{e}^{\{\frac{x_n}{h}\ln[1 - C(Ah)^{p+1} + O(h^{p+2})]\}} = \\ &= \mathrm{e}^{\{-x_n CA^{p+1}h^p + O(h^{p+1})\}} = [1 - x_n CA^{p+1}h^p + O(h^{p+1})][1 + O(h^{p+1})]. \end{aligned} \tag{4.170}$$

Nyní už máme představu, jak závisí na h mocniny kořenů $\tilde{\xi}_s$ ve vzorci (4.147). Dalším naším úkolem je vyjádřit koeficienty A_s v tomto vzorci pomocí počátečních podmínek. Čísla A_s splňují soustavu lineárních algebraických rovnic (4.148). Abychom tuto soustavu rozřešili, definujme čísla δ_r, $r = 0, \ldots, k-1$, rovnicemi

$$\delta_r = y_r - \tilde{\xi}_1^r \tag{4.171}$$

a polynomy $\tilde{\varrho}_\nu$, $\nu = 1, \ldots, k$, rovnicemi

$$\tilde{\varrho}_\nu(\xi) = \frac{\tilde{\varrho}(\xi)}{\xi - \tilde{\xi}_\nu} = \tilde{\alpha}_{\nu 0} + \tilde{\alpha}_{\nu 1}\xi + \ldots + \tilde{\alpha}_{\nu,k-1}\xi^{k-1}. \tag{4.172}$$

Protože je $\tilde{\varrho}(\tilde{\xi}_\nu) = 0$ pro $\nu = 1, \ldots, k$, platí

$$\tilde{\varrho}_\nu(\tilde{\xi}_s) = \begin{cases} 0, & \nu \neq s, \\ \tilde{\varrho}'(\tilde{\xi}_\nu), & \nu = s. \end{cases} \tag{4.173}$$

Dosadíme-li za pravou stranu v rovnici (4.148) podle (4.171), vynásobíme-li každou z takto vzniklých rovnic koeficientem $\tilde{\alpha}_{\nu r}$ a sečteme od $r = 0$ do $r = k - 1$,

dostaneme

$$\sum_{s=1}^{k} A_s \tilde{\varrho}_\nu(\tilde{\xi}_s) = \tilde{\varrho}_\nu(\tilde{\xi}_1) + \tilde{\Lambda}_\nu, \quad \nu = 1, \ldots, k, \tag{4.174}$$

kde

$$\tilde{\Lambda}_\nu = \sum_{r=0}^{k-1} \tilde{\alpha}_{\nu r}\delta_r. \tag{4.175}$$

Použijeme-li rovnic (4.173), přejdou rovnice (4.174) v

$$A_\nu \tilde{\varrho}'(\tilde{\xi}_\nu) = \tilde{\varrho}_\nu(\tilde{\xi}_1) + \tilde{\Lambda}_\nu, \quad \nu = 1, \ldots, k. \tag{4.176}$$

Protože je $\tilde{\varrho}'(\tilde{\xi}_\nu) \neq 0$ pro $\nu = 1, \ldots, k$ (kořeny polynomu $\tilde{\varrho}$ jsou jednoduché), plyne z rovnic (4.176), že platí

$$A_1 = 1 + \frac{\tilde{\Lambda}_1}{\tilde{\varrho}'(\tilde{\xi}_1)}, \quad A_\nu = \frac{\tilde{\Lambda}_\nu}{\tilde{\varrho}'(\tilde{\xi}_\nu)}, \quad \nu = 2, \ldots, k. \tag{4.177}$$

Podařilo se nám tedy jednoduše vyjádřit koeficienty A_ν pomocí veličin $\tilde{\Lambda}_\nu$. Koeficienty $\tilde{\alpha}_{\nu r}$, které vystupují v definici těchto $\tilde{\Lambda}_\nu$ však závisí na h. Pokusme se tuto závislost blíže osvětlit. Definujme proto pro $\nu = 1, \ldots, k$ veličiny Λ_ν rovnicemi

$$\Lambda_\nu = \sum_{r=0}^{k-1} \alpha_{\nu r}\delta_r, \tag{4.178}$$

kde

$$\varrho_\nu(\xi) = \frac{\varrho(\xi)}{\xi - \xi_\nu} = \alpha_{\nu 0} + \alpha_{\nu 1}\xi + \ldots + \alpha_{\nu,k-1}\xi^{k-1}, \tag{4.179}$$

takže koeficienty $\alpha_{\nu r}$ už na h nezávisí. Zřejmě platí

$$|\tilde{\Lambda}_\nu - \Lambda_\nu| \leqq \delta \sum_{r=0}^{k-1} |\tilde{\alpha}_{\nu r} - \alpha_{\nu r}|, \tag{4.180}$$

kde

$$\delta = \max_{s=0,\ldots,k-1} |\delta_s|. \tag{4.181}$$

K posouzení rozdílu $\tilde{\Lambda}_\nu - \Lambda_\nu$ je tedy třeba odhadnout rozdíl $\tilde{\alpha}_{\nu r} - \alpha_{\nu r}$. Koeficienty $\alpha_{\nu r}$ polynomu ϱ_ν lze počítat jednoduše z koeficientů α_r polynomu ϱ Hornerovým schématem, tj. z rekurence

$$\begin{aligned} &\alpha_{\nu k} = 0, \\ &\alpha_{\nu r} = \alpha_{r+1} + \xi_\nu \alpha_{\nu,r+1}, \quad r = k-1, \ldots, 0, \end{aligned} \tag{4.182}$$

a podobné rekurence platí pro koeficienty $\tilde{\alpha}_{\nu r}$ polynomu $\tilde{\varrho}_\nu$. Platí tedy

$$\tilde{\alpha}_{\nu r} - \alpha_{\nu r} = \tilde{\alpha}_{r+1} - \alpha_{r+1} + (\tilde{\xi}_\nu - \xi_\nu)\tilde{\alpha}_{\nu,r+1} + \xi_\nu(\tilde{\alpha}_{\nu,r+1} - \alpha_{\nu,r+1}) \tag{4.183}$$

pro $r = k-1, \dots, 0$. Protože však platí podle lemmatu 4.5, že je

$$\tilde{\xi}_\nu - \xi_\nu = O(h^{1/s}), \tag{4.184}$$

kde s je násobnost kořene ξ_ν polynomu ϱ a protože je $\tilde{\alpha}_r = \alpha_r - hA\beta_r$, plyne z rovnic (4.183) snadno úplnou indukcí, že platí

$$\tilde{\alpha}_{\nu r} - \alpha_{\nu r} = O(h^{1/s}), \tag{4.185}$$

a tedy, že je

$$\tilde{\Lambda}_\nu = \Lambda_\nu + O(\delta h^{1/s}). \tag{4.186}$$

Vyšetřeme konečně, jak závisí na h čísla $\tilde{\varrho}'(\tilde{\xi}_\nu)$. Podle definice polynomu $\tilde{\varrho}$ platí

$$\tilde{\varrho}'(\tilde{\xi}_\nu) = \varrho'(\tilde{\xi}_\nu) - hA\sigma'(\tilde{\xi}_\nu). \tag{4.187}$$

Podle Taylorova vzorce platí pro každé $s \leqq k$

$$\begin{aligned}\varrho'(\tilde{\xi}_\nu) = \varrho'(\xi_\nu) + \frac{1}{1!}\varrho''(\xi_\nu)(\tilde{\xi}_\nu - \xi_\nu) + \dots + \\ + \frac{1}{(s-1)!}\varrho^{(s)}(\xi_\nu)(\tilde{\xi}_\nu - \xi_\nu)^{s-1} + \frac{1}{s!}\varrho^{(s+1)}(\xi_\nu^*)(\tilde{\xi}_\nu - \xi_\nu)^s.\end{aligned} \tag{4.188}$$

Je-li ξ_ν s-násobný kořen polynomu ϱ, plyne ze vzorců (4.188) a (4.184), že platí

$$\varrho'(\tilde{\xi}_\nu) = \frac{1}{(s-1)!}\varrho^{(s)}(\xi_\nu)(\tilde{\xi}_\nu - \xi_\nu)^{s-1} + O(h). \tag{4.189}$$

Pro jednoduchý kořen ξ_ν polynomu ϱ tedy odtud a z (4.187) dostáváme

$$\tilde{\varrho}'(\tilde{\xi}_\nu) = \varrho'(\xi_\nu) + O(h). \tag{4.190}$$

Buď nyní ξ_ν kořen polynomu ϱ násobnosti s. Holomorfní funkce Φ vystupující v lemmatu 4.5 má v bodě $t = 0$ derivaci od nuly různou. Platí tedy nejen (4.184), ale existuje dokonce konstanta $\gamma_\nu \neq 0$ taková, že je

$$\tilde{\xi}_\nu - \xi_\nu = \gamma_\nu h^{1/s}[1 + O(h^{1/s})]. \tag{4.191}$$

Použijeme-li rovnice (4.187) a (4.189), plyne odtud, že pro kořen ξ_ν polynomu ϱ násobnosti s platí

$$\begin{aligned}\tilde{\varrho}'(\tilde{\xi}_\nu) &= \varepsilon_\nu h^{(s-1)/s}[1 + O(h^{1/s})] + O(h) = \\ &= \varepsilon_\nu h^{(s-1)/s}[1 + O(h^{1/s})],\end{aligned} \tag{4.192}$$

kde $\varepsilon_\nu \neq 0$ je vhodná konstanta.

Získané odhady už umožňují posoudit závislost koeficientů A_ν na h. Dosadíme-li do vzorce (4.177) podle (4.186) a (4.190), dostaneme

$$A_1 = 1 + \frac{\Lambda_1}{\varrho'(1)} + O(\delta h), \quad A_\nu = \frac{\Lambda_\nu}{\varrho'(\xi_\nu)} + O(\delta h), \quad \nu = 2, \dots, m, \tag{4.193}$$

neboť kořeny ξ_ν jsou pro $\nu = 1, \dots, m$ jednoduché. Pro $\nu = m+1, \dots, k$ o násobnosti kořenů nic nevíme, a musíme proto vyjít ze vzorce (4.192) platného pro libovolnou násobnost. Z něj snadno plyne, že je

$$\begin{aligned}\frac{1}{\tilde{\varrho}'(\tilde{\xi}_\nu)} &= \frac{1}{\varepsilon_\nu h^{(s-1)/s}[1 + O(h^{1/s})]} = \\ &= \frac{1}{\varepsilon_\nu h^{1-1/s}}[1 + O(h^{1/s})] = O\Big(\frac{1}{h}\Big).\end{aligned} \tag{4.194}$$

Pro $\nu = m+1, \dots, k$ tedy máme

$$A_\nu = [\Lambda_\nu + O(\delta h^{1/s})]O\Big(\frac{1}{h}\Big) = O\Big(\frac{\delta}{h}\Big). \tag{4.195}$$

Dosaďme konečně do vzorce (4.147) odhady (4.150), (4.159), (4.169) (4.193) a (4.195). Dostaneme

$$\begin{aligned}y_n = \Big[1 + \frac{\Lambda_1}{\varrho'(1)} + O(\delta h)\Big]\mathrm{e}^{Ax_n}[1 - x_n CA^{p+1}h^p + O(h^{p+1})] + \\ + \sum_{\nu=2}^{m}\Big[\frac{\Lambda_\nu}{\varrho'(\xi_\nu)} + O(\delta h)\Big]\mathrm{e}^{\mathrm{i}n\varphi_\nu}[\mathrm{e}^{\lambda_\nu Ax_n} + O(h)] + \sum_{\nu=m+1}^{k} O\Big(\frac{\delta}{h}\Big)t^{\frac{x_n}{h}}.\end{aligned} \tag{4.196}$$

Odečteme-li od levé i pravé strany rovnice (4.196) přesné řešení $y(x_n) = \mathrm{e}^{Ax_n}$, dostaneme po jednoduchých úpravách

$$\begin{aligned}y_n - y(x_n) = -x_n CA^{p+1}\mathrm{e}^{Ax_n}h^p + O(h^{p+1}) + \\ + \sum_{\nu=1}^{m}\frac{\Lambda_\nu}{\varrho'(\xi_\nu)}\mathrm{e}^{\mathrm{i}n\varphi_\nu}\mathrm{e}^{\lambda_\nu Ax_n} + O(\delta h),\end{aligned} \tag{4.197}$$

neboť $\mathrm{e}^{x_n/h}/h^2 \to 0$ pro $h \to 0$.

Čísla Λ_ν jsou dána vzorci (4.178), a závisí tedy na daných počátečních podmínkách dosti nepřehledně. Vyšetřeme proto podrobněji tuto závislost, samozřejmě za vhodných doplňujících předpokladů. Předně podle (4.171) platí pro $s = 0, \dots, k-1$

$$\delta_s = y_s - \tilde{\xi}_1^s = y(x_s) - \tilde{\xi}_1^s + y_s - y(x_s). \tag{4.198}$$

Pro $\tilde{\xi}_1$ však platí (4.168). Pro $s = 0, \dots, k-1$ tedy platí

$$\tilde{\xi}_1^s = \mathrm{e}^{Ash}[1 - (Ah)^{p+1}C + O(h^{p+2})]^s = \mathrm{e}^{Ash} + O(h^{p+1}). \tag{4.199}$$

Dosazení do (4.198) tedy máme

$$\delta_s = y_s - y(x_s) + O(h^{p+1}). \tag{4.200}$$

Předpokládejme dále, že počáteční hodnoty $y_0, \dots, y_{k-1}$ jsou dány s přesností h^q (q je nějaké přirozené číslo) a že pro ně platí asymptotické vzorce

$$y_s = y(x_s) + \gamma_s h^q + O(h^{q+1}), \quad s = 0, \dots, k-1, \tag{4.201}$$

kde γ_s jsou konstanty. Položíme-li

$$D_\nu = \sum_{s=0}^{k-1} \alpha_{\nu s}\gamma_s, \quad s = 1, \dots, k, \tag{4.202}$$

můžeme psát

$$\Lambda_\nu = D_\nu h^q + O(h^{r+1}), \quad \nu = 1, \dots, k, \tag{4.203}$$

kde

$$r = \min(p, q). \tag{4.204}$$

Protože je za uvedených předpokladů $O(\delta h) = O(h^{q+1})$, dostáváme dosazením do (4.197) konečný výsledek, který zformulujeme do věty.

Věta 4.7. *Buď dána diferenciální rovnice $y' = Ay$ s počáteční podmínkou $y(0) = 1$ a buď y_n její přibližné řešení získané lineární k-krokovou metodou řádu p, které je určené počátečními podmínkami, pro něž platí* (4.201). *Pak platí*

$$y_n - y(x_n) = -x_n C A^{p+1} \mathrm{e}^{Ax_n} h^p + \\ + h^q \sum_{\nu=1}^{m} \frac{D_\nu}{\varrho'(\xi_\nu)} \mathrm{e}^{\mathrm{i}n\varphi_\nu} \mathrm{e}^{\lambda_\nu A x_n} + O(h^{r+1}), \tag{4.205}$$

kde $\xi_\nu = \mathrm{e}^{\mathrm{i}\varphi_\nu}$ jsou kořeny charakteristického polynomu ϱ, které jsou v absolutní hodnotě rovny jedné, C a λ_ν jsou konstanty dané vzorci (4.167) a (4.156), D_ν jsou funkce počátečních podmínek definované vzorci (4.202) a pro r platí (4.204).

Ze vzorce (4.205) je vidět, že celková diskretizační chyba se skládá ze dvou podstatně odlišných částí. První část daná členem $-x_n C A^{p+1} \mathrm{e}^{Ax_n} h^p$ plně odpovídá celkové diskretizační chybě u obecné jednokrokové metody, neboť funkce $-xCA^{p+1}\mathrm{e}^{Ax}$ je řešením diferenciální rovnice $e' = f_y(x, y(x))e + \varphi(x, y(x))$, kde $\varphi(x, y(x)) = -Cy^{(p+1)}(x)$. Je to tedy celková diskretizační chyba v užším smyslu, neboť závisí pouze na *hlavní části lokální diskretizační chyby*, tj. na prvním nenulovém členu v jejím rozvoji do mocninné řady podle integračního kroku h.

Druhá část celkové diskretizační chyby je ve vzorci (4.205) reprezentována součtem a závisí na počátečních hodnotách přibližného řešení. Ze zmíněného vzorce je vidět, že struktura této startovní chyby je podstatně složitější, než je tomu u vlastní diskretizační chyby. Nicméně v případě $m = 1$ (tj. nemá-li polynom jiné kořeny, které jsou v absolutní hodnotě rovny jedné, než ξ_1) není ani tato část chyby nebezpečná. Růstový parametr je totiž v tomto případě pouze jeden, a to rovný jedné, a startovní chyba je úměrná přesnému řešení, což je přijatelná situace. Je-li však $m > 1$, objeví se v součtu (4.205) dodatečné oscilační členy (oscilační v důsledku přítomnosti činitele $\mathrm{e}^{\mathrm{i}n\varphi_\nu}$), jejichž amplituda je úměrná číslu $\mathrm{e}^{(\lambda_\nu A)x}$. Je-li $\lambda_\nu A < A$, nejsou rovněž nebezpečné. V opačném případě však jejich amplituda roste rychleji než přesné řešení, a mohou tedy velmi rychle počítané řešení úplně znehodnotit.

Na první pohled se může zdát, že všech potíží se startovní chybou se zbavíme, zadáme-li počáteční podmínky pro přibližné řešení přesně. Nehledě na to, že z praktického hlediska je takové opatření obecně těžko proveditelné, není tato úvaha ani věcně správná. Ze vzorce (4.198) je totiž vidět, že veličiny δ_ν nejsou v tomto případě rovny nule. Je tedy $\delta \neq 0$ a vrátíme-li se ke vzorci (4.197), z něhož jsme vzorec (4.205) odvodili, vidíme, že startovní chyba se tak jako tak objeví.

Ani volba $y_\nu = \tilde{\xi}_1^\nu$ vedoucí k $\delta_\nu = 0$ mnoho nespraví, neboť v důsledku zaokrouhlování nejsme schopni splnit tyto podmínky přesně.

Ilustrujme popsané chování jednoduchým příkladem:

Příklad 4.2. Uvažujme metodu

$$y_{n+2} = y_n + 2hf_{n+1}. \tag{4.206}$$

Tato explicitní metoda je řádu 2 a je pro ni $\varrho(\xi) = \xi^2 - 1$, $\sigma(\xi) = 2\xi$, $\xi_1 = 1$ a $\xi_2 = -1$. Podstatné kořeny jsou tedy dva ($m = 2$) a příslušné růstové parametry jsou $\lambda_1 = 1$ a $\lambda_2 = -1$. Řešíme-li touto metodou modelovou diferenciální rovnici $y' = Ay$ s $A < 0$, osciluje chyba s amplitudou úměrnou e^{-Ax}, a tedy s délkou intervalu, v němž hledáme řešení, exponenciálně roste.

Řešme metodou (4.206) nelineární diferenciální rovnici

$$y' = 1 - y^2 \tag{4.207}$$

s počáteční podmínkou $y(0) = 5$. Výsledky jsou pro $h = 1/64$ uvedeny v tab. 4.9, která pro srovnání obsahuje také výsledky získané dvoukrokovou Adamsovou-Bashforthovou metodou. V obou případech jsme za y_1 vzali hodnotu přesného řešení. I když rovnice (4.207) je nelineární, jev, který jsme předpověděli pro modelovou rovnici, nastává i zde.

Metody, které se chovají jako metoda (4.206), tj. metody, pro něž je $m > 1$ a některý z růstových parametrů je záporný, se nazývají *slabě nestabilní*. Tyto metody jsou sice D-stabilní, a tedy konvergentní (tj. při daném intervalu, ve kterém hledáme řešení, lze volbou dostatečně malého integračního kroku a dostatečně přesnou aproximací počátečních hodnot dosáhnout toho, že chyba je libovolně malá), závislost integračního kroku a přesnosti aproximace počátečních podmínek však není stejnoměrná vzhledem k intervalu, v němž hledáme řešení. Tato skutečnost je z praktického hlediska značně nepříjemná, neboť integrační krok a přesnost počátečních podmínek, které jsou dostačující v daném intervalu, nemusí stačit, objeví-li se nutnost řešit danou rovnici na delším intervalu.

I když všechny zde provedené úvahy platí přísně vzato pouze pro modelovou diferenciální rovnici $y' = Ay$, dávají kvalitativně velmi přesnou indikaci o chování chyby v obecném případě. To je intuitivně jasné a dá se to i teoreticky dokázat. Do dalších podrobností (které jsou formálně značně složité) však nebudeme zacházet.

Závěrem ještě upozorněme, že užít existenci asymptotického vzorce k získání odhadu chyby metodou poloviního kroku (podobně jako v případě obecné jedno-

Tabulka 4.9

Řešení diferenciální rovnice (4.207) dvěma dvoukrokovými metodami

n	Metoda (4.206)		Adamsova-Bashforthova metoda	
	y_n	e_n	y_n	e_n
0	5,000 000	0	5,000 000	0
1	4,652 200	0	4,652 200	0
2	4,354 907	0,003 373	4,355 881	0,004 347
⋮				
63	1,159 287	−0,048 115	1,206 536	0,001 252
64	1,246 457	0,048 115	1,199 552	0,001 210
65	1,141 986	−0,049 670	1,192 825	0,001 170
⋮				
127	0,599 828	−0,425 689	1,025 688	0,000 171
128	1,409 513	0,384 790	1,024 888	0,000 166
129	0,568 993	−0,454 960	1,024 114	0,000 161
⋮				
191	−1,606 091	−2,609 506	1,003 441	0,000 026
192	1,827 522	0,824 211	1,003 336	0,000 025
193	−1,679 211	−2,682 419	1,003 233	0,000 024

krokové metody) lze jen tehdy, je-li zaručeno, že hlavní část celkové diskretizační chyby se chová „rozumně“, tj. nenastávají-li jevy slabé nestability.

4.2.4 Problematika zaokrouhlovacích chyb

V tomto odstavci si velice stručně všimneme problematiky šíření zaokrouhlovacích chyb. Zavedeme-li opět zcela analogicky jako u jednokrokové metody lokální zaokrouhlovací chybu ε_n, splňuje skutečně počítané přibližné řešení $\tilde{y}_n$ rovnici

$$\sum_{\nu=0}^{k} \alpha_\nu \tilde{y}_{n+\nu} = h \sum_{\nu=0}^{k} \beta_\nu f(x_{n+\nu}, \tilde{y}_{n+\nu}) + \varepsilon_n \tag{4.208}$$

s počátečními podmínkami $\tilde{y}_0, \dots, \tilde{y}_{k-1}$. Použijeme-li lemma 4.3, dostáváme ihned, že platí následující věta.

Věta 4.8. *Buď y_n přibližné řešení vypočtené D-stabilní mnohokrokovou metodou a buď $\tilde{y}_n$ řešení rovnice* (4.208). *Buď konečně $|\varepsilon_n| \leqq \varepsilon$ pro $n = 0, \dots, N$. Pak*

platí

$$|\tilde{y}_n - y_n| \leqq \left[\Gamma \max_{s=0,\dots,k-1} |\tilde{y}_s - y_s| + \Gamma^*(x_n - a)\frac{\varepsilon}{h}\right] e^{(x_n - a)L^*}. \tag{4.209}$$

Vidíme tedy, že celková zaokrouhlovací chyba bez ohledu na řád metody roste při $h \to 0$ opět lineárně. O důsledcích, které z tohoto poznatku plynou, platí tak doslovně totéž, co bylo řečeno u Eulerovy metody nebo u obecné jednokrokové metody.

4.2.5 Stabilita při pevném integračním kroku

Charakter vět týkajících se konvergence metod, které jsme až dosud uvedli, je asymptotický, tj. příslušná tvrzení platí pro dostatečně malá h. Jen pro dostatečně malé h je např. zaručeno, že při užití D-stabilní a konzistentní mnohokrokové metody (ale také např. Rungovy-Kuttovy metody) nedojde při rekurentním užívání příslušného vzorce ke katastrofálnímu nakupení lokálních chyb a celková diskretizační chyba zůstane přijatelně malá. Při konkrétním výpočtu však vždy pracujeme s pevným h, které nemusí být dostatečně malé. V této souvislosti je důležitý pojem tzv. intervalu absolutní stability dané metody. Zhruba řečeno, chceme, aby to byl takový interval, že leží-li v něm číslo hA, kde h je integrační krok a A je odhad pro derivace $\partial f/\partial y$, je více méně zaručeno, že lokální chyby, kterých se dopouštíme v jednotlivých krocích, se v průběhu výpočtu tlumí. Jinými slovy žádáme, aby přibližné řešení modelové diferenciální rovnice $y' = Ay$ při pevném h konvergovalo pro $n \to \infty$ k nule.

Přistupme nyní k přesné definici. Zaveďme k tomu cíli tzv. *třetí charakteristický polynom* π dané k-krokové metody rovnicí

$$\pi(\xi, z) = \varrho(\xi) - z\sigma(\xi). \tag{4.210}$$

Definice 4.5. Řekneme, že lineární k-kroková metoda daná polynomy ϱ a σ je pro dané z *absolutně stabilní*, platí-li pro všechny kořeny $\tilde{\xi}_s$, $s = 1, \dots, k$, polynomu π nerovnosti $|\tilde{\xi}_s| < 1$. V opačném případě říkáme, že daná metoda je pro dané z absolutně nestabilní. *Intervalem absolutní stability* pak rozumíme největší takový interval, že pro z v něm ležící je metoda absolutně stabilní.

Poznámka 4.5. Je zřejmé, že analogickou definici lze vyslovit i pro jednokrokovou metodu.

Na základě motivace předcházející tuto definici se lze domnívat, že při užití integračního kroku h takového, že z určené rovnicí

$$z = hA \tag{4.211}$$

leží v intervalu absolutní stability, nedojde v průběhu řešení k nepřípustnému růstu chyby.

Interval absolutní stability dané mnohokrokové metody je určen pouze jejími koeficienty. Přípustný integrační krok je však rovnicí (4.211) vázán na konkrétní diferenciální rovnici. V případě obecné nelineární rovnice klademe, jak už jsme řekli, $A = \partial f/\partial y$. Veličina $\partial f/\partial y$ není samozřejmě konstantní; za A se bere v tomto případě odhad derivace $\partial f/\partial y$ nebo některá její typická hodnota, která se v průběhu výpočtu eventuálně mění. Změnám ve volbě A pak přirozeně odpovídají i změny v maximální přípustné velikosti integračního kroku.

Snadno se zjistí (viz cvič. 9), že každá D-stabilní a konzistentní lineární k-kroková metoda je absolutně nestabilní pro každé dostatečně malé kladné z. Odtud plyne, že užije-li se mnohokroková metoda k řešení diferenciální rovnice, pro níž je $\partial f/\partial y > 0$, chyba v průběhu výpočtu roste. To však není tak vážné, jak by se mohlo na první pohled zdát. Uvedené tvrzení platí totiž přesně pouze za předpokladu, že daná diferenciální rovnice je typu $y' = Ay + q(x)$, kde A je konstanta. Řešení této rovnice obsahuje člen e^{Ax}, který je aproximován členem $\tilde{\xi}_1^n$, kde $\tilde{\xi}_1$ je ten kořen polynomu π, pro který platí $\tilde{\xi}_1 \to 1$ pro $h \to 0$ (srv. vzorec (4.169)). Zdá se tedy, že převládne-li v přibližném řešení tento člen, chyba sice v průběhu řešení poroste, ale tento růst bude úměrný řešení. To však je přijatelná situace. Úvahy tohoto typu vedou k vyslovení alternativní definice stability, kdy řekneme, že pro dané z je metoda *relativně stabilní*, platí-li pro kořeny polynomu π nerovnosti $|\tilde{\xi}_s| < |\tilde{\xi}_1$ pro $s = 2, \ldots, k$. *Interval relativní stability* je pak maximální interval takový, že pro z v něm ležící je metoda relativně stabilní.

V souvislosti s popisovanou problematikou se v literatuře vyskytuje řada mírně se navzájem lišících definic — je to např. požadavek, že je $|\tilde{\xi}_s| \leqq |\tilde{\xi}_1|$ a ty kořeny $\tilde{\xi}_s$, pro něž nastává rovnost, jsou jednoduché, nebo požadavek, že je $|\tilde{\xi}_s| \leqq e^z$ apod. — a vždy se přitom mluví o relativní stabilitě. Rozdíly sice nejsou většinou podstatné, je však třeba při porovnávání výsledků si vždy uvědomit příslušnou přesnou definici.

I když se zdá, že pojem relativní stability byl zaveden pouze proto, abychom byli schopni pracovat s případem $z > 0$, je užitečný i v obecném případě. Existují dokonce dobré důvody domnívat se, že relativní stabilita je pro posouzení chování chyby výstižnější než absolutní stabilita. Na druhé straně je však těžší ji podrobně analyzovat. Upozorněme ještě, že přestože intervaly stability představují velice užitečnou charakteristiku dané metody, při utváření příliš kategorických soudů na jejich základě je třeba být opatrný. Veškeré úvahy, které odtud vycházejí, jsou totiž ovlivněny tím, že se vlastně předpokládá, že $\partial f/\partial y$ je konstantní. Pro silně nelineární problémy, kdy se hodnota funkce $\partial f/\partial y$ rychle mění, dává naznačená teorie jen hrubou orientaci a je nebezpečné brát velikost integračního kroku příliš blízko největší přípustné odhadnuté hodnotě.

4.2.6 Optimální metody

V tomto odstavci si všimneme přirozené otázky, jak sestrojit lineární k-krokovou metodu, která je maximálního možného řádu. Polynomy ϱ a σ, jimiž je metoda

definována, mají dohromady $2k + 2$ koeficientů. Jeden z nich lze libovolně volit, neboť polynomy $\alpha\varrho$ a $\alpha\sigma$, kde $\alpha = \text{konst.} \neq 0$, definují zřejmě tutéž metodu, takže podstatných koeficientů je pouze $2k + 1$. Uvažujeme-li tedy pouze podmínky určující, že metoda je řádu p, je možno obecně dosáhnout, že je $p = 2k$ (viz definice 4.2). Konvergentní metoda však musí být navíc stabilní. Tento požadavek patrně nedovolí dosáhnout výše zmíněného maximálního řádu.

Abychom relaci mezi p a k blíže objasnili, začneme poněkud jiným, jednodušším problémem. Buď dán polynom ϱ stupně k, který splňuje podmínky stability a pro nějž platí $\varrho(1) = 0$. Máme sestrojit polynom σ stupně nejvýše k tak, aby k-kroková metoda daná polynomy ϱ a σ měla maximální řád. Abychom takový polynom sestrojili, definujme nejprve řád mnohokrokové metody poněkud jiným způsobem než v definici 4.2.

Buď tedy dána lineární k-kroková metoda, tj. polynomy ϱ a σ. Definujme funkci φ komplexní proměnné ξ vzorcem

$$\varphi(\xi) = \frac{1}{\ln \xi}\varrho(\xi) - \sigma(\xi), \tag{4.212}$$

přičemž jednoznačnou větev funkce $\ln \xi$ zvolíme tak, že rozřízneme komplexní rovinu podél záporné reálné osy a položíme $\ln 1 = 0$. Funkce $1/\ln \xi$ má v bodě $\xi = 1$ pól prvního řádu. Je-li daná metoda konzistentní, je $\varrho(1) = 0$, a tedy funkce $\varrho(\xi)/\ln \xi$ je holomorfní v bodě $\xi = 1$. Totéž platí samozřejmě i pro funkci φ. Zmíněná alternativní definice řádu mnohokrokové metody se opírá o následující větu.

Věta 4.9. *K tomu, aby mnohokroková metoda daná polynomy ϱ a σ, byla řádu p, je nutné a stačí, aby funkce φ definovaná rovnicí* (4.212) *měla v bodě $\xi = 1$ kořen násobnosti p.*

D ů k a z . Nechť daná metoda je řádu p. Pak pro každou dostatečně hladkou funkci y platí

$$L(y(x); h) = C_{p+1}y^{(p+1)}(x)h^{(p+1)} + O(h^{p+2}) \tag{4.213}$$

a je $C_{p+1} \neq 0$. Tedy speciálně, položíme-li ve vztahu (4.213) $y(x) = e^x$, máme

$$e^x \sum_{\nu=0}^{k} \alpha_\nu e^{\nu h} - he^x \sum_{\nu=0}^{k} \beta_\nu e^{\nu h} = C_{p+1}e^x h^{p+1} + O(h^{p+2}), \tag{4.214}$$

neboli

$$\varrho(e^h) - h\sigma(e^h) = C_{p+1}h^{p+1} + O(h^{p+2}). \tag{4.215}$$

Odtud ale plyne, že holomorfní funkce $f(h) = \varrho(e^h) - h\sigma(e^h)$ má v bodě $h = 0$ kořen násobnosti $p + 1$. Funkce $f(h)/h$ má tedy v bodě $h = 0$ kořen násobnosti p. Protože transformace $\xi = e^h$ zobrazuje jednoznačně okolí bodu $h = 0$ na okolí bodu $\xi = 1$, platí totéž pro funkci φ a pro bod $\xi = 1$.

Naopak, nechť funkce φ má v bodě $\xi = 1$ kořen násobnosti p. Pak funkce $h\varphi(\mathrm{e}^h)$ má v bodě $h = 0$ kořen násobnosti $p+1$. Platí tedy

$$\varrho(\mathrm{e}^h) - h\sigma(\mathrm{e}^h) = \gamma h^{p+1} + O(h^{p+2}), \tag{4.216}$$

kde γ je nenulová konstanta. Porovnáním koeficientů u stejných mocnin proměnné h na levé i pravé straně rovnice (4.216) ihned dostaneme, že pro čísla C_q definovaná v definici 4.2 platí $C_0 = C_1 = \ldots = C_p = 0$ a $C_{p+1} = \gamma \neq 0$. Metoda je tedy řádu p a věta je dokázána.

Vzhledem k tomu, že tvrzení věty 4.9 je nutná a postačující podmínka pro to, aby metoda byla daného řádu, lze je použít jako alternativní definici tohoto pojmu. Tato nová definice je na rozdíl od algebraické definice 4.2 analytická.

Na základě věty 4.9 už snadno odpovíme na otázku, kterou jsme si před její formulací položili.

Věta 4.10. *Buď ϱ polynom stupně k, pro který platí $\varrho(1) = 0$ a buď $\tilde{k}$ celé číslo, pro něž platí $0 \leqq \tilde{k} \leqq k$. Pak existuje právě jeden polynom σ stupně nejvýše $\tilde{k}$ takový, že mnohokroková metoda daná polynomy ϱ a σ je řádu nejméně $\tilde{k}+1$.*

D ů k a z . Funkce $\varrho(\xi)/\ln\xi$ je holomorfní v bodě $\xi = 1$; v okolí tohoto bodu tedy platí

$$\frac{\varrho(\xi)}{\ln\xi} = c_0 + c_1(\xi-1) + c_2(\xi-1)^2 + \ldots . \tag{4.217}$$

Položme

$$\sigma(\xi) = c_0 + c_1(\xi-1) + \ldots + c_{\tilde{k}}(\xi-1)^{\tilde{k}}. \tag{4.218}$$

Pak funkce $\varphi(\xi) = \varrho(\xi)/\ln\xi - \sigma(\xi)$ má v bodě $\xi = 1$ kořen násobnosti nejméně $\tilde{k}+1$ a příslušná mnohokroková metoda je řádu nejméně $\tilde{k}+1$.

Nechť naopak ϱ a σ definují metodu řádu nejméně $\tilde{k}+1$. Pak σ musí být tvaru (4.218), jak plyne z jednoznačnosti Taylorova rozvoje. Polynom σ je tedy určen jednoznačně a věta je dokázána.

Poznámka 4.6. Metoda daná polynomy ϱ a σ z věty 4.10 je řádu $p > \tilde{k}+1$, platí-li pro koeficienty v rozvoji (4.217) $c_{\tilde{k}+1} = c_{\tilde{k}+2} = \ldots = c_{p-1} = 0$.

Tato poznámka, i když je triviální, bude v dalších úvahách velmi užitečná.

Vraťme se nyní ke studiu k-krokové stabilní metody mající nejvyšší možný řád. Z věty 4.10 plyne, že tato v právě uvedeném smyslu optimální metoda má řád nejméně $k+1$. Toto číslo však už nelze o mnoho překročit, jak vyplývá z následujících vět.

Věta 4.11. *Stabilní k-kroková metoda s lichým k nemůže být řádu vyššího než $k+1$.*

Věta 4.12. *Stabilní k-kroková metoda se sudým k je řádu nejvýše $k+2$. Řádu právě $k+2$ je tehdy a jen tehdy, má-li polynom ϱ všechny kořeny v absolutní hodnotě rovny jedné, je-li $+1$ a -1 mezi nimi a je-li polynom σ určen podle věty 4.10.*

Při lichém k tedy existuje nekonečně mnoho optimálních metod. Abychom některou z nich dostali, stačí zřejmě vzít libovolný polynom ϱ splňující podmínky stability a takový, že je $\varrho(1) = 0$ a určit polynom σ podle věty 4.10. Optimálních k-krokových metod při sudém $k > 2$ je sice také nekonečně mnoho, jejich množina je však podstatně užší než v případě lichého k a její struktura je zcela popsána ve větě 4.12.

K důkazu vět 4.11 a 4.12 budeme potřebovat řadu pomocných tvrzení.

Lemma 4.6. *Buď ϱ polynom stupně k a buď r polynom definovaný rovnicí*

$$r(z) = \Big(\frac{1-z}{2}\Big)^k \varrho\Big(\frac{1+z}{1-z}\Big). \tag{4.219}$$

Buď dále $\xi_0 \neq -1$ kořen polynomu ϱ násobnosti m. Pak

$$z_0 = \frac{\xi_0 - 1}{\xi_0 + 1} \tag{4.220}$$

je m-násobným kořenem polynomu r. Naopak, buď $z_0 \neq 1$ kořen polynomu r násobnosti m. Pak

$$\xi_0 = \frac{1+z_0}{1-z_0} \tag{4.221}$$

je m-násobný kořen polynomu

$$\varrho(\xi) = (\xi+1)^k r\Big(\frac{\xi-1}{\xi+1}\Big). \tag{4.222}$$

D ů k a z . Je-li $\xi_0 \neq -1$ m-násobný kořen polynomu ϱ, platí

$$\varrho(\xi) = \frac{1}{m!}\varrho^{(m)}(\xi_0)(\xi-\xi_0)^m + \ldots \tag{4.223}$$

a $\varrho^{(m)}(\xi_0) \neq 0$. Dosazením do (4.219) dostáváme, že je

$$\begin{aligned} r(z) &= \Big(\frac{1-z}{2}\Big)^k \Big[\frac{1}{m!}\varrho^{(m)}(\xi_0)\Big(\frac{1+z}{1-z} - \frac{1+z_0}{1-z_0}\Big)^m + \ldots\Big] = \\ &= \Big(\frac{1-z}{2}\Big)^k \Big[\frac{1}{m!}\varrho^{(m)}(\xi_0)\Big(\frac{2}{1-z}\Big)^m \Big(\frac{1}{1-z_0}\Big)^m (z-z_0)^m + \ldots\Big]. \end{aligned} \tag{4.224}$$

Z vyjádření (4.224) je však vidět, že koeficient u $(z-z_0)^m$ je od nuly různý. První část našeho tvrzení je tedy dokázána. Druhá část se dokáže zcela analogicky.

Lemma 4.7. *Buď -1 jednoduchý kořen polynomu ϱ. Pak polynom r definovaný rovnicí (4.219) je stupně $k-1$. Naopak, nechť polynom r je stupně $k-1$. Pak polynom ϱ daný vzorcem (4.222) má jednoduchý kořen -1.*

D ů k a z . Pro polynom ϱ platí

$$\varrho(\xi) = \frac{1}{1!}\varrho'(-1)(\xi+1) + \ldots + \frac{1}{k!}\varrho^{(k)}(-1)(\xi+1)^k. \tag{4.225}$$

Je tedy

$$r(z) = \Big(\frac{1-z}{2}\Big)^k \Big[\frac{1}{1!}\varrho'(-1)\frac{2}{1-z} + \ldots + \frac{1}{k!}\varrho^{(k)}(-1)\Big(\frac{z}{1-z}\Big)^k\Big]. \tag{4.226}$$

Odtud však už plyne první část tvrzení. Důkaz druhé části tvrzení je stejně snadný.

Lemma 4.8. *Nechť polynom ϱ splňuje podmínky stability. Pak pro kořeny z_i polynomu r definovaného rovnicí* (4.219) *platí* $\mathrm{Re} z_i \leqq 0$ *a ty kořeny z_i, pro které je* $\mathrm{Re} z_i = 0$, *jsou jednoduché.*

D ů k a z plyne ihned z lemmat 4.6 a 4.7.

Lemma 4.9. *Nechť jsou splněny předpoklady lemmatu 4.8 a nechť je $\varrho(1) = 0$. Pak platí*

$$r(z) = a_1 z + \ldots + a_k z^k, \tag{4.227}$$

kde $a_1 \neq 0$ a všechny ostatní koeficienty a_i mají stejná znaménka jako a_1 nebo jsou to nuly.

D ů k a z . Podle lemmatu 4.6 platí $r(0) = 0$, $r'(0) \neq 0$, tedy platí (4.227) a $a_1 \neq 0$. Abychom dokázali platnost dalšího tvrzení, označme z_i kořeny polynomu r. Položíme-li $z_j = -x_j + \mathrm{i}y_j$, je $x_j \geqq 0$ podle lemmatu 4.8 (polynom splňuje podmínky stability). Platí tedy

$$r(z) = a_l z \prod_i (z + x_i) \prod_j [(z + x_j)^2 + y_j^2], \tag{4.228}$$

kde index i probíhá hodnoty příslušné reálným kořenům polynomu r, j hodnoty příslušné párům komplexně sdružených kořenů polynomu a a_l je nejvyšší nenulový koeficient v rozvoji (4.227). Všechny koeficienty v rozvoji (4.227) mají tedy stejná znaménka jako a_l.

Lemma 4.10. *Buďte*

$$f(t) = \sum_{k=0}^{\infty} A_k t^k \tag{4.229}$$

a

$$g(t) = \sum_{k=0}^{\infty} B_k t^k \tag{4.230}$$

dvě mocninné řady s kladnými poloměry konvergence a nechť platí

$$f(t)g(t) = 1. \tag{4.231}$$

Nechť dále platí $A_k > 0$ pro $k = 0, 1, \ldots$ a $A_{k+1}A_{k-1} - A_k^2 > 0$ pro $k = 1, 2, \ldots$. Pak platí $B_k < 0$ pro $k = 1, 2, \ldots$.

D ů k a z . Z předpokladů $A_k > 0$, $k = 0, 1, \ldots$ a $A_{k+1}A_{k-1} - A_k^2 > 0$, $k = 1, 2, \ldots$ plyne, že je $A_{k+1}/A_k > A_k/A_{k-1}$ pro $k = 1, 2, \ldots$. Posloupnost A_{k+1}/A_k je tedy rostoucí, takže platí

$$\frac{A_{n+1}}{A_n} > \frac{A_{n-k+1}}{A_{n-k}} \tag{4.232}$$

pro $k = 1, 2, \ldots, n$. Z identity (4.231) plyne, že je $A_0B_0 = 1$, $A_1B_0 + A_0B_1 = 0, \ldots$, obecně

$$A_nB_0 + A_{n-1}B_1 + \ldots + A_0B_n = 0. \tag{4.233}$$

Je tedy $B_0 = 1/A_0 > 0$, $B_1 = -A_1B_0/A_0 < 0$. Vynásobíme-li rovnici (4.233) číslem A_{n+1} a rovnici, která vznikne z rovnice (4.233) záměnou n za $n+1$ číslem $-A_n$, a vzniklé rovnice sečteme, dostaneme

$$\begin{aligned} B_{n+1} = {} & \frac{1}{A_0A_n}\big[(A_{n+1}A_{n-1} - A_nA_n)B_1 + (A_{n+1}A_{n-2} - A_nA_{n-1})B_2 + \\ & + \ldots + (A_{n+1}A_{n-k} - A_nA_{n-k+1})B_k + \ldots + (A_{n+1}A_0 - A_nA_1)B_n\big] = \\ = {} & \frac{A_{n-1}}{A_0}\Big(\frac{A_{n+1}}{A_n} - \frac{A_n}{A_{n-1}}\Big)B_1 + \ldots + \frac{A_{n-k}}{A_0}\Big(\frac{A_{n+1}}{A_n} - \frac{A_{n-k+1}}{A_{n-k}}\Big) + \\ & + \ldots + \frac{A_0}{A_0}\Big(\frac{A_{n+1}}{A_n} - \frac{A_1}{A_0}\Big)B_n. \end{aligned} \tag{4.234}$$

Nyní už dokážeme požadované tvrzení snadno úplnou indukcí. Předpokládáme-li totiž, že platí $B_k < 0$ pro $k = 1, \ldots, n$, plyne ihned z (4.232) a (4.234), že je $B_{n+1} < 0$. Lemma je dokázáno.

Lemma 4.11. *Pro $|z| < 1$ je*

$$\frac{z}{\ln\frac{1+z}{1-z}} = c_0 + c_2 z^2 + \ldots \tag{4.235}$$

a platí

$$c_{2i} < 0, \quad i = 1, 2, \ldots \tag{4.236}$$

D ů k a z . Funkce na levé straně rovnice (4.235) je zřejmě holomorfní v okolí bodu $z = 0$, takže ji lze v okolí tohoto bodu rozvinout v mocninou řadu. Protože je to navíc funkce sudá, je rozvoj tvaru (4.235). Abychom dokázali nerovnosti (4.236), položme v lemmatu 4.10

$$f(t) = \frac{1}{t^{1/2}} \ln\frac{1+t^{1/2}}{1-t^{1/2}} = 2\sum_{k=0}^{\infty}\frac{1}{2k+1}t^k \tag{4.237}$$

a

$$g(t) = \frac{t^{1/2}}{\ln \frac{1+t^{1/2}}{1-t^{1/2}}} = c_0 + c_2 t + c_4 t^2 + \dots . \tag{4.238}$$

Nyní už se snadno zjistí přímým výpočtem, že všechny předpoklady lemmatu 4.10 jsou splněny, takže požadované tvrzení plyne z tohoto lemmatu.

Uvedená lemmata už umožní poměrně snadno dokázat věty 4.11 a 4.12. Důkaz provedeme pro obě tyto věty zároveň. Buď tedy dána lineární k-kroková stabilní metoda řádu $p \geqq 1$. Definujeme-li funkci f předpisem

$$f(z) = \Big(\frac{1-z}{2}\Big)^k \varphi\Big(\frac{1+z}{1-z}\Big) = \frac{r(z)}{\ln \frac{1+z}{1-z}} - s(z), \tag{4.239}$$

kde

$$s(z) = \Big(\frac{1-z}{2}\Big)^k \sigma\Big(\frac{1+z}{1-z}\Big) \tag{4.240}$$

a φ je funkce definovaná vzorcem (4.212), plyne ihned z věty 4.9, že nula je p-násobným kořenem funkce f. Pro polynom s musí proto podle věty 4.10 platit

$$s(z) = b_0 + b_1 z + \dots + b_{p-1} z^{p-1}, \tag{4.241}$$

kde b_i jsou koeficienty rozvoje funkce $r(z)/\ln \frac{1+z}{1-z}$, tj.

$$b_0 + b_1 z + \dots = \frac{r(z)}{\ln \frac{1+z}{1-z}}. \tag{4.242}$$

Stupeň polynomu s však smí být nejvýše k. Proto, je-li $p \leqq k+1$, lze polynom s zvolit k libovolnému polynomu r. Je-li však $p > k+1$, musí být polynom r takový, že platí

$$b_{k+1} = b_{k+2} = \dots = b_{p-1} = 0 \tag{4.243}$$

(srv. poznámku 4.6). Na otázku po existenci stabilní k-krokové metody řádu většího než $k+1$ tedy odpovíme, sestrojíme-li polynom r tak, aby splňoval podmínky stability a aby pro něj platily rovnice (4.243). Vyjádřeme proto koeficienty b_i pomocí koeficientů polynomu r (které označíme jako v lemmatu 4.9 a_i). Položíme-li ještě $a_i = 0$ pro $i > k$, platí podle lemmatu 4.9 (je $p \geqq 1$, a tedy $\varrho(1) = 0$) a podle vzorce (4.242)

$$\sum_{i=0}^{\infty} a_{i+1} z^i \sum_{j=0}^{\infty} c_{2j} z^{2j} = \sum_{\nu=0}^{\infty} b_\nu z^\nu, \tag{4.244}$$

kde koeficienty c_{2j} jsou dány rovnicí (4.235). Porovnáním koeficientů u stejných mocnin na levé a pravé straně identity (4.244) dostaneme

$$\begin{aligned} b_{2\nu} &= a_{2\nu+1} c_0 + a_{2\nu-1} c_2 + \dots + a_1 c_{2\nu}, \\ b_{2\nu+1} &= a_{2\nu+2} c_0 + a_{2\nu} c_2 + \dots + a_2 c_{2\nu}. \end{aligned} \tag{4.245}$$

Podle lemmatu 4.11 je $c_{2i} < 0$ pro $i = 1, 2, \dots$. Vzhledem k lemmatu 4.9 můžeme předpokládat, že platí $a_1 > 0$ a $a_i \geqq 0$ pro $i = 2, 3, \dots$, neboť v opačném případě bychom místo metody dané polynomy ϱ a σ uvažovali metodu danou polynomy $-\varrho$ a $-\sigma$.

Buď nejprve k liché. Pak vzhledem k tomu, co jsme právě řekli, je

$$b_{k+1} = a_k c_2 + \dots + a_1 c_{k+1} < 0. \tag{4.246}$$

Žádnou volbou polynomu r tedy nelze dosáhnout toho, aby platilo $b_{k+1} = 0$. Odtud však už ihned plyne věta 4.11.

Buď nyní k sudé. Pak platí

$$b_{k+1} = a_k c_2 + \dots + a_2 c_k. \tag{4.247}$$

K tomu, aby mohlo být $b_{k+1} = 0$, je tedy nutné a stačí, aby $a_2 = a_4 = \dots = a_k = 0$. V tomto případě je polynom r lichá funkce, tj. platí $r(z) = -r(-z)$. Protože polynom r nemůže mít v důsledku stability kořeny v pravé polorovině, musí tedy být jeho kořeny ryze imaginární (a nula). Polynom ϱ musí mít proto kořeny v absolutní hodnotě rovny jedné (viz lemma 4.6). Protože je $a_k = 0$, je polynom r stupně $k-1$; -1 je tedy kořenem polynomu ϱ (viz lemma 4.7). Pro koeficient b_{k+2} platí

$$b_{k+2} = a_{k-1} c_4 + \dots + a_1 c_{k+2} < 0, \tag{4.248}$$

takže žádnou další speciální volbou polynomu r nelze docílit, aby řád byl větší než $k+2$. Tím je dokončen důkaz věty 4.12.

Z věty 4.12 vidíme, že všechny kořeny polynomu ϱ optimální k-krokové metody se sudým k jsou podstatné ve smyslu odst. 4.2.3. Může se tedy stát, že některé z těchto metod jsou slabě nestabilní. Následující věta ukazuje, že je tomu tak pro všechny tyto metody.

Věta 4.13. *Buď dána k-kroková metoda se sudým k, která je optimální ve smyslu věty 4.12. Pak pro růstový parametr λ příslušný ke kořenu $\xi = -1$ polynomu ϱ platí*

$$\lambda \leqq -\tfrac{1}{3}. \tag{4.249}$$

Důkaz. Je

$$\lambda = -\frac{\sigma(-1)}{\varrho'(-1)}. \tag{4.250}$$

Vyjádřeme hodnoty $\varrho'(\xi)$ a $\sigma(\xi)$ pomocí hodnot polynomů r a s definovaných vzorci (4.219) a (4.240). Protože zřejmě platí $\varrho(\xi) = (1+\xi)^k r\big((\xi-1)/(\xi+1)\big)$ a $\sigma(\xi) = (1+\xi)^k s\big((\xi-1)/(\xi+1)\big)$, je $\varrho'(\xi) = k(1+\xi)^{k-1} r\big((\xi-1)/(\xi+1)\big) + 2(1+\xi)^{k-2} r'\big((\xi-1)/(\xi+1)\big)$. Odtud máme

$$\varrho'(-1) = \lim_{\xi \to -1} \Big[k(1+\xi)^{k-1} r\Big(\frac{\xi-1}{\xi+1}\Big) + 2(1+\xi)^{k-2} r'\Big(\frac{\xi-1}{\xi+1}\Big)\Big]. \tag{4.251}$$

Protože polynom r je stupně $k-1$, dostáváme z tohoto vzorce, že je

$$\varrho'(-1) = (-1)^{k-1}2^{k-1}ka_{k-1} + (-1)^{k-2}2^{k-1}(k-1)a_{k-1} = \\ = (-1)^{k-1}2^{k-1}a_{k-1}. \tag{4.252}$$

Analogicky dostáváme, že je

$$\sigma(-1) = (-1)^k 2^k b_k. \tag{4.253}$$

Celkem je tedy podle (4.250)

$$\lambda = \frac{2b_k}{a_{k-1}}. \tag{4.254}$$

Dosadíme-li do této rovnice podle (4.245), máme

$$\lambda = \frac{2}{a_{k-1}}(c_2 a_{k-1} + c_4 a_{k-3} + \ldots + c_k a_1) \leqq 2c_2, \tag{4.255}$$

neboť čísla c_{2i} jsou záporná a a_i nezáporná. Tvrzení věty nyní plyne z toho, že je $c_2 = -1/6$, jak se snadno zjistí ze vzorce (4.235).

4.3 Užití lineárních mnohokrokových metod

Užijeme-li k řešení dané diferenciální rovnice implicitní metodu, je nutné přibližné řešení počítat z obecně nelineární rovnice. Proto otázka, jak to učinit co nejefektivněji, je značně důležitá. Než se na ni pokusíme odpovědět, všimneme si poněkud podrobněji, k čemu vůbec implicitní metody potřebujeme, když u explicitních metod zmíněné problémy vůbec nenastanou. Z předchozího výkladu víme, že maximální dosažitelný řád D-stabilní k-krokové metody je u explicitní metody nižší než u implicitní metody. Vypočítat více počátečních hodnot, které vyžaduje užití metody s větší hodnotou k, však nepředstavuje žádný vážný problém. Proto není vidět žádný na první pohled patrný důvod, proč nedat přednost explicitní metodě s větší hodnotou k, která má stejný řád jako uvažovaná implicitní metoda. Implicitní metody však mají další výhody proti explicitním metodám, než jen to, že mají při daném k vyšší řád. Jako příklad porovnáme z několika dalších hledisek explicitní Adamsovy-Bashforthovy metody a implicitní Adamsovy-Moultonovy metody. Výsledky jsou shrnuty v tabulce 4.10, v níž p značí řád metody, C_{p+1} konstantu z definice 4.2 (což je v tomto případě konstanta chyby ve smyslu vzorce (4.167), neboť u Adamsových metod je $\varrho'(1) = 1$) a $(\alpha, 0)$ je interval absolutní stability. Z tabulky je vidět, že implicitní Adamsova metoda téhož řádu jako explicitní metoda má menší konstantu chyby a větší interval absolutní stability. Tak např. tříkroková implicitní metoda má interval absolutní stability 10krát větší a konstantu chyby asi 13krát menší než čtyřkroková explicitní metoda, která je stejného řádu. Toto pozorování je typické i pro daleko obecnější případy a pro explicitní metody

natolik nevýhodné, že se jako samostatné metody užívají v praxi jen zřídka. Hrají však důležitou roli v souvislosti s tzv. metodami prediktor-korektor, které nyní popíšeme.

Tabulka 4.10

Porovnání Adamsových explicitních a implicitních metod

	Adams-Bashforth			Adams-Moulton		
p	2	3	4	2	3	4
k	2	3	4	1	2	3
C_{p+1}	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{120}$
α	-1	$-\frac{6}{11}$	$-\frac{3}{10}$	$-\infty$	-6	-3

4.3.1 Metody prediktor-korektor

Při užití implicitní mnohokrokové metody je třeba v každém kroku řešit rovnici (4.58) pro neznámou y_{n+k}, což se většinou provede postupnými aproximacemi. V každém kroku těchto iterací je třeba vypočítat hodnotu pravé strany dané diferenciální rovnice. Protože výpočet hodnoty funkce f spotřebuje většinou mnohem více výpočetního času než všechny ostatní početní operace, které je třeba v jednom kroku provést, snažíme se minimalizovat počet potřebných iterací, tj. snažíme se volit počáteční aproximaci $y_{n+k}^{(0)}$ tak přesnou, jak je to jen možné. Toho dosáhneme tak, že k jejímu určení použijeme vhodnou explicitní metodu. Explicitní metoda užitá k tomuto cíli se nazývá *prediktor* a původní implicitní metoda, kterou užijeme v iteracích, se nazývá *korektor*.

Postupovat můžeme dvojím způsobem. První způsob spočívá v tom, že v iteracích pokračujeme až do „konvergence", což v praxi znamená tak dlouho, až je splněno nějaké kritérium typu $|y_{n+k}^{(s+1)} - y_{n+k}^{(s)}| < \varepsilon$, kde ε je předem daná tolerance (srovnatelná nejčastěji s velikostí lokální zaokrouhlovací chyby). Získanou hodnotu $y_{n+k}^{(s+1)}$ pak pokládáme za přijatelnou aproximaci přesné hodnoty y_{n+k}. Při tomto způsobu nelze předem udat počet hodnot funkce f potřebných k provedení jednoho kroku metody. Získaná hodnota $y_{n+k}^{(s+1)}$ však nezávisí na počáteční aproximaci, takže lokální chyba i charakter stability metody jsou určeny pouze korektorem.

Při druhém způsobu provádíme v každém kroku vždy pevný počet iterací, takže počet funkčních hodnot potřebných k provedení jednoho kroku je předem znám. Při tomto postupu už není pravda, že stabilita a lokální chyba jsou určeny pouze korektorem, což může mít podstatný vliv na výběr dvojice prediktor-korektor.

Nyní popíšeme naznačený postup podrobněji. Nechť explicitní metoda užitá jako prediktor je metoda

$$\sum_{\nu=0}^{k} \alpha_\nu^* y_{n+\nu} = h \sum_{\nu=0}^{k-1} \beta_\nu^* f(x_{n+\nu}, y_{n+\nu}) \tag{4.256}$$

a implicitní metoda užitá jako korektor je metoda (4.57). Protože se často ukazuje výhodné užít prediktor stejného řádu, jako je řád korektoru a protože, jak víme, u implicitní metody lze dosáhnout při daném k vyššího řádu než u explicitní metody, je často vhodné použít prediktor s větší hodnotou čísla k, než je příslušná hodnota u korektoru. Abychom mohli formálně jak v prediktoru, tak v korektoru užívat tutéž hodnotu čísla k, dohodneme se, že v tomto odstavci upustíme od požadavku, že čísla α_0 a β_0 nejsou současně rovna nule, takže nyní připouštíme v korektoru i případ, že je $\alpha_0 = \beta_0 = 0$. Nechť dále $m \geqq 1$ je pevně zvolené číslo. Pak za přibližné řešení v bodě x_{n+k} získané užitím prediktoru a m-násobným užitím korektoru budeme pokládat číslo $y_{n+k}^{(m)}$ vypočtené pomocí rekurencí

$$\begin{aligned} &\alpha_k^* y_{n+k}^{(0)} + \sum_{\nu=0}^{k-1} \alpha_\nu^* y_{n+\nu}^{(m)} = h \sum_{\nu=0}^{k-1} \beta_\nu^* f(x_{n+\nu}, y_{n+\nu}^{(m)}), \\ &\alpha_k y_{n+k}^{(s+1)} + \sum_{\nu=0}^{k-1} \alpha_\nu y_{n+\nu}^{(m)} + h\beta_k f(x_{n+k}, y_{n+k}^{(s)}) + \\ &\qquad + h \sum_{\nu=0}^{k-1} \beta_\nu f(x_{n+\nu}, y_{n+\nu}^{(m)}), \quad s = 0, \ldots, m-1. \end{aligned} \tag{4.257}$$

V praxi se zřídka užívá hodnoty m větší než 2.

Pod metodou prediktor-korektor se většinou v literatuře rozumí právě popsaná metoda, někdy dokonce jen její speciální případ s $m = 1$. Žádná z těchto metod nespadá do třídy lineárních k-krokových metod, neboť veličiny $y^{(m)}$ a $f(x, y^{(m)})$ v nich nevystupují lineárně. Předchozí teorii tedy nelze bezprostředně užít a má-li mít metoda (4.257) rozumný smysl, musíme vyšetřit její konvergenci. Omezíme se přitom na případ $m = 1$.

Věta 4.14. *Buď dána D-stabilní mnohokroková metoda* (4.57) *řádu $p \geqq 1$ a explicitní metoda* (4.256) *řádu $p^* \geqq 0$. Nechť jsou dále splněny předpoklady* (i) *a* (ii) *z čl. 1 a nechť přesné řešení je dostatečně hladké. Nechť konečně počáteční podmínky jsou dány s přesností $O(h^r)$, kde $r = \min(p^* + 1, p)$, nebo vyšší. Pak pro přibližné řešení $y_n^{(1)}$ vypočtené metodou prediktor-korektor* (4.257) *s $m = 1$ platí pro $h \to 0$, $x_n = x$*

$$y_n^{(1)} - y(x_n) = O(h^r). \tag{4.258}$$

D ů k a z . Protože prediktor je řádu p^* a korektor řádu p, platí pro každou

dostatečně hladkou funkci y, a tedy také pro přesné řešení

$$\sum_{\nu=0}^{k} \alpha_\nu^* y(x_{n+\nu}) - h \sum_{\nu=0}^{k-1} \beta_\nu^* y'(x_{n+\nu}) = C_{p^*+1}^* y^{(p^*+1)}(x_n) h^{p^*+1} + O(h^{p^*+2}) \tag{4.259}$$

a

$$\sum_{\nu=0}^{k} \alpha_\nu y(x_{n+\nu}) - h \sum_{\nu=0}^{k} \beta_\nu y'(x_{n+\nu}) = C_{p+1} y^{(p+1)}(x_n) h^{p+1} + O(h^{p+2}). \tag{4.260}$$

Zavedeme-li veličiny e_n a $e_n^{(0)}$ rovnicemi $e_n = y_n^{(1)} - y(x_n)$ a $e_n^{(0)} = y_n^{(0)} - y(x_n)$, je

$$\begin{aligned} &\alpha_k^* e_{n+k}^{(0)} + \sum_{\nu=0}^{k-1} \alpha_\nu^* e_{n+\nu} = \\ &\quad = h \sum_{\nu=0}^{k-1} \beta_\nu^* g_{n+\nu} e_{n+\nu} - C_{p^*+1}^* y^{(p^*+1)}(x_n) h^{p^*+1} + O(h^{p^*+2}) \end{aligned} \tag{4.261}$$

a

$$\begin{aligned} \sum_{\nu=0}^{k} \alpha_\nu e_{n+\nu} &= h\beta_k g_{n+k}^{(0)} e_{n+k}^{(0)} + h \sum_{\nu=0}^{k-1} \beta_\nu g_{n+\nu} e_{n+\nu} - \\ &\quad - C_{p+1} y^{(p+1)}(x_n) h^{p+1}(x_n) h^{p+1} + O(h^{p+2}), \end{aligned} \tag{4.262}$$

kde jsme položili

$$g_n = \begin{cases} \frac{f(x_n, y_n^{(1)}) - f(x_n, y(x_n))}{e_n} & \text{pro } e_n \neq 0 \\ 0 & \text{pro } e_n = 0 \end{cases} \tag{4.263}$$

a

$$g_n^{(0)} = \begin{cases} \frac{f(x_n, y_n^{(0)}) - f(x_n, y(x_n))}{e_n^{(0)}} & \text{pro } e_n^{(0)} \neq 0 \\ 0 & \text{pro } e_n^{(0)} = 0. \end{cases} \tag{4.264}$$

Vyloučíme-li z rovnic (4.261) a (4.262) veličinu $e_{n+k}^{(0)}$, dostaneme

$$\begin{aligned} &\sum_{\nu=0}^{k} \alpha_\nu e_{n+\nu} - h \sum_{\nu=0}^{k-1} \left[\beta_\nu g_{n+\nu} + \frac{\beta_k}{\alpha_k^*} g_{n+k}^{(0)} (-\alpha_\nu^* + h\beta_\nu^* g_{n+\nu}) \right] e_{n+\nu} = \\ &\quad = -\frac{\beta_k}{\alpha_k^*} g_{n+k}^{(0)} C_{p^*+1}^* y^{(p^*+1)}(x_n) h^{p^*+2} - C_{p+1} y^{(p+1)}(x_n) h^{p+1} + \\ &\quad + O(h^{p^*+3}) + O(h^{p+2}) = O(h^{r+1}). \end{aligned} \tag{4.265}$$

Protože polynom daný koeficienty α_ν splňuje podmínky D-stability a protože chyba v počátečních podmínkách je řádu $O(h^r)$, plyne tvrzení věty ihned z lemmatu 4.3.

Z právě dokázané věty vidíme, že ke konvergenci metody prediktor-korektor v námi uvažovaném případě $m = 1$ není vůbec třeba, aby prediktor byl D-stabilní.

Dále je rovněž vidět, že k tomu, abychom dosáhli touto metodou stejnou řádovou přesnost, jako kdybychom použili pouze implicitní metodu danou korektorem (tj. kdybychom iterovali do konvergence), může být řád prediktoru o jedničku menší než řád korektoru.

Poznámka 4.7. V případě obecného m zůstává věta 4.14 v platnosti, položíme-li $r = \min(p^* + m, p)$. Důkaz je úplně stejný.

Uvědomíme-li si, že technika, kterou jsme užili k důkazu věty 4.14, je zcela analogická, jako technika užitá v podobné situaci v případě mnohokrokové metody, je zřejmé, že pro metodu prediktor-korektor platí i tvrzení analogická k ostatním tvrzením, která jsme dokázali v případě lineární mnohokrokové metody. Speciálně aplikujeme-li metodu (4.257), opět s $m = 1$, na rovnici $y' = Ay$, dostaneme pro přibližné řešení $y_n^{(1)}$ rovnici

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu}^{(1)} = hA \sum_{\nu=0}^{k-1} \left[\beta_\nu + \frac{\beta_k}{\alpha_k^*}(-\alpha_\nu^* + hA\beta_\nu^*)\right] y_{n+\nu}^{(1)}. \tag{4.266}$$

To je opět lineární diferenční rovnice s konstantními koeficienty. Platí tedy vzorec zcela analogický ke vzorci (4.205), jen růstové parametry jsou nyní jiné. Čísla λ_i ve vzorci (4.205) jsou totiž definována rovnicí (4.155), kde $\tilde{\xi}_i$ je kořen charakteristického polynomu $\tilde{\varrho}$ diferenční rovnice (4.145) a ξ_i je kořen polynomu ϱ. Růstové parametry $\lambda_i^{(1)}$ uvažované metody prediktor-korektor jsou tedy definovány rovnicí

$$\tilde{\xi}_i^{(1)} = \xi_i[1 + \lambda_i^{(1)} hA + O(h^2)], \quad i = 1, \ldots, m, \tag{4.267}$$

kde $\tilde{\xi}_i^{(1)}$ je kořen polynomu $\tilde{\varrho}^{(1)}$ daného rovnicí

$$\begin{aligned}\tilde{\varrho}^{(1)}(\xi) &= \varrho(\xi) - hA\left\{\sigma(\xi) - \beta_k\xi^k + \frac{\beta_k}{\alpha_k^*}[-\varrho^*(\xi) + \alpha_k^*\xi^k + hA\sigma^*(\xi)]\right\} = \\ &= \varrho(\xi) - hA\left[\sigma(\xi) - \frac{\beta_k}{\alpha_k^*}\varrho^*(\xi) + hA\frac{\beta_k}{\alpha_k^*}\sigma^*(\xi)\right]\end{aligned} \tag{4.268}$$

a m je počet podstatných kořenů polynomu ϱ. Postupem zcela analogickým jako v odst. 4.2.3 plyne z rovnice (4.267) a (4.268), že platí

$$\lambda_i^{(1)} = \frac{\sigma(\xi_i) - \frac{\beta_k}{\alpha_k^*}\varrho^*(\xi_i)}{\xi_i\varrho'(\xi_i)}, \tag{4.269}$$

Přítomnost prediktoru mění tedy růstové parametry korektoru, a tím také asymptotické chování chyby. V následujícím příkladě ukážeme, jak lze této skutečnosti využít.

Příklad 4.3. Uvažujeme Milnovu-Simpsonovu metodu (srv. odst. 4.1.5)

$$y_{n+2} - y_n = \tfrac{1}{3}h(f_{n+2} + 4f_{n+1} + f_n). \tag{4.270}$$

Tato metoda je řádu 4, je tedy optimální ve smyslu věty 4.12 a podle věty 4.13 je slabě nestabilní. (Přímým výpočtem zjistíme, že je $\lambda_2 = -1/3$.) Její užití může být v případě některých rovnic nebezpečné. Přibereme-li k metodě (4.270) metodu (4.67) jako prediktor a korigujeme-li pouze jednu, dostaneme metodu prediktor-korektor řádu 4, neboť metoda (4.67) je řádu 3. To, že metoda (4.67) není samostatně použitelná, neboť není D-stabilní, u prediktoru podle věty 4.14 nevadí. Ze vzorce (4.269) snadno vypočteme, že růstový parametr $\lambda_2^{(1)}$ právě sestrojené metody je roven jedné. Přítomnost vhodného prediktoru odstranila tedy slabou nestabilitu Milnovy-Simpsonovy metody (4.270). V tabulce 4.11 jsou uvedeny výsledky konkrétního numerického řešení diferenciální rovnice $y' = 1 - y^2$ s počáteční podmínkou $y(0) = 5$ metodou (4.270) a metodou prediktor-korektor (4.67) a (4.270).

Tabulka 4.11

Řešení diferenciální rovnice $y' = 1 - y^2$

	Milnova-Simpsonova metoda		Metoda prediktor-korektor (4.67) a (4.270)	
	y_n	e_n	y_n	e_n
⋮				
190	1,003 473 460 3	−0,000 050 896 6	1,003 524 370 8	0,000 000 013 9
191	1,003 467 147 6	0,000 051 409 0	1,003 415 750 4	0,000 000 011 8
192	1,003 258 502 6	−0,000 051 970 9	1,003 310 486 5	0,000 000 013 1
⋮				
318	0,999 871 182 1	−0,000 193 257 3	1,000 064 439 6	0,000 000 000 3
319	1,000 257 714 5	0,000 195 257 8	1,000 062 456 9	0,000 000 000 2
320	0,999 863 209 2	−0,000 197 325 9	1,000 060 535 3	0,000 000 000 2

V obou případech byl užit integrační krok $h = 1/64$ a hodnota y_1 byla vzata z přesného řešení. Vidíme, že výsledky jsou i v nelineárním případě takové, jak jsme teoreticky očekávali.

Závěrem tohoto odstavce ještě poznamenejme, že metoda daná rovnicemi

$$\begin{aligned}\alpha_k^* y_{n+k}^{(0)} + \sum_{\nu=0}^{k-1} \alpha_\nu^* y_{n+\nu}^{(m)} &= h\sum_{\nu=0}^{k-1} \beta_\nu^* f(x_{n+\nu}, y_{n+\nu}^{(m-1)}), \\ \alpha_k y_{n+k}^{(s+1)} + \sum_{\nu=0}^{k-1} \alpha_\nu y_{n+\nu}^{(m)} &= h\beta_k f(x_{n+k}, y_{n+k}^{(s)}) + \\ &\quad + h\sum_{\nu=0}^{k-1} \beta_\nu f(x_{n+\nu}, y_{n+\nu}^{(m-1)}), \quad s = 0, \ldots, m-1,\end{aligned} \tag{4.271}$$

se rovněž nazývá metoda prediktor-korektor. Rozdíl mezi oběma popsanými variantami je v tom, že v každém kroku užíváme v bodech, v nichž už je přibližné řešení vypočteno, v případě metody (4.257) hodnoty $f(x_r, y_r^{(m)})$, zatímco v případě metody (4.271) hodnoty $f(x_r, y_r^{(m-1)})$. Při způsobu daném rovnicemi (4.257) je tedy nutno vypočítávat o jednu hodnotu pravé strany dané diferenciální rovnice více než při způsobu (4.271).

Abychom byli schopni stručně a jednoznačně popsat, který z uvedených způsobů máme v konkrétní situaci na mysli, zavedeme následující symbolické značení (které je v literatuře týkající se problematiky řešení diferenciálních rovnic už zcela standardní). Symbolem P označíme užití prediktoru, symbolem E výpočet hodnoty pravé strany a symbolem C užití korektoru (důvody k zavedení těchto symbolů jsou mnemotechnické a pocházejí od anglických slov prediction, evaluation a correction). Postup (4.271) s $m = 1$ se tedy užitím této symboliky zapíše jako PEC, s $m = 2$ jako $PECEC = P(EC)^2$ atd. obecně $P(EC)^m$. Postup (4.257) je pak označen $PECE$, $P(EC)^2E$ atd. Z uvedeného zápisu je také hned vidět, kolikrát je třeba v jednom kroku vypočítávat hodnotu pravé strany dané diferenciální rovnice.

Teoretické zkoumání metod prediktor-korektor v režimu $P(EC)^m$ je obdobné, jako tomu bylo u metod $P(EC)^mE$.

4.3.2 Volba integračního kroku

Jeden z nejnesnadnějších problémů, který je třeba rozřešit při konkrétní aplikaci mnohokrokové metody, je volba integračního kroku. Odhady celkové diskretizační chyby neposkytují k této volbě vhodné vodítko, neboť jsou skoro vždy natolik pesimistické, že vedou k zbytečně a často také neúnosně malé hodnotě integračního kroku. Proto se většinou užívá při řešení této otázky přístup, kdy integrační krok se volí tak, aby lokální diskretizační chyba byla přijatelně malá a aby číslo $z = h\partial f/\partial y$ leželo v intervalu stability užité metody, což zabrání nebezpečnému nakupení lokálních chyb v dalším průběhu výpočtu. Při odhadu lokální diskretizační chyby podle (4.141) vzniká značná praktická obtíž spočívající v nutnosti odhadu $(p+1)$-ní derivace hledaného řešení. Užijeme-li však vhodným způsobem metodu prediktor-korektor, můžeme se nutnosti odhadu vyšších derivací vyhnout užitím postupu, který pochází od W. E. Milneho. Popišme jej v případě metody $PECE$ a za zjednodušujícího předpokladu, že hodnoty $y_n, \dots, y_{n+k-1}$ vypočtené v předešlých krocích jsou rovny hodnotám přesného řešení. Za tohoto předpokladu je přibližné řešení dáno rovnicemi

$$\alpha_k^* y_{n+k}^{(0)} + \sum_{\nu=0}^{k-1} \alpha_\nu^* y(x_{n+\nu}) = h \sum_{\nu=0}^{k-1} \beta_\nu^* f\big(x_{n+\nu}, y(x_{n+\nu})\big), \tag{4.272}$$

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

$$+ h \sum_{\nu=0}^{k-1} \beta_\nu f\big(x_{n+\nu}, y(x_{n+\nu})\big).$$

Předpokládáme-li, že prediktor a korektor mají stejný řád p, platí (pro každé dostatečně hladké řešení dané diferenciální rovnice)

$$\sum_{\nu=0}^{k} \alpha_\nu^* y(x_{n+\nu}) = h \sum_{\nu=0}^{k-1} \beta_\nu^* f\big(x_{n+\nu}, y(x_{n+\nu})\big) + \tag{4.273}$$

$$+ C_{p+1}^* y^{(p+1)}(x_n) h^{p+1} + O(h^{p+2}),$$

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

$$+ C_{p+1} y^{(p+1)}(x_n) h^{p+1} + O(h^{p+2}).$$

Odečteme-li první rovnici (4.273) od první rovnice (4.272) a druhou rovnici (4.273) od druhé rovnice (4.272), dostáváme

$$\alpha_k^* [y_{n+k}^{(1)} - y(x_{n+k})] = -C_{p+1}^* y^{(p+1)}(x_n) h^{p+1} + O(h^{p+2}) \tag{4.274}$$

a

$$\alpha_k [y_{n+k}^{(1)} - y(x_{n+k})] = -C_{p+1} y^{(p+1)}(x_n) h^{p+1} + O(h^{p+2}) + \tag{4.275}$$

$$+ h\beta_k \big[f(x_{n+k}, y_{n+k}^{(0)}) - f\big(x_{n+k}, y(x_{n+k})\big)\big] =$$

$$= -C_{p+1} y^{(p+1)}(x_n) h^{p+1} + O(h^{p+2}).$$

Všimněme si, že vzorec (4.275) znovu dokazuje skutečnost známou nám už z důkazu věty 4.14, totiž, že hlavní část lokální chyby metody $PECE$ je za předpokladu rovnosti řádu prediktoru a korektoru stejná jako hlavní část lokální chyby korektoru.

Vyloučíme-li nyní z rovnice (4.274) a (4.275) hodnotu přesného řešení $y(x_{n+k})$, máme

$$y_{n+k}^{(1)} - y_{n+k}^{(0)} = \Big(\frac{C_{p+1}^*}{\alpha_k^*} - \frac{C_{p+1}}{\alpha_k}\Big) y^{(p+1)}(x_n) h^{p+1} + O(h^{p+2}). \tag{4.276}$$

Platí tedy celkem

$$y^{(p+1)}(x_n) = \Big(\frac{C_{p+1}^*}{\alpha_k^*} - \frac{C_{p+1}}{\alpha_k}\Big)^{-1} h^{-p-1} [y_{n+k}^{(1)} - y_{n+k}^{(0)}] + O(h), \tag{4.277}$$

což je zmíněný *Milnův vzorec*. Podtrhněme ještě jednou, že rovnost řádů prediktoru a korektoru je zde podstatný předpoklad.

Abychom byli schopni zvolit integrační krok tak, aby byla zaručena stabilita, potřebujeme odhadnout velikost čísla $\partial f/\partial y$ v každém kroku. To pak umožní vypočítat číslo $z = h\partial f/\partial y$ a posoudit, zda leží v předem známém intervalu absolutní (relativní) stability. Jedna z možností, jak to provést, vychází opět z metody

prediktor-korektor v režimu $PECE$ a spočívá v užití takřka zřejmého přibližného vzorce

$$z = h\frac{\partial f}{\partial y} \approx h\frac{f(x_{n+k}, y_{n+k}^{(1)}) - f(x_{n+k}, y_{n+k}^{(0)})}{y_{n+k}^{(1)} - y_{n+k}^{(0)}} \tag{4.278}$$

Tento postup však nelze jednoduše přenést na případ soustav diferenciálních rovnic.

4.3.3 Změna integračního kroku

V předešlém odstavci jsme stručně shrnuli požadavky, které je třeba klást na přípustný integrační krok. Aby výpočet probíhal co nejekonomičtěji, je žádoucí, aby se integrační krok volil co největší. Připomenutá kritéria však na něj mohou klást v různé fázi výpočtu nestejně přísná omezení. A nalezneme-li tedy možnost, jak v průběhu výpočtu integrační krok měnit, bude patrně výsledný algoritmus ekonomičtějším, než když se této možnosti vzdáme.

U Rungových-Kuttových metod změna integračního kroku nepředstavuje žádný technický problém, neboť provést jeden krok této metody znamená vlastně řešit novou úlohu. V případě k-krokových metod je situace odlišná, neboť k výpočtu přibližné hodnoty y_{n+k} v bodě x_{n+k} je zapotřebí znát přibližné hodnoty hledané funkce y a funkce f v bodech $x_{n+k-1}, \dots, x_n$. Ty mohou být už vypočteny v předchozích krocích, ale také tomu tak nemusí být. Jestliže např. před výpočtem přibližné hodnoty y_{n+k} integrační krok zdvojnásobíme (tak se to často dělá, dovolují-li velikost lokální chyby a požadavky stability zvětšení integračního kroku), jsou všechny potřebné hodnoty k dispozici, neboť jde o hodnoty v bodech $x_{n+k-1}, x_{n+k-3}, \dots, x_{n-k+1}$ a ve všech těchto bodech se přibližné řešení v předchozím průběhu výpočtu už počítalo. Některé z těchto hodnot si však nebylo třeba pamatovat při výpočtu s pevným integračním krokem, a tedy i při zvětšování integračního kroku mohou vzniknout jisté nevelké programovací obtíže. Jestliže však integrační krok před výpočtem hodnoty y_{n+k} rozpůlíme, potřebujeme znát hodnoty funkce y a f v bodech $x_{n+k-1}, x_{n+k-3/2}, \dots, x_{n+(k-1)/2}$, přičemž ty z nich, kde příslušný index není celé číslo, se v předchozím výpočtu nepočítaly. V dalším výkladu naznačíme tři způsoby, jak tyto dodatečné hodnoty získat.

Patrně nejjednodušší z programátorského hlediska je způsob, kdy se na tento problém díváme jako na speciální problém získání počátečních hodnot pro užití mnohokrokové metody. Pro výpočet potřebných hodnot funkce y použijeme tedy např. Rungovu-Kuttovu metodu a potřebné hodnoty funkce f pak dopočteme vypočtením hodnot pravé strany dané diferenciální rovnice.

Druhý způsob spočívá v dopočítání potřebných hodnot funkce y pomocí standardní interpolace. Samozřejmě je třeba volit interpolační vzorec tak, aby chyba interpolace byla stejného řádu, jako je lokální diskretizační chyba použité metody. Upozorněme, že v případě, že dvojice prediktor-korektor je tvořena Adamsovou-Bashforthovou a Adamsovou-Moultonovou metodou, je potřebná funkční hodnota

funkce y vždy známá, neboť jde o hodnotu y_{n+k-1}, takže lze interpolovat přímo funkci f, což ušetří počet potřebných výpočtů funkčních hodnot pravé strany dané diferenciální rovnice.

Konečně se používá postup, který pochází od A. Nordsiecka a jehož základní myšlenka spočívá v tom, že místo hodnot funkce y a f v několika bodech (které potřebujeme v dalším kroku metody) se uchovávají v paměti derivace lokálního interpolačního polynomu, který představuje přibližné řešení, v jediném bodě. Do dalších podrobností zde už nebudeme zacházet; upozorněme pouze, že tento postup je ve srovnání s předešlými z programátorského hlediska komplikovanější.

5 Porovnání mnohokrokových metod a Rungových-Kuttových metod

V tomto článku se velice stručně dotkneme závažného a z praktického hlediska nepominutelného problému výběru konkrétní metody k přibližnému řešení konkrétní matematické úlohy. Tento problém prolíná celou numerickou matematikou a chceme-li řešit jakoukoliv matematickou úlohu až od konce (tj. až do získání konkrétních čísel), musíme se s ním tak či onak vyrovnat. Poněkud paradoxně na první pohled je situace svým způsobem jednodušší v těch případech, kdy je úloha natolik složitá, že jsme rádi, že máme k jejímu řešení k dispozici vůbec nějakou metodu Existuje-li však celá třída metod a to ještě pro celou třídu úloh, je otázka výběru nejvhodnější metody nesmírně složitá a v současné době zatím matematicky obecně nerozřešená. Při prakticky nezbytném výběru konkrétní metody jsme tedy nuceni se řídit pouze určitými vodítky, která se budou ovšem opírat o matematicky přesně definované pojmy a rigorózně dokázaná tvrzení. To také je důvod důležitosti podrobného teoretického zkoumání přibližných metod a tomuto účelu je také podřízen obsah této knihy. Následující porovnání Rungových-Kuttových metod a metod prediktor-korektor užitých v režimu $PECE$ je proto nutno chápat pouze jako příklad postupu při praktické interpretaci teoretických výsledků, který ani zdaleka nevyčerpává zmíněnou problematiku.

Metody obou zmíněných skupin budeme porovnávat z hlediska lokální přesnosti (k jejíž hrubé charakteristice slouží řád metody), z hlediska charakteru stability, z hlediska pracnosti (měřené množstvím výpočtů hodnot pravé strany) a z hlediska složitosti algoritmu, a tím také složitosti programování.

Všimněme si nejprve lokální přesnosti metod, které mají stejný řád p. Má-li prediktor a korektor řád p, je hlavní část lokální diskretizační chyby metody prediktor-korektor dána výrazem

$$Cy^{(p+1)}(x_n)h^{p+1}, \tag{5.1}$$

kde $C = C_{p+1}/\varrho(1)$ je konstanta chyby korektoru. Hlavní část lokální diskretizační

chyby Rungovy-Kuttovy metody je tvaru

$$\varphi(x_n, y(x_n))h^{p+1}, \tag{5.2}$$

přičemž φ závisí na přesném řešení dané diferenciální rovnice podstatně komplikovaněji než v případě metody prediktor-korektor. Abychom to konkrétně dokumentovali, stačí se podívat na vzorec (3.65), v němž je uveden tvar této funkce pro standardní Rungovu-Kuttovu metodu čtvrtého řádu. Z uvedených vzorců je vidět, že přímé srovnání hlavních částí lokální diskretizační chyby není možné. Praktické zkušenosti však ukazují,k že lokální chyba Rungových-Kuttových metod čtvrtého řádu je většinou menší než lokální chyba metod prediktor-korektor téhož řádu. Toto srovnání se však může změnit, učiníme-li předpoklad, že budeme porovnávat metody stejného řádu za doplňující podmínky, že obě srovnávané metody užívají tentýž počet hodnot pravé strany dané diferenciální rovnice. Protože pro $p \leqq 4$ existují Rungovy-Kuttovy metody řádu p vyžadující p hodnot pravé strany a neexistují metody, u nichž je tento počet nižší, a protože metody prediktor-korektor užité v režim $PECE$ vyžadují dvě funkční hodnoty na jeden krok, lze při zachování stejné pracnosti užít u metod $PECE$ integrační krok, který je $(2/p)$-násobkem integračního kroku Rungovy-Kuttovy metody. Za těchto podmínek je třeba výraz pro hlavní část lokální diskretizační chyby metody $PECE$ nahradit výrazem

$$Cy^{(p+1)}(x_n)\Big(\frac{2h}{p}\Big)^{p+1}, \tag{5.3}$$

ze kterého plyne, že při tomto způsobu porovnání je lokální chyba metod prediktor-korektor už většinou menší než chyba Rungových-Kuttových metod.

Snaha porovnávat přesnost Rungovy-Kuttovy metody a metody prediktor-korektor pouze z hlediska počtu potřebných hodnot pravé strany bez ohledu na jejich řád nemá dobrý smysl, neboť u metod prediktor-korektor lze řád libovolně zvýšit, aniž by bylo třeba výpočtu dalších funkčních hodnot.

Velikost intervalů absolutní stability se mění s řádem u Rungových-Kuttových metod úplně jinak než u lineárních k-krokových metod. Snadno se zjistí, že interval absolutní stability Rungovy-Kuttovy metody řádu p užívající p hodnot pravé strany (takové metody existují jen pro $p \leqq 4$) závisí pouze na p a nikoliv na konkrétní metodě. Tyto intervaly jsou po řadě $(-2; 0)$, $(-2; 0)$, $(-2, 51; 0)$, $(-2, 78; 0)$ pro metodu prvního až čtvrtého řádu, a tedy se s rostoucím řádem poněkud zvětšují. Z tabulky 4.10 naproti tomu je vidět, že intervaly stability Adamsových metod se se zvětšováním řádu dosti podstatně zmenšují. Intervaly absolutní stability u Rungových-Kuttových metod jsou přitom podstatně příznivější než u explicitních Adamsových metod. I u jiných explicitních mnohokrokových metod je tomu podobně a to je také důvod, proč se tyto metody samostatně většinou nepoužívají. Dále je třeba si uvědomit, že užijeme-li implicitní metodu jako část metody prediktor-korektor, je tím její interval absolutní stability rovněž ovlivněn, většinou pak nepříznivě. Tak např. $PECE$ Adamsova-Bashforthova a Adamsova-Moultonova metoda čtvrtého řádu

má interval absolutní stability $(-1, 25; 0)$ oproti intervalu $(-3; 0)$ u čisté Adamsovy-Moultonovy metody. Interval absolutní stability této metody prediktor-korektor je tedy zhruba poloviční než u Rungovy-Kuttovy metody 4. řádu. Uvědomíme-li si však, že při stejné pracnosti metoda $PECE$ umožňuje užití poloviního integračního kroku než u standardní Rungovy-Kuttovy metody, je interval absolutní stability takové metody vlastně dvojnásobný. Potom jsou však standardní Rungova-Kuttova metoda a Adamsova-Bashforthova-Moultonova $PECE$ metoda z hlediska stability v podstatě rovnocenné. Kromě toho, co bylo právě řečeno, je třeba při porovnávání Rungových-Kuttových metod a mnohokrokových metod z hlediska stability vzít také v úvahu, že je-li daná úloha takového charakteru, že při jejím řešení vznikají vážné problémy se stabilitou, umožňují mnohokrokové metody mnohem více prostoru pro modifikace zlepšující stabilitu než Rungovy-Kuttovy metody. Tak např. iterujeme-li s Adamsovým-Moultonovým korektorem druhého řádu do konvergence, je příslušný interval absolutní stability $(-\infty; 0)$. Tedy i porovnání z hlediska stability spíše mluví pro metody prediktor-korektor.

Porovnání Rungových-Kuttových metod a mnohokrokových metod z hlediska složitosti programování vyzní naproti tomu dosti jednoznačně ve prospěch Rungových-Kuttových metod. Nepotřebují žádné speciální procedury na začátku výpočtu a změna velikosti integračního kroku v kterékoliv fázi výpočtu rovněž nepředstavuje technicky žádný problém. Chceme-li však řídit velikost integračního kroku na základě kritérií zformulovaných v odst. 4.3.2, je třeba mít možnost odhadnout velikost lokální diskretizační chyby a velikost derivace $\partial f/\partial y$. U metod prediktor-korektor se tyto odhady získají relativně jednoduše a hlavně lacino co do počtu potřebných operací (viz odst. 4.3.2). U Rungových-Kuttových metod je to sice také možné (hlavní část lokální chyby lze např. odhadnout metodou poloviního kroku analogicky, jako se to dělalo pro celkovou diskretizační chybu, viz cvič. 12, nebo lze k tomu cíli užít dvou metod různých řádů, viz Fehlbergova metoda (3.25)), zaplatí se však za to většinou dosti vysoká cena ve vzrůstu nároků na počet operací.

Shrneme-li výsledky porovnání, lze soudit, že zejména z hlediska univerzálnosti a možnosti řízení stability a velikosti lokální chyby je třeba dát metodám prediktor-korektor přednost před Rungovými-Kuttovými metodami. Jednoduchost logické struktury příslušného algoritmu je pak významnou předností Rungových-Kuttových metod, takže i tyto metody mohou být dosti užitečné zejména v případech malých nároků na přesnost.

Poznamenejme ještě, že v praktické situaci je výběr konkrétní metody často podstatně ovlivněn také tím, že je k dispozici skutečně fungující a ověřený program. Tento program pak většinou použijeme a spokojíme se tím, že výběr metody za nás vlastně už provedl autor programu. Předchozí text měl sloužit zejména k tomu, aby čtenáři alespoň naznačil, jakými směry by se měly ubírat úvahy při vytváření nových programů.

6 Soustavy diferenciálních rovnic a problematika silného tlumení

Máme-li řešit úlohu s počátečními podmínkami pro soustavu diferenciálních rovnic (1.1), lze v principu užít všechny metody, které jsme až dosud popsali a stačí přitom pouze předpokládat, že veličiny y_n a f_n jsou m-dimenzionální vektory. Vznikají však také některé rozdíly a některé nové problémy, které stručně popíšeme v následujícím odstavci.

6.1 Lineární mnohokrokové metody

Lineární k-kroková metoda pro řešení soustavy (1.1) je dána vzorcem

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = h \sum_{\nu=0}^{k} \beta_\nu f_{n+\nu}, \tag{6.1}$$

kde

$$y_n = ({}^1y_n, \ldots, {}^my_n)^T, \quad f_n = ({}^1f_n, \ldots, {}^mf_n)^T \tag{6.2}$$

a

$${}^if_n = {}^if(x_n, {}^1y_n, \ldots, {}^my_n). \tag{6.3}$$

Všechny základní teoretické výsledky, které jsme zformulovali v čl. 4 pro jednu diferenciální rovnici, jako jsou např. věty o konvergenci, o řádu metody atd., zůstávají v platnosti i v případě soustavy diferenciálních rovnic. Jen je třeba si uvědomit, že lokální chyba je nyní m-dimenzionální vektor a nikoliv skalár, a že tam, kde se vyskytují v různých odhadech absolutní hodnoty, je třeba užít nějakou normu v m-dimenzionálním vektorovém prostoru.

V problematice stability při pevném integračním kroku však dochází k podstatným rozdílům. V definici absolutní nebo relativní stability se totiž vyskytoval parametr $z = hA$, kde A byl odhad derivace $\partial f/\partial y$ pravé strany dané diferenciální rovnice. Roli tohoto čísla hrají nyní vlastní čísla Jacobiovy matice $\partial f/\partial y$, tj. matice

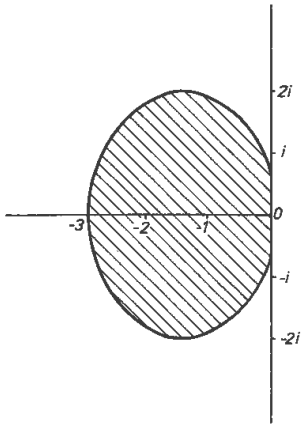
$$\frac{\partial f}{\partial y} = \begin{bmatrix} \frac{\partial {}^1f}{\partial {}^1y} & \frac{\partial {}^1f}{\partial {}^2y} & \cdots & \frac{\partial {}^1f}{\partial {}^my} \\ \vdots & & & \\ \frac{\partial {}^mf}{\partial {}^1y} & \frac{\partial {}^mf}{\partial {}^2y} & \cdots & \frac{\partial {}^mf}{\partial {}^my} \end{bmatrix} \tag{6.4}$$

Protože tato matice není obecně symetrická, mohou být její vlastní čísla i komplexní a za parametr z v definici intervalu stability je nyní třeba vzít komplexní číslo. V souvislosti se soustavou diferenciálních rovnic tedy mluvíme o *oblastech stability*, které se zavádějí obdobně jako intervaly stability.

Kromě případu nejjednodušších metod nelze většinou popsat oblast stability jinak než obrázkem. Jako příklad může sloužit obr. 6.1, v němž je znázorněna oblast absolutní stability Adamsovy-Moultonovy metody čtvrtého řádu.

Obr. 6.1

Oblast stability Adamsovy-Moultonovy metody čtvrtého řádu



6.2 Rungovy-Kuttovy metody

Metody této třídy lze rovněž aplikovat úplně beze změny i v případě soustavy diferenciálních rovnic. Je jen třeba pokládat veličiny Φ a k ve vzorcích (3.17) a (3.18) za m-dimenzionální vektory. Upozorněme však na to, že řád některých Rungových-Kuttových metod může být v případě soustavy rovnic nižší než v případě jedné rovnice. Tento jev nastává pouze pro některé metody řádu vyššího než 4, a ty se v praxi málo užívají. U všech metod vyššího řádu než 4, které jsme uvedli v odst. 3.1, však tento jev nenastává a jejich řád nezávisí na tom, zda jsou užity v případě jedné nebo více diferenciálních rovnic.

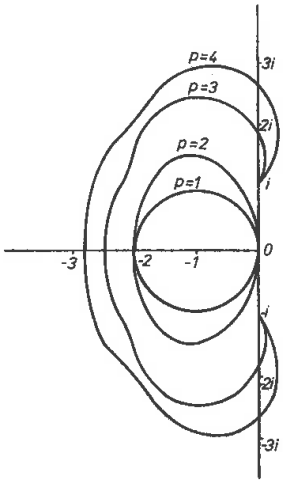
Pokud jde o stabilitu při pevném integračním kroku, platí totéž, co bylo řečeno v souvislosti s lineárními k-krokovými metodami. Ve shodě o tom, co víme o intervalech stability Rungových-Kuttových metod prvního až čtvrtého řádu, které užívají jednu až čtyři hodnoty pravé strany, jsou pro tyto metody i oblasti stability nezávislé na konkrétní metodě. Tyto oblasti jsou znázorněny na obr. 6.2.

6.3 Problematika řešení diferenciálních rovnic se silným tlumením

V mnoha oblastech aplikací, zejména v problémech chemického inženýrství a v problémech regulace, vznikají soustavy diferenciálních rovnic, které vykazují jev, který je v anglicky psané literatuře znám pod jménem „stiffness“. Doslovný překlad tohoto termínu „tuhost“ se nám v češtině příliš nelíbí a budeme jej většinou opisovat tak, že budeme mluvit o soustavách diferenciálních rovnic se silným tlumením. Podstatu tohoto jevu objasníme na následujícím jednoduchém příkladě.

Obr. 6.2

Oblasti stability Rungových-Kuttových metod



Mějme soustavu

$$y' = Ay + b(x) \tag{6.5}$$

m lineárních diferenciálních rovnic, kde A je čtvercová konstantní matice řádu m s navzájem různými vlastními čísly λ_i a odpovídajícími vlastními vektory u_i. Obecné řešení této soustavy se dá v tomto případě psát ve tvaru

$$y(x) = \sum_{i=1}^{m} c_i e^{\lambda_i x} u_i + p(x), \tag{6.6}$$

kde $p(x)$ je partikulární řešení nehomogenní soustavy (6.5). Předpokládejme, že je $\mathrm{Re}\lambda_i < 0$ pro $i = 1, \ldots, m$. Pak platí

$$\sum_{i=1}^{m} c_i e^{\lambda_i x} u_i \to O \quad \text{pro } x \to \infty. \tag{6.7}$$

Z tohoto důvodu mluvíme o kombinaci ve vzorci (6.7) jako o přechodovém řešení a o vektoru $p(x)$ jako o ustáleném řešení. Buďte dále $\lambda_{\max}$ a $\lambda_{\min}$ taková vlastní čísla, že platí

$$|\mathrm{Re}\lambda_{\max}| \geqq |\mathrm{Re}\lambda_i| \geqq |\mathrm{Re}\lambda_{\min}|, \quad i = 1, \ldots, m. \tag{6.8}$$

Je-li naším úkolem vypočítat ustálené řešení, je třeba řešit soustavu (6.5) nejméně

až do toho bodu, v němž je nejpomaleji klesající exponenciála v přechodovém řešení už zanedbatelná. Čím menší je tedy $|\mathrm{Re}\lambda_{\min}|$, tím na delším intervalu je třeba soustavu (6.5) řešit. Na druhé straně z důvodu zachování stabilního průběhu výpočtu je žádoucí volit integrační krok h tak, aby číslo $z = h\lambda_{\max}$ leželo v oblasti absolutní stability užité metody. Je-li tato oblast omezená, je omezení, které zmíněný požadavek klade na velikost integračního kroku tím přísnější, čím větší je $|\mathrm{Re}\lambda_{\max}|$. Je-li $|\mathrm{Re}\lambda_{\max}|$ podstatně větší než $|\mathrm{Re}\lambda_{\min}|$, můžeme se tedy dostat do obtížné situace řešit na dlouhém intervalu diferenciální rovnici a užít při tom integrační krok, který je v kterékoliv fázi výpočtu velice malý vzhledem k délce příslušného intervalu.

O soustavě, která vykazuje právě popsané chování, mluvíme jako o *soustavě se silným tlumením*. Poměr

$$S = \left| \frac{\mathrm{Re}\lambda_{\max}}{\mathrm{Re}\lambda_{\min}} \right| \tag{6.9}$$

zvaný *S-poměr* dané soustavy přitom slouží za míru zmíněných obtíží. Čím je číslo S větší, tím hůře se soustava chová.

V případě, že soustava (6.5) je nelineární, je třeba vlastní čísla matice A nahradit vlastními čísly Jacobiovy matice (6.4). Protože ta už nejsou konstantní, může se charakter uvažovaného jevu měnit v závislosti na nezávisle proměnné x.

Za předpokladu, že pravá strana dané soustavy diferenciálních rovnic má v nějaké vhodné oblasti omezené všechny parciální derivace podle proměnných ${}^1y, \ldots, {}^my$, lze za její Lipschitzovu konstantu vzít číslo $L = \|\partial f/\partial y\|$. Tato konstanta bude při velkém čísle $|\mathrm{Re}\lambda_{\max}|$ rovněž velká. Proto se o soustavě se silným tlumením mluví také jako o soustavě s velkou Lipschitzovou konstantou.

Je třeba podtrhnout, že v uvedené definici jsme se záměrně vyhnuli tomu, abychom udali, co přesně kvantitativně znamenají slova „$|\mathrm{Re}\lambda_{\max}|$ je podstatně větší než $|\mathrm{Re}\lambda_{\min}|$“. Jinými slovy neuvedli jsme žádnou konkrétní hranici, kdy se už soustava chová špatně a kdy ještě ne. To vzhledem k heuristické povaze zmíněné definice není ani možné; je jen jasné, že čím je větší S-poměr, tím je popsaný jev závažnější. Uveďme pouze, že soustava s S-poměrem 20 se pokládá ještě za velmi rozumnou, a že se nezřídka v praxi vyskytují soustavy, kde je toto číslo 10^6.

Následující příklad ukazuje, že obtíže spojené s řešením soustav diferenciálních rovnic se silným tlumením se mohou drasticky projevit už ve velmi jednoduchém případě.

Příklad 6.1. Řešme soustavu diferenciálních rovnic $y' = z$, $z' = -10^6 y - (10^6 + 1)z$ s počátečními podmínkami $y(0) = 1$, $z(0) = -1$ a s přesným řešením $y(x) = e^{-x}$ a $z(x) = -e^{-x}$ Eulerovou metodou. Interval absolutní stability této metody je $(-2, 0)$, iterační krok je tedy třeba volit menší než $2 \cdot 10^{-6}$ (vlastní čísla matice uvažované soustavy jsou -1 a -10^6). Tak např. při výpočtu s integračním krokem $h = 10^{-5}$ dostaneme zcela nesmyslné výsledky už po provedení pouhých několika desítek kroků a to v situaci, kdy se přesné řešení chová jako e^{-x}. Stejné přesné řešení má i diferenciální rovnice $y' = -y$ s počáteční podmínkou $y(0) = 1$.

Řešíme-li však tuto rovnici Eulerovou metodou, není v intervalu $\langle 0, 1)$ celková diskretizační chyba větší než 2.10^{-5}, použijeme-li dokonce podstatně větší integrační krok $h = 10^{-4}$. Zvolíme-li na druhé straně k řešení dané soustavy tzv. implicitní Eulerovu metodu $y_{n+1} = y_n + hf(x_{n+1}, y_{n+1})$, která má interval absolutní stability $(-\infty, 0)$ a položíme-li $h = 10^{-4}$, vypočteme řešení na intervalu $\langle 0, 1\rangle$ s přesností na 4 desetinná místa, tedy stejnou, jako v případě explicitní Eulerovy metody a rovnice $y' = -y$.

Řešíme-li soustavu diferenciálních rovnic se silným tlumením, je tedy žádoucí užít k tomu takovou metodu, že omezení, která z ní plynou pro integrační krok z důvodu zachování stability, nejsou pokud možno příliš vážná. Z tohoto hlediska se zdá rozumné omezit se v této situaci na takové metody, jejichž oblast absolutní stability obsahuje celou levou polorovinu komplexní roviny. V tomto případě požadavky stability nekladou na velikost integračního kroku žádné omezení bez ohledu na to, jak velké je číslo $|\mathrm{Re}\lambda_{\max}|$. Metody mající tuto vlastnost se nazývají *A-stabilní* a pro jejich zřejmou důležitost si jich všimněme poněkud podrobněji.

Vzpomeneme-li si, že interval stability byl definován na základě chování přibližného řešení modelové diferenciální rovnice

$$y' = Ay \tag{6.10}$$

při $n \to \infty$, je skoro na první pohled patrné, že žádná Rungova-Kuttova metoda nemůže být A-stabilní.

Skutečně, přibližné řešení diferenciální rovnice (6.10) libovolnou Rungovou-Kuttovou metodou je totiž dáno vzorcem

$$y_n = [P(z)]^n y(0), \tag{6.11}$$

kde $z = hA$ a $P(z)$ je polynom stupně nejméně 1. Nutná a postačující podmínka A-stability této metody je tedy platnost nerovnosti $|P(z)| < 1$ pro všechna z, pro něž je $\mathrm{Re}z < 0$. To však není možné, neboť funkce P je polynom, a má tedy v nekonečnu pól.

Komplikovanější situace je v případě lineární k-krokové metody. Podle definice 4.5 je tato metoda A-stabilní právě tehdy, má-li třetí charakteristický polynom

$$\pi(z, \xi) = \varrho(\xi) - z\sigma(\xi) \tag{6.12}$$

při libovolném z takovém, že $\mathrm{Re}z < 0$ všechny kořeny uvnitř jednokrokového kruhu. K dalšímu vyšetřování A-stability mnohokrokových metod bude užitečná jiná nutná a postačující podmínka. Její formulace je obsahem následujícího pomocného tvrzení.

Lemma 6.1. *Buď dána lineární k-kroková metoda, jejíž charakteristické polynomy ϱ a σ nemají společné činitele. Buď dále R racionální funkce definovaná*

rovnicí

$$R(\xi) = \frac{\varrho(\xi)}{\sigma(\xi)}. \tag{6.13}$$

Pak je daná metoda A-stabilní tehdy a jen tehdy, je-li funkce R holomorfní vně jednotkového kruhu a platí-li pro ni

$$\mathrm{Re}R(\xi) \geqq 0 \quad \text{pro } |\xi| > 1. \tag{6.14}$$

D ů k a z . Předpokládejme nejprve, že daná metoda je A-stabilní a buď ξ_0 takový bod ležící vně jednotkového kruhu, v němž je funkce R regulární. Pak v důsledku toho, že polynomy ϱ a σ nemají společné činitele, platí $\sigma(\xi_0) \neq 0$. Položíme-li $z = R(\xi_0) = \varrho(\xi_0)/\sigma(\xi_0)$, je ξ_0 kořenem polynomu (6.12) pro toto z. Protože je $|\xi_0| > 1$, je v důsledku A-stability $\mathrm{Re}z \geqq 0$. Buď dále $|\xi_1| > 1$ takový bod, že je $\sigma(\xi_1) = 0$. Pak opět v důsledku nesoudělnosti polynomů ϱ a σ platí $\varrho(\xi_1) \neq 0$ a funkce R se tedy chová v okolí bodu $\xi = \xi_1$ jako $a(\xi - \xi_1)^{-m}$, kde $a \neq 0$ a m je přirozené číslo. V okolí bodu $\xi = \xi_1$ tedy existují body, které stále ještě zůstávají vně jednotkového kruhu a v nichž nabývá funkce $\mathrm{Re}R$ jak kladných, tak záporných hodnot. To je ale ve sporu s právě dokázanou nerovností $\mathrm{Re}R(\xi) \geqq 0$, která platí ve všech zmíněných bodech. Funkce R je tedy regulární pro každé $|\xi| > 1$ a platí (6.14).

Nechť naopak podmínka (6.14) je splněna a nechť existuje pro nějaké z, pro něž je $\mathrm{Re}z < 0$ kořen ξ_1 polynomu (6.12), pro který platí $|\xi_1| \geqq 1$. Pak je $\sigma(\xi_1) \neq 0$, neboť v opačném případě by byl bod ξ_1 společným kořenem polynomů ϱ a σ, což není možné. Platí-li dokonce $|\xi_1| > 1$, dostáváme spor s (6.14). Je-li $|\xi_1| = 1$, plyne ze spojitosti funkce $\mathrm{Re}R$ v bodě $\xi = \xi_1$ existence bodu ξ_2, pro který platí $|\xi_2| > 1$ a $\mathrm{Re}R(\xi_2) < 0$. Došli jsme tedy opět ke sporu. Lemma je dokázáno.

Pomocí tohoto lemmatu už snadno dokážeme základní a typickou vlastnost A-stabilních mnohokrokových metod zformulovanou v následující větě.

Věta 6.1. *A-stabilní lineární k-kroková metoda je implicitní (tj. platí pro ni $\beta_k \neq 0$).*

D ů k a z . Buď dána explicitní lineární k-kroková metoda. Je tedy $\beta_k = 0$ a pro funkci R definovanou rovnicí (6.13) platí

$$R(\xi) = \frac{\alpha_k \xi^k + \ldots + \alpha_0}{\beta_{k-s}\xi^{k-s} + \ldots + \beta_0}, \tag{6.15}$$

kde $\alpha_k \neq 0$, $\beta_{k-s} \neq 0$ a $s \geqq 1$. V okolí bodu v nekonečnu se tedy chová funkce R jako ξ^s a nabývá tak hodnot jak z levé, tak z pravé poloroviny. Podle lemmatu 6.1 není tedy uvažovaná metoda A-stabilní.

K tomu, abychom dokázali ještě závažnější vlastnost A-stabilních mnohokrokových metod, totiž že jejich řád je silně limitován, budeme potřebovat jedno pomocné

tvrzení, které pojednává o integrální reprezentaci holomorfní funkce s hodnotami v pravé polorovině komplexní roviny.

Lemma 6.2. *Buď funkce φ holomorfní v polorovině* $\mathrm{Re}\, z > 0$, *a nechť platí*

$$\sup\{|x\varphi(x)|;\ 0 < x < \infty\} < \infty \tag{6.16}$$

a

$$\mathrm{Re}\,\varphi(z) \geqq 0 \quad \text{pro } \mathrm{Re}\, z \geqq 0. \tag{6.17}$$

Pak existuje omezená neklesající funkce ω taková, že platí

$$\varphi(z) = \int_{-\infty}^{\infty} \frac{\mathrm{d}\,\omega(t)}{z - \mathrm{i}t} \tag{6.18}$$

pro každé z takové, že je $\mathrm{Re}\, z > 0$.

D ů k a z lemmatu přesahuje elementární rámec této knihy a lze jej nalézt např. v knize Achiezer, Glazman (1950), str. 206.

Věta 6.2. *Maximální řád A-stabilní a D-stabilní lineární k-krokové metody takové, že polynomy ϱ a σ nemají společné činitele, je 2.*

D ů k a z . Buď dána lineární k-kroková metoda řádu $p \geqq 2$. Podle věty 4.9 má funkce $\varrho(\xi) - \sigma(\xi)\ln\xi$ v bodě $\xi = 1$ kořen násobnosti $p+1$, tj. pro $\xi \to 1$ platí

$$\varrho(\xi) - \sigma(\xi)\ln\xi = -c(\xi - 1)^{p+1} + O\big((\xi-1)^{p+2}\big). \tag{6.19}$$

V důsledku D-stability a konzistence je $\sigma(1) \neq 0$; platí tedy pro $\xi \to 1$

$$\frac{1}{\sigma(\xi)} = \frac{1}{\sigma(1) + \frac{1}{1!}\sigma'(1)(\xi-1) + \ldots + \frac{1}{k!}\sigma^{(k)}(1)(\xi-1)^k} = \tag{6.20}$$
$$= \frac{1}{\sigma(1)}[1 + O(\xi - 1)].$$

Dělíme-li rovnici (6.19) funkcí σ, dostáváme odtud, že pro $\xi \to 1$ platí

$$\ln\xi - \frac{\varrho(\xi)}{\sigma(\xi)} = c^*(\xi-1)^{p+1} + O\big((\xi-1)^{p+2}\big), \tag{6.21}$$

kde jsme položili $c^* = c/\sigma(1)$.

Zaveďme nyní novou proměnnou z pomocí transformace

$$z = \frac{\xi+1}{\xi-1}, \quad \xi = \frac{z+1}{z-1} \tag{6.22}$$

a položme

$$r(z) = \Big(\frac{z-1}{2}\Big)^k \varrho\Big(\frac{z+1}{z-1}\Big), \quad s(z) = \Big(\frac{z-1}{2}\Big)^k \sigma\Big(\frac{z+1}{z-1}\Big). \tag{6.23}$$

Protože transformace (6.22) převádí vnějšek jednotkového kruhu roviny ξ na pravou polorovinu roviny z, lze lemma 6.1 vyslovit také tak, že uvažovaná lineární k-kroková metoda je A-stabilní právě tehdy, je-li funkce r/s holomorfní v polorovině $\mathrm{Re}\, z > 0$ a má-li tam nezápornou reálnou část. Dále, protože je $\varrho(1) = 0$, je polynom r v (6.23) stupně nejvýše $k-1$ a protože je $\sigma(1) \neq 0$, je polynom s stupně k. Funkce $xr(x)/s(x)$ je proto na intervalu $(0,\infty)$ omezená. Je-li navíc daná metoda A-stabilní, jsou pro funkci r/s splněny předpoklady lemmatu 6.2. Platí tedy

$$\frac{r(z)}{s(z)} = \int_{-\infty}^{\infty} \frac{\mathrm{d}\,\omega(t)}{z - \mathrm{i}t} \tag{6.24}$$

pro $\mathrm{Re}\, z > 0$, kde ω je vhodná ohraničená a neklesající funkce. Pro kladná x je však výraz $xr(x)/s(x)$ reálný, neboť daná metoda má reálné koeficienty. Podle (6.24) tak dostáváme

$$\frac{xr(x)}{s(x)} = \int_{-\infty}^{\infty} x\frac{\mathrm{d}\,\omega(t)}{x - \mathrm{i}t} = \int_{-\infty}^{\infty} \frac{x(x+\mathrm{i}t)}{x^2+t^2}\,\mathrm{d}\,\omega(t) = \int_{-\infty}^{\infty} \frac{x^2}{x^2+t^2}\,\mathrm{d}\,\omega(t). \tag{6.25}$$

Při pevném t je funkce $x \to x^2/(x^2+t^2)$ neklesající, a tedy také funkce $x \to$ $\to xr(x)/s(x)$ je neklesající.

Přejděme konečně ve vzorci (6.21) k proměnné z. Dostaneme

$$\ln\frac{z+1}{z-1} - \frac{r(z)}{s(z)} = c^*\Big(\frac{2}{z}\Big)^{p+1} + O\Big(\frac{1}{z^{p+2}}\Big) \tag{6.26}$$

pro $z \to \infty$. Protože je zřejmě

$$\ln\frac{z+1}{z-1} = \ln\frac{1+\frac{1}{z}}{1-\frac{1}{z}} = \frac{2}{z} + \frac{2}{3z^3} + O\Big(\frac{1}{z^5}\Big) \tag{6.27}$$

pro $z \to \infty$, plyne z (6.26), že pro $z \to \infty$ platí

$$\frac{r(z)}{s(z)} = \frac{2}{z} + (\tfrac{2}{3} - 8\tilde{c})\frac{1}{z^3} + O\Big(\frac{1}{z^5}\Big), \tag{6.28}$$

kde

$$\tilde{c} = \begin{cases} c^*, & \text{je-li } p = 2 \\ 0, & \text{je-li } p > 2. \end{cases} \tag{6.29}$$

Je-li daná metoda A-stabilní, je funkce $xr(x)/s(x)$ neklesající. Funkce $2 + (2/3 - -8\tilde{c})/x^2$ musí být tedy vzhledem k rovnici (6.28) také neklesající. To však pro $p > 2$ není možné a pro $p = 2$ je to možné právě tehdy, je-li koeficient u x^{-2} nekladný, tj. je-li $c^* \geqq 1/12$. Věta je dokázána.

Pro Adamsovu-Moultonovu metodu řádu 2

$$y_{n+1} = y_n + \tfrac{1}{2}h(f_{n+1} + f_n), \tag{6.30}$$

která se nazývá lichoběžníkové pravidlo, je $c^* = 1/12$. Protože pro tuto metodu je $r(z)/s(z) = 2/z$, je A-stabilní. Adamsova-Moultonova metoda je tedy příkladem A-stabilní metody řádu dvě. Věta 6.2 říká, že žádná z eventuálně dalších A-stabilních metod nemá vyšší řád a z jejího důkazu dokonce plyne, že metoda (6.30) má mezi těmito metodami nejmenší konstantu chyby.

Omezení na řád plynoucí z A-stability je velmi závažné a proto byla navržena řada dalších méně omezujících definic stability. Jeden z těchto pojmů je např. tzv. *$A(\alpha)$-stabilita*, kdy se požaduje, aby oblast stability obsahovala nekonečný úhel $U_\alpha = \{z; -\alpha < \pi - \arg z < \alpha\}$. $A(\alpha)$-stabilita sice pořád ještě vyžaduje implicitnost dané metody, omezení na její řád však není tak přísné jako u A-stability. Do dalších podrobností nebudeme zacházet a spokojíme se konstatováním, že ve speciálních případech mohou prokázat $A(\alpha)$-stabilní metody stejně cennou službu jako A-stabilní metody.

Na druhé straně nelze zamlčet, že ani požadavek A-stability není v jistém smyslu dostačující. Přibližné řešení y_n modelové diferenciální rovnice (6.10) získané libovolnou jednokrokovou metodou, kterou jsme až dosud zavedli, je tvaru $R^n(z)y(a)$, kde $z = hA$ a R je racionální funkce aproximující funkci e^z. Požadavek A-stability proto znamená, že tato funkce musí splňovat podmínku

$$|R(z)| < 1 \quad \text{pro } \mathrm{Re}\, z < 0. \tag{6.31}$$

Např. pro lichoběžníkové pravidlo (6.30) je funkce R dána rovnicí

$$R(z) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}. \tag{6.32}$$

Podmínka (6.31) je sice pro tuto funkci splněna, pro $\mathrm{Re}\, z \to -\infty$ však platí $R(z) \to -1$. Silně tlumené složky tedy klesají k nule, ale velmi pomalu.

Právě provedená úvaha vede k zavedení pojmu *L-stability* jednokrokové metody, kdy kromě A-stability požadujeme, aby pro funkci R platilo

$$R(z) \to 0 \quad \text{pro } \mathrm{Re}\, z \to -\infty. \tag{6.33}$$

Příkladem L-stabilní metody je implicitní Eulerova metoda

$$y_{n+1} = y_n + hf_{n+1}. \tag{6.34}$$

Zakončeme tento odstavec a vůbec celou kapitolu o obyčejných diferenciálních rovnicích užitečným a důležitým upozorněním. V souvislosti s řešením soustav diferenciálních rovnic se silným tlumením jde vždy o metody implicitní, takže v každém kroku je třeba řešit soustavu

$$y_{n+k} - h\frac{\beta_k}{\alpha_k} f(x_{n+k}, y_{n+k}) - c = 0. \tag{6.35}$$

Postupy prediktor-korektor $P(EC)^m$ nebo $P(EC)^m E$ s konečným m, které jsme doporučili v případě „normálně“ se chovajících diferenciálních rovnic, zde nepřicházejí v úvahu, neboť nepřípustně mění oblast stability použité metody. Rovněž tak

iterování do konvergence zde není vhodné, neboť k tomu, aby tento proces skutečně konvergoval, je třeba splnění podmínky

$$h\left|\frac{\beta_k}{\alpha_k}\right| L < 1, \tag{6.36}$$

kde L je Lipschitzova konstanta pravé strany dané soustavy diferenciálních rovnic. Tato konstanta je však u soustavy diferenciálních rovnic se silným tlumením natolik velká, že podmínka (6.36) klade na velikost integračního kroku stejně vážné omezení jako podmínky stability nevhodné metody. Užijeme-li však k řešení nelineární soustavy (6.35) Newtonovu metodu, je tato metoda většinou konvergentní, aniž by bylo nutno klást příliš vážná omezení na velikost integračního kroku, jen když počáteční aproximace je dostatečně přesná. K jejímu získání je možné užít vhodný prediktor. Je však třeba vzít na vědomí, že tento postup je značně náročný na počet potřebných operací, neboť v každé Newtonově iteraci je třeba invertovat vždy novou matici.

CVIČENÍ

1. Dokažte, že odhad rychlosti konvergence Eulerovy metody daný větou 2.2 nelze obecně zlepšit. Návod: Řešte např. diferenciální rovnici $y' = 2x$, $y(0) = 0$ Eulerovou metodou, vypočtěte přibližné řešení vzorcem a ukažte, že je $e_n = y_n - y(x_n) = -x_n h$.

2. Odvoďte rovnice (3.19).

3. Odvoďte aposteriorní odhad typu (3.71) pro obecné integrační kroky h_1 a h_2.

4. Odvoďte rekurence (4.37), (4.41) a (4.44) pro koeficienty Adamsovy-Moultonovy, Nyströmovy a Milnovy-Simpsonovy metody.

5. Odvoďte vzorec (4.56) pro lokální chybu metody numerického derivování.

6. Nechť φ_n, ψ_n a χ_n jsou posloupnosti reálných čísel definované pro $n = 0, \ldots, N$ takové, že je $\chi_n \geqq 0$, $n = 0, \ldots, N$, a

$$\varphi_n \leqq \psi_n + \sum_{\nu=0}^{n-1} \chi_\nu \varphi_\nu, \quad n = 0, \ldots, N.$$

   Pak platí

$$\varphi_n \leqq \psi_n + \sum_{\nu=0}^{n-1} \chi_\nu \psi_\nu \prod_{s=\nu+1}^{n-1} (1 + \chi_s).$$

   Dokažte!

7. Dokažte poznámku 4.3 ze str. 69.

8. Dokažte nerovnosti (4.115).

9. Dokažte, že D-stabilní lineární k-kroková metoda řádu nejméně jedna je absolutně nestabilní pro každé dostatečně malé kladné z. Návod: Použijte toho, že růstový parametr λ_1 je roven 1.

10. Vypočtěte intervaly absolutní stability metod z tab. 4.10.

11. Dokažte poznámku 4.7 ze str. 100.

12. Odvoďte vzorec pro hlavní část lokální diskretizační chyby Rungovy-Kuttovy metody metodou polovičního kroku. Návod: Buďte y_{n+1} a y^*_{n+1} přibližná řešen vypočtená Rungovou-Kuttovou metodou s integračním krokem h, resp. $2h$ v bodě $x = x_{n+1}$ za předpokladu, že výchozí hodnoty jsou přesné. Pak platí

$$\begin{aligned} y(x+h) - y_{n+1} &= L\big(y(x);h\big) = \varphi\big(x,y(x)\big)h^{p+1} + O(h^{p+2}),\\ y(x+h) - y^*_{n+1} &= L\big(y(x),2h\big) = \varphi\big(x-h,y(x-h)\big)(2h)^{p+1} + O(h^{p+2}) = \\ &= \big[\varphi\big(x,y(x)\big) + O(h)\big](2h)^{p+1} + O(h^{p+2}) = \\ &= \varphi\big(x,y(x)\big)2^{p+1}h^{p+1} + O(h^{p+2}). \end{aligned}$$

Odtud vyloučením přesného řešení dostanete

$$L\big(y(x);h\big) = \frac{1}{2^{p+1}-1}(y_{n+1} - y^*_{n+1}) + O(h^{p+2}).$$

13. Ukažte, že interval absolutní stability Rungovy-Kuttovy metody řádu p, která užívá p hodnot pravé strany, nezávisí na její konkrétní podobě. Návod: Tvrzení plyne z toho, že přibližné řešení diferenciální rovnice $y' = Ay$ vypočtené zmíněnou Rungovou-Kuttovou metodou řádu p je dáno vzorcem $y_n = (1 + z + z^2/2! + \ldots + z^p/p!)^n y_0$, kde $z = hA$.

POZNÁMKY K LITERATUŘE

Čl. 1. Publikací, které pojednávají o teorii obyčejných diferenciálních rovnic je celá řada. K velmi dobrým a pro českého čtenáře dostupným pramenům patří kniha Coddingtona a Levinsona (1955). Literatura věnovaná problematice numerického řešení obyčejných diferenciálních rovnic je rovněž neobyčejně obsáhlá. Základní informace čtenář získá v celé řadě příruček: viz např. Collatz (1951), Henrici (1964), Berezin, Židkov (1966), Isaacson, Keller (1966), Shampine, Allen (1973), Dahlquist, Björck (1974), Stoer, Bulirsch (1980), Vitásek (1987). Z celé řady specializovaných knih, z nichž některé jsou uvedeny v seznamu literatury bez dalšího komentáře, jmenujme knihu Henriciovu (1962), patrně jeden z nejlepších a nejúplnějších pramenů zejména v teoretických otázkách souvisejících s numerickým řešením obyčejných diferenciálních rovnic. Většina materiálu první kapitoly je na ní také založena a čtenáři hledajícímu hlubší porozumění této problematice ji lze co nejvřeleji doporučit. Rovněž kniha Stetterova (1973) i když pro čtenáře dosti náročná, zaslouží

být jmenována. Praktičtěji než uvedené dvě publikace je zaměřena vynikající kniha Lambertova (1973), která nepomíjí snad žádný aspekt spojený s praktickou realizací popisovaných metod. Na mnoha místech této kapitoly jsme z ní vydatně čerpali. Podobný charakter má také kniha Lapiduse a Seinfelda (1971). Konečně upozorněme na sborník vydaný Hallem a Wattem (1976), v němž nalezne čtenář popis a hodnocení celé řady algoritmů založených nejen na metodách popisovaných v této knize, ale i na tzv. extrapolačních metodách, které jsou rovněž dosti populární.

Čl. 2. Nejlepším pramenem k obsahu tohoto článku je kniha Henriciova (1962).

Čl. 3. Rungovy-Kuttovy metody jsou klasické a základní informace o nich je v každé učebnici. Většinu materiálu zde uvedeného nalezne čtenář v už zmíněné knize Lambertově (1973), která také obsahuje odkazy na originální literaturu o jednotlivých metodách. Kniha Forsytha, Malcoma a Molera (1977) obsahuje velmi dobře implementovaný program pro tzv. Rungovu-Kuttovu-Fehlbergovu metodu.

Čl. 4. Všechny otázky spojené s problematikou konvergence a konstrukcí většiny základních tříd mnohokrokových metod jsou vyčerpávajícím způsobem řešeny v knize Henriciově (1962). Problémy stability při pevném integračním kroku jsou v souvislosti s mnohokrokovými metodami a s metodami prediktor-korektor podrobně studovány v knize Lambertově (1973). Výhradně metodám Adamsova typu je věnována velmi přístupně psaná kniha Shampinova a Gordonova (1975), která obsahuje mimo jiné dobře osvědčený program užívající Adamsovy metody proměnného řádu a automatickou volbou integračního kroku. Na Adamsových metodách proměnného řádu je založen také program publikovaný Gearem (1971b), který realizuje změnu integračního kroku na základě Nordsieckovy (1962) myšlenky.

Čl. 5. Tento článek se opírá hlavně o Lambertovu knihu (1973).

Čl. 6. Průkopnická práce o problematice řešení soustav diferenciálních rovnic se silným tlumením je práce Dahlquistova (1963). Velmi mnoho materiálu o této problematice obsahuje sborník vydaný Willoughbym (1974). Gearův (1971b) program, o němž jsme se už zmínili, obsahuje opci pro integraci stiff systémů. Příslušné metody jsou založeny na metodách numerického derivování.

LITERATURA

ACHIJEZER, N.I. – GLAZMAN, I.M.: Teorija linejnych operatorov. Moskva, GITTL 1950.

BABUŠKA, I. – PRÁGER, M. – VITÁSEK, E.: Numerical Processes in Differential Equations. Praha–London–New York–Sydney, SNTL + Interscience Publishers 1966. (Překlad do ruštiny: Moskva, Mir 1969.).

BACHVALOV, N.S.: Čislennyje metody. Moskva, Nauka 1975.

CODDINGTON, E.A. – LEVINSON, N.: Theory of Ordinary Differential Equations. New York, McGraw-Hill 1955. (Překlad do ruštiny: Moskva, IL 1958.)

COLLATZ, L.: Numerische Behandlung von Differentialgleichungen. Berlin–

Göttingen–Heidelberg, Springer–Verlag 1951. (Překlad do ruštiny: Moskva, IL 1953.)

ČERNÝ, I.: Základy analýzy v komplexním oboru. Praha, Academia 1967.

DAHLQUIST, G.: A Special Stability Problem for Linear Multistep Methods. BIT, *3*, 1963, s. 27 – 43.

DAHLQUIST, G. – BJÖRCK, A.: Numerical Methods. Englewood Cliffs, N.J., Prentice–Hall 1974.

DANIEL, J.W. – MOORE, R.E.: Computation and Theory in Ordinary Differential Equations. San Francisco, W.H. Freeman and Co. 1970.

FORSYTHE, G.E. – MALCOLM, M.A. – MOLER, C.B.: Computer Methods for Mathematical Computations. Englewood Cliffs, N.J., Prentice–Hall 1977. (Překlad do ruštiny: Moskva, Mir 1980.)

FOX, L.: Numerical Solution of Ordinary and Partial Differential Equations. Oxford, Pergamon Press 1962.

GEAR, C.W.: Numerical Initial Value Problems in Ordinary Differential Equations. Englewood Cliffs, N.J., Prentice–Hall 1971a.

GEAR, C.W.: Algorithm 407, DIFSUB for Solution of Ordinary Differential Equations. Com. ACM, *14*, 1971b, s. 185 – 190.

GRIGORIEFF, R.D.: Numerik gewöhnlicher Differentialgleichungen. Stuttgart, B.G. Teubner 1972 – 1977, 2 sv.

HALL, G. – WATT, J.M. (ed.): Modern Numerical Methods for Ordinary Differential Equation. Oxford, Clarendon Press 1976. (Překlad do ruštiny: Moskva, Mir 1979.)

HAMMING, R.W.: Numerical Methods for Scientists and Engineers. New York, McGraw-Hill 1962. (Překlad do ruštiny: Moskva, Nauka 1968.)

HENRICI, P.: Discrete Variable Methods in Ordinary Differential Equations. New York–London, J. Wiley and Sons 1962.

HENRICI, P.: Error Propagation for Difference Methods. London, J. Wiley and Sons 1963.

HENRICI, P.: Elements of Numerical Analysis. New York, J. Wiley and Sons 1964.

ISAACSON, E. – KELLER, H.B.: Analsysis of Numerical Methods. New York–London–Sydney, J. Wiley and Sons 1966.

LAMBERT, J.D.: Computational Methods in Ordinary Differential Equations. London-New York-Sydney-Toronto, J. Wiley and Sons 1973.

LAPIDUS, L. – SEINFELD, J.H.: Numerical Solution of Ordinary Differential Equations. New York–London, Academic Press 1971.

NORDSIECK, A.: On Numerical Integration of Ordinary Differential Equations. Math. Comp., *16*, 1962, s. 22 – 49.

RICE, J.R. (ed.): Mathematical Software. New York–London, Academic Press 1971.

SHAMPINE, L.F. – ALLEN, R.: Numerical Computing. Philadelphia, Saunders 1973.

SHAMPINE, L.F. – GORDON, M.K.: Computer Solution of Ordinary Differential Equations. San Francisco, W.H. Freeman and Co. 1975.

STETTER, H.J.: Analysis of Discretization Methods for Ordianary Differential Equations. Berlin–Heidelberg–New York, Springer–Verlag 1973. (Překlad do ruštiny: Moskva, Mir 1978.)

STOER, J. – BULIRSCH, R.: Introduction to Numerical Analysis. New York–Heidelberg–Berlin, Springer–Verlag 1980.

VITÁSEK, E.: Numerické metody. Praha, SNTL 1987.

WILLOUGHBY, R.A. (ED.): Stiff Differential Systems. New York–London, Plenum Press 1974.

Kapitola II.

Obyčejné diferenciální rovnice – okrajové úlohy

1 Úvod

V této kapitole popíšeme některé základní metody pro numerické řešení okrajových úloh pro obyčejné diferenciální rovnice. *Okrajovou úlohou* nebo podrobněji dvoubodovou okrajovou úlohou pro soustavu diferenciálních rovnic prvního řádu

$$y' = f(x, y), \tag{1.1}$$

kde

$$y = \begin{bmatrix} {}^1y \\ {}^2y \\ \vdots \\ {}^my \end{bmatrix}, \quad f(x,y) = \begin{bmatrix} {}^1f(x, {}^1y, \dots, {}^my) \\ {}^2f(x, {}^1y, \dots, {}^my) \\ \vdots \\ {}^mf(x, {}^1y, \dots, {}^my) \end{bmatrix} \tag{1.2}$$

rozumíme úlohu nalézt takové řešení soustavy (1.1), pro něž platí

$$r(y(a), y(b)) = 0, \tag{1.3}$$

kde r je m-složková vektorová funkce $2m$ proměnných a a a b jsou dva navzájem různé body intervalu, v němž hledáme řešení. Nejčastěji to jsou jeho krajní body.

Dají-li se podmínky (1.3), které nazýváme *okrajovými podmínkami*, psát ve tvaru

$$r_1(y(a)) = 0, \quad r_2(y(b)) = 0, \tag{1.4}$$

kde r_1, resp. r_2 jsou m_1-, resp. m_2-složkové vektorové funkce m proměnných ($m = m_1 + m_2$), mluvíme o *separovaných okrajových podmínkách*. Ještě speciálnější typ okrajových podmínek jsou *lineární okrajové podmínky*

$$Uy(a) + Vy(b) = c, \tag{1.5}$$

kde U a V jsou čtvercové matice řádu m a c je m-dimenzionální vektor a *lineární separované okrajové podmínky*

$$V_1 y(a) = v_1, \quad V_2 y(b) = v_2, \tag{1.6}$$

kde V_1, resp. V_2, jsou obdélníkové matice typu $m_1 \times m$, resp. $m_2 \times m$, a v_1, resp. v_2 jsou m_1-, resp. m_2-dimenzionální vektory.

Protože každou diferenciální rovnici m-tého řádu lze psát zavedením nových neznámých funkcí jako soustavu rovnic prvního řádu (srv. vzorec (1.4) a (1.5) z kap. 1), je zřejmé, co rozumíme dvoubodovou okrajovou úlohou pro rovnici m-tého řádu.

Položíme-li v rovnici (1.5) $U = I$ a $V = 0$, vidíme, že úloha s počátečními podmínkami je speciálním případem okrajové úlohy. Tedy už nejjednodušší separované okrajové podmínky jsou značně obecnější než počáteční podmínky. To má velmi závažné důsledky. Zatímco řešení úlohy s počátečními podmínkami existuje a je jediné pro značně širokou třídu nelineárních rovnic typu (1.1), u okrajové úlohy se může i v případě jednoduché lineární rovnice snadno stát, že řešení neexistuje, nebo naopak, že řešení je nekonečně mnoho. Tuto skutečnost si dobře uvědomíme na příkladě jednoduché diferenciální rovnice 2. řádu

$$y'' + y = 0 \tag{1.7}$$

nebo ekvivalentní soustavě prvního řádu

$$\begin{bmatrix} {}^1y \\ {}^2y \end{bmatrix}' = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} {}^1y \\ {}^2y \end{bmatrix}, \tag{1.8}$$

která se dostane z rovnice (1.7) standardní substitucí. Každé řešení rovnice (1.7) se dá psát ve tvaru

$$y(x) = C_1 \cos x + C_2 \sin x, \tag{1.9}$$

kde C_1 a C_2 jsou libovolné konstanty. Odtud ihned plyne, že každá funkce tvaru $A \sin x$, kde A je libovolné číslo, je řešením diferenciální rovnice (1.7) s okrajovými podmínkami

$$y(0) = 0, \quad y(\pi) = 0. \tag{1.10}$$

Okrajová úloha (1.7), (1.10) má tedy nekonečně mnoho řešení. Zaměníme-li naproti tomu okrajové podmínky (1.10) okrajovými podmínkami

$$y(0) = 0, \quad y(\pi) = 1, \tag{1.11}$$

neexistuje vůbec žádné řešení.

Uvedený jednoduchý příklad napovídá, že teorie existence a jednoznačnosti řešení okrajových úloh je značně komplikovanější než odpovídající teorie úloh s počátečními podmínkami. Odrazem této skutečnosti v numerických metodách pro řešení okrajových úloh je pak to, že pro ně doposud nebyla vypracována natolik univerzální teorie, jako tomu bylo u metod pro řešení úloh s počátečními podmínkami. Proto také popis jednotlivých konkrétních metod bude mít v této kapitole značně jiný charakter, než tomu bylo v kap. I. Tam jsme vlastně všude vystačili s tím, že

jsme se zabývali jedním obecným problémem, totiž diferenciální rovnicí (1.1) s příslušnou počáteční podmínkou a v podstatě všechny uváděné metody měly pro tuto úlohu smysl. Zde je tomu z naznačených důvodů jinak, a proto další text bude vlastně do značné míry souhrn jednotlivých více méně konkrétních úloh a příslušných algoritmů pro jejich řešení.

Jednou z nejvíce frekventovaných úloh v této kapitole bude úloha řešit lineární samoadjungovanou diferenciální rovnici

$$-(p(x)y')' + q(x)y = f(x) \tag{1.12}$$

se separovanými lineárními okrajovými podmínkami

$$\begin{aligned} -\alpha_1 p(a)y'(a) + \beta_1 y(a) &= \gamma_1, \\ \alpha_2 p(b)y'(b) + \beta_2 y(b) &= \gamma_2. \end{aligned} \tag{1.13}$$

Většinu výsledků, které dostaneme pro tuto okrajovou úlohu, lze bez podstatných obtíží přenést na diferenciální rovnici čtvrtého řádu

$$(p(x)y'')'' + q(x)y = f(x) \tag{1.14}$$

s okrajovými podmínkami např.

$$\begin{aligned} y(a) &= \gamma_1, \quad y'(a) = \delta_1, \\ y(b) &= \gamma_2, \quad y'(b) = \delta_2. \end{aligned} \tag{1.15}$$

Rovněž touto úlohou se budeme v dalším zabývat.

Konečně si také všimneme, zejména proto, abychom ilustrovali problémy, které vznikají při řešení nelineárních úloh, diferenciální rovnice

$$y'' = f(x, y) \tag{1.16}$$

s okrajovými podmínkami

$$y(a) = \gamma_1, \quad y(b) = \gamma_2. \tag{1.17}$$

Většinu prostoru budeme tedy věnovat lineárním úlohám, které jsme v případě úloh s počátečními podmínkami ani separátně neprobírali. Důvody pro to jsou v podstatě dva. O jednom z nich — teoretických obtížích spjatých s nelineárními okrajovými úlohami — jsme se už zmínili. Druhý důvod je ten, že nelineární úlohy se často, opět na rozdíl od úloh s počátečními podmínkami, řeší iteračně tak, že se konstruuje posloupnost lineárních úloh, jejichž řešení konvergují k řešení původní úlohy. K provedení tohoto postupu je uspokojivé a efektivní řešení lineárních úloh nezbytným předpokladem.

Z konkrétních jednotlivých metod pro řešení výše popsaných problémů se všimneme metod, které převádějí řešení okrajové úlohy na řešení úloh s počátečními podmínkami, dále pak metody sítí a konečně velmi stručně variačních metod, zejména jejich důležitého speciálního případu, totiž metody konečných prvků.

2 Metody založené na převodu na úlohy s počátečními podmínkami

V kapitole 1 jsme viděli, že úlohy s počátečními podmínkami pro obyčejné diferenciální rovnice umíme celkem uspokojivě řešit. Je proto přirozené snažit se převést řešení okrajové úlohy na řešení úloh s počátečními podmínkami. V tomto článku si všimněme několika způsobů, jak toho lze docílit.

2.1 Metoda střelby

Základní myšlenka této metody je velmi jednoduchá a nabízí se na první pohled. Popíšeme ji nejprve na příkladě diferenciální rovnice druhého řádu.

2.1.1 Okrajová úloha pro lineární rovnici druhého řádu

Uvažujme diferenciální rovnici (1.12) s okrajovými podmínkami (1.17), které jsou zřejmě speciálním případem podmínek (1.13). Často se jim také říká *Dirichletovy okrajové podmínky*. Předpokládejme přitom, že funkce p, p', q f jsou spojité v intervalu $\langle a, b\rangle$, že existuje kladná konstanty p_0 taková, že platí

$$p(x) \geqq p_0\ (> 0), \quad x \in \langle a, b\rangle\,, \tag{2.1}$$

a že je

$$q(x) \geqq 0, \quad x \in \langle a, b\rangle\,. \tag{2.2}$$

Později uvidíme, že tyto předpoklady zaručují existenci a jednoznačnost řešení úlohy (1.12), (1.17). (Ve skutečnosti je možno tyto předpoklady ještě podstatně zeslabit, nám však nejde o to vyslovit příslušná tvrzení za co nejobecnějších předpokladů, ale o to ukázat základní myšlenku a použít přitom pokud možno co nejelementárnější matematický aparát.)

Z věty 1.1 z kap. I (vyslovené pro případ soustav diferenciálních rovnic) a z výše zformulovaných předpokladů o koeficientech diferenciální rovnice (1.12) plyne, že existuje právě jedna funkce $y(x) = y(x;\alpha)$, která je v intervalu $\langle a, b\rangle$ řešením diferenciální rovnice (1.12) a pro niž platí

$$y(a) = \gamma_1, \quad y'(a) = \alpha, \tag{2.3}$$

kde α je libovolné pevně zvolené reálné číslo. Funkce y splňuje tedy danou diferenciální rovnici, jednu z okrajových podmínek (1.17) (v bodě $x = a$) a závisí na parametru α. Podaří-li se určit takovou hodnotu α^* parametru α, aby platilo

$$y(b;\alpha^*) = \gamma_2, \tag{2.4}$$

bude funkce $y(x;\alpha^*)$ řešením okrajové úlohy (1.12), (1.17). Hodnoty funkce $y(b;\alpha^*)$ proměnné α jsme schopni určovat pomocí řešení diferenciální rovnice (1.12) s počátečními podmínkami (2.3) a převedli jsme tedy původní úlohu na řešení úloh

s počátečními podmínkami a na řešení rovnice (2.4) pro α^*. Je také dostatečně jasné, proč jsme tuto metodu nazvali metodou střelby.

Všimněme si nyní řešitelnosti rovnice (2.4). Označme symbolem $z(x)$ řešení diferenciální rovnice

$$-(p(x)z')' + q(x)z = 0 \tag{2.5}$$

(tj. homogenní rovnice příslušné k rovnici (1.12)) s počátečními podmínkami

$$z(a) = 0, \quad z'(a) = 1. \tag{2.6}$$

Pak platí pro libovolné α

$$y(x;\alpha) = y(x;0) + \alpha z(x) \tag{2.7}$$

a speciálně pro $x = b$ máme $y(b;\alpha) = y(b;0) + \alpha z(b)$. Rovnice (2.4) pro určení α^* je tedy lineární a za předpokladu, že je $z(b) \neq 0$, platí

$$\alpha^* = \frac{1}{z(b)}[\gamma_2 - y(b;0)]. \tag{2.8}$$

Dosadíme-li konečně do rovnice (2.7) podle (2.8), dostáváme pro řešení $y(x;\alpha^*)$ dané okrajové úlohy vzorec

$$y(x;\alpha^*) = y(x;0) + \alpha^* z(x) \tag{2.9}$$

platný samozřejmě za předpokladu, že je $z(b) \neq 0$. O něco později ukážeme, že předpokládáme-li o koeficientech diferenciální rovnice (1.12) nerovnosti (2.1) a (2.2), je tento požadavek skutečně splněn. Řešit lineární okrajovou úlohu (1.12), (1.17) metodou střelby tedy znamená vyřešit jednu homogenní rovnici, jednu nehomogenní rovnici, obě s počátečními podmínkami a získané funkce zkombinovat podle rovnice (2.9).

Takřka na první pohled je vidět, že metoda střelby je použitelná i při obecných okrajových podmínkách (1.13). Je-li např. $\beta_1 \neq 0$, stačí za funkci $y(x;\alpha)$ vzít řešení diferenciální rovnice (1.12) splňující počáteční podmínky

$$y(a) = \frac{1}{\beta_1}[\gamma_1 + \alpha_1 p(a)\alpha], \quad y'(a) = \alpha. \tag{2.10}$$

Pak tato funkce splňuje při libovolném α první okrajovou podmínku (1.13) a k tomu, aby daná okrajová úloha byla vyřešena, stačí vypočítat α^* z rovnice

$$\alpha_2 p(b) y(b;\alpha^*) + \beta_2 y(b;\alpha^*) = \gamma_2. \tag{2.11}$$

Tato rovnice je podobně jako rovnice (2.4) opět lineární, neboť funkce $y(x;\alpha)$ se dá psát ve tvaru

$$y(x;\alpha) = y(x;0) + \alpha\frac{p(a)}{\beta_1} z(x), \tag{2.12}$$

kde z je řešením homogenní rovnice (2.5), tentokrát s počátečními podmínkami

$$z(a) = \alpha_1, \quad p(a)z'(a) = \beta_1. \tag{2.13}$$

Rovnice (2.11) tedy přejde v rovnici

$$\alpha_2 p(b)\left[y'(b;0) + \alpha\frac{p(a)}{\beta_1} z'(b)\right] + \beta_2\left[y(b;0) + \alpha\frac{p(a)}{\beta_1} z(b)\right] = \gamma_2, \tag{2.14}$$

a je-li $\alpha_2 p(b)z'(b) + \beta_2 z(b) \neq 0$ (což opět později za vhodných předpokladů dokážeme), můžeme odtud α^* vypočítat.

2.1.2 Obecná okrajová úloha

Podobně jako v případě modelové rovnice druhého řádu lze postupovat i při řešení soustavy diferenciálních rovnic (1.1) s okrajovými podmínkami (1.3). Zvolíme libovolný m-dimenzionální vektor $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_m)^T$ a řešíme soustavu (1.1) s počátečními podmínkami

$$y(a) = \boldsymbol{\alpha}. \tag{2.15}$$

Získané řešení je opět jako v případě rovnice druhého řádu funkcí parametru $\boldsymbol{\alpha}$ a označíme je $y(x;\boldsymbol{\alpha})$. Další krok spočívá v řešení soustavy m (obecně nelineárních) rovnic

$$F(\boldsymbol{\alpha}) = 0, \tag{2.16}$$

kde vektorová funkce F je definována předpisem

$$F(\boldsymbol{\alpha}) = r(\boldsymbol{\alpha}, y(b;\boldsymbol{\alpha})). \tag{2.17}$$

Označíme-li $\boldsymbol{\alpha}^*$ řešení soustavy (2.16), nalezneme řešení dané okrajové úlohy řešením soustavy diferenciálních rovnic (1.1) s počáteční podmínkou

$$y(a) = \boldsymbol{\alpha}^*. \tag{2.18}$$

Jsou-li okrajové podmínky dané okrajové úlohy separované, tj. tvaru (1.4), volíme parametr $\boldsymbol{\alpha}$ apriori tak, aby pro něj platila rovnice

$$r_1(\boldsymbol{\alpha}) = 0. \tag{2.19}$$

Funkce F na levé straně rovnice (2.16) je pak definována vztahem

$$F(\boldsymbol{\alpha}) = r_2(y(b;\boldsymbol{\alpha})), \tag{2.20}$$

takže soustava, kterou je třeba řešit, má méně rovnic a také méně neznámých než v obecném případě.

Jedním z kroků, který je třeba provést při užití této metody, je tedy řešení soustavy nelineárních rovnic (2.16), což je obecně značně obtížný problém. Tento problém odpadá v případě soustavy lineárních rovnic

$$y' = Ay + f \tag{2.21}$$

s lineárními okrajovými podmínkami (1.5). Zde A je čtvercová matice řádu m, jejíž prvky jsou spojité funkce a f je m-dimenzionální vektor se spojitými složkami. V tomto případě je funkce F tvaru

$$F(\alpha) = U\alpha + Vy(b;\alpha) - c, \tag{2.22}$$

kde vektor $y(x;\alpha)$ je řešením diferenciální rovnice (2.21) s počáteční podmínkou (2.15). Tento vektor však lze psát ve tvaru

$$y(x;\alpha) = \Phi(x)\alpha + y(x;0), \tag{2.23}$$

kde matice Φ je tzv. *fundamentální matice* soustavy diferenciálních rovnic (2.21) v bodě $x = a$, tj. je to matice, jejíž i-tý sloupec je řešením homogenní diferenciální rovnice

$$y' = Ay \tag{2.24}$$

s počátečními podmínkami

$$^{i}y(a) = 1, \quad ^{j}y(a) = 0, \quad j \neq i. \tag{2.25}$$

Řešení rovnice (2.16) je tedy dáno vzorcem

$$\alpha^* = [U + V\Phi(b)]^{-1}[c - Vy(b;0)] \tag{2.26}$$

(samozřejmě za předpokladu, že příslušná inverzní matice existuje) a k jeho získání stačí řešit jednu soustavu lineárních algebraických rovnic.

Sloupce matice Φ lze vypočítat řešením m úloh s počátečními podmínkami. Požadované řešení $y(x;\alpha)$ dané okrajové úlohy pak už není třeba získávat řešením rovnice (2.21) s počáteční podmínkou $y(a) = \alpha^*$, ale obdobně jako v případě rovnice druhého řádu lze použít vzorec

$$y(x;\alpha^*) = \Phi(x)\alpha^* + y(x;0), \tag{2.27}$$

neboť funkce, které se v něm vyskytují, jsme už v předchozím průběhu museli stejně počítat.

2.1.3 Obtíže spojené s metodou střelby

V předchozích odstavcích jsme viděli, že metoda střelby je aspoň formálně použitelná i v případě velmi obecných okrajových úloh. Upozornili jsme již, že řešení nelineární soustavy (2.16), které je při jejím užití nezbytné, může být neobyčejně obtížné. I když však od tohoto problému odhlédneme (to je možné například, řešíme-li lineární úlohu), mohou se při provádění metody střelby objevit další velmi vážné obtíže vyplývající z toho, že aritmetické operace nejsme schopni provádět přesně. Jejich podstatu nejlépe pochopíme na jednoduchém příkladě.

Příklad 2.1. Řešme metodou střelby diferenciální rovnici druhého řádu

$$-y'' + 100y = 0 \tag{2.28}$$

s lineárními separovanými okrajovými podmínkami

$$y(0) = y(10) = 1. \tag{2.29}$$

Řešení této jednoduché okrajové úlohy lze tedy nalézt ze vzorce (2.9), kde je

$$y(x;0) = \tfrac{1}{2}e^{10x} + \tfrac{1}{2}e^{-10x}, \tag{2.30}$$

$$z(x) = \tfrac{1}{20}e^{10x} - \tfrac{1}{20}e^{-10x} \tag{2.31}$$

a

$$\alpha^* = -10 + 20\frac{1 - e^{-100}}{e^{100} - e^{-100}} \tag{2.32}$$

(jde o řešení diferenciální rovnice (1.12) s počátečními podmínkami (2.3) s $\alpha = 0$, o řešení diferenciální rovnice (2.5) s počátečními podmínkami (2.6) a o určení α^* z rovnice (2.8)). Vypočteme-li α^* s relativní chybou ε, dostaneme jako aproximaci hledaného řešení funkci

$$y_\varepsilon(x) = -\tfrac{1}{2}\varepsilon e^{10x} + (1 + \tfrac{1}{2}\varepsilon)e^{-10x}. \tag{2.33}$$

Položíme-li zde např. $\varepsilon = 10^{-10}$, nabývá toto přibližné řešení v bodě $x = 10$ hodnotu

$$y_\varepsilon(10) \approx -\tfrac{1}{2}10^{-10}e^{100} \approx -1,34.10^{33}. \tag{2.34}$$

Z uvedeného jednoduchého příkladu vidíme, že i když vypočítáme hledanou počáteční hodnotu na plnou strojovou přesnost, ještě zdaleka odtud neplyne, že je možné vypočítat i přibližné řešení s přijatelnou přesností. Zároveň také vidíme, že vzniklé obtíže budou v našem jednoduchém příkladě tím větší, čím větší bude koeficient u y v rovnici (2.28) nebo čím delší bude interval, v němž hledáme řešení. Konečně je z tohoto příkladu patrné, že podstata obtíží spočívá v tom, že malá změna v počáteční podmínce způsobuje velkou změnu v řešení.

Jev, který jsme pozorovali v uvedeném příkladě, je naneštěstí pro „rozumné“ okrajové úlohy typický. „Rozumnou“ úlohou zde rozumíme úlohu, v níž malé změny ve vstupních datech mají malý vliv na řešení. Úlohy vzniklé správnou aplikací fyzikálních zákonů většinou tuto vlastnost mají. Praktická zkušenost přitom ukazuje, že je-li pro danou rovnici „rozumná“ okrajová úloha, je pro tuto rovnici úloha s počátečními podmínkami nepřirozená, a proto se tato úloha chová většinou „nerozumně“.

Obtíže metody střelby, které jsme právě popsali, i když jsou značné a mohou často její užití úplně znemožnit, nejsou jediné. V případě nelineární okrajové úlohy, i když její řešení existuje (a soustava (2.16) musí mít tedy řešení), může metoda střelby selhat z toho důvodu, že řešení diferenciální rovnice (1.1) s počáteční podmínkou $y(a) = \alpha$ existuje v celém intervalu $\langle a, b\rangle$ pouze pro ty hodnotu parametru α, které

leží v nějakém — často velice malém — okolí bodu α^*. Určení tohoto okolí však představuje většinou velmi vážný problém, takže lze jen těžko odhadnout přesnost, s níž je třeba řešit soustavu (2.16), abychom byli schopni nalézt vůbec nějaké řešení.

Důvody, které jsme uvedli, nemají čtenáři vnutit představu, že metoda střelby je v zásadě špatná. Chtěli jsme jen upozornit na některé typické problémy, které se při jejím užití mohou vyskytnout a které také vedly k tomu, že byl navržena řada jiných postupů převádějících okrajovou úlohu na úlohu s počátečními podmínkami, které si kladou za cíl aspoň některé ze zmíněných obtíží odstranit. Jeden takový postup, který navazuje přímo na metodu střelby, popíšeme v následujícím odstavci, dalších si pak všimneme v odst. 2.2 a 2.3.

2.1.4 Střelba na více cílů

V předešlém odstavci jsme ukázali, že při metodě střelby může docházet k podstatným obtížím, které jsou způsobeny často tím, že jsme nuceni řešit úlohu s počátečními podmínkami pro danou diferenciální rovnici na příliš dlouhém intervalu. Metoda, kterou popíšeme v tomto odstavci, možnost vzniku těchto obtíží většinou dosti podstatně snižuje. Její základní myšlenka spočívá v tom, že se hodnoty hledaného řešení počítají ne v jednom bodě, ale v několika bodech najednou.

Buď tedy dána okrajová úloha (1.1), (1.3) a buďte $x_0, \dots, x_n$ takové body intervalu $\langle a, b\rangle$, pro něž platí

$$a = x_0 < x_1 < \dots < x_n = b. \tag{2.35}$$

Označme $y(x; x_k, \alpha_k)$ řešení diferenciální rovnice (1.1) s počáteční podmínkou

$$y(x_k) = \alpha_k = (\alpha_{k1}, \dots, \alpha_{kn})^T. \tag{2.36}$$

Podaří-li se nalézt $n+1$ vektorů α_k, $k = 0, \dots, n$, tak, aby platilo

$$y(x_{k+1}; x_k, \alpha_k) = \alpha_{k+1}, \quad k = 0, \dots, n-1, \tag{2.37}$$

a

$$r(\alpha_0, \alpha_n) = 0, \tag{2.38}$$

je funkce $y = y(x)$ definovaná předpisem

$$y(x) = y(x; x_k, \alpha_k) \quad \text{pro } x \in \langle x_k, x_{k+1}\rangle, \ k = 0, \dots, n-1, \tag{2.39}$$

řešením dané okrajové úlohy.

Zavedeme-li označení

$$\alpha = (\alpha_0, \dots, \alpha_n)^T \tag{2.40}$$

a

$$F(\alpha) = (F_0(\alpha_0, \alpha_1)^T, F_1(\alpha_1, \alpha_2)^T, \dots, F_{n-1}(\alpha_{n-1}, \alpha_n)^T, F_n(\alpha_0, \alpha_n)^T)^T, \tag{2.41}$$

kde

$$\begin{aligned} F_0(\alpha_0, \alpha_1) &= y(x_1; x_0, \alpha_0) - \alpha_1, \\ F_1(\alpha_1, \alpha_2) &= y(x_2; x_1, \alpha_1) - \alpha_2, \\ &\vdots \\ F_{n-1}(\alpha_{n-1}, \alpha_n) &= y(x_n; x_{n-1}, \alpha_{n-1}) - \alpha_n, \\ F_n(\alpha_0, \alpha_n) &= r(\alpha_0, \alpha_n), \end{aligned} \tag{2.42}$$

dá se soustava (2.37), (2.38) $(n+1)m$ rovnice pro $(n+1)m$ neznámých $\alpha_0, \dots, \alpha_n$ psát ve tvaru

$$F(\alpha) = 0. \tag{2.43}$$

Ústředním bodem metody střelby na více cílů je tedy řešení obecně nelineární soustavy (2.43). Jejím řešením získáme hodnoty hledaného řešení dané okrajové úlohy v bodech $x_0, \dots, x_n$. Zajímá-li nás řešení ještě v dalších bodech intervalu $\langle a, b\rangle$, získáme je řešením úloh s počátečními podmínkami, jak plyne z rovnic (2.39).

Praktické zkušenosti ukazují, že metoda střelby na více cílů má ve srovnání s metodou prosté střelby většinou podstatně příznivější vlastnosti. I přesto, že o ní není teoreticky zatím mnoho známo, představuje často v případě složitých nelineárních úloh jedinou alternativu, jak se pokusit je numericky řešit.

2.2 Metoda přesunu okrajové podmínky

Tato metoda převodu okrajové úlohy na úlohy s počátečními podmínkami je vhodná především v případě lineárních okrajových úloh se separovanými okrajovými podmínkami. Popíšeme ji nejprve ve speciálním případě diferenciální rovnice druhého řádu a pak v případě obecné lineární soustavy se separovanými okrajovými podmínkami. V odst. 2.2.3 pak ukážeme, jak lze užít metodu přesunu v případě složitějších okrajových podmínek a v odst. 2.2.4 stručně upozorníme na některé problémy spojené s její numerickou realizací.

2.2.1 Diferenciální rovnice druhého řádu

Buď dána okrajová úloha (1.12), (1.13). Popisovaná metoda vychází z následující úvahy: Obecné řešení rovnice (1.12) závisí na dvou parametrech. Splňuje-li toto řešení kromě toho ještě jednu z okrajových podmínek (1.13), představuje tato podmínka vazbu mezi zmíněnými parametry. Množina řešení diferenciální rovnice (1.12), které splňují jednu z podmínek (1.13), závisí tedy na jedné konstantě. Vyloučíme-li tuto konstantu derivováním, dostaneme jako charakteristiku zmíněné množiny diferenciální rovnici prvního řádu. Následující tvrzení ukazuje, že tomu tak skutečně je.

Lemma 2.1. *Nechť funkce y splňuje v intervalu $\langle\xi_1,\xi_2\rangle$ diferenciální rovnici* (1.12), *kde p, p' a q jsou spojité funkce a p splňuje nerovnost* (2.1). *Nechť kromě toho v nějakém bodě $\xi_0 \in \langle\xi_1,\xi_2\rangle$ platí*

$$\alpha p(\xi_0)y'(\xi_0) + \beta y(\xi_0) = \gamma. \tag{2.44}$$

Nechť konečně funkce z je v intervalu $\langle\xi_1,\xi_2\rangle$ řešením diferenciální rovnice (2.5) s počátečními podmínkami

$$z(\xi_0) = -\alpha, \quad p(\xi_0)z'(\xi_0) = \beta \tag{2.45}$$

a funkce c řešením diferenciální rovnice

$$c' = fz \tag{2.46}$$

s počáteční podmínkou

$$c(\xi_0) = \gamma. \tag{2.47}$$

Pak platí

$$-z(x)p(x)y'(x) + p(x)z'(x)y(x) = c(x) \tag{2.48}$$

pro každé $x \in \langle\xi_1,\xi_2\rangle$.

D ů k a z . Funkce z, resp. c jsou diferenciálními rovnicemi (2.5), resp. (2.46) a počátečními podmínkami (2.45), resp. (2.47) jednoznačně určeny, jak plyne z modifikace věty 1.1 kap. I pro soustavy diferenciálních rovnic prvního řádu. Vynásobíme-li tedy rovnici (1.12) funkcí z a dosadíme-li do vzniklé rovnice z rovnic (2.5) a (2.46), máme

$$-[p(x)y'(x)]'z(x) + y(x)[p(x)z'(x)]' - c'(x) = 0 \tag{2.49}$$

pro $x \in \langle\xi_1,\xi_2\rangle$. Položíme-li

$$\eta(x) = -z(x)p(x)y'(x) + p(x)z'(x)y(x) - c(x), \tag{2.50}$$

plyne z rovnice (2.49), že pro každé $x \in \langle\xi_1,\xi_2\rangle$ je $\eta'(x) = 0$. Funkce η je tedy v intervalu $\langle\xi_1,\xi_2\rangle$ konstantní. Protože platí $\eta(\xi_0) = 0$, jak plyne z rovnic (2.44), (2.45) a (2.47), je $\eta(x) \equiv 0$. Lemma je dokázáno. ∎

Vzhledem k tomu, že diferenciální rovnice prvního řádu (2.48) je obdobného tvaru jako podmínka (2.44), můžeme tvrzení lemmatu 2.1 interpretovat také tak, že lineární podmínku (2.44) platnou v jednom bodě intervalu, v němž je splněna daná diferenciální rovnice, lze přesunout do libovolného jiného bodu tohoto intervalu, přičemž přesunutá podmínka bude mít tvar (2.48).

Lemma 2.1 nabízí postup řešení okrajové úlohy (1.12), (1.13): Každou z podmínek (1.13) přesuneme do téhož bodu intervalu $\langle a,b\rangle$. Dostaneme tak dvě rovnice, které svazují lineárně hodnotu hledané funkce s hodnotou její derivace v tomtéž bodě. Z těchto rovnic můžeme, alespoň za vhodných předpokladů, vypočíst funkční

hodnotu a hodnotu první derivace, a získat tak počáteční podmínky pro rovnici (1.12).

V následující větě zformulujeme základní vlastnosti právě popsané metody.

Věta 2.1. *Nechť funkce p, p', q a f jsou spojité na intervalu $\langle a,b\rangle$ a nechť platí* (2.1). *Nechť dále funkce z, resp. $\hat{z}$ jsou řešením diferenciální rovnice* (2.5) *s počátečními podmínkami*

$$z(a) = \alpha_1, \quad p(a)z'(a) = \beta_1, \tag{2.51}$$

resp.

$$\hat{z}(b) = -\alpha_2, \quad p(b)\hat{z}'(b) = \beta_2 \tag{2.52}$$

a funkce c, resp. $\hat{c}$ řešením diferenciální rovnice (2.46), *resp. diferenciální rovnice*

$$\hat{c}' = f\hat{z} \tag{2.53}$$

s počáteční podmínkou

$$c(a) = \gamma_1, \tag{2.54}$$

resp.

$$\hat{c}(b) = \gamma_2, \tag{2.55}$$

Pak lze pro každé $x_0 \in \langle a,b\rangle$ sestavit soustavu lineárních algebraických rovnic

$$\begin{bmatrix} p(x_0)z'(x_0) & -p(x_0)z(x_0) \\ p(x_0)\hat{z}'(x_0) & -p(x_0)\hat{z}(x_0) \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} c(x_0) \\ \hat{c}(x_0) \end{bmatrix} \tag{2.56}$$

pro niž platí: Má-li okrajová úloha (1.12), (1.13) *řešení y, je vektor $(y(x_0), y'(x_0))^T$ řešením soustavy* (2.56), *a naopak, má-li soustava* (2.56) *řešení $(k_1,k_2)^T$, je funkce y, kterou získáme řešením diferenciální rovnice* (1.12) *s počátečními podmínkami $y(x_0) = k_1$, $y'(x_0) = k_2$, řešením okrajové úlohy* (1.12), (1.13). *Má-li okrajová úloha* (1.12), (1.13) *jediné řešení, má i soustavy* (2.56) *jediné řešení a naopak.*

D ů k a z . Existence řešení soustavy (2.56) za předpokladu existence řešení okrajové úlohy (1.12), (1.13) plyne ihned z lemmatu 2.1, neboť první rovnice soustavy (2.56) je první podmínka (1.13) přesunutá do bodu $x = x_0$ a druhá rovnice této soustavy je druhá podmínka (1.13) přesunutá do téhož bodu.

Naopak, nechť soustava (2.56) má řešení $(k_1,k_2)^T$. Pak existuje funkce y, která je řešením diferenciální rovnice (1.12) a pro niž platí $y(x_0) = k_1$, $y'(x_0) = k_2$. Pro tuto funkci platí dále

$$-z(x_0)p(x_0)y'(x_0) + p(x_0)z'(x_0)y(x_0) = c(x_0) \tag{2.57}$$

a

$$-\hat{z}(x_0)p(x_0)y'(x_0) + p(x_0)\hat{z}'(x_0)y(x_0) = \hat{c}(x_0) \tag{2.58}$$

Přesuňme podmínku (2.57) do bodu a. Podle lemmatu 2.1 platí

$$-p(a)r(a)y'(a) + p(a)r'(a)y(y) = s(a), \tag{2.59}$$

kde r je řešením diferenciální rovnice (2.5) s počátečními podmínkami

$$r(x_0) = z(x_0), \quad r'(x_0) = \frac{p(x_0)z'(x_0)}{p(x_0)} = z'(x_0) \tag{2.60}$$

a s řešením diferenciální rovnice

$$s' = fr \tag{2.61}$$

s počáteční podmínkou

$$s(x_0) = c(x_0). \tag{2.62}$$

Protože funkce r splňuje tutéž diferenciální rovnici jako funkce z a protože pro ni platí (2.60), plyne z věty o jednoznačnosti řešení úlohy s počátečními podmínkami pro obyčejné diferenciální rovnice, že je $r(x) \equiv z(x)$. Odtud však už plyne, že je také $s(x) \equiv c(x)$. Námi sestrojená funkce y splňuje tedy vzhledem k podmínkám (2.51), (2.54) levou okrajovou podmínku. Splnění pravé okrajové podmínky se dokáže úplně stejně. Za předpokladu řešitelnosti soustavy (2.56) se nám tedy skutečně podařilo sestrojit řešení okrajové úlohy (1.12), (1.13).

Nechť nyní má okrajová úloha (1.12), (1.13) jediné řešení a předpokládejme, že soustava (2.56) má dvě různá řešení $(k_1, k_2)^T$ a $(\tilde{k}_1, \tilde{k}_2)^T$. Sestrojme funkce y_1, resp. y_2 jako řešení diferenciální rovnice (1.12) s počátečními podmínkami $y_1(x_0) = k_1$, $y_1'(x_0) = k_2$, resp. $y_2(x_0) = \tilde{k}_1$, $y_2'(x_0) = \tilde{k}_2$. Obě tyto funkce jsou podle toho, co už jsme dokázali, řešeními dané okrajové úlohy, a musí se tedy sobě identicky rovnat. To ale odporuje předpokladu, že je $(k_1, k_2)^T \neq (\tilde{k}_1, \tilde{k}_2)^T$.

Nechť konečně má soustava (2.56) právě jedno řešení a nechť okrajová úloha (1.12), (1.13) má dvě různá řešení y_1 a y_2. Pak opět z věty o jednoznačnosti řešení úlohy s počátečními podmínkami pro obyčejnou diferenciální rovnici plyne, že $\big(y_1(x_0), y_1'(x_0)\big)^T \neq \big(y_2(x_0), y_2'(x_0)\big)^T$, což je znovu spor. Věta je dokázána. ∎

Z věty 2.1 plyne, že řešitelnost okrajové úlohy (1.12), (1.13) a řešitelnost soustavy (2.56) jsou úlohy ekvivalentní.

Z hlediska získání metody pro řešení okrajové úlohy (1.12), (1.13) lze interpretovat právě dokázanou větu dvojím způsobem. Přesunem obou okrajových podmínek do téhož bodu intervalu $\langle a, b\rangle$ získáme pro hledanou funkci počáteční podmínky a hodnotu hledaného řešení v libovolném bodě intervalu $\langle a, b\rangle$ pak vypočteme řešením původní diferenciální rovnice. Tento postup jsme také měli až dosud na mysli. Pokud však povaha řešeného problému je taková, že nás řešení zajímá jen v jednom nebo několika málo bodech intervalu $\langle a, b\rangle$, může být výhodnější (tj. ekonomičtější) takový postup, že žádanou hodnotu nebo hodnoty vypočteme přímo ze soustavy (2.56). Řešení původní diferenciální rovnice pak odpadá. I tuto variantu nazveme metodou přesunu.

Upozorněme ještě, že při řešení okrajové úlohy (1.12), (1.13) metodou přesunu okrajových podmínek založeném na větě 2.1 se předem nemusíme starat o její řešitelnost. Soustavu (2.56) lze sestavit vždy a má-li řešení, nalezneme též řešení původní okrajové úlohy. Nemá-li soustava (2.56) řešení, nemá ani okrajová úloha (1.12), (1.13) řešení. Tato skutečnost také umožní nalézt elementárně podmínky, za kterých je úloha (1.12), (1.13) řešitelná při libovolných γ_1 a γ_2. K tomu cíli dokážeme nejprve následující lemmata.

Lemma 2.2. *Nechť funkce p, p' a q jsou spojité a nechť platí nerovnosti* (2.1) a (2.2). *Nechť dále α_1 a β_1 jsou nezáporná čísla, pro něž platí*

$$\alpha_1 + \beta_1 > 0. \tag{2.63}$$

Buď konečně funkce z definovaná diferenciální rovnicí (2.5) *a počátečními podmínkami* (2.51). *Pak pro každé $x \in \langle a, b\rangle$ platí*

$$z(x) \geqq \alpha_1 + \beta_1 \int_a^x \frac{\mathrm{d}\,t}{p(t)} \tag{2.64}$$

a

$$p(x)z'(x) \geqq \beta_1 + \alpha_1 \int_a^x q(t)\,\mathrm{d}\,t. \tag{2.65}$$

D ů k a z . Dokažme nejprve, že platí $z(x) \geqq 0$ a $z'(x) \geqq 0$ pro $x \in \langle a, b\rangle$. Důkaz tohoto tvrzení provedeme sporem. Položme k tomu cíli $\mathfrak{M} = \{x;\ a < x \leqq b,\ z(x) < 0\}$ a předpokládejme, že množina $\mathfrak{M}$ je neprázdná. Za tohoto předpokladu je $x_0 = \inf \mathfrak{M}$ reálné číslo ležící v intervalu $\langle a, b\rangle$. Dále z počátečních podmínek pro funkci z ihned plyne, že je $a < x_0 \leqq b$. Skutečně, je-li $\alpha_1 > 0$, plyne to ze spojitosti funkce z; je-li $\alpha_1 = 0$, je v důsledku předpokladu (2.63) $\beta_1 > 0$ a funkce z je v bodě $x = a$ rostoucí. Existuje tedy $\delta > 0$ takové, že je $z(x) > 0$ pro $x \in (a, a + \delta)$. Konečně je $z(x_0) = 0$, neboť z definice čísla x_0 plyne, že je $z(x_0) \leqq 0$ a ostrá nerovnost odporuje definici čísla x_0 a spojitosti funkce z. Pro $x \leqq x_0$ zřejmě platí $z(x) \geqq 0$. Integrací rovnice (2.5) v mezích od a do x dostáváme

$$p(x)z'(x) = \beta_1 + \int_a^x q(t)z(t)\,\mathrm{d}\,t. \tag{2.66}$$

Odtud však ihned plyne, že je i $z'(x) \geqq 0$ pro $x \leqq x_0$. Je tedy speciálně $z'(x_0) \geqq 0$. Kdyby bylo $z'(x_0) > 0$, byla by funkce z v bodě $x = x_0$ rostoucí, což odporuje definici čísla x_0. Celkem je tedy $z(x_0) = z'(x_0) = 0$. Odtud ale plyne, že je $z(x) \equiv 0$, neboť funkce z je řešením lineární diferenciální rovnice (2.5), pro niž platí věta o jednoznačnosti. To je však ve sporu s předpokladem $\alpha_1 \geqq 0$, $\beta_1 \geqq 0$, $\alpha_1 + \beta_1 > 0$. Obdržený spor dokazuje, že množina $\mathfrak{M}$ je prázdná, a tedy je skutečně $z(x) \geqq 0$ pro $x \in \langle a, b\rangle$. Z rovnice (2.66) pak plyne, že na tomto intervalu je také $z'(x) \geqq 0$.

Protože je $z'(x) \geqq 0$, je funkce z neklesající a platí tedy

$$z(x) \geqq \alpha_1. \tag{2.67}$$

Použijeme-li tuto nerovnost v (2.66), dostáváme nerovnost (2.65).

Z nerovnosti (2.65) však plyne, že je

$$z'(x) \geqq \frac{\beta_1}{p(x)}, \tag{2.68}$$

neboť funkce q je nezáporná. Z této nerovnosti dostáváme integrací ihned nerovnost (2.64). Lemma je dokázáno.

Lemma 2.3. *Nechť funkce p, p' a q splňují tytéž předpoklady jako v lemmatu 2.2 a nechť α_2 a β_2 jsou nezáporná čísla, pro něž platí*

$$\alpha_2 + \beta_2 > 0. \tag{2.69}$$

Pak pro řešení $\hat{z}$ diferenciální rovnice (2.5) s počátečními podmínkami (2.52) platí pro každé $x \in \langle a, b\rangle$ nerovnosti

$$\hat{z}(x) \leqq -\alpha_2 - \beta_2 \int_x^b \frac{\mathrm{d}t}{p(t)} \tag{2.70}$$

a

$$p(x)\hat{z}'(x) \geqq \beta_2 + \alpha_2 \int_x^b q(t)\,\mathrm{d}t. \tag{2.71}$$

Důkaz se provede zcela analogicky jako důkaz lemmatu 2.2.

Lemma 2.4. *Nechť jsou splněny předpoklady lemmat 2.2 a 2.3. Pak platí implikace*

$$\det \begin{bmatrix} p(b)z'(b) & -z(b) \\ \beta_2 & \alpha_2 \end{bmatrix} = 0 \implies \beta_1 = \beta_2 = 0 \text{ a } q(x) \equiv 0 \tag{2.72}$$

a

$$\det \begin{bmatrix} p(a)\hat{z}'(a) & -\hat{z}(a) \\ \beta_1 & -\alpha_1 \end{bmatrix} = 0 \implies \beta_1 = \beta_2 = 0 \text{ a } q(x) \equiv 0. \tag{2.73}$$

D ů k a z . Pro determinant na levé straně implikace (2.72) platí podle lemmatu 2.2

$$\det \begin{bmatrix} p(b)z'(b) & -z(b) \\ \beta_2 & \alpha_2 \end{bmatrix} = \alpha_2 p(b)z'(b) + \beta_2 z(b) \geqq \tag{2.74}$$

$$\geqq \alpha_2\beta_1 + \alpha_1\beta_2 + \alpha_1\alpha_2 \int_a^b q(t)\,\mathrm{d}t + \beta_1\beta_2 \int_a^b \frac{\mathrm{d}t}{p(t)},$$

což je součet nezáporných čísel. Je-li tedy determinant na levé straně nerovnosti (2.74) roven nule, rovnají se nule všechny sčítance na pravé straně, tj. platí

$$\alpha_2\beta_1 = \alpha_1\beta_2 = \alpha_1\alpha_2 \int_a^b q(t)\,\mathrm{d}t = \beta_1\beta_2 \int_a^b \frac{1}{p(t)}\,\mathrm{d}t = 0. \tag{2.75}$$

Podle (2.63) a (2.69) je $0 < (\alpha_1+\beta_1)(\alpha_2+\beta_2) = \alpha_1\alpha_2+\alpha_1\beta_2+\beta_1\alpha_2+\beta_1\beta_2 = \alpha_1\alpha_2$. Je tedy $\alpha_1 > 0$ a $\alpha_2 > 0$ a z rovnic (2.75) plyne, že $\beta_1 = \beta_2 = 0$ a $\int_a^b q(t)\,\mathrm{d}t = 0$. Z poslední rovnosti však plyne, že je $q \equiv 0$, neboť funkce q je nezáporná a spojitá. Implikace (2.73) se dokáže stejně snadno. Lemma je dokázáno.

Věta 2.2. *Nechť funkce p, p', q a f jsou spojité v intervalu $\langle a, b\rangle$ a nechť platí nerovnosti (2.1) a (2.2). Nechť koeficienty α_i a β_i jsou nezáporné a nechť pro ně platí nerovnosti (2.63) a (2.69). Konečně nechť není současně $\beta_1 = \beta_2 = 0$ a $q \equiv 0$. Pak okrajová úloha (1.12), (1.13) má při libovolných γ_1 a γ_2 právě jedno řešení.*

D ů k a z . Podle věty 2.1 je řešitelnost dané okrajové úlohy ekvivalentní s řešitelností soustavy

$$\begin{bmatrix} p(b)z'(b) & -p(b)z(b) \\ \beta_2 & p(b)\alpha_2 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} c(b) \\ \gamma_2 \end{bmatrix}. \tag{2.76}$$

(Ve větě 2.1 jsme položili $x_0 = b$.) Podle lemmatu 2.4 je však determinant této soustavy od nuly různý; soustava (2.76) má tedy při libovolné pravé straně právě jedno řešení. Tvrzení věty tedy plyne z věty 2.1.

Věta 2.2 udává jednoduché postačující podmínky pro existenci a jednoznačnost řešení okrajové úlohy (1.12), (1.13). Abychom se v dalším nemuseli příliš často opakovat, umluvíme se, že pokud budeme mluvit o této úloze, budeme předpokládat, že předpoklady věty 2.2 jsou splněny, aniž to vždy budeme konkrétně uvádět.

Z věty 2.1 také snadno plyne následující poznámka, která pokrývá případ vyloučený ve větě 2.2.

Poznámka 2.1. Řešení diferenciální rovnice

$$-(p(x)y')' = f(x) \tag{2.77}$$

s okrajovými podmínkami

$$\begin{aligned} -p(a)y'(a) &= \gamma_1, \\ p(b)y'(b) &= \gamma_2 \end{aligned} \tag{2.78}$$

(tzv. *Neumannovými podmínkami*) existuje tehdy a jen tehdy, platí-li

$$\gamma_1 + \gamma_2 + \int_a^b f(x)\,\mathrm{d}x = 0. \tag{2.79}$$

Je-li podmínka (2.79) splněna, je řešení určeno jednoznačně až na aditivní konstantu.

Než ukážeme, jak lze myšlenku přesunu okrajové podmínky přenést na obecnou soustavu lineárních diferenciálních rovnic, všimneme si ještě, jak překlenout mezeru, která zůstala v odvození metody střelby pro rovnici druhého řádu v odst. 2.1.1. Tato metoda byla odvozena za předpokladu nenulovosti čísel $z(b)$, resp. $\alpha_2 p(b)z'(b) + \beta_2 z(b)$, kde funkce z byla v intervalu $\langle a, b\rangle$ řešením diferenciální rovnice (2.5)

s počátečními podmínkami (2.6), resp. (2.13). Jsou-li však splněny předpoklady věty 2.2, plyne tato skutečnost ihned z lemmatu 2.4.

Z této úvahy také vidíme, že metoda střelby a metoda přesunu okrajové podmínky jsou si značně blízké, neboť v obou případech počítáme podobné veličiny. K této poznámce se ještě vrátíme v odst. 2.2.4.

2.2.2 Obecná soustava lineárních diferenciálních rovnic

V tomto odstavci si všimneme, jak lze jednoduše formulovat metodu přesunu okrajových podmínek pro soustavu lineárních diferenciálních rovnic (2.21) se separovanými lineárními okrajovými podmínkami (1.6). O prvcích matice A a o složkách vektoru f budeme přitom předpokládat, že jsou to funkce spojité v intervalu $\langle a, b\rangle$, takže jsme oprávněni užívat větu o existenci a jednoznačnosti pro úlohu s počátečními podmínkami pro tuto soustavu. Kromě toho budeme předpokládat, že okrajové podmínky jsou nezávislé, což znamená, že hodnosti matic V_1 a V_2 (typů $m_1 \times m$ a $m_2 \times m$ s $m_1 + m_2 = m$) jsou rovny počtu jejich řádků.

Metoda přesunu okrajových podmínek se opírá o následující lemma, které je přímým zobecnění lemmatu 2.1.

Lemma 2.5. *Nechť vektor y splňuje v intervalu $\langle \xi_1, \xi_2\rangle \subset \langle a, b\rangle$ diferenciální rovnici* (2.21) *a nechť v nějakém bodě $\xi_0 \in \langle \xi_1, \xi_2\rangle$ platí*

$$V_0 y(\xi_0) = v_0, \tag{2.80}$$

kde V_0 je obdélníková matice typu $m_0 \times m$ ($m_0 \leqq m$) a v_0 je daný m_0-dimenzionální vektor. Buď dále $R(x)$, $x \in \langle \xi_1, \xi_2\rangle$, matice typu $m_0 \times m$, která je definovaná v intervalu $\langle \xi_1, \xi_2\rangle$ (maticovou) diferenciální rovnicí

$$R' = -RA(x) \tag{2.81}$$

s počáteční podmínkou

$$R(\xi_0) = V_0. \tag{2.82}$$

Buď konečně $r(x)$ m_0-dimenzionální vektor definovaný v intervalu $\langle \xi_1, \xi_2\rangle$ diferenciální rovnicí

$$r' = R(x) f(x) \tag{2.83}$$

s počáteční podmínkou

$$r(\xi_0) = v_0. \tag{2.84}$$

Pak pro každé $x \in \langle \xi_1, \xi_2\rangle$ platí

$$R(x) y(x) = r(x). \tag{2.85}$$

D ů k a z . Předně rovnice (2.81) s podmínkou (2.82) skutečně jednoznačně definuje matici R, neboť jde vlastně o soustavu $m_0 m$ lineárních diferenciálních rovnic

prvního řádu pro $m_0 m$ neznámých prvků této matice. Totéž platí i o vektoru r. Vynásobíme-li rovnici (2.21) maticí R a použijeme-li rovnic (2.81) a (2.83), dostaneme

$$[R(x) y(x) - r(x)]' = 0. \tag{2.86}$$

Vektor $R(x) y(x) - r(x)$ je tedy v intervalu $\langle \xi_1, \xi_2\rangle$ konstantní a vzhledem k podmínce (2.80) je nulový. Lemma je dokázáno.

Postupem zcela analogickým, jakým jsme dokázali větu 2.1, se na základě lemmatu 2.5 dokáže i následující věta.

Věta 2.3. *Nechť prvky matice A a složky vektoru f jsou spojité na intervalu $\langle a, b\rangle$. Nechť dále matice $R(x)$, resp. $\hat{R}(x)$ typu $m_1 \times m$, resp. $m_2 \times m$ jsou definovány diferenciální rovnicí*

$$R' = -RA(x) \tag{2.87}$$

s počáteční podmínkou

$$R(a) = V_1, \tag{2.88}$$

resp. diferenciální rovnicí

$$\hat{R}' = -\hat{R}A(x) \tag{2.89}$$

s počáteční podmínkou

$$\hat{R}(b) = V_2. \tag{2.90}$$

Nechť konečně $r(x)$, resp. $\hat{r}(x)$ jsou m_1-, resp. m_2-dimenzionální vektory definované diferenciální rovnicí

$$r' = R(x) f(x) \tag{2.91}$$

s počáteční podmínkou

$$r(a) = v_1, \tag{2.92}$$

resp. diferenciální rovnicí

$$\hat{r}' = \hat{R}(x) f(x) \tag{2.93}$$

s počáteční podmínkou

$$\hat{r}(b) = v_2. \tag{2.94}$$

Utvoříme-li pak pro libovolné $x_0 \in \langle a, b\rangle$ soustavu m lineárních algebraických rovnic

$$\begin{bmatrix} R(x_0) \\ \hat{R}(x_0) \end{bmatrix} k = \begin{bmatrix} r(x_0) \\ \hat{r}(x_0) \end{bmatrix}, \tag{2.95}$$

platí: má-li okrajová úloha (2.21), (1.6) *řešení* $y(x)$, *je vektor* $y(x_0)$ *řešením soustavy* (2.95), a *naopak, je-li vektor* k *řešením soustavy* (2.95), *je vektor* $y(x)$, *který získáme řešením soustavy* (2.21) *s počáteční podmínkou* $y(x_0) = k$, *řešením okrajové úlohy* (2.21), (1.6). *Má-li okrajová úlohy* (2.21), (1.6) *jediné řešení, má i soustava* (2.95) *jediné řešení a naopak.*

Poznámka 2.2. Z právě uvedené věty plyne následující jednoduché, často však velmi užitečné tvrzení: Nechť okrajová úloha (2.21), (1.6) má při libovolném f a při libovolných v_1 a v_2 nanejvýš jedno řešení a nechť je $m_1 + m_2 = m$. Pak má tato úloha při libovolných f, v_1 a v_2 právě jedno řešení. Skutečně, podle věty 2.3 stačí ukázat, že determinant soustavy (2.95) je od nuly různý. Kdyby však byl tento determinant roven nule, měla by homogenní soustavy příslušná k soustavě (2.95) netriviální řešení a okrajová úloha (2.21), (1.6) s $f = 0$ a $v_1 = v_2 = 0$ by měla kromě triviálního řešení ještě také netriviální řešení.

Na základě věty 3.2 se dají formulovat algoritmy pro řešení okrajové úlohy (2.21), (1.6), které jsou zcela paralelní k algoritmům pro rovnici druhého řádu popsaným v odst. 2.2.1.

2.2.3 Svázané a integrální okrajové podmínky

V tomto odstavci ukážeme, jak lze některé úlohy s obecnějšími okrajovými podmínkami, než jsou lineární separované podmínky, převést na úlohu vyšetřovanou v předešlém odstavci.

Nejprve si všimneme diferenciální rovnice (2.21) s obecnými lineárními okrajovými podmínkami (1.5). Tato úloha se převede na úlohu se separovanými podmínkami následujícím jednoduchým obratem.

Řešíme v intervalu $\langle a, (a+b)/2\rangle$ soustavu $2m$ diferenciálních rovnic

$$u' = \begin{bmatrix} A(x) & 0_m \\ 0_m & -A(a+b-x) \end{bmatrix} u + \begin{bmatrix} f(x) \\ -f(a+b-x) \end{bmatrix}, \tag{2.96}$$

kde u je $2m$-dimenzionální vektor a 0_m je čtvercová nulová matice řádu m, se separovanými lineárními okrajovými podmínkami

$$\begin{aligned} [U, V]u(a) &= c, \\ [I_m, -I_m]u\big((a+b)/2\big) &= 0, \end{aligned} \tag{2.97}$$

kde I_m značí jednotkovou matici řádu m. Hledaný vektor $y(x) = [{}^1y(x), \ldots, {}^my(x)]^T$ je pak dán rovnicemi

$${}^iy(x) = {}^iu(x), \quad i = 1, \ldots, m, \tag{2.98}$$

pro $x \in \langle a, (a+b)/2\rangle$ a rovnicemi

$${}^iy(x) = {}^{m+i}u(a+b-x), \quad i = 1, \ldots, m, \tag{2.99}$$

pro $x \in \langle (a+b)/2, b\rangle$.

Okrajovou úlohu se svázanými okrajovými podmínkami se tedy podařilo převést na okrajovou úlohu se separovanými okrajovými podmínkami. Počet rovnic však přitom vzrostl na dvojnásobek.

Jak druhý příklad uvažujme okrajovou úlohu, kdy kromě okrajových podmínek typu (1.6) je zadána ještě podmínka tvaru

$$\int_a^b D(x) y(x)\, \mathrm{d}x = c, \tag{2.100}$$

kde $D(x)$ je daná obdélníková matice typu $m_0 \times m$ se spojitými prvky a c je daný konstantní m_0-dimenzionální vektor. Doplňující podmínka tohoto typu se praxi vyskytuje např. při úlohách na vlastní čísla.

Nalezneme-li řešení soustavy $m_0 + m$ diferenciálních rovnic

$$\begin{bmatrix} y \\ u \end{bmatrix}' = \begin{bmatrix} A(x) & 0 \\ D(x) & 0 \end{bmatrix} \begin{bmatrix} y \\ u \end{bmatrix} + \begin{bmatrix} f \\ 0 \end{bmatrix} \tag{2.101}$$

se separovanými lineárními okrajovými podmínkami

$$\begin{aligned} \begin{bmatrix} V_1 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} y(a) \\ u(a) \end{bmatrix} &= \begin{bmatrix} v_1 \\ 0 \end{bmatrix}, \\ \begin{bmatrix} V_2 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} y(b) \\ u(b) \end{bmatrix} &= \begin{bmatrix} v_2 \\ c \end{bmatrix} \end{aligned} \tag{2.102}$$

kde y je m-dimenzionální vektor, u je m_0-dimenzionální vektor a typy nulových a jednotkových matic 0 a I jsou zřejmé ze souvislosti, je složka y nalezeného řešení řešením původní diferenciální rovnice (2.21), které splňuje okrajové podmínky (1.6) a integrální podmínku (2.100).

2.2.4 Obtíže spojené s metodou přesunu okrajové podmínky

Metoda přesunu okrajových podmínek je jednoduchá, elegantní a má, alespoň teoreticky, významnou přednost v tom, že zároveň dává odpověď na otázku, zda řešení dané okrajové úlohy existuje. Praktické zkušenosti však ukazují, že při její konkrétní realizaci dochází často k značným obtížím. Kořen těchto obtíží je stejný, jako tomu bylo u metody střelby, jak je vlastně více méně zřejmé, neboť při metodě přesunu okrajových podmínek řešíme v podstatě tytéž úlohy s počátečními podmínkami jako u metody střelby.

Ilustrujme typické jevy, které mohou nastat při metodě přesunu na následujícím jednoduchém příkladě.

Příklad 2.2. Řešme v intervalu $\langle 0, 1\rangle$ diferenciální rovnici

$$-[(1+x)y']' + qy = [q + \pi^2(1+x)] \sin \pi x - \pi \cos \pi x \tag{2.103}$$

s Dirichletovými okrajovými podmínkami $y(0) = y(1) = 0$. Předpoklady existenční věty 2.2 jsou zřejmě splněny a přesné řešení této okrajové úlohy je funkce $\sin \pi x$.

V tab. 2.1 jsou uvedeny výsledky výpočtu pro $q = 100$ a $q = 500$. Potřebné úlohy s počátečními podmínkami byly řešeny standardní Rungovou-Kuttovou metodou s integračním krokem $h = 0,0025$. Z tabulky vidíme, že zejména v případě $q = 500$ jsou výsledky zcela neuspokojivé. Pro srovnání uvádíme v tab. 2.2 ještě řešení téže okrajové úlohy metodou střelby. I touto metodou dostáváme výsledky stejně špatné.

Tabulka 2.1

Řešení okrajové úlohy pro diferenciální rovnici (2.103) metodou přesunu

x	$q = 100$		$q = 500$	
	přibližné řešení	chyba	přibližné řešení	chyba
0,0	0,000 275	0,000 275	163,484 000	163,484 000
0,1	0,309 114	0,000 097	18,324 700	18,015 700
0,2	0,587 819	0,000 034	2,779 730	2,191 950
0,3	0,809 028	0,000 011	1,099 890	0,290 869
0,4	0,951 059	0,000 003	0,992 730	0,041 673
0,5	1,000 000	0,000 000	1,006 387	0,006 387
0,6	0,951 056	0,000 000	0,952 091	0,001 034
0,7	0,809 017	0,000 000	0,809 189	0,000 172
0,8	0,587 785	0,000 000	0,587 813	0,000 028
0,9	0,309 017	0,000 000	0,309 021	0,000 003
1,0	0,000 000	0,000 000	0,000 000	0,000 000

V následujícím odstavci ukážeme, jak lze neuspokojivé chování metody přesunu, které jsme ilustrovali v právě uvedeném příkladě, podstatně zlepšit.

2.3 Metoda normalizovaného přesunu

Jde o modifikaci metody přesunu okrajové podmínky, která je motivována snahou odstranit obtíže spojené s praktickou realizací této metody. Při jejím popisu začneme opět s diferenciální rovnicí druhého řádu.

2.3.1 Diferenciální rovnice druhého řádu

Buď dána okrajová úloha (1.12), (1.13). V odst. 2.2.1 jsme ukázali, že přesuneme-li levou a pravou okrajovou podmínku pomocí rovnic (2.56) do téhož bodu intervalu $\langle a, b\rangle$, jsme schopni získat počáteční podmínky hledaného řešení řešením soustavy dvou rovnic o dvou neznámých. Máme-li na mysli jako konkrétní příklad okrajovou úlohu (2.28), (2.29), kterou jsme užili v odst. 2.1.3 k ilustraci potíží spojených

Tabulka 2.2

Řešení okrajové úlohy pro diferenciální rovnici (2.103) metodou střelby

x	$b = 100$		$b = 500$	
	přibližné řešení	chyba	přibližné řešení	chyba
0,0	0	0	0	0
0,1	0,309 016	−0,000 001	0,309 012	−0,000 005
0,2	0,587 781	−0,000 004	0,587 774	−0,000 011
0,3	0,809 011	−0,000 005	0,809 006	−0,000 011
0,4	0,951 039	−0,000 017	0,950 562	−0,000 495
0,5	0,999 936	−0,000 064	0,997 559	−0,002 441
0,6	0,950 989	−0,000 068	0,953 125	0,002 068
0,7	0,808 945	−0,000 072	0,710 938	−0,098 080
0,8	0,587 387	−0,000 398	0,625 000	0,037 215
0,9	0,307 861	−0,001 156	0	−0,309 018
1,0	−0,000 244	−0,000 224	−1,000 000	−1,000 000

s metodou střelby, a přesouváme-li např. pravou okrajovou podmínku, jsou příslušné funkce $\hat{z}$ a $\hat{c}$ dány vzorci

$$\hat{z}(x) = \tfrac{1}{20}\mathrm{e}^{10(x-10)} - \tfrac{1}{20}\mathrm{e}^{-10(x-10)} \tag{2.104}$$

a

$$\hat{c}(x) = 1. \tag{2.105}$$

Soustava (2.56) (s $x_0 = a = 0$) je tedy tvaru

$$\begin{bmatrix} 1 & 0 \\ \frac{1}{2}\mathrm{e}^{100} + \frac{1}{2}\mathrm{e}^{-100} & \frac{1}{20}\mathrm{e}^{100} - \frac{1}{20}\mathrm{e}^{-100} \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{2.106}$$

Neznámé k_1 a k_2, tj. hodnoty přesného řešení dané okrajové úlohy a jeho derivace v bodě $x = 0$ vypočteme z této soustavy sice v rámci běžné strojové přesnosti přesně ($k_1 = 1$ a $k_2 = -10$ dokonce na více než 42 platných dekadických cifer), tato přesnost však ani zdaleka nestačí k výpočtu uspokojivé aproximace, užijeme-li je jako počáteční podmínky pro rovnici (2.28) (srv. příklad 2.1 na str. 129). Čtenář se už sám může snadno přesvědčit, že obtíže zůstanou stejné, budeme-li přesouvat levou okrajovou podmínku.

Je zde však ještě další okolnost, která může realizaci metody přesunu okrajové podmínky (a ostatně i metodu střelby, protože se v ní řeší tytéž rovnice) znemožnit ještě dříve, než sestrojíme soustavu (2.56). Tato okolnost spočívá v tom, že prvky matice této soustavy mohou být značně velké (u soustavy (2.106) jsou řádu 10^{43}),

takže může dojít k přeplnění dříve než dospějeme s řešením příslušných rovnic do žádaného bodu. Následující věty udávají vodítko, jak se těchto obtíží zbavit. Navíc pak nabídnou možnost, jak obejít nutnost řešit úlohu s počátečními podmínkami pro rovnici (1.12), která může být neúnosně citlivá na malé chyby ve vstupních datech.

Věta 2.4. *Nechť p, p', q a f jsou spojité funkce v intervalu $\langle a, b\rangle$ a nechť platí* (2.1) a (2.2). *Nechť funkce y splňuje v intervalu $\langle a, b\rangle$ diferenciální rovnici* (1.12) *a první okrajovou podmínku* (1.13) *a nechť platí $\alpha_1 \geqq 0$, $\beta_1 \geqq 0$ a $\alpha_1 + \beta_1 > 0$. Pak platí:*

(i) *Je-li $\alpha_1 > 0$, je*

$$-p(x)y'(x) + \varphi(x)y(x) = v(x) \tag{2.107}$$

pro každé $x \in \langle a, b\rangle$, kde funkce φ je řešením diferenciální rovnice

$$\varphi' = -\frac{1}{p(x)}\varphi^2 + q(x) \tag{2.108}$$

s počáteční podmínkou

$$\varphi(a) = \frac{\beta_1}{\alpha_1} \tag{2.109}$$

a funkce v je řešením diferenciální rovnice

$$v' = -\frac{\varphi(x)}{p(x)} + f(x) \tag{2.110}$$

s počáteční podmínkou

$$v(a) = \frac{\gamma_1}{\alpha_1}. \tag{2.111}$$

(ii) *Je-li $\beta_1 > 0$, je*

$$-\psi(x)p(x)y'(x) + y(x) = u(x) \tag{2.112}$$

pro každé $x \in \langle a, b\rangle$, kde funkce ψ je řešením diferenciální rovnice

$$\psi' = -q(x)\psi^2 + \frac{1}{p(x)} \tag{2.113}$$

s počáteční podmínkou

$$\psi(a) = \frac{\alpha_1}{\beta_1} \tag{2.114}$$

a funkce u řešením diferenciální rovnice

$$u' = -q(x)\psi(x)u + f(x)\psi(x) \tag{2.115}$$

s počáteční podmínkou

$$u(a) = \frac{\gamma_1}{\beta_1}. \tag{2.116}$$

D ů k a z . Buď z řešení diferenciální rovnice (2.5) s počátečními podmínkami (2.51) a c řešení diferenciální rovnice (2.46) s počáteční podmínkou (2.54). Je-li $\alpha_1 > 0$, pak podle lemmatu 2.2 je $z(x) \geqq \alpha_1 > 0$ pro každé $x \in \langle a, b\rangle$, a je tedy možno rovnici $-z(x)p(x)y'(x) + p(x)z'(x)y(x) = c(x)$ platnou podle lemmatu 2.1 pro každé $x \in \langle a, b\rangle$ dělit funkcí z. Položíme-li tedy $\varphi(x) = p(x)z'(x)/z(x)$ a $v(x) = c(x)/z(x)$, platí rovnice (2.107). Funkce φ a v jsou však zřejmě v intervalu $\langle a, b\rangle$ spojitě diferencovatelné a přímým výpočtem se snadno zjistí, že splňují diferenciální rovnice (2.108), resp. (2.110) s počátečními podmínkami (2.109), resp. (2.111).

Je-li $\beta_1 > 0$, dává lemma 2.2 nerovnost $p(x)z'(x) \geqq \beta_1 > 0$. Položíme-li $\psi(x) = z(x)/[p(x)z'(x)]$ a $u(x) = c(x)/[p(x)z'(x)]$, dokážeme stejně snadno jako výše platnost tvrzení (ii). Věta je dokázána.

Věta 2.5. *Nechť p, p', q a f jsou spojité funkce v intervalu $\langle a, b\rangle$ a nechť platí* (2.1) a (2.2). *Nechť funkce y splňuje v intervalu $\langle a, b\rangle$ diferenciální rovnici* (1.12) *a druhou okrajovou podmínku* (1.13) *a nechť $\alpha_2 \geqq 0$, $\beta_2 \geqq 0$, $\alpha_2 + \beta_2 > 0$. Pak platí*

(i) *Je-li $\alpha_2 > 0$, je*

$$-p(x)y'(x) + \hat{\varphi}(x)y(x) = \hat{v}(x) \tag{2.117}$$

pro každé $x \in \langle a, b\rangle$, kde funkce $\hat{\varphi}$ je řešením diferenciální rovnice

$$\hat{\varphi}' = -\frac{1}{p(x)}\hat{\varphi}^2 + q(x) \tag{2.118}$$

s počáteční podmínkou

$$\hat{\varphi}(b) = -\frac{\beta_2}{\alpha_2} \tag{2.119}$$

a funkce $\hat{v}$ je řešením diferenciální rovnice

$$\hat{v}' = -\frac{\hat{\varphi}(x)}{p(x)}v + f(x) \tag{2.120}$$

s počáteční podmínkou

$$\hat{v}(b) = -\frac{\gamma_2}{\alpha_2}. \tag{2.121}$$

(ii) *Je-li $\beta_2 > 0$, je*

$$-\hat{\psi}(x)p(x)y'(x) + y(x) = \hat{u}(x) \tag{2.122}$$

pro každé $x \in \langle a, b\rangle$, kde funkce $\hat{\psi}$ je řešením diferenciální rovnice

$$\hat{\psi}' = -q(x)\hat{\psi}^2 + \frac{1}{p(x)} \tag{2.123}$$

s počáteční podmínkou

$$\hat{\psi}(b) = -\frac{\alpha_2}{\beta_2} \tag{2.124}$$

a funkce $\hat{u}$ je řešením diferenciální rovnice

$$\hat{u}' = -q(x)\hat{\psi}(x)\hat{u} + f(x)\hat{\psi}(x) \tag{2.125}$$

s počáteční podmínkou

$$\hat{u}(a) = \frac{\gamma_2}{\beta_2}. \tag{2.126}$$

D ů k a z . Větu 2.5 dokážeme pomocí lemmatu 2.3 stejně snadno, jako jsme dokázali větu 2.4 pomocí lemmatu 2.2.

Z věty 2.4 je vidět že lineární okrajovou podmínku typu (1.13) lze přesouvat pomocí rovnice (2.107), resp. (2.112), v níž je koeficient u $p(x)y'(x)$, resp. u $y(x)$ roven minus jedné nebo jedné. Tento tvar přesunuté podmínky vznikl normalizací rovnice $-z(x)p(x)y'(x)+p(x)z'(x)y(x) = c(x)$, a dá se proto očekávat, že koeficient φ, resp. ψ neporoste tak nebezpečně, jako tomu bylo u koeficientů z a pz'. Totéž samozřejmě platí i pro druhou okrajovou podmínku. Jen je třeba vycházet místo z věty 2.4 z věty 2.5.

Nyní už je snadné sestavit algoritmus pro řešení dané okrajové úlohy založený na normalizovaném přesunu okrajové podmínky. Přitom lze vycházet stejně dobře z věty 2.4 jako z věty 2.5.

Začněme prvním případem. Nechť jsou splněny předpoklady existenční věty 2.2 a buď nejprve $\alpha_1 > 0$. Pak podle věty 2.4 platí pro řešení dané okrajové úlohy speciálně

$$-p(b)y'(b) + \varphi(b)y(b) = v(b). \tag{2.127}$$

Protože platí zároveň i druhá okrajová podmínka (1.13), můžeme z rovnic (1.13) a (2.127) vypočítat hodnotu hledaného řešení v bodě $x = b$. Dostaneme

$$y(b) = \frac{\gamma_2 + \alpha_2 v(b)}{\beta_2 + \alpha_2\varphi(b)} \tag{2.128}$$

a tento vzorec má smysl, neboť je $\beta_2 + \alpha_2\varphi(b) = [\beta_2 z(b) + \alpha_2 p(b)z'(b)]/z(b) \neq 0$ podle věty 2.2. Kdybychom ze zmíněných rovnic vypočetli ještě hodnotu $y'(b)$, měli bychom k dispozici počáteční podmínky pro původní diferenciální rovnici (1.12). Jejím řešením se získanými počátečními podmínkami bychom dostali, aspoň teoreticky, řešení okrajové úlohy. Provedením tohoto postupu bychom se však zbavili pouze jedné nepříjemné vlastnosti metody prostého přesunu, totiž růstu prvků matice (2.56). Druhá potíž spočívající v citlivosti původní diferenciální rovnice na počátečních podmínkách by zůstala jako dříve. Protože však řešení dané okrajové úlohy splňuje rovnici (2.107) pro každé $x \in \langle a, b\rangle$, což je diferenciální rovnice prvního řádu, zdá se, že řešení dané okrajové úlohy je možné dostat také tak, že řešíme

diferenciální rovnici (2.107) s počáteční podmínkou (2.128). Později uvidíme, že tento postup bude dokonce v jistém smyslu výhodný.

Podobně postupujeme i v případě, že je $\beta_1 > 0$. Z rovnice (2.112) psané pro $x = b$ a z druhé okrajové podmínky (1.13) dostaneme

$$y(b) = \frac{\gamma_2\psi(b) + \alpha_2 u(b)}{\beta_2\psi(b) + \alpha_2}. \tag{2.129}$$

Použít tuto hodnotu jako počáteční podmínku přímo pro rovnici (2.112) je sice v zásadě možné, mohou však přitom vzniknout jisté problémy, protože na základě dosud uvedených tvrzení není jasné, zda nemůže nastat degenerace rovnice (2.112) v důsledku anulování koeficientu ψ. Proto je rozumnější počítat místo funkce y funkci

$$p(x)y'(x) = w(x) \tag{2.130}$$

z diferenciální rovnice

$$-w' + q(x)\psi(x)w = f(x) - q(x)u(x) \tag{2.131}$$

s počáteční podmínkou

$$w(b) = \frac{\gamma_2 - \beta_2 u(b)}{\alpha_2 + \beta_2\psi(b)} \tag{2.132}$$

a funkci y pak vypočítat z rovnice

$$y(x) = u(x) + \psi(x)w(x). \tag{2.133}$$

Diferenciální rovnice (2.131) vznikla z diferenciální rovnice (1.12) dosazením za $p(x)z'(x)$ podle (2.130) a za y podle (2.133) a počáteční podmínka (2.132) vznikla řešením rovnice (2.112) psané pro $x = b$ a druhé rovnice (1.13) pro neznámou $p(b)y'(b) = w(b)$.

Následující věta právě popsanou metodu ospravedlňuje.

Věta 2.6. *Nechť jsou splněny předpoklady věty 2.2. Buď $\alpha_1 > 0$ a buď funkce φ definovaná diferenciální rovnicí* (2.108) *s počáteční podmínkou* (2.109), *funkce v diferenciální rovnicí* (2.110) *s počáteční podmínkou* (2.111) *a funkce y diferenciální rovnicí* (2.107) *s počáteční podmínkou* (2.128). *Pak y je řešením okrajové úlohy* (1.12), (1.13).

Buď $\beta_1 > 0$ a buď funkce ψ definovaná diferenciální rovnicí (2.113) *s počáteční podmínkou* (2.114), *funkce u diferenciální rovnicí* (2.115) *s počáteční podmínkou* (2.116) *a funkce w diferenciální rovnicí* (2.131) *s počáteční podmínkou* (2.132). *Pak funkce y definovaná rovnicí* (2.133) *je řešením okrajové úlohy* (1.12), (1.13).

D ů k a z . Vyšetřujeme nejprve případ $\alpha_1 > 0$. Podle věty 2.2 existuje jediné řešení problému (1.12), (1.13). Označme je pro potřeby tohoto důkazu $\tilde{y}$. Podle věty

2.4 platí

$$-p(b)\tilde{y}'(b) + \varphi(b)\tilde{y}(b) = v(b) \tag{2.134}$$

a zároveň

$$\alpha_2 p(b)\tilde{y}'(b) + \beta_2\tilde{y}(b) = \gamma_2. \tag{2.135}$$

Vypočteme-li z těchto dvou rovnic $\tilde{y}(b)$, dostaneme

$$\tilde{y}(b) = \frac{\gamma_2 + \alpha_2 v(b)}{\beta_2 + \alpha_2\varphi(b)} = y(b). \tag{2.136}$$

Dosadíme-li však za $\tilde{y}(b)$ do rovnice (2.134) hodnotu $y(b)$ a vzniklou rovnici odečteme od rovnice (2.107) psané pro $x = b$, dostaneme, že platí

$$p(b)\tilde{y}'(b) = p(b)y'(b). \tag{2.137}$$

Z rovnice (2.107) je na první pohled vidět, že funkce y má v intervalu $\langle a, b)$ spojitou nejen první derivaci, ale i druhou derivaci. Derivujeme-li rovnici (2.107) a dosadíme-li za φ' podle (2.108), za v' podle (2.110) a za y' znovu podle (2.107), dostaneme, že funkce y splňuje diferenciální rovnici (1.12). Funkce y tedy splňuje stejnou diferenciální rovnici a stejné počáteční podmínky v bodě b jako funkce $\tilde{y}$. Podle věty o jednoznačnosti je $y(x) \equiv \tilde{y}(x)$.

Všimněme si nyní případu $\beta_1 > 0$. Je-li $\tilde{y}$ opět řešení problému (1.12), (1.13), platí znovu podle věty 2.4

$$-\psi(b)p(b)\tilde{y}'(b) + \tilde{y}(b) = u(b) \tag{2.138}$$

a zároveň (2.135). Z těchto dvou rovnic však ihned plyne, že je

$$p(b)\tilde{y}'(b) = w(b). \tag{2.139}$$

Píšeme-li však rovnici (2.133) pro $x = b$ a dosadíme-li do rovnice (2.138) podle (2.139) a vzniklé rovnice odečteme, dostáváme, že je

$$\tilde{y}(b) = y(b). \tag{2.140}$$

Derivováním rovnice (2.133) a dosazením podle (2.115), (2.113) a (2.131) dostaneme, že platí

$$p(x)y'(x) = w(x). \tag{2.141}$$

Odtud dalším derivováním plyne, že funkce y splňuje diferenciální rovnici (1.12). Toto zjištění spolu s podmínkami (2.140) a (2.139) však dává, že je $y(x) \equiv \tilde{y}(x)$. Věta je dokázána.

Metodou popsanou ve větě 2.6, která tedy spočívá v tom, že okrajová podmínka se přesouvá v normalizovaném tvaru a že hledané řešení se dostane řešením diferenciální rovnice nižšího řádu, než je řád původní diferenciální rovnice, nazveme *metodou normalizovaného přesunu* okrajové podmínky.

Při metodě normalizovaného přesunu okrajové podmínky tedy řešíme nejprve zleva doprava dvě rovnice prvního řádu, z nichž jedna je nelineární (Ricattiova) a druhá lineární (v případě $\alpha_1 > 0$ jsou to rovnice (2.108) a (2.110)), a hledané řešení dostaneme řešením lineární diferenciální rovnice zprava doleva. Řešení příslušných úloh s počátečními podmínkami přitom provedeme některou z metod popsaných v kap. I. Tam jsme sice všude předpokládali, že počáteční podmínka je zadaná v levém krajním bodě intervalu, v němž hledáme řešení, metody, které jsme uvedli, však lze beze změny užíti k řešení úloh, v nichž počáteční podmínka je zadaná v pravém krajním bodě daného intervalu. Stačí k tomu pouze předpokládat, že příslušný integrační krok je záporný.

V popsaném postupu lze samozřejmě zaměnit roli krajních bodů intervalu $\langle a, b\rangle$. Postup je zřejmý, je jen třeba užít místo věty 2.4 větu 2.5. Tak za předpokladu, že je $\alpha_2 > 0$, se řešení okrajové úlohy (1.12), (1.13) dostane řešením diferenciální rovnice (2.118) s počáteční podmínkou (2.119), diferenciální rovnice (2.120) s počáteční podmínkou (2.121) (obě tyto rovnice se řeší zprava doleva) a diferenciální rovnice (2.117) s počáteční podmínkou

$$y(a) = \frac{\gamma_1 - \alpha_1\hat{v}(a)}{\beta_1 - \alpha_1\hat{\varphi}(a)}. \tag{2.142}$$

V případě, že je $\beta_2 > 0$, se řeší diferenciální rovnice (2.123) a (2.125) s počátečními podmínkami (2.124) a (2.126) zprava doleva, pak se řeší diferenciální rovnice

$$-\hat{w}' + q(x)\hat{\psi}(x)\hat{w} = f(x) - q(x)\hat{u}(x) \tag{2.143}$$

s počáteční podmínkou

$$\hat{w}(a) = \frac{\gamma_1 - \beta_1\hat{u}(a)}{-\alpha_1 + \beta_1\hat{\psi}(a)} \tag{2.144}$$

a hledané řešení se dostane z rovnice

$$y(x) = \hat{u}(x) + \hat{\psi}(x)\hat{w}(x). \tag{2.145}$$

V případě, že nás zajímá řešení dané okrajové úlohy jen v jednom nebo několika málo bodech, je obvykle výhodnější postupovat obdobně jako u metody prostého přesunu. Řešením rovnic (2.108) a (2.110), resp. rovnic (2.113) a (2.115) přesuneme levou okrajovou podmínku do žádaného bodu x_0 a pomocí rovnic (2.118) a (2.120), resp. (2.123) a (2.125) přesuneme pravou okrajovou podmínku do téhož bodu. Hodnotu $y(x_0)$ a $p(x_0)y'(x_0)$ pak vypočteme ze vzniklé soustavy. I tuto variantu nazveme metodou normalizovaného přesunu.

O metodě normalizovaného přesunu bychom rádi tvrdili, že se v nějakém smyslu chová lépe než metoda střelby nebo metoda prostého přesunu. Doposud ovšem víme pouze, že tato metoda odstraňuje jen ten menší problém spojený s metodou prostého přesunu nebo metodou střelby, totiž přílišný růst veličin, s nimiž jsme nuceni počítat. Daleko vážnější nedostatek zmíněných metod spočívající v neúnosně

vysoké citlivosti výchozí rovnice na počátečních podmínkách jsme zatím odstranili pouze tak, že jsme příslušnou úlohu z algoritmu normalizovaného přesunu prostě vyřadili a nahradili ji úlohou s počátečními podmínkami pro jinou diferenciální rovnici. Snažme se proto nalézt nějaké matematické důvody, které by vhodnost tohoto postupu podepřely. Začněme následující definicí:

Definice 2.1. Řekneme, že řešení y diferenciální rovnice

$$y' = f(x, y) \tag{2.146}$$

je *stabilní vzhledem k trvale působícím poruchám*, jestliže ke každému $\varepsilon > 0$ existuje $\Delta > 0$ tak, že pro každé řešení diferenciální rovnice

$$\tilde{y}' = f(x, \tilde{y}) + \delta(x), \quad x \in \langle a, \infty), \tag{2.147}$$

pro něž platí

$$|\tilde{y}(a) - y(a)| \leqq \Delta \tag{2.148}$$

a

$$\sup |\delta(x)| \leqq \Delta, \tag{2.149}$$

platí

$$|\tilde{y}(x) - y(x)| < \varepsilon \tag{2.150}$$

pro každé $x \in \langle a, \infty)$.

Je zřejmé, že stejně dobře můžeme mluvit o stabilitě vzhledem k trvale působícím poruchám i v případě diferenciálních rovnic vyšších řádů nebo soustav diferenciálních rovnic. Stačí k tomu pouze pokládat v předchozí definici veličiny y a f nikoli za skaláry ale za vektory.

Definice 2.1 postihuje zejména to, že řešení diferenciální rovnice, které je stabilní vzhledem k trvale působícím poruchám, závisí spojitě na malých změnách počáteční podmínky a pravé strany stejnoměrně vzhledem k délce intervalu, v němž je uvažujeme. Stabilita vzhledem k trvale působícím poruchám tedy vylučuje chování, které je při metodě střelby nebo metodě prostého přesunu typické, totiž že chyby podstatně závisí na délce intervalu, v němž hledáme řešení. Proto kdyby se nám podařilo dokázat, že rovnice normalizovaného přesunu jsou stabilní vzhledem k trvale působícím poruchám, byl by to velmi významný matematický argument v jejich prospěch.

V dalším textu ukážeme, že za určitých doplňujících předpokladů tomu tak skutečně je. Nejdříve však zformulujeme a dokážeme pomocné tvrzení citované obvykle jako *Bellmanovo* nebo *Grönwallovo* lemma.

Lemma 2.6. *Nechť funkce φ, ψ a χ jsou spojité v intervalu $\langle a, b\rangle$ a nechť χ je v tomto intervalu nezáporná. Nechť dále platí*

$$\varphi(x) \leqq \psi(x) + \int_a^x \chi(t)\varphi(t)\,\mathrm{d}\,t \tag{2.151}$$

pro každé $x \in \langle a, b\rangle$. Pak platí

$$\varphi(x) \leqq \psi(x) + \int_a^x \chi(\tau)\psi(\tau)\mathrm{e}^{\int_\tau^x \chi(t)\,\mathrm{d}\,t}\,\mathrm{d}\,\tau \tag{2.152}$$

pro každé $x \in \langle a, b\rangle$.

Důkaz. Položme

$$R(x) = \int_a^x \chi(t)\varphi(t)\,\mathrm{d}\,t. \tag{2.153}$$

Z předpokladu (2.151) a z nezápornosti funkce χ snadno plyne, že platí

$$R'(x) - \chi(x)R(x) \leqq \chi(x)\psi(x). \tag{2.154}$$

Definujeme-li funkci V diferenciální rovnicí

$$V' - \chi(x)V = \chi(x)\psi(x) \tag{2.155}$$

s počáteční podmínkou

$$V(a) = 0, \tag{2.156}$$

je

$$R(x) \leqq V(x) \tag{2.157}$$

pro každé $x \in \langle a, b\rangle$. Skutečně, položíme-li $Z(x) = V(x) - R(x)$, je

$$Z'(x) - \chi(x)Z(x) = \gamma(x), \tag{2.158}$$

kde γ je nezáporná funkce; rovnice (2.158) plyne ihned odečtením nerovnosti (2.154) od rovnice (2.155). Řešení diferenciální rovnice (2.158) s počáteční podmínkou $Z(a) = 0$, která je zřejmě splněna, však lze psát vzorcem

$$Z(x) = \int_a^x \gamma(\tau)\mathrm{e}^{\int_\tau^x \chi(t)\,\mathrm{d}\,t}\,\mathrm{d}\,\tau, \tag{2.159}$$

jak se snadno zjistí přímým výpočtem. Ze vzorce (2.159) je však patrno, že funkce Z je nezáporná v intervalu $\langle a, b\rangle$. Použijeme-li vzorec (2.159) ještě jednou, tentokrát na funkci V, dostáváme tvrzení lemmatu z nerovnosti (2.157).

Věta 2.7. *Nechť funkce p, p', q a f jsou spojité v intervalu $\langle a, \infty)$ a nechť v tomto intervalu platí*

$$0 < p_0 \leqq p(x) \leqq P_0 \tag{2.160}$$

a

$$0 < q_0 \leqq q(x), \tag{2.161}$$

kde p_0, q_0 a P_0 jsou konstanty. Nechť dále platí $\alpha_1 > 0$ a $\beta_1 > 0$. Pak řešení diferenciální rovnice (2.108) a (2.110) *s počátečními podmínkami* (2.109) a (2.111) *jsou stabilní vzhledem k trvale působícím poruchám.*

D ů k a z . Začneme rovnicí (2.108). Buď tedy $\tilde{\varphi}$ řešení diferenciální rovnice

$$\tilde{\varphi}' + \frac{1}{p(x)}\tilde{\varphi}^2 = q(x) + \delta(x) \tag{2.162}$$

a položme

$$e(x) = \tilde{\varphi}(x) - \varphi(x). \tag{2.163}$$

Odečteme-li rovnici (2.108) od rovnice (2.162), dostaneme po snadných úpravách, že funkce e splňuje diferenciální rovnici

$$e' + \frac{2\varphi(x)}{p(x)}e + \frac{1}{p(x)}e^2 = \delta(x). \tag{2.164}$$

Máme dokázat, že řešení této diferenciální rovnice je pro každé $x \in \langle a, \infty)$ malé za předpokladu, že jeho počáteční podmínka je malá a že pravá strana je malá. Abychom toto tvrzení dokázali, ukažme nejprve, že existuje konstanta $\varphi_0 > 0$ taková, že pro řešení φ diferenciální rovnice (2.108) s počáteční podmínkou (2.109) platí

$$\varphi(x) \geqq \varphi_0 \tag{2.165}$$

pro každé $x \in \langle a, \infty)$. Definujme k tomu cíli v intervalu $\langle a, \infty)$ funkci ψ diferenciální rovnicí

$$\psi' + \frac{1}{p_0}\psi^2 = q_0 \tag{2.166}$$

s počáteční podmínkou

$$\psi(a) = \varphi(a). \tag{2.167}$$

Přímým výpočtem se zjistí, že je

$$\psi(x) = (p_0q_0)^{1/2}\frac{\varphi(a) + (p_0q_0)^{1/2}\operatorname{tgh}\left[\left(\frac{q_0}{p_0}\right)^{1/2}(x-a)\right]}{\varphi(a)\operatorname{tgh}\left[\left(\frac{q_0}{p_0}\right)^{1/2}(x-a)\right] + (p_0q_0)^{1/2}}. \tag{2.168}$$

Položíme-li $\omega(x) = \varphi(x) - \psi(x)$, zjistíme snadno, že platí

$$\omega'(x) + \frac{\varphi(x)+\psi(x)}{p(x)}\omega(x) = q(x) - q_0 + \psi^2(x)\left(\frac{1}{p_0} - \frac{1}{p(x)}\right) \geqq 0. \tag{2.169}$$

Odtud a ze vzorce (2.167) však plyne, že je

$$\varphi(x) \geqq \psi(x) \tag{2.170}$$

pro každé $x \in \langle a, \infty)$. Je-li nyní

$$\varphi(a) \geqq (p_0q_0)^{1/2}, \tag{2.171}$$

je

$$\varphi(a)\left\{1 - \operatorname{tgh}\left[\left(\frac{q_0}{p_0}\right)^{1/2}(x-a)\right]\right\} \geqq (p_0q_0)^{1/2}\left\{1 - \operatorname{tgh}\left[\left(\frac{q_0}{p_0}\right)^{1/2}(x-a)\right]\right\}, \tag{2.172}$$

neboť platí $\operatorname{tgh} z \leqq 1$ pro každé z. Z nerovnosti (2.172) však ihned plyne, že je

$$\psi(x) \geqq (p_0q_0)^{1/2}. \tag{2.173}$$

Je-li naopak

$$\varphi(a) < (p_0q_0)^{1/2}, \tag{2.174}$$

je zřejmě $\varphi^2(a) < p_0q_0$, a tedy

$$\begin{aligned} \varphi^2(a)\operatorname{tgh}\left[\left(\frac{q_0}{p_0}\right)^{1/2}(x-a)\right] + (p_0q_0)^{1/2}\varphi(a) &\leqq \\ \leqq p_0q_0\operatorname{tgh}\left[\left(\frac{g_o}{p_0}\right)^{1/2}(x-a)\right] + (p_0q_0)^{1/2}\varphi(a) &= \\ = (p_0q_0)^{1/2}\left\{(p_0q_0)^{1/2}\operatorname{tgh}\left[\left(\frac{q_0}{p_0}\right)^{1/2}(x-a)\right] + \varphi(a)\right\} & \end{aligned} \tag{2.175}$$

Z nerovnosti (2.175) už snadno plyne, že je

$$\psi(x) \geqq \varphi(a). \tag{2.176}$$

Za číslo φ_0 v nerovnosti (2.165) tedy stačí vzít $\min\left((p_0q_0)^{1/2}, \varphi(a)\right)$, protože vzhledem k předpokladu $\beta_1 > 0$ je $\varphi(a) > 0$.

Za předpokladu, že rovnice (2.164) má řešení, platí

$$\begin{aligned} e(x) = e(a)\exp\left[-\int_a^x \frac{2\varphi(t)}{p(t)}\,\mathrm{d}t\right] + \\ + \int_a^x \left[\delta(\tau) - \frac{e^2(\tau)}{p(\tau)}\right]\exp\left[-\int_\tau^x \frac{2\varphi(t)}{p(t)}\,\mathrm{d}t\right]\mathrm{d}\tau. \end{aligned} \tag{2.177}$$

Zde jsme použili vzorce analogického ke vzorci (2.159) pro řešení lineární diferenciální rovnice s nenulovou počáteční podmínkou. Položme dále

$$\sigma = \frac{2\varphi_0}{P_0}, \tag{2.178}$$

takže platí

$$0 < \sigma \leqq \frac{2\varphi(x)}{p(x)} \tag{2.179}$$

pro každé $x \in \langle a, \infty)$. Odhadneme-li integrály z funkce $2\varphi/p$ ve vzorci (2.177) pomocí tohoto čísla σ, dostaneme

$$|e(x)| \leqq |e(a)|\mathrm{e}^{-\sigma(x-a)} + \int_a^x \left|\delta(\tau) - \frac{e^2(\tau)}{p(\tau)}\right| \mathrm{e}^{-\sigma(x-\tau)}\,\mathrm{d}\,\tau. \tag{2.180}$$

Buď nyní ε libovolné pevně zvolené kladné číslo, pro něž platí

$$\varepsilon \leqq \tfrac{1}{2} p_0 \sigma \tag{2.181}$$

a buď $\beta(\varepsilon) = \sup\{\beta;$ řešení e rovnice (2.164) existuje na intervalu $\langle a, \beta\rangle$ a je $|e(x)| \leqq \leqq \varepsilon\}$. Podle definice čísla $\beta(\varepsilon)$ tedy platí pro $x \in \langle a, \beta(\varepsilon)\rangle$

$$\left|\frac{e^2(x)}{p(x)}\right| \leqq \frac{\varepsilon}{p_0}|e(x)|. \tag{2.182}$$

Dosadíme-li tento odhad do nerovnosti (2.180) a položíme-li $\delta = \max|\delta(x)|$, dostaneme po snadných úpravách, že je

$$|e(x)|\mathrm{e}^{\sigma(x-a)} \leqq |e(a)| + \frac{\delta}{\sigma}[\mathrm{e}^{\sigma(x-a)} - 1] + \frac{\varepsilon}{p_0}\int_a^x |e(\tau)|\mathrm{e}^{\sigma(\tau-a)}\,\mathrm{d}\,\tau. \tag{2.183}$$

Použijeme-li na nerovnost (2.183) lemma 2.6, dostaneme

$$|e(x)|\mathrm{e}^{\sigma(x-a)} \leqq |e(a)| + \frac{\delta}{\sigma}[\mathrm{e}^{\sigma(x-a)} - 1] + \frac{\varepsilon}{p_0}\int_a^x \left\{|e(a)| + \frac{\delta}{\sigma}[\mathrm{e}^{\sigma(\tau-a)} - 1]\right\}\mathrm{e}^{\frac{\varepsilon}{p_0}(x-\tau)}\,\mathrm{d}\,\tau. \tag{2.184}$$

Výpočtem integrálů a snadnou úpravou odtud plyne, že platí

$$|e(x)| \leqq |e(a)|\mathrm{e}^{-(\sigma-\frac{\varepsilon}{p_0})(x-a)} + \frac{\delta}{\sigma - \frac{\varepsilon}{p_0}}[1 - \mathrm{e}^{-(\sigma-\frac{\varepsilon}{p_0})(x-a)}]. \tag{2.185}$$

Z nerovnosti (2.181) plyne, že je $\sigma - \varepsilon/p_0 \geqq \sigma - \sigma/2 > 0$ a $1/(\sigma - \varepsilon/p_0) \leqq 2/\sigma$. Tedy pro $x \in \langle a, \beta(\varepsilon)\rangle$ a pro $|\delta(x)| \leqq \delta$ platí

$$|e(x)| \leqq |e(a)| + \frac{2\delta}{\sigma}. \tag{2.186}$$

Buď konečně

$$\Delta = \min\left(\frac{\varepsilon}{4}, \frac{\sigma\varepsilon}{8}\right) \tag{2.187}$$

a nechť platí

$$|\tilde{\varphi}(a) - \varphi(a)| \leqq \Delta \tag{2.188}$$

a

$$|\delta(x)| \leqq \Delta. \tag{2.189}$$

Tvrdíme, že pak řešení e existuje na intervalu $\langle a, \infty)$ a platí

$$|e(x)| < \varepsilon. \tag{2.190}$$

Skutečně, ze spojitosti funkce e a z nerovnosti $|e(a)| \leqq \varepsilon/4 < \varepsilon$ plyne, že číslo $\beta(\varepsilon)$ je větší než a a nerovnost (2.190) zřejmě platí pro $x \in \langle a, \beta(\varepsilon)\rangle$. Předpokládejme, že je $\beta(\varepsilon) < \infty$. Podle (2.186) platí pro $x \in \langle a, \beta(\varepsilon)\rangle$, že je

$$|e(x)| \leqq \Delta + \frac{2\Delta}{\sigma} \leqq \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2} < \varepsilon. \tag{2.191}$$

To je však spor s definicí čísla $\beta(\varepsilon)$. Řešení φ rovnice (2.108) je tedy skutečně stabilní vzhledem k trvale působícím poruchám.

Platnost stejného tvrzení pro rovnici (2.110) se už dokáže podstatně snadněji, neboť tato rovnice je lineární. Buď tedy $\tilde{v}$ řešení diferenciální rovnice

$$\tilde{v}' + \frac{\varphi(x)}{p(x)}\tilde{v} = f(x) + \delta(x) \tag{2.192}$$

a položme

$$e(x) = \tilde{v}(x) - v(x). \tag{2.193}$$

Pak funkce e splňuje diferenciální rovnici

$$e' + \frac{\varphi(x)}{p(x)}e = \delta(x). \tag{2.194}$$

Řešení této diferenciální rovnice existuje v celém intervalu $\langle a, \infty)$ a dá se psát pomocí vzorce

$$e(x) = e(a)\mathrm{e}^{-\int_a^x \frac{\varphi(t)}{p(t)}\,\mathrm{d}\,t} + \int_a^x \delta(\tau)\mathrm{e}^{-\int_\tau^x \frac{\varphi(t)}{p(t)}\,\mathrm{d}\,t}\,\mathrm{d}\,t. \tag{2.195}$$

Použijeme-li nerovnost (2.179), plyne odtud, že je

$$|e(x)| \leqq |e(a)|\mathrm{e}^{\frac{\sigma}{2}(x-a)} + \frac{\delta}{\sigma}[1 - \mathrm{e}^{-\frac{\sigma}{2}(x-a)}], \tag{2.196}$$

kde $\delta = \max|\delta(x)|$. Odtud však už požadované tvrzení plyne bezprostředně. Věta je dokázána.

Snadno lze též dokázat, že za předpokladů analogických jako ve větě 2.7, ovšem tentokrát formulovaných pro interval $(-\infty, b\rangle$, je i rovnice (2.107) stabilní vzhledem k trvale působícím poruchám.

Na závěr tohoto odstavce ilustrujme metodu normalizovaného přesunu řešením téže okrajové úlohy jako v příkladě 2.2. Výsledky, které byly získány za stejných podmínek jako v citovaném příkladě, totiž užitím standardní Rungovy-Kuttovy metody s integračním krokem $h = 0,025$, jsou shrnuty v tab. 2.3. Vidíme, že jsou podstatně uspokojivější než u metody střelby nebo prostého přesunu.

Tabulka 2.3

Řešení okrajové úlohy pro diferenciální rovnici (2.103) metodou normalizovaného přesunu

x	$q = 100$		$q = 500$	
	přibližné řešení	chyba	přibližné řešení	chyba
0,0	0,000 000	0,000 000	0,000 000	0,000 000
0,1	0,309 012	−0,000 005	0,309 007	−0,000 010
0,2	0,587 778	−0,000 007	0,587 770	−0,000 015
0,3	0,809 009	−0,000 008	0,808 999	−0,000 018
0,4	0,951 048	−0,000 009	0,951 037	−0,000 019
0,5	0,999 989	−0,000 011	0,999 979	−0,000 021
0,6	0,951 045	−0,000 012	0,951 038	−0,000 018
0,7	0,809 009	−0,000 008	0,809 007	−0,000 010
0,8	0,587 784	−0,000 001	0,587 780	−0,000 005
0,9	0,309 018	0,000 001	0,309 016	−0,000 001
1,0	0,000 000	0,000 000	0,000 000	0,000 000

2.3.2 Obecná soustava lineárních diferenciálních rovnic

V tomto odstavci stručně ukážeme, jak lze přenést myšlenku normalizovaného přesunu na případ obecné soustavy lineárních diferenciálních rovnic. Začneme několika pomocnými tvrzeními.

Lemma 2.7. *Buď $\boldsymbol{\Phi}(x) = \{\varphi_{ij}(x)\}$ čtvercová matice řádu m, jejíž prvky jsou diferencovatelné funkce v intervalu I. Pak platí*

$$[\det \boldsymbol{\Phi}(x)]' = \operatorname{tr}[\boldsymbol{\Phi}'(x)\tilde{\boldsymbol{\Phi}}(x)] \tag{2.197}$$

pro každé $x \in I$, kde $\tilde{\boldsymbol{\Phi}} = \{\tilde{\varphi}_{ij}\}$ a $\tilde{\varphi}_{ij}$ je algebraický doplněk prvku φ_{ji} matice $\boldsymbol{\Phi}$, tj. determinant matice řádu $m-1$, která vznikne z matice $\boldsymbol{\Phi}$ vypuštěním j-tého řádku a i-toho sloupce opatřený znaménkem $(-1)^{i+j}$ a $\operatorname{tr}(.)$ značí stopu matice, tj. součet jejích diagonálních prvků.

D ů k a z . Z definice determinantu plyne, že platí

$$[\det \boldsymbol{\Phi}(x)]' = \det \begin{bmatrix} \varphi'_{11} & \varphi_{12} & \dots & \varphi_{1m} \\ \varphi'_{21} & \varphi_{22} & \dots & \varphi_{2m} \\ \vdots & \vdots & & \vdots \\ \varphi'_{m1} & \varphi_{m2} & \dots & \varphi_{mm} \end{bmatrix} + \tag{2.198}$$

$$+ \det \begin{bmatrix} \varphi_{11} & \varphi'_{12} & \dots & \varphi_{1m} \\ \varphi_{21} & \varphi'_{22} & \dots & \varphi_{2m} \\ \vdots & \vdots & & \vdots \\ \varphi_{m1} & \varphi'_{m2} & \dots & \varphi_{mm} \end{bmatrix} + \dots + \det \begin{bmatrix} \varphi_{11} & \varphi_{12} & \dots & \varphi'_{1m} \\ \varphi_{21} & \varphi_{22} & \dots & \varphi'_{2m} \\ \vdots & \vdots & & \vdots \\ \varphi_{m1} & \varphi_{m2} & \dots & \varphi'_{mm} \end{bmatrix}.$$

Rozvineme-li první determinant na pravé straně rovnice (2.198) podle prvků prvního sloupce, druhý determinant podle prvků druhého sloupce atd., dostaneme

$$[\det \boldsymbol{\Phi}(x)]' = \sum_{k=1}^{m} \varphi'_{k1}\tilde{\varphi}_{1k} + \sum_{k=1}^{m} \varphi'_{k2}\tilde{\varphi}_{2k} + \dots + \sum_{k=1}^{m} \varphi'_{km}\tilde{\varphi}_{mk} = \sum_{k=1}^{m}\sum_{j=1}^{m} \varphi'_{kj}\tilde{\varphi}_{jk}, \tag{2.199}$$

což dokazuje lemma.

Lemma 2.8. *Buď $\boldsymbol{\Phi}(x) = \{\varphi_{ij}(x)\}$ čtvercová matice, jejíž prvky $\varphi_{ij}(x)$ jsou diferencovatelné pro $x \in I$ a nechť existuje pro $x \in I$ inverzní matice $\boldsymbol{\Phi}^{-1}(x)$. Pak prvky matice $\boldsymbol{\Phi}^{-1}(x)$ jsou také diferencovatelné v I a platí*

$$[\boldsymbol{\Phi}^{-1}(x)]' = -\boldsymbol{\Phi}^{-1}(x)\boldsymbol{\Phi}'(x)\boldsymbol{\Phi}^{-1}(x). \tag{2.200}$$

D ů k a z . Diferencovatelnost matice $\boldsymbol{\Phi}^{-1}(x)$ plyne ihned z rovnosti $\boldsymbol{\Phi}^{-1}(x) = \tilde{\boldsymbol{\Phi}}(x)/\det \boldsymbol{\Phi}(x)$, kde $\tilde{\boldsymbol{\Phi}}$ je matice z lemmatu 2.7. Rovnici (2.200) pak dostaneme derivování identity $\boldsymbol{\Phi}^{-1}(x)\boldsymbol{\Phi}(x) = \boldsymbol{I}$.

Lemma 2.9. *Buď $A(x)$ čtvercová matice řádu m se spojitými prvky v intervalu I a buď $\boldsymbol{\Phi}(x)$ čtvercová matice řádu m definovaná maticovou diferenciální rovnicí*

$$\boldsymbol{\Phi}' = A(x)\boldsymbol{\Phi}, \quad x \in I, \tag{2.201}$$

s počáteční podmínkou

$$\boldsymbol{\Phi}(a) = \boldsymbol{I}, \tag{2.202}$$

kde $\boldsymbol{I}$ je jednotková matice řádu m a $a \in I$. Pak matice $\boldsymbol{\Phi}(x)$ je regulární pro každé $x \in I$.

D ů k a z . Podle lemmatu 2.7 platí pro $x \in I$

$$[\det \boldsymbol{\Phi}(x)]' = \operatorname{tr}[\boldsymbol{\Phi}'(x)\tilde{\boldsymbol{\Phi}}(x)] = \operatorname{tr}[A(x)\boldsymbol{\Phi}(x)\tilde{\boldsymbol{\Phi}}(x)] = [\det \boldsymbol{\Phi}(x)] \operatorname{tr}[A(x)], \tag{2.203}$$

neboť $\boldsymbol{\Phi}(x)\tilde{\boldsymbol{\Phi}}(x) = [\det \boldsymbol{\Phi}(x)]\boldsymbol{I}$. Pro libovolné $x \in I$ tedy platí

$$\det \boldsymbol{\Phi}(x) = \mathrm{e}^{\int_a^x \operatorname{tr}[A(\tau)]\,\mathrm{d}\tau}, \tag{2.204}$$

protože je $\det \boldsymbol{\Phi}(a) = 1$.

Připomeňme, že matici $\boldsymbol{\Phi}$ zavedenou v lemmatu 2.9 jsme v odst. 2.1.2 nazvali fundamentální maticí soustavy $y' = Ay$.

Zatímco lemmata 2.7 – 2.9 mají obecnou platnost, následující lemma už se bezprostředně týká metody přesunu okrajových podmínek pro soustavu lineárních diferenciálních rovnic (2.11) s lineárními separovanými okrajovými podmínkami (1.6). Předpokládejme tedy v dalším, že pro tuto okrajovou úlohu jsou splněny předpoklady zformulované na začátku odst. 2.2.2.

Lemma 2.10. *Nechť matice $R(x)$ typu $m_0 \times m$ je definovaná v intervalu $\langle \xi_1, \xi_2 \rangle$ diferenciální rovnicí* (2.81) *s počáteční podmínkou* (2.82). *Pak hodnost matice $R(x)$ je pro každé $x \in \langle \xi_1, \xi_2 \rangle$ rovna hodnosti matice V_0.*

D ů k a z . Buď $\Phi(x)$ fundamentální matice soustavy $y' = Ay$ určená počáteční podmínkou $\Phi(\xi_0) = I$. Podle lemmatu 2.9 je matice $\Phi(x)$ regulární pro každé $x \in \in \langle \xi_1, \xi_2 \rangle$. Matice $V_0\Phi^{-1}(x)$ má tedy smysl a splňuje zřejmě podmínku (2.82). Na základě lemmatu 2.8 se snadno vypočte, že splňuje i diferenciální rovnici (2.81). Platí tedy

$$R(x) = V_0\Phi^{-1}(x) \tag{2.205}$$

pro každé $x \in \langle \xi_1, \xi_2 \rangle$. Odtud však už požadované tvrzení plyne, neboť matice $\Phi^{-1}(x)$ je regulární pro každé $x \in \langle \xi_1, \xi_2 \rangle$.

Obraťme se nyní k řešení okrajové úlohy (2.21), (1.6). V odst. 2.2.2 jsme viděli, jak lze m_1 okrajových podmínek tvaru

$$V_1 y(a) = v_1, \tag{2.206}$$

kde V_1 je obdélníková matice typu $m_1 \times m$, přesunout do libovolného bodu intervalu $\langle a, b \rangle$. Předpokládáme-li, že hodnost matice V_1 je rovna číslu m_1, lze v ní vybrat m_1 lineárně nezávislých sloupců. Budeme předpokládat, že je to prvních m_1 sloupců, neboť toho lze vhodným přečíslováním složek vektoru y, a tedy přerovnáním sloupců matice V_1, vždy dosáhnout. Dá se tedy psát

$$V_1 = [V_1^{(1)}, V_1^{(2)}], \tag{2.207}$$

kde $V_1^{(1)}$ je regulární čtvercová matice řádu m_1, a okrajovou podmínku (2.206) lze uvést na tvar

$$[I, -G_0]y(a) = g_0, \tag{2.208}$$

kde matice G_0 typu $m_1 \times (m - m_1)$ je dána rovnicí

$$G_0 = -(V_1^{(1)})^{-1} V_1^{(2)} \tag{2.209}$$

a m_1-dimenzionální vektor g_0 rovnicí

$$g_0 = (V_0^{(1)})^{-1} v_1. \tag{2.210}$$

Nechť matice $R(x)$ a vektor $r(x)$ realizují přesun okrajové podmínky (2.208). Matice R tedy splňuje diferenciální rovnici (2.87) s počáteční podmínkou

$$R(a) = [I, -G_0] \tag{2.211}$$

a vektor r diferenciální rovnici (2.91) s počáteční podmínkou

$$r(a) = g_0, \tag{2.212}$$

přičemž platí

$$R(x)y(x) = r(x) \tag{2.213}$$

pro každé řešení dané okrajové úlohy (srv. větu 2.3 z odst. 2.2.2). Položme

$$R(x) = [R_1(x), R_2(x)], \tag{2.214}$$

kde bloky matice R jsou utvořeny stejně jako bloky matice V_1. Položíme-li ještě

$$y(x) = [y_1(x), y_2(x)]^T, \tag{2.215}$$

kde

$$\begin{aligned} y_1(x) &= [{}^1y(x), \ldots, {}^{m_1}y(x)]^T, \\ y_2(x) &= [{}^{m_1+1}y(x), \ldots, {}^{m}y(x)]^T \end{aligned} \tag{2.216}$$

(tj. vektor y_1 má za složky prvních m_1 složek vektoru y a vektor y_2 je tvořen zbývajícími složkami tohoto vektoru), lze rovnici (2.213) psát ve tvaru

$$R_1(x)y_1(x) + R_2(x)y_2(x) = r(x), \quad x \in \langle a, b \rangle. \tag{2.217}$$

Mimoto pro matici $R_1(x)$ platí podmínka $R_1(a) = I$. V důsledku spojitosti prvků této matice tedy existuje $\delta > 0$ takové, že matice $R_1(x)$ je regulární na intervalu $\langle a, a + \delta \rangle$. Předpokládáme-li, že matice $R_1(x)$ je regulární na *celém* intervalu $\langle a, b \rangle$, lze přesunutou podmínku (2.217) psát ve tvaru

$$y_1(x) - G(x)y_2(x) = g(x), \tag{2.218}$$

kde matice $G(x)$ je dána rovnicí

$$G(x) = -[R_1(x)]^{-1}R_2(x) \tag{2.219}$$

a vektor $g(x)$ rovnicí

$$g(x) = [R_1(x)]^{-1}r(x). \tag{2.220}$$

Podmínka (2.218) už je v normalizovaném tvaru v tom smyslu, že koeficienty vystupující u prvních m_1 složek vektoru $y(x)$ tvoří v celém intervalu $\langle a, b \rangle$ jednotkovou matici.

Abychom tedy mohli přesouvat danou okrajovou podmínku v normalizovaném tvaru, musí být především matice $R_1(x)$ regulární pro každé $x \in \langle a, b \rangle$ a musí být

možné počítat prvky matice $G(x)$ a složky vektoru $g(x)$ přímo a nikoliv pomocí matice $R(x)$ a vektoru $r(x)$ z metody prostého přesunu. Položíme-li

$$A(x) = \begin{bmatrix} A_{11}(x) & A_{12}(x) \\ A_{21}(x) & A_{22}(x) \end{bmatrix} \tag{2.221}$$

a

$$f(x) = [f_1(x), f_2(x)]^T, \tag{2.222}$$

kde

$$A_{11}(x) = \begin{bmatrix} a_{11}(x) & \dots & a_{1m_1}(x) \\ \vdots & & \vdots \\ a_{m_1 1}(x) & \dots & a_{m_1 m_1}(x) \end{bmatrix}, \tag{2.223}$$

$$A_{12}(x) = \begin{bmatrix} a_{1,m_1+1}(x) & \dots & a_{1m}(x) \\ \vdots & & \vdots \\ a_{m_1,m_1+1}(x) & \dots & a_{m_1 m}(x) \end{bmatrix},$$

$$A_{21}(x) = \begin{bmatrix} a_{m_1+1,1}(x) & \dots & a_{m_1+1,m_1}(x) \\ \vdots & & \vdots \\ a_{m1}(x) & \dots & a_{mm_1}(x) \end{bmatrix},$$

$$A_{22}(x) = \begin{bmatrix} a_{m_1+1,m_1+1}(x) & \dots & a_{m_1+1,m}(x) \\ \vdots & & \vdots \\ a_{m,m_1+1}(x) & \dots & a_{mm}(x) \end{bmatrix},$$

$$f_1(x) = [{}^1f(x), \dots, {}^{m_1}f(x)]^T,$$
$$f_2(x) = [{}^{m_1+1}f(x), \dots, {}^{m}f(x)]^T,$$

určí se matice $G(x)$ řešením maticové diferenciální rovnice

$$G' = A_{11}(x)G - GA_{22}(x) - GA_{21}(x)G + A_{12}(x) \tag{2.224}$$

s počáteční podmínkou

$$G(a) = G_0 \tag{2.225}$$

a vektor g řešením diferenciální rovnice

$$g' = [A_{11}(x) - G(x)A_{21}(x)]g + f_1(x) - G(x)f_2(x) \tag{2.226}$$

s počáteční podmínkou

$$g(a) = g_0, \tag{2.227}$$

jak se snadno zjistí.

Maticová diferenciální rovnice (2.224) je nelineární (viz člen $GA_{21}G$), takže její řešení nemusí existovat v celém intervalu $\langle a, b\rangle$. Předpoklad, že matice $R_1^{-1}(x)$ existuje pro každé $x \in \langle a, b\rangle$, je vlastně totožný s předpokladem, že soustava diferenciálních rovnic (2.224) má řešení v celém intervalu $\langle a, b\rangle$. O opatřeních, která je třeba učinit, je-li tento předpoklad porušen (což se pozná podle toho, že se nepodaří dojít s řešením rovnic (2.224) až do bodu b), se zmíníme později.

Pomocí rovnic (2.218) přesuneme tedy okrajovou podmínku až do bodu b. Dostaneme tak rovnice

$$y_1(b) - G(b)y_2(b) = g(b). \tag{2.228}$$

Přidáme-li k těmto rovnicím druhou okrajovou podmínku (1.6), kterou zapíšeme pro názornost ve tvaru

$$V_2^{(1)}y_1(b) + V_2^{(2)}y_2(b) = v_2, \tag{2.229}$$

kde jsme položili

$$V_2 = [V_2^{(1)}, V_2^{(2)}], \tag{2.230}$$

dostaneme řešením takto vzniklé soustavy rovnic počáteční podmínku $y(b) = [y_1(b), y_2(b)]^T$ hledaného řešení. Abychom utvořili úplnou analogii metody normalizovaného přesunu z odst. 2.3.1, nesmíme použít tuto počáteční podmínku pro původní soustavu diferenciálních rovnic, ale použijeme pouze její složku $y_2(b)$ jako počáteční podmínku pro soustavu diferenciálních rovnic

$$y_2' = [A_{22}(x) + A_{21}G(x)]y_2 + f_2 + A_{21}g. \tag{2.231}$$

Tato soustava diferenciálních rovnic vznikla vyloučením složky $y_1(x)$ hledaného řešení z původní soustavy (2.11) užitím rovnic (2.218). Řešením diferenciální rovnice (2.231) dostaneme složku y_2 hledaného řešení. Potřebujeme-li znát i složku y_1, určíme ji z (2.218).

Metoda normalizovaného přesunu spočívá i pro uvažovanou obecnou okrajovou úlohu v řešení dvou úloh s počátečními podmínkami v bodě a, z nichž jedna je nelineární, a lineární úlohy s počátečními podmínkami v bodě b.

Není-li splněn předpoklad regularity matice $R_1(x)$ v celém intervalu $\langle a, b\rangle$, za něhož jsme metodu normalizovaného přesunu okrajové podmínky odvodili, projeví se to tak, že uvnitř intervalu $\langle a, b\rangle$ existuje bod, v němž absolutní hodnoty některých prvků matice $G(x)$ rostou nade všechny meze. Z lemmatu 2.10 však plyne, že matice $R(x) = [R_1(x), R_2(x)]^T$ má hodnost m_1 pro každé $x \in \langle a, b\rangle$, a tedy má vždy některých m_1 sloupců lineárně nezávislých. Můžeme tedy postupovat tak, že v tom bodě intervalu $\langle a, b\rangle$, v němž absolutní hodnota některého prvku matice $G(x)$ převýší nějakou předem danou toleranci, provedeme nový výběr lineárně nezávislých sloupců matice $R(x)$. V dalším výpočtu pak bude roli matice $R_1(x)$ hrát matice sestavená z těchto sloupců. Příslušné diferenciální rovnice se samozřejmě změní, neboť nový výběr lineárně nezávislých sloupců znamená nové přečíslování složek

hledaného vektoru. Do dalších podrobností, které mohou být značně komplikované, už nebudeme zacházet, naším cílem bylo ukázat pouze základní myšlenku možného postupu.

Pravou okrajovou podmínku lze samozřejmě také přesouvat v normalizovaném tvaru pomocí diferenciálních rovnic, které jsou analogické k rovnicím (2.224) a (2.226). Přesuneme-li tedy levou okrajovou podmínku a pravou okrajovou podmínku do téhož bodu ξ, můžeme vypočítat složky hledaného vektoru řešení v tomto bodě řešením soustavy lineárních algebraických rovnic. Dostaneme tak modifikaci metody normalizovaného přesunu, při níž řešení soustavy diferenciálních rovnic typu (2.231) odpadá.

Praktické zkušenosti ukazují, že popsaná metoda normalizovaného přesunu okrajové podmínky vykazuje např. ve srovnání s metodou střelby většinou podstatně stabilnější chování. Je však dosti náročná na paměť počítače a kromě toho soustavy diferenciálních rovnic, které se v ní vyskytují, jsou často soustavami se silným tlumením, z čehož mohou vzniknout určité problémy při jejich řešení. Toto vše spolu s tím, že jde o metodu určenou výhradně k řešení lineárních problémů, může představovat v některých případech její citelný nedostatek.

3 Metoda sítí

Základní myšlenka metody sítí, zvané také diferenční metoda, je velice jednoduchá. V intervalu $\langle a, b\rangle$, v němž hledáme řešení dané okrajové úlohy, se zvolí konečná množina bodů zvaná síť — může to být např. množina ekvidistantních bodů $\{x_k; k = 0, \dots, n\}$, tj. množina bodů x_k daných předpisem

$$x_k = a + kh, \quad k = 0, \dots, n, \tag{3.1}$$

kde

$$h = \frac{b-a}{n} \tag{3.2}$$

a n je přirozené číslo — a v bodech této množiny se splní daná diferenciální rovnice (eventuálně i okrajové podmínky) přibližně, a to tak, že derivace, které se v nich vyskytují, se nahradí diferenčními podíly (tj. lineárními kombinacemi funkčních hodnot v okolních bodech), které je aproximují. Zanedbáme-li chyby, které přitom vzniknou, dostaneme pro hodnoty v bodech sítě soustavu konečně mnoha rovnic (v případě lineárního problému lineárních).

Abychom mluvili konkrétně, uvažujme jako příklad jednoduchou diferenciální rovnici

$$y'' = f(x, y), \quad x \in (a, b), \tag{3.3}$$

s okrajovými podmínkami

$$y(a) = \gamma_1, \quad y(b) = \gamma_2. \tag{3.4}$$

V odst. 3.4 ukážeme, že tato okrajová úlohy má za předpokladu dostatečné hladkosti funkce f a za předpokladu, že platí $f_y(x, y) \geq 0$, právě jedno řešení. Provedeme-li postup, který jsme naznačili, tj. aproximujeme-li druhou derivaci funkce y v bodě x_k podílem $[y(x_{k-1}) - 2y(x_k) + y(x_{k+1})]/h^2$, dostaneme, že pro $k = 1, \dots, n-1$ platí

$$\frac{y(x_{k-1}) - 2y(x_k) + y(x_{k+1})}{h^2} = f(x_k, y_k) + \varepsilon_k, \tag{3.5}$$

kde ε_k je chyba, které jsme se dopustili náhradou druhé derivace zmíněným podílem. Hodnoty přibližného řešení v bodech x_k, $k = 0, \dots, n$, které označíme y_k, budeme tedy hledat ze soustavy

$$\begin{gathered} y_{k-1} - 2y_k + y_{k-1} = h^2 f(x_k, y_k), \quad k = 1, \dots, n-1, \\ y_0 = \gamma_1, \quad y_n = \gamma_2. \end{gathered} \tag{3.6}$$

Vidíme, že soustavu rovnic (3.6), která nahrazuje původní okrajovou úlohu (3.3), (3.4) konečnědimenzionálním problém, jsme získali formálně velice jednoduše. Abychom však mohli mluvit o skutečné numerické metodě pro řešení dané okrajové úlohy, musíme ještě dát odpověď na tři otázky: (i) Musíme zjistit, zda soustava (3.6) má vůbec nějaké řešení. (ii) Musíme určit numerickou metodu pro řešení této soustavy a vyšetřit její vlastnosti (např. otázku konvergence iterací apod.). (iii) Musíme vyšetřit vztah mezi veličinami y_k získanými řešením soustavy (3.6) a hodnotami $y(x_k)$ přesného řešení při $h \to 0$; jinými slovy, je třeba vyšetřit konvergenci popsaného postupu. Dát univerzální návody na řešení bodů (i), (ii) a (iii) není za současného stavu našich znalostí možné, upozorňujeme pouze na nutnost provedení patřičných úvah pro každý jednotlivý případ zvlášť.

Z předchozího textu by mohl vzniknout dojem, že „nultý“ bod při užití metody sítí, tj. postup, kterým získáme soustavu typu (3.6), je natolik jednoduchý, že se jím není třeba už dále zabývat. Není však tomu úplně tak, neboť je známo, že zdánlivě nepatrné modifikace mohou mít dosti podstatný vliv na vyšetřování problémů (i), (ii) a (iii). Proto si v dalších odstavcích, v nichž popíšeme metodu sítí pro některé konkrétní úlohy, všimneme všech čtyř zmíněných aspektů. Ještě předtím však věnujeme zvláštní odstavec tzv. monotónním maticím, neboť tato třída speciálních matic hraje v souvislosti s metodou sítí značně důležitou roli.

3.1 Monotónní matice

Definice 3.1. Řekněme, že čtvercová matice $A = \{a_{ij}\}$ řádu n je *monotónní*, platí-li implikace

$$Ax \geq 0 \Longrightarrow x \geq 0. \tag{3.7}$$

Nerovnostem mezi vektory, případně maticemi, zde rozumíme tak, že příslušné nerovnosti platí pro všechny složky, případně prvky. Podobně symbolem $|x|$ rozumíme vektor, jehož složky jsou absolutní hodnoty složek vektoru x.

Monotónní matice je tedy taková matice A, že z nezápornosti všech složek vektoru Ax plyne nezápornost všech složek samotného vektoru x. V několika následujících tvrzeních ukážeme některé základní vlastnosti monotónních matic, které v dalším textu použijeme.

Lemma 3.1. *Matice A je monotónní tehdy a jen tehdy, je-li regulární a je-li inverzní matice A^{-1} nezáporná (tj. má-li všechny prvky nezáporné).*

D ů k a z . Předpokládejme nejprve, že matice A je monotónní a nechť vektor x je řešením soustavy $Ax = 0$. Z definice 3.1 pak ihned plyne, že je $x = 0$ ($Ax \geqq 0$ a $A(-x) \geqq 0$). Homogenní soustava $Ax = 0$ má tedy pouze triviální řešení, což dokazuje regulárnost matice A. Nezápornost matice A^{-1} plyne ihned z rovnice $AA^{-1} = I$ a definice 3.1. Obrácené tvrzení je zřejmé na první pohled. Lemma je dokázáno.

Lemma 3.2. *Nechť A je monotónní matice a nechť pro vektory y a z platí*

$$|Ay| \leqq Az. \tag{3.8}$$

Pak platí

$$|y| \leqq z. \tag{3.9}$$

D ů k a z . Z nerovností (3.8) ihned plyne, že pro vektory $z - y$ a $z + y$ platí $A(z - y) \geqq 0$ a $A(z + y) \geqq 0$. Použijeme-li nyní implikaci (3.7), dostaneme ihned požadované tvrzení.

Lemma 3.3. *Nechť pro monotónní matice A a B platí nerovnost*

$$A \leqq B. \tag{3.10}$$

Pak platí

$$A^{-1} \geqq B^{-1}. \tag{3.11}$$

D ů k a z . Tvrzení plyne ihned z identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, na jejíž pravé straně je součin tří nezáporných matic, tj. nezáporná matice.

Z uvedených lemmat vidíme, že monotónní matice mají řadu sice jednoduchých, ale, jak uvidíme, velmi důležitých vlastností. Podtrhněme zejména lemma 3.2, které dává možnost odhadnout řešení soustavy lineárních rovnic s monotónní maticí řešením soustavy, jejíž pravá strana majorizuje pravou stranu původní soustavy, a které tak představuje účinný aparát ke konstrukci nejrůznějších odhadů. Vzhledem k tomu, že zkoumání monotónie dané matice přímo z definice 3.1 je většinou nepohodlné (a často také vůbec prakticky neproveditelné) a ani ekvivalentní podmínka vyjádřená v lemmatu 3.1 obecně také mnoho nepomůže, uvedeme jednoduchou postačující podmínku uváděnou v literatuře pod názvem Collatzovo lemma. Před jeho formulací je třeba zavést ještě některé další pojmy.

Definice 3.2. Řekněme, že čtvercová matice $A = \{a_{ij}\}$ řádu n je *reducibilní*, je-li možné přerovnat indexy $1, \ldots, n$ v posloupnost $\varrho_1, \ldots, \varrho_r$, $\sigma_1, \ldots, \sigma_{n-r}$, $1 \leqq r < n$, tak, že platí $a_{\varrho_\nu \sigma_\mu} = 0$ pro $\nu = 1, \ldots, r$ a $\mu = 1, \ldots, n - r$. Matici která není reducibilní nazveme *ireducibilní*.

Poznamenejme, že kromě slov reducibilní a ireducibilní se užívá také pojmů *rozložitelná* a *nerozložitelná*.

Lemma 3.4. *Matice A řádu aspoň 2 je ireducibilní tehdy a jen tehdy, existuje-li ke každé dvojici indexů i, j, $i \neq j$ posloupnost indexů $i_1, \ldots, i_s$ taková, že platí $a_{ii_1} \neq 0, a_{i_1 i_2} \neq 0, \ldots, a_{i_s j} \neq 0$.*

D ů k a z . Nechť za prvé tvrzení lemmatu není pravda. Pak existuje dvojice indexů i_0, j_0, $i_0 \neq j_0$, taková, že ať zvolíme posloupnost $i_1, \ldots, i_s$ jakkoliv, vždy je některé z čísel $a_{i_0 i_1}, a_{i_1 i_2}, \ldots, a_{i_s j_0}$ rovno nule. Znakem $\mathscr{J}$ označme množinu indexů, do které patří index i_0 a dále každý index $j \neq i_0$ takový, že k němu existuje posloupnost indexů $i_1, \ldots, i_m$ taková, že je $a_{i_0 i_1} \neq 0, a_{i_1 i_2} \neq 0, \ldots, a_{i_m j} \neq 0$. Množina $\mathscr{J}$ je neprázdná ($i_0 \in \mathscr{J}$) a neobsahuje všechny indexy, neboť index j_0 do ní nepatří. Je tedy tvořena indexy $\varrho_1, \ldots, \varrho_r$, $1 \leqq r < n$. Zbývající indexy označme $\sigma_1, \ldots, \sigma_{n-r}$. Tvrdíme, že platí $a_{\varrho_\nu \sigma_\mu} = 0$ pro $\nu = 1, \ldots, r$ a $\mu = 1, \ldots, n - r$. Kdyby to totiž nebyla pro některé $\tilde{\nu}$, $\tilde{\mu}$ pravda, platilo by $a_{\varrho_{\tilde{\nu}} \sigma_{\tilde{\mu}}} \neq 0$. Podle definice množiny $\mathscr{J}$ však existuje posloupnost $i_1, \ldots, i_m$ taková, že platí $a_{i_0 i_1} \neq 0, a_{i_1 i_2} \neq 0, \ldots, a_{i_m \varrho_{\tilde{\nu}}} \neq 0$. Celkem by tedy platilo $a_{i_0 i_1} \neq 0, a_{i_1 i_2} \neq 0, \ldots, a_{i_m \varrho_{\tilde{\nu}}} \neq 0, a_{\varrho_{\tilde{\nu}} \sigma_{\tilde{\mu}}} \neq 0$, tj. $\sigma_{\tilde{\mu}} \in \mathscr{J}$. To je však spor dokazující, že skutečně platí $a_{\varrho_\nu \sigma_\mu} = 0$ pro $\nu = 1, \ldots, r$ a $\mu = 1, \ldots, n - r$. Odtud však plyne, že matice A je reducibilní.

Nechť naopak A je reducibilní. Pak tvrzení lemmatu neplatí pro libovolnou dvojici indexů ϱ_ν, σ_μ.

Lemma je dokázáno.

Tvrzení lemmatu 3.4 se nejsnáze pamatuje a také prověřuje pomocí pojmů teorie grafů. Pojem *graf* je značně obecný topologický pojem a je základem velmi obsáhlé teorie. My se zde omezíme na *konečné orientované grafy*. Konečným orientovanými grafem rozumíme uspořádanou dvojici konečných množin V a H, kde množina H je tvořena některými uspořádanými dvojicemi bodů z množiny V (je tedy $H \subset V \times V$). Prvky množiny V se nazývají *uzly* nebo *vrcholy* grafu a prvky množiny H *hrany* daného grafu. Je-li počet uzlů grafu malý, můžeme jej znázornit diagramem tak, že uzly znázorníme body v rovině a hrany oblouky označenými šipkami. Tyto oblouky vycházejí z uzlu, v němž hrana začíná a směřují do uzlu, v němž hrana končí. Posloupnost neopakujících se uzlů $u_1, \ldots, u_k$ taková, že dvojice $(u_1, u_2), (u_2, u_3), \ldots, (u_{k-1}, u_k)$ jsou hrany grafu, se nazývá *dráha*. Existuje-li z každého uzlu daného grafu dráha do libovolného dalšího uzlu, nazveme graf *silně souvislým*.

Přiřaďme dané matici $A = \{a_{ij}\}$ řádu n orientovaný graf tak, že za uzly vezmeme množinu indexů $\{1, 2, \ldots, n\}$ a za jeho hrany ty dvojice indexů (i, k), po něž platí

$a_{ij} \neq 0$. Nyní se dá lemma 3.4 parafrázovat takto: Matice A je ireducibilní tehdy a jen tehdy, je-li její graf silně souvislý. Pravdivost tohoto tvrzení je zřejmá.

Definice 3.3. Řekněme, že čtvercová matice $A = \{a_{ij}\}$ řádu n je *diagonálně dominantní*, platí-li pro $i = 1, \ldots, n$

$$|a_{ii}| \geqq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \tag{3.12}$$

řekneme, že je *ostře diagonálně dominantní*, platí-li v (3.12) ostrá nerovnost pro každý index i a konečně řekneme, že je *ireducibilně diagonálně dominantní*, je-li ireducibilní, diagonálně dominantní a platí-li v (3.12) ostrá nerovnost alespoň pro jeden index i.

Nyní máme vše připraveno k formulaci Collatzova lemmatu.

Lemma 3.5. *Nechť matice $A = \{a_{ij}\}$ řádu n má kladné diagonální prvky a nekladné nediagonální prvky a nechť je ostře diagonálně dominantní nebo ireducibilně diagonálně dominantní. Pak je monotónní.*

D ů k a z . Tvrzení dokážeme sporem, a to pro případ ireducibilně diagonální matice. V průběhu důkazu uvidíme, že je tím pokryt i případ ostře diagonálně dominantní matice. Předpokládejme tedy, že matice A s kladnými diagonálními prvky a nekladnými nediagonálními prvky je ireducibilně diagonálně dominantní, ale není monotónní. Pak existuje vektor x takový, že je $Ax \geqq o$ a přitom aspoň jedna jeho složka je záporná. Položíme-li $x_j = \min(x_1, \ldots, x_n)$, je $x_j < 0$. Nechť pro tento index platí za prvé, že je

$$|a_{jj}| - \sum_{\substack{k=1 \\ k \neq j}}^{n} |a_{jk}| = \sum_{k=1}^{n} a_{jk} > 0. \tag{3.13}$$

Pak platí

$$0 \leqq \sum_{k=1}^{n} a_{jk} x_k = a_{jj} x_j + \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk} x_k \leqq \tag{3.14}$$

$$\leqq a_{jj} x_j + \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk} x_j = x_j \sum_{k=1}^{n} a_{jk} < 0.$$

To je však spor dokazující v tomto případě lemma.

Nechť za druhé pro výše definovaný index j platí

$$|a_{jj}| - \sum_{\substack{k=1 \\ k \neq j}}^{n} |a_{jk}| = \sum_{k=1}^{n} a_{jk} = 0. \tag{3.15}$$

Pak je

$$0 \leqq \sum_{k=1}^{n} a_{jk} x_k = \sum_{k=1}^{n} a_{jk}(x_k - x_j) + x_j \sum_{k=1}^{n} a_{jk} = \sum_{k=1}^{n} a_{jk}(x_k - x_j) \tag{3.16}$$

součet vesměs nekladných čísel $a_{jk}(x_k - x_j)$, $k = 1, \ldots, n$, má být tedy nezáporný. To je možné jen tak, že platí $a_{jk}(x_k - x_j) = 0$ pro $k = 1, \ldots, n$. Je tedy $x_k = x_j$ pro každý takový index k, pro který je $a_{jk} \neq 0$. Buď i_0 index, pro který platí v (3.12) ostrá nerovnost. Protože matice A je ireducibilní, existuje posloupnost indexů $i_1, \ldots, i_s$ taková, že je $a_{ji_1} \neq 0, a_{i_1 i_2} \neq 0, \ldots, a_{i_s i_0} \neq 0$. Podle právě dokázaného je tedy $x_{i_1} = x_j$ a celý postup můžeme zopakovat pro index i_1, pro který také platí, že je $x_{i_1} = \min(x_1, \ldots, x_n)$. Platí-li pro tento index ostrá nerovnost (3.12), jsme hotovi, v opačném případě dostaneme, že je $x_{i_2} = x_{i_1} = x_j$. Pokračujeme-li v tomto postupu dále, nutně dojdeme k indexu, pro který už platí (3.12) ostře. Důkaz je hotov.

Po této krátké exkurzi do teorie speciálních matic se vraťme znovu k vyšetřování metody sítí.

3.2 Lineární diferenciální rovnice druhého řádu

V tomto odstavci podrobně probereme problematiku užití metody sítí k řešení okrajové úlohy (1.12), (1.13). Aby nedocházelo k zbytečným obtížím a nedorozuměním, dohodneme se, že budeme předpokládat, aniž to vždy budeme znovu opakovat, že jsou splněny předpoklady existenční věty 2.2.

3.2.1 Sestavení diferenčních rovnic

Základní myšlenka metody sítí vychází, jak vidíme, z náhrady derivací v daných rovnicích diferenčními podíly. Začneme proto tím, že uvedeme několik nejjednodušších příkladů takových podílů.

Lemma 3.6. *Nechť funkce z má v intervalu I derivace do řádu dvě včetně a nechť platí $\langle x, x+h \rangle \subset I$. Pak existuje bod $\xi \in (x, x+h)$ takový, že platí*

$$\frac{z(x+h) - z(x)}{h} = z'(x) + \tfrac{1}{2} h z''(\xi). \tag{3.17}$$

D ů k a z . Vztah (3.17) je speciálním případem Taylorova vzorce.

Lemma 3.7. *Nechť funkce z má v intervalu I spojité derivace až do třetího řádu včetně a nechť platí $\langle x-h, x+h \rangle \subset I$. Pak existuje bod $\xi \in \langle x-h, x+h \rangle$ takový, že platí*

$$\frac{z(x+h) - z(x-h)}{2h} = z'(x) + \tfrac{1}{6} h^2 z'''(\xi). \tag{3.18}$$

D ů k a z . Za uvedených předpokladů zaručuje Taylorův vzorec existenci bodů $\xi_1 \in (x, x+h)$ a $\xi_2 \in (x-h, x)$ takových, že platí

$$z(x+h) = z(x) + z'(x)h + \tfrac{1}{2}z''(x)h^2 + \tfrac{1}{6}z'''(\xi_1)h^3 \tag{3.19}$$

a

$$z(x-h) = z(x) - z'(x)h + \tfrac{1}{2}z''(x)h^2 - \tfrac{1}{6}z'''(\xi_2)h^3. \tag{3.20}$$

Odečtením rovnice (3.20) od rovnice (3.19) a dělením činitelem $2h$ dostaneme

$$\frac{z(x+h) - z(x-h)}{2h} = z'(x) + \tfrac{1}{12}[z'''(\xi_1) + z'''(\xi_2)]h^3. \tag{3.21}$$

Funkce z''' je však spojitá v uzavřeném intervalu $\langle x-h, x+h\rangle$ a číslo $[z'''(\xi_1) + z'''(\xi_2)]/2$ leží zřejmě mezi jejím maximem a minimem. Z darbouxovské vlastnosti funkcí spojitých na uzavřeném intervalu plyne existence bodu $\xi \in \langle x-h, x+h\rangle$ takového že platí

$$\frac{z'''(\xi_1) + z'''(\xi_2)}{2} = z'''(\xi). \tag{3.22}$$

Dosadíme-li (3.22) do (3.21), dostaneme požadovaný výsledek.

Lemma 3.8. *Nechť funkce z má v I čtyři spojité derivace a nechť platí $\langle x-h, x+h\rangle \subset I$. Pak existuje bod $\xi \in \langle x-h, x+h\rangle$ takový, že platí*

$$\frac{z(x+h) - 2z(x) + z(x-h)}{h^2} = z''(x) + \tfrac{1}{12}z''''(\xi)h^2. \tag{3.23}$$

D ů k a z je zcela analogický důkazu lemmatu 3.7, a proto jej přenecháme čtenáři.

Zanedbáním chybových členů v uvedených lemmatech dostaneme výrazy, které aproximují první a druhou derivaci uvažované funkce. Tyto a podobné výrazy jsme také měli na mysli, když jsme v dřívějším textu hovořili o diferenčních podílech.

Vzhledem k tomu, že na vznik aproximace první derivace dané např. vzorcem (3.21) se lze dívat také tak, že se sestrojí polynom $P(t)$ stupně dvě, který v bodech $t = x-h$, x a $x+h$ nabývá hodnot $z(x-h)$, $z(x)$ a $z(x+h)$ a vypočte se derivace tohoto polynomu v bodě $t = x$ a vzhledem k tomu, že podobná tvrzení platí i pro aproximace z ostatních lemmat, je zřejmé, jak by se sestrojily další diferenční podíly, které by aproximovaly příslušné derivace eventuálně s vyšší přesností.

Rozdělme nyní interval $\langle a, b\rangle$ na n dílů délky h dělicími body (3.1) a pokusme se aproximovat diferenciální operátor vystupující na levé straně rovnice (1.12). Použijeme-li přímočaře lemmat 3.7 a 3.8, tj. provedeme-li v diferenciálním operátoru

$$(Ly)(x) = -[p(x)y'(x)]' + q(x)y(x) \tag{3.24}$$

naznačené derivování a derivace y', resp. y'' v bodě $x = x_k$ nahradíme diferenčními podíly ze vzorců (3.21) a (3.23), dospějeme k operátoru $L_h^{(0)}$, který $(n+1)$-dimenzionálnímu vektoru $y = (y_0, \dots, y_n)^T$ přiřazuje $(n+1)$-dimenzionální vektor $L^{(0)}y$, který je definován předpisem

$$\begin{aligned}(L_h^{(0)}y)_k = \frac{1}{h^2}\{ &- [p(x_k) - \tfrac{1}{2}hp'(x_k)]y_{k-1} + \\ &+ [2p(x_k) + h^2q(x_k)]y_k - [p(x_k) + \tfrac{1}{2}hp'(x_k)]y_{k+1}\}.\end{aligned} \tag{3.25}$$

Tento operátor aproximuje diferenciální operátor L ve smyslu následující věty.

Věta 3.1. *Nechť funkce p má v $\langle a, b\rangle$ spojitou derivaci a nechť funkce q je spojitá. Nechť dále funkce y má v $\langle a, b\rangle$ čtyři spojité derivace. Položme $y^{(\mathrm{pr})} = [y(x_0), \dots, y(x_n)]^T$. Pak platí*

$$(L_h^{(0)}y^{(\mathrm{pr})})_k = (Ly)(x_k) + O(h^2), \quad k = 1, \dots, n-1. \tag{3.26}$$

D ů k a z plyne ihned z lemmat 3.7 a 3.8.

Nahradíme-li derivace v okrajových podmínkách (1.13) diferenčními podíly podle lemmatu 3.6 dostaneme operátory $l_h^{(1)}$ a $l_h^{(2)}$, které vektoru $y = (y_0, \dots, y_n)^T$ přiřazují čísla $l_h^{(1)}y$ a $l_h^{(2)}y$ definované předpisem

$$\begin{aligned} l_h^{(1)}y &= -\alpha_1 p(a)\frac{y_1 - y_0}{h} + \beta_1 y_0, \\ l_h^{(2)}y &= \alpha_2 p(b)\frac{y_n - y_{n-1}}{h} + \beta_2 y_n. \end{aligned} \tag{3.27}$$

Užitím lemmatu 3.6 dostaneme snadno následující větu.

Věta 3.2. *Nechť koeficient p je spojitý v $\langle a, b\rangle$. Pak pro každou funkci y, která má v $\langle a, b\rangle$ spojité dvě derivace, platí*

$$\begin{aligned} l_h^{(1)}y^{(\mathrm{pr})} &= l^{(1)}y + O(h), \\ l_h^{(2)}y^{(\mathrm{pr})} &= l^{(2)}y + O(h). \end{aligned} \tag{3.28}$$

Zde $y^{(\mathrm{pr})}$ značí stejně jako ve větě 3.1 vektor o složkách $y(x_k)$ a $l^{(1)}y$, resp. $l^{(2)}y$ je stručný zápis levých stran okrajových podmínek (1.13).

Jsou-li nyní splněny předpoklady věty 2.2 a předpokládáme-li navíc, že funkce p má v intervalu $\langle a, b\rangle$ tři a funkce q a f dvě spojité derivace, existuje nejen jediné řešení okrajové úlohy (1.12), (1.13), ale toto řešení má dokonce čtyři spojité derivace v $\langle a, b\rangle$. Pro vektor $y^{(\mathrm{pr})}$, jehož složky jsou nyní hodnoty přesného řešení v bodech x_k, $k = 0, \dots, n$, tedy podle vět 3.1 a 3.2 platí

$$(L_h^{(0)}y^{(\mathrm{pr})})_k = f(x_k) + O(h^2), \quad k = 1, \dots, n-1, \tag{3.29}$$

a

$$l_h^{(1)} y^{(\mathrm{pr})} = \gamma_1 + O(h), \tag{3.30}$$
$$l_h^{(2)} y^{(\mathrm{pr})} = \gamma_2 + O(h).$$

Je tedy přirozené hledat přibližné řešení $y = (y_0, \dots, y_n)^T$ ze soustavy rovnic

$$(L_h^{(0)} y)_k = f(x_k), \quad k = 1, \dots, n-1, \tag{3.31}$$
$$l_h^{(1)} y = \gamma_1,$$
$$l_h^{(2)} y = \gamma_2.$$

Úvaha, která nás přivedla k soustavě (3.31) (na níž je zatím třeba dívat se pouze formálně, neboť doposud nevíme vůbec nic o její řešitelnosti), není zdaleka jediná možná a dokonce není asi ani nepřirozenější. Praktické zkušenosti ukazují, že při aproximaci nekonečnědimenzionálních problémů konečnědimenzionálními je žádoucí zachovat co nejvíce vlastností původního problému. V našem případě je diferenciální operátor (3.24) vystupující v diferenciální rovnici (1.12) samoadjungovaný, což by mělo vyústit v symetrii matice aproximující soustavy lineárních rovnic. Tuto vlastnost však matice soustavy (3.31) nemá. K její symetrii je totiž třeba, aby koeficient $p(x_k) + (h/2)p'(x_k) = p(x_k + h/2) + O(h^2)$ u neznámé y_{k+1} v k-té rovnici (3.31) se rovnal koeficientu $p(x_{k+1}) - (h/2)p'(x_{k+1}) = p(x_k + h/2) + O(h^2)$ u neznámé y_k v $(k+1)$-ní rovnici. Tato čísla se však sobě rovnají pouze přibližně. Matice soustavy (3.31) je tedy sice blízká symetrické matici, ale obecně symetrická není.

Při odvození rovnic (3.31) jsme postupovali tak, že jsme naznačené derivování v operátoru L provedli, a tím jsme vlastně vůbec jeho samoadjungovanost nezužitkovali. Položíme-li v tomto operátoru $p(x)y'(x) = z(x)$, aproximujeme-li derivaci $[p(x)y'(x)]' = z'(x)$ v bodě $x = x_k$ podílem $[z(x_k + h/2) - z(x_k - h/2)]/h = [p(x_k + h/2)y'(x_k + h/2) - p(x_k - h/2)y'(x_k - h/2)]/h$ (ve vzorci (3.21)jsme užili $h/2$ místo h) a hodnotu funkce py' v bodě $x = x_k + h/2$, resp. $x = x_k - h/2$ podílem $p(x_k + h/2)[y(x_{k+1}) - y(x_k)]/h$, resp. $p(x_k - h/2)[y(x_k) - y(x_{k-1})]/h$, dostaneme operátor $L_h y$ ($y = (y_0, \dots, y_n)^T$), definovaný předpisem

$$(L_h y)_k = \frac{1}{h^2}\{ -p(x_k - h/2)y_{k-1} + \tag{3.32}$$
$$+ [p(x_k - h/2) + p(x_k + h/2) + h^2 q(x_k)]y_k -$$
$$- p(x_k + h/2)y_{k+1}\}.$$

Aproximační vlastnosti tohoto operátoru jsou popsány v následující větě.

Věta 3.3. *Nechť koeficient p má v intervalu $\langle a, b\rangle$ tři spojité derivace a nechť koeficient q je v tomto intervalu spojitý. Pak pro každou funkci y, která má v intervalu $\langle a, b\rangle$ čtyři spojité derivace, platí*

$$(L_h y^{(\mathrm{pr})})_k = (Ly)(x_k) + O(h^2), \quad k = 1, \dots, n-1. \tag{3.33}$$

D ů k a z . Přímočaré použití lemmatu 3.7, které se nabízí, vede pouze k slabšímu tvrzení $(L_h y^{(\mathrm{pr})})_k = (Ly)(x_k) + O(h)$. Proto je třeba provést důkaz podrobněji. Položme $p(x)y'(x) = z(x)$. Pak funkce z má v intervalu $\langle a, b\rangle$ tři spojité derivace a podle lemmatu 3.7 dostáváme, že pro každé $x \in \langle a + h, b - h\rangle$ platí

$$\frac{z(x + h/2) - z(x - h/2)}{h} = z'(x) + O(h^2). \tag{3.34}$$

Z Taylorova vzorce plyne, že je

$$y(x + h) = y(x + h/2) + y'(x + h/2)\frac{h}{2} + \tag{3.35}$$
$$+ \tfrac{1}{2}y''(x + h/2)\frac{h^2}{4} + \tfrac{1}{6}y'''(x + h/2)\frac{h^3}{8} + O(h^4)$$

a

$$y(x) = y(x + h/2) - y'(x + h/2)\frac{h}{2} + \tag{3.36}$$
$$+ \tfrac{1}{2}y''(x + h/2)\frac{h^2}{4} - \tfrac{1}{6}y'''(x + h/2)\frac{h^3}{8} + O(h^4).$$

Odečtením rovnice (3.36) od rovnice (3.35) dostáváme, že platí

$$\frac{y(x + h) - y(x)}{h} = y'(x + h/2) + \tfrac{1}{24}y'''(x + h/2)h^2 + O(h^3). \tag{3.37}$$

Podobně se dokáže, že je

$$\frac{y(x) - y(x - h)}{h} = y'(x - h/2) + \tfrac{1}{24}y'''(x - h/2)h^2 + O(h^3). \tag{3.38}$$

Vynásobíme-li rovnici (3.37) číslem $p(x + h/2)/h$, rovnici (3.38) číslem $-p(x - h/2)/h$ a vzniklé rovnice sečteme, dostaneme

$$\frac{1}{h}\Big[p(x + h/2)\frac{y(x + h) - y(x)}{h} - p(x - h/2)\frac{y(x) - y(x - h)}{h}\Big] = \tag{3.39}$$
$$= \frac{z(x + h/2) - z(x - h/2)}{h} +$$
$$+ \tfrac{1}{24}[p(x + h/2)y'''(x + h/2) - p(x - h/2)y'''(x - h/2)]h + O(h^2).$$

Protože však funkce py''' má v intervalu $\langle a, b\rangle$ spojitou derivaci, plyne tvrzení (3.33) ihned z (3.39), z věty o střední hodnotě a z (3.34). Věta je dokázána.

Aproximační vlastnosti operátoru L_h jsou tedy stejné jako u operátoru $L_h^{(0)}$ a navíc přidáme-li k rovnicím

$$(L_h y)_k = f(x_k), \quad k = 1, \dots, n-1, \tag{3.40}$$

okrajové podmínky

$$l_h^{(1)} y = \gamma_1, \tag{3.41}$$
$$l_h^{(2)} y = \gamma_2$$

vynásobené vhodnými konstantami, dostaneme soustavu rovnic se symetrickou maticí.

Až dosud jsme generovali diferenční rovnice tak, že jsme nahrazovali derivace v diferenciální rovnici (1.12) a v okrajových podmínkách (1.13) diferenčními podíly. Velmi užitečný, v některých případech asi dokonce lepší postup je založen na tzv. *Marčukově integrální identitě*, která je obsahem následujících věty.

Věta 3.4. *Nechť funkce y splňuje v intervalu $\langle a, b\rangle$ diferenciální rovnici* (1.12) *a nechť koeficienty této diferenciální rovnice splňují předpoklady zformulované v existenční větě* 2.2. *Nechť x_{k-1}, $x_{k-1/2}$, x_k, $x_{k+1/2}$, x_{k+1} jsou libovolné body z intervalu $\langle a, b\rangle$, pro něž platí $x_{k-1} \leqq x_{k-1/2} < x_k < x_{k+1/2} \leqq x_{k+1}$. Pak platí*

$$(3.42)\quad \frac{y(x_{k+1})-y(x_k)}{\int_{x_k}^{x_{k+1}} \frac{1}{p(x)}\,\mathrm{d}\,x} - \frac{y(x_k)-y(x_{k-1})}{\int_{x_{k-1}}^{x_k} \frac{1}{p(x)}\,\mathrm{d}\,x} = \int_{x_{k-1/2}}^{x_{k+1/2}} [q(x)y(x)-f(x)]\,\mathrm{d}\,x +$$

$$+ \frac{1}{\int_{x_k}^{x_{k+1}} \frac{1}{p(x)}\,\mathrm{d}\,x} \int_{x_k}^{x_{k+1}} \left\{ \frac{1}{p(x)} \int_x^{x_{k+1/2}} [q(t)y(t)-f(t)]\,\mathrm{d}\,t \right\} \mathrm{d}\,x +$$

$$+ \frac{1}{\int_{x_{k-1}}^{x_k} \frac{1}{p(x)}\,\mathrm{d}\,x} \int_{x_{k-1}}^{x_k} \left\{ \frac{1}{p(x)} \int_{x_{k-1/2}}^{x} [q(t)y(t)-f(t)]\,\mathrm{d}\,t \right\} \mathrm{d}\,x = 0,$$

$$(3.43)\quad p(x_{k-1})y'(x_{k-1}) = \frac{1}{\int_{x_{k-1}}^{x_k} \frac{1}{p(x)}\,\mathrm{d}\,x} \left[y(x_k) - y(x_{k-1} - \right.$$

$$\left. - \int_{x_{k-1}}^{x_k} \left\{ \frac{1}{p(x)} \int_{x_{k-1}}^{x} [q(t)y(t)-f(t)]\,\mathrm{d}\,t \right\} \mathrm{d}\,x \right]$$

a

$$(3.44)\quad p(x_{k+1})y'(x_{k+1}) = \frac{1}{\int_{x_k}^{x_{k+1}} \frac{1}{p(x)}\,\mathrm{d}\,x} \left[y(x_{k+1}) - y(x_k) + \right.$$

$$\left. + \int_{x_k}^{x_{k+1}} \left\{ \frac{1}{p(x)} \int_x^{x_{k+1}} [q(t)y(t)-f(t)]\,\mathrm{d}\,t \right\} \mathrm{d}\,x \right].$$

D ů k a z . Položme $z(x) = p(x)y'(x)$. Integrací diferenciální rovnice (1.12) v intervalech $(x_{k-1/2}, x_{k+1/2})$, $(x_{k-1/2}, x)$ a $(x, x_{k+1/2})$ dostáváme postupně

$$(3.45)\quad z(x_{k+1/2}) - z(x_{k-1/2}) = \int_{x_{k-1/2}}^{x_{k+1/2}} [q(x)y(x)-f(x)]\,\mathrm{d}\,x,$$

$$(3.46)\quad p(x)y'(x) = z(x_{k-1/2}) + \int_{x_{k-1/2}}^{x} [q(t)y(t)-f(t)]\,\mathrm{d}\,t$$

a

$$(3.47)\quad p(x)y'(x) = z(x_{k+1/2}) - \int_x^{x_{k+1/2}} [q(t)y(t)-f(t)]\,\mathrm{d}\,t.$$

Dělením rovnice (3.46) funkcí p a integrací v intervalu (x_{k-1}, x_k) nalezneme

$$(3.48)\quad y(x_k) - y(x_{k-1}) = z(x_{k-1/2}) \int_{x_{k-1}}^{x_k} \frac{1}{p(x)}\,\mathrm{d}\,x +$$

$$+ \int_{x_{k-1}}^{x_k} \left\{ \frac{1}{p(x)} \int_{x_{k-1/2}}^{x} [q(t)y(t)-f(t)]\,\mathrm{d}\,t \right\} \mathrm{d}\,x$$

neboli

$$(3.49)\quad z(x_{k-1/2}) = \frac{y(x_k)-y(x_{k-1})}{\int_{x_{k-1}}^{x_k} \frac{1}{p(x)}\,\mathrm{d}\,x} -$$

$$- \frac{1}{\int_{x_{k-1}}^{x_k} \frac{1}{p(x)}\,\mathrm{d}\,x} \int_{x_{k-1/2}}^{x_k} \left\{ \frac{1}{p(x)} \int_{x_{k-1/2}}^{x} [q(t)y(t)-f(t)]\,\mathrm{d}\,t \right\} \mathrm{d}\,x.$$

Podobnými manipulacemi odvodíme z rovnice (2.47) vztah

$$(3.50)\quad z(x_{k+1/2}) = \frac{y(x_{k+1})-y(x_k)}{\int_{x_k}^{x_{k+1}} \frac{1}{p(x)}\,\mathrm{d}\,x} +$$

$$+ \frac{1}{\int_{x_k}^{x_{k+1}} \frac{1}{p(x)}\,\mathrm{d}\,x} \int_{x_k}^{x_{k+1}} \left\{ \frac{1}{p(x)} \int_x^{x_{k+1/2}} [q(t)y(t)-f(t)]\,\mathrm{d}\,t \right\} \mathrm{d}\,x.$$

Položíme-li nyní v rovnici (3.49), resp. (3.50) $x_{k-1/2} = x_{k-1}$, resp. $x_{k+1/2} = x_{k+1}$, dostáváme ihned identity (3.43) a (3.44). Dosazení do rovnice (3.45) podle rovnic (3.49) a (3.50) pak dává identitu (3.42). Věta je dokázána.

Položíme-li v identitě (3.42) $x_k = a + kh$, $x_{k+1/2} = x_k + h/2$ a aproximujeme-li integrály následujícími jednoduchými kvadraturními vzorci

$$(3.51)\quad \int_{x_k}^{x_{k+1}} \frac{1}{p(x)}\,\mathrm{d}\,x \approx h \frac{1}{p(x_k + h/2)},$$

$$\int_{x_{k-1}}^{x_k} \frac{1}{p(x)}\,\mathrm{d}\,x \approx h \frac{1}{p(x_k - h/2)},$$

$$\int_{x_{k-1/2}}^{x_{k+1/2}} [q(x)y(x)-f(x)]\,\mathrm{d}\,x \approx h[q(x_k)y(x_k)-f(x_k)],$$

$$\int_{x_k}^{x_{k+1}} \left\{ \frac{1}{p(x)} \int_x^{x_{k+1/2}} [q(t)y(t)-f(t)]\,\mathrm{d}\,t \right\} \mathrm{d}\,x \approx 0,$$

$$\int_{x_{k-1}}^{x_k} \left\{ \frac{1}{p(x)} \int_{x_{k-1/2}}^{x} [q(t)y(t) - f(t)]\,\mathrm{d}t \right\} \mathrm{d}x \approx 0$$

a chyby, kterých se přitom dopustíme, zanedbáme, dostaneme rovnice (3.40), tj. rovnice, v nichž vystupuje už dříve definovaný operátor L_h. Podobně, nahradíme-li integrály ve vzorci (3.43) s $k = 1$ a ve vzorci (3.44) s $k = n - 1$ kvadraturními vzorci

(3.52)
$$\int_{x_0}^{x_1} \frac{1}{p(x)}\,\mathrm{d}x \approx h\frac{1}{p(x_0)},$$

$$\int_{x_{n-1}}^{x_n} \frac{1}{p(x)}\,\mathrm{d}x \approx h\frac{1}{p(x_n)},$$

$$\int_{x_0}^{x_1} \left\{ \frac{1}{p(x)} \int_{x_0}^{x} [q(t)y(t) - f(b)]\,\mathrm{d}t \right\} \mathrm{d}x \approx 0,$$

$$\int_{x_{n-1}}^{x_n} \left\{ \frac{1}{p(x)} \int_{x}^{x_n} [q(t)y(t) - f(t)]\,\mathrm{d}t \right\} \mathrm{d}x \approx 0,$$

a užijeme-li vzniklé aproximace v okrajových podmínkách studovaného problému, dostaneme operátory $l_h^{(1)}$ a $l_h^{(1)}$. Užijeme-li v těchto vzorcích první dva kvadraturní vzorce (3.51) a vzorce

(3.53)
$$\int_{x_0}^{x_1} \left\{ \frac{1}{p(x)} \int_{x_0}^{x} [q(t)y(t) - f(t)]\,\mathrm{d}t \right\} \mathrm{d}x \approx$$

$$\approx \tfrac{1}{2}h^2 \frac{1}{p(x_0 + h/2)} [q(x_0)y(x_0) - f(x_0)],$$

$$\int_{x_{n-1}}^{x_n} \left\{ \frac{1}{p(x)} \int_{x}^{x_n} [q(t)y(t) - f(t)]\,\mathrm{d}t \right\} \mathrm{d}x \approx$$

$$\approx \tfrac{1}{2}h^2 \frac{1}{p(x_n - h/2)} [q(x_n)y(x_n) - f(x_n)],$$

dostaneme operátory $l_h^{(1,\mathrm{d})}$ a $l_h^{(2,\mathrm{d})}$ dané rovnicemi

(3.54)
$$l_h^{(1,\mathrm{d})} y = -\alpha_1 p(x_0 + h/2)\frac{y_1 - y_0}{h} + [\tfrac{1}{2}\alpha_1 h q(x_0) + \beta_1]y_0,$$
$$l_h^{(2,\mathrm{d})} y = \alpha_2 p(x_n - h/2)\frac{y_n - y_{n-1}}{h} + [\tfrac{1}{2}\alpha_2 h q(x_n) + \beta_2]y_n.$$

Pro tyto operátory platí následující věta.

Věta 3.5. *Nechť funkce p má v intervalu $\langle a, b\rangle$ dvě spojité derivace a nechť funkce q je v $\langle a, b\rangle$ spojitá. Pak pro každou funkci y, která má v $\langle a, b\rangle$ tři spojité derivace, platí*

(3.55)
$$l_h^{(1,\mathrm{d})} y^{(\mathrm{pr})} = l^{(1)} y + \tfrac{1}{2}\alpha_1 h(Ly)(x_0) + O(h^2),$$
$$l_h^{(2,\mathrm{d})} y^{(\mathrm{pr})} = l^{(2)} y + \tfrac{1}{2}\alpha_2 h(Ly)(x_n) + O(h^2),$$

kde $y^{(\mathrm{pr})}$ je vektor o složkách $y(x_0), \ldots, y(x_n)$.

D ů k a z . Za učiněných předpokladů platí podle Taylorova vzorce

(3.56)
$$p(x_0 + h/2)\frac{y(x_1) - y(x_0)}{h} =$$
$$= [p(x_0) + \tfrac{1}{2}hp'(x_0) + O(h^2)][y'(x_0) + \tfrac{1}{2}hy''(x_0) + O(h^2)] =$$
$$= p(x_0)y'(x_0) + \tfrac{1}{2}hp'(x_0)y'(x_0) + \tfrac{1}{2}hp(x_0)y''(x_0) + O(h^2) =$$
$$= p(x_0)y'(x_0) + \tfrac{1}{2}h[p(x_0)y'(x_0)]' + O(h^2) =$$
$$= p(x_0)y'(x_0) + \tfrac{1}{2}h[q(x_0)y(x_0) - (Ly)(x_0)] + O(h^2).$$

Je tedy

(3.57)
$$l_h^{(1,\mathrm{d})} y^{(\mathrm{pr})} = -\alpha_1 p(x_0)y'(x_0) - \tfrac{1}{2}\alpha_1 h[q(x_0)y(x_0) - (Ly)(x_0)] +$$
$$+ [\tfrac{1}{2}\alpha_1 h q(x_0) + \beta_1]y(x_0) + O(h^2).$$

Z rovnice (3.57) však už plyne první rovnice (3.55). Druhá rovnice (3.55) se dokáže úplně analogicky. Věta je dokázána.

Nahradíme-li okrajové podmínky rovnicemi

(3.58)
$$l_h^{(1,\mathrm{d})} y = \gamma_1 + \tfrac{1}{2}\alpha_1 h f(x_0),$$
$$l_h^{(2,\mathrm{d})} y = \gamma_2 + \tfrac{1}{2}\alpha_2 h f(x_n),$$

($y = (y_0, \ldots, y_n)^T$), dopustíme se tedy na rozdíl od rovnic (3.41) chyby velikosti $O(h^2)$. Rovnice (3.58) přitom nejsou o nic složitější než rovnice (3.41). V následujícím odstavci uvidíme, že lepší aproximační vlastnosti rovnic (3.58) oproti rovnicím (3.41) se projeví tak, že jejich užití jako okrajových podmínek pro operátor L_h vede k celkové chybě velikosti $O(h^2)$, zatímco užití rovnic (3.41) pouze k chybě řádu $O(h)$. Intuitivně není tento výsledek překvapující, neboť při odvození operátorů $l_h^{(1)}$ a $l_h^{(2)}$, ať už užitím diferenčních podílů nebo kvadraturních vzorců v integrální identitě z věty 3.4, jsme užili méně kvalitní aproximace než při konstrukci operátoru L_h a operátorů $l_h^{(1,\mathrm{d})}$ a $l_h^{(2,\mathrm{d})}$. Z tohoto hlediska je tedy odvození rovnic (3.58) z integrální identity pomocí kvadraturních vzorců (3.51) a (3.53) dokonce přirozenější, než odvození rovnic (3.41) pomocí kvadraturních vzorců (3.52), neboť vzorce (3.53) aproximují příslušné integrály se stejnou řádovou chybou jako vzorce (3.51). Zakončeme stručný výklad o generování konečnědimenzionálních aproximací daného diferenciálního operátoru poznámkou, že užitím přesnějších kvadraturních vzorců v integrálních identitách z věty 3.4, než které jsme dosud užili, lze dosáhnout

aproximací daného operátoru, které jsou vyššího řádu, než tomu bylo u operátoru L_h. K této otázce se v odst. 3.2.2 ještě vrátíme.

Až dosud jsme sestrojovali konečnědimenzionální soustavy, jejichž řešení má být přibližným řešením dané okrajové úlohy formálně, aniž bychom studovali, zda tyto soustavy mají vůbec rozumný smysl, tj. zda jejich matice jsou regulární. Nyní si všimneme této otázky. Při jejím řešení a i při studiu konvergence v dalším odstavci nám prokáže velmi cenné služby následující lemma. Před jeho formulací se však ještě dohodněme, že o integračním kroku $h = (b-a)/n$ budeme všude v odst. 3.2 předpokládat, že je volen v případě, že je $q(x) \equiv 0$ v $\langle a, b\rangle$ libovolně a v opačném případě tak, aby existoval bod $x_{\tilde{k}} = a + \tilde{k}h$, $0 < \tilde{k} < n$, takový, že platí $q(x_{\tilde{k}}) > 0$. K splnění tohoto požadavku stačí v důsledku spojitosti funkce q zřejmě, aby toto h bylo dostatečně malé.

Lemma 3.9. *Nechť funkce p a q jsou spojité a nechť platí nerovnosti* (2.1) a (2.2). *Nechť dále $\eta = (\eta_0, \ldots, \eta_n)^T$ je libovolný vektor, pro který platí*

$$(L_h\eta)_k \leqq 0, \quad k = 1, \ldots, n-1. \tag{3.59}$$

Nechť konečně je

$$M = \max_{k=0,\ldots,n} \eta_k > 0. \tag{3.60}$$

Pak platí

$$M = \max(\eta_0, \eta_n) \tag{3.61}$$

a existuje-li index k_0, $0 < k_0 < n$ takový, že je $\eta_{k_0} = M$, je $q(x) \equiv 0$ v $\langle a, b\rangle$ a $\eta_k = M$ pro $k = 0, \ldots, n$.

D ů k a z . Předpokládejme, že existuje index k_0, $1 \leqq k_0 \leqq n-1$, takový, že platí $\eta_{k_0} = M$. Pak je

$$\begin{aligned} 0 \geqq & -p(x_{k_0} - h/2)\eta_{k_0-1} + [p(x_{k_0} - h/2) + p(x_{k_0} + h/2) + h^2 q(x_{k_0})]M - \\ & - p(x_{k_0} + h/2)\eta_{k_0+1} \geqq h^2 q(x_{k_0})M \geqq 0. \end{aligned} \tag{3.62}$$

Protože M je kladné a $q(x_{k_0})$ nezáporné, plyne z nerovnosti (3.62) především, že je $q(x_{k_0}) = 0$. Kdyby platilo $\eta_{k_0-1} < M$ nebo $\eta_{k_0+1} < M$, bylo by tedy

$$\begin{aligned} 0 \geqq & -p(x_{k_0} - h/2)\eta_{k_0-1} + [p(x_{k_0} - h/2) + p(x_{k_0} + h/2)]M - \\ & - p(x_{k_0} + h/2)\eta_{k_0+1} > 0, \end{aligned} \tag{3.63}$$

což není možné. Platí proto $\eta_{k_0-1} = \eta_{k_0+1} = M$ a celou úvahu je možno zopakovat pro indexy $k_0 - 1$ a $k_0 + 1$. Postupně tak dostaneme, že platí

$$q(x_k) = 0, \quad k = 1, \ldots, n-1, \tag{3.64}$$

a

$$\eta_k = M, \quad k = 0, \ldots, n. \tag{3.65}$$

Není-li funkce q identicky rovna nule v intervalu $\langle a, b\rangle$, není rovnice (3.64) pro $k = \tilde{k}$ splněna. Tento spor dokazuje, že v případě, že funkce q není rovna nule identicky, platí $\eta_k < M$ pro $k = 1, \ldots, n-1$. V tomto případě je tedy lemma dokázáno. V případě, že je $q(x) \equiv 0$ v $\langle a, b\rangle$, však plyne tvrzení lemmatu z rovnic (3.65).

Stručně můžeme vyjádřit podstatu tvrzení lemmatu 3.9 tak, že funkce η_k definovaná na diskrétní množině bodů x_k a splňující nerovnosti (3.59) nabývá kladného maxima v krajním bodě. To je také důvod, proč se toto tvrzení nazývá *princip maxima*. Pomocí něj už snadno dokážeme regularitu matic námi sestrojených soustav. Jako příklad uveďme následující větu.

Věta 3.6. *Nechť jsou splněny předpoklady existenční věty 2.2 a nechť h je dostatečně malé. Pak soustava* (3.40), (3.41) *má při libovolné funkci f a při libovolných γ_1 a γ_2 právě jedno řešení.*

D ů k a z . Buď η_k, $k = 0, \ldots, n$, řešením příslušné homogenní soustavy, tj.

$$(L_h\eta)_k = 0, \quad k = 1, \ldots, n-1, \tag{3.66}$$

a

$$l_h^{(1)}\eta = 0, \quad l_h^{(2)}\eta = 0. \tag{3.67}$$

Položme $M = \max_{k=0,\ldots,n} \eta_k$ a předpokládejme, že je $M > 0$. Pak podle lemmatu 3.9 platí, že $M = \max(\eta_0, \eta_n)$ a zřejmě bez újmy na obecnosti můžeme předpokládat, že je $\eta_0 \geqq \eta_n$, takže je $M = \eta_0$. Píšeme-li první rovnici (3.67) podrobněji, máme

$$[\alpha_1 p(a) + h\beta_1]\eta_0 - \alpha_1 p(a)\eta_1 = 0. \tag{3.68}$$

Je-li nyní $\alpha_1 = 0$, je $\beta_1 > 0$ a z rovnice (3.68) plyne, že je $\eta_0 = 0$. To však znamená, že je $M = 0$ a to je spor dokazující, že je $M \leqq 0$. Je-li $\alpha_1 > 0$, je též $\beta_1 > 0$. Kdyby tomu tak totiž nebylo, plynulo by z rovnice (3.68), že je $\eta_1 = \eta_0$, a tedy podle lemmatu 3.9 by bylo $\eta_k = M$ pro $k = 0, \ldots, n$ a $q(x) \equiv 0$ v $\langle a, b\rangle$. Z druhé rovnice (3.67) by tedy plynulo, že je

$$0 = [\alpha_2 p(b) + h\beta_2]M - \alpha_2 p(b)M = h\beta_2 M, \tag{3.69}$$

a tedy $\beta_2 = 0$. To však není možné, neboť případ $\beta_1 = \beta_2 = 0$ a $q(x) \equiv 0$ je předpoklady věty 2.2 vyloučen. Při $\alpha_1 > 0$ je tedy skutečně $\beta_1 > 0$ a z rovnice (3.68) dostáváme

$$0 = [\alpha_1 p(a) + h\beta_1]M - \alpha_1 p(a)\eta_1 \geqq h\beta_1 M > 0, \tag{3.70}$$

tedy opět spor, dokazující, že je $M \leqq 0$. Celkem jsme tak dostali, že je $\eta_k \leqq 0$ pro $k = 0, \ldots, n$. Zopakujeme-li celou úvahu pro funkci $-\eta_k$, dostaneme, že je $\eta_k \geqq 0$ pro $k = 0, \ldots, n$. Celkem tedy platí $\eta_k = 0$ pro $k = 0, \ldots, n$. Soustava (3.66) a (3.67) má tedy pouze triviální řešení, což dokazuje větu.

Poznámka 3.1. Předpokládáme-li, že místo rovnic (3.66) a (3.67) platí nerovnosti $(L_h\eta)_k \leqq 0$, $k = 1, \ldots, n-1$, $l_h^{(1)}\eta \leqq 0$ a $l_h^{(2)}\eta \leqq 0$, dokážeme úplně stejným postupem jako v důkazu věty 3.6, že je $\eta_k \leqq 0$, $k = 0, \ldots, n$. Matice soustavy (3.40), (3.41) je tedy zřejmě monotónní (implikace $Ax \geqq 0 \Rightarrow x \geqq 0$ je zřejmě ekvivalentní implikaci $Ax \leqq 0 \Rightarrow \leqq 0$).

Monotónnost matice soustavy (3.40), (3.41) (a tedy speciálně její regularitu) je možno také dokázat pomocí Collatzova lemmatu 3.5. Zde je však třeba projevit určitou opatrnost. Uvažovaná matice má sice diagonální prvky kladné a nediagonální prvky nekladné, ostře diagonálně dominantní však obecně není, takže je třeba zkoumat její ireducibilní diagonální dominanci. Podle lemmatu 3.4 je třídiagonální matice ireducibilní právě tehdy, jsou-li její nediagonální prvky nenulové. To je však pro uvažovanou matici splněno pouze v případě, jsou-li oba koeficienty α_1 a α_2 nenulové. Diagonální dominance se v tomto případě už prověří snadno: Všechny řádkové součty jsou nezáporné, jak je vidět na první pohled a kladné jsou v nultém nebo n-tém řádku podle toho, je-li β_1 nebo β_2 různé od nuly. Jsou-li obě tato čísla nulová, nesmí se koeficient q rovnat nule identicky a ostrá nerovnost pak platí pro k-tý řádek, kde k je určeno tak, že pro něj je $q(x_k) > 0$. Je-li α_1 nebo α_2 rovno nule, je matice soustavy reducibilní. V tomto případě je však β_1 nebo β_2 nenulové a nultá nebo n-tá rovnice udává přímo hodnotu nulté nebo n-té neznámé. Dosadíme-li tuto hodnotu do první nebo $(n-1)$-ní rovnice, dostaneme soustavu, jejíž matice je řádu n nebo $n-1$ a je zřejmě ireducibilně diagonálně dominantní. Collatzovo lemma pak aplikujeme na tuto redukovanou matici.

Monotónnost matice soustavy (3.31) se dokáže stejně snadno. Rovněž tak snadno se dokáže monotónnost soustavy, která vznikne tak, že se okrajové podmínky (3.41) nahradí okrajovými podmínkami (3.58). Lze k tomu použít opět Collatzovo lemma nebo princip maxima, který platí i pro operátor $L_h^{(0)}$.

3.2.2 Konvergence

V předešlém odstavci jsme na základě rovnic (3.33) a (3.30) usoudili, že je rozumné počítat přibližné řešení $y_0, \ldots, y_n$ dané okrajové úlohy z rovnic (3.40) a (3.41) a ukázali jsme, že tento postup má smysl, neboť vektor $y_0, \ldots, y_n$ je rovnicemi (3.40), (3.41) jednoznačně určen. V tomto odstavci ukážeme konvergenci popsaného postupu, tj. ukážeme, že rozdíl $\eta_k = y_k - y(x_k)$ (který se i zde podobně jako v kap. I nazývá *celková diskretizační chyba*) lze volbou dostatečně malého h učinit libovolně malý.

Nechť jsou splněny předpoklady věty 2.2 zaručující existenci a jednoznačnost přesného řešení a položme

$$(L_h\eta)_k = \varepsilon_k, \quad k = 1, \ldots, n-1, \tag{3.71}$$

a

$$l_h^{(1)}\eta = \varepsilon_0, \qquad l_h^{(2)}\eta = \varepsilon_n. \tag{3.72}$$

Právě zavedené veličiny ε_k se nazývají *lokální chyby* metody sítí a vlastně udávají, jaké chyby se dopouštíme aproximací daných diferenciálních operátorů konečnědimenzionálními operátory. Předpokládáme-li navíc, že funkce p, q a f jsou dostatečně hladké, plyne z vět 3.3 a 3.2 existence konstanty M takové, že pro dostatečně malé h platí

$$|\varepsilon_k| \leqq Mh^2, \quad k = 1, \ldots, n-1, \tag{3.73}$$

a

$$|\varepsilon_0| \leqq Mh, \qquad |\varepsilon_n| \leqq Mh. \tag{3.74}$$

Z důvodu dalšího zestručnění zápisu se dohodneme, že konstantu M zde a v podobných případech pokládáme za tzv. generickou konstantu, tj. za konstantu, která v jednotlivých případech může nabývat různých hodnot.

Ze vztahů (3.71) až (3.74) vidíme, že chyba splňuje soustavu lineárních algebraických rovnic s malou pravou stranou. V následující větě ukážeme, že to má za následek i malost příslušného řešení.

Věta 3.7. *Nechť jsou splněny předpoklady věty 2.2 a nechť navíc funkce p má v intervalu $\langle a, b\rangle$ tři spojité derivace a funkce q a f dvě spojité derivace. Buď dále y_k přibližné řešení vypočtené ze soustavy* (3.40), (3.41). *Pak existují konstanty M a $h_0 > 0$ takové, že platí*

$$|y_k - y(x_k)| \leqq Mh \tag{3.75}$$

pro $k = 0, \ldots, n$ a $h \leqq h_0$.

D ů k a z . Buď z funkce, která je v intervalu $\langle a, b\rangle$ řešením diferenciální rovnice

$$-\big(p(x)z'\big)' + q(x)z = 1 \tag{3.76}$$

s okrajovými podmínkami

$$-\alpha_1 p(a)z'(a) + \beta_1 z(a) = 1, \qquad \alpha_2 p(b)z'(b) + \beta_2 z(b) = 1. \tag{3.77}$$

Z předpokladů plyne, že tato funkce nejen existuje a je jediná, ale má v intervalu $\langle a, b\rangle$ čtyři spojité derivace. Podle vět 3.3 a 3.2 tedy existuje konstanta K taková, že platí

$$|(L_h z)_k - 1| \leqq Kh^2, \quad k = 1, \ldots, n-1, \tag{3.78}$$

a

$$|l_h^{(1)} z - 1| \leqq Kh, \tag{3.79}$$
$$|l_h^{(2)} z - 1| \leqq Kh.$$

(Položili jsme $z = \big(z(x_0), \dots, z(x_k)\big)^T$.) Z nerovností (3.78), (3.79) plyne existence čísel θ_k, $k = 0, \dots, n$, takových, že platí $|\theta_k| \leqq 1$ pro $k = 0, \dots, n$ a že je

$$(L_h z)_k = 1 + \theta_k K h^2, \quad k = 1, \dots, n-1, \tag{3.80}$$

a

$$l_h^{(1)} z = 1 + \theta_0 K h, \tag{3.81}$$
$$l_h^{(2)} z = 1 + \theta_n K h.$$

Zvolme nyní h_0 pevně tak, aby platilo

$$h_0 < \min\Big(\frac{1}{K}, \frac{1}{K^{1/2}}\Big) \tag{3.82}$$

a položme

$$N = \max\Big(\frac{M}{1 - Kh_0}, \frac{Mh_0}{1 - Kh_0^2}\Big), \tag{3.83}$$

kde M je konstanta z nerovností (3.73) a (3.74). Pak pro $h \leqq h_0$ platí

$$0 < 1 - Kh_0 \leqq 1 + \theta_0 K h, \tag{3.84}$$

a tedy je

$$Nh(1 + \theta_0 K h) \geqq N(1 - Kh_0)h \geqq Mh. \tag{3.85}$$

Analogicky se dokáže, že je též

$$Nh(1 + \theta_n K h) \geqq Mh \tag{3.86}$$

a

$$Nh(1 + \theta_k K h^2) \geqq Mh^2, \quad k = 1, \dots, n-1. \tag{3.87}$$

Položíme-li tedy

$$v_k = Nhz(x_k), \tag{3.88}$$

plyne z (3.71) až (3.74), (3.80), (3.81) a (3.85) až (3.87), že pro $h \leqq h_0$ platí

$$|(L_h \eta)_k| \leqq (L_h v)_k, \quad k = 1, \dots, n-1, \tag{3.89}$$

a

$$|l_h^{(1)} \eta| \leqq l_h^{(1)} v, \tag{3.90}$$
$$|l_h^{(2)} \eta| \leqq l_h^{(2)} v.$$

Protože matice soustavy (3.71), (3.72) je podle poznámky 3.1 monotónní, plyne z nerovností (3.89) a (3.90) a z lemmatu 3.2, že platí $|\eta_k| \leqq v_k$ pro $k = 0, \dots, n$. Protože pomocná funkce z je v intervalu $\langle a, b\rangle$ omezená, plyne odtud tvrzení věty.

Právě dokázaná věta dává tedy odpověď na otázku po konvergenci metody sítí (3.40), (3.41). Konvergence metody sítí (3.31) by se dokázala úplně stejně. Podtrhněme tu skutečnost, že k důkazu tvrzení věty 3.7 bylo třeba předpokládat větší hladkost hledaného řešení, než kterou zajišťuje existenční věta. To je v situaci, kdy dokazujeme vlastně tvrzení o rychlosti konvergence, a tedy více, než pouhou konvergenci, typické a není na tom nic překvapivého. S podobnými jevy jsme se setkali i v případě řešení úloh s počátečními podmínkami. Stejně jako tam pouhá konvergence se dá dokázat za podstatně slabších předpokladů o hladkosti hledaného řešení. Zde je však příslušný postup výrazně složitější, než tomu bylo v důkazu věty 3.7, a to je také důvod, proč jsme se omezili pouze na hladký případ.

Vzhledem k tomu, že při sestavování $n-1$ rovnic (3.40), jejichž počet se při $h \to 0$ bez omezení zvětšuje, jsme se dopustili lokální chyby velikosti $O(h^2)$ a pouze ve dvou rovnicích chyby $O(h)$, může vzniknout dojem, že lepší odhad celkové diskretizační chyby než pouze $O(h)$ nebyl odvozen pouze proto, že jsme užili ne dosti důmyslnou důkazovou techniku. Následující jednoduchá věta však ukazuje, že tomu tak není.

Věta 3.8. *Odhad chyby metody sítí* (3.40), (3.41) *daný vzorcem* (3.75) *nelze zlepšit.*

D ů k a z . Buď y řešením diferenciální rovnice

$$y'' = 2 \tag{3.91}$$

v intervalu $\langle -1, 1\rangle$ s okrajovými podmínkami

$$-y'(-1) + y(-1) = 3, \tag{3.92}$$
$$y'(1) + y(1) = 3,$$

takže jsou splněny předpoklady existenční věty a přesné řešení je funkce $y(x) = x^2$. Přibližné řešení metodou (3.40), (3.41) se vypočte z diferenční rovnice

$$-y_{k-1} + 2y_k - y_{k+1} = -2h^2, \quad k = 1, \dots, n-1, \tag{3.93}$$

s okrajovými podmínkami

$$-\frac{y_1 - y_0}{h} + y_0 = 3, \tag{3.94}$$
$$\frac{y_n - y_{n-1}}{h} + y_n = 3.$$

Z teorie lineárních diferenčních rovnic s konstantními koeficienty (nebo také přímým dosazením) ihned plyne, že řešení těchto rovnic je dáno vzorcem

$$y_k = (-1 + kh)^2 + h = x_k^2 + h. \tag{3.95}$$

Pro chybu η_k tedy platí $\eta_k = h$, což dokazuje větu.

Návod, jak sestrojit metodu sítí, jejíž chyba je řádu $O(h^2)$, je vlastně obsažen v důkazu věty 3.7. Z něj je vidět, že žádaný výsledek se dosáhne, podaří-li se volit funkci v_k ve tvaru $Nz(x_k)h^2$ a nikoliv pouze ve tvaru $Nz(x_k)h$. K tomu však zřejmě stačí, aby okrajové podmínky byly aproximovány s přesností $O(h^2)$ a aby přitom nebyla porušena monotónnost příslušné matice. Okrajové podmínky, které mají tuto vlastnost, jsou dány operátory $l_h^{(1,\mathrm{d})}$ a $l_h^{(2,\mathrm{d})}$ z předešlého odstavce. Připojením okrajových podmínek (3.58) k rovnicím (3.40) dostaneme tedy metodu sítí, jejíž celková chyba je řádu $O(h^2)$. Důkaz tohoto tvrzení se provede úplně stejně jako důkaz věty 3.7.

My zde zmíněné tvrzení dokážeme ještě jednou jiným postupem, který bude sice formálně komplikovanější, ale vzhledem k tomu, že se v něm nebude užívat princip maxima (a tedy ani s ním související teorie monotónních matic), bude jej možné přednést na vyšetřování rovnic vyšších řádů, kde princip maxima neplatí. Základní idea vlastně všech konvergenčních důkazů, které jsme až dosud v celé knize prováděli, je i zde samozřejmě stejná: vždy sestrojujeme apriorní odhad pro přibližné řešení v závislosti na pravých stranách rovnic, které je definují, neboť celková diskretizační chyba splňuje rovnice tohoto typu. Ve větě 3.7 jsme takový odhad konstruovali na základě principu maxima. Zde to provedeme pomocí tzv. *energetických nerovností*. Základní myšlenku příslušného postupu ukážeme nejprve na spojitém případě, kde je méně zastřena komplikovanou symbolikou.

Nechť funkce p, q a y jsou dostatečně hladké a nechť platí nerovnosti (2.1) a (2.2). Buď dána funkce y a položme

$$(Ly)(x) = -[p(x)y'(x)]' + q(x)y(x) \tag{3.96}$$

a

$$\begin{aligned} l^{(1)}y &= -\alpha_1 p(a)y'(a) + \beta_1 y(a), \\ l^{(2)}y &= \alpha_2 p(b)y'(b) + \beta_2 y(b). \end{aligned} \tag{3.97}$$

Naším úkolem je odhadnout nějakou normu funkce y pomocí vhodné normy funkce Ly a pomocí čísel $l^{(1)}y$ a $l^{(2)}y$. Pro určitost předpokládejme, že koeficienty α_i a β_i v rovnicích (3.97) jsou vesměs kladné. Tento případ lze pokládat za modelový a modifikace uvedeného postupu, které je nutno provést v jiných případech, jsou snadné. Čtenář si je provede podle návodu, který uvedeme později, bez problémů sám.

Vynásobme rovnici (3.96) funkcí y a integrujme v mezích od a do b. Užitím integrace per partes a dosazením za $p(a)y'(a)$ a $p(b)y'(b)$ z rovnic (3.97) dostaneme

$$\begin{aligned} \int_a^b (Ly)(x)y(x)\,\mathrm{d}x &= \frac{\beta_2}{\alpha_2}y^2(b) - \frac{1}{\alpha_2}(l^{(2)}y)y(b) + \\ &+ \frac{\beta_1}{\alpha_1}y^2(a) - \frac{1}{\alpha_1}(l^{(1)}y)y(a) + \int_a^b p(x)[y'(x)]^2\,\mathrm{d}x + \int_a^b q(x)y^2(x)\,\mathrm{d}x. \end{aligned} \tag{3.98}$$

Odtud vzhledem k předpokladům (2.1) a (2.2) ihned plyne, že platí

$$\begin{aligned} &\frac{1}{\alpha_1}y(a)l^{(1)}y + \frac{1}{\alpha_2}y(b)l^{(2)}y + \int_a^b (Ly)(x)y(x)\,\mathrm{d}x \geqq \\ &\geqq \frac{\beta_1}{\alpha_1}y^2(a) + \frac{\beta_2}{\alpha_2}y^2(b) + p_0\int_a^b [y'(x)]^2\,\mathrm{d}x. \end{aligned} \tag{3.99}$$

Pokusme se nyní odhadnout hodnoty funkce y pomocí výrazu na pravé straně nerovnosti (3.99) (který je v podstatě kvadrátem tzv. *energetické normy* funkce y; tato norma hraje významnou roli ve funkcionálně analytické teorii diferenciálních rovnic). Zřejmě platí

$$y(x) = \tfrac{1}{2}y(a) + \tfrac{1}{2}y(b) + \tfrac{1}{2}\int_a^x y'(t)\,\mathrm{d}t - \tfrac{1}{2}\int_x^b y'(t)\,\mathrm{d}t. \tag{3.100}$$

Použijeme-li nerovnost

$$(\alpha+\beta)^2 \leqq 2\alpha^2 + 2\beta^2, \tag{3.101}$$

která platí pro libovolná reálná α a β, dostáváme

$$\begin{aligned} [y(x)]^2 &\leqq 2\left[\tfrac{1}{2}y(a) + \tfrac{1}{2}y(b)\right]^2 + 2\left[\tfrac{1}{2}\int_a^x y'(t)\,\mathrm{d}t - \tfrac{1}{2}\int_x^b y'(t)\,\mathrm{d}t\right]^2 \leqq \\ &\leqq y^2(a) + y^2(b) + \left(\int_a^x y'(t)\,\mathrm{d}t\right)^2 + \left(\int_x^b y'(t)\,\mathrm{d}t\right)^2. \end{aligned} \tag{3.102}$$

Podle Schwartzovy nerovnosti je však

$$\left(\int_a^x 1.\,y'(t)\,\mathrm{d}t\right)^2 \leqq \int_a^x 1^2\,\mathrm{d}t\int_a^x [y'(t)]^2\,\mathrm{d}t \leqq (x-a)\int_a^b [y'(x)]^2\,\mathrm{d}x \tag{3.103}$$

a podobně

$$\left(\int_x^b y'(t)\,\mathrm{d}t\right)^2 \leqq (b-x)\int_a^b [y'(x)]^2\,\mathrm{d}x. \tag{3.104}$$

Dosazení (3.103) a (3.104) do (3.102) dává

$$[y(x)]^2 \leqq y^2(a) + y^2(b) + (b-a)\int_a^b [y'(x)]^2\,\mathrm{d}x. \tag{3.105}$$

Protože tato nerovnost platí pro libovolné $x \in \langle a,b\rangle$, plyne odtud ihned, že je

$$\|y\|_{\mathcal{L}_\infty}^2 \leqq \gamma\left\{\frac{\beta_1}{\alpha_1}y^2(a) + \frac{\beta_2}{\alpha_2}y^2(b) + p_0\int_a^b [y'(x)]^2\,\mathrm{d}x\right\}, \tag{3.106}$$

kde jsme položili

$$\gamma = \max\left(\frac{\alpha_1}{\beta_1}, \frac{\alpha_2}{\beta_2}, \frac{b-a}{p_0}\right) \tag{3.107}$$

a symbol $\|.\|_{\mathcal{L}_\infty}$ značí normu v prostoru $\mathcal{L}_\infty$, tj. $\|y\|_{\mathcal{L}_\infty} = \sup_{x\in\langle a,b\rangle} |y(x)|$. Z nerovností (3.106) a (3.99) plyne, že platí

$$\|y\|^2_{\mathcal{L}_\infty} \leqq \gamma\left[\frac{1}{\alpha_1}y(a)l^{(1)}y + \frac{1}{\alpha_2}y(b)l^{(2)}y + \int_a^b (Ly)(x)y(x)\,\mathrm{d}x\right] \leqq \tag{3.108}$$

$$\leqq \gamma\|y\|_{\mathcal{L}_\infty}\left[\frac{1}{\alpha_1}|l^{(1)}y| + \frac{1}{\alpha_2}|l^{(2)}y| + \int_a^b |(Ly)(x)|\,\mathrm{d}x\right].$$

Odtud však už plyne hledaný odhad

$$\|y\|_{\mathcal{L}_\infty} \leqq \gamma\left[\frac{1}{\alpha_1}|l^{(1)}y| + \frac{1}{\alpha_2}|l^{(2)}y| + \int_a^b |(Ly)(x)|\,\mathrm{d}x\right]. \tag{3.109}$$

Podařilo se nám tedy skutečně odhadnou $\mathcal{L}_\infty$ normu funkce y pomocí $\mathcal{L}_1$ normy funkce Ly. V několika následujících lemmatech ukážeme, jak lze popsaný postup přenést na konečnědimenzionální případ.

Lemma 3.10. *Nechť γ_k a δ_k jsou libovolné posloupnosti čísel definovaných pro $k = 0, \ldots, n-1$. Pak platí*

$$\sum_{k=1}^{n-1}(\gamma_k - \gamma_{k-1})\delta_k = \gamma_{n-1}\delta_{n-1} - \gamma_0\delta_0 - \sum_{k=1}^{n-1}\gamma_{k-1}(\delta_k - \delta_{k-1}). \tag{3.110}$$

D ů k a z . Vzorec (3.110) ihned plyne ze zřejmé identity $\gamma_k\delta_k - \gamma_{k-1}\delta_{k-1} = (\gamma_k - \gamma_{k-1})\delta_k + \gamma_{k-1}(\delta_k - \delta_{k-1})$.

Tvrzení tohoto lemmatu je obdobou vzorce pro integraci per partes.

Lemma 3.11. *Nechť koeficienty p a q jsou spojité a nechť platí nerovnosti* (2.1) a (2.2). *Nechť dále je $\alpha_i > 0$ pro $i = 1, 2$ a h je kladné číslo. Pak pro libovolný vektor $\boldsymbol{\eta} = (\eta_0, \ldots, \eta_n)^T$ platí*

$$h\frac{1}{\alpha_1}\eta_0 l_h^{(1,\mathrm{d})}\boldsymbol{\eta} + h\frac{1}{\alpha_2}\eta_n l_h^{(2,\mathrm{d})}\boldsymbol{\eta} + h^2\sum_{k=1}^{n-1}(L_y\boldsymbol{\eta})_k\eta_k \geqq \tag{3.111}$$

$$\geqq h\frac{\beta_1}{\alpha_1}\eta_0^2 + h\frac{\beta_2}{\alpha_2}\eta_n^2 + p_0\sum_{k=1}^{n}(\eta_k - \eta_{k-1})^2.$$

D ů k a z . Podle definice operátoru L_h a podle lemmatu 3.10 je

$$h^2\sum_{k=1}^{n-1}(L_h\boldsymbol{\eta})_k\eta_k = \tag{3.112}$$

$$= -\sum_{k=1}^{n-1}[p(x_k + h/2)(\eta_{k+1} - \eta_k) - p(x_k - h/2)(\eta_k - \eta_{k-1})]\eta_k +$$

$$+ h^2\sum_{k=1}^{n-1} q(x_k)\eta_k^2 =$$

$$= -p(x_n - h/2)(\eta_n - \eta_{n-1})\eta_{n-1} + p(x_0 + h/2)(\eta_1 - \eta_0)\eta_0 +$$

$$+ \sum_{k=1}^{n-1} p(x_k - h/2)(\eta_k - \eta_{k-1})^2 + h^2\sum_{k=1}^{n-1} q(x_k)\eta_k^2 =$$

$$= -p(x_n - h/2)(\eta_n - \eta_{n-1})\eta_n + p(x_0 + h/2)(\eta_1 - \eta_0)\eta_0 +$$

$$+ \sum_{k=1}^{n} p(x_k - h/2)(\eta_k - \eta_{k-1})^2 + h^2\sum_{k=1}^{n-1} q(x_k)\eta_k^2.$$

Použijeme-li rovnic (3.54) k vyjádření výrazů $p(x_0+h/2)(\eta_1-\eta_0)$ a $p(x_n-h/2)(\eta_n-\eta_{n-1})$ pomocí operátorů $l_h^{(1,\mathrm{d})}$ a $l_h^{(2,\mathrm{d})}$ a dosadíme-li do (3.112), dostaneme

$$h^2\sum_{k=1}^{n-1}(L_h\boldsymbol{\eta})_k\eta_k = -h\frac{1}{\alpha_2}\eta_n l_h^{(2,\mathrm{d})}\boldsymbol{\eta} + \frac{1}{2}h^2 q(x_n)\eta_n^2 + h\frac{\beta_2}{\alpha_2}\eta_n^2 - \tag{3.113}$$

$$- h\frac{1}{\alpha_1}\eta_0 l_h^{(1,\mathrm{d})}\boldsymbol{\eta} + \frac{1}{2}h^2 q(x_0)\eta_0^2 + h\frac{\beta_1}{\alpha_1}\eta_0^2 +$$

$$+ \sum_{k=1}^{n} p(x_k - h/2)(\eta_k - \eta_{k-1})^2 + h^2\sum_{k=1}^{n-1} q(x_k)\eta_k^2.$$

Užijeme-li ve vztahu (3.113) nerovnosti (2.1) a (2.2), dostáváme už tvrzení lemmatu.

Lemma 3.12. *Nechť platí $\alpha_i > 0$ a $\beta_i > 0$ pro $i = 1, 2$. Pak existuje konstanta $\gamma > 0$ taková, že pro libovolný vektor $\boldsymbol{\eta} = (\eta_0, \ldots, \eta_n)^T$ platí*

$$\|\boldsymbol{\eta}\|^2_{\mathcal{L}_\infty} \leqq \frac{1}{h}\gamma\left[h\frac{\beta_1}{\alpha_1}\eta_0^2 + h\frac{\beta_2}{\alpha_2}\eta_n^2 + p_0\sum_{k=1}^{n}(\eta_k - \eta_{k-1})^2\right], \tag{3.114}$$

kde $\|\boldsymbol{\eta}\|_{\mathcal{L}_\infty} = \max_{k=0,\ldots,n}|\eta_k|$.

D ů k a z . Přímým výpočtem se zjistí, že pro $k = 0, \ldots, n$ platí

$$\eta_k = \tfrac{1}{2}\eta_0 + \tfrac{1}{2}\eta_n + \tfrac{1}{2}\sum_{j=1}^{k}(\eta_j - \eta_{j-1}) - \tfrac{1}{2}\sum_{j=k+1}^{n}(\eta_j - \eta_{j-1}) \tag{3.115}$$

(činíme obvyklou konvenci, že součet je roven nule, pokud jeho dolní mez je větší než horní). Užijeme-li nerovnost (3.101), dostaneme odtud, že je

$$\eta_k^2 \leqq \eta_0^2 + \eta_n^2 + \left[\sum_{j=1}^{k}(\eta_j - \eta_{j-1})\right]^2 + \left[\sum_{j=k+1}^{n}(\eta_j - \eta_{j-1})^2\right]^2. \tag{3.116}$$

Podle Schwarzovy nerovnosti platí

$$\left[\sum_{j=1}^{k}(\eta_j - \eta_{j-1})\right]^2 \leqq k\sum_{j=1}^{k}(\eta_j - \eta_{j-1})^2 \leqq k\sum_{j=1}^{n}(\eta_j - \eta_{j-1})^2 \tag{3.117}$$

a

$$\left[\sum_{j=k+1}^{n}(\eta_j-\eta_{j-1})\right]^2 \leqq (n-k)\sum_{j=1}^{n}(\eta_j-\eta_{j-1})^2. \tag{3.118}$$

Dosadíme-li odhady (3.117) a (3.118) do nerovnosti (3.116), dostaneme, že pro $k=0,\dots,n$ platí

$$\eta_k^2 \leqq \frac{1}{h}\left[h\eta_0^2+h\eta_n^2+(b-a)\sum_{j=1}^{n}(\eta_j-\eta_{j-1})^2\right]. \tag{3.119}$$

Abychom dostali tvrzení lemmatu, stačí už jen zavést číslo γ rovnicí (3.107).

Lemma 3.13. *Nechť platí $\alpha_i>0$ a $\beta_i>0$ pro $i=1,2$, nechť koeficienty p a q jsou spojité a nechť jsou splněny nerovnosti* (2.1) a (2.2). *Pak pro libovolný vektor $\eta=(\eta_0,\dots,\eta_n)^T$ platí*

$$\|\eta\|_{\mathcal{L}_\infty} \leqq \gamma\left[\frac{1}{\alpha_1}|l^{(1,\mathrm{d})}\eta|+\frac{1}{\alpha_2}|l^{(2,\mathrm{d})}\eta|+h\sum_{k=1}^{n-1}|(L_h\eta)_k|\right]. \tag{3.120}$$

D ů k a z . Tvrzení lemmatu plyne ihned spojením nerovností (3.114) a (3.111).

Důkaz konvergenční věty pro metodu sítí (3.40), (3.58) je nyní už snadný.

Věta 3.9. *Nechť funkce p má v intervalu $\langle a,b\rangle$ tři spojité derivace a funkce q a f dvě spojité derivace. Nechť dále platí nerovnosti* (2.1) a (2.2) *a nechť je $\alpha_i>0$, $\beta_i>0$ pro $i=1,2$. Nechť konečně y_k je přibližné řešení získané metodou sítí* (3.40), (3.58). *Pak existuje konstanta M taková, že pro každé dostatečně malé h platí*

$$|y_k-y(x_k)| \leqq Mh^2. \tag{3.121}$$

D ů k a z . Za uvedených předpokladů existuje podle vět 3.3 a 3.5 konstanta K taková, že platí

$$\begin{aligned}(L_h y^{(\mathrm{pr})})_k &= \varepsilon_k, \quad k=1,\dots,n-1,\\ l_h^{(1,\mathrm{d})}y^{(\mathrm{pr})} &= \varepsilon_0,\\ l_h^{(2,\mathrm{d})}y^{(\mathrm{pr})} &= \varepsilon_n\end{aligned} \tag{3.122}$$

a

$$|\varepsilon_k| \leqq Kh^2, \quad k=0,\dots,n. \tag{3.123}$$

Položíme-li $\eta=y-y^{(\mathrm{pr})}$ a dosadíme-li do nerovnosti (3.120) z rovnic (3.122), máme

$$\|\eta\|_{\mathcal{L}_\infty} \leqq \gamma\left(\frac{1}{\alpha_1}|\varepsilon_0|+\frac{1}{\alpha_2}|\varepsilon_n|+h\sum_{k=1}^{n-1}|\varepsilon_k|\right) \tag{3.124}$$

Odhadneme-li nyní lokální chyby ε_k pomocí nerovností (3.123), dostáváme, že platí

$$\|\eta\|_{\mathcal{L}_\infty} \leqq \gamma K\left(\frac{1}{\alpha_1}+\frac{1}{\alpha_2}+b-a\right)h^2. \tag{3.125}$$

Protože konstanta γ nezávisí na h, dokazuje nerovnost (3.125) tvrzení věty.

Poznámka 3.2. Vynásobím-li každou z rovnic (3.40) číslem h^2, první rovnici (3.58) číslem h/α_1 a druhou rovnici (3.58) číslem h/α_2, dostaneme soustavu

$$\begin{aligned}h^2(L_hy)_k &= h^2f(x_k), \quad k=1,\dots,n-1,\\ \frac{1}{\alpha_1}hl_h^{(1,\mathrm{d})}y &= \frac{\gamma_1}{\alpha_1}h+\tfrac{1}{2}h^2f(x_0),\\ \frac{1}{\alpha_2}hl_h^{(2,\mathrm{d})}y &= \frac{\gamma_2}{\alpha_2}+\tfrac{1}{2}h^2f(x_n)\end{aligned} \tag{3.126}$$

$n+1$ rovnic o $n+1$ neznámých, která je ekvivalentní soustavě (3.40), (3.58) definující vyšetřované přibližné řešení. Matice této soustavy — označme ji A_h — je přitom symetrická a pozitivně definitní.

Skutečně, pro libovolný vektor $\eta=(\eta_0,\dots,\eta_n)^T$ je

$$\begin{aligned}(A_h\eta,\eta) &= \sum_{k=0}^{n}(A_h\eta)_k\eta_k = \\ &= \left[p(x_0+h/2)+\tfrac{1}{2}h^2q(x_0)+h\frac{\beta_1}{\alpha_1}\right]\eta_0^2-p(x_0+h/2)\eta_0\eta_1+\\ &\quad +h^2\sum_{k=1}^{n-1}(L_h\eta)_k\eta_k+\\ &\quad +\left[p(x_n-h/2)+\tfrac{1}{2}h^2q(x_n)+h\frac{\beta_2}{\alpha_2}\right]\eta_n^2-p(x_n-h/2)\eta_{n-1}\eta_n = \\ &= -p(x_0-h/2)(\eta_1-\eta_0)\eta_0+\left[\tfrac{1}{2}h^2q(x_0)+h\frac{\beta_1}{\alpha_1}\right]\eta_0^2+\\ &\quad +h^2\sum_{k=1}^{n-1}(L_h\eta)_k\eta_k+\\ &\quad +p(x_n-h/2)(\eta_n-\eta_{n-1})\eta_n+\left[\tfrac{1}{2}hq(x_n)+h\frac{\beta_2}{\alpha_2}\right]\eta_n^2 = \\ &= \frac{1}{\alpha_1}h\eta_0l_h^{(1,\mathrm{d})}\eta+\frac{1}{\alpha_2}h\eta_nl_h^{(2,\mathrm{d})}\eta+h^2\sum_{k=1}^{n-1}(L_h\eta)_k\eta_k.\end{aligned} \tag{3.127}$$

Odtud a z lemmatu 3.11 a 3.12 však plyne, že platí

$$(A_h\eta,\eta) \geqq \frac{h}{\gamma}\|\eta\|_{\mathcal{L}_\infty}^2. \tag{3.128}$$

Z poznámky 3.2 plyne jednoznačná řešitelnost soustavy (3.40), (3.58) (nebo (3.126)), aniž je zapotřebí užít monotónie matice A_h.

V právě popsané technice vyšetřování metody sítí (3.40), (3.58) jsme se omezili na typický případ, že je $\alpha_i > 0$ a $\beta_i > 0$ pro $i = 1, 2$. Zmiňme se nyní stručně o modifikacích postupu v případě, že některá z těchto podmínek není splněna. Je-li jedno nebo obě z čísel α_1 a α_2 rovno nule, udává příslušná okrajová podmínka přímo hodnotu hledaného řešení. Při přepisu této okrajové podmínky se tedy nedopustíme žádné chyby a příslušnou složku vektoru chyby lze položit rovnu nule. Nerovnosti (3.111) a (3.114) pak platí s tím, že je v nich třeba vypustit členy obsahující nulové α_1 nebo α_2. V případě, že je jedno z čísel β_1 a β_2 rovno nule, stačí při konstrukci maxima ve vzorci (3.107) příslušný člen vypustit. Je-li konečně $\beta_1 = \beta_2 = 0$, není funkce q identicky rovna nule a při odhadování členů na pravé straně rovnice (3.113) je třeba zachovat nenulovou část součtu $\sum q(x_k)\eta_k^2$.

Zakončeme tento odstavec několika poznámkami o možnostech zlepšení rychlosti konvergence popsané metody sítí. Toho lze podle předešlého výkladu dosáhnout tak, že nalezneme konečnědimenzionální náhradu příslušných diferenciálních výrazů takovou, že její chyba je řádově menší než u těch, které jsme až dosud užili. V literatuře se většinou doporučuje takový postup, že se derivace v dané diferenciální rovnici aproximují s větší přesností, než které je možno dosáhnout námi užitými prostými diferenčními podíly. Tak např. k náhradě druhé derivace lze užít vzorec

$$(3.129)\quad y''(x_k) = \\ = \frac{-y(x_{k-2}) + 16y(x_{k-1}) - 30y(x_k) + 16y(x_{k+1}) - y(x_{k+2})}{12h^2} + O(h^4),$$

který platí pro každou dostatečně hladkou funkci a který vznikl derivováním interpolačního polynomu čtvrtého stupně pro funkci y. Užijeme-li k sestavení diferenčních rovnic vzorce tohoto typu, je na první pohled patrné, že matice vzniklé soustavy bude pětidiagonální, a že tedy řešení této soustavy bude pracnější než v předešlých případech. Kromě toho, každou z takto vzniklých rovnic lze užít pouze pro $k = 2, \dots, n-2$, takže je třeba přidat celkem čtyři další rovnice, zatímco okrajové podmínky dávají přirozenou možnost k sestrojení pouze dvou takových rovnic. Tento problém zřejmě vznikne i v jednoduchém případě, kdy pro koeficienty α_1 a α_2 v okrajových podmínkách platí $\alpha_1 = \alpha_2 = 0$. Konečně je třeba také upozornit, že matice takto vzniklé soustavy ztrácí některé pro její vyšetřování příjemné vlastnosti, např. monotónnost.

Mnohým z těchto problémů se lze vyhnout, vycházíme-li i zde z Marčukovy integrální identity. Tak např. položíme-li

$$(3.130)\quad \int_{x_k}^{x_{k+1}} \frac{1}{p(x)}\,\mathrm{d}\,x \approx \tfrac{1}{6}h\Big[\frac{1}{p(x_k)} + \frac{4}{p(x_k + h/2)} + \frac{1}{p(x_{k+1})}\Big],$$

$$\int_{x_{k-1/2}}^{x_{k+1/2}} [q(x)y(x) - f(x)]\,\mathrm{d}\,x \approx h\{q(x_k)y(x_k) - f(x_k) + \\ + \tfrac{1}{24}[q(x_{k+1})y(x_{k+1}) - 2q(x_k)y(x_k) + q(x_{k-1})y(x_{k-1})] -$$

$$- \tfrac{1}{24}[f(x_{k+1}) - 2f(x_k) + f(x_{k-1})]\},$$

$$\int_{x_k}^{x_{k+1}} \Big\{\frac{1}{p(x)}\int_x^{x_{k+1/2}} [q(t)y(t) - f(t)]\,\mathrm{d}\,t\Big\}\,\mathrm{d}\,x \approx$$

$$\approx \tfrac{1}{48}h^2\Big(\frac{1}{p(x_k)}\{3[q(x_k)y(x_k) - f(x_k)] + [q(x_{k+1})y(x_{k+1}) - f(x_{k+1})]\} -$$

$$- \frac{1}{p(x_{k+1})}\{[q(x_k)y(x_k) - f(x_k)] + 3[q(x_{k+1})y(x_{k+1}) - f(x_{k+1})]\}\Big)$$

a užijeme-li analogické vzorce také pro aproximaci ostatních integrálů v identitě (3.42), dostaneme soustavu

$$(3.131)\quad \Big\{-\tilde{p}_{k-1/2} + \tfrac{1}{24}h^2 q(x_{k-1}) + \\ + \tfrac{1}{48}h^2\tilde{p}_{k-1/2}q(x_{k-1})\Big[\frac{3}{p(x_{k-1})} - \frac{1}{p(x_k)}\Big]\Big\}y_{k-1} + \\ + \Big\{\tilde{p}_{k-1/2} + \tilde{p}_{k+1/2} + \tfrac{11}{12}h^2 q(x_k) + \\ + \tfrac{1}{48}h^2\tilde{p}_{k-1/2}q(x_k)\Big[\frac{1}{p(x_{k-1})} - \frac{3}{p(x_k)}\Big] + \\ + \tfrac{1}{48}h^2\tilde{p}_{k+1/2}q(x_k)\Big[\frac{1}{p(x_{k+1})} - \frac{3}{p(x_k)}\Big]\Big\}y_k + \\ + \Big\{-\tilde{p}_{k+1/2} + \tfrac{1}{24}h^2 q(x_{k+1}) + \\ + \tfrac{1}{48}h^2\tilde{p}_{k+1/2}q(x_{k+1})\Big[\frac{3}{p(x_{k+1})} - \frac{1}{p(x_k)}\Big]\Big\}y_{k+1} = \\ = h^2\Big\{\tfrac{1}{24} + \tfrac{1}{48}\tilde{p}_{k-1/2}\Big[\frac{3}{p(x_{k-1})} - \frac{1}{p(x_k)}\Big]\Big\}f(x_{k-1}) + \\ + h^2\Big\{\tfrac{11}{12} + \tfrac{1}{48}\tilde{p}_{k-1/2}\Big[\frac{1}{p(x_{k-1})} - \frac{3}{p(x_k)}\Big] + \\ + \tfrac{1}{48}\tilde{p}_{k+1/2}\Big[\frac{1}{p(x_{k+1})} - \frac{3}{p(x_k)}\Big]\Big\}f(x_k) + \\ + h^2\Big\{\tfrac{1}{24} + \tfrac{1}{48}\tilde{p}_{k+1/2}\Big[\frac{3}{p(x_{k+1})} - \frac{1}{p(x_k)}\Big]\Big\}p(x_{k+1}),$$

kde

$$(3.132)\quad \tilde{p}_{k-1/2} = \frac{6}{\frac{1}{p(x_{k-1})} + \frac{4}{p(x_k - h/2)} + \frac{1}{p(x_k)}}$$

a analogický vzorec platí pro $\tilde{p}_{k+1/2}$. Přidáme-li k této soustavě (jejíž matice je pro dostatečně malá h monotónní) vhodné okrajové podmínky (např. v případě, že je $\alpha_1 = \alpha_2 = 0$, je to jednoduché, neboť pak známe krajní hodnoty hledaného řešení přesně), dostaneme soustavu rovnic s třídiagonální maticí. Její řešení je tedy

stejně pracné jako dříve, přesnost však je řádu $O(h^4)$. Je však poctivé upozornit, že přepis okrajových podmínek obsahujících derivace tak, aby se vysoká přesnost uvedené metody neznehodnotila, představuje dosti vážný problém.

3.2.3 Řešení vzniklých soustav lineárních rovnic

V tomto odstavci si stručně všimneme problematicky řešení soustav lineárních rovnic, které vznikají při řešení dané okrajové úlohy metodou sítí. Omezíme se opět na případ $\alpha_i > 0$, $\beta_i > 0$ pro $i = 1, 2$ a přepis okrajových podmínek pomocí operátorů $l_h^{(1,d)}$ a $l_h^{(2,d)}$. Řešíme tedy soustavu (3.126) z předešlého odstavce. Zapišme ji ve tvaru

$$A_h y = g. \tag{3.133}$$

Je tedy

$$\begin{aligned} g_0 &= h\frac{\gamma_1}{\alpha_1} + \tfrac{1}{2}h^2 f(x_0), \\ g_k &= h^2 f(x_k), \quad k = 1, \dots, n-1, \\ g_n &= h\frac{\gamma_2}{\alpha_2} + \tfrac{1}{2}h^2 f(x_n), \\ g &= (g_0, \dots, g_n)^T. \end{aligned} \tag{3.134}$$

Matice A_h je symetrická a pozitivně definitní. Pišme ji ve tvaru

$$A_h = \begin{bmatrix} a_0 & b_0 & 0 & \dots & 0 \\ b_0 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_{n-1} \\ 0 & \dots & 0 & b_{n-1} & a_n \end{bmatrix}. \tag{3.135}$$

Pak je

$$\begin{aligned} a_0 &= p(x_0 + h/2) + \tfrac{1}{2}h^2 q(x_0) + \frac{\beta_1}{\alpha_1}h, \\ a_k &= p(x_k - h/2) + p(x_k + h/2) + h^2 q(x_k), \quad k = 1, \dots, n-1, \\ a_n &= p(x_n - h/2) + \tfrac{1}{2}h^2 q(x_n) + \frac{\beta_2}{\alpha_2}h \end{aligned} \tag{3.136}$$

a

$$b_k = -p(x_k + h/2), \quad k = 0, \dots, n-1. \tag{3.137}$$

Řešme nejprve soustavu (3.133) eliminační metodou. Snadno se zjistí přímým

výpočtem, že jsou-li čísla d_k definována rekurencí

$$\begin{aligned} d_0 &= a_0, \\ d_k &= a_k - \frac{b_{k-1}^2}{d_{k-1}}, \quad k = 1, \dots, n, \end{aligned} \tag{3.138}$$

platí

$$A_h = L_h R_h, \tag{3.139}$$

kde

$$L_h = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ \frac{b_0}{d_0} & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{b_{n-1}}{d_{n-1}} & 1 \end{bmatrix} \tag{3.140}$$

a

$$R_h = \begin{bmatrix} d_0 & b_0 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & b_{n-1} \\ 0 & \dots & \dots & 0 & d_n \end{bmatrix} \tag{3.141}$$

Rovnice (3.139) platí samozřejmě pouze za předpokladu, že definice čísel d_k má smysl, tj. za předpokladu, že je $d_k \neq 0$ pro $k = 0, \dots, n-1$. V následující větě ukážeme, že tomu tak skutečně je.

Věta 3.10. *Nechť funkce p, p' a q jsou spojité, nechť platí nerovnosti* (2.1) a (2.2), *nechť je $\alpha_i > 0$, $\beta_i > 0$ pro $i = 1, 2$ a nechť a_k, b_k jsou definována vzorci* (3.136), (3.137). *Pak pro čísla d_k definovaná rekurencí* (3.138) *platí pro dostatečně malá h*

$$d_k \geqq p_0, \quad k = 0, \dots, n-1. \tag{3.142}$$

D ů k a z . Buď φ řešení diferenciální rovnice (2.108) v intervalu $\langle a, b\rangle$ s počáteční podmínkou (2.109). Z věty 2.4 víme, že funkce φ existuje a je jediná a z důkazu této věty dokonce víme, že funkce φ je v intervalu $\langle a, b\rangle$ kladná. Položíme-li dále

$$\Phi(x, \varphi, h) = -\frac{\varphi^2}{p(x + h/2) + h\varphi} + q(x + h), \tag{3.143}$$

je obecná jednokroková metoda daná funkcí Φ konzistentní s diferenciální rovnicí

(2.108) (srv. kap. I, odst. 3.2). Pro posloupnost φ_k definovanou rovnicemi

$$\varphi_0 = \frac{\beta_1}{\alpha_1} + \tfrac{1}{2}hq(x_0), \tag{3.144}$$
$$\varphi_k = \varphi_{k-1} + h\Phi(x_{k-1}, \varphi_{k-1}, h), \quad k = 1, \dots, n,$$

tedy platí podle věty 3.1 z kap. I

$$\varphi_k - \varphi(x_k) \to 0 \quad \text{pro } h \to 0,\ x_k = a + kh = \text{konst.} \tag{3.145}$$

Přesně vzato, toto tvrzení ze zmíněné věty přímo neplyne, neboť v ní se předpokládá, že počáteční podmínka přibližného řešení je rovna přesné počáteční podmínce. Zde tomu tak není, neboť počáteční podmínka pro přibližné řešení k přesné počáteční podmínce pouze konverguje pro $h \to 0$. Je však naprosto zřejmé, jak modifikovat větu 3.1 z kap. I, aby pokrývala i tento případ.

Ze vztahu (3.145) plyne, že pro všechna dostatečně malá h jsou čísla φ_k nezáporná (funkce φ je kladná). Položíme-li tedy

$$\tilde{d}_k = p(x_k + h/2) + h\varphi_k, \quad k = 0, \dots, n-1, \tag{3.146}$$

platí pro $k = 0, \dots, n-1$

$$\tilde{d}_k \geqq p_0. \tag{3.147}$$

Z rekurence (3.144) dostáváme ihned, že je

$$\tilde{d}_0 = p(x_0 + h/2) + \tfrac{1}{2}h^2 q(x_0) + h\frac{\beta_1}{\alpha_1} = a_0 \tag{3.148}$$

a

(3.149)

$$\begin{aligned}
\tilde{d}_k &= p(x_k + h/2) + h\varphi_k = \\
&= p(x_k + h/2) + h\varphi_{k-1} - h^2\frac{\varphi_{k-1}^2}{p(x_k - h/2) + h\varphi_{k-1}} + h^2 q(x_k) = \\
&= p(x_k - h/2) + p(x_k + h/2) + h^2 q(x_k) + h\varphi_{k-1} - p(x_k - h/2) - \\
&\quad - h^2\frac{\varphi_{k-1}^2}{p(x_k - h/2) + h\varphi_{k-1}} = \\
&= a_k + \\
&\quad + \frac{h\varphi_{k-1}p(x_k - \frac{h}{2}) + h^2\varphi_{k-1}^2 - p^2(x_k - \frac{h}{2}) - hp(x_k - \frac{h}{2})\varphi_{k-1} - h^2\varphi_{k-1}^2}{p(x_k - h/2) + h\varphi_{k-1}} = \\
&= a_k - \frac{b_{k-1}^2}{\tilde{d}_{k-1}}
\end{aligned}$$

pro $k = 0, \dots, n-1$. Platí tedy $\tilde{d}_k = d_k$ pro $k = 0, \dots, n-1$ a tvrzení věty plyne z nerovnosti (3.147).

Jsou-li tedy splněny předpoklady věty 3.10, rozklad (3.139) skutečně existuje a protože je rovněž $d_n > 0$, jak plyne ihned přímo z rovnic (3.138) (nebo také, chceme-li, ze skutečnosti, že matice A_h je regulární), lze řešit soustavu (3.133) tak, že nejprve vypočteme vektor $c = (c_0, \dots, c_n)^T$ ze soustavy

$$L_h c = g \tag{3.150}$$

a pak hledaný vektor $y = (y_0, \dots, y_n)^T$ ze soustavy

$$R_h y = c. \tag{3.151}$$

Vzhledem k speciálnímu tvaru matic L_h a R_h zřejmě platí

$$c_0 = g_0, \tag{3.152}$$
$$c_k = g_k - \frac{b_{k-1}c_{k-1}}{d_{k-1}}, \quad k = 1, \dots, n,$$

a

$$y_n = \frac{c_n}{d_n}, \tag{3.153}$$
$$y_k = \frac{1}{d_k}(c_k - b_k y_{k+1}), \quad k = n-1, \dots, 0.$$

Soustavu (3.133) lze tedy řešit eliminační metodou bez výběru hlavního prvku a znamená to vypočítat veličiny d_k a c_k z rekurencí (3.138) a (3.152) a hledané řešení z rekurence (3.153).

Pokud jde o jiné metody pro řešení soustavy (3.133), dá se jistě užít Gaussova-Seidelova metoda nebo optimalizovaná superrelaxační metoda, neboť matice A_h je pozitivně definitní. Všeobecně však nelze užití iteračních metod v tomto případě doporučit. Z rekurencí (3.138), (3.152) a (3.153) je totiž ihned vidět, že k provedení eliminační metody je třeba řádově pouze $O(n)$ operací, tj. řádově pouze tolik operací, kolik je neznámých. Žádná iterační metoda nemůže tedy eliminační metodě co do výpočetní ekonomie konkurovat.

Všimněme si ještě jedné zajímavé okolnosti spojené s eliminační metodou. Z důkazu věty 3.10 je vidět, že definujeme-li pro $k = 0, \dots, n-1$ posloupnost φ_k rovnicí $d_k = p(x_k + h/2) + h\varphi_k$, kde posloupnost d_k je definovaná rekurencí (3.138), představuje tato posloupnost přibližné řešení rovnice (2.108) z metody normalizovaného přesunu. Tak je tomu i pro ostatní rekurence, podle nichž se v eliminační metodě počítá. Skutečně, definujeme-li posloupnost v_k rovnicí $c_k = hv_k$, kde c_k jsou dány rekurencí (3.152), snadno vypočteme, že platí

$$v_0 = \frac{\gamma_1}{\alpha_1} + \tfrac{1}{2}hf(x_0), \tag{3.154}$$
$$v_k = v_{k-1} + h\Big[-\frac{\varphi_{k-1}v_{k-1}}{p(x_k - h/2) + h\varphi_{k-1}} + f(x_k)\Big],$$

takže posloupnost v_k aproximuje řešení diferenciální rovnice (2.110) s počáteční podmínkou (2.111). Vyjádříme-li konečně y_k definované rekurencí (3.153) pomocí

veličin φ_k a v_k, dostaneme

$$y_n = \frac{\gamma_2 + \alpha_2 v_{n-1}}{\beta_2 + \alpha_2 \varphi_{n-1}} + O(h), \tag{3.155}$$

$$y_k = y_{k+1} - h\frac{\varphi_k y_{k+1} - v_k}{p(x_k + h/2) + h\varphi_k},$$

takže posloupnost y_k představuje přibližné řešení rovnice (2.107) s počáteční podmínkou (2.128).

Právě zjištěná souvislost eliminační metody s metodou normalizovaného přesunu opravňuje domnívat se, že i eliminační metoda se chová stejně jako metoda normalizovaného přesunu numericky uspokojivě. Praktické zkušenosti to také plně potvrzují.

Zakončeme celý odstavec věnovaný problematice metody sítí pro rovnice druhého řádu dvěma jednoduchými příklady.

Tabulka 3.1

Řešení okrajové úlohy (3.156), (3.157) metodou sítí

x	Užitá metoda			Přesné řešení
	(3.40) $h = 1/20$	(3.40) $h = 1/40$	(3.131) $h = 1/20$	
0,10	1,005 181 88	1,005 061 31	1,005 020 25	1,005 020 91
0,20	1,020 651 27	1,020 417 30	1,020 337 52	1,020 338 83
0,30	1,047 205 57	1,046 865 62	1,046 749 61	1,046 751 59
0,40	1,086 287 94	1,085 850 99	1,085 701 74	1,085 704 42
0,50	1,140 190 24	1,139 668 80	1,139 490 55	1,139 493 91
0,60	1,212 411 16	1,211 824 88	1,211 624 28	1,211 620 30
0,70	1,308 283 55	1,307 666 27	1,307 454 66	1,307 459 25
0,80	1,436 106 36	1,435 520 74	1,435 319 38	1,435 324 20
0,90	1,609 298 85	1,608 869 91	1,608 721 01	1,608 725 79

Příklad 3.1. Řešme v intervalu $\langle 0, 1\rangle$ diferenciální rovnici

$$-y'' + (1 + 2\operatorname{tg}^2 x)y = 0 \tag{3.156}$$

s okrajovými podmínkami

$$y(0) = 1, \quad y(1) = \frac{1}{\cos 1} \tag{3.157}$$

a s přesným řešením $y(x) = 1/\cos x$. V tab. 3.1 jsou uvedeny hodnoty přibližného řešení získaného pomocí rovnic (3.40) s $h = 1/20$ a s $h = 1/40$ a hodnoty přibližného řešení získaného pomocí rovnic (3.131) s $h = 1/20$ spolu s hodnotami přesného

řešení. Z tabulky je vidět, že podstatně méně pracná metoda (3.131) s $h = 1/20$ dává skutečně dosti lepší výsledky než metoda (3.40) s $h = 1/40$.

Příklad 3.2. Řešme diferenciální rovnici (3.156) s okrajovými podmínkami

$$y'(0) = 0, \quad y'(1) = \frac{\operatorname{tg} 1}{\cos 1}. \tag{3.158}$$

Přesné řešení této úlohy je opět funkce $y(x) = 1/\cos x$. Tab. 3.2 uvádí příslušné výsledky získané ze soustavy (3.40), (3.41) a (3.40), (3.58). Vidíme z ní, že nevhodný přepis okrajových podmínek se může projevit dosti dramaticky.

Tabulka 3.2

Řešení okrajové úlohy (3.156), (3.158) metodou sítí

x	Užitá metoda		Přesné řešení
	(3.40), (3.41) $h = 1/40$	(3.40), (3.58) $h = 1/40$	
0,00	1,060 536 63	0,999 118 25	1,000 000 00
0,10	1,064 532 36	1,004 133 47	1,005 020 91
0,20	1,079 431 78	1,019 433 84	1,020 338 83
0,30	1,106 064 84	1,045 815 84	1,046 751 59
0,40	1,145 943 59	1,084 722 28	1,085 704 42
0,50	1,201 475 20	1,138 445 72	1,139 493 91
0,60	1,276 339 03	1,210 487 76	1,211 620 30
0,70	1,376 148 32	1,306 188 92	1,307 459 25
0,80	1,509 646 22	1,433 866 91	1,435 324 20
0,90	1,690 986 79	1,606 987 15	1,608 725 79
1,00	1,944 423 37	1,848 623 70	1,850 815 72

3.3 Lineární diferenciální rovnice čtvrtého řádu

V tomto odstavci si všimneme užití metody sítí k řešení okrajové úlohy (1.14) s okrajovými podmínkami (1.15). Začneme tím, že vyslovíme a dokážeme základní existenční větu.

Věta 3.11. *Nechť funkce p je dvakrát spojitě diferencovatelná v intervalu $\langle a, b\rangle$, nechť funkce q a f jsou spojité v $\langle a, b\rangle$ a nechť platí*

$$p(x) \geqq p_0 > 0, \quad q(x) \geqq 0, \quad x \in \langle a, b\rangle, \tag{3.159}$$

kde p_0 je konstanta. Pak existuje právě jedno řešení úlohy (1.14), (1.15).

D ů k a z . Podle poznámky 2.2 na str. 140 stačí ukázat, že příslušná homogenní úloha má pouze triviální řešení. Buď tedy η řešení diferenciální rovnice

$$(\mathcal{L}\eta)(x) \equiv \big(p(x)\eta''\big)'' + q(x)\eta = 0, \quad x \in \langle a, b\rangle \tag{3.160}$$

s okrajovými podmínkami

$$\eta(a) = \eta'(a) = \eta(b) = \eta'(b) = 0. \tag{3.161}$$

Máme dokázat, že funkce η je rovna v intervalu $\langle a, b\rangle$ identicky nule. Vynásobme rovnici (3.160) funkcí η a integrujme v mezích od a do b. Dvojí integrací per partes a užitím okrajových podmínek (3.161) dostaneme

$$\begin{aligned}\int_a^b \big\{[p(x)\eta''(x)]'' + q(x)\eta(x)\big\}\eta(x)\,\mathrm{d}\,x &= \\ = \int_a^b p(x)[\eta''(x)]^2\,\mathrm{d}\,x + \int_a^b q(x)\eta^2(x)\,\mathrm{d}\,x.\end{aligned} \tag{3.162}$$

Použijeme-li nerovnosti (3.159), máme

$$\int_a^b \big\{[p(x)\eta''(x)]'' + q(x)\eta(x)\big\}\eta(x)\,\mathrm{d}\,x \geqq p_0 \int_a^b [\eta''(x)]^2\,\mathrm{d}\,x. \tag{3.163}$$

Na druhé straně vzhledem k okrajovým podmínkám (3.161) zřejmě platí

$$\eta(x) = \int_a^x \left(\int_a^t \eta''(s)\,\mathrm{d}\,s\right)\mathrm{d}\,t. \tag{3.164}$$

Zaměníme-li ve vzorci (3.164) pořadí integrace, dostaneme

$$\eta(x) = \int_a^x (x-s)\eta''(s)\,\mathrm{d}\,s. \tag{3.165}$$

Odtud užitím Schwarzovy nerovnosti vypočteme, že platí

$$\eta^2(x) \leqq \int_a^x (x-s)^2\,\mathrm{d}\,s \int_a^x [\eta''(s)]^2\,\mathrm{d}\,s \leqq \tfrac{1}{3}(b-a)^3 \int_a^b [\eta''(s)]^2\,\mathrm{d}\,s. \tag{3.166}$$

Dosadíme-li z této nerovnosti do nerovnosti (3.163), dostáváme celkem, že pro každé $x \in \langle a, b\rangle$ platí

$$\eta^2(x) \leqq \tfrac{1}{3}(b-a)^3 \frac{1}{p_0} \int_a^b \big\{[p(x)\eta''(x)]'' + q(x)\eta(x)\big\}\eta(x)\,\mathrm{d}\,x. \tag{3.167}$$

Pravá strana této nerovnosti je však nulová; je tedy nulová i levá strana a $\eta(x) \equiv 0$ v $\langle a, b\rangle$. Věta je dokázána.

3.3.1 Sestavení diferenčních rovnic

Položíme-li $p(x)y''(x) = z(x)$ a aproximujeme-li derivaci $z''(x)$ v diferenciální rovnici (1.14) podílem $[z(x+h) - 2z(x) + z(x-h)]/h^2$ a chybu, které se přitom dopustíme, zanedbáme, jsme přirozeně přivedeni k zavedení operátoru L_h (užití téhož symbolu jako v předešlém odstavci nepovede k nedorozumění), který $(n+1)$-dimenzionálnímu vektoru $y = (y_0, \ldots, y_n)^T$ přiřazuje $(n-3)$-dimenzionální vektor $L_h y$ předpisem

$$\begin{aligned}h^4(L_h y)_k = {} & p(x_{k-1})y_{k-2} - 2[p(x_{k-1}) + p(x_k)]y_{k-1} + \\ & + [p(x_{k-1}) + 4p(x_k) + p(x_{k+1}) + h^4 q(x_k)]y_k - \\ & - 2[p(x_k) + p(x_{k+1})]y_{k+1} + \\ & + p(x_{k+1})y_{k+2}, \quad k = 2, \ldots, n-2.\end{aligned} \tag{3.168}$$

Tento operátor aproximuje operátor L na levé straně diferenciální rovnice (1.14) ve smyslu následující věty.

Věta 3.12. *Nechť funkce p má v intervalu $\langle a, b\rangle$ čtyři spojité derivace a nechť q je spojitá. Pak pro každou funkci y, která má v intervalu $\langle a, b\rangle$ šest spojitých derivací, platí*

$$(L_h y^{(\mathrm{pr})})_k = (Ly)(x_k) + O(h^2), \quad k = 2, \ldots, n-2, \tag{3.169}$$

kde $y^{(\mathrm{pr})} = \big(y(x_0), \ldots, y(x_n)\big)^T$.

D ů k a z . Buď $x \in \langle a, b\rangle$ a položme $z(x) = p(x)y''(x)$. Funkce z má v intervalu $\langle a, b\rangle$ čtyři spojité derivace a podle Taylorova vzorce (nebo podle lemmatu 3.8) platí

$$\frac{z(x+h) - 2z(x) + z(x-h)}{h^2} = z''(x) + O(h^2). \tag{3.170}$$

Užijeme-li znovu Taylorův vzorec, máme

$$\begin{aligned}\frac{y(x+2h) - 2y(x+h) + y(x)}{h^2} &= y''(x+h) + \tfrac{1}{12}y''''(x+h)h^2 + O(h^4), \\ \frac{y(x+h) - 2y(x) + y(x-h)}{h^2} &= y''(x) + \tfrac{1}{12}y''''(x)h^2 + O(h^4), \\ \frac{y(x) - 2y(x-h) + y(x-2h)}{h^2} &= y''(x-h) + \tfrac{1}{12}y''''(x-h)h^2 + O(h^4),\end{aligned} \tag{3.171}$$

neboť funkce y má podle předpokladu šest spojitých derivací. Vynásobíme-li první rovnici v (3.171) číslem $p(x+h)/h^2$, druhou číslem $-2p(x)/h^2$, třetí číslem $p(x-h)/h^2$ a sečteme, dostaneme

$$\begin{aligned}\frac{1}{h^4}\big\{ & p(x+h)[y(x+2h) - 2y(x+h) + y(x)] - \\ & - 2p(x)[y(x+h) - 2y(x) + y(x-h)] + \\ & + p(x-h)[y(x) - 2y(x-h) + y(x-2h)]\big\} =\end{aligned} \tag{3.172}$$

$$= \frac{z(x+h) - 2z(x) + z(x-h)}{h^2} +$$
$$+ \tfrac{1}{12}[p(x+h)y''''(x+h) - 2p(x)y''''(x) + p(x-h)y''''(x-h)] +$$
$$+ O(h^2).$$

Výraz v hranaté závorce na pravé straně rovnice (3.172) je však velikosti $O(h^2)$, neboť funkce $p(x)y''''(x)$ má v intervalu $\langle a, b\rangle$ dvě spojité derivace. Tvrzení věty tedy plyne z rovnic (3.172) a (3.170).

Na základě této věty je přirozené hledat přibližné řešení $y_0, \dots, y_n$ dané okrajové úlohy z rovnic

$$(L_h y)_k = f(x_k), \quad k = 2, \dots, n-2. \tag{3.173}$$

Počet těchto rovnic je $n-3$, zatímco neznámých je $n+1$. Je tedy třeba připojit k nim ještě čtyři rovnice, které se získají z okrajových podmínek. Dvě z nich dostaneme ihned tak, že položíme $y_0 = \gamma_1$ a $y_n = \gamma_2$. Další dvě rovnice lze získat snadno tak, že ve zbývajících okrajových podmínkách se nahradí derivace prostými diferenčními podíly (srv. odvození rovnic (3.27) v odst. 3.2.1). Tím se však dopustíme chyby řádu $O(h)$, zatímco v rovnicích (3.173) děláme chybu řádu $O(h^2)$. Proto použijeme postup, který se v podobných případech často doporučuje (a který jsme už vlastně mohli užít i v odst. 3.2.1). Derivace v příslušných okrajových podmínkách nahradíme výrazy $(y_1 - y_{-1})/(2h)$, resp. $(y_{n+1} - y_{n-1})/(2h)$, kde y_{-1}, resp. y_{n+1} značí aproximaci prodloužení hledaného řešení do bodů $x_{-1} = x_0 - h$, resp. $x_{n+1} = x_n + h$, neboť tyto podíly aproximují odpovídající derivace s přesností $O(h^2)$. Protože však hodnoty y_{-1} a y_{n+1} se v nich vyskytují oproti rovnicím (3.173) navíc, je třeba požadovat platnost těchto rovnic i pro $k = 1$ a $k = n-1$. Přidané hodnoty y_{-1} a y_{n+1} pak můžeme vyloučit, takže zůstanou zachovány původní neznámé $y_0, \dots, y_n$. Výsledky naznačeného postupu jsou shrnuty v následující větě.

Věta 3.13. *Nechť funkce p, p' a q jsou spojité v nějakém okolí bodu a a b. Pak pro každou funkci y, která má v okolí bodu a a b tři spojité derivace, platí*

$$l_h^{(1)} y^{(\mathrm{pr})} = 2hp(a)y'(a) + O(h^3),$$
$$l_h^{(2)} y^{(\mathrm{pr})} = -2hp(b)y'(b) + O(h^3), \tag{3.174}$$

kde $y^{(\mathrm{pr})} = [y(x_0), \dots, y(x_n)]^T$ a operátory $l_h^{(1)}$ a $l_h^{(2)}$ jsou definovány předpisem

$$\begin{aligned} l_h^{(1)} y^{(\mathrm{pr})} &= -2[p(x_0) + p(x_1)]y(x_0) + \\ &+ [2p(x_0) + 4p(x_1) + p(x_2) + h^4 q(x_1)]y(x_1) - \\ &- 2[p(x_1) + p(x_2)]y(x_2) + p(x_2)y(x_3), \\ l_h^{(2)} y^{(\mathrm{pr})} &= p(x_{n-2})y(x_{n-3}) - 2[p(x_{n-2}) + p(x_{n-1})]y(x_{n-2}) + \\ &+ [p(x_{n-2}) + 4p(x_{n-1}) + 2p(x_n) + h^4 q(x_{n-1})]y(x_{n-1}) - \\ &- 2[p(x_{n-1}) + p(x_n)]y(x_n). \end{aligned} \tag{3.175}$$

D ů k a z . Za uvedených předpokladů zřejmě platí

$$y(x_1) - y(x_{-1}) = 2hy'(x_0) + O(h^3) \tag{3.176}$$

(srv. lemma 3.7). Rovněž snadno se zjistí, že je

$$\begin{aligned} &p(x_0)y(x_{-1}) - 2[p(x_0) + p(x_1)]y(x_0) + [p(x_0) + 4p(x_1) + p(x_2)]y(x_1) - \\ &\quad - 2[p(x_1) + p(x_2)]y(x_2) + p(x_2)y(x_3) + h^4 q(x_1)y(x_1) = \\ &= p(x_2)[y(x_3) - 2y(x_2) + y(x_1)] - 2p(x_1)[y(x_2) - 2y(x_1) + y(x_0)] = \\ &\quad + p(x_0)[y(x_1) - 2y(x_0) + y(x_{-1})] + h^4 q(x_1)y(x_1) = \\ &= h^2[p(x_2)y''(x_2) - 2p(x_1)y''(x_1) + p(x_0)y''(x_0)] + O(h^4) = O(h^3), \end{aligned} \tag{3.177}$$

neboť funkce y má v okolí bodu $x = x_0 = a$ tři spojité derivace, a tedy funkce py'' má v okolí téhož bodu spojitou derivaci. Vyloučíme-li $y(x_{-1})$ z rovnic (3.176) a (3.177), dostaneme první rovnici (3.174). Druhá rovnice (3.174) se dokáže úplně stejně. Důkaz věty 3.13 je hotov.

Přibližné řešení okrajové úlohy (1.14), (1.15) budeme tedy počítat ze soustavy (3.173) doplněné rovnicemi

$$y_0 = \gamma_1, \quad y_n = \gamma_2,$$
$$l_h^{(1)} y = 2hp(x_0)\delta_1, \quad l_h^{(2)} y = -2hp(x_n)\delta_2. \tag{3.178}$$

Vzhledem k speciálnímu tvaru prvních dvou rovnic (3.178) jsou neznámé vlastně pouze složky $y_1, \dots, y_{n-1}$. Značí-li nyní y vektor o složkách $y_1, \dots, y_{n-1}$, je třeba jej vypočíst ze soustavy

$$A_h y = g, \tag{3.179}$$

kde A_h je pětidiagonální symetrická matice řádu $n-1$ a g je daný $(n-1)$-dimenzionální vektor. Položíme-li ještě $A_h = \{a_{ij}\}$ a $g = (g_1, \dots, g_{n-1})^T$, je

$$\begin{aligned} a_{11} &= 2p(x_0) + 4p(x_1) + p(x_2) + h^4 q(x_1), \\ a_{kk} &= p(x_{k-1}) + 4p(x_k) + p(x_{k+1}) + h^4 q(x_k), \quad k = 2, \dots, n-2, \\ a_{n-1,n-1} &= p(x_{n-2}) + 4p(x_{n-1}) + 2p(x_n) + h^4 q(x_{n-1}), \\ a_{k,k+1} &= -2[p(x_k) + p(x_{k+1})], \quad k = 1, \dots, n-2, \\ a_{k,k+2} &= p(x_{k+1}), \quad k = 1, \dots, n-3, \end{aligned} \tag{3.180}$$

a

$$\begin{aligned} q_1 &= 2[p(x_0) + p(x_1)]\gamma_1 + 2hp(x_0)\delta_1, \\ q_2 &= -p(x_1)\gamma_1 + h^4 f(x_2), \\ q_k &= h^4 f(x_k), \quad k = 3, \dots, n-3, \\ q_{n-2} &= -p(x_{n-1})\gamma_2 + h^4 f(x_{n-2}), \\ q_{n-1} &= 2[p(x_{n-1}) + p(x_n)]\gamma_2 - 2hp(x_n)\delta_2, \end{aligned} \tag{3.181}$$

jak plyne ihned z rovnic (3.173) a (3.178).

Chceme-li, aby právě sestrojená metoda sítí měla rozumný smysl, je třeba především ukázat, že soustava (3.179) má právě jedno řešení. To bude snadným důsledkem lemmatu, které představuje paralelu z lemmatu 3.13 a které bude hrát klíčovou roli v důkazu konvergence popsané metody. Proto úvahy o regularitě matice A_h odsuneme až do následujícího odstavce.

3.3.2 Konvergence

Základem všech úvah, které budou prováděny v tomto odstavci, je následující lemma.

Lemma 3.14. *Nechť funkce p a q jsou spojité v $\langle a, b\rangle$ a nechť platí nerovnosti* (3.159). *Pak existuje konstanta $\gamma > 0$ taková, že pro každý $(n-1)$-dimenzionální vektor $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{n-1})^T$ platí*

$$(A_h\boldsymbol{\eta}, \boldsymbol{\eta}) \geqq \gamma h^3 \|\boldsymbol{\eta}\|_{h,\mathscr{L}_\infty}^2 \tag{3.182}$$

kde

$$\|\boldsymbol{\eta}\|_{h,\mathscr{L}_\infty} = \max\left(|\eta_1| h^{-3/2}, \max_{k=2,\dots,n-2} |\eta_k|, |\eta_{n-1}| h^{-3/2}\right). \tag{3.183}$$

Důkaz. Buď $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{n-1})^T$ libovolný $(n-1)$-dimenzionální vektor. Položíme-li $\eta_0 = \eta_n = 0$, $\eta_{-1} = \eta_1$ a $\eta_{n+1} = \eta_{n-1}$, plyne z rovnic (3.180), že je

$$(A_h\boldsymbol{\eta}, \boldsymbol{\eta}) = h^4 \sum_{k=1}^{n-1} (L_h\boldsymbol{\eta})_k \eta_k. \tag{3.184}$$

Zavedeme-li ještě čísla z_k rovnicemi

$$z_k = \eta_{k+1} - 2\eta_k + \eta_{k-1}, \quad k = 0, \dots, n, \tag{3.185}$$

lze rovnici (3.184) psát ve tvaru

$$(A_h\boldsymbol{\eta}, \boldsymbol{\eta}) = S_n + h^4 \sum_{k=1}^{n-1} q(x_k)\eta_k^2, \tag{3.186}$$

kde

$$S_n = \sum_{k=1}^{n-1} \{[p(x_{k+1})z_{k+1} - p(x_k)z_k] - [p(x_k)z_k - p(x_{k-1})z_{k-1}]\}\eta_k. \tag{3.187}$$

K úpravě posledního součtu užijeme dvakrát lemma 3.10; dostaneme

$$\begin{aligned} S_n &= [p(x_n)z_n - p(x_{n-1})z_{n-1}]\eta_{n-1} - \\ &\quad - \sum_{k=1}^{n-1} [p(x_k)z_k - p(x_{k-1})z_{k-1}](\eta_k - \eta_{k-1}) = \\ &= p(x_n)z_n\eta_{n-1} - p(x_{n-1})z_{n-1}\eta_{n-1} - \\ &\quad - p(x_{n-1})z_{n-1}(\eta_{n-1} - \eta_{n-2}) - p(x_0)z_0\eta_1 + \sum_{k=1}^{n-1} p(x_{k-1})z_{k-1}^2 = \\ &= p(x_n)z_n\eta_{n-1} - p(x_{n-1})z_{n-1}(2\eta_{n-1} - \eta_{n-2}) - p(x_0)z_0\eta_1 + \\ &\quad + \sum_{k=1}^{n-1} p(x_k)z_k^2 + p(x_0)z_0^2 - p(x_{n-1})z_{n-1}^2. \end{aligned} \tag{3.188}$$

Protože je však $2\eta_{n-1} - \eta_{n-2} = -z_{n-1}$ a

$$\eta_1 = \tfrac{1}{2}z_0, \quad \eta_{n-1} = \tfrac{1}{2}z_n, \tag{3.189}$$

jak plyne ihned z definice čísel z_k, dostáváme celkem, že platí

$$S_n = \sum_{k=1}^{n-1} p(x_n)z_k^2 + \tfrac{1}{2}p(x_0)z_0^2 + \tfrac{1}{2}p(x_n)z_n^2. \tag{3.190}$$

Dosadíme-li tento výsledek do rovnice (3.186) a použijeme-li nerovnosti (3.159), máme

$$(A_h\boldsymbol{\eta}, \boldsymbol{\eta}) \geqq p_0\left(\sum_{k=1}^{n-1} z_k^2 + \tfrac{1}{2}z_0 + \tfrac{1}{2}z_n^2\right). \tag{3.191}$$

Odtud a z rovnic (3.189) plyne speciálně, že je

$$(A_h\boldsymbol{\eta}, \boldsymbol{\eta}) \geqq 2p_0\eta_1^2, \quad (A_h\boldsymbol{\eta}, \boldsymbol{\eta}) \geqq 2p_0\eta_{n-1}^2; \tag{3.192}$$

je tedy

$$(A_h\boldsymbol{\eta}, \boldsymbol{\eta}) \geqq 2p_0h^3\left[\max\left(|\eta_1|h^{-3/2}, |\eta_{n-1}|h^{-3/2}\right)\right]^2. \tag{3.193}$$

Píšeme-li rovnice (3.185) ve tvaru $z_k = (\eta_{k+1} - \eta_k) - (\eta_k - \eta_{k-1})$ a sečteme je pro $k = 1, \dots, \nu$, dostaneme, že je

$$\eta_{\nu+1} - \eta_\nu - \eta_1 = \sum_{k=1}^{\nu} z_k. \tag{3.194}$$

Sečtením rovnic (3.194) s $\nu = 1, \dots, j-1$ dostáváme dále, že platí

$$\eta_j - \eta_1 - (j-1)\eta_1 = \sum_{\nu=1}^{j-1} z_k \tag{3.195}$$

pro $j = 2, 3, \ldots, n$ neboli, zaměníme-li pořádek sčítání a dosadíme-li za η_1 podle rovnice (3.189),

$$\eta_j = \tfrac{1}{2} j z_0 + \sum_{k=1}^{j-1} (j-k) z_k. \tag{3.196}$$

Použijeme-li nyní nerovnost (3.101) s Schwarzovu nerovnost, plyne z rovnice (3.196), že pro $j = 2, \ldots, n$ platí

$$\eta_j^2 \leqq \tfrac{1}{2} j^2 z_0^2 + 2 \sum_{k=1}^{j-1} (j-k)^2 \sum_{k=1}^{j-1} z_k^2 \leqq n^3 z_0^2 + 2n^3 \sum_{k=1}^{n-1} z_k^2 + n^3 z_n^2. \tag{3.197}$$

Z nerovností (3.197) a (3.191) však už bezprostředně plyne, že je

$$\eta_j^2 \leqq \frac{2(b-a)^3}{h^3} \frac{1}{p_0} (A_h \boldsymbol{\eta}, \boldsymbol{\eta}) \tag{3.198}$$

pro $j = 2, \ldots, n$, neboť je $n = (b-a)/h$. Je tedy také

$$\left(\max_{n=2,\ldots,n-2} |\eta_j| \right)^2 \leqq \frac{2(b-a)^3}{h^3} \frac{1}{p_0} (A_h \boldsymbol{\eta}, \boldsymbol{\eta}). \tag{3.199}$$

Položíme-li konečně

$$\gamma = p_0 \min \left[2, \frac{1}{2(b-a)^3} \right], \tag{3.200}$$

plyne z tvrzení lemmatu už snadno z nerovností (3.199) a (3.192).

Z lemmatu 3.14 plyne bezprostředně, že jsou-li funkce p a q spojité a platí-li nerovnosti (3.159), je matice A_h pozitivně definitní, a soustava (3.179) má tedy při libovolné pravé straně právě jedno řešení. Skoro stejně snadno umožňuje lemma 3.14 dát odpověď i na otázku po konvergenci uvažované metody sítí.

Věta 3.14. *Nechť funkce p má čtyři spojité derivace funkce q a f dvě spojité derivace v intervalu $\langle a, b \rangle$ a nechť platí nerovnosti* (3.159). *Pak existuje konstanta M taková, že platí*

$$\|y - y^{(\mathrm{pr})}\|_{h, \mathscr{L}_\infty} \leqq M h^{3/2}, \tag{3.201}$$

kde $y = (y_1, \ldots, y_{n-1})^T$ je řešení soustavy (3.179) *a $y^{(\mathrm{pr})}$ je vektor, jehož složky jsou hodnoty přesného řešení y okrajové úlohy* (1.14), (1.15) *v bodech x_k.*

D ů k a z . Za uvedených předpokladů řešení y okrajové úlohy (1.14), (1.15) nejen existuje a je jediné, ale má mimo to v intervalu $\langle a, b \rangle$ šest spojitých derivací. Položíme-li

$$A_h y^{(\mathrm{pr})} - g = \boldsymbol{\epsilon}, \tag{3.202}$$

existuje podle vět 3.12 a 3.13 konstanta K taková, že pro složky ε_k vektoru $\boldsymbol{\epsilon}$ platí odhady

$$|\varepsilon_1| \leqq K h^3, \quad |\varepsilon_k| \leqq K h^6, \quad k = 2, \ldots, n-2, \quad |\varepsilon_{n-1}| \leqq K h^3. \tag{3.203}$$

Pro vektor $\boldsymbol{\eta} = y^{(\mathrm{pr})} - y$ tedy máme

$$A_h \boldsymbol{\eta} = \boldsymbol{\epsilon}. \tag{3.204}$$

Podle lemmatu 3.14 platí pro libovolný vektor $\boldsymbol{\eta}$ nerovnost (3.182). Dosadíme-li do této nerovnosti speciálně podle (3.204) a použijeme-li nerovnosti (3.203), snadno odvodíme, že platí

$$\begin{aligned} \gamma h^3 \|\boldsymbol{\eta}\|_{h,\mathscr{L}_\infty}^2 &\leqq (\boldsymbol{\epsilon}, \boldsymbol{\eta}) \leqq \sum_{k=1}^{n-1} |\varepsilon_k| |\eta_k| \leqq \\ &\leqq |\varepsilon_1| h^{3/2} |\eta_1| h^{-3/2} + \\ &\quad + \sum_{k=2}^{n-2} |\varepsilon_k| |\eta_k| + |\varepsilon_{n-1}| h^{3/2} |\eta_{n-1}| h^{-3/2} \leqq \\ &\leqq \|\boldsymbol{\eta}\|_{h,\mathscr{L}_\infty} \left(|\varepsilon_1| h^{3/2} + \sum_{k=2}^{n-2} |\varepsilon_k| + |\varepsilon_{n-1}| h^{3/2} \right) \leqq \\ &\leqq K h^3 [h^{3/2} + (b-a) h^2 + h^{3/2}] \|\boldsymbol{\eta}\|_{h,\mathscr{L}_\infty}. \end{aligned} \tag{3.205}$$

Odtud již požadovaná nerovnost (3.201) ihned plyne.

Zcela stejně se dokáže konvergence metody sítí i v případě obecné samoadjungované diferenciální rovnice 4. řádu.

$$\left(p(x) y''\right)'' - \left(s(x) y'\right)' + q(x) y = f(x). \tag{3.206}$$

Na vyšetřovaný speciální případ jsme se omezili hlavně z důvodu zjednodušení zápisu. Analogicky by se také vyšetřovaly jiné typy okrajových podmínek.

Upozorněme závěrem ještě na jednu okolnost. Rychlosti konvergence, které jsme uvedli v případě rovnic druhého řádu, se nedaly zlepšit. Numerické experimenty naznačují, že zde tomu patrně tak není a že exponent 3/2 v odhadu (3.201) lze pravděpodobně zlepšit na 2. Dokázat se to však doposud nepodařilo.

3.3.3 Řešení vzniklých soustav

V tomto odstavci uvedeme vzorce pro řešení soustavy (3.179) Gaussovou eliminační metodou bez výběru hlavního prvku. V případě rovnice druhého řádu jsme tento postup ospravedlnili tak, že jsme ukázali jeho souvislost s metodou normalizovaného přesunu. Zde nemáme vybudován potřebný aparát; protože však matice soustavy (3.179) je symetrická a pozitivně definitní, odvoláme se na obecné tvrzení o proveditelnosti Gaussovy eliminace bez výběru hlavního prvku pro tuto třídu matic (viz např. Wilkinson (1965)).

Řešit soustavu (3.179) Gaussovou eliminací znamená vzhledem k pětidiagonálnosti její matice vypočítat čísla s_i a t_i z rekurencí

$$(3.207)\qquad s_1 = a_{12} \quad t_1 = a_{11}, \quad t_2 = a_{22} - \frac{a_{12}^2}{a_{11}},$$
$$s_{i+1} = a_{i+1,i+2} - \frac{s_i a_{i,i+2}}{t_i},$$
$$t_{i+2} = a_{i+2,i+2} - \frac{s_{i+1}^2}{t_{i+1}} - \frac{a_{i,i+2}^2}{t_i}, \quad i = 1, \ldots, n-3,$$

čísla d_i z rekurence

$$(3.208)\qquad d_1 = g_1, \quad d_2 = g_2 - \frac{a_{12}}{a_{11}} g_1,$$
$$d_{i+2} = g_{i+2} - \frac{s_{i+1}}{t_{i+1}} d_{i+1} - \frac{a_{i,i+2}}{t_i} d_i, \quad i = 1, \ldots, n-3,$$

a složky y_i hledaného vektoru y z rekurence

$$(3.209)\qquad y_{n-1} = \frac{d_{n-1}}{t_{n-1}}, \quad y_{n-2} = \frac{1}{t_{n-2}}(d_{n-2} - s_{n-2} y_{n-1}),$$
$$y_i = \frac{1}{t_i}(d_i - s_i y_{i+1} - a_{i,i+2} y_{i+2}), \quad i = n-3, \ldots, 1.$$

Eliminační metoda je v tomto případě sice pracnější než v případě okrajové úlohy pro rovnici druhého řádu, počet potřebných operací je však opět úměrný počtu rovnic, takže patrně žádná jiná metoda nebude efektivnější.

3.4 Nelineární diferenciální rovnice

V tomto odstavci zakončíme studium metody sítí pro řešení okrajových úloh pro obyčejné diferenciální rovnice tím, že na příkladě diferenciální rovnice (3.3) s okrajovými podmínkami (3.4) ukážeme, s jakými problémy se můžeme setkat při řešení nelineárních okrajových úloh. Začneme tím, že zformulujeme předpoklady, které stačí k tomu, aby výše zmíněná úloha měla právě jedno řešení.

Věta 3.15. *Nechť platí*

(i) *funkce f je definovaná a spojitá v pásu $R = \{(x, y);\ a \leqq x \leqq b,\ -\infty < y < \infty\}$;*

(ii) *existuje konstanta L taková, že platí*

$$(3.210)\qquad |f(x, y) - f(x, z)| \leqq L|y - z|$$

pro libovolnou dvojici $(x, y) \in R$ a $(x, z) \in R$;

(iii) *pro každé $(x, y) \in R$ existuje derivace f_y a je spojitá a nezáporná v R.*

Pak okrajová úloha (3.3), (3.4) má při libovolných γ_1 a γ_2 právě jedno řešení.

D ů k a z . K důkazu použijeme myšlenku metody střelby. Buď tedy $y(x; \alpha)$ řešení diferenciální rovnice (3.3) s počátečními podmínkami $y(x; \alpha) = \gamma_1$ a $y'(x; \alpha) = \alpha$ (čárkou zde i v dalších částech důkazu se rozumí derivace podle proměnné x). Podaří-li se dokázat, že rovnice

$$(3.211)\qquad y(b; \alpha) = \gamma_2$$

má při libovolném γ_2 právě jedno řešení, bude existence a jednoznačnost řešení dané okrajové úlohy dokázána.

Abychom to dokázali, uvědomme si předně, že za našich předpokladů funkce $y(x; \alpha)$ nejen existuje a je dvakrát spojitě diferencovatelná podle proměnné x při libovolném (pevném) α — tento fakt plyne ihned z věty 1.1 kap. I — ale je to dokonce spojitá funkce proměnných x a α a její parciální derivace $y_\alpha(x; \alpha)$ podle α existuje a je spojitá pro $x \in \langle a, b\rangle$ a pro libovolné α. Toto tvrzení plyne snadno z lemmatu 2.6. Položme

$$(3.212)\qquad \eta(x) = y_\alpha(x; \alpha)$$

a dokažme, že při pevném α platí

$$(3.213)\qquad \eta(x) \geqq x - a, \quad x \in \langle a, b\rangle.$$

Důkaz tohoto tvrzení provedeme sporem. Předpokládejme tedy, že (3.213) neplatí. Pak existuje bod $\xi \in \langle a, b\rangle$ takový, že je

$$(3.214)\qquad \eta(\xi) < \xi - a.$$

Derivujeme-li identitu

$$(3.215)\qquad y''(x; \alpha) = f(x, y(x; \alpha))$$

platnou pro libovolné $x \in \langle a, b\rangle$ a pro libovolné α podle α, dostaneme

$$(3.216)\qquad \eta''(x) = f_y(x, y(x; \alpha))\eta(x).$$

Protože platí $y(a; \alpha) = \gamma_1$ a $y'(a; \alpha) = \alpha$, platí dále

$$(3.217)\qquad \eta(a) = 0, \quad \eta'(a) = 1.$$

Odtud plyne předně, že pro bod ξ v nerovnosti (3.214) platí $\xi > a$. Dále tvrdíme, že tento bod lze zvolit dokonce tak, aby platilo

$$(3.218)\qquad \eta(x) > 0 \quad \text{pro } a < x \leqq \xi.$$

Položíme-li totiž $\xi_0 = \inf\{x; a < x \leqq b,\ \eta(x) \leqq 0\}$, je $\xi_0 > a$, protože platí (3.217). Dále je $\eta(x) > 0$ pro $x \in (a, \xi_0)$ a $\eta(\xi_0) = 0$, jak plyne ihned z definice čísla ξ_0. Nyní mohou nastat dva případy: 1) $\xi < \xi_0$; pak (3.218) platí. 2) $\xi \geqq \xi_0$; v bodě ξ_0 platí $\eta(\xi_0) = 0 < \xi_0 - a$. Funkce $\eta(x) - (x - a)$ je tedy v bodě $x = \xi_0$ záporná. Protože jde o spojitou funkci, je možno nalézt bod ξ_1, $\xi_1 < \xi_0$ tak, že platí

$$(3.219)\qquad \eta(\xi_1) < \xi_1 - a.$$

Protože je však $\xi_1 < \xi_0$, platí

$$\eta(x) > 0 \quad \text{pro } a < x \leqq \xi_1 \tag{3.220}$$

a za bod ξ lze vzít bod ξ_1.
Není-li tedy tvrzení (3.213) pravda, existuje skutečně bod $\xi \in (a, b\rangle$ tak, že platí (3.214) a (3.218). Podle věty o střední hodnotě existuje bod $\xi_2 \in (a, \xi)$ tak, že platí

$$\eta(\xi) = (\xi - a)\eta'(\xi_2). \tag{3.221}$$

V důsledku nerovnosti (3.214) je tedy

$$\eta'(\xi_2) < 1. \tag{3.222}$$

Aplikujeme-li znovu větu o střední hodnotě, tentokrát na funkci η' a na interval $\langle a, \xi_2\rangle$, dostaneme, že existuje bod $\xi_3 \in (a, \xi_2)$ takový, že platí

$$\eta'(\xi_2) - 1 = (\xi_2 - a)\eta''(\xi_3). \tag{3.223}$$

Z rovnice (3.223) a z nerovnosti (3.222) plyne, že v bodě ξ_1 platí $\eta_2''(\xi_3) < 0$. To je však ve sporu s rovnicí (3.216), neboť je $f_y(x, y) \geqq 0$ podle předpokladu (iii) a $\eta(\xi_3) > 0$ podle (3.118). Tento spor dokazuje platnost nerovnosti (3.213).

Z nerovnosti (3.213) plyne speciálně, že je

$$y_\alpha(b; \alpha) \geqq b - a > 0, \tag{3.224}$$

takže funkce $y(b; \alpha)$ je v intervalu $-\infty < \alpha < \infty$ rostoucí. Zvolíme-li α_1 pevně, plyne z nerovnosti (3.224) ihned, že pro $\alpha \geqq \alpha_1$ platí

$$\begin{aligned} y(b; \alpha) = y(b; \alpha_1) + \int_{\alpha_1}^{\alpha} y_\alpha(b; t)\,\mathrm{d}t \geqq \\ \geqq y(b; \alpha_1) + (\alpha - \alpha_1)(b - a). \end{aligned} \tag{3.225}$$

Platí tedy

$$\lim_{\alpha \to \infty} y(b; \alpha) = \infty. \tag{3.226}$$

Úplně stejně se dokáže, že platí také

$$\lim_{\alpha \to -\infty} y(b; \alpha) = -\infty. \tag{3.227}$$

Funkce $y(b; \alpha)$ je tedy rostoucí funkce proměnné α, která zobrazuje interval $(-\infty, \infty)$ na interval $(-\infty, \infty)$. Rovnice (3.221) má tedy při libovolném γ_2 právě jedno řešení. Věta je dokázána.

Důkaz věty 3.15 je konstruktivní v tom smyslu, že udává současně postup, kterým lze hledané řešení sestrojit. Protože jsme však viděli, že už v lineárním případě může docházet při metodě střelby k nežádoucí ztrátě přesnosti, ukážeme v dalších dvou pododstavcích, jak lze danou okrajovou úlohu alternativně řešit metodou sítí.

3.4.1 Sestavení diferenčních rovnic a jejich řešitelnost

Řešit okrajovou úlohu (3.3), (3.4) metodou sítí značí, jak už víme z úvodu k tomuto článku, hledat přibližné řešení $y_1, \ldots, y_{n-1}$ ze soustavy rovnic (3.6). Položíme-li $y = (y_1, \ldots, y_{n-1})^T$, lze tuto soustavu psát ve tvaru

$$Ay + h^2 f(y) = g, \tag{3.228}$$

kde A je čtvercová matice řádu $(n-1)$ daná vzorcem

$$A = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix} \tag{3.229}$$

f je vektorová funkce $(n-1)$ proměnných daná předpisem

$$f(y) = [f(x_1, y_1), \ldots, f(x_{n-1}, y_{n-1})]^T \tag{3.230}$$

a $g = (\gamma_1, 0, \ldots, 0, \gamma_2)^T$ je daný $(n-1)$-dimenzionální vektor. Poznamenejme, že vektorová funkce, která je tvaru (3.230), tj. funkce, která $(n-1)$-dimenzionálnímu vektoru přiřazuje $(n-1)$-dimenzionální vektor takový, že jeho i-tá složka závisí pouze na i-té složce argumentu, se nazývá *diagonální zobrazení*.

Stejně jako dříve je první otázka, na kterou je třeba odpovědět, otázka po existenci řešení soustavy (3.228). Má-li však soustava, na kterou vede metoda sítí, popisovat algoritmus pro řešení dané úlohy, je nutno nejen vědět, že její řešení existuje, ale i udat postup, kterým lze toto řešení vypočítat. Proto na otázku po existenci řešení soustavy (3.228) odpovíme konstruktivně, a to tak, že existenční důkaz provedeme pomocí konstrukce určité iterační metody pro řešení této soustavy. K tomuto cíli vyslovíme a dokážeme několik pomocných tvrzení.

Lemma 3.15. *Buď B matice, jejíž spektrální poloměr $\varrho(B)$ je menší než 1. Pak řada*

$$\sum_{k=0}^{\infty} B^k \tag{3.231}$$

konverguje.

Důkaz. Buď $J = T^{-1}BT$ Jordanův kanonický tvar matice B. Protože zřejmě platí

$$\sum_{k=0}^{N} B^k = T\left(\sum_{k=0}^{N} J^k\right)T^{-1}, \tag{3.232}$$

stačí tvrzení dokázat pro matici v Jordanově kanonickém tvaru. Protože však taková matice je blokově diagonální, stačí se omezit na jeden Jordanův blok. Obecný prvek matice J_s^k, kde J_s je Jordanův blok řádu s, je tvaru $P(k)\lambda^{k-r}$, kde P je polynom stupně nejvýše $s-1$ a pro číslo r platí nerovnosti $0 \leqq r \leqq s-1$ (srv. důkaz lemmatu 4.2 v kap. I). Konvergence řady $\sum P(k)\lambda^{k-r}$ pak plyne ihned např. z podílového kritéria, neboť λ je vlastní číslo matice B a pro ně platí $|\lambda| < 1$ podle předpokladu.

Lemma 3.16. *Položme*

$$A = D - L - U, \tag{3.233}$$

kde D je diagonální matice, L dolní trojúhelníková matice, U horní trojúhelníková matice, obě s nulami na diagonále a A je matice daná vzorcem (3.229). *Pak platí*

$$\varrho(D^{-1}(L+U)) < 1. \tag{3.234}$$

D ů k a z . Je zřejmě

$$D^{-1}(L+U) = \begin{bmatrix} 0 & 1/2 & 0 & \dots & 0 \\ 1/2 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1/2 \\ 0 & \dots & 0 & 1/2 & 0 \end{bmatrix}. \tag{3.235}$$

Přímým výpočtem se snadno přesvědčíme, že soustava vektorů $v^{(\nu)} = (v_1^{(\nu)}, \dots, v_{n-1}^{(\nu)})^T$, $\nu = 1, \dots, n-1$, kde

$$v_k^{(\nu)} = \sin \frac{k\nu\pi}{n}, \quad k = 1, \dots, n-1, \tag{3.236}$$

tvoří úplnou soustavu vlastních vektorů matice $D^{-1}(L+U)$ a že odpovídající vlastní čísla $\lambda^{(\nu)}$ jsou dána vzorcem

$$\lambda^{(\nu)} = \cos \frac{\nu\pi}{n}, \quad \nu = 1, \dots, n-1. \tag{3.237}$$

Odtud už tvrzení lemmatu plyne.

Lemma 3.17. *Nechť jsou splněny předpoklady* (i) *až* (iii) *věty 3.15 a nechť $R = (R_1, \dots, R_{n-1})^T$ je vektorová funkce $(n-1)$ proměnných $y_1, \dots, y_{n-1}$ daná předpisem*

$$R(y) = y + Sf(y), \tag{3.238}$$

kde S je diagonální matice s nezápornými diagonálními prvky a $f(y)$ je vektor ze vzorce (3.230). Pak rovnice

$$R(y) = r, \tag{3.239}$$

kde $r = (r_1, \dots, r_{n-1})^T$ je libovolný $(n-1)$-dimenzionální vektor, má právě jedno řešení.

D ů k a z . Protože je

$$R_i(y) = y_i + s_i f(x_i, y_i), \quad i = 1, \dots, n-1, \tag{3.240}$$

kde s_i jsou diagonální prvky matice S, je R diagonální zobrazení a každá složka vektoru y je řešením skalární rovnice

$$R_i(y_i) \equiv y_i + s_i f(x_i, y_i) = r_i. \tag{3.241}$$

Tato rovnice má však při libovolném r_i právě jedno řešení, neboť funkce $R_i(y_i)$ je rostoucí funkce proměnné y_i a zobrazuje interval $(-\infty, \infty)$ na interval $(-\infty, \infty)$. Je tomu tak proto, že platí $\partial R_i(y_i)/\partial y_i = 1 + s_i f_y(x_i, y_i)$ a čísla s_i a $f_y(x_i, y_i)$ jsou vesměs nezáporná (srv. důkaz věty 3.15).

Věta 3.16. *Nechť jsou splněny předpoklady* (i) *až* (iii) *z věty 3.15 a nechť $y^{(0)}$ je libovolný $(n-1)$-dimenzionální vektor. Definujme posloupnost vektorů $y^{(\nu)}$ předpisem*

$$y^{(\nu+1)} + h^2 D^{-1} f(y^{(\nu+1)}) = D^{-1}(L+U)y^{(\nu)} + D^{-1}g, \quad \nu = 0, 1, \dots, \tag{3.242}$$

kde matice D, L a U jsou definovány rovnicí (3.233). *Pak posloupnost $y^{(\nu)}$ konverguje k jedinému řešení soustavy* (3.228).

D ů k a z . Položme speciálně

$$R(y) = y + h^2 D^{-1} f(y). \tag{3.243}$$

Protože matice D^{-1} má nezáporné prvky, jsou splněny předpoklady lemmatu 3.17 a posloupnost vektorů $y^{(\nu)}$ je tedy rovnicemi (3.242) skutečně definována. Buďte y a $\tilde{y}$ dva libovolné $(n-1)$-dimenzionální vektory a vyšetřujme rozdíl $R(y) - R(\tilde{y})$. Je

$$\begin{aligned} R_i(y_i) - R_i(\tilde{y}_i) &= y_i + \frac{h^2}{a_{ii}} f(x_i, y_i) - \tilde{y}_i - \frac{h^2}{a_{ii}} f(x_i, \tilde{y}_i) = \\ &= y_i - \tilde{y}_i + \frac{h^2}{a_{ii}} [f(x_i, y_i) - f(x_i, \tilde{y}_i)] \end{aligned} \tag{3.244}$$

(a_{ii} jsou diagonální prvky matice A). Protože podle předpokladu (iii) je $f(x_i, y_i) \geqq f(x_i, \tilde{y}_i)$ pro $y_i \geqq \tilde{y}_i$, plyne odtud, že platí

$$|y_i - \tilde{y}_i| \leqq |R_i(y_i) - R_i(\tilde{y}_i)|, \quad i = 1, \dots, n-1. \tag{3.245}$$

Nerovnosti (3.245) můžeme stručně zapsat ve tvaru

$$|y - \tilde{y}| \leqq |R(y) - R(\tilde{y})|. \tag{3.246}$$

Položíme-li v této nerovnosti $y = y^{(\nu+1)}$ a $\tilde{y} = y^{(\nu)}$, dostáváme

$$|y^{(\nu+1)} - y^{(\nu)}| \leqq |R(y^{(\nu+1)}) - R(y^{(\nu)})| = \\ = |D^{-1}(L+U)(y^{(\nu)} - y^{(\nu-1)})| \leqq D^{-1}(L+U)|y^{(\nu)} - y^{(\nu-1)}|. \tag{3.247}$$

Poslední nerovnost plyne z toho, že prvky matice $D^{-1}(L+U)$ jsou nezáporné. Z nerovnosti (3.247) však plyne okamžitě úplnou indukcí, že je

$$|y^{(\nu+1)} - y^{(\nu)}| \leqq [D^{-1}(L+U)]^{\nu}|y^{(1)} - y^{(0)}|. \tag{3.248}$$

Protože však je $\varrho\big(D^{-1}(L+U)\big) < 1$ (viz lemma 3.16), je podle lemmatu 3.15 řada

$$\sum_{\nu=0}^{\infty} |y^{(\nu+1)} - y^{(\nu)}| \tag{3.249}$$

konvergentní. Konverguje tedy také řada $\sum_{\nu=0}^{\infty}(y^{(\nu+1)} - y^{(\nu)})$, tj. také posloupnost $y^{(\nu)}$. Limitní vektor y je řešením soustavy (3.228), jak plyne ihned limitním přechodem v rovnici (3.242), neboť obě její strany jsou spojité.

Zbývá dokázat jednoznačnost řešení. Důkaz provedeme sporem. Nechť tedy existují dvě řešení y a z soustavy (3.228). Pro ně platí

$$Ay + h^2 f(y) = Az + h^2 f(z) \tag{3.250}$$

neboli

$$A(y-z) + h^2[f(y) - f(z)] = 0. \tag{3.251}$$

Podle věty o střední hodnotě však existují čísla θ_i, $i = 1, \ldots, n-1$, taková, že je

$$f(y) - f(z) = H(y-z) \tag{3.252}$$

a H je diagonální matice s čísly $f_y(x_i, \theta_i)$ na diagonále. Tato čísla jsou však nezáporná a matice $A + h^2 H$ je tedy podle Collatzova lemmatu 3.5 monotónní. To však dokazuje, že je $y - z = 0$. Tím jsme dokončili důkaz věty.

Metoda sítí pro řešení problému (3.3), (3.4) má tedy skutečně smysl a věta 3.16 dává navíc návod, jak přibližné řešení nalézt. K řešení $(n-1)$ skalárních nelineárních rovnic, na něž se rozpadá konstrukce každé iterace ve vzorci (3.242), lze užít kteroukoliv metodu pro řešení nelineárních rovnic; jako příklad vhodné metody může sloužit Newtonova metoda.

Poznamenejme, že Newtonovu metodu lze užít na soustavu (3.228) také přímo. V tomto případě se posloupnost aproximací řešení sestrojuje podle vzorce

$$y^{(\nu+1)} = y^{(\nu)} + [G(y^{(\nu)})]^{-1}\varphi(y^{(\nu)}), \quad \nu = 0, 1, \ldots, \tag{3.253}$$

kde jsme položili

$$\varphi(y) = Ay + h^2 f(y) - g \tag{3.254}$$

a $G(y)$ je Jaccobiova matice této funkce, tj. je

$$G(y) = A + h^2 F(y), \tag{3.255}$$

kde A je matice (3.229) a $F(y)$ je diagonální matice tvaru

$$F(y) = \begin{bmatrix} f_y(x_1, y_1) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & f_y(x_{n-1}, y_{n-1}) \end{bmatrix}. \tag{3.256}$$

Protože matice F je nezáporná, je matice G podle Collatzova lemmatu (viz lemma 3.5) regulární a posloupnost $y^{(\nu)}$ je rovnicemi (3.253) dobře definována. Je-li tedy $y^{(0)}$, dostatečně blízko přesnému řešení, iterace (3.253), jak je známo, konvergují.

Upozorněme ještě závěrem, že všechny provedené úvahy lze bezprostředně zobecnit na případ diferenciální rovnice (3.3), kde druhá derivace na levé straně je nahrazena takovým lineárním diferenciálním operátorem druhého řádu, že jeho aproximace metodou sítí vede na matici A, pro jejíž rozklad (3.233) platí, že diagonální matice D má kladné diagonální prvky, že matice $D^{-1}(L+U)$ je nezáporná a že platí nerovnost (3.234).

3.4.2 Konvergence

Důkaz konvergence navržené metody sítí je ve srovnání s důkazem řešitelnosti soustavy (3.228) už velice jednoduchý a je v podstatě pouhým důsledkem následujícího lemmatu.

Lemma 3.18. *Buď A matice řádu $n-1$ daná vzorcem* (3.229). *Pak A je regulární a označíme-li prvky inverzní matice a_{ij}^{-1}, platí*

$$a_{ij}^{-1} = \begin{cases} \frac{(n-j)i}{n}, & i \leqq j, \\ \frac{(n-i)j}{n}, & i > j. \end{cases} \tag{3.257}$$

D ů k a z . Regularita matice A plyne triviálně např. z Collatzova lemmatu.

Buď dále $a^{-1,j}$ j-tý sloupec matice A^{-1}, tj. $a^{-1,j} = (a_{1j}^{-1}, \ldots, a_{n-1,j}^{-1})^T$. Pak platí $Aa^{-1,j} = e^{(j)}$, kde vektor $e^{(j)}$ je j-tý sloupec jednotkové matice, neboli, položíme-li $a_{0j}^{-1} = a_{nj}^{-1} = 0$,

$$-a_{i-1,j}^{-1} + 2a_{ij}^{-1} - a_{i+1,j}^{-1} = 0, \quad i = 1, \ldots, n-1, \ i \neq j, \tag{3.258}$$

a

$$-a_{j-1,j}^{-1} + 2a_{jj}^{-1} - a_{j+1,j}^{-1} = 1. \tag{3.259}$$

Hledejme řešení soustavy (3.258), (3.259) ve tvaru

$$a_{ij}^{-1} = \alpha i \tag{3.260}$$

pro $i = 0, \ldots, j$ a

$$a_{ij}^{-1} = \beta(n-i) \tag{3.261}$$

pro $i = j, \ldots, n$, kde α a β jsou vhodné konstanty. Rovnice (3.258) jsou zřejmě splněny při libovolném α a β, je tedy třeba určit tyto konstanty tak, aby platilo

$$\alpha j = \beta(n-j) \tag{3.262}$$

a aby byla splněna rovnice (3.259), tj. aby platilo

$$-\alpha(j-1) + \alpha j + \beta(n-j) - \beta(n-j-1) = 1. \tag{3.263}$$

Vypočteme-li α a β z rovnic (3.262) a (3.263), dostaneme už rovnice (3.257). Lemma je dokázáno.

Věta 3.17. *Nechť pravá strana f diferenciální rovnice* (3.3) *splňuje požadavky* (i) až (iii) *z věty 3.15 a nechť navíc má v R spojité druhé parciální derivace podle y. Nechť dále y je přesné řešení úlohy* (3.3), (3.4) *a y_k, $k = 1, \ldots, n-1$, je řešení soustavy* (3.228). *Pak existuje konstanta M taková, že pro $k = 1, \ldots, n-1$ a pro každé dostatečně malé h platí*

$$|y_k - y(x_k)| \leqq Mh^2. \tag{3.264}$$

D ů k a z . Za uvedených předpokladů má přesné řešení y dané okrajové úlohy v intervalu $\langle a, b\rangle$ čtyři spojité derivace. Podle lemmatu 3.8 tedy platí

$$|-y(x_{k-1}) + 2y(x_k) - y(x_{k+1}) + h^2 f(x_k, y_k)| \leqq \tfrac{1}{12} Z h^4, \quad k = 1, \ldots, n-1, \tag{3.265}$$

kde

$$Z = \max_{x \in \langle a,b\rangle} |y''''(x)|. \tag{3.266}$$

Z nerovností (3.265) však plyne, že existují čísla θ_k, $|\theta_k| \leqq 1$, taková, že je

$$-y(x_{k-1}) + 2y(x_k) - y(x_{k+1}) + h^2 f(x_k, y_k) = \tfrac{1}{12}\theta_k Z h^4. \tag{3.267}$$

Položíme-li $\eta_k = y(x_k) - y_k$, odečteme-li k-tou rovnici (3.228) od rovnice (3.267) a užijeme-li větu o střední hodnotě, dostaneme

$$-\eta_{k-1} + 2\eta_k - \eta_{k+1} + h^2 g_k \eta_k = \tfrac{1}{12}\theta_k Z h^4, \quad k = 1, \ldots, n-1, \quad \eta_0 = \eta_n = 0, \tag{3.268}$$

kde g_k jsou hodnoty funkce $f_y(x_k, y)$ ve vhodných bodech $y = \tilde{y}_k$, tedy nezáporná čísla. Položíme-li jako obvykle $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{n-1})^T$, můžeme rovnice (3.268) zapsat maticově jako

$$(A + h^2 H)\boldsymbol{\eta} = \tfrac{1}{12} Z h^4 P, \tag{3.269}$$

kde H a P jsou diagonální matice s čísly g_k, resp. θ_k na diagonále. Matice H je tedy nezáporná a matice $A + h^2 H$ je podle Collatzova lemmatu monotónní. Podle lemmatu 3.3 tedy platí

$$(A + h^2 H)^{-1} \leqq A^{-1}. \tag{3.270}$$

Maticová norma, která je indukována $\mathscr{L}_\infty$-vektorovou normou, je rovna maximu ze součtů absolutních hodnot prvků v jednotlivých jejích řádcích. Užijeme-li tedy tuto maticovou normu, platí

$$\|(A + h^2 H)^{-1}\| \leqq \|A^{-1}\|, \tag{3.271}$$

protože matice $A + h^2 H$ a A jsou monotónní, a tedy matice $(A + h^2 H)^{-1}$ a A^{-1} jsou podle lemmatu 3.1 nezáporné. Je však

$$\|A^{-1}\| = \max_{i=1,\ldots,n-1} \sum_{j=1}^{n-1} a_{ij}^{-1} = \max_{i=1,\ldots,n-1} \frac{(n-i)i}{2} \leqq \frac{n^2}{8} \leqq \frac{(b-a)^2}{8}\frac{1}{h^2}. \tag{3.272}$$

Z nerovností (3.270), (3.271) a z rovnice (3.272) už tvrzení věty plyne bezprostředně.

4 Variační metody

V tomto článku si všimneme velice stručně takových metod pro řešení okrajových úloh, které souvisí s variačním počtem. Tyto metody vycházejí ze skutečnosti, že velmi mnoho fyzikálních zákonů lze vyjádřit pomocí minimálních principů. Tak např. rovnovážný stav mechanické soustavy je takový stav, který odpovídá minimu její potenciální energie. Z toho důvodu je okrajová úloha pro diferenciální rovnici popisující rovnováhu mechanické soustavy obecně ekvivalentní s úlohou nalezení funkce, pro níž integrál vyjadřující potenciální energii této soustavy nabývá minima. Matematicky řečeno to znamená, že řešení okrajové úlohy pro diferenciální rovnici je ekvivalentní úloze variačního počtu, tj. úloze nalézt extrém integrálu, jehož Eulerovou rovnicí je daná diferenciální rovnice. K řešení úloh variačního počtu lze užít metody, kdy se přímo sestrojuje aproximace extrému daného integrálu, aniž by se napřed přecházelo k jeho Eulerově rovnici. V důsledku naznačené ekvivalence jsou tyto metody zároveň přibližnými metodami řešení okrajových úloh pro diferenciální rovnice. Z důvodů, které byly naznačeny výše, se variační metody řešení okrajových úloh nazývají také *přímé metody*. Dříve než popíšeme nejdůležitější z nich, totiž Ritzovu a Galerkinovu metodu, ukážeme souvislost některých důležitých okrajových úloh s variačními úlohami. Protože variační metody velmi úzce souvisí s funkcionálně analytickou teorií okrajových úloh pro diferenciální rovnice, budeme v celém tomto článku předpokládat znalosti základních pojmů funkcionální analýzy, zejména pak teorie Hilbertových prostorů.

4.1 Variační formulace okrajových úloh

4.1.1 Lineární diferenciální rovnice druhého řádu

Buď dána diferenciální rovnice (1.12) s jednoduchými okrajovými podmínkami

$$y(a) = \gamma_1, \quad y(b) = \gamma_2 \tag{4.1}$$

a nechť jsou splněny předpoklady věty 2.2, takže okrajová úloha (1.12), (4.1) má právě jedno řešení. Položíme-li $v(x) = y(x) - l(x)$, kde funkce l je dána vzorcem

$$l(x) = \gamma_1 \frac{b-x}{b-a} + \gamma_2 \frac{a-x}{a-b} \tag{4.2}$$

(graf této funkce je přímka spojující body (a, γ_1) a (b, γ_2)), je funkce v řešením diferenciální rovnice $-\big(p(x)v'\big)' + q(x)v = g(x)$ s okrajovými podmínkami $v(a) = v(b) = 0$. Protože tato diferenciální rovnice je stejného typu jako diferenciální rovnice (1.12), jen její pravá strana je jiná, lze v dalším výkladu bez újmy na obecnosti předpokládat, že řešíme diferenciální rovnici (1.12) s homogenními okrajovými podmínkami

$$y(a) = 0, \quad y(b) = 0. \tag{4.3}$$

Zaveďme operátor L předpisem

$$Lv = -(pv')' + qv. \tag{4.4}$$

Tento operátor zobrazuje množinu $\mathscr{D}_L$, což je množina funkcí, které mají v intervalu $\langle a, b\rangle$ dvě spojité derivace a které splňují podmínky $v(a) = v(b) = 0$, do množiny $\mathscr{C}(\langle a, b\rangle)$ funkcí spojitých na intervalu $\langle a, b\rangle$. Řešit okrajovou úlohu je tedy totéž, jako řešit rovnici

$$Ly = f \tag{4.5}$$

v množině $\mathscr{D}_L$.

Následující pomocná tvrzení umožní formulovat variační úlohu, jejíž řešení je řešením dané okrajové úlohy.

Lemma 4.1. *Operátor L definovaný na množině $\mathscr{D}_L$ rovnicí* (4.4) *je symetrický operátor, tj. pro každé dvě funkce $u, v \in \mathscr{D}_L$ platí*

$$\int_a^b (Lu)(x)v(x)\,\mathrm{d}x = \int_a^b u(x)(Lv)(x)\,\mathrm{d}x. \tag{4.6}$$

D ů k a z . Tvrzení lemmatu dostaneme ihned dvojí integrací per partes.

Lemma 4.2. *Nechť funkce p, p' a q jsou spojité a nechť platí nerovnosti*

$$p(x) \geqq p_0 > 0, \quad q(x) \geqq 0, \quad x \in \langle a, b\rangle, \tag{4.7}$$

Pak pro každou funkci $u \in \mathscr{D}_L$ platí

$$\int_a^b (Lu)(x)v(x)\,\mathrm{d}x \geqq \frac{p_0}{b-a}\|u\|^2_{\mathscr{L}_\infty}, \tag{4.8}$$

kde symbol $\|u\|_{\mathscr{L}_\infty}$ značí normu funkce u v prostoru $\mathscr{L}_\infty$, tj. $\|u\|_{\mathscr{L}_\infty} = \max\limits_{x\in\langle a,b\rangle} |u(x)|$.

D ů k a z . Tvrzení bylo už v podstatě dokázáno v odst. 3.2.2, když jsme studovali konvergenci metody sítí pomocí energetických nerovností (srv. vzorce (3.98), (3.99) a (3.106)). Protože však tam vyšetřovaná rovnice má jiné okrajové podmínky, zopakujeme příslušný důkaz ještě jednou. Pro každou funkci $u \in \mathscr{D}_L$ platí pro každé $x \in \langle a, b\rangle$

$$u(x) = \int_a^x u'(\xi)\,\mathrm{d}\xi, \tag{4.9}$$

protože je $u(a) = 0$. Odtud pomocí Schwarzovy nerovnosti dostáváme

$$u^2(x) \leqq (x-a)\int_a^x [u'(\xi)]^2\,\mathrm{d}\xi \leqq (b-a)\int_a^b [u'(\xi)]^2\,\mathrm{d}\xi, \tag{4.10}$$

tj.

$$\|u\|^2_{\mathscr{L}_\infty} \leqq (b-a)\int_a^b [u'(x)]^2\,\mathrm{d}x. \tag{4.11}$$

Na druhé straně, integrace per partes a nerovnosti (4.7) dávají

$$\int_a^b (Lu)(x)u(x)\,\mathrm{d}x = \int_a^b p(x)[u'(x)]^2\,\mathrm{d}x + \int_a^b q(x)u^2(x)\,\mathrm{d}x \geqq \geqq p_0\int_a^b [u'(x)]^2\,\mathrm{d}x. \tag{4.12}$$

Z nerovností (4.12) a (4.11) však už tvrzení lemmatu plyne bezprostředně.

Položme nyní pro $u \in \mathscr{D}_L$

$$F(u) = \int_a^b (Lu)(x)u(x)\,\mathrm{d}x - 2\int_a^b f(x)u(x)\,\mathrm{d}x \tag{4.13}$$

nebo alternativně

$$F(u) = \int_a^b \{p(x)[u'(x)]^2 + q(x)u^2(x)\}\,\mathrm{d}x - 2\int_a^b f(x)u(x)\,\mathrm{d}x, \tag{4.14}$$

kde vztah (4.14) vznikl ze (4.13) integrací per partes. Zavedené označení dovoluje zformulovat následující větu.

Věta 4.1. *Nechť funkce p, p', q a f jsou spojité v intervalu $\langle a, b\rangle$ a nechť jsou splněny nerovnosti* (4.7). *Nechť dále funkce $y \in \mathscr{D}_L$ je řešením rovnice* (4.5). *Pak platí*

$$F(u) > F(y) \tag{4.15}$$

pro každou funkci $u \in \mathscr{D}_L$, $u \neq y$. Naopak, nechť existuje funkce $y \in \mathscr{D}_L$ taková, že pro každou funkci $u \in \mathscr{D}_L$, $u \neq y$, platí nerovnost (4.15). *Pak y je řešením rovnice* (4.5).

D ů k a z . Buď za prvé $y \in \mathscr{D}_L$ řešením rovnice (4.5) a buď u libovolná funkce taková, že je $u \in \mathscr{D}_L$, $u \neq y$. Pak podle lemmatu 4.1 platí

$$\begin{aligned} F(u) &= \int_a^b (Lu)(x)u(x)\,\mathrm{d}x - 2\int_a^b f(x)u(x)\,\mathrm{d}x = \\ &= \int_a^b (Lu)(x)u(x)\,\mathrm{d}x - 2\int_a^b (Ly)(x)u(x)\,\mathrm{d}x + \\ &\quad + \int_a^b (Ly)(x)y(x)\,\mathrm{d}x - \int_a^b (Ly)(x)y(x)\,\mathrm{d}x = \\ &= \int_a^b L(u-y)(x)[u(x)-y(x)]\,\mathrm{d}x - \int_a^b (Ly)(x)y(x)\,\mathrm{d}x. \end{aligned} \tag{4.16}$$

Podle lemmatu 4.2 však pro $u \neq y$ platí

$$\int_a^b L(u-y)(x)[u(x)-y(x)]\,\mathrm{d}x > 0, \tag{4.17}$$

takže odtud a z rovnice (4.16) plyne, že je

$$F(u) > -\int_a^b (Ly)(x)y(x)\,\mathrm{d}x = F(y). \tag{4.18}$$

Nechť naopak je y funkce, ve které nabývá funkcionál F absolutního minima v množině $\mathscr{D}_L$. Buď u libovolná pevně zvolená funkce z množiny $\mathscr{D}_L$ a položme

$$\varphi(t) = F(y + tu). \tag{4.19}$$

Funkce φ jedné reálné proměnné t je zřejmě diferencovatelná na celé reálné ose a nabývá v bodě $t = 0$ minima. Musí tedy pro ni platit

$$\varphi'(0) = 0. \tag{4.20}$$

Protože je

$$\begin{aligned} \varphi'(t) &= \frac{\mathrm{d}}{\mathrm{d}t}F(y+tu) = \\ &= 2\int_a^b \{p(x)[y'(x)+tu'(x)]u'(x) + q(x)[y(x)+tu(x)]u(x)\}\,\mathrm{d}x - \\ &\quad - 2\int_a^b f(x)u(x)\,\mathrm{d}x, \end{aligned} \tag{4.21}$$

je

$$\begin{aligned} \varphi'(0) &= 2\int_a^b p(x)y'(x)u'(x)\,\mathrm{d}x + 2\int_a^b q(x)y(x)u(x)\,\mathrm{d}x - \\ &\quad - 2\int_a^b f(x)u(x)\,\mathrm{d}x = 2\int_a^b \{-[p(x)y'(x)]' + q(x)y(x) - f(x)\}u(x)\,\mathrm{d}x. \end{aligned} \tag{4.22}$$

Pro každou funkci $u \in \mathscr{D}_L$ tedy platí

$$\int_a^b \{-[p(x)y'(x)]' + q(x)y(x) - f(x)\}u(x)\,\mathrm{d}x = 0. \tag{4.23}$$

Protože však množina $\mathscr{D}_L$ je zřejmě hustá v prostoru $\mathscr{L}_2(a,b)$ funkcí, které jsou měřitelné a integrovatelné v intervalu $\langle a, b\rangle$ s druhou mocninou, je $-[p(x)y'(x)]' + q(x)y(x) - f(x) = 0$ skoro všude v $\langle a, b\rangle$. Protože je však $y \in \mathscr{D}_L$, je funkce $-[p(x)y'(x)]' + q(x)y(x) - f(x)$ vzhledem k předpokladům o funkcích p, p', q a f spojitá, a tedy rovna nule pro každé $x \in \langle a, b\rangle$. Věta je dokázána.

Ve větě 4.1 jsme tedy nalezli hledanou ekvivalenci mezi řešením okrajové úlohy (1.12), (4.3) a mezi hledáním extrému funkcionálu definovaného rovnicí (4.13) nebo (4.14).

Poznamenejme, že variační formulace okrajové úlohy, která je obsahem věty 4.1, tvoří výchozí bod k podstatnému zobecnění pojmu okrajové úlohy, a to nejen v případě obyčejných diferenciálních rovnic, ale i v případě parciálních diferenciálních rovnic. Aby čtenář získal aspoň přibližnou orientaci, popíšeme, aniž budeme zacházet do podrobností, základní rysy příslušného postupu.

Množina $\mathscr{D}_L$, která zřejmě tvoří reálný vektorový prostor (tj. takovou množinu, že patří-li do ní prvky u a v, patří do ní také lineární kombinace $\alpha u + \beta v$, kde α a β jsou libovolná reálná čísla), se pokládá za podprostor Hilbertova prostoru $\mathscr{L}_2(a,b)$ funkcí integrovatelných v intervalu $\langle a, b\rangle$ s kvadrátem, v němž je skalární součin zaveden rovnicí

$$(u, v) = \int_a^b u(x)v(x)\,\mathrm{d}x \tag{4.24}$$

a norma rovnicí

$$\|u\| = (u, u)^{1/2}. \tag{4.25}$$

Operátor L definovaný rovnicí (4.4) je pak lineární operátor, který zobrazuje množinu $\mathscr{D}_L \subset \mathscr{L}_2$ do prostoru $\mathscr{L}_2$, a tento operátor je podle lemmat 4.1 a 4.2 symetrický a pozitivně definitní v tom smyslu, že platí $(Lu, u) > 0$ pro každou funkci $u \in \mathscr{D}_L$, $u \neq 0$. Definujeme-li tedy na $\mathscr{D}_L$ bilineární formu $[u, v]$ (tj. funkci dvou proměnných u a v, která je lineární v každé proměnné zvlášť) rovnicí

$$[u, v] = (Lu, v) \tag{4.26}$$

nebo alternativně vzhledem k možnosti integrace per partes rovnicí

$$[u, v] = \int_a^b [p(x)u'(x)v'(x) + q(x)u(x)v(x)]\,\mathrm{d}\,x, \tag{4.27}$$

je tato forma skalárním součinem. Zavedeme-li ještě normu rovnicí

$$\|u\|_L = [u, u]^{1/2}, \tag{4.28}$$

stane se vektorový prostor $\mathscr{D}_L$ lineárním normovaným prostorem, který není obecně úplný. Jeho úplný obal označíme $\mathscr{D}$; je to Hilbertův prostor, pro nějž je zřejmě $\mathscr{D} \supset \supset \mathscr{D}_L$, a množina $\mathscr{D}_L$ je v něm hustá v metrice dané normou (4.28). Prvky prostoru $\mathscr{D}$ jsou tvořeny právě těmi funkcemi z tzv. Sobolevova prostoru $\mathscr{H}^1$ které splňují okrajové podmínky (4.3). *Sobolevovým prostorem* $\mathscr{H}^k$ (běžné značení je také $\mathscr{W}_2^k$) se přitom rozumí Hilbertův prostor funkcí, které mají v daném omezeném intervalu absolutně spojitou $(k-1)$-ní derivaci takovou, že platí

$$\int_a^b [v^{(k)}(x)]^2\,\mathrm{d}\,x < \infty, \tag{4.29}$$

a v němž je skalární součin dán rovnicí

$$(u, v)_k = \sum_{j=0}^{k} (u^{(j)}, v^{(j)}). \tag{4.30}$$

Rozšíříme-li definici funkcionálu F na funkce $u \in \mathscr{D}$ rovnicí

$$F(u) = [u, u] - 2(f, u), \tag{4.31}$$

což má smysl, neboť skalární součin $[u, v]$ je pro prvky z prostoru $\mathscr{D}$ dán rovnicí (4.27), zůstává věta 4.1 v platnosti, bereme-li v ní funkce u z množiny $\mathscr{D}$ místo z množiny $\mathscr{D}_L$. Až dosud jsme předpokládali, že koeficient p v diferenciálním operátoru L je spojitě diferencovatelná funkce v intervalu $\langle a, b\rangle$. K tomu, aby integrál na pravé straně rovnice (4.27) měl smysl, je tento požadavek příliš přísný a lze jej podstatně zeslabit, stačí např. předpokládat, že koeficient je omezená měřitelná funkce, pro niž platí nerovnost $p(x) \geqq p_0 > 0$. Podobně k tomu, aby měl smysl výraz (f, u), není třeba předpokládat, že funkce f je spojitá; stačí např., aby platilo $f \in \mathscr{L}_2(a, b)$. Pro funkce f a p právě zmíněných vlastností a pro $u \in \mathscr{D}$ nemá rovnice (4.5) obecně vůbec smysl, nicméně funkcionál F je rovnicí (4.31) stále dobře definován a nadto, jak hned uvidíme, nabývá v množině $\mathscr{D}$ svého minima. V tomto

případě je rozumné pokládat toto minimum za *zobecněné řešení* úlohy, která není nyní charakterizována operátorem L (ten nemá obecně smysl), ale bilineární formou (4.27).

Zobecněné řešení lze charakterizovat také tak, že je to takový prvek y prostoru $\mathscr{D}$, pro který platí rovnice

$$[y, v] = (f, v) \tag{4.32}$$

pro každou funkci $v \in \mathscr{D}_L$. V této souvislosti se zobecněné řešení dané okrajové úlohy nazývá také *slabé řešení*. Důvod je ten, že v případě, že řešení y leží v množině $\mathscr{D}_L$ a že jak koeficient p, tak pravá strana f jsou dostatečně hladké, takže jde o klasické řešení rovnice (4.5), lze rovnici (4.32) psát ve tvaru

$$\int_a^b (Ly)(x)v(x)\,\mathrm{d}\,x = \int_a^b f(x)v(x)\,\mathrm{d}\,x. \tag{4.33}$$

Tato rovnice vznikla vynásobením obou stran rovnice (4.5) funkcí $v \in \mathscr{D}_L$ a integrací v mezích od a do b. Vzorec (4.33) tedy vyjadřuje splnění rovnice (4.5) v integrálním, tj. slabém smyslu.

Závěrem uveďme existenční větu pro takto přeformulovanou okrajovou úlohu.

Věta 4.2. *Buď $\mathscr{H}$ reálný Hilbertův prostor se skalárním součinem (u, v) a normou $\|u\|$. Buď dále $[u, v]$ symetrická bilineární forma definovaná na vektorovém prostoru $\mathscr{D}_L \subset \mathscr{H}$, která je pozitivně definitní, tj. pro niž existuje konstanta $\gamma > 0$ taková, že platí*

$$[u, u] \geqq \gamma(u, u), \quad u \in \mathscr{D}_L, \tag{4.34}$$

takže $[u, v]$ je skalární součin na $\mathscr{D}_L$. Buď konečně $\mathscr{D}$ úplný obal prostoru $\mathscr{D}_L$ v normě $\|u\|_L = [u, u]^{1/2}$ a f libovolný prvek z $\mathscr{H}$. Pak existuje právě jeden prvek $y \in \mathscr{D}$ takový, že platí

$$F(y) = \min_{u \in \mathscr{D}} F(u), \tag{4.35}$$

kde funkcionál F je definován rovnicí (4.31).

D ů k a z . Za uvedených předpokladů je výraz (f, u) lineární ohraničený funkcionál na Hilbertově prostoru $\mathscr{D}$, neboť ze Schwarzovy nerovnosti a z nerovnosti (4.34) plyne, že je

$$|(f, u)| \leqq \|f\|\,\|u\| \leqq \frac{1}{\gamma^{1/2}}\|f\|\,\|u\|_L. \tag{4.36}$$

Podle Rieszovy věty o reprezentaci lineárního spojitého funkcionálu existuje tedy právě jeden prvek $y \in \mathscr{D}$ takový, že platí

$$(f, u) = [y, u], \quad u \in \mathscr{D}. \tag{4.37}$$

Dosadíme-li do (4.31) podle (4.37), máme

$$F(u) = [u,u] - 2[y,u] = [u,u] - 2[y,u] + [y,y] - [y,y] = \tag{4.38}$$
$$= [u-y, u-y] - [y,y].$$

Protože však je $[u-y, u-y] > 0$ pro $u \neq y$ podle (4.34), plyne požadované tvrzení z identity (4.38).

Poznámka 4.1. Rovnice (4.37), kterou jsme odvodili v průběhu důkazu věty 4.2, je totožná s rovnicí (4.32). Zobecněné řešení je tedy skutečně zároveň slabé řešení úlohy charakterizované bilineární formou $[u,v]$.

V předešlém textu jsme ukázali, jak lze definovat bilineární formu z věty 4.2 v případě okrajové úlohy (1.12), (4.3). V dalším odstavci ukážeme, že i v případě diferenciálních rovnic vyšších řádů je postup principiálně stejný.

4.1.2 Lineární diferenciální rovnice vyšších řádů

Uvažujme nejprve okrajovou úlohu (3.206), (1.15). Ze stejných důvodů, jako jsme uvedli v odst. 4.1.1, můžeme předpokládat, že dané okrajové podmínky jsou homogenní, tj. že jsou tvaru

$$y(a) = 0, \quad y'(a) = 0, \tag{4.39}$$
$$y(b) = 0, \quad y'(b) = 0.$$

Kdyby totiž tomu tak nebylo, sestrojíme snadno polynom třetího stupně, který splňuje podmínky (1.15) a pomocí něj převedeme řešení dané okrajové úlohy s nehomogenními okrajovými podmínkami na řešení okrajové úlohy s homogenními okrajovými podmínkami a jinou pravou stranou.

Značí-li nyní $\mathscr{D}_L$ množinu funkcí čtyřikrát spojitě diferencovatelných v intervalu $\langle a,b\rangle$ a splňujících okrajové podmínky (4.39) a označíme-li L operátor definovaný na množině $\mathscr{D}_L$ rovnicí

$$Lv = (pv'')'' - (sv')' + qv, \tag{4.40}$$

lze okrajovou úlohu (3.206), (4.39) zapsat jako rovnici

$$Lu = f, \tag{4.41}$$

jejíž řešení hledáme v množině $\mathscr{D}_L$. Zavedeme-li v množině $\mathscr{D}_L$ skalární součin $[u,v]$ rovnicí

$$[u,v] = \int_a^b (pu''v'' + su'v' + quv)\,\mathrm{d}\,x \tag{4.42}$$

a funkcionál F opět rovnicí (4.31), je úloha (4.41) ekvivalentní úloze nalezení minima funkcionálu F v množině $\mathscr{D}$, která vznikne zúplněním vektorového prostoru $\mathscr{D}_L$ v normě dané skalárním součinem (4.42). Množina $\mathscr{D}$ je tvořena právě těmi funkcemi ze Sobolevova prostoru $\mathscr{H}^2$, které splňují okrajové podmínky (4.39). Úlohu

(4.41) lze opět jako v případě rovnice druhého řádu zobecnit na případ, že koeficienty p, s a q jsou omezené nezáporné měřitelné funkce, že platí $p(x) \geqq p_0 > 0$ a že pravá strana f je z prostoru $\mathscr{L}_2$. Stačí k tomu pokládat rovnici (4.42) za definici bilineární formy z věty 4.2 a za řešení vzít takovou funkci $y \in \mathscr{D}$, pro niž funkcionál (4.31) nabývá minima nebo pro niž platí rovnice $[y,v] = (f,v)$ pro každou funkci $v \in \mathscr{D}_L$.

Popsaný postup lze snadno zobecnit na okrajovou úlohu $2m$-tého řádu, kde m je přirozené číslo. Buď dáno $m+1$ nezáporných měřitelných omezených funkcí a_j $(j = 0,\ldots,m)$ a nechť platí $a_m(x) \geqq \alpha > 0$. Označme $\mathscr{D}_L$ množinu funkcí, které jsou v intervalu $\langle a,b\rangle$ $2m$-krát spojitě diferencovatelné a které splňují homogenní okrajové podmínky

$$y^{(i)}(a) = 0, \quad i = 0,\ldots,m-1, \tag{4.43}$$
$$y^{(i)}(b) = 0, \quad i = 0,\ldots,m-1.$$

Definujme na množině $\mathscr{D}_L$ bilineární formu $[u,v]$ rovnicí

$$[u,v] = \sum_{j=0}^{m} \int_a^b a_j(x) u^{(j)}(x) v^{(j)}(x)\,\mathrm{d}\,x. \tag{4.44}$$

Za uvedených předpokladů je tato bilineární forma skalární součin a Hilbertův prostor $\mathscr{D}$, který vznikne zúplněním množiny $\mathscr{D}_L$ v normě dané tímto skalárním součinem, je tvořen právě těmi prvky Sobolevova prostoru $\mathscr{H}^m$, které splňují okrajové podmínky (4.43). Zobecněná okrajová úloha příslušná k bilineární formě (4.44) je pak stejně jako výše úloha nalézt takový prvek y prostoru $\mathscr{D}$, který minimalizuje funkcionál (4.31). Podle věty 4.2 řešení této úlohy existuje a je jediné pro každou pravou stranu $f \in \mathscr{L}_2$. Jsou-li koeficienty $a_j(x)$ dostatečně hladké a pravá strana f je spojitá, leží toto zobecněné řešení dokonce v $\mathscr{D}_L$; je tedy klasickým řešením diferenciální rovnice

$$(Ly)(x) \equiv \sum_{j=0}^{m} (-1)^j [a_j(x) y^{(j)}(x)]^{(j)} = f(x), \quad x \in \langle a,b\rangle \tag{4.45}$$

s okrajovými podmínkami (4.43).

4.1.3 Jiné typy okrajových podmínek

Variační formulace okrajových úloh není vázána pouze na okrajové podmínky typu (4.3) nebo (4.43). Je-li např. dána okrajová úloha (1.12), (1.13), můžeme stejně jako v odst. 4.1.1 bez újmy na obecnosti předpokládat, že okrajové podmínky jsou homogenní, tj. že jsou tvaru

$$-\alpha_1 p(a) y'(a) + \beta_1 y(a) = 0 \tag{4.46}$$
$$\alpha_2 p(b) y'(b) + \beta_2 y(b) = 0.$$

V opačném případě snadno sestrojíme funkci l, která splňuje okrajové podmínky (1.13); funkce $v = y - l$ pak splňuje diferenciální rovnici typu (1.12) s homogenními okrajovými podmínkami.

Předpokládáme-li, že je $\alpha_1 > 0$ a $\alpha_2 > 0$, je okrajová úloha (1.12), (4.46) ekvivalentní s úlohou nalézt v prostoru $\mathscr{H}^1$ minimum funkcionálu (4.31), kde bilineární forma $[u, v]$ je dána rovnicí

$$[u,v] = \int_a^b [p(x)u'(x)v'(x) + q(x)u(x)v(x)]\,\mathrm{d}\,x + \frac{\beta_1}{\alpha_1}u(a)v(a) + \frac{\beta_2}{\alpha_2}u(b)v(b). \tag{4.47}$$

Protože pravá strana vzorce (4.47) vznikla z výrazu

$$\int_a^b \{-[p(x)u'(x)]'v(x) + q(x)u(x)v(x)\}\,\mathrm{d}\,x \tag{4.48}$$

integrací per partes za využití podmínek (4.45), je také naznačeno, jak se postupuje v případě jiných okrajových úloh.

4.2 Základní přibližné metody

4.2.1 Ritzova metoda

V odstavci 4.1 jsme ukázali, že řešení mnoha okrajových úloh pro diferenciální rovnice je ekvivalentní s hledáním minima jistého funkcionálu. Tak např. řešit lineární diferenciální rovnici (1.12) s okrajovými podmínkami (4.3) je totéž jako nalézt minimum funkcionálu F daného rovnicí (4.14) na prostoru funkcí $\mathscr{D}$, které patří do Sobolevova prostoru $\mathscr{H}^1$ a které splňují okrajové podmínky (4.3). Princip Ritzovy metody spočívá v tom, že se zvolí konečnědimenzionální podprostor $\mathscr{D}_N$ prostoru $\mathscr{D}$ a za aproximaci řešení dané okrajové úlohy se pokládá funkce y_N, která minimalizuje funkcionál F na tomto konečnědimenzionálním prostoru.

Tvoří-li funkce $\Phi_1, \ldots, \Phi_N$ bázi prostoru $\mathscr{D}_N$ (funkce $\Phi_1, \ldots, \Phi_N$ jsou tedy takové lineárně nezávislé funkce z prostoru $\mathscr{D}_N$, že každá jiná funkce z tohoto prostoru se dá psát jako jejich lineární kombinace), je hledaná aproximace tvaru

$$y_N = \sum_{k=1}^{N} c_k^* \Phi_k. \tag{4.49}$$

Koeficienty $c^* = (c_1^*, \ldots, c_N^*)^T$ je přitom třeba určit tak, aby funkce F_N proměnných $c_1, \ldots, c_N$ definovaná rovnicí

$$F_N(c_1, \ldots, c_N) = F\left(\sum_{k=1}^{N} c_k \Phi_k\right) \tag{4.50}$$

nabývala v bodě $(c_1^*, \ldots, c_N^*)$ svého minima. Protože zřejmě platí

$$F_N(c_1, \ldots, c_N) = \sum_{i=1}^{N}\sum_{k=1}^{N} c_i c_k [\Phi_i, \Phi_k] - 2\sum_{i=1}^{N} c_i (f, \Phi_i), \tag{4.51}$$

je funkce F_N jako funkce proměnných $c_1, \ldots, c_N$ diferencovatelná. Pokud tedy funkce F_N nabývá v nějakém bodě extrému, musí v tomto bodě platit rovnice

$$\frac{\partial F_N}{\partial c_j} = 0, \quad j = 1, \ldots, N. \tag{4.52}$$

Dosadíme-li však do rovnic (4.52) podle definice funkce F_N, dostáváme, že tyto rovnice představují soustavu lineárních algebraických rovnic

$$Ac = g, \tag{4.53}$$

kde A je čtvercová matice řádu N, jejíž prvky a_{ij} jsou dány rovnicemi

$$a_{ij} = [\Phi_i, \Phi_j], \quad i, j = 1, \ldots, N, \tag{4.54}$$

a g je N-dimenzionální vektor o složkách g_i daných vzorci

$$g_i = (f, \Phi_i), \quad i = 1, \ldots, N. \tag{4.55}$$

Soustava (4.53) tedy představuje nutné podmínky pro existenci extrému funkcionálu F na podprostoru $\mathscr{D}_N$. V následující větě dokážeme, že tyto podmínky jsou také postačující.

Věta 4.3. *Nechť jsou splněny předpoklady věty 4.2, nechť $\mathscr{D}_N$ je konečnědimenzionální podprostor prostoru $\mathscr{D}$ a nechť funkce $\Phi_1, \ldots, \Phi_N$ tvoří jeho bázi. Pak soustava* (4.53) *má při libovolném vektoru g právě jedno řešení a navíc toto řešení minimalizuje funkci F_N definovanou vzorcem* (4.50).

D ů k a z . Matice A, jejíž prvky jsou dány rovnicemi (4.54), je zřejmě symetrická. Dokážeme, že je také pozitivně definitní. Buď tedy $c = (c_1, \ldots, c_N)^T$ libovolný nenulový vektor, takže funkce $u = \sum_{i=1}^{N} c_i \Phi_i$ představuje nenulový prvek prostoru $\mathscr{D}_N$. Je tedy

$$c^T A c = \sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j [\Phi_i, \Phi_j] = [u, u] > 0, \tag{4.56}$$

neboť bilineární forma $[u, v]$ je pozitivně definitní.

Soustava (4.53) má tedy právě jedno řešení c^*. Protože je $F_N(c) = c^T A c - 2g^T c$, je

$$\begin{aligned} F_N(c) &= c^T A c - 2(c^*)^T A c = \\ &= c^T A c - 2(c^*)^T A c + (c^*)^T A c^* - (c^*)^T A c^* = \\ &= (c - c^*)^T A (c - c^*) - (c^*)^T A c^* > -(c^*)^T A c^* = F_N(c^*) \end{aligned} \tag{4.57}$$

pro $c \neq c^*$. Tato nerovnost však dokazuje, že funkce F_N skutečně nabývá v bodě $c = c^*$ svého minima. Věta je dokázána.

Vypočíst Ritzovou metodou přibližné řešení okrajové úlohy charakterizované bilineární formou $[u, v]$ a pravou stranou f značí tedy utvořit funkci y_N tvaru (4.49), kde vektor c^* je řešením soustavy lineárních algebraických rovnic (4.53). Matice této soustavy se v matematické literatuře nazývá *Gramova matice*. V technické literatuře, zejména pak v souvislosti s metodou konečných prvků, se matice A nazývá většinou *matice tuhosti* příslušná k bázi $\Phi_1, \ldots, \Phi_N$ a vektor g na pravé straně soustavy (4.53) se nazývá *vektor zatížení*.

Abychom mohli posoudit velikost chyby, dokážeme ještě následující větu.

Věta 4.4. *Nechť jsou splněny předpoklady věty 4.3. Pak funkce y_N tvaru* (4.49), *kde vektor c^* je řešením soustavy* (4.53), *je ortogonální projekcí v prostoru $\mathscr{D}$ přesného řešení dané okrajové úlohy do podprostoru $\mathscr{D}_N$.*

D ů k a z . Buď y přesné řešení okrajové úlohy charakterizované bilineární formou $[u, v]$ a funkcí f. Z identity (4.38) platné pro libovolný prvek $u \in \mathscr{D}$ plyne speciálně, že je

$$\|y_N - y\|_L^2 = [y_N - y, y_N - y] = F(y_N) + [y, y] = \tag{4.58}$$
$$= \min_{u \in \mathscr{D}_N} F(u) + [y, y] = \min_{u \in \mathscr{D}_N} [u - y, u - y] = \min_{u \in \mathscr{D}_N} \|u - y\|_L^2.$$

Přibližné řešení y_N je tedy prvek podprostoru $\mathscr{D}_N$, který leží nejblíže přesnému řešení. Odtud a ze známých vět o nejlepší aproximaci v Hilbertově prostoru už tvrzení plyne.

Ze vzorce (4.58) vyplývá, že každý horní odhad výrazu $\min_{u \in \mathscr{D}_N} \|u - y\|_L$ dává odhad chyby přibližného řešení získaného Ritzovou metodou. Chyba Ritzovy metody tedy závisí na tom, jak dobře jsme schopni aproximovat přesné řešení prvky konečnědimenzionálního podprostoru $\mathscr{D}_N$. Volba prostorů $\mathscr{D}_N$ a bázových funkcí $\Phi_1, \ldots, \Phi_N$ má tedy zřejmě při užití Ritzovy metody zásadní důležitost. V odst. 4.3 popíšeme jednu zejména v poslední době populární strategii pro systematickou konstrukci bázových funkcí. Tato metoda je známá pod názvem metoda konečných prvků a má pro efektivní užití variačních metod prvořadou důležitost.

4.2.2 Galerkinova metoda

Základní myšlenka této metody je ještě jednodušší, než je tomu u Ritzovy metody. Vychází z toho, že zobecněné řešení okrajové úlohy se dá definovat rovnicí (4.32). Opět, jako u Ritzovy metody, se zvolí konečnědimenzionální podprostor $\mathscr{D}_N$ prostoru $\mathscr{D}$ a za přibližné řešení se pokládá taková funkce $y_N \in \mathscr{D}_N$, pro niž platí

$$[y_N, v_N] = (f, v_N) \tag{4.59}$$

pro každou funkci $v_N \in \mathscr{D}_N$. Tvoří-li funkce $\Phi_1, \ldots, \Phi_N$ bázi v prostoru $\mathscr{D}_N$, je přibližné řešení y_N dáno úplně stejnými vzorci jako u Ritzovy metody.

V případě, že okrajovou úlohu lze interpretovat jako minimalizační úlohu, jsou tedy metody Ritzova a Galerkinova totožné. Proto se často v této souvislosti mluví o Ritzově-Galerkinově metodě. Galerkinova metoda je však obecnější než Ritzova metoda, neboť je ji možné užít i v případě zobecněné úlohy typu (4.32), kde výraz $[u, v]$ je bilineární forma, která nemusí být nutně symetrická. V tomto případě obecně nelze udat funkcionál typu (4.31), jehož extrém by bylo možno interpretovat jako zobecněné řešení. Úloha (4.32) však přesto může mít rozumný smysl.

O volbě bázových funkcí Galerkinovy metody platí samozřejmě totéž, co bylo řečeno u Ritzovy metody.

4.3 Metoda konečných prvků

Při popisu Ritzovy a Galerkinovy metody v odst. 4.2 jsme neudali žádný systematický způsob konstrukce konečnědimenzionálních podprostorů $\mathscr{D}_N$ prostoru $\mathscr{D}$. Této otázky si všimneme teprve v tomto odstavci, a proto objasníme nejprve, jaká hlediska je přitom záhodno respektovat. Pro rozdíl přibližného řešení y_N získaného Ritzovou-Galerkinovou metodou a přesného řešení y platí nerovnost

$$\|y_N - y\|_L \leqq \|u - y\|_L, \tag{4.60}$$

kde u je libovolný prvek z prostoru $\mathscr{D}_N$ (viz věta 4.4). Podtrhněme, že index L v nerovnosti (4.60) značí, že jde o tzv. energetickou normu, tj. o normu definovanou pomocí bilineární formy $[u, v]$. Pro konkrétní příklady bilineárních forem uvedené v odst. 4.1.1 a 4.1.2 je však známo, že příslušné energetické normy jsou ekvivalentní s normami ve vhodných Sobolevových prostorech. Místo nerovnosti (4.60) lze tedy eventuálně užít nerovnost

$$\|y_N - y\|_k \leqq M \|u - y\|_k, \tag{4.61}$$

kde M je konstanta, která nezávisí na volbě podprostoru $\mathscr{D}_N$ a index k značí, že jde o normu v Sobolevově prostoru $\mathscr{H}^k$.

Otázka přesnosti Ritzovy aproximace je tedy v podstatě otázkou, jak přesně lze aproximovat libovolnou funkci z prostoru $\mathscr{D}$ funkcí z konečnědimenzionálního podprostoru $\mathscr{D}_N$. První hledisko, kterým je rozumné řídit se při konstrukci prostoru $\mathscr{D}_N$, je proto požadavek co nejlepší aproximovatelnosti funkcí z prostoru $\mathscr{D}$ funkcemi z prostoru $\mathscr{D}_N$.

Druhé, neméně důležité hledisko je efektivnost vzniklého algoritmu. Podstatnou část tohoto algoritmu tvoří řešení soustavy obecně velice mnoha lineárních rovnic (4.53). Je-li matice této soustavy plná (tj. má-li všechny nebo alespoň převážnou většinu prvků různých od nuly), mohou vznikat dosti závažné problémy s pamětí počítače a s počtem prováděných operací. Tyto problémy se velmi podstatně zmírní, je-li zmíněná matice řídká (tj. má-li pouze $O(N)$ nenulových prvků), nebo je-li navíc dokonce pásová.

Je tedy žádoucí řídit se při konstrukci prostoru $\mathscr{D}_N$ a zejména jeho báze i tímto hlediskem. Řídkosti matice soustavy (4.53) se obvykle dosáhne tím způsobem, že bázové funkce se volí tak, aby měly malé nosiče (tj. množiny bodů, v nichž jsou od nuly různé) ve srovnání s intervalem $\langle a, b\rangle$.

Je také důležité, aby všechny výpočty potřebné k sestavení konečnědimenzionálního problému, tj. výpočty potřebné k určení prvků matice A a složek vektoru g, byly snadno proveditelné. S touto otázkou souvisí také to, že je žádoucí, aby koeficienty c_k nebo aspoň některé z nich měly význam hodnot hledaného řešení, resp. jeho derivací v některých bodech intervalu $\langle a, b\rangle$, aby nebylo nutné ještě navíc počítat součet (4.49).

Všechny tyto požadavky do značné míry splňuje *metoda konečných prvků*, která dává dostatečně obecný a systematický návod pro konstrukci bázových funkcí pro Ritzovu-Galerkinovu aproximaci řešení okrajových úloh. Její základní myšlenka vychází z polynomiální interpolace, nikoliv však jedním polynomem na celém intervalu, ale různými polynomy na podintervalech, na které daný interval rozdělíme. Příslušné aproximace jsou pak na jednotlivých podintervalech rovny polynomům nízkých stupňů.

Při konstrukci podprostoru $\mathscr{D}_N$ prostoru $\mathscr{D}$ budeme tedy postupovat takto: Daný interval $\langle a, b\rangle$ rozdělíme body (zvanými nejčastěji *uzly*) na podintervaly, které mají společné jen tyto uzly. V uzlech budeme zadávat tzv. *uzlové parametry*, tj. hodnoty polynomu a jeho derivací. Protože některé uzly jsou společné pro dva podintervaly rozkladu, užijí se hodnoty odpovídajících uzlových parametrů pro interpolaci na obou těchto sousedních podintervalech. Podinterval rozkladu, na něm definované uzlové parametry a příslušný interpolační polynom nazveme *konečným prvkem*. Souhrn všech funkcí, které jsou polynomiální na jednotlivých podintervalech rozkladu, mají na sousedních prvcích společné uzlové parametry a jsou dostatečně hladké, tvoří při všech možných hodnotách uzlových parametrů konečnědimenzionální podprostor $\mathscr{D}_h$ Sobolevova prostoru $\mathscr{H}^k$. O podprostoru $\mathscr{D}_h$ budeme hovořit jako o *prostoru konečných prvků*. Index h, který jsme užili místo dřívějšího indexu N, který udával dimenzi příslušného podprostoru, se rovná délce největšího podintervalu rozkladu a budeme jím měřit kvalitu aproximačních vlastností příslušného podprostoru.

V dalších odstavcích si všimneme několika speciálních případů této konstrukce podrobněji.

4.3.1 Aproximace po částech lineárními funkcemi

Rozdělme interval $\langle a, b\rangle$ na n podintervalů dělicími body x_i, pro něž platí

$$a = x_0 < x_1 < \ldots < x_n = b. \tag{4.62}$$

Je-li dáno $n+1$ čísel $c_0, \ldots, c_n$ a sestrojíme-li funkci u, která je na každém podin-

tervalu $\langle x_{i-1}, x_i\rangle$ lineární a pro níž platí

$$u(x_i) = c_i, \quad i = 0, \ldots, n, \tag{4.63}$$

je tato funkce spojitá a její derivace je v každém podintervalu (x_{i-1}, x_i) konstantní. Tato funkce zřejmě patří do Sobolevova prostoru $\mathscr{H}^1$ a kromě toho je parametry $c_0, \ldots, c_n$ jednoznačně určena. Vezmeme-li tedy krajní body každého podintervalu za uzly a hodnoty funkce za uzlové parametry, tvoří soustava všech funkcí tohoto typu prostor končených prvků, který je podprostorem prostoru $\mathscr{H}^1$. Abychom tuto skutečnost zdůraznili, označíme jej $\mathscr{D}_{h1}$. Je-li $f \in \mathscr{H}^1$, existuje právě jedna funkce $f_h \in \mathscr{D}_{h1}$ taková, že pro ni platí $f_h(x_i) = f(x_i)$ pro $i = 0, \ldots, n$.

Aproximační vlastnosti právě sestrojeného prostoru jsou popsány v následující větě.

Věta 4.5. *Nechť je $f \in \mathscr{H}^r$ pro $r = 1$ nebo 2. Pak existuje konstanta $M > 0$ tak, že pro chybu interpolace $f_h \in \mathscr{D}_{h1}$ platí odhady*

$$\|f - f_h\|_p \leqq M h^{r-p} \|f^{(r)}\|_0, \quad r = 1, 2,\ p = 0, \ldots, r-1, \tag{4.64}$$

kde

$$h = \max_{i=0,\ldots,n-1} (x_{i+1} - x_i) \tag{4.65}$$

a index 0 značí, že jde o normu v prostoru $\mathscr{L}_2$.

D ů k a z . Zvolme pevně interval $\langle x_i, x_{i+1}\rangle$, položme $x_{i+1} - x_i = h_i$ a buď nejprve $f \in \mathscr{H}^1$. Definujeme-li funkci φ proměnné ξ na intervalu $\langle 0, 1\rangle$ předpisem

$$\varphi(\xi) = f(h_i \xi + x_i), \tag{4.66}$$

je zřejmě $\varphi \in \mathscr{H}^1(0, 1)$. Definujme podobně funkci φ_h předpisem

$$\varphi_h(\xi) = f_h(h_i \xi + x_i). \tag{4.67}$$

Protože funkce f_h je lineární na intervalu $\langle x_i, x_{i+1}\rangle$, platí totéž pro funkci φ_h na intervalu $\langle 0, 1\rangle$; je tedy

$$\varphi_h(\xi) = \varphi(0)(1 - \xi) + \varphi(1)\xi. \tag{4.68}$$

Položíme-li konečně

$$K(\xi, \tau) = \begin{cases} 1 - \xi, & \tau < \xi, \\ -\xi, & \tau > \xi, \end{cases} \tag{4.69}$$

snadno vypočteme, že platí

$$\varphi(\xi) - \varphi_h(\xi) = \int_0^1 K(\xi, \tau) \varphi'(\tau)\, \mathrm{d}\tau. \tag{4.70}$$

Odtud pomocí Schwarzovy nerovnosti snadno odvodíme, že je

$$[\varphi(\xi)-\varphi_h(\xi)]^2 \leqq \int_0^1 K^2(\xi,\tau)\,\mathrm{d}\tau \int_0^1 [\varphi'(\tau)]^2\,\mathrm{d}\tau. \tag{4.71}$$

Integrací obou stran nerovnosti (4.71) podle ξ v mezích od $\xi = 0$ do $\xi = 1$ dostáváme odhad

$$\int_0^1 [\varphi(\xi)-\varphi_h(\xi)]^2\,\mathrm{d}\xi \leqq \gamma \int_0^1 [\varphi'(\tau)]^2\,\mathrm{d}\tau, \tag{4.72}$$

kde

$$\gamma = \int_0^1 \int_0^1 K^2(\xi,\tau)\,\mathrm{d}\xi\,\mathrm{d}\tau, \tag{4.73}$$

a je to tedy absolutní konstanta.

Dosaďme konečně do odhadu (4.72) za funkce φ a φ_h z rovnic (4.66) a (4.67). Protože je zřejmě $\varphi'(\tau) = h_i f'(h_i\tau + x_i)$, je

$$\int_0^1 [f(h_i\xi + x_i) - f_h(h_i\xi + x_i)]^2\,\mathrm{d}\xi \leqq \gamma h_i^2 \int_0^1 [f'(h_i\tau + x_i)]^2\,\mathrm{d}\tau. \tag{4.74}$$

Provedeme-li v integrálech na levé i pravé straně této nerovnosti substituce $h_i\xi + x_i = x$, resp. $h_i\tau + x_i = x$, dostaneme

$$\int_{x_i}^{x_{i+1}} [f(x) - f_h(x)]^2\,\mathrm{d}x \leqq \gamma h_i^2 \int_{x_i}^{x_{i+1}} [f'(x)]^2\,\mathrm{d}x \tag{4.75}$$

(krátili jsme činitelem $1/h_i$). Sečtením přes všechny intervaly dělení a použitím odhadu $h_i \leqq h$ však už dostaneme odhad (4.64) pro $r = 1$.

Předpokládejme nyní, že je $f \in \mathscr{H}^2$. Pak platí totéž i pro funkci φ a ve vyjádření (4.70) můžeme integrovat per partes. Potřebnou primitivní funkci jádra K vzhledem k proměnné τ označíme K_1 a zvolíme ji ve tvaru

$$K_1(\xi,\tau) = \int_0^\tau K(\xi,\eta)\,\mathrm{d}\eta. \tag{4.76}$$

Zřejmě je $K_1(\xi,0) = 0$. Položíme-li však ve vzorci (4.70) speciálně $\varphi(\xi) = \xi$, zjistíme, že je i $K_1(\xi,1) = 0$, neboť v tomto případě je také $\varphi_h(\xi) = \xi$. Zmíněná integrace per partes tedy dá vzorec

$$\varphi(\xi) - \varphi_h(\xi) = \int_0^1 K_1(\xi,\tau)\varphi''(\tau)\,\mathrm{d}\tau. \tag{4.77}$$

Odtud úplně stejným postupem jako výše dostaneme, že platí

$$\int_0^1 [\varphi(\xi) - \varphi_h(\xi)]^2\,\mathrm{d}\xi \leqq \gamma_1 \int_0^1 [\varphi''(\tau)]^2\,\mathrm{d}\tau, \tag{4.78}$$

kde

$$\gamma_1 = \int_0^1 \int_0^1 K_1^2(\xi,\tau)\,\mathrm{d}\xi\,\mathrm{d}\tau, \tag{4.79}$$

což je opět absolutní konstanta. Dosadíme-li však do nerovnosti (4.78) za φ a φ_h podle vzorců (4.66) a (4.67), a provedeme-li substituci $h_i\xi + x_i = x$, resp. $\tau_i\xi + x_i = x$, dostaneme nerovnost

$$\int_{x_i}^{x_{i+1}} [f(x) - f_h(x)]^2\,\mathrm{d}x \leqq \gamma_1 h_i^4 \int_{x_i}^{x_{i+1}} [f''(x)]^2\,\mathrm{d}x. \tag{4.80}$$

Odtud už bezprostředně plyne platnost nerovnosti (4.64) s $r = 2$ a $p = 0$.

Abychom dokázali zbývající případ, derivujme identitu (4.77) podle ξ. Dostaneme

$$\varphi'(\xi) - \varphi_h'(\xi) = \int_0^1 K_2(\xi,\tau)\varphi''(\tau)\,\mathrm{d}\tau, \tag{4.81}$$

kde $K_2(\xi,\tau) = \partial K_1(\xi,\tau)/\partial\xi$. Z vyjádření (4.81) zopakováním předešlého postupu dostaneme nerovnost

$$\int_{x_i}^{x_{i+1}} [f'(x) - f_h'(x)]^2\,\mathrm{d}x \leqq \gamma_2 h_i^2 \int_{x_i}^{x_{i+1}} [f''(x)]^2\,\mathrm{d}x, \tag{4.82}$$

kde γ_2 je absolutní konstanta (rovná integrálu z kvadrátu jádra K_2). Protože je $\|f - f_h\|_1^2 = \|f - f_h\|_0^2 + \|f' - f_h'\|_0^2$, plyne tvrzení věty pro $r = 2$ a $p = 1$ z nerovností (4.80) a (4.82). Věta je dokázána.

Na pravé straně nerovností (4.64) jsme stejně dobře mohli psát normy funkce f v prostoru $\mathscr{H}^1$ nebo $\mathscr{H}^2$, neboť zřejmě platí $\|f'\|_0 \leqq \|f\|_1$ a $\|f''\|_0 \leqq \|f\|_2$. Někdy však může být užitečné, že na pravé straně těchto nerovností stačí brát pouze normu nejvyšších derivací. Poznamenejme v této souvislosti, že $\mathscr{L}_2$ norma k-té derivace funkce $f \in \mathscr{H}^k$ se nazývá k-tá *seminorma* této funkce a značí se $|f|_k$.

Poznámka 4.2. Přibližné řešení modelové okrajové úlohy (1.12), (4.3) získané Ritzovou metodou v prostoru $\mathscr{D}_{h1}^0$ těch funkcí z $\mathscr{D}_{h1}$, které splňují podmínky (4.3), konverguje při $h \to 0$ k přesnému řešení.

Tvrzení poznámky plyne ihned z nerovnosti (4.61), v níž je $k = 1$, y_N je přibližné řešení, y přesné řešení a u je libovolný prvek z prostoru $\mathscr{D}_{h1}^0$. Pravou stranu této nerovnosti lze totiž majorizovat výrazem $M(\|y_\varepsilon - y\|_1 + \|y_{\varepsilon h} - y_\varepsilon\|_1)$, kde $y_\varepsilon \in \mathscr{H}^2$ a $y_{\varepsilon h}$ je interpolace funkce y_ε. Normu prvního rozdílu lze učinit libovolně malou, neboť prostor $\mathscr{H}^2$ je hustý v prostoru $\mathscr{H}^1$ a norma druhého rozdílu je malá v důsledku věty 4.5. O rychlosti konvergence se nedá obecně nic říci; v případě, že je $y \in \mathscr{H}^2$, je rychlost konvergence (v normě prostoru $\mathscr{H}^1$) ovšem $O(h)$.

Sestrojme ještě bázi v prostoru $\mathscr{D}_{h1}$. Tento prostor má dimenzi rovnou zřejmě počtu uzlových parametrů, tj. číslu $n+1$. Jednu z možných (a jak hned uvidíme také vhodných) bází tvoří soustava funkcí Φ_j, které vzniknou tak, že j-tý uzlový

parametr položíme rovný jedné a ostatní uzlové parametry rovné nule. Provedeme-li naznačený postup, jsou prvky báze dány rovnicemi

$$\Phi_0(x) = \begin{cases} \frac{x_1 - x}{x_1 - x_0}, & x_0 \leqq x \leqq x_1, \\ 0, & x_1 \leqq x \leqq x_n, \end{cases} \tag{4.83}$$

$$\Phi_i(x) = \begin{cases} 0, & x_0 \leqq x \leqq x_{i-1}, \\ \frac{x - x_{i-1}}{x_i - x_{i-1}}, & x_{i-1} \leqq x \leqq x_i, \\ \frac{x_{i+1} - x}{x_{i+1} - x_i}, & x_i \leqq x \leqq x_{i+1}, \\ 0, & x_{i+1} \leqq x \leqq x_n, \end{cases} \quad i = 1, \dots, n-1,$$

$$\Phi_n(x) = \begin{cases} 0, & x_0 \leqq x \leqq x_{n-1}, \\ \frac{x - x_{n-1}}{x_n - x_{n-1}}, & x_{n-1} \leqq x \leqq x_n. \end{cases}$$

Chceme-li sestrojit prostor konečných prvků, který je podprostorem prostoru $\mathscr{D}$ těch funkcí z prostoru $\mathscr{H}^1$, jež splňují okrajové podmínky (4.3), stačí vzít $(n-1)$-dimenzionální prostor $\mathscr{D}_{h1}^0$, který je lineární obal funkcí $\Phi_1, \dots, \Phi_{n-1}$, neboť uzlové parametry pro uzly x_0 a x_n jsou trvale rovny nule. Užijeme-li tuto bázi pro přibližné řešení okrajové úlohy (1.12), (4.3), je matice A v soustavě lineárních rovnic (4.53) třídiagonální, neboť nosiče bázových funkcí Φ_i a Φ_j jsou pro $|i-j| \geqq 2$ disjunktní, takže integrály ve vzorci (4.27) jsou pro tato i a j rovny nule. Složky vektoru řešení soustavy (4.53) mají přitom přímo význam hodnot přibližného řešení, neboť každá bázová funkce je rovna jedné v právě jednom uzlu a v ostatních uzlech je rovna nule. Přibližné řešení modelové úlohy (1.12), (4.3) Ritzovou-Galerkinovou metodou při užití právě sestrojeného prostoru konečných prvků vede tedy na řešení soustavy lineárních algebraických rovnic s třídiagonální maticí podobně jako v metodě sítí. Kdybychom k přibližnému výpočtu integrálů, které je nutno počítat při výpočtu prvků matice soustavy (4.53), zvolili vhodné kvadraturní vzorce, byly by dokonce obě tyto metody totožné. Metodu sítí lze tedy uvést do velmi úzké souvislosti s variačními metodami.

Stejným způsobem jako výše lze užít metodu konečných prvků užívající lineární prvky $\mathscr{D}_{h1}$ i k řešení jiných okrajových úloh druhého řádu. Předpoklad, že problém je druhého řádu, je zde podstatný, neboť prostor $\mathscr{D}_{h1}$ je podprostorem Sobolevova prostoru $\mathscr{H}^1$, v němž je přirozené hledat řešení rovnic druhého řádu. Protože první derivace funkcí z $\mathscr{D}_{h1}$ nejsou obecně spojité, tím méně absolutně spojité, není tento prostor podprostorem žádného prostoru $\mathscr{H}^k$ pro $k > 1$; k řešení úloh vyššího řádu jej tedy nelze užít.

4.3.2 Aproximace Hermitova typu

Jedna z možností konstrukce prostorů konečných prvků, které jsou podprostory Sobolevova prostoru $\mathscr{H}^k$ pro libovolné k, vychází z myšlenky Hermitovy interpolace. Z teorie této interpolace plyne, že jsou-li c_{ij}, $i = 0, \dots, n$, $j = 0, \dots, k-1$, libovolná čísla, existuje právě jedna funkce u, která je na každém podintervalu $\langle x_{i-1}, x_i \rangle$

polynomem stupně nejvýše $2k-1$ a pro niž platí

$$u^{(j)}(x_i) = c_{ij}, \quad i = 0, \dots, n, \ j = 0, \dots, k-1. \tag{4.84}$$

Každá taková funkce má spojité derivace až do řádu $k-1$ a k-tá derivace má nespojitosti pouze v uzlech x_i; patří tedy do prostoru $\mathscr{H}^k$. Vezmeme-li krajní body podintervalů, na něž jsme rozdělili interval $\langle a, b \rangle$, za uzly a hodnoty funkce spolu s hodnotami jejích derivací až do řádu $k-1$ za uzlové parametry, dostaneme prostor konečných prvků $\mathscr{D}_{hk} \subset \mathscr{H}^k$. Každou funkci $f \in \mathscr{H}^k$ můžeme opět jako v odst. 4.3.1 aproximovat funkcí $f_h \in \mathscr{D}_{hk}$, která je jednoznačně určena rovnicemi

$$f_h^{(j)}(x_i) = f^{(j)}(x_i), \quad i = 0, \dots, n, \ j = 0, \dots, k-1. \tag{4.85}$$

V dalším se zejména pro zjednodušení zápisu omezíme na ekvidistantní uzly, tj. na uzly dané vzorcem

$$x_i = a + ih, \quad i = 0, \dots, n, \ h = (b-a)/n. \tag{4.86}$$

Následující věta popisuje aproximační vlastnosti prostoru $\mathscr{D}_{hk}$.

Věta 4.6. *Nechť je $f \in \mathscr{H}^r$ pro $r \in \{k, \dots, 2k\}$. Pak pro chybu aproximace f_h určené rovnicemi* (4.85) *platí odhad*

$$\|f - f_h\|_p \leqq M h^{r-p} \|f^{(r)}\|_0, \quad r = k, \dots, 2k, \ p = 0, \dots, \min(r-1, k). \tag{4.87}$$

Důkaz provedeme pouze pro případ $k = 2$. Budeme vycházet stejně jako v důkazu věty 4.5 z integrálního vyjádření chyby interpolace. Definujme funkce φ a φ_h opět rovnicemi (4.66) a (4.67). Funkce φ_h je nyní polynom třetího stupně a dá se psát ve tvaru

$$\begin{aligned} \varphi_h(\xi) = {} & \varphi(0)(1+2\xi)(1-\xi)^2 + \varphi(1)(3-2\xi)\xi^2 + \\ & + \varphi'(0)\xi(1-\xi)^2 + \varphi'(1)(\xi-1)\xi^2. \end{aligned} \tag{4.88}$$

Dosadíme-li do tohoto vzorce ze zřejmých identit

$$\begin{aligned} \varphi(0) &= \varphi(\xi) - \xi\varphi'(\xi) + \int_0^\xi \tau\varphi''(\tau)\,\mathrm{d}\tau, \\ \varphi(1) &= \varphi(\xi) + (1-\xi)\varphi'(\xi) + \int_\xi^1 (1-\tau)\varphi''(\tau)\,\mathrm{d}\tau, \\ \varphi'(0) &= \varphi'(\xi) - \int_0^\xi \varphi''(\tau)\,\mathrm{d}\tau, \\ \varphi'(1) &= \varphi'(\xi) + \int_\xi^1 \varphi''(\tau)\,\mathrm{d}\tau \end{aligned} \tag{4.89}$$

platných pro $\varphi \in \mathscr{H}^2(0,1)$, dostaneme po elementárních úpravách

$$\varphi_h(\xi) = \varphi(\xi) - \int_0^\xi [\xi(1-\xi)^2 - (1+2\xi)(1-\xi)^2\tau]\varphi''(\tau)\,\mathrm{d}\tau - \tag{4.90}$$
$$- \int_\xi^1 [(3-2\xi)\xi^2(\tau-1) + (1-\xi)\xi^2]\varphi''(\tau)\,\mathrm{d}\tau,$$

neboli

$$\varphi(\xi) - \varphi_h(\xi) = \int_0^1 K(\xi,\tau)\varphi''(\tau)\,\mathrm{d}\tau, \tag{4.91}$$

kde

$$K(\xi,\tau) = \begin{cases} (\xi-1)^2[\xi-(1+2\xi)\tau], & \tau < \xi, \\ \xi^2[1-\xi+(2\xi-3)(1-\tau)], & \tau > \xi. \end{cases} \tag{4.92}$$

Další postup je identický jako v důkazu věty 4.5, jen případů, které je třeba vyšetřit, je více. Z identity (4.91) ihned plyne odhad

$$\int_{x_i}^{x_{i+1}} [f(x) - f_h(x)]^2\,\mathrm{d}x \leqq \gamma_1 h^4 \int_{x_i}^{x_{i+1}} [f''(x)]^2\,\mathrm{d}x, \tag{4.93}$$

kde γ_1 je absolutní hodnota. Z nerovnosti (4.93) však už plyne tvrzení věty pro $r = 2$ a $p = 0$. Derivováním identity (4.91) odvodíme nerovnost

$$\int_{x_i}^{x_{i+1}} [f'(x) - f_h'(x)]^2\,\mathrm{d}x \leqq \gamma_2 h^2 \int_{x_i}^{x_{i+1}} [f''(x)]^2\,\mathrm{d}x, \tag{4.94}$$

která spolu s tím, co už jsme dokázali výše, dokazuje nerovnost (4.87) pro $r = 2$ a $p = 1$.

Abychom vyšetřili případ $f \in \mathscr{H}^3$, uvažme, že je $\varphi = \varphi_h$, pokud funkce φ je polynom třetího nebo nižšího stupně. Dosadíme-li tedy do identity (4.91) za φ funkci $\xi^2/2$, dostaneme, že je

$$\int_0^1 K(\xi,\tau)\,\mathrm{d}\tau = 0. \tag{4.95}$$

Primitivní funkci $K_1(\xi,\tau)$ k funkci $K(\xi,\tau)$ vzhledem k proměnné τ lze proto volit tak, že pro ni platí $K_1(\xi,0) = K(\xi,1) = 0$. Integrace per partes v pravé straně identity (4.91) pak dá vyjádření

$$\varphi(\xi) - \varphi_h(\xi) = \int_0^1 K_1(\xi,\tau)\varphi'''(\tau)\,\mathrm{d}\tau. \tag{4.96}$$

Odtud a z identit, které se dostanou dvojím derivováním této identity podle ξ, dostáváme tvrzení věty pro $r = 3$ a pro $p = 0, 1, 2$.

Konečně v případě, že je $f \in \mathscr{H}^4$, můžeme integrovat per partes ve vzorci (4.96).

Výsledné vyjádření bude tvaru

$$\varphi(\xi) - \varphi_h(\xi) = \int_0^1 K_2(\xi,\tau)\varphi''''(\tau)\,\mathrm{d}\tau, \tag{4.97}$$

protože primitivní funkci $K_2(\xi,\tau)$ k funkci $K_1(\xi,\tau)$ lze zvolit ze stejných důvodů jako výše tak, že pro ni platí $K_2(\xi,0) = K_2(\xi,1) = 0$ (tentokrát je třeba dosadit do identity (4.96) $\varphi(\xi) = \xi^3/6$). Dvojím derivováním identity (4.97) podle ξ dostaneme tvrzení věty pro $r = 4$ a $p = 0, 1, 2$. Důkaz je hotov.

Poznamenejme, že třetí derivování identity (4.97) dá odhad

$$\int_{x_i}^{x_{i+1}} [f'''(x) - f_h'''(x)]\,\mathrm{d}x \leqq \gamma h^2 \int_{x_i}^{x_{i+1}} [f''''(x)]^2\,\mathrm{d}x. \tag{4.98}$$

Tuto nerovnost však ke konstrukci odhadu chyby aproximace v prostoru $\mathscr{H}^3$ použít nelze, neboť funkce f_h obecně nemusí být vůbec prvkem tohoto prostoru.

Důkaz tvrzení věty 4.6 pro obecné k vychází z identity

$$\varphi(\xi) - \varphi_h(\xi) = \int_0^1 K(\xi,\tau)\varphi^{(k)}(\tau)\,\mathrm{d}\tau \tag{4.99}$$

platné pro $\varphi \in \mathscr{H}^k$. Odvození toto identity je však dosti komplikované, a proto jsme se omezili na uvedený speciální případ.

Prvky prostorů $\mathscr{D}_{hk}$ se nazývají *Hermitovy spline-funkce* stupně $2k-1$. Abychom mohli těchto prostorů efektivně využívat k variačnímu řešení okrajových úloh, je třeba v nich zavést vhodné báze. Všimněme si proto nyní této otázky. Začneme případem $k = 2$. V tomto případě je dimenze prostoru $\mathscr{D}_{h2}$ rovna $2(n+1)$, neboť v $n+1$ uzlech zadáváme vždy dva uzlové parametry, totiž hodnotu funkce a hodnotu její derivace. Je tedy třeba sestrojit $2(n+1)$ bázových funkcí $\Phi_0, \ldots, \Phi_{2n+1}$. Jedna z možností, jak je získat, je zcela analogická jako v případě aproximace po částech lineárními funkcemi. Funkci Φ_{2i}, $i = 0, \ldots, n$, sestrojíme tak, že uzlový parametr, který udává funkční hodnotu v i-tém uzlu, položíme rovný jedné a všechny ostatní uzlové parametry položíme rovny nule; u funkcí Φ_{2i+1}, $i = 0, \ldots, n$, to provedeme obráceně v tom smyslu, že uzlový parametr, který udává hodnotu první derivace v i-tém uzlu, položíme rovný jedné a všechny ostatní uzlové parametry rovny nule. Abychom mohli takto vzniklé bázové funkce snadno zapsat, položme

$$\varphi_0(x) = \begin{cases} 0, & x \leqq -1, \\ 1 - 3x^2 - 2x^3, & -1 \leqq x \leqq 0, \\ 1 - 3x^2 + 2x^3, & 0 \leqq x \leqq 1, \\ 0, & 1 \leqq x \end{cases} \tag{4.100}$$

a

$$\varphi_1(x) = \begin{cases} 0, & x \leqq -1, \\ x + 2x^2 + x^3, & -1 \leqq x \leqq 0, \\ x - 2x^2 + x^3, & 0 \leqq x \leqq 1, \\ 0, & 1 \leqq x. \end{cases} \tag{4.101}$$

Funkce φ_0 je tedy v bodě nula rovna jedné a má tam nulovou derivaci a ve všech ostatních bodech o celočíselných souřadnicích je spolu se svou první derivací rovna nule. U funkce φ_1 je oproti funkci φ_0 role funkčních hodnot a hodnot první derivace obrácená. Užijeme-li právě zavedené funkce, jsou bázové funkce Φ_i dány jednoduchými vzorci

$$\begin{aligned} \Phi_{2i}(x) &= \varphi_0\Big(\frac{x - x_i}{h}\Big), \\ \Phi_{2i+1}(x) &= \varphi_1\Big(\frac{x - x_i}{h}\Big), \quad i = 0, \ldots, n. \end{aligned} \tag{4.102}$$

Vypustíme-li v soustavě bázových funkcí (4.102) funkce Φ_0, Φ_1, Φ_{2n}, Φ_{2n+1}, dostaneme prostor konečných prvků $\mathscr{D}^0_{h2}$, který je podprostorem prostoru $\mathscr{D} = \mathscr{H}^2_0$ těch funkcí ze Sobolevova prostoru $\mathscr{H}^2$, které splňují okrajové podmínky (4.39). Užijeme-li tuto bázi k přibližnému řešení okrajové úlohy čtvrtého řádu (3.206), (4.39), bude matice příslušné soustavy lineárních rovnic pásová se sedmi nenulovými diagonálami, neboť z toho, že každá bázová funkce Φ_{2i}, Φ_{2i+1} je různá od nuly pouze v intervalu $\langle x_{i-1}, x_{i+1}\rangle$, plyne, že platí $[\Phi_i, \Phi_j] = 0$ pro $|i - j| \geqq 4$.

Prostor konečných prvků $\mathscr{D}_{h2}$ lze užít i k řešení úloh druhého řádu, neboť podmínka, aby byl podprostorem prostoru $\mathscr{H}^1$, je samozřejmě splněna. V tomto případě plyne ze vzorců (4.61) a (4.87), že za předpokladu dostatečné hladkosti řešení lze dosáhnout rychlosti konvergence měřené v normě prostoru $\mathscr{H}^1$ až $O(h^3)$. Cena, která se za to zaplatí, je však ta, že matice soustavy (4.53) je nyní sedmidiagonální a je zhruba dvojnásobného řádu než v případě prostoru $\mathscr{D}_{h1}$.

Při konstrukci báze v prostoru $\mathscr{D}_{hk}$ při obecném k lze postupovat obdobně. Za bázi lze vzít soustavu funkcí $\Phi_0, \ldots, \Phi_{k(n+1)-1}$, z nichž každá vznikne tak, že je pro ni právě jeden z uzlových parametrů rovný jedné a ostatní jsou rovny nule. Je-li funkce φ_j $(j = 0, \ldots, k-1)$ rovna v intervalu $\langle -1, 0\rangle$ a v intervalu $\langle 0, 1\rangle$ polynomu $(2k-1)$-ního stupně (v každém intervalu jinému) určenému podmínkami $\varphi_j^{(j)}(0) = 1$, $\varphi_j^{(r)}(0) = 0$ pro $r = 0, \ldots, k-1$, $r \neq j$ a $\varphi_j^{(r)}(\pm 1) = 0$ pro $r = 0, \ldots, k-1$ a všude jinde na reálné ose je rovna nule, jsou bázové funkce Φ_i dány vzorci

$$\Phi_{ki+j}(x) = \varphi_j\Big(\frac{x - x_i}{h}\Big), \quad j = 0, \ldots, k-1, \ i = 0, \ldots, n. \tag{4.103}$$

Vypustíme-li v soustavě bázových funkcí (4.103) funkce $\Phi_0, \ldots, \Phi_{k-1}$ a funkce $\Phi_{nk}, \ldots, \Phi_{(n+1)k-1}$, dostaneme prostor konečných prvků $\mathscr{D}^0_{hk}$, který je podprostorem těch funkcí ze Sobolevova prostoru $\mathscr{H}^k$, které splňují okrajové podmínky

(4.43) s $m = k$. Tento prostor je tedy vhodný k přibližnému řešení diferenciální rovnice (4.45) s okrajovými podmínkami (4.43).

Prostor konečných prvků $\mathscr{D}_{hk}$ lze také užít k řešení okrajových úloh pro diferenciální rovnice řádů nižších než $2k$. Vhodnou volbou parametru k (tj. stupně aproximačních polynomů) lze dosáhnout, že rychlost konvergence je řádu $O(h^p)$, kde p je libovolně vysoké číslo (samozřejmě za předpokladu dostatečné hladkosti řešení). Tento vzrůst v přesnosti je však zaplacen komplikovanější strukturou matice soustavy (šíře jejího pásu je větší) a větší počtem rovnic.

4.3.3 Některé praktické otázky spojené s metodou konečných prvků

V předchozích odstavcích jsme popsali základní myšlenky užití metody konečných prvků k řešení okrajových úloh pro obyčejné diferenciální rovnice. Zde si všimneme některých aspektů praktické realizace popsaných algoritmů. Pro určitost budeme mít v tomto odstavci na mysli diferenciální rovnici (1.12) s okrajovými podmínkami (4.3), protože i na tomto jednoduchém příkladě budou patrné všechny jevy, které jsou důležité a podstatné v obecném případě.

Při řešení dané okrajové úlohy metodou konečných prvků musíme vždy provést čtyři kroky: (i) zvolit prostor konečných prvků a jeho bázi, (ii) sestavit příslušnou matici tuhosti (Gramovu matici) a vektor zatížení, (iii) řešit vzniklou soustavu rovnic a (iv) vypočítat hodnoty přibližného řešení v bodech, v nichž nás zajímá.

Začneme bodem (iv) jako nejsnazším. Už dříve jsme konstatovali, že báze v prostorech konečných prvků je žádoucí volit tak, aby aspoň některé neznámé, které dostaneme řešením soustavy z bodu (iii), byly už přímo hodnoty hledaného řešení. Pokud tento požadavek splníme — v případech prostorů konečných prvků, které jsme uvedli, tomu tak vždy bylo — krok (iv) vlastně odpadá. Poznamenejme, že v některých případech nás zajímají spíše hodnoty některých derivací hledaného řešení než hodnoty samotného řešení. V tomto případě může být užití Hermitových prvků výhodné, protože při něm část vypočítaných neznámých udává přímo hodnoty derivací řešení.

Probereme nyní body (i) až (iii). Pokud se týká bodu (i), k problematice volby prostoru konečných prvků se dá říci obecně jen velmi málo. Udat nějaká matematicky podložená exaktní kritéria, která by udávala, jaký konkrétní prostor konečných prvků v daném případě zvolit, je obtížné. Zkušenost s řešením problémů podobného typu zde může být velmi užitečná. Zdůrazněme, že při volbě báze v už zvoleném prostoru konečných prvků je třeba postupovat tak, aby vzniklá matice tuhosti měla co možná nejmenší šíři pásu. Tento požadavek je pro ekonomii výpočtu natolik důležitý, že může ovlivnit zpětně i volbu samotného prostoru konečných prvků.

Všimněme si dále bodu (ii), tj. problematiky sestavení matice tuhosti a vektoru zatížení. Pro určitost zvolme za příslušný prostor konečných prvků prostor $\mathscr{D}^0_{h1}$ z odst. 4.3.1 s bází tvořenou po částech lineárními funkcemi $\Phi_1, \ldots, \Phi_{n-1}$ definovanými rovnicemi (4.83). Píšeme-li v případě modelové úlohy integrál udávající prvky

matice tuhosti A jako součet integrálů přes jednotlivé intervaly dělení, dostaneme

$$a_{ij} = \sum_{r=1}^{n} a_{ij}^{(r)}, \tag{4.104}$$

kde

$$a_{ij}^{(r)} = \int_{x_{r-1}}^{x_r} [p(x)\Phi_i'(x)\Phi_j'(x) + q(x)\Phi_i(x)\Phi_j(x)]\,\mathrm{d}\,x, \quad r = 1, \ldots, n. \tag{4.105}$$

Matici $A^{(r)} = \{a_{ij}^{(r)}\}$ nazveme *elementární maticí tuhosti* příslušnou k podintervalu $\langle x_{r-1}, x_r\rangle$. Pro vektor zatížení (tj. pro pravou stranu soustavy (4.53)) platí podobně

$$g_i = \sum_{r=1}^{n} g_i^{(r)}, \tag{4.106}$$

kde

$$g_i^{(r)} = \int_{x_{r-1}}^{x_r} \Phi_i(x) f(x)\,\mathrm{d}\,x, \quad r = 1, \ldots, n, \tag{4.107}$$

přičemž vektor $g^{(r)} = (g_1^{(r)}, \ldots, g_{n-1}^{(r)})^T$ nazveme *elementárním vektorem zatížení* pro podinterval $\langle x_{r-1}, x_r\rangle$.

Prvky matice tuhosti a složky vektoru zatížení můžeme tedy vypočítat jako součet příspěvků od jednotlivých podintervalů rozkladu. Tato okolnost, přestože je zcela elementární, má pro efektivní užití metody konečných prvků zásadní význam. Vzhledem k tomu, že bázové funkce se na jednotlivých prvcích rovnají jen několika málo velmi jednoduchým funkcím (ve všech příkladech prostorů konečných prvků, které jsme uvedli, jsou tyto funkce polynomy nízkých stupňů), je možné sestrojit matici A a vektor g tak, že se v podstatě vypočítá elementární matice $A^{(r)}$ a elementární vektor $g^{(r)}$ jen pro typický prvek a pak se užijí vzorce (4.104) a (4.106). V našem konkrétním případě je každá bázová funkce na podintervalu $\langle x_{r-1}, x_r\rangle$ rovna buď přímce $(x - x_r)/h$, nebo přímce $(x_r - x)/h$, nebo je nulová. K sestavení elementární matice tuhosti je tedy třeba vypočítat integrály pouze tří typů:

$$\int_{x_{r-1}}^{x_r} \Big[\frac{1}{h^2}p(x) + q(x)\Big(\frac{x - x_{r-1}}{h}\Big)^2\Big]\,\mathrm{d}\,x = h\int_0^1 \Big[\frac{1}{h^2}p(x_{r-1} + hs) + q(x_{r-1} + hs)s^2\Big]\,\mathrm{d}\,s, \tag{4.108}$$

$$\int_{x_{r-1}}^{x_r} \Big[-\frac{1}{h^2}p(x) + q(x)\frac{x - x_{r-1}}{h}\frac{x_r - x}{h}\Big]\,\mathrm{d}\,x = h\int_0^1 \Big[-\frac{1}{h^2}p(x_{r-1} + hs) + q(x_{r-1} + hs)s(1 - s)\Big]\,\mathrm{d}\,s \tag{4.109}$$

a

$$\int_{x_{r-1}}^{x_r} \Big[\frac{1}{h^2}p(x) + q(x)\Big(\frac{x_r - x}{h}\Big)^2\Big]\,\mathrm{d}\,x = h\int_0^1 \Big[\frac{1}{h^2}p(x_r - hs) + q(x_r - hs)s^2\Big]\,\mathrm{d}\,s. \tag{4.110}$$

Podobně k výpočtu elementárního zatížení je třeba počítat pouze integrál

$$\int_{x_{r-1}}^{x_r} f(x)\frac{x - x_{r-1}}{h}\,\mathrm{d}\,x = h\int_0^1 f(x_{r-1} + hs)s\,\mathrm{d}\,s \tag{4.111}$$

a integrál

$$\int_{x_{r-1}}^{x_r} f(x)\frac{x_r - x}{h}\,\mathrm{d}\,x = h\int_0^1 f(x_r - hs)s\,\mathrm{d}\,s. \tag{4.112}$$

Pravé strany vzorců (4.108) až (4.112) ukazují, že výpočet prvků elementární matice tuhosti a elementárního vektoru zatížení, se vlastně provádí na referenčním prvku $\langle 0, 1\rangle$. Upozorněme také, že integrály uvedených typů nelze v obecném případě počítat analyticky. K jejich výpočtu je tedy třeba užít kvadraturní vzorce. Přitom je třeba si uvědomit, že tento postup zavádí do výpočtu další chybu.

Provedení kroku (iii) nepředstavuje při užití metody konečných prvků k řešení jednodimenzionálních problémů principiálně žádný příliš závažný problém. Vzniklé soustavy rovnic totiž mají pásové matice, jejichž šíře pásu nezávisí na počtu rovnic. Užijeme-li k řešení takových soustav Gaussovu eliminaci bez výběru hlavního prvku (a to je možné, neboť příslušné matice jsou pozitivně definitní), je počet potřebných operací úměrný počtu rovnic. Z tohoto důvodu nelze očekávat, že by některá iterační metoda mohla být výhodnější než eliminační metoda.

CVIČENÍ

1. Odvoďte rovnici (2.48) tak, že napíšete obecné řešení rovnice (1.12), jednu konstantu vyloučíte z podmínky (2.45) a druhou derivováním.
2. Dokažte lemma 2.3 (viz str. 136).
3. Dokažte implikaci (2.73).
4. Ukažte, že (2.79) je nutná a postačující podmínka řešitelnosti okrajové úlohy (2.77), (2.78).
5. Dokažte podrobně větu 2.3 (viz str. 139).
6. Ukažte, že okrajové úlohy (2.21), (1.5) a (2.96), (2.97) jsou ekvivalentní.
7. Dokažte ekvivalenci okrajových úloh (2.21), (1.6), (2.100) a (2.101), (2.102).
8. Proveďte podrobný důkaz věty 2.5 ze str. 145.

9. Dokažte lemma 3.8 (viz str. 168).

10. Odvoďte řády chyb kvadraturních vzorců (3.51), (3.52) a (3.53) užitých v Marčukově identitě a zformulujte potřebné hladkostní požadavky.

11. Proveďte podrobný důkaz poznámky 3.1 (viz str. 178).

12. Dokažte princip maxima pro operátor $L_h^{(0)}$ definovaný rovnicí (3.25).

13. Prověřte monotónnost matice soustavy (3.31). Návod: Užijte výsledku cvičení 12.

14. Dokažte konvergenci metody sítí (3.31). Návod: Postupujte obdobně jako při důkazu věty 3.7 na str. 179.

15. Buď $\boldsymbol{\eta} = (\eta_0, \ldots, \eta_n)^T$ vektor, pro který platí $\eta_0 = \eta_n = 0$. Dokažte, že platí nerovnosti

$$h^2 \sum_{k=1}^{n-1} (L_h \boldsymbol{\eta})_k \eta_k \geqq p_0 \sum_{k=1}^{n} (\eta_k - \eta_{k-1})^2$$

a

$$\|\boldsymbol{\eta}\|_{\mathscr{L}_\infty} \leqq \frac{1}{h} \gamma p_0 \sum_{k=1}^{n} (\eta_k - \eta_{k-1})^2.$$

Návod: Postupujte analogicky jako v důkazu lemmat 3.11 a 3.12 na str. 184 až 186.

16. Dokažte, že metoda sítí (3.131) má v případě řešení okrajové úlohy (1.12), (1.13) s $\alpha_1 = \alpha_2 = 0$ chybu řádu $O(h^4)$. Návod: Ukažte, že operátor definovaný levou stranou rovnic (3.131) splňuje princip maxima.

17. Zformulujte metodu sítí pro řešení diferenciální rovnice (1.14) s okrajovými podmínkami $y(a) = \gamma_1$, $y''(a) = \delta_1$, $y(b) = \gamma_2$, $y'' = \delta_2$.

18. Dokažte konvergenci metody sítí z předešlého cvičení.

19. Ukažte, že operátor L definovaný na množině $\mathscr{D}_L \subset \mathscr{L}_2$ rovnicí (4.40) je symetrický a pozitivně definitní.

20. Dokažte symetrii a pozitivní definitnost operátoru L daného rovnicí (4.45) na množině $\mathscr{D}_L$ dostatečně hladkých funkcí, které splňují okrajové podmínky (4.43). Návod: Vyjádřete $y(x)$ pomocí integrálu z m-té derivace.

POZNÁMKY K LITERATUŘE

Čl. 1. Základy teorie okrajových úloh pro obyčejné diferenciální rovnice lze nalézt např. v knize Reidově (1971) a v řadě dalších standardních příruček o obyčejných diferenciálních rovnicích (viz např. Coddington, Levinson (1955)). Literatura pojednávající o numerických aspektech této problematiky je skoro stejně rozsáhlá jako literatura o úlohách s počátečními podmínkami. Kromě příslušných kapitol v obecných příručkách o numerické matematice, viz např. Collatz (1951), Babuška, Práger, Vitásek (1964), Berezin, Židkov (1966), Babuška, Práger, Vitásek (1966), Stoer, Bulirsch (1973), Vitásek (1987), existuje řada monografií věnovaných této problematice. V seznamu literatury se tato díla poznají většinou už podle názvu; proto uvedeme jmenovitě pouze sborník redigovaný Azizem (1975), protože obsahuje velmi obsáhlou bibliografii.

Čl. 2. O metodě střelby pojednává už např. Fox (1957) a Keller (1960). Metoda střelby na více cílů pochází od Osborna (1969) a spolu s ostatními variantami metody střelby ji podrobně popisuje Stoer a Bulirsch (1980). Robertsova a Shipmanova (1972) monografie je věnována výhradně metodě střelby. Metoda přesunu a normalizovaného přesunu okrajové podmínky v podobě, jak ji uvádíme, pochází od Taufera (1973). Tyto metody velice úzce souvisí s metodami invariantního vnoření, kterým je věnována poměrně rozsáhlá literatura; viz např. Meyer (1973) a Scott (1977). V ruské literatuře se podobné metody nazývají „metod progonki“ (viz např. Babuška, Práger, Vitásek (1964)).

Čl. 3. Metoda sítí patří k základním metodám pro numerické řešení okrajových úloh a aspoň její princip je vyložen ve většině příruček zabývajících se přibližným řešením diferenciálních rovnic. O monotónních maticích, jakož i o dalších důležitých třídách matic, které vznikají při metodě sítí, nalezne čtenář poučení např. v knize Vargově (1962) a Fiedlerově (1981). Myšlenka užití integrálních identit k sestavení diferenčních schémat pochází od Marčuka (1961) a byla dále rozpracována Babuškou, Prágerem a Vitáskem (1966) a Marčukem (1980). Jiné integrální identity než ty, které jsou zde uvedeny, užívá Varga (1962). Mnoho materiálu o metodě sítí lze nalézt v obsáhlé monografii Samarského (1971). Velmi přístupný výklad o řešení nelineárních úloh metodou sítí podává Isaacson a Keller (1966).

Čl. 4. Teoretický základ, na němž spočívají variační metody řešení okrajových úloh, je podrobně a hlavně také s ohledem na čtenáře nespecialistu vyložen v knize Rektorysově (1974). Nezbytná funkcionálně-analytická teorie okrajových úloh pro obyčejné i parciální diferenciální rovnice je vyložena také v knize Strangově a Fixově (1973), kterou lze čtenáři, který chce hlouběji proniknout do teorie metody konečných prvků, co nejvřeleji doporučit. Rovněž tak Michlinovy knihy (1957 a 1966) mohou být pro čtenáře, který se zajímá o problematiku tohoto článku, velmi užitečné. Metoda konečných prvků prodělala v posledním dvacetiletí bouřlivý rozvoj a v současné době je jí věnována už značně rozsáhlá literatura. Protože však její hlavní uplatnění je zejména při řešení parciálních diferenciálních rovnic, uvedeme hlavní prameny pro její studium v následující kapitole a zde se zmíníme jen o ně-

kterých příručkách, které obsahují elementární úvod do této problematiky. Takový přístupný výklad se nalezne kromě v už zmíněné knize Strangově a Fixově (1973), např. také v knize Stoerově a Bulirschově (1980), Beckerově, Careyově a Odenově (1981), která uvádí čtenáře také do problémů spojených s implementací metody konečných prvků na počítači a v knize Axelssonově a Barkerově (1984).

LITERATURA

AXELSSON, O. – BARKER, V.A.: Finite Element Solution of Boundary Value Problems, Theory and Computation. New York–San Francisco–London, Academic Press 1984.

AZIZ, A.K., ed.: Numerical Solution fo Boundary Value Problems for Ordinary Differential Equations. New York–San Francisco–London, Academic Press 1975.

BABUŠKA, I. – PRÁGER, M. – VITÁSEK, E.: Numerické řešení diferenciálních rovnic. Praha, SNTL 1964.

BABUŠKA, I. – PRÁGER, M. – VITÁSEK, E.: Numerical Processes in Differential Equations. London–New York–Sydney, Interscience Publishers 1966 (Překlad do ruštiny: Moskva, Mir 1969.).

BAILEY, P.B. – SHAMPINE, L.F. – WALTMAN, P.E.: Nonlinear Two Point Boundary Value Problem. New York, Academic Press 1968.

BECKER, E.D. – CAREY, G.F. – ODEN, I.T.: Finite Elements. An Introduction, Vol. I. Englewood Cliffs, N.J., Prentice–Hall 1981.

BEREZIN, I.S. – ŽIDKOV, N.P.: Metody vyčislenij. 3. vyd. Moskva, Nauka 1966, 2 sv.

CODDINGTON, E.A. – LEVINSON, N.: Theory of Ordinary Differential Equations. New York, McGraw–Hill 1955. (Překlad do ruštiny: Moskva, IL 1958.)

COLLATZ, L.: Numerische Behandlung von Differentialgleichungen. Berlin–Göttingen–Heidelberg, Springer–Verlag 1951. (Překlad do ruštiny: Moskva, IL 1953.)

FIEDLER, M.: Speciální matice a jejich použití v numerické matematice. Praha, SNTL 1981.

FOX, L.: The Numerical Solution of Two Point Boundary Value Problems. Oxford, Clarendon Press 1957.

ISAACSON, E. – KELLER, H.B.: Analysis of Numerical Methods. New York–London–Sydney, J. Wiley and Sons 1966.

KELLER, H.B.: Numerical Methods for Two Point Boundary Value Problems. Waltham, Mass., Blaisdell 1968.

MARČUK, G.I.: Metody rasčota jadernych reaktorov. Moskva, Gosatomizdat 1961.

MARČUK, G.I.: Metody vyčislitel'noj matematiky. 2 vyd. Moskva, Nauka 1980. (Překlad do češtiny: Praha, Academia 1987.)

MEYER, G.H.: Initial Value Methods for Boundary Value Problems: Theory and Application of Invariant Imbedding. New York, Academic Press 1973.

MICHLIN, S.G.: Variacionnyje metody v matematičeskoj fizike. Moskva, Gostechizdat 1957.

MICHLIN, S.G.: Čislennaja realizacija variacionnych metodov. Moskva, Nauka 1966.

OSBORNE, M.R.: On Shooting Methods for Boundary Value Problems. J. Math. Anal. Appl., *27*, 1969, s. 417 – 433.

REID, W.T.: Ordinary Differential Equations. New York–London–Sydney–Toronto, Interscience Publishers 1971.

REKTORYS, K.: Variační metody v inženýrských problémech a v problémech matematické fyziky. Praha, SNTL 1974.

ROBERTS, S.M. – SHIPMAN, I.S.: Two Point Boundary Value Problems: Shooting Methods. New York, American Elsevier 1972.

SAMARSKIJ, A.A.: Vvedenije v teoriju raznostnych schem. Moskva, Nauka 1971.

SCOTT, M.R.: Invariant Imbedding and its Applications for Ordinary Differential Equations: An Introduction. Reeding, Mass., Addision–Wesley 1977.

STOER, J. – BULIRSCH, R.: Introduction to Numerical Analysis. New York–Heidelberg–Berlin, Springer–Verlag 1980.

STRANG, G. – FIX, G.J.: An Analysis of the Finite Element Method. Englewood Cliffs, N.J., Prentice–Hall 1973. (Překlad do ruštiny: Moskva, Mir 1977.)

TAUFER, J.: Lösung der Randwertprobleme von linearen Differentialgleichungen. Praha, Rozpravy ČSAV, Řada mat. přír. věd, *83*, Academia 1973. (Překlad do ruštiny: Moskva, Nauka 1981.)

VITÁSEK, E.: Numerické metody. Praha, SNTL 1987.

VARGA, R.S.: Matrix Iterative Analysis. Englewood Cliffs, N.J., Prentice–Hall 1962.

WILKINSON, J.H.: The Algebraic Eigenvalue Problem. Oxford, Clarendon Press 1965.

Kapitola III.

Parciální diferenciální rovnice eliptického typu

1 Úvod

Základní úloha pro parciální diferenciální rovnici eliptického typu je *okrajová úloha*, tj. úloha při níž je třeba nalézt funkci, která splňuje uvnitř dané oblasti danou diferenciální rovnici a na celé hranici této oblasti ještě další doplňující podmínky, zvané *okrajové podmínky*. Typickým příkladem okrajové úlohy pro eliptickou lineární parciální diferenciální rovnici druhého řádu je úloha nalézt funkci u (m reálných proměnných $x = (x_1, \dots, x_m)$), která v dané omezené oblasti $\Omega \subset \mathbf{E}^m$ ($\mathbf{E}^m$ je m-dimenzionální euklidovský prostor) splňuje diferenciální rovnici

$$Lu \equiv -\sum_{i,j=1}^{m} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) + q(x)u = f(x) \tag{1.1}$$

a pro niž na hranici Γ této oblasti platí

$$\alpha(x)\frac{\partial u}{\partial n_c} + \beta(x)u = \gamma(x). \tag{1.2}$$

Zde $a_{ij} = a_{ji}$, q a f jsou funkce zadané v oblasti Ω, funkce α, β a γ na její hranici a symbol $\partial u/\partial n_c$ značí derivaci ve směru tzv. *konormály* a je definován rovnicí

$$\frac{\partial u}{\partial n_c} = \sum_{i,j=1}^{m} a_{ij}(x)\frac{\partial u}{\partial x_j}\cos(\nu, x_j), \tag{1.3}$$

kde $\cos(\nu, x_j)$ je kosinus úhlu, který svírá vnější normála k hranici v příslušném bodě s osou x_j.

Eliptičnost diferenciální rovnice (1.1) je charakterizována zpravidla nerovností

$$\sum_{i,j=1}^{m} a_{ij}(x)\xi_i\xi_j \geqq p_0 \sum_{i=1}^{m} \xi_i^2, \tag{1.4}$$

kde p_0 je předem daná kladná konstanta a $\boldsymbol{\xi} = (\xi_1, \dots, \xi_m)^T$ je libovolný m-dimenzionální vektor. Platnost nerovnosti (1.4) se přitom požaduje pro každý bod $x \in \Omega$.

O koeficientech α a β v okrajové podmínce se většinou předpokládá, že splňují nerovnosti

$$\alpha(x) \geqq 0, \quad \beta(x) \geqq 0, \quad \alpha(x) + \beta(x) > 0. \tag{1.5}$$

V případě, že je $\alpha(x) = 0$ pro $x \in \Gamma$, hovoříme o *Dirichletově okrajové podmínce* v bodě $x \in \Gamma$, v případě, že je $\beta(x) = 0$, o *Neumannově podmínce*; obecná podmínka (1.2) se nazývá *Newtonova okrajová podmínka*.

V první a druhé kapitole jsme viděli, že už pro obyčejné diferenciální rovnice je okrajová úloha podstatně komplikovanější než úloha s počátečními podmínkami, a to nejen po stránce teoretické, ale i po stránce prakticky početní. Je tomu tak zejména proto, že zatímco řešení úloh s počátečními podmínkami vede většinou na algoritmy rekurentního charakteru, řešení okrajových úloh vede na soustavy rovnic, které je nutno (někdy i velmi pracně) řešit. Je přirozené, že u parciálních diferenciálních rovnic s růstem počtu dimenzí tyto obtíže porostou. Dimenze prostoru, v němž je definována hledaná funkce, přináší však ještě další problém. Zatímco u jednodimenzionálních úloh jsme měli co dělat zejména s dvoubodovými okrajovými úlohami, a tedy s případy, kdy definiční obor hledané funkce byl jednorozměrný interval, u parciálních diferenciálních rovnic pracujeme v prostoru, jehož dimenze je větší než jedna, čímž je dána nesrovnatelně větší pestrost v možných definičních oblastech a okrajových podmínkách. Tím spíše jsme tedy nuceni omezit se v této kapitole při popisu uváděných metod na vyšetřování jednotlivých konkrétních příkladů.

Odlišný bude v této kapitole přístup k teoretickým otázkám spojeným s řešenými problémy. V předchozích dvou kapitolách jsme nemuseli o řešitelnosti úloh, jimiž jsme se zabývali, nic předpokládat předem, neboť způsob, kterým jsme vyšetřovali zavedené přibližné metody, nám umožnil zároveň také odpovědět i na tyto existenční otázky. Zde by tento postup byl sice také možný, byl by však tak náročný a zdlouhavý, že by přesahoval rámec i rozsahové možnosti této elementární příručky. Pokud tedy půjde o otázku existence řešení studovaných úloh, odvoláme se na obecnou teorii eliptických rovnic, která je rozpracována do značných podrobností, a to jak moderními funkcionálně analytickými metodami (viz např. Nečas (1967), Lions-Magenes (1968)), tak i klasickými metodami (viz např. Bers, John, Schechter (1964), Miranda (1955)).

Pojem eliptičnosti parciální diferenciální rovnice není ovšem vázán na jednu rovnici daného řádu (ani rovnice (1.1) není nejobecnější lineární eliptická parciální diferenciální rovnice druhého řádu; jde pouze o tzv. samoadjungovanou rovnici) a dá se zavést i pro rovnice vyšších řádů nebo pro soustavy rovnic. Typickým příkladem eliptické diferenciální rovnice čtvrtého řádu je *biharmonická rovnice*

$$\left(\sum_{i=1}^{m} \frac{\partial^2}{\partial x_i^2}\right)^2 u = f(x), \tag{1.6}$$

pro niž mohou okrajové podmínky nabývat velice různých forem. Vždy však bude

k jednoznačnému řešení zapotřebí dvou okrajových podmínek; za příklad mohou sloužit okrajové podmínky $u = g_1(x)$ a $\partial u/\partial n = g_2(x)$ pro $x \in \Gamma$, které jsou v jistém smyslu obdobou Dirichletových okrajových podmínek pro rovnici druhého řádu.

Jako příklad eliptické soustavy diferenciálních rovnic druhého řádu mohou sloužit v dvoudimenzionálním případě rovnice

$$-\frac{\partial}{\partial x}\Big(\frac{\partial u_1}{\partial x}+\frac{\partial u_2}{\partial y}\Big)-\frac{\partial^2 u_1}{\partial x^2}-\frac{\partial^2 u_1}{\partial y^2}=f_1(x,y), \tag{1.7}$$

$$-\frac{\partial}{\partial y}\Big(\frac{\partial u_1}{\partial x}+\frac{\partial u_2}{\partial y}\Big)-\frac{\partial^2 u_2}{\partial x^2}-\frac{\partial^2 u_2}{\partial y^2}=f_2(x,y),$$

které se vyskytují v matematické teorii pružnosti.

Z metod pro přibližné řešení eliptických parciálních diferenciálních rovnic si zde všimneme dvou základních skupin. Především to budou metody založené na aproximaci derivací diferenčními podíly. Nahradíme-li všechny derivace, které se v rovnici vyskytují, těmito podíly, dostaneme metodu sítí, ponecháme-li derivace podle jedné z proměnných a pouze derivace podle zbývajících proměnných nahradíme diferenčními podíly, dostaneme metodu, jejíž běžný název je *metoda přímek*. V této metodě se tedy řešení dané okrajové úlohy aproximuje okrajovou úlohou pro soustavu obyčejných diferenciálních rovnic. Metodou přímek se nebudeme v dalším detailně zabývat, pokládáme však za užitečné, aby se čtenář seznámil aspoň s její základní myšlenkou.

Další skupina metod, jichž si také stručně všimneme, je tvořena variačními metodami založenými na stejném principu jako metody v čl. 4 z kap. II. Zde je důležitý zejména speciální případ, totiž metoda konečných prvků.

Je užitečné upozornit, že metody všech zmíněných skupin spolu úzce souvisí. Tak např. metoda přímek přechází ve variantu metody sítí, řešíme-li aproximující soustavu obyčejných diferenciálních rovnic znovu metodou sítí. Podobně metoda konečných prvků při určité volbě bázových funkcí přechází v metodu sítí.

Pokud jde o konkrétní realizaci zmíněných metod na počítači, má zde podstatný význam rychlost a kapacita paměti stroje. Současný stav je takový, že řada trojdimenzionálních úloh přesahuje praktické možnosti. Proto se v této kapitole, v níž si postupně všimneme zmíněných metod, omezíme na dvoudimenzionální a lineární problémy. Přitom budeme také věnovat pozornost problematice řešení náhradního problému, který aproximuje původní problém. Pokud jde o nelineární úlohy, nebudeme se jimi zabývat, neboť tyto problémy přinášejí často už i teoretické těžkosti při důkazu existence a jednoznačnosti řešení a je nutno většinou vyšetřovat každý případ jednotlivě. Jejich byť jen částečné pojednání by tedy neúnosně zvětšilo rozsah příručky.

2 Metoda sítí

Metoda sítí je velmi oblíbená metoda pro numerické řešení nejen parciálních diferenciálních rovnic eliptického typu, ale parciálních diferenciálních rovnic vůbec. Popularita metody konečných prvků, která v současnosti stále roste, její význam v poslední době sice poněkud zmenšila, přesto však řada důležitých fyzikálních a technických problémů, které vedou na parciální diferenciální rovnice, se dodnes pomocí ní řeší. Oblíbenost metody sítí má zřejmě své kořeny v tom, že její základní myšlenka je velice prostá. Spočívá v tom, jak už bylo ostatně řečeno v čl. 3 z kap. II. a v úvodu k této kapitole, že v oblasti, ve které hledáme řešení, zvolíme nějakou konečnou množinu bodů, kterou nazveme *sítí* a příslušné body jejími *uzly*, a nahradíme derivace, které se vyskytují v dané diferenciální rovnici a v jejích okrajových podmínkách, diferenčními podíly (tj. lineárními kombinacemi funkčních hodnot), které je aproximují. Tím dostaneme místo původního problému soustavu konečně mnoha rovnic pro hodnoty hledané funkce v uzlech. Zde se hned setkáváme s podstatnou odchylkou od jednodimenzionálního případu. Zatímco v jedné dimenzi bylo možno volit uzly v podstatě pouze dvěma způsoby, a to ekvidistantní a neekvidistantní, ve dvou- a vícedimenzionálním případě i při zachování pravidelnosti v rozmístění uzlů je zřejmě možná podstatně větší rozmanitost. Tak např. v rovině lze vzít za uzly sítě vrcholy čtverců, rovnostranných trojúhelníků, pravidelných šestiúhelníků apod.

Z uvedeného výkladu je vidět, že metoda sítí je v principu použitelná pro libovolný typ diferenciální rovnice, ať už lineární nebo nelineární. V dalším textu se však z důvodů, o nichž jsme se zmínili v úvodu, omezíme na některé příklady lineárních dvoudimenzionálních problémů.

2.1 Lineární rovnice druhého řádu

2.2.1 Sestavení diferenčních rovnic

V tomto odstavci popíšeme různé možnosti odvození diferenčních rovnic pro lineární rovnice druhého řádu ve dvou prostorových proměnných. Většinou se přitom omezíme na speciální případ lineární samoadjungované rovnice, totiž na rovnici

$$Lu \equiv -\frac{\partial}{\partial x}\Big(p(x,y)\frac{\partial u}{\partial x}\Big)-\frac{\partial}{\partial y}\Big(p(x,y)\frac{\partial u}{\partial y}\Big)+q(x,y)u=f(x,y) \tag{2.1}$$

v omezené oblasti Ω, jejíž hranice Γ je tvořena konečným počtem po částech hladkých oblouků. O funkcích p, q a f budeme přitom předpokládat, že jsou dostatečně hladké v nějaké oblasti $\tilde{\Omega}$ takové, že pro ni platí $\bar{\Omega} \subset \tilde{\Omega}$ (pruhem značíme uzávěr příslušné množiny), a navíc, že koeficient p je v této oblasti kladný a koeficient q nezáporný. O jaký konkrétní stupeň hladkosti jde, bude v jednotlivých situacích buď specifikováno, nebo to bude patrné ze souvislosti.

Podrobně probereme zejména případ pravidelné čtvercové sítě $K^{(h)}$. Tato síť vznikne tak, že se sestrojí soustava rovnoběžek $x = x_k = x_0 + kh$, $y = y_s = y_0 + sh$, $k, s = 0, \pm 1, \pm 2, \ldots$, kde (x_0, y_0) je libovolný pevně zvolený bod v rovině (x, y) a $h > 0$ je parametr zvaný *integrační krok* nebo *oko sítě*, a za uzly sítě se vezmou všechny průsečíky těchto rovnoběžek, tj. body o souřadnicích $(x_k, y_s) = (x_0 + kh, y_0 + sh)$. Při zavádění metody sítí v jednodimenzionálním případě jsme postupovali tak, že v uzlech, které ležely uvnitř daného intervalu, jsme příslušné lineární rovnice získali z dané diferenciální rovnice a v krajních bodech jsme využili okrajové podmínky. Totéž je třeba provést i ve dvou dimenzích. Abychom mohli příslušný postup snadno popsat, zavedeme nejprve několik pojmů a označení.

Uzly sítě budeme značit buď jejich souřadnicemi, tj. (x_k, y_s), nebo také, zejména tam, kde na konkrétních hodnotách souřadnic nezáleží, pouhými velkými písmeny. Dva uzly sítě nazveme *sousedními*, je-li jejich vzdálenost ve směru osy x nebo ve směru osy y rovna oku sítě h. Každý pevně zvolený uzel sítě $K^{(h)}$ má tedy právě čtyři sousední uzly. Uzel (x_k, y_s) nazveme *vnitřním uzlem* sítě $K^{(h)}$ vzhledem k množině Ω, leží-li všechny čtyři k němu sousední uzly v množině $\bar{\Omega}$. Množinu všech vnitřních uzlů (vzhledem k množině Ω) označíme $\Omega^{(h)}$. Uzel $(x_k, y_s) \in \bar{\Omega} \setminus \Omega^{(h)}$ takový, že aspoň jeden jeho soused je vnitřním uzlem, nazveme *hraničním uzlem*. Množinu všech hraničních uzlů označíme $\Gamma^{(h)}$. Kromě toho budeme pro množinu $\Omega^{(h)} \cup \Gamma^{(h)}$ užívat označení $\bar{\Omega}^{(h)}$. Dále budeme předpokládat, že integrační krok h je zvolen tak malý, že ke každým dvěma vnitřním uzlům P a Q existuje konečná posloupnost vnitřních uzlů $P_0, P_1, \ldots, P_r$ taková, že $P_0 = P$, $P_r = Q$ a libovolné dva uzly P_{i-1} a P_i jsou sousední. To je vzhledem k souvislosti množiny Ω možné. O této souvislosti budeme mluvit jako o souvislosti množiny $\Omega^{(h)}$.

Nyní už můžeme snadno sestavit diferenční rovnice, které aproximují diferenciální rovnici (2.1). Proveďme v rovnici (2.1) naznačené derivování a nahraďme derivace $\partial^2 u/\partial x^2$, $\partial^2 u/\partial y^2$, $\partial u/\partial x$ a $\partial u/\partial y$ podíly $[u(x_k - h, y_s) - 2u(x_k, y_s) + u(x_k + h, y_s)]/h^2$, $[u(x_k, y_s - h) - 2u(x_k, y_s) + u(x_k, y_s + h)]/h^2$, $[u(x_k + h, y_s) - u(x_k - h, y_s)]/(2h)$, $[u(x_k, y_s + h) - u(x_k, y_s - h)]/(2h)$, které aproximují příslušné derivace s přesností $O(h^2)$. Dostaneme tak operátor $L_h^{(1)}$, který funkci u definované na množině $\bar{\Omega}^{(h)}$ (tedy vlastně vektoru, který má tolik složek, kolik prvků má množina $\bar{\Omega}^{(h)}$) přiřazuje funkci $L_h^{(1)}u$ definovanou na množině $\Omega^{(h)}$ a který je dán předpisem

$$
\begin{aligned}
(2.2)\qquad h^2(L_h^{(1)}u)_{ks} = {} & [4p(x_k, y_s) + h^2 q(x_k, y_s)]u_{ks} - \\
& - [p(x_k, y_s) + \tfrac{1}{2}h\frac{\partial p}{\partial x}(x_k, y_s)]u_{k+1,s} - \\
& - [p(x_k, y_s) - \tfrac{1}{2}h\frac{\partial p}{\partial x}(x_k, y_s)]u_{k-1,s} - \\
& - [p(x_k, y_s) + \tfrac{1}{2}h\frac{\partial p}{\partial y}(x_k, y_s)]u_{k,s+1} - \\
& - [p(x_k, y_s) - \tfrac{1}{2}h\frac{\partial p}{\partial y}(x_k, y_s)]u_{k,s-1}, \quad (x_k, y_s) \in \Omega^{(h)}.
\end{aligned}
$$

Všimněme si, že operátor $L_h^{(1)}$ definovaný touto rovnicí svazuje funkce u v uzlu (x_k, y_s) a ve všech jeho sousedech. Zavedení pojmu sousedního uzlu bylo touto skutečností motivováno.

Z Taylorova vzorce ihned plyne následující věta popisující aproximační vlastnosti operátoru $L_h^{(1)}$.

Věta 2.1. *Nechť funkce p má v $\bar{\Omega}$ spojité první parciální derivace podle obou proměnných a nechť funkce q je v Ω spojitá. Nechť dále funkce u má v $\bar{\Omega}$ spojité parciální derivace podle x a y až do čtvrtého řádu včetně. Pak platí*

$$
(2.3)\qquad (L_h^{(1)}u^{(\mathrm{pr})})_{ks} = (Lu)(x_k, y_s) + O(h^2), \quad (x_k, y_s) \in \Omega^{(h)},
$$

kde $u^{(\mathrm{pr})}$ je funkce definovaná na množině $\bar{\Omega}^{(h)}$ předpisem $u_{ks}^{(\mathrm{pr})} = u(x_k, y_s)$ a operátor L je definovaný v rovnici (2.1).

Položíme-li $p(x, y)(\partial u/\partial x) = z(x, y)$, aproximujeme-li derivaci $\partial(p(x, y)\partial u/\partial x)/\partial x = \partial z(x, y)/\partial x$ v uzlu (x_k, y_s) podílem $[z(x_k + h/2, y_s) - z(x_k - h/2, y_s)]/h$ a hodnotu funkce z, tj. hodnotu funkce $p(\partial u/\partial x)$ v bodě $(x_k + h/2, y_s)$, resp. $(x_k - h/2, y_s)$ podílem $p(x_k + h/2)[u(x_k + h, y_s) - u(x_k, y_s)]/h$, resp. podílem $p(x_k - h/2, y_s)[u(x_k, y_s) - u(x_k - h, y_s)]/h$ a provedeme-li analogické operace i pro derivaci podle proměnné y, dostaneme operátor $L_h^{(2)}$ definovaný rovnicí

$$
\begin{aligned}
(2.4)\qquad h^2(L_h^{(2)}u)_{ks} = {} & [p(x_k + h/2, y_s) + p(x_k - h/2, y_s) + \\
& + p(x_k, y_s + h/2) + p(x_k, y_s - h/2) + h^2 q(x_k, y_s)]u_{ks} - \\
& - p(x_k + h/2, y_s)u_{k+1,s} - p(x_k - h/2, y_s)u_{k-1,s} - \\
& - p(x_k, y_s + h/2)u_{k,s+1} - p(x_k, y_s - h/2)u_{k,s-1}.
\end{aligned}
$$

Aproximační vlastnosti operátoru $L_h^{(2)}$ jsou podobné jako aproximační vlastnosti operátoru $L_h^{(1)}$ a jsou popsány v následující větě.

Věta 2.2. *Nechť koeficient p má v $\bar{\Omega}$ spojité parciální derivace až do třetího řádu včetně a nechť koeficient q je v $\bar{\Omega}$ spojitý. Pak pro každou funkci u, která má v $\bar{\Omega}$ spojité parciální derivace až do čtvrtého řádu, platí*

$$
(2.5)\qquad (L_h^{(2)}u^{(\mathrm{pr})})_{ks} = (Lu)(x_k, y_s) + O(h^2), \quad (x_k, y_s) \in \Omega^{(h)},
$$

kde $u^{(\mathrm{pr})}$ je funkce definovaná na množině $\bar{\Omega}^{(h)}$ předpisem $u_{ks}^{(\mathrm{pr})} = u(x_k, y_s)$ a L je operátor z rovnice (2.1).

Důkaz je podobný důkazu věty 3.3 z kap. II a plyne ihned z Taylorova vzorce.

Operátor $L_h^{(2)}$ svazuje opět hodnoty funkce u v uzlu (x_k, y_s) a v jeho čtyřech sousedech a splňuje stejně jako operátor $L_h^{(1)}$ princip maxima. Užití operátoru $L_h^{(2)}$ však vede ve spojení s vhodným předpisem okrajových podmínek (o těchto otázkách budeme hovořit v dalším odstavci) k soustavě lineárních algebraických rovnic

se symetrickou maticí. Vzhledem k tomu, že původní diferenciální rovnice je samoadjungovaná, je třeba dát operátoru $L_h^{(2)}$ přednost před operátorem $L_h^{(1)}$, který vede k soustavě, jejíž matice je pouze „skoro" symetrická (srv. také odst. 3.2.1 v kap. II). Proto také přesně zformulujeme a dokážeme zmíněný princip maxima pouze pro operátor $L_h^{(2)}$. Dříve se však ještě dohodněme, že o integračním kroku, jehož velikost jsme už omezili požadavkem souvislosti množiny $\Omega^{(h)}$, budeme navíc předpokládat, že je tak malý, že v případě, že funkce q není v Ω identicky rovna nule, existuje uzel $(x_{\tilde{k}}, y_{\tilde{s}}) \in \Omega^{(h)}$ takový, že v něm platí $q(x_{\tilde{k}}, y_{\tilde{s}}) > 0$. To je vzhledem k spojitosti funkce q jistě možné.

Lemma 2.1. *Nechť funkce p a q jsou spojité v $\bar{\Omega}$ a nechť je p kladná a q nezáporná. Nechť dále u je libovolná funkce definovaná na množině $\bar{\Omega}^{(h)}$, pro niž platí*

$$(L_h^{(2)}u)_{ks} \leqq 0 \tag{2.6}$$

pro každý uzel $(x_k, y_s) \in \Omega^{(h)}$. Nechť konečně je

$$M = \max_{(x_k,y_s)\in\bar{\Omega}^{(h)}} u_{ks} \tag{2.7}$$

a buď $M > 0$. Pak platí

$$M = \max_{(x_k,y_s)\in\Gamma^{(h)}} u_{ks} \tag{2.8}$$

a existuje-li uzel $(x_{k_0}, y_{k_0}) \in \Omega^{(h)}$ takový, že je $u_{k_0,s_0} = M$, je $q(x) \equiv 0$ v Ω a $u_{ks} = M$ pro každý uzel $(x_k, y_s) \in \bar{\Omega}^{(h)}$.

D ů k a z . Při důkazu tohoto tvrzení budeme postupovat stejně jako při důkazu lemmatu 3.9 z kap. II. Předpokládejme tedy, že existuje uzel $(x_{k_0}, y_{s_0}) \in \Omega^{(h)}$ takový, že platí $u_{k_0,s_0} = M$. Pak je

$$\begin{aligned} 0 \geqq\ & [p(x_{k_0}+h/2, y_{s_0}) + p(x_{k_0}-h/2, y_{s_0}) + \\ & + p(x_{k_0}, y_{s_0}+h/2) + p(x_{k_0}, y_{s_0}-h/2) + h^2 q(x_{k_0}, y_{s_0})]M - \\ & - p(x_{k_0}+h/2, y_{s_0})u_{k_0+1,s_0} - p(x_{k_0}-h/2, y_{s_0})u_{k_0-1,s_0} - \\ & - p(x_{k_0}, y_{s_0}+h/2)u_{k_0,s_0+1} - p(x_{k_0}, y_{s_0}-h/2)u_{k_0,s_0-1} \geqq \\ \geqq\ & h^2 q(x_{k_0}, y_{s_0})M \geqq 0. \end{aligned} \tag{2.9}$$

Protože M je kladné a $q(x_{k_0}, y_{s_0})$ nezáporné, plyne z nerovnosti (2.9) především, že je $q(x_{k_0}, y_{s_0}) = 0$. Kdyby aspoň jedno z čísel u_{k_0-1,s_0}, u_{k_0+1,s_0}, u_{k_0,s_0-1}, u_{k_0,s_0+1} bylo menší než M, platilo by

$$\begin{aligned} 0 \geqq\ & [p(x_{k_0}+h/2, y_{s_0}) + p(x_{k_0}+h/2, y_{s_0}) + \\ & + p(x_{k_0}, y_{s_0}-h/2) + p(x_{k_0}, y_{s_0}+h/2)]M - \\ & - p(x_{k_0}-h/2, y_{s_0})u_{k_0-1,s_0} - p(x_{k_0}+h/2, y_{s_0})u_{k_0+1,s_0} - \\ & - p(x_{k_0}, y_{s_0}-h/2)u_{k_0,s_0-1} - p(x_{k_0}, y_{s_0}+h/2)u_{k_0,s_0+1} > 0, \end{aligned} \tag{2.10}$$

což není možné. Platí tedy $u_{k_0-1,s_0} = u_{k_0+1,s_0} = u_{k_0,s_0-1} = u_{k_0,s_0+1} = M$ a celou úvahu je možno zopakovat pro uzly x_{k_0-1,s_0}, x_{k_0+1,s_0}, x_{k_0,s_0-1} a x_{k_0,s_0+1}. Protože množina $\Omega^{(h)}$ je souvislá, dostaneme tak postupně, že platí

$$q(x_k, y_s) = 0 \quad \text{pro} \quad (x_k, y_s) \in \Omega^{(h)} \tag{2.11}$$

a

$$u_{ks} = M \quad \text{pro} \quad (x_k, y_s) \in \bar{\Omega}^{(h)}. \tag{2.12}$$

Není-li funkce q identicky rovna nule, není rovnice (2.11) splněna pro uzel $(x_{\tilde{k}}, y_{\tilde{s}})$. Tento spor dokazuje, že v případě, že funkce q není rovna nule identicky, platí $u_{ks} < M$ pro $(x_k, y_s) \in \Omega^{(h)}$. V tomto případě je tedy lemma dokázáno. V případě, že je $q(x) \equiv 0$ v Ω, však plyne tvrzení lemmatu ihned z (2.12).

Při zavedení operátorů $L_h^{(1)}$ a $L_h^{(2)}$ jsme postupovali tak, že jsme derivace, které se vyskytují v aproximovaném operátoru nahradili diferenčními podíly. Odtud také pocházejí názvy *diferenční metody* a *diferenční rovnice*, které často v souvislosti s metodou sítí užíváme. Vyskytuje-li se v dané diferenciální rovnici kromě derivací $\partial^2 u/\partial x^2$ a $\partial^2 u/\partial y^2$ i smíšená derivace $\partial^2 u/(\partial x \partial y)$, lze ji aproximovat pomocí rovnice

$$\begin{aligned} \frac{\partial^2 u}{\partial x \partial y}(x_k, y_s) = \frac{1}{2h^2}[& u(x_k+h, y_s+h) - u(x_k-h, y_s+h) + \\ & + u(x_k-h, y_s-h) - u(x_k+h, y_s-h)] + O(h^2). \end{aligned} \tag{2.13}$$

Tuto rovnici dostaneme tak, že položíme $z = \partial u/\partial y$, derivaci $\partial z/\partial x$ v bodě (x_k, y_s) aproximujeme podílem $[z(x_k+h, y_s) - z(x_k-h, y_s)]/(2h)$ a hodnoty funkce z v bodě (x_k+h, y_s) a (x_k-h, y_s), tj. hodnoty derivace $\partial u/\partial y$ v těchto bodech, podíly $[u(x_k+h, y_s+h) - u(x_k+h, y_s-h)]/(2h)$ a $[u(x_k-h, y_s+h) - u(x_k-h, y_s-h)]/(2h)$. V tomto případě pak vzniklý operátor svazuje hodnoty přibližného řešení nejen v uzlu (x_k, y_s) a v uzlech (x_{k+1}, y_s), (x_{k-1}, y_s), (x_k, y_{s+1}) a (x_k, y_{s-1}), ale navíc ještě v uzlech (x_{k+1}, y_{s+1}), (x_{k-1}, y_{s+1}), (x_{k+1}, y_{s-1}) a (x_{k-1}, y_{s-1}). Všechny tyto uzly je tedy třeba pokládat za sousedy uzlu (x_k, y_s). Samozřejmě, že se pak změní množina vnitřních uzlů $\Omega^{(h)}$ i množina hraničních uzlů $\Gamma^{(h)}$, neboť obě tyto množiny jsou definovány pomocí pojmu sousedního uzlu.

Operátory $L_h^{(1)}$ a $L_h^{(2)}$ lze také sestrojit následujícím postupem, zvaným *metoda neurčitých koeficientů*. Tento postup je obecnější a je použitelný v podstatě pro libovolnou síť, ať pravidelnou či nepravidelnou. Při tomto postupu stanovíme ke každému uzlu P skupinu uzlů Q_i, $i = 1, \ldots, r_P$, které nazveme sousedními k uzlu P. Dále hledáme operátor L_h jako výraz tvaru

$$(L_h u)_P = \sum_{i=1}^{r_P} \sigma(P, Q_i) u(Q_i), \quad (Q_0 = P), \tag{2.14}$$

přičemž koeficienty $\sigma(P, Q_i)$ určíme tak, aby Taylorův rozvoj výrazu na pravé straně rovnice (2.14) vzhledem k bodu P souhlasil až na členy vyššího řádu s hodnotou

daného diferenciálního operátoru v bodě P. Je-li např. diferenciální operátor, který chceme aproximovat, dán rovnicí

$$Lu = -a_{11}\frac{\partial^2 u}{\partial x^2} - 2a_{12}\frac{\partial^2 u}{\partial x \partial y} - a_{22}\frac{\partial^2 u}{\partial y^2} + b_1\frac{\partial u}{\partial x} + b_2\frac{\partial u}{\partial y} + cu, \tag{2.15}$$

mají-li uzly Q_i souřadnice (x_i, y_i) a předpokládáme-li, že uzel P má souřadnice $(0,0)$ (což zřejmě není na újmu na obecnosti), platí

$$\begin{aligned} u(Q_i) = u(x_i, y_i) = u(P) + x_i\frac{\partial u}{\partial x}(P) + y_i\frac{\partial u}{\partial y}(P) + \\ + \tfrac{1}{2}x_i^2\frac{\partial^2 u}{\partial x^2}(P) + x_i y_i\frac{\partial^2 u}{\partial x \partial y}(P) + \tfrac{1}{2}y_i^2\frac{\partial^2 u}{\partial y^2}(P) + \dots, \end{aligned} \tag{2.16}$$

takže po dosazení do rovnice (2.14) máme

$$\begin{aligned} (L_h u)(P) = c^{(P)}u(P) + b_1^{(P)}\frac{\partial u}{\partial x}(P) + b_2^{(P)}\frac{\partial u}{\partial y}(P) - \\ - a_{11}^{(P)}\frac{\partial^2 u}{\partial x^2}(P) - 2a_{12}^{(P)}\frac{\partial^2 u}{\partial x \partial y}(P) - a_{22}^{(P)}\frac{\partial^2 u}{\partial y^2}(P) + \dots, \end{aligned} \tag{2.17}$$

kde

$$\begin{aligned} c^{(P)} &= \sum_{i=0}^{r_p}\sigma(P,Q_i), \quad b_1^{(P)} = \sum_{i=0}^{r_p} x_i\sigma(P,Q_i), \\ b_2^{(P)} &= \sum_{i=0}^{r_p} y_i\sigma(P,Q_i), \quad a_{11}^{(P)} = -\tfrac{1}{2}\sum_{i=0}^{r_p} x_i^2\sigma(P,Q_i), \\ a_{12}^{(P)} &= -\tfrac{1}{2}\sum_{i=0}^{r_p} x_i y_i\sigma(P,Q_i), \quad a_{22}^{(P)} = -\tfrac{1}{2}\sum_{i=0}^{r_p} y_i^2\sigma(P,Q_i). \end{aligned} \tag{2.18}$$

K tomu, aby operátor L_h daný rovnicí (2.14) aproximoval operátor L ze vzorce (2.15) stačí, aby platily rovnice

$$\begin{aligned} c^{(P)} &= c(P), \quad b_1^{(P)} = b_1(P), \quad b_2^{(P)} = b_2(P), \\ a_{11}^{(P)} &= a_{11}(P), \quad a_{12}^{(P)} = a_{12}(P), \quad a_{22}^{(P)} = a_{22}(P). \end{aligned} \tag{2.19}$$

Soustava (2.18) je soustava šesti rovnic pro neznámé $\sigma(P,Q_i)$. K jejich splnění je tedy obecně třeba šesti hodnot $\sigma(P,Q_i)$. To tedy znamená, že k aproximaci operátoru L ze vzorce (2.15) je třeba, aby každý uzel měl nejméně pět sousedů. V případě konstrukce operátorů $L_h^{(1)}$ a $L_h^{(2)}$ jsme vystačili pouze se čtyřmi sousedy. Bylo to proto, že diferenciální rovnice (2.1) neobsahuje smíšenou derivaci $\partial^2 u/(\partial x \partial y)$.

Řád aproximace operátoru L_h sestrojeného právě popsaným postupem je třeba vypočítat pro konkrétní volbu sousedů uzlu P. Obecně se dá říci, že nebude horší než $O(h)$, označíme-li zde písmenem h nejdelší z úseček $\overline{PQ_i}$. K jeho zlepšení je třeba vzít více bodů Q_i nebo je umístit symetricky. Tím totiž můžeme dosáhnout

toho, že několik dalších (nevypsaných) členů v rozvoji (2.17) vymizí. Jako příklad uveďme aproximaci Laplaceova operátoru

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \tag{2.20}$$

operátorem Δ_h definovaným rovnicí

$$\begin{aligned} (\Delta_h)_{ks} = \frac{1}{6h^2}[4(u_{k+1,s} + u_{k-1,s} + u_{k,s+1} + u_{k,s-1}) \\ + (u_{k+1,s+1} + u_{k+1,s-1} + u_{k-1,s+1} + u_{k-1,s-1}) - 20u_{ks}]. \end{aligned} \tag{2.21}$$

Pro tento operátor totiž platí (za předpokladu dostatečné hladkosti funkce u)

$$\begin{aligned} (\Delta_h u^{(\mathrm{pr})})_{ks} = (\Delta u)(x_k, y_s) + \tfrac{1}{12}h^2(\Delta\Delta u)(x_k, y_s) + \\ + \tfrac{1}{360}h^4\Big[(\Delta^3 u)(x_k, y_s) + 2\frac{\partial^4 \Delta u}{\partial x^2 \partial y^2}(x_k, y_s)\Big] + O(h^6). \end{aligned} \tag{2.22}$$

Řád aproximace je zde $O(h^6)$, za sousedy uzlu (x_k, y_s) je však třeba vzít osm uzlů (x_{k+1}, y_s), (x_{k-1}, y_s), (x_k, y_{s+1}), (x_k, y_{s-1}), (x_{k+1}, y_{s+1}), (x_{k-1}, y_{s+1}), (x_{k-1}, y_{s-1}) a (x_{k+1}, y_{s-1}), tj. všechny uzly, které mají od uzlu (x_k, y_s) vzdálenost menší nebo rovnu číslu $2^{1/2}h$.

Při řešení okrajových úloh pro obyčejné diferenciální rovnice metodou sítí jsme pro sestavování diferenčních rovnic doporučovali také postup opírající se o integrální identity platící pro řešení dané diferenciální rovnice. I zde je analogický postup možný, příslušné identity jsou však formálně podstatně komplikovanější. Jako příklad uveďme takovou identitu pro diferenciální rovnici (2.1).

Věta 2.3. *Nechť funkce u splňuje v oblasti Ω diferenciální rovnici* (2.1) *a nechť pro uzavřený čtverec R_{ks} o vrcholech (x_{k+1}, y_{s+1}), (x_{k-1}, y_{s+1}), (x_{k-1}, y_{s-1}), (x_{k+1}, y_{s-1}) platí $R_{ks} \subset \bar{\Omega}$. Pak platí*

$$\begin{aligned} &\int_{y_{s-1}}^{y_{s+1}} \Big\{ \frac{Q(x_{k-1},y)u(x_{k-1},y)}{\int_{x_{k-1}}^{x_k}\frac{1}{p(x,y)}\,\mathrm{d}x} + \frac{Q(x_{k+1},y)u(x_{k+1},y)}{\int_{x_k}^{x_{k+1}}\frac{1}{p(x,y)}\,\mathrm{d}x} - \\ &- Q(x_k,y)u(x_k,y)\Big[\frac{1}{\int_{x_{k-1}}^{x_k}\frac{1}{p(x,y)}\,\mathrm{d}x} + \frac{1}{\int_{x_k}^{x_{k+1}}\frac{1}{p(x,y)}\,\mathrm{d}x}\Big]\Big\}\,\mathrm{d}y + \\ &+ \int_{x_{k-1}}^{x_{k+1}} \Big\{ \frac{P(x,y_{s-1})u(x,y_{s-1})}{\int_{y_{s-1}}^{y_s}\frac{1}{p(x,y)}\,\mathrm{d}y} + \frac{P(x,y_{s+1})u(x,y_{s+1})}{\int_{y_s}^{y_{s+1}}\frac{1}{p(x,y)}\,\mathrm{d}y} - \\ &- P(x,y_s)u(x,y_s)\Big[\frac{1}{\int_{y_{s-1}}^{y_s}\frac{1}{p(x,y)}\,\mathrm{d}y} + \frac{1}{\int_{y_s}^{y_{s+1}}\frac{1}{p(x,y)}\,\mathrm{d}y}\Big]\Big\}\,\mathrm{d}x + \end{aligned} \tag{2.23}$$

$$+\iint_{R_{ks}}[f(x,y)-q(x,y)u(x,y)]P(x,y)Q(x,y)\,\mathrm{d}x\,\mathrm{d}y=$$

$$=\iint_{R_{ks}}p(x,y)\Big[P(x,y)\frac{\partial Q(x,y)}{\partial x}\frac{\partial u(x,y)}{\partial x}+Q(x,y)\frac{\partial P(x,y)}{\partial y}\frac{\partial u(x,y)}{\partial y}\Big]\,\mathrm{d}x\,\mathrm{d}y+$$

$$+\int_{x_{k-1}}^{x_{k+1}}\Bigg[\frac{1}{\int_{y_s}^{y_{s+1}}\frac{1}{p(x,y)}\,\mathrm{d}y}\int_{y_s}^{y_{s+1}}\frac{\partial P(x,y)}{\partial y}u(x,y)\,\mathrm{d}y-$$

$$-\frac{1}{\int_{y_{s-1}}^{y_s}\frac{1}{p(x,y)}\,\mathrm{d}y}\int_{y_{s-1}}^{y_s}\frac{\partial P(x,y)}{\partial y}u(x,y)\,\mathrm{d}y\Bigg]\,\mathrm{d}x+$$

$$+\int_{y_{s-1}}^{y_{s+1}}\Bigg[\frac{1}{\int_{x_k}^{x_{k+1}}\frac{1}{p(x,y)}\,\mathrm{d}x}\int_{x_k}^{x_{k+1}}\frac{\partial Q(x,y)}{\partial x}u(x,y)\,\mathrm{d}x-$$

$$-\frac{1}{\int_{x_{k-1}}^{x_k}\frac{1}{p(x,y)}\,\mathrm{d}x}\int_{x_{k-1}}^{x_k}\frac{\partial Q(x,y)}{\partial x}u(x,y)\,\mathrm{d}x\Bigg]\,\mathrm{d}y,$$

kde funkce P a Q jsou definovány vzorci

$$(2.24)\quad P(x,y)=\begin{cases}\dfrac{\int_x^{x_{k+1}}\frac{1}{p(\xi,y)}\,\mathrm{d}\xi}{\int_{x_k}^{x_{k+1}}\frac{1}{p(x,y)}\,\mathrm{d}x}, & x_k\leqq x\leqq x_{k+1},\ y_{s-1}\leqq y\leqq y_{s+1},\\[2ex] \dfrac{\int_{x_{k-1}}^{x}\frac{1}{p(\xi,y)}\,\mathrm{d}\xi}{\int_{x_{k-1}}^{x_k}\frac{1}{p(x,y)}\,\mathrm{d}x}, & x_{k-1}\leqq x\leqq x_k,\ y_{s-1}\leqq y\leqq y_{s+1},\end{cases}$$

$$Q(x,y)=\begin{cases}\dfrac{\int_y^{y_{s+1}}\frac{1}{p(x,\eta)}\,\mathrm{d}\eta}{\int_{y_s}^{y_{s+1}}\frac{1}{p(x,y)}\,\mathrm{d}y}, & x_{k-1}\leqq x\leqq x_{k+1},\ y_s\leqq y\leqq y_{s+1},\\[2ex] \dfrac{\int_{y_{s-1}}^{y}\frac{1}{p(x,\eta)}\,\mathrm{d}\eta}{\int_{y_{s-1}}^{y_s}\frac{1}{p(x,y)}\,\mathrm{d}y}, & x_{k-1}\leqq x\leqq x_{k+1},\ y_{s-1}\leqq y\leqq y_s.\end{cases}$$

D ů k a z . Vynásobíme-li rovnici (2.1) funkcí v definovanou vzorcem

$$(2.25)\qquad v(x,y)=P(x,y)Q(x,y)$$

a integrujeme přes čtverec R_{ks}, dostaneme identitu

$$(2.26)\quad -\iint_{R_{ks}}\Big\{\frac{\partial}{\partial x}\Big[p(x,y)\frac{\partial u(x,y)}{\partial x}\Big]+\frac{\partial}{\partial y}\Big[p(x,y)\frac{\partial u(x,y)}{\partial y}\Big]\Big\}v(x,y)\,\mathrm{d}x\,\mathrm{d}y=$$
$$=\iint_{R_{ks}}[f(x,y)-q(x,y)u(x,y)]v(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Abychom dostali identitu (2.23), stačí v rovnici (2.26) dvakrát integrovat per partes a uvědomit si, že funkce definovaná vzorcem (2.25) je rovna nule na hranici čtverce R_{ks} a že její derivace mají skoky na úsečkách $x=x_k$, $y_{s-1}\leqq y\leqq y_{s+1}$ a $y=y_s$, $x_{k-1}\leqq x\leqq x_{k+1}$. Výpočty jsou sice zdlouhavé, ale natolik triviální, že je není třeba podrobně provádět.

2.1.2 Přepis okrajových podmínek a konvergence vzniklých metod

V předešlém odstavci jsme ukázali několik možností, jak sestrojit konečnědimenzionální operátory, které aproximují lineární diferenciální operátory druhého řádu. Z úvah tam provedených se ukázalo rozumné hledat přibližné řešení modelové diferenciální rovnice (2.1) ze soustavy rovnic

$$(2.27)\qquad (L_h^{(2)}u)_{ks}=f(x_k,y_s),\quad (x_k,y_s)\in\Omega^{(h)},$$

kde operátor $L_h^{(2)}$ je definován rovnicí (2.4). Těchto rovnic je tolik, kolik je vnitřních uzlů, zatímco počet neznámých je roven počtu vnitřních a hraničních uzlů dohromady. K soustavě (2.27) je tedy třeba přidat ještě tolik rovnic, kolik prvků má množina hraničních uzlů $\Gamma^{(h)}$, a tyto rovnice je nutno získat z okrajových podmínek. Symbolicky je budeme zapisovat ve tvaru

$$(2.28)\qquad (l_hu)_{ks}=(\Lambda_h\gamma)_{ks},\quad (x_k,y_s)\in\Gamma^{(h)}.$$

Zde l_h je operátor, který funkci definované na množině $\bar\Omega^{(h)}$ přiřazuje funkci definovanou na množině $\Gamma^{(h)}$, γ je pravá strana okrajové podmínky 1.2 a Λ_h je operátor, který funkci definované na Γ přiřazuje funkci definovanou na $\Gamma^{(h)}$. Některé způsoby, jak konkrétně sestrojit tyto operátory l_h a Λ_h ukážeme v tomto odstavci.

Začneme Dirichletovou okrajovou podmínkou, tj. podmínkou tvaru

$$(2.29)\qquad u(x,y)=\gamma(x,y),\quad (x,y)\in\Gamma.$$

Okrajová podmínka tohoto typu nepředstavovala v jednodimenzionálním případě žádný problém, neboť v hraničních uzlech jsme znali přibližné řešení přesně. Zde tomu obecně tak není, neboť hraniční uzly nemusí ležet na hranici oblasti Ω. V každém případě však leží blízko hranice v tom smyslu, že ke každému hraničnímu uzlu existuje bod na hranici oblasti Ω, jehož vzdálenost od uvažovaného uzlu je nanejvýš h. Tato skutečnost nabízí elementární způsob přepisu Dirichletových okrajových podmínek, při němž položíme

$$(2.30)\qquad (l_hu)_{ks}=u_{ks}\quad\text{pro}\quad (x_k,y_s)\in\Gamma^{(h)}$$

a

$$(2.31)\qquad (\Lambda_h\gamma)_{ks}=\varphi(x_k,y_s)\quad\text{pro}\quad (x_k,y_s)\in\Gamma^{(h)}.$$

kde φ je libovolná spojitá a spojitě diferencovatelná funkce v $\bar{\Omega}$ a taková, že pro ni platí

$$\varphi(x,y) = \gamma(x,y), \quad (x,y) \in \Gamma. \tag{2.32}$$

Ukažme předně, že soustava (2.27) doplněná právě sestrojenými rovnicemi typu (2.28) skutečně jednoznačně určuje přibližné řešení, tj. že její matice je regulární. Toto tvrzení je bezprostředním důsledkem následujícího lemmatu.

Lemma 2.2. *Matice soustavy* (2.27), (2.28), *kde operátory* $L_h^{(2)}$, l_h *a* Λ_h *jsou definovány rovnicemi* (2.4), (2.30) *a* (2.31), *je monotónní.*

D ů k a z . Buď η funkce definovaná na množině $\bar{\Omega}^{(h)}$ a taková, že pro ni platí

$$(L_h^{(2)}\eta)_{ks} \leqq 0, \quad (x_k, y_s) \in \Omega^{(h)}, \tag{2.33}$$

a

$$(l_h\eta)_{ks} \leqq 0, \quad (x_k, y_s) \in \Gamma^{(h)}. \tag{2.34}$$

Máme dokázat, že je $\eta_{ks} \leqq 0$ pro $(x_k, y_s) \in \bar{\Omega}^{(h)}$. To však plyne ihned z principu maxima (viz lemma 2.1), neboť rovnice (2.34) neznamenají nic jiného, než že funkce η je na množině $\Gamma^{(h)}$ nekladná. Lemma je dokázáno.

O konvergenci sestrojené metody sítí platí následující věta.

Věta 2.4. *Nechť řešení diferenciální rovnice* (2.1) *s okrajovou podmínkou* (2.29) *existuje a má v* $\bar{\Omega}$ *spojité parciální derivace podle* x *a* y *až do čtvrtého řádu včetně. Nechť dále existuje funkce* z, *která je v* Ω *řešením diferenciální rovnice*

$$-\frac{\partial}{\partial x}\Big(p(x,y)\frac{\partial u}{\partial x}\Big) - \frac{\partial}{\partial y}\Big(p(x,y)\frac{\partial u}{\partial y}\Big) = 1 \tag{2.35}$$

s okrajovou podmínkou

$$z(x,y) = 1, \quad (x,y) \in \Gamma, \tag{2.36}$$

a která má v $\bar{\Omega}$ *spojité parciální derivace až do čtvrtého řádu včetně. Nechť konečně* u_{ks} *je přibližné řešení získané ze soustavy* (2.27) *a* (2.28), *kde operátory* $L_h^{(2)}$, l_h *a* Λ_h *jsou definovány rovnicemi* (2.4), (2.30) *a* (2.31). *Pak existují konstanty* M *a* $h_0 > 0$ *takové, že platí*

$$|u_{ks} - u(x_k, y_s)| \leqq Mh \tag{2.37}$$

pro každé $h \leqq h_0$ *a pro každý uzel* $(x_k, y_s) \in \bar{\Omega}^{(h)}$.

D ů k a z . Položíme-li $\eta_{ks} = u_{ks} - u(x_k, y_s)$, existuje za uvedených předpokladů podle věty 2.2 konstanta K_1 taková, že platí

$$(L_h^{(2)}\eta)_{ks} = \varepsilon_{ks}, \quad (x_k, y_s) \in \Omega^{(h)}, \tag{2.38}$$

a

$$|\varepsilon_{ks}| \leqq K_1 h^2. \tag{2.39}$$

Podle (2.30) je

$$(l_h\eta)_{ks} = \varphi(x_k, y_s) - u(x_k, y_s), \quad (x_k, y_s) \in \Gamma^{(h)}. \tag{2.40}$$

V důsledku definice hraničních uzlů existuje ke každému uzlu $(x_k, y_s) \in \Gamma^{(h)}$ bod $(\tilde{x}, \tilde{y}) \in \Gamma$ takový, že jeho vzdálenost od uzlu (x_k, y_s) je nanejvýš h. Protože však funkce u a φ mají spojité derivace, plyne z věty o střední hodnotě, že platí

$$u(x_k, y_s) = u(\tilde{x}, \tilde{y}) + O(h) \tag{2.41}$$

a

$$\varphi(x_k, y_s) = \varphi(\tilde{x}, \tilde{y}) + O(h). \tag{2.42}$$

Protože však je $\varphi(\tilde{x}, \tilde{y}) = \gamma(\tilde{x}, \tilde{y})$, dostáváme odtud celkem, že je

$$(l_h\eta)_{ks} = \varepsilon_{ks}, \quad (x_k, y_s) \in \Gamma^{(h)} \tag{2.43}$$

a

$$|\varepsilon_{ks}| \leqq K_1 h. \tag{2.44}$$

Ze stejných důvodů však existuje konstanta K_2 taková, že pro funkci z platí

$$\begin{aligned} (L_h^{(2)} z^{(\mathrm{pr})})_{ks} &= 1 + \tilde{\varepsilon}_{ks}, \quad (x_k, y_s) \in \Omega^{(h)}, \\ (l_h z^{(\mathrm{pr})})_{ks} &= 1 + \tilde{\varepsilon}_{ks}, \quad (x_k, y_s) \in \Gamma^{(h)}, \end{aligned} \tag{2.45}$$

kde $z^{(\mathrm{pr})}$ je funkce definovaná na množině $\bar{\Omega}^{(h)}$ předpisem $z_{ks}^{(\mathrm{pr})} = z(x_k, y_s)$ a

$$\begin{aligned} |\tilde{\varepsilon}_{ks}| &\leqq K_2 h^2, \quad (x_k, y_s) \in \Omega^{(h)}, \\ |\tilde{\varepsilon}_{ks}| &\leqq K_2 h, \quad (x_k, y_s) \in \Gamma^{(h)}, \end{aligned} \tag{2.46}$$

Postupem úplně stejným jako v důkazu věty 3.7 z kap. II (viz str. 179) nyní už snadno dokážeme, že existuje konstanta N taková, že pro funkci $v_{ks} = Nz(x_k, y_s)h$ platí

$$\begin{aligned} (L_h^{(2)} v)_{ks} &\geqq K_1 h^2 \geqq |(L_h^{(2)}\eta)_{ks}|, \quad (x_k, y_s) \in \Omega^{(h)}, \\ (l_h v)_{ks} &\geqq K_1 h \geqq |(l_h\eta)_{ks}|, \quad (x_k, y_s) \in \Gamma^{(h)}, \end{aligned} \tag{2.47}$$

pro každé dostatečně malé h. Z těchto nerovností a z lemmatu 2.2 (srv. také lemma 3.2 z kap. II) však už plyne, že funkce v_{ks} majorizuje funkci η_{ks}. Protože funkce z je za uvedených předpokladů omezená v $\bar{\Omega}$, plyne odtud tvrzení věty.

Poznamenejme, že existence funkce z ve větě požadovaných vlastností není podstatná pro tvrzení této věty a váže se pouze na zvolenou techniku důkazu. Jiným postupem, podstatně však komplikovanějším, by bylo možno dokázat větu 2.4 i bez funkce z.

Aproximace řešení dané okrajové úlohy získané popsaným postupem tedy konvergují při $h \to 0$ k přesnému řešení, rychlost konvergence je však obecně pouze $O(h)$ a chyba může navíc dosti podstatně záviset na charakteru prodloužení funkce γ. Z těchto důvodů se tento postup užívá jen zřídka. Když už je použit, doporučuje se volit množinu $\Omega^{(h)}$ ne striktně tak, jak jsme ji definovali v odst. 2.1.1, ale připustit do ní ještě některé další body ležící v oblasti Ω. Příslušné hraniční uzly pak mohou samozřejmě ležet i vně množiny $\bar{\Omega}$, můžeme však dosáhnout toho, že množina $\Gamma^{(h)}$ lépe vystihuje hranici Γ dané oblasti. Následující příklad ukazuje, že to může přinést až překvapivý efekt.

Příklad 2.1. Řešme Dirichletův problém pro rovnici $\Delta u = 0$ na jednotkovém kruhu $\Omega = \{(x, y);\ x^2 + y^2 < 1\}$ pro okrajovou podmínku $\gamma(\cos\theta, \sin\theta) = \sin 4\theta$, $0 < \theta \leqq 2\pi$. Přesné řešení této okrajové úlohy je funkce $u(x, y) = 4xy(x^2 - y^2)$. Protože v tomto případě je $\partial^4 u/\partial x^4 = \partial^4 u/\partial y^4 = 0$, jsou veličiny ε_{ks} pro $(x_k, y_s) \in \in \Omega^{(h)}$ rovny nule a celková chyba přibližného řešení je způsobena pouze nepřesným splněním okrajových podmínek. Funkci γ prodloužíme do bodů blízkých k hranici předpisem $\varphi(r\cos\theta, r\sin\theta) = \sin 4\theta$, $0 < \theta \leqq 2\pi$ a množiny $\Omega^{(h)}$ a $\Gamma^{(h)}$ zvolíme podle obr. 2.1a, 2.1b a 2.1c, v nichž jsou body z $\Gamma^{(h)}$ označeny křížky. Výsledky v bodech, které jsou v obrázcích označeny čísly 1 až 10 jsou uspořádány v tab. 2.1. Pro porovnání jsou uvedeny také hodnoty přesného řešení a hodnoty přibližného řešení s množinou $\Omega^{(h)}$ zvolenou podle obr. 2.1a a přepisem okrajových podmínek podle Collatze, který bude popsán v dalším textu.

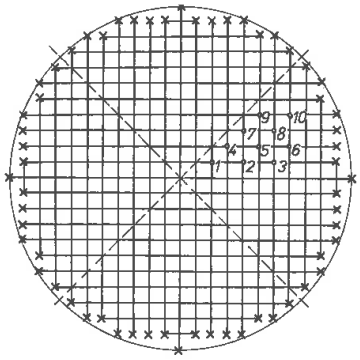
Tabulka 2.1

Závislost řešení na volbě hraničních uzlů

Bod	Užitá metoda			Přesné řešení	Přibližné řešení s přepisem okrajových podmínek dle Collatze
	a	*b*	*c*		
1	0,002 046 0	0,001 446 7	0,001 801 4	0,001 639 2	0,000 464 8
2	0,020 475 9	0,014 471 1	0,018 018 2	0,016 392 3	0,015 268 2
3	0,072 209 8	0,050 778 5	0,063 196 7	0,057 373 1	0,058 897 3
4	0,010 213 6	0,007 229 7	0,009 002 3	0,008 196 2	0,006 972 4
5	0,071 511 7	0,050 611 8	0,063 039 0	0,057 373 1	0,058 565 6
6	0,216 174 5	0,152 205 9	0,190 102 1	0,172 119 4	0,177 270 1
7	0,028 480 9	0,020 216 0	0,025 155 4	0,022 948 3	0,023 316 0
8	0,164 399 3	0,116 874 6	0,145 546 4	0,132 777 8	0,136 529 9
9	0,060 620 2	0,043 225 4	0,053 530 7	0,049 177 0	0,050 521 8
10	0,309 469 8	0,221 543 2	0,273 353 2	0,252 441 8	0,259 730 8

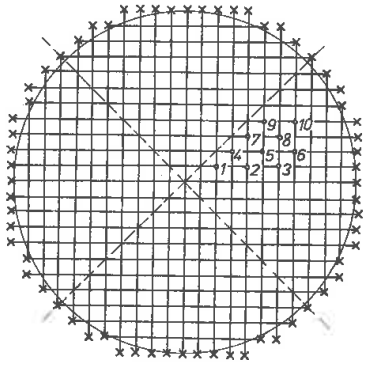
Obr. 2.1a

Speciální volba množiny $\Gamma^{(h)}$ pro příklad 2.1



Obr. 2.1b

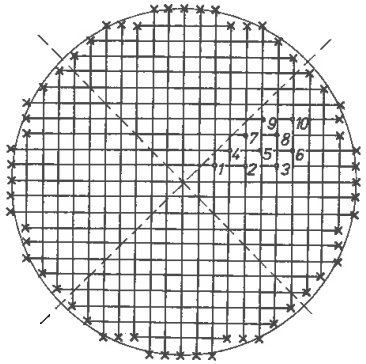
Speciální volba množiny $\Gamma^{(h)}$ pro příklad 2.1



Z důkazu věty 2.4 je zřejmé, že platí-li speciálně $\Gamma^{(h)} \subset \Gamma$, je možno volit funkci z tak, že pro ni platí $z(x, y) = 0$ pro $(x, y) \in \Gamma$. Tato volba pak umožní brát funkci v ve tvaru $Nz(x_k, y_s)h^2$ a dokázat tak rychlost konvergence $O(h^2)$. Nepříznivější výsledek, který jsme obdrželi v obecném případě, je tedy způsoben ne dosti přesným předpisem okrajových podmínek. Zároveň je také vidět, že k rychlosti konvergence $O(h^2)$ vede každý takový přepis okrajových podmínek, který nenaruší monotónnost matice vzniklé soustavy a jehož *lokální chyba* (jak se veličina ε_{ks} zavedená v dů-

Obr. 2.1c

Speciální volba množiny $\Gamma^{(h)}$ pro příklad 2.1



kazu věty 2.4 nazývá) je nejméně řádu $O(h^2)$. V dalším textu uvedeme dvě takové možnosti.

Nejprve popíšeme nejběžnější a nejčastěji používaný přepis Dirichletových okrajových podmínek, který je založen na lineární interpolaci a bývá spojován se jménem *Collatzovým*. Buď tedy A hraniční uzel. Pak musí existovat aspoň jeden uzel B, který je sousední k uzlu A a který je vnitřní a uzel C', který je sousední k uzlu A a který leží vně oblasti Ω. Uvedené uzly můžeme přitom zřejmě vybrat tak, že leží na přímce, a bez újmy na obecnosti můžeme předpokládat, že tato přímka je rovnoběžná s osou x. Buď dále C průsečík úsečky $C'A$ s hranicí Γ dané oblasti a položme $\sigma h = CA$, takže je $0 \leqq \sigma < 1$ (viz obr. 2.2a). Nechť konečně u je dostatečně hladká funkce v $\bar{\Omega}$. Položíme-li počátek souřadnic do bodu A a osu x orientujeme kladně ve směru od bodu A k bodu B, je funkce $u(A)(h-x)/h + u(B)x/h$ Lagrangeovým interpolačním polynomem pro funkci u. Platí tedy

$$u(x,y) = u(A)\frac{h-x}{h} + u(B)\frac{x}{h} + r_A \tag{2.48}$$

a je

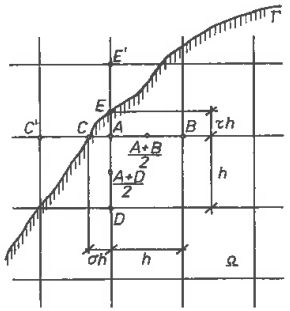
$$|r_A| \leqq Mh^2 \tag{2.49}$$

pro body ležící na úsečce $C'B$. Speciálně je tedy (pro $x = -\sigma h$ v našem souřadném systému)

$$u(C) = (1+\sigma)u(A) - \sigma u(B) + r_A. \tag{2.50}$$

Obr. 2.2a

Přepis Dirichletových okrajových podmínek



Vyjdeme-li z této rovnice, je přirozené definovat operátory l_h^C a Λ_h^C předpisem

$$(l_h^C u)_A = u_A - \frac{\sigma}{1+\sigma}u_B \tag{2.51}$$

a

$$(\Lambda_h^C \gamma)_A = \frac{1}{1+\sigma}\gamma(C). \tag{2.52}$$

Při Collatzově přepisu okrajových podmínek tedy doplníme soustavu (2.27) rovnicemi typu

$$(l_h^C u)_A = \frac{1}{1+\sigma}\gamma(C), \tag{2.53}$$

které zapíšeme pro každý hraniční uzel.

Z lemmatu 2.2 opět plyne, že matice metody sítí dané rovnicemi (2.27), (2.53) je monotónní, takže tato metoda má smysl. Její rychlost konvergence je příznivější než u metody sítí, kdy okrajové podmínky přepisujeme na základě pouhé spojitosti, a je popsána v následující větě.

Věta 2.5. *Nechť řešení u diferenciální rovnice (2.1) s okrajovou podmínkou (2.29) existuje a má čtyři spojité parciální derivace v $\bar{\Omega}$. Nechť dále existuje funkce z vlastností popsaných ve větě 2.4. Buď konečně u_{ks} řešení soustavy (2.27), (2.53). Pak existují konstanty M a $h_0 > 0$ takové, že pro každý uzel $(x_k, y_s) \in \bar{\Omega}^{(h)}$ a pro každé $h \leqq h_0$ platí*

$$|u_{ks} - u(x_k, y_s)| \leqq Mh^2. \tag{2.54}$$

D ů k a z je přesným opakování důkazu věty 2.4 s tím jediným rozdílem, že pro

operátor l_h^C platí rovnice (2.43) s lepším odhadem

$$|\varepsilon_{ks}| \leqq Mh^2. \tag{2.55}$$

Tato skutečnost umožňuje brát funkci v_{ks} splňující nerovnosti (2.47) ve tvaru $Nz(x_k, y_s)h^2$.

Postulování existence funkce z požadovaných vlastností opět není nutné a důmyslnější důkazovou technikou se mu lze vyhnout.

Přepis Dirichletových okrajových podmínek založený na lineární interpolaci je tedy z hlediska řádové přesnosti výhodný a ve srovnání s přepisem založeným na spojitosti není o mnoho komplikovanější. Proto se jej v praxi také velice často užívá. Přesto má však podle našeho názoru určité nevýhody. Předně při sestavování rovnic typu (2.53) není v některých případech jasné, jak vybrat směr, v němž máme interpolovat. Tak např. nastává-li situace jako na obr. 2.2a, můžeme stejně oprávněně jako rovnici (2.53) užít rovnici

$$u_A - \frac{\tau}{1+\tau} u_D = \frac{1}{1+\tau}\gamma(E), \tag{2.56}$$

která svazuje uzly A a D. Kromě toho, už vícekrát jsme zdůraznili, že při diskretizaci kteréhokoliv problému je žádoucí zachovat co nejvíce jeho původních vlastností. Tak např. operátory $L_h^{(1)}$ a $L_h^{(2)}$ jsme v předchozím odstavci sestrojili tak, aby byla zachována platnost principu maxima. Přednost jsme však dali operátoru $L_h^{(2)}$, který vedl na symetrickou soustavu lineárních rovnic, takže lépe odpovídal tomu, že původní rovnice je samoadjungovaná. Připojíme-li však k rovnicím (2.27) rovnice typu (2.53), je matice takto vzniklé soustavy nesymetrická. Proto se zdá přirozené vyjít sice z myšlenky lineární interpolace, a tím dosáhnout i stejné řádové rychlosti konvergence jako výše, modifikovat však užitý postup tak, abychom získali symetrii matice soustavy pro určení neznámých u_{ks}. Takový postup nyní popíšeme. Rozlišujme přitom tři případy charakteru hraničního uzlu tak, jak jsou vyznačeny na obr. 2.2a, 2.2b a 2.2c.

(a) Charakter hraničního uzlu podle obr. 2.2a: V tomto případě užijeme rovnici

$$\begin{aligned}\Big[\frac{1+\sigma}{\sigma} p\Big(\frac{A+B}{2}\Big) + \frac{1+\tau}{\tau} p\Big(\frac{A+D}{2}\Big)\Big] u_A - \\ - p\Big(\frac{A+B}{2}\Big) u_B - p\Big(\frac{A+D}{2}\Big) u_D = \\ = \frac{1}{\sigma} p\Big(\frac{A+B}{2}\Big)\varphi(C) + \frac{1}{\tau} p\Big(\frac{A+D}{2}\Big)\varphi(E),\end{aligned} \tag{2.57}$$

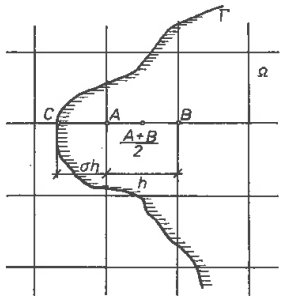
kterou dostaneme užitím rovnice (2.53) vynásobené činitelem $(1+\sigma)p((A+B)/2)/\sigma$ a rovnice (2.56) vynásobené číslem $(1+\tau)p((A+D)/2)/\tau$. Lokální chyba této rovnice je tedy zřejmě opět $O(h^2)$.

(b) Charakter hraničního uzlu podle obr. 2.2b: Zde užijeme rovnici

$$\frac{1+\sigma}{\sigma} p\Big(\frac{A+B}{2}\Big) u_A - p\Big(\frac{A+B}{2}\Big) u_B = \frac{1}{\sigma} p\Big(\frac{A+B}{2}\Big)\varphi(C), \tag{2.58}$$

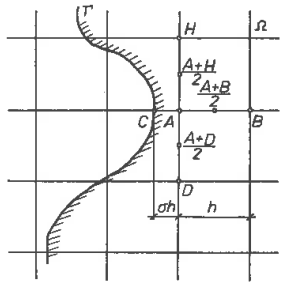
Obr. 2.2b

Přepis Dirichletových okrajových podmínek



Obr. 2.2c

Přepis Dirichletových okrajových podmínek



která je pouhým násobkem rovnice (2.53).

(c) Charakter hraničního uzlu podle obr. 2.2c: V tomto případě užijeme rovnici

$$\begin{aligned}\Big[\frac{1+\sigma}{\sigma} p\Big(\frac{A+B}{2}\Big) + p\Big(\frac{A+D}{2}\Big) + p\Big(\frac{A+H}{2}\Big)\Big] u_A - \\ - p\Big(\frac{A+B}{2}\Big) u_B - p\Big(\frac{A+D}{2}\Big) u_D - p\Big(\frac{A+H}{2}\Big) u_H = \frac{1}{\sigma} p\Big(\frac{A+B}{2}\Big)\varphi(C).\end{aligned} \tag{2.59}$$

Rovnice (2.59) vznikla tak, že jsme k rovnici (2.53) vynásobené činitelem $(1+\sigma)p((A+B)/2)/\sigma$ přičetli rovnici

$$-p\Big(\frac{A+D}{2}\Big) u_D - p\Big(\frac{A+H}{2}\Big) u_H + \Big[p\Big(\frac{A+D}{2}\Big) + p\Big(\frac{A+H}{2}\Big)\Big] u_A = 0. \tag{2.60}$$

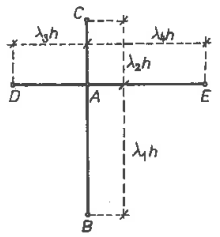
Pro přesné řešení dá totiž operace na levé straně rovnice (2.60) výsledek $O(h^2)$, jak

plyne ihned z Taylorova vzorce, takže lokální chyba rovnice (2.59) je stejně jako u předešlých rovnic řádu $O(h^2)$.

Soustava rovnic (2.27), k níž přidáme rovnice typu (2.57) až (2.59), má zřejmě symetrickou matici a pro přibližné řešení vypočtené pomocí ní platí opět věta 2.5.

Obr. 2.3

Nepravidelná síť pro aproximaci rovnice druhého řádu



Jiný často užívaný přepis Dirichletovy okrajové podmínky je založen na myšlence aproximace dané diferenciální rovnice i v hraničním uzlu, a to tak, že se užije nepravidelná síť. Popišme stručně i tento postup. Buď tedy A uzel, v němž chceme aproximovat diferenciální rovnici (2.1) pomocí bodů B, C, D a E, které jsou rozmístěny jako na obr. 2.3. Pomocí Taylorova vzorce nebo metodou neurčitých koeficientů snadno zjistíme, že rovnice

$$
\begin{aligned}
(2.61)\qquad & -\frac{\lambda_3+\lambda_4}{2\lambda_1}p\Big(\frac{A+B}{2}\Big)u_B-\frac{\lambda_3+\lambda_4}{2\lambda_2}p\Big(\frac{A+C}{2}\Big)u_C- \\
& -\frac{\lambda_1+\lambda_2}{2\lambda_3}p\Big(\frac{A+D}{2}\Big)u_D-\frac{\lambda_1+\lambda_2}{2\lambda_4}p\Big(\frac{A+E}{2}\Big)u_E+ \\
& +\Big[\frac{\lambda_3+\lambda_4}{2\lambda_1}p\Big(\frac{A+B}{2}\Big)+\frac{\lambda_3+\lambda_4}{2\lambda_2}p\Big(\frac{A+C}{2}\Big)+ \\
& +\frac{\lambda_1+\lambda_2}{2\lambda_3}p\Big(\frac{A+D}{2}\Big)+\frac{\lambda_1+\lambda_2}{2\lambda_4}p\Big(\frac{A+E}{2}\Big)+ \\
& +\tfrac{1}{4}(\lambda_1+\lambda_2)(\lambda_3+\lambda_4)h^2q(A)\Big]u_A= \\
& =\tfrac{1}{4}(\lambda_1+\lambda_2)(\lambda_3+\lambda_4)h^2f(A)
\end{aligned}
$$

aproximuje diferenciální rovnici (2.1) s přesností $O(h)$. Rovnici tohoto typu užijeme v každém hraničním uzlu. Za body B, C, D a E z obr. 2.3 vezmeme přitom ty uzly, které jsou sousední k uzlu A a jsou vnitřní, a ty body, v nichž protínají vlákna sítě procházející uzlem A hranici. V těchto posledních bodech jsou tedy hodnoty přibližného řešení známé a dané okrajovou podmínkou. Soustava lineárních algebraických rovnic, která vznikne připojením rovnic typu (2.61) k rovnicím (2.27)

má opět monotónní (a tedy i regulární) matici a celková diskretizační chyba je opět řádu $O(h^2)$. Upozorněme však, že zmíněná matice není obecně symetrická.

Všimněme si závěrem tohoto odstavce ještě stručně problematiky přepisu okrajových podmínek obsahujících derivace. Začneme Neumannovou úlohou pro rovnici (2.1), tj. úlohou nalézt funkci u, která splňuje v Ω diferenciální rovnici (2.1), je spolu s prvními parciálními derivacemi spojitá v $\bar{\Omega}$ a pro niž platí na hranici Γ oblasti Ω

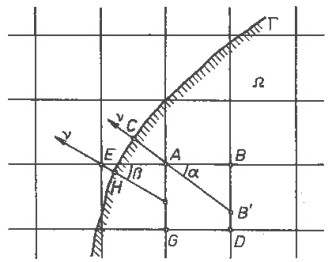
$$
(2.62)\qquad \frac{\partial u}{\partial \nu}=\gamma,
$$

kde ν je vnější normála ke hranici v daném bodě a γ je funkce definovaná a spojitá na Γ. Vzhledem k tomu, že přepis okrajových podmínek obsahujících derivace přinášel už i v jednodimenzionálním případě určité obtíže, čtenáře jistě nepřekvapí, že se vzrůstem dimenzí tyto obtíže rostou. Protože celá řada otázek týkajících se nejvhodnější formy tohoto přepisu není dosud úplně uspokojivě vyřešena, omezíme se na popis jedné velmi jednoduché možnosti, která se opírá opět o myšlenku lineární interpolace. Buď tedy A hraniční uzel. Označme C průsečík normály ke křivce Γ vedené tímto uzlem s Γ a B' ten průsečík sestrojené normály s některou přímkou sítě, který leží nejblíže uzlu A a zároveň leží uvnitř oblasti Ω. Jsou-li tyto body rozmístěné jako na obr. 2.4, platí pro úhel α, který svírá normála CA s osou x zřejmě $0 \leqq \alpha \leqq \pi/4$ a pro každou dostatečně hladkou funkci je

$$
(2.63)\qquad \frac{\partial u}{\partial \nu}(C)=\frac{\partial u}{\partial \nu}(A)+O(h)=\frac{u(A)-u(B')}{\frac{h}{\cos\alpha}}+O(h).
$$

Obr. 2.4

Přepis Neumannových okrajových podmínek



Aproximujeme-li hodnotu $u(B')$ v bodě, který nepatří do sítě, pomocí lineární interpolace, dostaneme

$$
(2.64)\qquad u(B')=(\operatorname{tg}\alpha)u(D)+(1-\operatorname{tg}\alpha)u(B)+O(h^2),
$$

takže celkem platí

$$\frac{\partial u}{\partial \nu}(C) = \frac{\cos\alpha}{h}[u(A) - (1 - \operatorname{tg}\alpha)u(B) - (\operatorname{tg}\alpha)u(D)] + O(h). \tag{2.65}$$

Zanedbáme-li v této rovnici člen $O(h)$ a dosadíme-li podle ní do okrajové podmínky (2.62), dostaneme pro hraniční uzel A rovnici

$$u_A - (1 - \operatorname{tg}\alpha)u_B - (\operatorname{tg}\alpha)u_D = \frac{h}{\cos\alpha}\gamma(C). \tag{2.66}$$

I zde se užívá modifikace popsaného postupu, kterou jsme doporučili již při přepisu Dirichletových podmínek a která spočívá v tom, že množina uzlů, ve které hledáme aproximaci daného problému, se rozšíří o některé uzly, které leží vně oblasti Ω. Tak např. v obr. 2.4 můžeme za hraniční uzly pokládat uzly E a F. Uzel A se pak stane vnitřním a např. v uzlu E dostaneme zcela analogicky jako v předchozím výkladu rovnici

$$u_E - (1 - \operatorname{tg}\beta)u_A - (\operatorname{tg}\beta)u_G = \frac{h}{\cos\beta}\gamma(H). \tag{2.67}$$

Uvedeným postupem se může podařit opět lépe vystihnout hranici dané oblasti, než když postupujeme striktně podle definic z odst. 2.1.1. Ať už užijeme podmínku typu (2.66) nebo (2.67), vznikne v případě, že funkce g v rovnici (2.1) není identicky rovna nule, soustava rovnic s monotónní maticí. Je-li funkce g identicky rovna nule, je řešitelnost vzniklé soustavy vázána na splnění jistých podmínek pro funkci γ. Jsou-li tyto podmínky splněny, je řešení určeno až na konstantu. Jednoznačným je můžeme učinit, zvolíme-li pevně hodnotu funkce u v některém uzlu. Touto volbou pak dostaneme ze soustavy (2.27) a (2.66), resp. (2.67) novou soustavu, jejíž matice je monotónní.

Pokud jde o Newtonovu okrajovou podmínku, tj. o podmínku

$$\frac{\partial u}{\partial \nu} = -k(u - \gamma) \quad \text{na } \Gamma, \tag{2.68}$$

kde k a γ jsou zadané funkce na Γ a k je kladná, je možno užít analogický postup. Tak např. v uzlu A z obr. 2.4 dostaneme rovnici

$$u_A - \frac{1 + \overline{AC}k(C)}{1 + \overline{B'C}k(C)}(1 - \operatorname{tg}\alpha)u_B - \frac{1 + \overline{AC}k(C)}{1 + \overline{B'C}k(C)}(\operatorname{tg}\alpha)u_D = \frac{\overline{AB'}k(C)}{1 + \overline{B'C}k(C)}\gamma(C), \tag{2.69}$$

jejíž lokální chyba je opět $O(h)$.

Postup pro přepis okrajových podmínek obsahujících derivace, který jsme právě popsali, vede (samozřejmě za předpokladu dostatečné hladkosti přesného řešení) k rychlosti konvergence řádu $O(h)$. Důkaz příslušných tvrzení, jejichž přesnou formulaci si snadno provede čtenář sám, je úplně analogický jako důkaz věty 2.4, a proto jej nebudeme detailně provádět. Pokud jde o postupy, pomocí nichž lze

docílit rychlosti konvergence řádu $O(h^2)$, není současná situace plně uspokojivá. Je známa sice řada formulí, pomocí nichž se této rychlosti konvergence dosáhne (je-li např. hranice složená z úseček sítě, vychází se obvykle z faktu, že výraz $[u(x+h) - u(x-h)]/(2h)$ aproximuje první derivaci s přesností $O(h^2)$), univerzální předpis však není v případě obecné křivočaré hranice dosud znám.

2.1.3 Metody zvýšené přesnosti, jiné tvary sítí

Zatím jsme při náhradě okrajové úlohy konečnědimenzionálním problémem postupovali tím nejjednodušším způsobem. K aproximaci diferenciálního operátoru jsme totiž používali diferenční operátory svazující hodnoty přibližného řešení v nejmenším možném počtu bodů. Použijeme-li k této aproximaci více bodů než je nezbytně nutné, dosáhneme přesnější aproximace daného operátoru. Metody, které takto vzniknou, se nazývají *metody zvýšené přesnosti*. Za příklad takové metody může sloužit náhrada Laplaceova operátoru operátorem Δ_h definovaným rovnicí (2.21), při níž se dopouštíme chyby řádu $O(h^6)$. Aby celková diskretizační chyba byla rovněž řádu $O(h^6)$, je třeba i okrajové podmínky aproximovat se zvýšenou přesností. Zde však vznikají obtíže, zejména pak v případě, že hranice je křivočará a že okrajové podmínky obsahují derivace. Obtíže tohoto druhu jsou pro metody zvýšené přesnosti typické a protože otázky s nimi související nejsou dosud uspokojivě řešeny, užívají se tyto metody dosti zřídka.

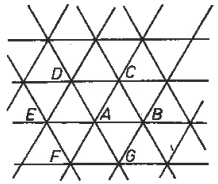
Druhá otázka, které je třeba se alespoň stručně dotknout, je problematika tvarů sítí užívaných pro řešení eliptických parciálních diferenciálních rovnic. Daleko nejběžnější jsou čtvercové sítě, jimiž jsme se zabývali až dosud. Kromě nich se někdy užívají, zejména v případech speciálních tvarů oblasti Ω, jiné pravidelné sítě jako např. trojúhelníkové, šestiúhelníkové apod. Tak např. *Poissonovu rovnici*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y), \tag{2.70}$$

(která je zřejmě speciálním případem diferenciální rovnice (2.1)) lze aproximovat na trojúhelníkové síti znázorněné na obr. 2.5 diferenční rovnicí

Obr. 2.5

Trojúhelníková síť



$$u_A - \tfrac{1}{6}(u_B + u_C + u_D + u_E + u_F + u_G) = \\ = -\tfrac{1}{4}h^2 f(A) - \tfrac{1}{64}h^4(\Delta f)(A) \tag{2.71}$$

s přesností $O(h^4)$ a na šestiúhelníkové síti z obr. 2.6 rovnicí

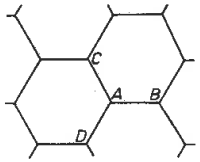
$$u_A - \tfrac{1}{3}(u_B + u_C + u_D) = -\tfrac{1}{4}h^2 f(A) \tag{2.72}$$

s přesností $O(h)$.

Nepravidelné sítě (kosoúhlé, eventuálně křivočaré) se používají nesmírně zřídka; jejich užití může být někdy ospravedlněno zjednodušením diferenčních rovnic.

Obr. 2.6

Šestiúhelníková síť



2.2 Lineární rovnice čtvrtého řádu

Problematiku užití metody sítí pro řešení okrajových úloh pro diferenciální rovnice vyšších řádů budeme ilustrovat na příkladě biharmonické rovnice

$$\Delta\Delta u \equiv \frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} = f(x,y) \tag{2.73}$$

ve čtverci $\Omega = \{(x,y);\ 0 < x < 1, 0 < y < 1\}$ s okrajovými podmínkami

$$u(x,y) = \varphi_1(x,y), \\ \frac{\partial u}{\partial \nu}(x,y) = \varphi_2(x,y) \tag{2.74}$$

na hranici Γ tohoto čtverce (ν je vnější normála).

Sestrojme stejně jako v odst. 2.1 v oblasti Ω čtvercovou síť a položme $h = 1/n$, kde n je přirozené číslo, což zaručí, že hranice Γ splyne s některými vlákny sítě. Protože operátor L_h definovaný předpisem

$$h^2(L_h u)_{ks} = 4u_{ks} - u_{k+1,s} - u_{k-1,s} - u_{k,s+1} - u_{k,s-1} \tag{2.75}$$

aproximuje Laplaceův operátor $-\Delta$ (jde o speciální případ operátoru $L_h^{(2)}$ z odst. 2.1), dá se očekávat, že operátor $L_h L_h = L_h^2$, který je tedy definován rovnicí

$$h^4(L_h^2 u)_{ks} = 20u_{ks} - 8(u_{k+1,s} + u_{k-1,s} + u_{k,s+1} + u_{k,s-1}) + \\ + 2(u_{k+1,s+1} + u_{k-1,s+1} + u_{k+1,s-1} + u_{k-1,s-1}) + \\ + (u_{k+2,s} + u_{k-2,s} + u_{k,s+2} + u_{k,s-2}), \tag{2.76}$$

aproximuje biharmonický operátor. Z následující věty plyne, že tomu tak skutečně je.

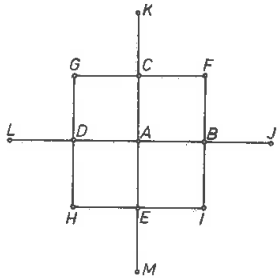
Věta 2.6. *Buď u funkce, která má v nějakém okolí U bodu (x,y) šest spojitých parciálních derivací podle obou proměnných a buď h tak malé, že body $(x+h,y)$, $(x-h,y)$, $(x,y+h)$, $x,y-h)$, $(x+h,y+h)$, $(x-h,y+h)$, $(x+h,y-h)$, $(x-h,y-h)$, $(x+2h,y)$, $(x-2h,y)$, $(x,y+2h)$ a $(x,y-2h)$ leží v U. Pak platí*

$$20u(x,y) - 8[u(x+h,y) + u(x-h,y) + u(x,y+h) + u(x,y-h)] + \\ + 2[u(x+h,y+h) + u(x-h,y+h) + u(x+h,y-h) + u(x-h,y-h)] + \\ + [u(x+2h,y) + u(x-2h,y) + u(x,y+2h) + u(x,y-2h)] = \\ = h^4(\Delta\Delta)u(x,y) + O(h^6). \tag{2.77}$$

D ů k a z tohoto tvrzení plyne ihned z Taylorova vzorce, a nebudeme jej proto podrobně provádět.

Obr. 2.7

Síťové schéma pro biharmonickou rovnici



Přibližné řešení u_{ks} budeme tedy počítat z rovnic

$$(L_h^2 u)_{ks} = f(x_k, y_s), \tag{2.78}$$

kde uzel (x_k, y_s) probíhá jako dříve množinu vnitřních uzlů $\Omega^{(h)}$, doplněných rovnicemi psanými pro každý hraniční uzel $(x_k, y_s) \in \Gamma^{(h)}$, které se získají z okrajových

podmínek. Množiny $\Omega^{(h)}$ a $\Gamma^{(h)}$ se přitom definují stejně jako v odst. 2.1 pomocí pojmu sousedních uzlů. Je však třeba mít na paměti, že za sousedy uzlu A z obr. 2.7 je třeba pokládat dvanáct uzlů $B, \dots, M$ (srv. vzorec (2.76)). V našem speciálním případě je tedy množina $\Omega^{(h)}$ tvořena uzly (x_k, y_s), pro něž je $k, s = 2, \dots, n-2$, a pro množinu $\Gamma^{(h)}$ platí $\Gamma^{(h)} = \Gamma_1^{(h)} \cup \Gamma_2^{(h)}$, kde $\Gamma_1^{(h)}$ obsahuje uzly ležící na hranici Γ čtverce Ω a $\Gamma_2^{(h)}$ obsahuje uzly, jejichž vzdálenost od Γ je právě h. Vzhledem k první okrajové podmínce (2.74) hodnoty přibližného řešení v uzlech z $\Gamma_1^{(h)}$ známe. V uzlech z $\Gamma_2^{(h)}$ zužitkujeme druhou okrajovou podmínku úplně stejně, jako jsme to učinili u okrajové úlohy pro obyčejnou diferenciální rovnici čtvrtého řádu v předešlé kapitole. Tak např. v uzlu (x_1, y_s) píšeme rovnici

$$u_{-1,s} = u_{1,s} + 2h\varphi_2(x_0, y_s) \tag{2.79}$$

a hodnotu $u_{-1,s}$ vypočtenou z této rovnice dosadíme do rovnice (2.78) psané pro uzel (x_1, y_s). Tímto postupem dostaneme zřejmě soustavu $(n-1)^2$ rovnic o $(n-1)^2$ neznámých u_{ks} v uzlech $(x_k, y_s) \in \Omega_1^{(h)}$, kde $\Omega_1^{(h)} = \Omega^{(h)} \cup \Gamma_2^{(h)}$. Každá rovnice z této soustavy je přitom buď přímo rovnice (2.78), nebo — v případě, že je $(x_k, y_s) \in \Gamma_2^{(h)}$ — rovnice (2.78), do níž se za hodnotu resp. hodnoty ležící vně Ω dosadilo z rovnic typu (2.79). Vynásobme ještě všechny rovnice číslem h^4 a matici takto vzniklé soustavy lineárních rovnic označme B_h. Tato matice je zřejmě řádu $(n-1)^2$. Je-li η libovolný $(n-1)^2$-dimenzionální vektor o složkách η_{ks}, položíme-li $\eta_{ks} = 0$ pro $(x_k, y_s) \in \Gamma_1^{(h)}$, $\eta_{-1,s} = \eta_{1s}$ a analogicky pro ostatní uzly z $\Gamma_2^{(h)}$, platí pro každou složku $(B_h\eta)_{ks}$ vektoru $B_h\eta$

$$(B_h\eta)_{ks} = h^4(L_h^2\eta)_{ks}, \quad (x_k, y_s) \in \Omega_1^{(h)}, \tag{2.80}$$

jak se snadno zjistí.

Symbolem $u^{(h)}$ označíme $(n-1)^2$-dimenzionální vektor přibližného řešení. Tento vektor je třeba vypočíst ze soustavy

$$B_h u^{(h)} = g, \tag{2.81}$$

kde g je daný vektor (jehož složky závisí na pravé straně dané diferenciální rovnice a na pravých stranách okrajových podmínek). Aby námi sestrojená metoda sítí měla vůbec smysl, je třeba odpovědět na otázku po řešitelnosti této soustavy. K tomu, ale i k úvahám o konvergenci použijeme následující lemma.

Lemma 2.3. *Existuje konstanta $\gamma > 0$ taková, že pro libovolný $(n-1)^2$-dimenzionální vektor η o složkách η_{ks} platí*

$$(B_h\eta, \eta) \geqq \gamma h^2 \|\eta\|_{h,\mathscr{L}_\infty}, \tag{2.82}$$

kde

$$\|\eta\|_{h,\mathscr{L}_\infty} = \max\Big[\max_{(x_k,y_s)\in\Omega^{(h)}} |\eta_{ks}|, \max_{(x_k,y_s)\in\Gamma_2^{(h)}} |\eta_{ks}|h^{-1}\Big]. \tag{2.83}$$

a kulaté závorky v (2.82) značí skalární součin v $(n-1)^2$-dimenzionálním euklidovském prostoru.

D ů k a z. Podle vzorce (2.80) je

$$(B_h\eta, \eta) = h^4 \sum_{(x_k,y_s)\in\Omega_1^{(h)}} (L_h^2\eta)_{ks} = h^4 \sum_{k=1}^{n-1}\sum_{s=1}^{n-1} (L_h^2\eta)_{ks}\eta_{ks}. \tag{2.84}$$

Dvojím užitím lemmatu 3.10 z kap. II na každý z dvojnásobného součtu na pravé straně (2.84) snadno zjistíme (srv. také důkaz lemmatu 3.14 z kap. II), že platí

$$(B_h\eta, \eta) = h^4 \sum_{(x_k,y_s)\in\Omega_1^{(h)}} [(L_h\eta)_{ks}]^2 + 2 \sum_{(x_k,y_s)\in\Gamma_2^{(h)}} \eta_{ks}^2 + \\ + 2\eta_{11}^2 + 2\eta_{n-1,1}^2 + 2\eta_{1,n-1}^2 + 2\eta_{n-1,n-1}^2. \tag{2.85}$$

Čísla $(L_h\eta)_{ks}$ jsou však zřejmě složky vektoru $A_h\eta$, kde A_h je matice soustavy rovnic

$$h^2(L_h\eta)_{ks} = g_{ks}, \quad (x_k, y_s) \in \Omega_1^{(h)}. \tag{2.86}$$

Je tedy

$$(B_h\eta, \eta) = (A_h\eta, A_h\eta) + 2 \sum_{(x_k,y_s)\in\Gamma_2^{(h)}} \eta_{ks}^2 + \\ + 2\eta_{11}^2 + 2\eta_{n-1,1}^2 + 2\eta_{1,n-1}^2 + 2\eta_{n-1,n-1}^2. \tag{2.87}$$

Z rovnice (2.87) však ihned plyne, že platí

$$(B_h\eta, \eta) \geqq 2 \sum_{(x_k,y_s)\in\Gamma_2^{(h)}} \geqq 2\Big[\max_{(x_k,y_s)\in\Gamma_2^{(h)}} |\eta_{ks}|\Big]^2 = 2h^2\Big[\max_{(x_k,y_s)\in\Gamma_2^{(h)}} |\eta_{ks}h^{-1}|\Big]^2. \tag{2.88}$$

Buď nyní $e^{(p,q)}$, $p, q = 1, \dots, n-1$, vektor, který má (p,q)-tou složku rovnu jedné a všechny ostatní složky rovny nule. Pak zřejmě platí

$$\eta_{pq} = (\eta, e^{(p,q)}). \tag{2.89}$$

Buď dále $\alpha^{(p,q)}$ vektor, který je řešením soustavy

$$A_h\alpha^{(p,q)} = e^{(p,q)}. \tag{2.90}$$

Abychom mohli řešení této soustavy jednoduše zapsat, uvažujme soustavu vektorů $v^{(\nu,\mu)}$, $\nu, \mu = 1, \dots, n-1$, jejichž složky jsou dány vzorci

$$v_{ks}^{(\nu,\mu)} = 2h \sin\frac{\nu\pi k}{n} \sin\frac{\mu\pi s}{n}, \quad k, s = 1, \dots, n-1. \tag{2.91}$$

Přímým výpočtem snadno zjistíme, že vektory $v^{(\nu,\mu)}$ tvoří úplnou ortonormovanou soustavu vlastních vektorů matice A_h a že příslušná vlastní čísla $\lambda_{\nu\mu}$ jsou dána

vzorci

$$\lambda_{\nu\mu} = 4 - 2\Big(\cos\frac{\nu\pi}{n} + \cos\frac{\mu\pi}{n}\Big) = \tag{2.92}$$
$$= 4\Big(\sin^2\frac{\nu\pi}{2n} + \sin^2\frac{\nu\pi}{2n}\Big), \quad \nu,\mu = 1,\dots,n-1.$$

Odtud mimo jiné plyne, že existují konstanty $c_{\nu\mu}^{(p,q)}$ takové, že platí

$$\alpha^{(p,q)} = \sum_{(x_\nu,y_\mu)\in\Omega_1^{(h)}} c_{\nu\mu}^{(p,q)} v^{(\nu,\mu)}. \tag{2.93}$$

Vynásobíme-li rovnici (2.93) zleva maticí A_h, dostaneme

$$A_h\alpha^{(p,q)} = e^{(p,q)} = \sum_{(x_\nu,y_\mu)\in\Omega_1^{(h)}} \lambda_{\nu\mu} c_{\nu\mu}^{(p,q)} v^{(\nu,\mu)}. \tag{2.94}$$

Čísla $\lambda_{\nu\mu}c_{\nu\mu}^{(p,q)}$, $\nu,\mu = 1\dots,n-1$, jsou tedy Fourierovy koeficienty vektoru $e^{(p,q)}$ vzhledem k ortonormální bázi $v^{(\nu,\mu)}$, a proto platí

$$\lambda_{\nu\mu}c_{\nu\mu}^{(p,q)} = (e^{(p,q)}, v^{(\nu,\mu)}) = v_{pq}^{(\nu,\mu)} \tag{2.95}$$

neboli

$$c_{\nu\mu}^{(p,q)} = \frac{1}{\lambda_{\nu\mu}} v_{pq}^{(\nu,\mu)}. \tag{2.96}$$

Z rovnice (2.92) a ze zřejmé nerovnosti $\sin^2 z \geqq 4z^2/\pi^2$ platné pro libovolné reálné z z intervalu $\langle 0, \pi/2\rangle$ však plyne existence konstanty γ_1 (nezávislé na n) takové, že platí

$$\frac{1}{\lambda_{\nu\mu}} \leqq \frac{\gamma_1}{h^2}\frac{1}{\nu^2+\mu^2}. \tag{2.97}$$

Dosadíme-li tento výsledek do (2.96) a použijeme-li rovnic (2.91), máme

$$|c_{\nu\mu}^{(p,q)}| \leqq \frac{2\gamma_1}{h}\frac{1}{\nu^2+\mu^2}. \tag{2.98}$$

Pro euklidovskou normu vektoru $\alpha^{(p,q)}$ tedy dostáváme podle (2.93) a (2.98)

$$\|\alpha^{(p,q)}\|^2 \leqq \frac{4\gamma_1^2}{h^2} \sum_{(x_\nu,y_\mu)\in\Omega_1^{(h)}} \frac{1}{(\nu^2+\mu^2)^2} \leqq \frac{\gamma_2}{h^2}, \tag{2.99}$$

neboť řada $\sum 1/(\nu^2+\mu^2)^2$, kde sčítáme přes všechna celá čísla, je konvergentní. Dosadíme-li do rovnice (2.89) za vektor $e^{(p,q)}$ podle rovnice (2.90) a použijeme-li toho, že matice A_h je symetrická, máme

$$\eta_{pq} = (A_h\eta, \alpha^{(p,q)}). \tag{2.100}$$

Podle Schwarzovy nerovnosti je tedy

$$|\eta_{pq}|^2 \leqq \|A_h\eta\|^2\|\alpha^{(p,q)}\|^2 \leqq \frac{\gamma_2}{h^2}(A_h\eta, A_h\eta) \tag{2.101}$$

a tato nerovnost platí pro všechny dvojice (p,q) takové, že je $(x_p, y_q) \in \Omega_1^{(h)}$. Odtud a z rovnice (2.87) však ihned plyne, že platí

$$(B_h\eta,\eta) \geqq \frac{h^2}{\gamma_2}[\max_{(x_p,y_q)\in\Omega^{(h)}} |\eta_{pq}|]^2. \tag{2.102}$$

K důkazu nerovnosti (2.82) stačí už jen položit $\gamma = \min(2, 1/\gamma_2)$ a použít nerovností (2.88) a (2.102). Důkaz lemmatu je hotov.

Z právě dokázaného lemmatu plyne předně, že matice B_h je pozitivně definitní. Soustava (2.81) má tedy právě jedno řešení. Z lemmatu 2.3 však plyne snadno i konvergence uvažované metody sítí.

Věta 2.7. *Nechť řešení u diferenciální rovnice* (2.73) *s okrajovými podmínkami* (2.74) *má ve čtverci $\bar\Omega$ spojité parciální derivace až do šestého řádu včetně a nechť u_{ks} je přibližné řešení vypočtené ze soustavy* (2.81). *Nechť η je vektor o složkách $u_{ks} - u(x_k, y_s)$, $(x_k, y_s) \in \Omega_1^{(h)}$. Pak existuje konstanta M taková, že pro každé dostatečně malé h platí*

$$\|\eta\|_{h,\mathscr{L}_\infty} \leqq Mh. \tag{2.103}$$

D ů k a z . Buď $u^{(\mathrm{pr})}$ vektor o složkách $u(x_k, y_s)$ a položme

$$\varepsilon = B_h u^{(\mathrm{pr})} - g, \tag{2.104}$$

kde g je pravá strana rovnice (2.81). Vzhledem ke konstrukci matice B_h a k hladkostním předpokladům kladeným na funkci u existuje konstanta M taková, že je

$$|\varepsilon_{ks}| \leqq Mh^6, \quad (x_k, y_s) \in \Omega^{(h)}, \tag{2.105}$$

a

$$|\varepsilon_{ks}| \leqq Mh^3, \quad (x_k, y_s) \in \Gamma_2^{(h)}. \tag{2.106}$$

Pro vektor η platí

$$B_h\eta = \varepsilon \tag{2.107}$$

a podle nerovnosti (2.82) máme

$$\|\eta\|_{h,\mathscr{L}_\infty}^2 \leqq \frac{1}{\gamma h^2}\Big[\sum_{(x_k,y_s)\in\Omega^{(h)}} \varepsilon_{ks}\eta_{ks} + \sum_{(x_k,y_s)\in\Gamma_2^{(h)}} \varepsilon_{ks}h\eta_{ks}h^{-1}\Big] \leqq \tag{2.108}$$
$$\leqq \frac{1}{\gamma h^2}\|\eta\|_{h,\mathscr{L}_\infty}\Big[\sum_{(x_k,y_s)\in\Omega^{(h)}} |\varepsilon_{ks}| + \sum_{(x_k,y_s)\in\Gamma_2^{(h)}} h|\varepsilon_{ks}|\Big].$$

Protože množina $\Omega^{(h)}$ má řádově $O(1/h^2)$ prvků a množina $\Gamma_2^{(h)}$ pouze $O(1/h)$ prvků, plyne tvrzení věty už snadno z nerovnosti (2.108) a z odhadů (2.105) a (2.106).

Při přepisu okrajových podmínek jsme v námi vyšetřovaném případě využili speciálního tvaru hranice dané oblasti. V obecném případě se při přepisu okrajových podmínek vychází z obdobných principů, jak bylo uvedeno u okrajových podmínek pro rovnici druhého řádu. Obtíže, které tam nastávaly zejména v případě, že okrajová podmínka obsahovala derivace, se zde vyskytují v případě jakékoliv okrajové podmínky, neboť u rovnice čtvrtého řádu jsou zadány okrajové podmínky dvě, a aspoň jedna z nich musí vždy obsahovat derivaci.

2.3 Řešení vzniklých soustav lineárních rovnic

V předchozím textu jsme viděli, že řešit okrajovou úlohu pro lineární parciální diferenciální rovnici metodou sítí znamená řešit jistou soustavu lineárních algebraických rovnic. Algoritmus pro řešení dané okrajové úlohy je tedy dán teprve tehdy, je-li udána konkrétní metoda pro řešení zmíněné soustavy rovnic. Protože tato poslední část celkového algoritmu metody sítí je z praktického hlediska značně důležitá — řešení náhradní soustavy lineárních rovnic spotřebuje prakticky všechen potřebný výpočetní čas — všimneme si i této problematiky. Z důvodu zachování přijatelného rozsahu této knížky budeme velice struční a v podstatě se omezíme na výčet nejtypičtějších možností, aniž budeme zacházet do nějakého podrobnějšího rozboru uváděných metod. Jako modelová úloha nám pro většinu výkladu v tomto odstavci poslouží Dirichletova úloha pro Poissonovu rovnici

$$-\Delta u = f(x,y) \tag{2.109}$$

na čtverci $\Omega = \{(x,y);\, 0 < x < 1, 0 < y < 1\}$. Laplaceův operátor $-\Delta$ budeme přitom aproximovat operátorem L_h definovaným rovnicí (2.75). Budeme se tedy zabývat soustavou rovnic

$$4u_{ks} - u_{k+1,s} - u_{k-1,s} - u_{k,s+1} - u_{k,s-1} = h^2 f(x_k, y_s), \quad k,s = 1,\dots,n-1, \tag{2.110}$$

kde hodnoty u_{0s}, u_{ns}, u_{k0}, u_{kn} jsou známé a pro integrační krok h platí $h = 1/n$, kde n je přirozené číslo. Jde tedy o soustavu $N = (n-1)^2$ rovnic o N neznámých. Očíslujeme-li neznámé jedním indexem, a to tak, že postupujeme po řádcích sítě odleva doprava a odshora dolů, lze soustavu (2.110) zapsat v maticovém tvaru

$$Au = g, \tag{2.111}$$

kde A je čtvercová matice řádu N (která je symetrická a ireducibilně diagonálně dominantní) a g je daný vektor pravé strany určený pravou stranou dané diferenciální rovnice a okrajovými podmínkami. Matice A je při daném očíslování neznámých

blokově třídiagonální, tj. je tvaru

$$A = \begin{bmatrix} A_1 & B_1 & 0 & \dots & 0 \\ B_1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & B_{n-2} \\ 0 & \dots & 0 & B_{n-2} & A_{n-1} \end{bmatrix}, \tag{2.112}$$

kde A_k a B_k jsou čtvercové matice řádu $n-1$ dané rovnicemi

(2.113)

$$A_k = \begin{bmatrix} 4 & -1 & 0 & \dots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 4 \end{bmatrix}, \quad B_k = \begin{bmatrix} -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 \end{bmatrix}$$

a symbol 0 ve vzorci (2.112) značí nulovou matici řádu $n-1$. Matice A_k jsou tedy třídiagonální a matice B_k diagonální (platí dokonce $B_k = -I$, kde I je jednotková matice). Hledaný N-dimenzionální vektor lze přitom psát blokově ve tvaru

$$u = [u_1^T, \dots, u_{n-1}^T]^T, \tag{2.114}$$

kde u_k jsou $(n-1)$-dimenzionální vektory dané rovnicemi

$$u_k = [u_{1,n-k}, \dots, u_{n-1,n-k}]^T, \quad k = 1,\dots,n-1. \tag{2.115}$$

V případě obecnějších diferenciálních rovnic druhého řádu bude struktura matice A v podstatě zachována, matice A_k a B_k ve vyjádření (2.112) však nebudou konstantní a v případě obecnějších oblastí bude i jejich řád záviset na indexu k. To byl také důvod, proč jsme psali ve vzorci (2.112) indexy, i když v našem speciálním případě je to zbytečné, neboť všechny matice A_k i B_k jsou zde stejné. V případě rovnice čtvrtého řádu lze většinou matici A uvést na blokově pětidiagonální tvar. Ať už jde o rovnici druhého či vyššího řádu, je volba očíslování neznámých podstatná a kdybychom ji provedli náhodně, dostali bychom matici, která by byla sice řídká, tj. obsahovala by mnoho nulových prvků, avšak ty by v ní byly rozmístěny náhodně. V případě námi zvoleného očíslování je však matice A nejen řídká, ale dokonce pásová a přitom žádným jiným očíslováním neznámých nelze dosáhnout menší šíři pásu. Možnost takového očíslování neznámých je pro matice vznikající v metodě sítí typická.

V následujících odstavcích stručně popíšeme některé konkrétní metody řešení soustavy (2.111).

2.3.1 Přímé metody

K řešení soustavy (2.111) lze v zásadě užít všechny běžné varianty Gaussovy eliminační metody. Vzhledem k tomu, že matice A je symetrická a pozitivně definitní, lze užít eliminační metodu bez výběru hlavního prvku a využít tak skutečnost, že matice je pásová. To vede k podstatné redukci počtu potřebných operací i počtu paměťových míst. (Počet násobení je např. řádově roven $O(1/h^4)$ oproti $O(1/h^6)$ bez využití pásovosti matice A.)

Speciální strukturu matice A však můžeme využít také přímo: Se soustavou (2.111), kterou lze psát blokově ve tvaru

$$\text{(2.116)}\qquad \begin{bmatrix} A_1 & B_1 & 0 & \dots & \dots & 0 \\ B_1 & \ddots & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & B_{n-2} \\ 0 & \dots & \dots & 0 & B_{n-2} & A_{n-1} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ \\ u_{n-1} \end{bmatrix} = \begin{bmatrix} g_1 \\ \vdots \\ \\ g_{n-1} \end{bmatrix},$$

budeme manipulovat přesně tak, jako kdyby veličiny A_i, B_i, resp. u_i, g_i nebyly matice, resp. vektory, ale čísla. Tento postup vede k následujícím rekurencím (srv. vzorce (3.138), (3.152) a (3.153) z kap. II). Nejprve vypočteme matice D_k (řádu $n-1$) z rovnic

$$\text{(2.117)}\qquad \begin{aligned} D_1 &= A_1, \\ D_k &= A_k - B_{k-1}D_{k-1}^{-1}B_{k-1}, \quad k = 2, \dots, n-1, \end{aligned}$$

pak $(n-1)$-dimenzionální vektory c_k z rovnic

$$\text{(2.118)}\qquad \begin{aligned} c_0 &= g_0, \\ c_k &= g_k - B_{k-1}D_{k-1}^{-1}c_{k-1}, \quad k = 2, \dots, n-1, \end{aligned}$$

a konečně neznámé vektory u_k z rovnic

$$\text{(2.119)}\qquad \begin{aligned} u_{n-1} &= D_{n-1}^{-1}c_{n-1}, \\ u_k &= D_k^{-1}(c_k - B_k u_{k+1}), \quad k = n-2, \dots, 1. \end{aligned}$$

Postup popsaný rovnicemi (2.117) je co do počtu operací ekvivalentní Gaussově eliminační metodě a lze jej bez obtíží provést, neboť postupně vznikající matice D_k jsou pozitivně definitní, takže matice k nim inverzní lze konstruovat trojúhelníkovým rozkladem bez výběru hlavního prvku. Zároveň je také vidět, že tento postup lze s mírnými modifikacemi užít i u obecnějších úloh druhého řádu. V případě rovnic čtvrtého řádu je přirozeně třeba užít rekurence typu (3.207) až (3.209) z kap. II.

Užití eliminační metody k řešení velkých soustav typu (2.111) často usnadní, lze-li matici A psát blokově ve tvaru

$$\text{(2.120)}\qquad A = \begin{bmatrix} P_1 & 0 & \dots & \dots & 0 & Q_1^T \\ 0 & \ddots & \ddots & & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & \dots & 0 & P_{n-2} & Q_{n-2}^T \\ Q_1 & \dots & \dots & \dots & Q_{n-2} & P_{n-1} \end{bmatrix},$$

kde P_k jsou symetrické pozitivně definitní matice. Pro matici A pak platí

$$\text{(2.121)}\qquad A = MM^T,$$

kde M je dolní trojúhelníková matice tvaru

$$\text{(2.122)}\qquad M = \begin{bmatrix} M_1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & M_{n-2} & 0 \\ N_1 & \dots & \dots & \dots & N_{n-2} & M_{n-1} \end{bmatrix}$$

a matice M_k a N_k splňují rovnice

$$\text{(2.123)}\qquad M_k M_k^T = P_k, \quad k = 1, \dots, n-2,$$

$$\text{(2.124)}\qquad N_k M_k^T = Q_k, \quad k = 1, \dots, n-2,$$

a

$$\text{(2.125)}\qquad M_{n-1}M_{n-1}^T = P_{n-1} - N_1 N_1^T - \dots - N_{n-2}N_{n-2}^T.$$

Matice M_k lze tedy pro $k = 1, \dots, n-2$ sestrojit Choleského metodou nezávisle jednu na druhé. Z rovnic (2.124) se pak určí matice N_k pro $k = 1, \dots, n-2$ a konečně z rovnice (2.125) — opět Choleského metodou — matice M_{n-1}. Řešení původní soustavy se pak zřejmě (viz rovnici (2.121)) dostane řešením rovnic

$$\text{(2.126)}\qquad \begin{aligned} Mv &= g, \\ M^T u &= v. \end{aligned}$$

Popišme jednu z možností, jak očíslovat uzly sítě tak, aby matice A soustavy (2.111) byla tvaru (2.120). Tento postup lze užít i v případě obecného pětibodového schématu (2.4). Předpokládejme, že platí $n - 1 = rm$, kde r a m jsou celá čísla, a rozdělme nejprve všechny vnitřní uzly do $r^2 + 1$ skupin takto: Do první skupiny

dáme uzly (kh, sh), pro něž je $k = 1, \dots, m-1$ a $s = 1, \dots, m-1$; do druhé skupiny dáme uzly, pro něž je $k = 1, \dots, m-1$ a $s = m+1, \dots, 2m-1$ atd. až do r-té skupiny dáme uzly, pro něž je $k = 1, \dots, m-1$ a $s = (r-1)m+1, \dots, rm-1$; do $(r+1)$-ní skupiny dáme uzly, pro něž je $k = m+1, \dots, 2m-1$ a $s = 1, \dots, m-1$ atd. až do $(2r)$-té skupiny dáme uzly, pro něž je $k = m+1, \dots, 2m-1$ a $s = (r-1)m+1, \dots, rm-1$ atd. až do r^2-té skupiny dáme uzly, pro něž je $k = (r-1)m+1, \dots, rm-1$ a $s = (r-1)m+1, \dots, rm-1$; do (r^2+1)-ní skupiny dáme všechny zbývající uzly. Uzly z první skupiny očíslujeme prvními $(m-1)^2$ indexy, uzly z druhé skupiny druhými $(m-1)^2$ indexy atd.

Za předpokladu, že platí $A_k = D$ a $B_k = -T$ pro $k = 1, \dots, n-1$, kde D a T jsou třídiagonální matice, že je $n = 2^{s+1}$ a že matice D a T komutují, lze soustavu (2.111) řešit velmi efektivně metodou zvanou *metoda cyklické redukce*. Postup je následující:

Položme pro $i = 0, \dots, s$

$$D_i = \prod_{r=1}^{2^i} \left[D - 2\cos\frac{(2r-1)\pi}{2^{i+1}} T \right], \qquad T_i = T^{2^i} \tag{2.127}$$

a vypočtěme pro $i = 1, \dots, s$ rekurentně vektory $g_j^{(i)}$

$$g_j^{(i)} = T_{i-1} g_{2j-1}^{(i-1)} + D_{i-1} g_{2j}^{(i-1)} + T_{i-1} g_{2j+1}^{(i-1)}, \quad j = 1, \dots, 2^{s+1-i} - 1, \tag{2.128}$$

přičemž klademe $g_j^{(0)} = g_j$ pro $j = 1, \dots, 2^{s+1} - 1$.

Dále řešme soustavu

$$D_s u_1^{(s)} = g_1^{(s)}. \tag{2.129}$$

Konečně pro $i = s-1, \dots, 0$ položme

$$u_{2j}^{(i)} = u_j^{(i+1)}, \quad j = 1, \dots, 2^{s-i} - 1, \qquad u_0^{(i)} = u_{2^{s+1-i}}^{(i)} = 0 \tag{2.130}$$

a řešme soustavy

$$D_i u_{2j+1}^{(i)} = T_i u_{2j}^{(i)} + g_{2j+1}^{(i)} + T_i u_{2j+2}^{(i)}, \quad j = 0, \dots, 2^{s-i} - 1. \tag{2.131}$$

Výsledné vektory $u_1^{(0)}, \dots, u_{2^{s+1}-1}^{(0)}$ jsou řešením soustavy (2.111).

Provádíme-li součiny matic a vektorů na pravé straně rovnic (2.128) a (2.131) opakovaným násobením třídiagonálními maticemi $D - 2\cos\left((2r-1)\pi/2^{i+1}\right) T$ a T a řešíme-li soustavy (2.131) tak, že opakovaně řešíme soustavy s třídiagonálními maticemi $D - 2\cos\left((2r-1)\pi/2^{i+1}\right) T$, je počet potřebných operací řádově roven číslu $(n-1)^2 \ln(n-1)$, tedy až na faktor $\ln(n-1)$ počtu rovnic. Popsaný postup

redukuje tedy počet potřebných operací ve srovnání např. s Gaussovou eliminací velice výrazně.

Upozorněme, že uvedené vzorce jsou správné i v případě, že D a T jsou libovolné, ne nutně třídiagonální matice řádu $(n-1)$, jen když komutují. V tomto případě už však nebude počet potřebných operací tak výhodný jako výše.

Jedna z obtíží spojených s užitím přímých metod k řešení rovnic vznikajících v metodě sítí spočívá v potřebě vysokého počtu paměťových míst i v případě, že příslušná matice má velmi mnoho nulových prvků. Při těchto metodách se totiž, ať už přímo nebo nepřímo, konstruuje trojúhelníkový rozklad dané matice, a vzniklé trojúhelníkové matice mohou obsahovat nenulové prvky i v těch místech, kde výchozí matice nuly měla. *Stoneova metoda neúplného rozkladu* redukuje podstatně počet potřebných paměťových míst. Vychází z toho, že se postupuje přesně podle vzorců pro trojúhelníkový rozklad obecné matice, prvky příslušných trojúhelníkových matic se však počítají pouze v těch polohách, které odpovídají polohám nenulových prvků původní matice, a jinde se kladou nuly. Součin takto vzniklých matic se pak přirozeně nerovná původní matici, přesto však mohou tyto matice sloužit za východisko k velmi efektivním algoritmům. Do dalších podrobností nebudeme zacházet a odkazujeme čtenáře na specializovanou literaturu (viz např. Stone (1968)).

Nakonec se zmíníme ještě o jedné skupině přímých metod, které se v poslední době intenzivně vyvíjejí a které se snaží zužitkovat tu skutečnost, že jednoduché rovnice vznikající v metodě sítí umíme velmi efektivně řešit, je-li i daná oblast jednoduchá (viz např. metoda cyklické redukce). Základní myšlenku těchto metod popíšeme jen ve velmi hrubých rysech. Nechť je např. třeba řešit soustavu (2.110), avšak na složitější oblasti, než je čtverec. Pak tuto oblast vnoříme do vhodného čtverce a v uzlech, které přitom přibudou, dodáme další rovnice typu (2.110). Pravé strany pro ně získáme řešením vhodné soustavy lineárních rovnic, která sice bude mít plnou matici, ta však bude nízkého řádu. Řešení původního problému se tak převede na řešení malé soustavy s plnou maticí a na řešení soustavy, která má zhruba tolik neznámých (ve skutečnosti o něco více) jako původní soustava, avšak kterou je možné vyřešit velice rychle. Podrobnější poučení o metodách tohoto typu nalezne čtenář např. v článku Proskurowského a Widlunda (1976).

2.3.2 Iterační metody

Přímé metody jsou pro řešení diferenčních rovnic většinou uspokojivé, jsou-li vůbec proveditelné, tj. zejména tehdy, nezabrání-li jejich užití nedostatečná paměťová kapacita počítače, který máme k dispozici. Proto se v souvislosti s metodou sítí (a také v souvislosti s metodou konečných prvků, jak uvidíme) užívají podstatně častěji než v případě jednodimenzionálních úloh iterační metody. Ze skupiny tzv. maticových iteračních metod jsou nejběžnější *Gaussova-Seidelova metoda*, *superrelaxační metoda* (metoda SOR) a různé varianty *metod střídavých* směrů. Z *gradientních metod*,

tj. metod, které vycházejí z myšlenky minimalizace kvadratické formy

$$Q(u) = \tfrac{1}{2}u^T Au - u^T g, \tag{2.132}$$

(jejíž minimum je v případě, že matice A je symetrická a pozitivně definitní, řešením soustavy (2.111)), je to pak zejména *metoda sdružených gradientů*, která je v poslední době značně populární. Všimněme si proto zmíněných metod poněkud podrobněji.

Připomeňme, že superrelaxační metoda je dána vzorcem

$$u^{(k+1)} = H_\omega u^{(k)} + C_\omega g, \tag{2.133}$$

s maticemi H_ω a C_ω definovanými rovnicemi

$$H_\omega = (D - \omega L)^{-1}[(1-\omega)D + \omega U] \tag{2.134}$$

a

$$C_\omega = \omega(D - \omega L)^{-1}, \tag{2.135}$$

přičemž je

$$A = D - L - U, \tag{2.136}$$

kde D je diagonální matice, L dolní trojúhelníková a U horní trojúhelníková matice, obě s nulami na hlavní diagonále a ω je číselný parametr zvaný relaxační faktor. Gaussovou-Seidelovu metodu dostaneme pak pro $\omega = 1$.

Matice soustavy (2.111) má v každém řádku kromě diagonálního prvku už jen nanejvýš čtyři nenulové prvky. K výpočtu jedné iterace je tedy zapotřebí řádově tolik operací, kolik složek má hledaný vektor. Protože pracujeme stále s toutéž původní maticí, počet potřebných paměťových míst je dán počtem jejích nenulových prvků a přirozeně dimenzí hledaného vektoru. To je také hlavní výhoda nejen superrelaxační metody, ale vlastně všech iteračních metod oproti přímým metodám.

Připomeňme také, že je-li $0 < \omega < 2$ a je-li matice A symetrická a pozitivně definitní, je superrelaxační metoda konvergentní. Dále připomeňme, že je-li matice A *dvoucyklická* (tj. existuje-li permutační matice P taková, že matice $P^T D^{-1}(L + U)P$ je tvaru

$$\begin{bmatrix} 0 & B_{12} \\ B_{21} & 0 \end{bmatrix}, \tag{2.137}$$

kde nuly znamenají čtvercové nulové matice) a *konzistentně uspořádaná* (tj. nezávisí-li vlastní čísla matice $B(\alpha) = \alpha D^{-1}L + \alpha^{-1}D^{-1}U$, $\alpha \neq 0$ na parametru α), je možno superrelaxační metodu optimalizovat, tj. je možno nalézt takové číslo ω_{opt}, $0 < \omega_{\text{opt}} < 2$, že platí

$$\varrho(H_{\omega_{\text{opt}}}) = \inf_{\omega \in (0,2)} \varrho(H_\omega), \tag{2.138}$$

přičemž optimální relaxační faktor je dán vzorcem

$$\omega_{\text{opt}} = \frac{1}{1 + \left[1 - \varrho^2\left(D^{-1}(L+U)\right)\right]^{1/2}}. \tag{2.139}$$

Konečně připomeňme, že platí

$$\varrho^2\left(D^{-1}(L+U)\right) = \varrho(H_1), \tag{2.140}$$

takže číslo $\varrho^2\left(D^{-1}(L+U)\right)$ je spektrální poloměr Gaussovy-Seidelovy iterační matice a že pro spektrální poloměr optimalizované superrelaxační metody je

$$\varrho(H_{\omega_{\text{opt}}}) = \frac{1 - \left[1 - \varrho^2\left(D^{-1}(L+U)\right)\right]^{1/2}}{1 + \left[1 - \varrho^2\left(D^{-1}(L+U)\right)\right]^{1/2}}. \tag{2.141}$$

Ukažme, že v případě naší modelové úlohy příslušná matice (tj. matice soustavy (2.111)) do výše popsané třídy patří. Matice $D^{-1}(L+U)$ vznikne z matice A zřejmě tak, že se její diagonální prvky nahradí nulami, u nediagonálních prvků se změní znaménko a vydělí se čtyřmi. Vynásobit matici $D^{-1}(L+U)$ zleva maticí P^T a zprava maticí P, kde P je permutační matice, znamená přečíslovat neznámé v příslušné soustavě rovnic a změnit odpovídajícím způsobem pořadí rovnic. Přečíslujme neznámé „šachovnicově“, tj. v případě, že n je liché, číslujeme ty neznámé, kde součet $k+s$ je lichý, postupně zleva doprava a shora dolů indexy $1, 2, \ldots, (n-1)^2/2$ a ty neznámé, kde součet $k+s$ je sudý, indexy $(n-1)^2/2+1, \ldots, (n-1)^2$; v případě, že n je sudé, užijeme pro ty neznámé, kde součet $k+s$ je sudý, indexy $1, \ldots, [(n-1)^2+1]/2$ a pro ty neznámé, kde součet $k+s$ je lichý, indexy $[(n-1)^2+1]/2+1, \ldots, (n-1)^2$. Přerovnáme-li odpovídajícím způsobem i rovnice, dostaneme soustavu, jejíž matice je zřejmě tvaru (2.137). Matice A je tedy dvoucyklická.

Abychom dokázali konzistentní uspořádanost matice A, uvažujme libovolné vlastní číslo λ matice $D^{-1}(L+U)$. Je-li v_{ks} jemu odpovídající vlastní vektor a položíme-li $v_{0s} = v_{ns} = 0$ pro $s = 1, \ldots, n-1$ a $v_{k0} = v_{kn} = 0$ pro $k = 1, \ldots, n-1$, platí

$$v_{k-1,s} + v_{k+1,s} + v_{k,s+1} + v_{k,s-1} = 4\lambda v_{ks} \tag{2.142}$$

pro $k, s = 1, \ldots, n-1$. Vynásobíme-li každou z těchto rovnic číslem α^{k-s}, dostaneme

$$\begin{aligned} &\alpha\alpha^{(k-1)-s}v_{k-1,s} + \alpha\alpha^{k-(s+1)}v_{k,s+1} + \\ &+ \frac{1}{\alpha}\alpha^{(k+1)-s}v_{k+1,s} + \frac{1}{\alpha}\alpha^{k-(s-1)}v_{k,s-1} = 4\lambda\alpha^{k-s}v_{ks}. \end{aligned} \tag{2.143}$$

Číslo λ je tedy také vlastním číslem matice $B(\alpha) = \alpha D^{-1}L + \alpha^{-1}D^{-1}U$ (a odpovídající vlastní vektor je $\alpha^{k-s}v_{ks}$). Snadno se už zjistí, že každé vlastní číslo matice $B(\alpha)$ se dostane popsaným způsobem. Vlastní čísla matice $B(\alpha)$ tedy skutečně nezávisí na α a matice A je konzistentně uspořádaná.

Při řešení soustavy (2.110) superrelaxační metodou platí tedy vzorce (2.139) a (2.141). Speciální případ této soustavy umožňuje také ilustrovat, co se skutečně získá na rychlosti konvergence užitím optimalizované superrelaxační metody.

Rychlost konvergence budeme přitom měřit, jak je zvykem, velikostí spektrálního poloměru příslušné iterační matice. Snadno se zjistí, že čísla $\tilde{\lambda}_{\nu\mu}$ daná vzorcem

$$\tilde{\lambda}_{\nu\mu} = \tfrac{1}{2}\Big(\cos\frac{\nu\pi}{n} + \cos\frac{\mu\pi}{n}\Big), \quad \nu,\mu = 1,\dots,n-1, \tag{2.144}$$

jsou všechna vlastní čísla matice $D^{-1}(L+U)$. Plyne to totiž z toho, že vektory $v^{(\nu,\mu)}$, jejichž souřadnice jsou dány vzorci (2.91) a které tvoří úplnou soustavu vlastních vektorů matice A_h z odst. 2.2, jsou zároveň vlastními vektory matice $D^{-1}(L+U)$. Je tedy

$$\varrho(D^{-1}(L+U)) = \cos\frac{\pi}{n} = \cos\pi h = 1 - O(h^2), \tag{2.145}$$

a tedy také

$$\varrho(H_1) = \varrho^2(D^{-1}(L+U)) = 1 - O(h^2). \tag{2.146}$$

Odtud a ze vzorce (2.141) však plyne, že platí

$$\varrho(H_{\omega_{\mathrm{opt}}}) = 1 - O(h). \tag{2.147}$$

Spektrální poloměr optimalizované superrelaxační metody je tedy pro malá h skutečně podstatně menší než spektrální poloměr Gaussovy-Seidelovy metody.

Při praktickém užití superrelaxační metody se postupuje obvykle tak, že se výpočet zahájí Gaussovou-Seidelovou metodou a po výpočtu několika iterací se odhadne spektrální poloměr příslušné iterační matice H_1 mocninnou metodou. Tato aproximace se pak užije ve vzorci (2.139), do něhož se dosadí ze vzorce (2.140).

Řešíme-li obecnější eliptickou rovnici druhého řádu na čtverci nebo obdélníku, jehož strany jsou tvořeny přímkami sítě, je příslušná matice rovněž dvoucyklická a konzistentně uspořádaná, takže vzorec (2.141) lze užít i v tomto případě. Tohoto vzorce se však v praxi užívá i v těch případech, kdy nevíme, zda skutečně vede k minimální hodnotě spektrálního poloměru. Alternativně se při užití superrelaxační metody postupuje také tak, že se v situaci, kde se řeší celá velká skupina problémů podobného charakteru, určí optimální relaxační faktor experimentálně v jednom případě a v dalších případech se užije takto získaná hodnota.

Gaussovu-Seidelovu i superrelaxační metodu (a i mnohé další maticové iterační metody) lze realizovat v tzv. ***blokové podobě***. Stručně řečeno to znamená, že matice dané soustavy se rozdělí na bloky tak, aby diagonální bloky byly regulární čtvercové matice a iterace se konstruují podle stejných vzorců, jaké byly uvedeny výše s tím, že skaláry, které se v nich vyskytují, je třeba pokládat za vektory, resp. matice. To mimo jiné znamená, že kdekoliv se v těchto vzorcích vyskytuje dělení, je třeba je chápat jako násobení příslušnou inverzní maticí. Tyto blokové iterační metody mohou být v naší speciální situaci výhodné, neboť bloková struktura matice A soustavy (2.111) je velice jednoduchá. Předpokládáme-li, že tato soustava je zapsána ve tvaru (2.116), Gaussova-Seidelova bloková metoda dána vzorcem

$$u_k^{(i+1)} = A_k^{-1}(g_k - B_{k-1}u_{k-1}^{(i+1)} - B_k u_k^{(i)}), \quad k = 1,\dots,n-1. \tag{2.148}$$

Podobně bloková superrelaxační metoda je dána vzorcem

$$u_k^{(i+1)} = \omega A_k^{-1}(g_k - B_{k-1}u_{k-1}^{(i+1)} - B_k u_k^{(i)}) + (1-\omega)u_k^{(i)}, \quad k = 1,\dots,n-1. \tag{2.149}$$

Pro výpočet je výhodnější přepsat rovnice (2.149) v ekvivalentním tvaru

$$\begin{aligned} A_k \tilde{u}_k^{(i+1)} &= g_k - B_{k-1}u_{k-1}^{(i+1)} - B_k u_k^{(i)}, \\ u_k^{(i+1)} &= \omega(\tilde{u}_k^{(i+1)} - u_k^{(i)}) + u_k^{(i)}. \end{aligned} \tag{2.150}$$

Je-li bloková matice A dvoucyklická a konzistentně uspořádaná — tyto pojmy se definují úplně stejně jako dříve — platí pro optimální relaxační faktor vzorec (2.139). Matice D, L a U v tomto vzorci jsou samozřejmě matice dané vztahy

$$D = \begin{bmatrix} A_1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & A_{n-1} \end{bmatrix}, \quad L = \begin{bmatrix} 0 & \dots & \dots & \dots & 0 \\ B_1 & \ddots & & & \vdots \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & B_{n-2} & 0 \end{bmatrix}, \tag{2.151}$$

$$U = \begin{bmatrix} 0 & B_1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & B_{n-2} \\ 0 & \dots & \dots & \dots & 0 \end{bmatrix}.$$

Vzhledem k tomu, že popsané blokové metody lze upravit tak, že provedení jedné iterace u nich nevyžaduje více početních operací než výpočet jedné iterace u dřívějších metod (kterým se v této souvislosti říká ***bodové iterační metody***), a vzhledem k tomu, že konvergují většinou rychleji než metody bodové, jsou dosti vážnými kandidáty na univerzálně použitelné metody pro řešení rovnic, které vznikají při metodě sítí.

Další metoda, která se zdá vhodná pro řešení diferenčních rovnic, je Peacemanova-Rachfordova metoda střídavých směrů.

Abychom ji popsali v případě modelové soustavy (2.110), pišme matici A této soustavy ve tvaru

$$A = H + V, \tag{2.152}$$

kde H a V jsou matice řádu $(n-1)^2$ definované vzorci

$$H = \begin{bmatrix} C & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & C \end{bmatrix}, \tag{2.153}$$

$$V = \begin{bmatrix} 2I_r & -I_r & 0 & \dots & \dots & 0 \\ -I_r & \ddots & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -I_r \\ 0 & \dots & \dots & 0 & -I_r & 2I_r \end{bmatrix}, \tag{2.154}$$

v nichž I_r je jednotková matice řádu $n-1$ a C je třídiagonální matice řádu $n-1$ daná rovnicí

$$C = \begin{bmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & \ddots & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & \dots & 0 & -1 & 2 \end{bmatrix}, \tag{2.155}$$

Peacemanova-Rachfordova metoda je pak dána dvojicí vzorců

$$\begin{aligned} (\omega_i I + H)u^{(i+1/2)} &= (\omega_i I - V)u^{(i)} + g, \\ (\omega_i I + V)u^{(i+1)} &= (\omega_i I - H)u^{(i+1/2)} + g, \quad i = 0, 1, \dots, \end{aligned} \tag{2.156}$$

kde I je jednotková matice řádu $(n-1)^2$ a na vektor $u^{(i+1/2)}$ je třeba se dívat jako na mezivýsledek.

Rozklad (2.152) vznikne tak, že píšeme levou stranu rovnice (2.110) ve tvaru $(-u_{k-1,s}+2u_{ks}-u_{k+1,s})+(-u_{k,s-1}+2u_{ks}-u_{k,s+1})$, matici H vytvoříme z prvního sčítance a matici V z druhého sčítance.

Pro matici $\omega_i I + H$ na levé straně první rovnice (2.156) platí

$$\omega_i I + H = \begin{bmatrix} \omega_i I_r + C & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \omega_i I_r + C \end{bmatrix}, \tag{2.157}$$

takže je to blokově diagonální matice. Položíme-li tedy

$$\begin{aligned} u^{(i+1/2)} &= [u_1^{(i+1/2)T}, \dots, u_{n-1}^{(i+1/2)T}]^T, \\ (\omega_i I - V)u^{(i)} + g &= [v_1^T, \dots, v_{n-1}^T]^T, \end{aligned} \tag{2.158}$$

rozpadne se první soustava ve vzorci (2.156) na $n-1$ soustav

$$(\omega_i I_r + C)u_k^{(i+1/2)} = v_k, \quad k = 1, \dots, n-1, \tag{2.159}$$

s třídiagonálními maticemi. Vektory $u_k^{(i+1/2)}$, z nichž je sestaven vektor $u^{(i+1/2)}$, můžeme tedy pohodlně a velmi ekonomicky vypočítat Gaussovou eliminací pro třídiagonální matici. Výpočet vektoru $u^{(i+1/2)}$ tak vede na řešení $n-1$ soustav s třídiagonálními maticemi pro $n-1$ neznámých. Všimněme si přitom, že vzhledem k použitému očíslování složek hledaného vektoru u představuje vektor $u_k^{(i+1/2)}$ hodnoty přibližného řešení na přímce $y = (n-k)h$ rovnoběžné s osou x. Uvědomíme-li si, že matice $\omega_i I + V$ je blokově třídiagonální, přičemž jednotlivé bloky jsou diagonální matice, je zřejmé, že očíslujeme-li nyní na rozdíl od předchozího výkladu neznámé v soustavě (2.110) po sloupcích odshora dolů a odleva doprava, lze matici $\omega_i I + V$ psát ve tvaru

$$\omega_i I + V = \begin{bmatrix} \omega_i I_r + C & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \omega_i I_r + C \end{bmatrix}, \tag{2.160}$$

tj. zcela stejně jako matici $\omega_i I + H$. Druhá soustava ve vzorci (2.156) se tedy rozpadne na $n-1$ soustav

$$(\omega_i I + C)u_k^{(i+1)} = w_k, \quad k = 1, \dots, n-1, \tag{2.161}$$

s třídiagonálními maticemi a výpočet vektoru $u^{(i+1)}$ vede opět na řešení $n-1$ soustav o $n-1$ neznámých s třídiagonálními maticemi. Zdůrazněme však ještě jednou, že vektor $u^{(i+1)}$ je rozdělen na bloky jinak než vektor $u^{(i+1/2)}$ a jeho složka $u_k^{(i+1)}$ značí nyní hodnoty přibližného řešení na přímce $x = kh$ rovnoběžné s osou y. Složky vektoru přibližného řešení tedy počítáme v prvním půlkroku metody (2.156) postupně po přímkách rovnoběžných s osou x a v druhém půlkroku po přímkách rovnoběžných s osou y. Odtud také pochází název metody.

Vezmeme-li v úvahu způsob, kterým jsme došli k maticím V a H v rozkladu (2.152), je snadné zobecnit metodu střídavých směrů i na podstatně obecnější diferenciální rovnici (2.1) s okrajovou podmínkou (2.68) v obecné oblasti Ω. Dokonce není ani nutné omezit se na rovnoměrnou síť. Matice C a I_r v rovnicích (2.153) a (2.154) nebudou ovšem konstantní a v případě obecné oblasti nebudou dokonce ani stejných řádů, jejich třídiagonální, resp. diagonální struktura však zůstane zachována. Z algoritmického hlediska bude tedy výpočet jednotlivých iterací probíhat stejně jako v popsaném příkladě.

Rovnice (2.156) představují nestacionární iterační metodu, kde čísla ω_i (zvaná iterační parametry) se mohou od iterace k iteraci měnit a je třeba je volit tak, aby konvergence byla co možná nejrychlejší. Obvykle se postupuje tak, že pro každých m po sobě jdoucích iteracích se za ně bere konečná posloupnost čísel ω_i, $i = 0, \dots, m-1$. Iterační parametry se tedy periodicky opakují. Pro ilustraci uveďme bez zdůvodnění některé volby iteračních parametrů, které byly v literatuře navrženy. Položíme-li $m = 1$, je $\omega_i = \omega$ pro každé i a optimální hodnota iteračního parametru je dána vzorcem

$$\omega_{\text{opt}} = 4 \sin \frac{\pi}{2n} \cos \frac{\pi}{2n}. \tag{2.162}$$

Metoda střídavých směrů konverguje v tomto případě asi tak rychle jako optimalizovaná superrelaxační metoda. Protože však její provedení vyžaduje více početních operací, než je tomu u superrelaxační metody, nelze tuto její variantu příliš doporučit. Podstatné zlepšení rychlosti konvergence Peacemanovy-Rachfordovy metody střídavých směrů lze dosáhnout až teprve tehdy, připustíme-li $m > 1$, tj. užijeme-li více iteračních parametrů než jeden. Byla navržena celá řada postupů, jak tyto parametry volit, a v současné době je této problematice věnována už rozsáhlá literatura. Jedna z možností, jak určit nejen dobré parametry, které jsou v případě výše uvažované modelové úlohy blízké optimálním parametrům, ale i jejich počet, tj. číslo m, pochází od Douglase. Číslo m se při tomto postupu určí jako největší celé číslo, pro něž platí nerovnost

$$\left(\frac{2^{1/2}+1}{2^{1/2}-1}\right)^m \sin^2 \frac{\pi}{2n} \leqq \cos^2 \frac{\pi}{2n}, \tag{2.163}$$

a iterační parametry ω_i se pro $i = 0, \dots, m-1$ počítají ze vzorců

$$\omega_i = \frac{4}{2^{1/2}-1} \left(\frac{2^{1/2}+1}{2^{1/2}-1}\right)^i \sin^2 \frac{\pi}{2n} \tag{2.164}$$

nebo ze vzorců

$$\omega_i = \frac{4}{2^{1/2}+1} \left(\frac{2^{1/2}-1}{2^{1/2}+1}\right)^i \cos^2 \frac{\pi}{2n}. \tag{2.165}$$

Ačkoliv z teoretického hlediska jsou obě tyto možnosti rovnocenné, praktické zkušenosti naznačují, že druhá volba vede k větší rychlosti konvergence.

Jiná posloupnost iteračních parametrů pochází od Wachspresse a je dána vzorcem

$$\omega_i = \frac{4 \cos^2 \frac{\pi}{2n}}{\left(\operatorname{cotg}^2 \frac{\pi}{2n}\right)^{i/(m-1)}}, \quad i = 0, \dots, m-1, \tag{2.166}$$

kde m je nejmenší celé číslo, pro něž platí nerovnosti

$$m \geqq 2, \quad (2^{1/2}-1)^{m-1} \leqq \operatorname{tg} \frac{\pi}{2n}. \tag{2.167}$$

Příznivé vlastnosti všech uvedených voleb iteračních parametrů jsou dokázány za předpokladu, že matice H a V v rozkladu (2.152) komutují. Je-li tento předpoklad porušen (např. proto, že daná oblast není čtverec, nebo že koeficienty dané rovnice nejsou konstantní), bere se často za posloupnost iteračních parametrů posloupnost, která se získá pro modelovou úlohu na nejmenším čtverci obklopujícím danou oblast. I když teoretické zdůvodnění zde není k dispozici, dochází při tomto postupu ve srovnání např. se superrelaxační metodou k podstatnému urychlení konvergence.

Poznamenejme ještě, že existuje celá řada variant metody střídavých směrů. Tak např. poměrně častá je tzv. *lokálně jednorozměrná metoda*, která vychází opět z rozkladu (2.152) a v níž se iterace počítají z dvojic vzorců

$$\begin{aligned} (\omega_i I + H) u^{(i+1/2)} &= (\omega_i I - H) u^{(i)} + g_1, \\ (\omega_i I + V) u^{(i+1)} &= (\omega_i I - V) u^{(i+1/2)} + g_2. \end{aligned} \tag{2.168}$$

Při užití těchto vzorců se postupuje podobně jako u Peacemanovy-Rachfordovy metody.

Upozorněme také, že metodu střídavých směrů lze užít i pro řešení diferenčních rovnic, které vzniknou diskretizací m-dimenzionálních úloh, kde $m > 2$. V tomto případě se rozkládá matice na m sčítanců, z nichž každý odpovídá náhradě derivací ve směru jedné souřadnicové osy, a jedna iterace se skládá z m dílčích kroků.

Uveďme končeně ještě dvě iterační metody, jejichž teoretický výzkum je v současné době velmi intenzívní a není ani zdaleka ještě ukončen, pro něž však existuje řada indikací, že by mohlo jít o vůbec nejefektivnější metody pro řešení soustav lineárních rovnic, které vznikají v metodě sítí nebo v metodě konečných prvků.

První z nich, *metoda sdružených gradientů*, slouží k přibližnému řešení soustavy $Au = g$ se symetrickou pozitivně definitní maticí. Je to, jak už jsme uvedli výše, gradientní metoda, v níž se kvadratická forma Q daná rovnicí (2.132) minimalizuje postupně ve směrech $v^{(i)}$, které vzniknou A-ortogonalizací vektorů reziduí. Příslušné vzorce jsou

$$\begin{aligned} v^{(1)} &= r^{(1)} = g - Au^{(1)}, \\ \alpha_i &= \frac{v^{(i)T} r^{(i)}}{v^{(i)T} A v^{(i)}}, \\ u^{(i+1)} &= u^{(i)} + \alpha_i v^{(i)}, \\ r^{(i+1)} &= r^{(i)} - \alpha_i A v^{(i)}, \end{aligned} \tag{2.169}$$

$$\beta_i = \frac{v^{(i)T} A r^{(i+1)}}{v^{(i)T} A v^{(i)}},$$
$$v^{(i+1)} = r^{(i+1)} + \beta_i v^{(i)}.$$

Tato metoda má pozoruhodnou vlastnost, že za předpokladu, že se nedopouštíme zaokrouhlovacích chyb, vede po N krocích (N je řád matice A) k přesnému řešení. Nejde tedy v přesném slova smyslu o iterační metodu, přesto se však jako taková užívá, neboť ve speciálních případech, zejména při užití na diferenční rovnice vznikajících v metodě sítí a v metodě končených prvků, dává často řešení s požadovanou přesností mnohem dříve než po N iteracích. Efektivitu metody sdružených gradientů je možné podstatně zvýšit úpravami matice A, které mají zhruba řečeno za cíl vhodně modifikovat její spektrum. V tomto případě se pak mluví o *předpodmíněné metodě sdružených gradientů*. Jde však o postupy dosti komplikované, takže do podrobností nelze v tomto elementárním textu zacházet.

Další iterační metoda, o níž se zde chceme zmínit, je *metoda více sítí* (multigrid method). Tuto metodu nelze jednoduše zařadit do žádné z výše zmíněných skupin iteračních metod a ani její teoretické fundování není dosud zdaleka ukončeno a v mnoha jejích variantách se zatím postupuje intuitivně. Proto naznačíme její základní myšlenku jen v těch nejhrubších obrysech, aniž bychom se sebeméně snažili udat nějaký konkrétní návod ke skutečnému počítání. Důvod, proč o ní vůbec mluvíme, je ten, že chceme čtenáře upozornit, že vůbec nějaká taková metoda existuje. Metoda vychází z toho, že řešení soustavy (2.111) představuje přibližné řešení dané diferenciální rovnice na síti $K^{(h)}$, které pro $h \to 0$ konverguje k přesnému řešení. Proto, je-li h malé, jsou řešení této soustavy při integračním kroku $h, 2h, \ldots$ v těch uzlech, které jsou společné sítím $K^{(h)}, K^{(2h)}, \ldots$, blízká. Této skutečnosti metoda více sítí využívá tak, že ke konstrukci zlepšené aproximace na síti $K^{(h)}$ se užije kombinace nějaké jednoduché iterační metody (např. Gaussovy-Seidelovy metody nebo metody sdružených gradientů) a oprav, které se vypočtou interpolací přibližného řešení z hrubší sítě na jemnější síť. Přibližné řešení na hrubší síti se získá snáze a lze to učinit např. zase tak, že se naznačeným způsobem užije ještě hrubší síť. Zakončeme tuto stručnou poznámku o metodě více sítí, že k tomu, aby byla skutečně efektivní, je třeba její dobrá implementace na počítači.

2.4 Obecné otázky konvergence a odhadů chyb při metodě sítí

V tomto odstavci popíšeme stručně princip konvergenčních důkazů pro metodu sítí. Začneme tím, že zformulujeme problém řešení okrajové úlohy metodou sítí abstraktně.

2.4.1 Základní pojmy teorie diferenčních schémat

Buď dána oblast Ω s hranicí Γ. Vyšetřujme diferenciální rovnici

$$Lu = f \quad \text{v } \Omega, \tag{2.170}$$

kde u je hledaná funkce, f daná funkce a L lineární diferenciální operátor, s okrajovými podmínkami

$$l_i u = \varphi_i \text{ na } \Gamma_i, \quad i = 1, \ldots, s. \tag{2.171}$$

Zde Γ_i jsou nějaké části hranice Γ, φ_i funkce zadané na Γ_i a l_i lineární operátory přiřazující dané funkci u funkci definovanou na množině Γ_i. Poznamenejme, že množiny Γ_i nemusí být pro různá i disjunktní. Stejně tak není nutné, aby sjednocení množin Γ_i vyčerpalo celou hranici Γ oblasti Ω. Tak např. v případě okrajových podmínek (2.74) pro biharmonickou rovnici (2.73) je $s = 2$, $\Gamma_1 = \Gamma_2 = \Gamma$, $l_1 u = u$ a $l_2 u = \partial u / \partial \nu$.

Buď dále k libovolnému $h > 0$ dána v uzavřené oblasti $\bar{\Omega}$ nějaká konečná množina bodů, kterou nazveme *sítí* a označíme $\bar{\Omega}^{(h)}$, a buď L_h lineární operátor, který funkci $u^{(h)}$ definované na síti $\bar{\Omega}^{(h)}$ přiřazuje funkci $L_h u^{(h)}$ definovanou na nějaké podmnožině $\Omega^{(h)}$ množiny $\bar{\Omega}^{(h)}$ Množinu $\Omega^{(h)}$ nazveme množinou *vnitřních uzlů* sítě $\bar{\Omega}^{(h)}$. Buď konečně $l_i^{(h)}$, $i = 1, \ldots, s$, operátor, který funkci $u^{(h)}$ definované na $\bar{\Omega}^{(h)}$ přiřazuje funkci $l_i^{(h)} u^{(h)}$ definovanou na množině $\Gamma_i^{(h)} \subset \bar{\Omega}^{(h)}$, $\Gamma_i^{(h)} \cap \Omega^{(h)} = \emptyset$, a $\Lambda_i^{(h)}$ operátor, který funkci φ_i uvažované na množině $\Gamma_{0i}^{(h)}$, což je konečná podmnožina množiny Γ_i, přiřazuje funkci $\Lambda_i^{(h)} \varphi_i$ definovanou na množině $\Gamma_i^{(h)}$. Uzly z množiny $\bigcup_{i=1}^{s} \Gamma_i^{(h)}$ nazveme *hraničními uzly*. V dalším budeme předpokládat, že platí $\bigcup_{i=1}^{s} \Gamma_i^{(h)} \cup \Omega^{(h)} = \bar{\Omega}^{(h)}$.

Uvažujme nyní diferenční rovnici

$$L_h u^{(h)} = f, \tag{2.172}$$

jejíž pravá strana je definována na množině $\Omega^{(h)}$ vnitřních uzlů a je tam rovna hodnotám pravé strany diferenciální rovnice (2.170), s okrajovými podmínkami

$$l_i^{(h)} u^{(h)} = \Lambda_i^{(h)} \varphi_i, \quad i = 1, \ldots, s. \tag{2.173}$$

Každá rovnice (2.173) představuje končeně mnoho rovnic pro uzly z množiny $\Gamma_i^{(h)}$ a rovnice (2.172) konečně mnoho rovnic pro uzly z množiny $\Omega^{(h)}$, takže soustava (2.172) s okrajovými podmínkami (2.173) tvoří soustavu lineárních rovnic pro určení hodnot neznámé funkce $u^{(h)}$ v uzlech sítě $\bar{\Omega}^{(h)}$. Diferenční rovnice (2.172) s okrajovými podmínkami (2.173) je tedy konečnědimenzionálním analogem diferenciální rovnice (2.170) s okrajovými podmínkami (2.171). Operátory $\Lambda_i^{(h)}$ v rovnicích (2.173) vlastně představují způsob, jakým se převádějí dané okrajové podmínky do uzlů sítě $\bar{\Omega}^{(h)}$. Tak např. pro přepis Dirichletových okrajových podmínek pro rovnici (2.1) pomocí lineární interpolace je $s = 1$, množina $\Gamma_1^{(h)}$ je množina uzlů typu A z obr. 2.2a, množina $\Gamma_{01}^{(h)}$ je množina bodů typu C z téhož obrázku a je

$$(l_1^{(h)} u)(A) = u^{(h)}(A) - \frac{\sigma}{1+\sigma} u^{(h)}(B), \tag{2.174}$$
$$(\Lambda_1^{(h)} \varphi)(A) = \frac{1}{1+\sigma} \varphi(C).$$

Poznamenejme, že podobně obecně jako v případě okrajových podmínek bychom mohli převádět do sítě i pravou stranu dané diferenciální rovnice. Na výše uvedený speciální případ, kdy pravé straně dané diferenciální rovnice přiřazujeme funkci definovanou na síti jejími funkčními hodnotami v uzlech sítě, jsme se omezili proto, že je v praxi daleko nejčastější a nejužívanější.

2.4.2 Obecné věty o konvergenci metody sítí

Máme-li dokázat konvergenci metody sítí nebo sestrojit odhad její chyby, je třeba posoudit velikost rozdílu $u^{(h)} - u$. K tomu cíli je třeba nějakým způsobem normovat funkce, které vystupují v našich úvahách. Buď tedy U normovaný prostor funkcí definovaných v $\bar{\Omega}$ a takových, že pro každou funkci $u \in U$ mají smysl výrazy Lu a $l_i u$. Buďte dále F, resp. Φ_i, $i = 1, \dots, s$, normované prostory funkcí definovaných na Ω, resp. na Γ_i a takových, že pro $u \in U$ platí $Lu \in F$ a $l_i u \in \Phi_i$. Analogicky $U^{(h)}$ je normovaný prostor funkcí definovaných na $\bar{\Omega}^{(h)}$ a $F^{(h)}$, resp. $\Phi_i^{(h)}$ jsou normované prostory funkcí definovaných na $\Omega^{(h)}$, resp. $\Gamma_i^{(h)}$, přičemž pro $u^{(h)} \in U^{(h)}$ a $\varphi_i \in \Phi_i$ platí $L_h u^{(h)} \in F^{(h)}$, $l_i^{(h)} u^{(h)} \in \Phi_i^{(h)}$ a $\Lambda_i^{(h)} \varphi_i \in \Phi_i^{(h)}$. Předpokládejme, že každá funkce $u \in U$, resp. $f \in F$ je jako funkce definovaná pouze na množině $\bar{\Omega}^{(h)}$, resp. $\Omega^{(h)}$ prvkem prostoru $U^{(h)}$, resp. $F^{(h)}$. Pak mají smysl výrazy $L_h u$ a $l_i^{(h)} u$. Předpokládejme dále, že normy zavedené v prostorech, které jsme právě definovali, jsou takové, že pro každou funkci $u \in U$, $f \in F$ a $\varphi_i \in \Phi_i$ platí

$$\begin{aligned} &\|u\|_{U^{(h)}} \to \|u\|_U, \quad \|f\|_{F^{(h)}} \to \|f\|_F, \\ &\|\Lambda_i^{(h)} \varphi_i\|_{\Phi_i^{(h)}} \to \|\varphi_i\|_{\Phi_i} \end{aligned} \tag{2.175}$$

pro $h \to 0$. Následující definice shrnují požadavky, které je nutno klást na danou metodu sítí, aby bylo možné odvodit základní informace o její konvergenci.

Definice 2.1. Řekneme, že diferenční rovnice (2.172) s okrajovými podmínkami (2.173) *aproximuje* diferenciální rovnici (2.170) s okrajovými podmínkami (2.171), platí-li pro každou funkci $u \in U$

$$\begin{aligned} &\|Lu - L_h u\|_{F^{(h)}} \to 0, \\ &\|\Lambda_i^{(h)}(l_i u) - l_i^{(h)} u\|_{\Phi_i^{(h)}} \to 0 \end{aligned} \tag{2.176}$$

pro $h \to 0$. Výrazy na levé straně rovnic (2.176) se nazývají *lokální chyby* dané metody.

Definice 2.2. Řekneme, že aproximace diferenciální rovnice (2.170) s okrajovými podmínkami (2.171) diferenční rovnicí (2.172) s okrajovými podmínkami (2.173) je *řádu* p, existují-li pro každou funkci $u \in U$ konstanty M a M_i tak, že platí

$$\begin{aligned} &\|Lu - L_h u\|_{F^{(h)}} \leqq M h^p, \\ &\|\Lambda_i^{(h)}(l_i u) - l_i^{(h)} u\|_{\Phi_i^{(h)}} \leqq M_i h^p \end{aligned} \tag{2.177}$$

pro každé dostatečně malé h.

Definice 2.3. Řekneme, že diferenční rovnice (2.172) s okrajovými podmínkami (2.173) je *korektní* (nebo *stabilní vzhledem k vstupním datům*), má-li při libovolných pravých stranách f a φ_i právě jedno řešení a existují-li konstanty N a N_i takové, že pro libovolnou funkci $u^{(h)} \in U^{(h)}$ a libovolné dostatečně malé h platí

$$\|u^{(h)}\|_{U^{(h)}} \leqq N \|L_h u^{(h)}\|_{F^{(h)}} + \sum_{i=1}^{s} N_i \|l_i^{(h)} u^{(h)}\|_{\Phi_i^{(h)}}. \tag{2.178}$$

Korektnost diferenční rovnice (2.172) s okrajovými podmínkami (2.173) tedy vlastně značí spojitou závislost jejího řešení na pravé straně rovnice a okrajových podmínek. Poznamenejme, že uvažujeme-li diferenční rovnici (2.173) s homogenními okrajovými podmínkami

$$l_i^{(h)} u^{(h)} = 0 \tag{2.179}$$

a platí-li nerovnost

$$\|u^{(h)}\|_{U^{(h)}} \leqq N \|L_h u^{(h)}\|_{F^{(h)}}, \tag{2.180}$$

mluvíme o *stabilitě vzhledem k pravé straně*; podobně, uvažujeme-li homogenní rovnici

$$L_h u^{(h)} = 0 \tag{2.181}$$

s nehomogenními okrajovými podmínkami (2.173) a platí-li nerovnost

$$\|u^{(h)}\|_{U^{(h)}} \leqq \sum_{i=1}^{s} N_i \|l_i^{(h)} u^{(h)}\|_{\Phi_i^{(h)}}, \tag{2.182}$$

mluvíme o *stabilitě vzhledem k okrajovým podmínkám*. Analogicky lze také zavést pojem stability pouze vzhledem k některé okrajové podmínce. Diferenční rovnice je tedy korektní, je-li stabilní vzhledem ke všem okrajovým podmínkám a vzhledem k pravé straně.

Zavedené pojmy dávají možnost zformulovat konvergenční větu pro abstraktní metodu sítí.

Věta 2.8. *Nechť $u \in U$ je řešení diferenciální rovnice* (2.170) *s okrajovými podmínkami* (2.171). *Nechť diferenční rovnice* (2.172) *s okrajovými podmínkami* (2.173) *aproximuje diferenciální rovnici* (2.170) *s okrajovými podmínkami* (2.171). *Konečně nechť diferenční rovnice* (2.172) *s okrajovými podmínkami* (2.173) *je korektní. Pak platí*

$$\lim_{h \to 0} \|u^{(h)} - u\|_{U^{(h)}} = 0. \tag{2.183}$$

Je-li navíc řád aproximace roven číslu p, platí pro chybu přibližného řešení odhad

$$\|u^{(h)} - u\|_{U^{(h)}} \leqq h^p \left(MN + \sum_{i=1}^{s} M_i N_i \right). \tag{2.184}$$

D ů k a z . Z rovnic (2.170) až (2.173) plyne, že je

$$L_h(u^{(h)} - u) = L_h u^{(h)} - L_h u + Lu - Lu = Lu - L_h u \tag{2.185}$$

a

$$l_i^{(h)}(u^{(h)} - u) = l_i^{(h)} u^{(h)} - l_i^{(h)} u + \Lambda_i^{(h)}(l_i u) - \Lambda_i^{(h)}(l_i u) = \Lambda_i^{(h)}(l_i u) - l_i^{(h)} u. \tag{2.186}$$

Odtud vzhledem ke korektnosti dané diferenční rovnice s danými okrajovými podmínkami dostáváme

$$\|u^{(h)} - u\|_{U^{(h)}} \leqq N \|L_h(u^{(h)} - u)\|_{F^{(h)}} + \sum_{i=1}^{s} N_i \|l_i^{(h)}(u^{(h)} - u)\|_{\Phi_i^{(h)}} = \tag{2.187}$$

$$= N \|Lu - L_h u\|_{F^{(h)}} + \sum_{i=1}^{s} N_i \|\Lambda_i^{(h)}(l_i u) - l_i^{(h)} u\|_{\Phi_i^{(h)}}.$$

Tvrzení věty nyní už snadno dostaneme užitím rovnice (2.187) a rovnic (2.176), resp. (2.177) z definic 2.1, resp. 2.2. Věta je dokázána.

Předpoklady této věty lze ve speciálních situacích zeslabit. Je-li např. některá podmínka přepsána do diferenčního tvaru přesně, tj. je-li $\Gamma_i^{(h)} \subset \Gamma_i$, $l_i^{(h)} = l_i$ a $\Lambda_i^{(h)} \varphi_i = \varphi_i$ pro některý index i, není třeba v předpokladu o korektnosti požadovat spojitou závislost na této podmínce. Upozorněme také, že bez podstatných obtíží lze zformulovat obdobnou větu pro obecný případ nelineární soustavy diferenciálních rovnic s nelineárními okrajovými podmínkami.

Teoretické schéma, které jsme právě uvedli, se v podstatě ať už v té či oné podobě vždy používá při důkazu konvergence libovolné diferenční metody a při odhadu její chyby. Vždy je tedy třeba prověřit platnost požadavků (2.177) a (2.178), tj. vždy je třeba učinit si představu o velikosti lokální chyby metody — to většinou není příliš obtížné a často na to stačí pouhý Taylorův vzorec — a vyšetřit korektnost příslušného diferenčního přepisu, tj. vyšetřit spojitou závislost tohoto přepisu na okrajových podmínkách a na pravé straně. Tento poslední problém je už většinou podstatně obtížnější. U rovnic druhého řádu zde velmi pomáhá už vícekrát připomenutý princip maxima. U rovnic čtvrtého řádu, kde věta o maximu neplatí, si vypomáháme pozitivní definitností diferenčního problému, nerovnostmi, které, fyzikálně řečeno, bilancují energii soustavy popsané danou diferenciální rovnicí apod. Podaří-li se všechny tyto obtíže překlenout, dostáváme ve vzorci (2.184) odhad chyby užité metody. Tento odhad je však prakticky málo cenný, neboť bývá značně obtížné určit konkrétní velikost v něm vystupujících konstant (některé z nich závisí na vyšších derivacích hledané funkce), kromě toho bývá vždy značně pesimistický. V praxi se proto k posouzení chyby užívá velice často *metoda polovičního kroku*,

o níž jsme hovořili v souvislosti s obyčejnými diferenciálními rovnicemi v kap. I. Teoretický podklad k jejímu užití dává následující věta.

Věta 2.9. *Nechť jsou splněny předpoklady věty 2.8 a nechť existují funkce ψ a ψ_i nezávislé na h tak, že pro dané řešení diferenciální rovnice* (2.170) *s okrajovými podmínkami* (2.171) *platí*

$$\lim_{h \to 0} \|h^{-p}(Lu - L_h u) - \psi\|_{F^{(h)}} = 0, \tag{2.188}$$

$$\lim_{h \to 0} \|h^{-p}[\Lambda_i^{(h)}(l_i u) - l_i^{(h)} u] - \Lambda_i^{(h)} \psi_i\|_{\Phi_i^{(h)}} = 0.$$

Nechť dále v nějaké třídě V, na které diferenční rovnice (2.172) *s okrajovými podmínkami* (2.173) *aproximuje diferenciální rovnici* (2.170) *s okrajovými podmínkami* (2.171), *existuje řešení okrajové úlohy*

$$Lw = \psi, \quad l_i w = \psi_i, \quad i = 1, \ldots, s. \tag{2.189}$$

Pak platí

$$\lim_{h \to 0} \|h^{-p}(u^{(h)} - u) - w\|_{U^{(h)}} = 0. \tag{2.190}$$

D ů k a z . Za uvedených předpokladů platí pro $h \to 0$

$$\begin{aligned} \|L_h[h^{-p}(u^{(h)} - u) - w]\|_{F^{(h)}} &= \|h^{-p} L_h(u^{(h)} - u) - L_h w + Lw - Lw\|_{F^{(h)}} \leqq \\ &\leqq \|h^{-p} L_h(u^{(h)} - u) - \psi\|_{F^{(h)}} + \|Lw - L_h w\|_{F^{(h)}} = \\ &= \|h^{-p}(L_h u^{(h)} - L_h u + Lu - Lu) - \psi\|_{F^{(h)}} + \|Lw - L_h w\|_{F^{(h)}} = \\ &= \|h^{-p}(Lu - L_h u) - \psi\|_{F^{(h)}} + \|Lw - L_h w\|_{F^{(h)}} \to 0 \end{aligned} \tag{2.191}$$

a

$$\begin{aligned} &\|l_i^{(h)}[h^{-p}(u^{(h)} - u) - w]\|_{\Phi_i^{(h)}} = \\ &= \|h^{-p} l_i^{(h)}(u^{(h)} - u) - l_i^{(h)} w + \Lambda_i^{(h)}(l_i w) - \Lambda_i^{(h)}(l_i w)\|_{\Phi_i^{(h)}} \leqq \\ &\leqq \|h^{-p} l_i^{(h)}(u^{(h)} - u) - \Lambda_i^{(h)} \psi_i\|_{\Phi_i^{(h)}} + \\ &\quad + \|\Lambda_i^{(h)}(l_i w) - l_i^{(h)} w\|_{\Phi_i^{(h)}} = \\ &= \|h^{-p}[l_i^{(h)} u^{(h)} - l_i^{(h)} u + \Lambda_i^{(h)}(l_i u) - \Lambda_i^{(h)}(l_i u)] - \Lambda_i^{(h)} \psi_i\|_{\Phi_i^{(h)}} + \\ &\quad + \|\Lambda_i^{(h)}(l_i w) - l_i^{(h)} w\|_{\Phi_i^{(h)}} = \\ &= \|h^{-p}[\Lambda_i^{(h)}(l_i u) - l_i^{(h)} u] - \Lambda_i^{(h)} \psi_i\|_{\Phi_i^{(h)}} + \\ &\quad + \|\Lambda_i^{(h)}(l_i w) - l_i^{(h)} w\|_{\Phi_i^{(h)}} \to 0 \end{aligned} \tag{2.192}$$

Z těchto rovnic a z korektnosti vyšetřovaného diferenčního schématu však už tvrzení věty ihned plyne.

Předpoklady této věty jsou obvykle splněny, je-li řešení daného diferenciálního problému dostatečně hladké.

Vyjdeme-li z rovnice (2.190) a jsou-li $u^{(h_1)}$ a $u^{(h_2)}$ dvě řešení diferenční rovnice (2.172) s okrajovými podmínkami (2.173) získaná užitím sítí $\bar{\Omega}^{(h_1)} \subset \bar{\Omega}^{(h_2)}$, kde $h_1 = ch_2$ a $c > 1$, platí (jsou-li splněny předpoklady věty 2.9)

$$u^{(h_1)} - u = wh_1^p + o(h_1^p) \tag{2.193}$$

a

$$u^{(h_2)} - u = c^{-p} wh_1^p + o(h_1^p). \tag{2.194}$$

Vyloučíme-li z těchto rovnic přesné řešení u tak, že odečteme rovnici (2.194) od rovnice (2.193), dostaneme

$$u^{(h_1)} - u^{(h_2)} = \frac{c^p - 1}{c^p} wh_1^p + o(h_1^p). \tag{2.195}$$

Platí tedy

$$u^{(h_1)} - u = \frac{c^p}{c^p - 1}[u^{(h_1)} - u^{(h_2)}] + o(h_1^p). \tag{2.196}$$

Zanedbáme-li ve vzorci (2.196) člen $o(h_1^p)$, který je vyššího řádu, než je řád chyby, dostáváme odhad chyby metodou polovičního kroku. K jeho získání je tedy třeba řešit daný problém dvakrát se dvěma různými hodnotami parametru h.

Alternativně lze vzorce (2.196) užít také tak, že za přibližné řešení se bere funkce $u^{(e)}$ daná vzorcem

$$u^{(e)} = \frac{c^p}{c^p - 1} u^{(h_2)} - \frac{1}{c^p - 1} u^{(h_1)}. \tag{2.197}$$

Vzorec (2.197) dostaneme tak, že z rovnice (2.196) vypočteme přesné řešení u a zanedbáme člen $o(h_1^p)$. Tato funkce dá patrně přesnější aproximaci než kterákoliv z funkcí $u^{(h_1)}$ a $u^{(h_2)}$, zbavíme se však možnosti odhadu chyby.

Poznamenejme ještě, že postup, který vedl k rovnici (2.197), představuje první krok tzv. *Richardsonovy extrapolace*, jíž se někdy užívá v případě, že pro přibližné řešení lze odvodit asymptotický vzorec typu

$$u^{(h)} = u + w_1 h^p + w_2 h^{2p} + \dots. \tag{2.198}$$

3 Variační metody

Variační metody, které stručně popíšeme v tomto článku, jsou založeny na úplně stejných principech jako metody v čl. 4 z kap. II. Stejně jako tam se vychází ze skutečnosti, že řešení mnohých okrajových úloh pro parciální diferenciální rovnice eliptického typu je ekvivalentní úloze nalezení extrému integrálu, jehož Eulerovou rovnicí je daná diferenciální rovnice. Proto si nejprve stručně všimneme souvislosti

některých základních úloh pro eliptické parciální diferenciální rovnice s variačními úlohami. Omezíme se přitom převážně na dvoudimenzionální problémy.

3.1 Variační formulace okrajových úloh

3.1.1 Diferenciální rovnice druhého řádu

Nechť je dána diferenciální rovnice (2.1) s Dirichletovou okrajovou podmínkou (2.29). Předpokládejme přitom, že koeficient p je kladný, koeficient q nezáporný a že hranice oblasti a funkce p, q, f a γ jsou tak hladké, že existuje klasické řešení úlohy (2.1), (2.29). Při hledání funkcionálu, jehož extrémem je řešení této okrajové úlohy se opět jako v odst. 4.1.1 z kap. II omezíme na homogenní Dirichletovu okrajovou podmínku, tj. na podmínku

$$u(x, y) = 0, \quad (x, y) \in \Gamma. \tag{3.1}$$

Okrajová úloha s obecnou Dirichletovou okrajovou podmínkou se převede na úlohu typu (2.1), (3.1) stejně jako v případě okrajové úlohy pro obyčejnou diferenciální rovnici, a to tak, že se nalezne dostatečně hladká funkce w, která splňuje podmínku (2.29), a místo funkce u hledáme funkci $u - w$. Tato funkce pak splňuje diferenciální rovnici (2.1) s jinou pravou stranou a s homogenní okrajovou podmínkou. Funkce w se většinou nazývá *přípustná funkce*. Upozorněme, že zatímco v případě obyčejné diferenciální rovnice bylo určení přípustné funkce triviální, v případě více proměnných to už může být podstatně složitější problém. Dostatečnou hladkostí přípustné funkce zde rozumíme takovou hladkost, která dovoluje dosadit ji do dané diferenciální rovnice. Z dalších úvah vyplyne, že se lze spokojit s hladkostí menší a že stačí, aby přípustná funkce byla prvkem prostoru, v němž je přirozené hledat zobecněné řešení dané okrajové úlohy.

Zaveďme diferenciální operátor L předpisem

$$Lv = -\frac{\partial}{\partial x}\Big(p(x, y)\frac{\partial v}{\partial x}\Big) - \frac{\partial}{\partial y}\Big(p(x, y)\frac{\partial v}{\partial y}\Big) + q(x, y)v \tag{3.2}$$

a buď $\mathscr{D}_L$ množina funkcí spojitých v $\bar{\Omega}$, které mají v Ω spojité parciální derivace až do druhého řádu včetně a které splňují podmínku (3.1). Operátor L pak zřejmě zobrazuje tuto množinu do množiny funkcí spojitých v Ω. Při zavedení tohoto označení je řešení okrajové úlohy (2.1), (3.1) totéž, jako řešení rovnice

$$Lu = f \tag{3.3}$$

v množině $\mathscr{D}_L$. Definujme dále na množině $\mathscr{D}_L$ funkcionál F předpisem

$$F(u) = \int_\Omega (Lu)(x, y)u(x, y)\,\mathrm{d}x\,\mathrm{d}y - 2\int_\Omega f(x, y)u(x, y)\,\mathrm{d}x\,\mathrm{d}y \tag{3.4}$$

nebo alternativně předpisem

$$F(u) = \int_\Omega \left\{ p(x,y)\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] + q(x,y)u^2 \right\} \mathrm{d}x\,\mathrm{d}y - 2\int_\Omega f(x,y)u\,\mathrm{d}x\,\mathrm{d}y, \tag{3.5}$$

přičemž vyjádření funkcionálu F ve tvaru (3.5) vzniklo z vyjádření (3.4) užitím Greenova vzorce. Nyní už můžeme zformulovat následující větu.

Věta 3.1. *Buď $u \in \mathscr{D}_L$ řešení rovnice* (3.3). *Pak platí*

$$F(v) > F(u) \tag{3.6}$$

pro každou funkci $v \in \mathscr{D}_L$, $v \neq u$. Naopak, nechť existuje funkce $u \in \mathscr{D}_L$ taková, že pro ni nabývá funkcionál F minima v množině $\mathscr{D}_L$. Pak je tato funkce řešením rovnice (3.3).

Tato věta je obdobou k větě 4.1 z kap. II a její tvrzení plyne stejně snadno ze symetrie a pozitivní definitnosti operátoru L, jako tomu bylo u zmíněné věty. Symetrie operátoru L se bez obtíží dokáže užitím Greenova vzorce. Jeho pozitivní definitnost plyne z nerovnosti

$$\int_\Omega (Lu)(x,y)u(x,y)\,\mathrm{d}x\,\mathrm{d}y \geqq \gamma\|u\|^2_{\mathscr{L}_2}, \tag{3.7}$$

kde γ je kladná konstanta a symbol $\|u\|_{\mathscr{L}_2}$ značí normu funkce $u \in \mathscr{D}_L$ v prostoru $\mathscr{L}_2(\Omega)$ funkcí integrovatelných s kvadrátem v oblasti Ω. Důkaz této nerovnosti však už není zdaleka tak elementární, jako tomu bylo u analogické nerovnosti v jednodimenzionálním případě. Nebudeme jej proto provádět a odkazujeme čtenáře na specializovanou literaturu (viz např. Nečas (1967)).

Věta 3.1 ukazuje ekvivalenci mezi řešením okrajové úlohy (2.1), (3.1) a hledáním extrému funkcionálu F definovaného rovnicí (3.4) nebo (3.5). Situace je zde tedy stejná jako v odst. 4.1 v kap. II. Tato paralela jde však i dále a variační formulace okrajové úlohy tvoří stejně jako v jednodimenzionálním případě podklad k zobecnění pojmu okrajové úlohy. Myšlenkový postup je v podstatě stejný jako ve zmíněném jednodimenzionálním případě a uvedeme jej bez důkazů s odvoláním např. na už zmíněnou knihu Nečasovu (1967). Množina $\mathscr{D}_L$, která zřejmě tvoří vektorový prostor, se pokládá za podprostor Hilbertova prostoru $\mathscr{L}_2(\Omega)$ a operátor L definovaný rovnicí (3.2) představuje tak v $\mathscr{L}_2(\Omega)$ lineární symetrický operátor s definičním oborem $\mathscr{D}_L$. Protože platí pro každé $u \in \mathscr{D}_L$ také nerovnost (3.7), kterou lze stručně psát jako

$$(Lu, u) \geqq \gamma(u, u) \tag{3.8}$$

(kulaté závorky zde i v dalším značí skalární součin v prostoru $\mathscr{L}_2(\Omega)$), je možné

ve vektorovém prostoru $\mathscr{D}_L$ zavést nový skalární součin rovnicí

$$[u, v] = (Lu, v), \tag{3.9}$$

nebo, vzhledem ke Greenově vzorci, alternativně rovnicí

$$[u, v] = \int_\Omega \left[p(x,y)\left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right) + q(x,y)uv \right] \mathrm{d}x\,\mathrm{d}y. \tag{3.10}$$

Zavedeme-li ještě ve vektorovém prostoru $\mathscr{D}_L$ normu obvyklým způsobem, tj. rovnicí

$$\|u\|_L = [u, u]^{1/2}, \tag{3.11}$$

stane se vektorový prostor $\mathscr{D}_L$ lineárním normovaným prostorem. Poznamenejme, že právě zavedená norma se nazývá *energetická norma* a příslušný skalární součin *energetický skalární součin*. Zúplněním prostoru $\mathscr{D}_L$ v této normě dostaneme Hilbertův prostor $\mathscr{D}$, v němž je zřejmě podprostor $\mathscr{D}_L$ hustý. V jednodimenzionálním případě byl prostor $\mathscr{D}$ tvořen právě těmi funkcemi ze Sobolevova prostoru $\mathscr{H}^1$, které splňovaly dané homogenní okrajové podmínky. Ve vícedimenzionálním případě je situace složitější a platnost obdobného tvrzení závisí na topologických vlastnostech hranice Γ dané oblasti. Důležitá třída oblastí, pro níž zmíněné tvrzení platí, tj. pro niž je prostor $\mathscr{D}$ tvořen právě těmi funkcemi ze Sobolevova prostoru $\mathscr{H}^1(\Omega)$, které splňují okrajovou podmínku (3.1) ve smyslu stop (tento podprostor prostoru $\mathscr{H}^1(\Omega)$ se značí $\mathscr{H}_0^1(\Omega)$), je tvořena oblastmi s tzv. lipschitzovskou hranicí. Přitom řekneme, že hranice Γ dané oblasti je *lipschitzovská*, jestliže existuje m kartézských soustav souřadnic, čísla $\alpha > 0$ a $\beta > 0$ a m funkcí a_r definovaných a lipschitzovských v intervalech $(-\alpha, \alpha)$ tak, že platí (a) každý bod hranice Γ lze psát aspoň v jedné z uvedených soustav souřadnic ve tvaru $(x_r, a_r(x_r))$; (b) body (x_r, y_r), pro něž platí nerovnosti $-\alpha < x_r < \alpha$, $a_r(x_r) < y_r < a_r(x_r) + \beta$, leží v Ω a body (x_r, y_r), pro něž platí nerovnosti $-\alpha < x_r < \alpha$, $a_r(x_r) - \beta < y_r < a_r(x_r)$, leží vně množiny Ω (viz obr. 3.1). V dalším budeme předpokládat, že uvažovaná oblast do této třídy patří.

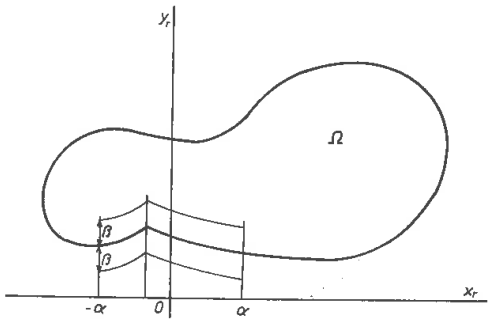
Samotný pojem Sobolevova prostoru $\mathscr{H}^k(\Omega)$ není také možno definovat tak jednoduše jako v případě jedné dimenze, kdy k jeho zavedení stačil pojem absolutní spojitosti a běžný pojem derivace. Ve vícedimenzionálním případě se prostor $\mathscr{H}^k(\Omega)$ nejobvykleji definuje (definici vyslovíme pro obecný případ funkcí m proměnných) jako množina měřitelných funkcí definovaných v oblasti $\Omega \subset \mathbf{E}^m$ (přesněji řečeno jako množina tříd funkcí, které se navzájem liší na množině míry nula), které mají zobecněné derivace až do k-tého řádu včetně, přičemž tyto derivace jsou integrovatelné s kvadrátem, a v níž je skalární součin definován rovnicí

$$(u, v)_{\mathscr{H}^k(\Omega)} = \sum_{\substack{\alpha_1+\ldots+\alpha_m \leqq k \\ \alpha_i \geqq 0}} \left(\frac{\partial^{\alpha_1+\ldots+\alpha_m} u}{\partial x_1^{\alpha_1} \ldots \partial x_m^{\alpha_m}}, \frac{\partial^{\alpha_1+\ldots+\alpha_m} v}{\partial x_1^{\alpha_1} \ldots \partial x_m^{\alpha_m}} \right). \tag{3.12}$$

Přitom řekneme, že funkce g definovaná a lokálně integrovatelná na množině Ω

Obr. 3.1

Oblast s lipschitsovskou hranicí



je *zobecněnou derivací* $\partial^{\alpha_1+\ldots+\alpha_m}/\partial x_1^{\alpha_1}\ldots\partial x_m^{\alpha_m}$ lokálně integrovatelné funkce f, platí-li

$$\int_\Omega f\frac{\partial^{\alpha_1+\ldots+\alpha_m}\varphi}{\partial x_1^{\alpha_1}\ldots\partial x_m^{\alpha_m}}\,\mathrm{d}\,x = (-1)^{\alpha_1+\ldots+\alpha_m}\int_\Omega g\varphi\,\mathrm{d}\,x \tag{3.13}$$

pro každou funkci φ, která je definovaná v Ω, má derivace všech řádů a jejíž nosič K (tj. uzávěr největší podmnožiny množiny Ω, na níž je funkce φ různá od nuly) je kompaktní množina, pro niž je $K \subset \Omega$. Chápeme-li derivace ve smyslu Schwarzových distribucí (srv. např. Schwarz (1957)), má každá lokálně integrovatelná funkce derivace všech řádů. V této souvislosti pak má funkce zobecněnou derivaci, je-li příslušná distributivní derivace ekvivalentní lokálně integrovatelné funkci. Proto se zobecněné derivace také nazývají derivacemi v *distributivním smyslu.*

Výrok, že prostor $\mathscr{D}$ je tvořen těmi funkcemi ze Sobolevova prostoru $\mathscr{H}^1(\Omega)$, které splňují okrajovou podmínku ve smyslu stop, který jsme užili výše, je ve dvou (i více) proměnných rovněž nutný. Na rozdíl od jednodimenzionálního případu, kdy každá funkce z prostoru $\mathscr{H}^1(a,b)$ je spojitá na intervalu $\langle a,b\rangle$, zde tomu tak není. Dá se pouze říci, že je integrovatelná s libovolně vysokou mocninou, a nemusí být dokonce ani omezená. Hraniční hodnotu v normálním smyslu lze tedy přiřadit jen takovým funkcím z prostoru $\mathscr{H}^1(\Omega)$, které jsou spojité v $\bar{\Omega}$. Dá se však ukázat, že operátor, označme jej S, který jsme vlastně právě zavedli a který spojité funkci z $\mathscr{H}^1(\Omega)$ přiřazuje její hodnoty na hranici, lze rozšířit na celý prostor $\mathscr{H}^1(\Omega)$ jako spojitý lineární operátor do prostoru $\mathscr{L}_2(\Gamma)$. Každé funkci $u \in \mathscr{H}^1(\Omega)$ je tedy přiřazena funkce $Su \in \mathscr{L}_2(\Gamma)$, která se nazývá její *stopa* a která se v hladkém případě rovná její hraniční hodnotě. Stopa funkce je proto přirozeným rozšířením pojmu hraniční hodnoty pro ty funkce z prostoru $\mathscr{H}^1(\Omega)$, které ji v normálním smyslu nemají.

Vraťme se k Hilbertově prostoru $\mathscr{D}$. Rovnice (3.10) definuje skalární součin $[u,v]$ nejen pro prvky z množiny $\mathscr{D}_L$ ale i pro libovolné dva prvky z prostoru $\mathscr{D}$, neboť platí $\mathscr{D} \subset \mathscr{H}^1(\Omega)$. Rozšíříme-li definici funkcionálu F pomocí vzorce

$$F(u) = [u,u] - 2(f,u) \tag{3.14}$$

na celý prostor $\mathscr{D}$, zůstává věta 3.1 v platnosti, budeme-li v ní brát funkce v místo z množiny $\mathscr{D}_L$ z celého prostoru $\mathscr{D}$. Opět stejně jako v jednodimenzionálním případě můžeme připustit v diferenciální rovnici (2.1) podstatně obecnější funkce než dosud. Stačí na ně klást jen takové požadavky, které zaručí, že výrazy na pravé straně rovnice (3.14) mají smysl. V tomto případě se může stát, že funkcionál F nenabývá minima v množině $\mathscr{D}_L$, ale v množině $\mathscr{D}$. Funkci $u \in \mathscr{D}$ udílící tomuto funkcionálu extrém, nazveme opět *zobecněným řešením.*

Za zobecněné řešení dané okrajové úlohy můžeme také pokládat, opět analogicky jako v jednodimenzionálním případě, funkci u, pro niž platí

$$[u,v] = (f,v) \tag{3.15}$$

pro každou funkci $v \in \mathscr{D}_L$. V této souvislosti mluvíme také o *slabém řešení.* Opět je slabé řešení totožné se zobecněným a pokud jde o jejich existenci, platí úplně beze změny věta 4.2 z kap. II. Poznamenejme také, že pojem slabého řešení, ne však zobecněného řešení, lze převést i na případ, že okrajová úloha je charakterizovaná bilineární formou, která není symetrická.

V podstatě úplně stejně jako výše lze formulovat okrajové úlohy pro eliptické parciální diferenciální rovnice druhého řádu s libovolným počtem nezávisle proměnných.

3.1.2 Diferenciální rovnice čtvrtého řádu

Uvažujme jako typický příklad biharmonickou rovnici (2.73) s homogenními okrajovými podmínkami

$$u = 0, \quad \frac{\partial u}{\partial n} = 0 \text{ na } \Gamma. \tag{3.16}$$

Od nehomogenních okrajových podmínek přejdeme k homogenním opět pomocí přípustné funkce. Její sestrojení je zde však obecně obtížnější než v případě rovnic druhého řádu, neboť je třeba splnit dvě okrajové podmínky.

Schéma dalšího postupu bude v podstatě stejné jako u rovnic druhého řádu a uvedeme je opět bez důkazu.

Nechť $\mathscr{D}_L$ značí nyní množinu funkcí čtyřikrát spojitě diferencovatelných v $\bar{\Omega}$ a splňujících okrajové podmínky (3.16). Definujeme-li na množině $\mathscr{D}_L$, která tvoří zřejmě vektorový prostor, operátor L rovnicí

$$Lu = \frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4}, \tag{3.17}$$

lze okrajovou úlohu (2.73), (3.16) stručně zapsat jako rovnici

$$Lu = f \tag{3.18}$$

v množině $\mathscr{D}_L$. Zavedeme-li v množině $\mathscr{D}_L$ skalární součin rovnicí

$$[u,v] = \int_\Omega \Big(\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 v}{\partial x^2} + 2\frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 v}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2}\frac{\partial^2 v}{\partial y^2}\Big)\,\mathrm{d}x\,\mathrm{d}y \tag{3.19}$$

a funkcionál F rovnicí

$$F(u) = [u,u] - 2(f,u), \tag{3.20}$$

je úloha (3.18) ekvivalentní úloze nalezení minima funkcionálu F v prostoru $\mathscr{D}$, který vznikne zúplněním vektorového prostoru $\mathscr{D}_L$ v normě dané skalárním součinem (3.19). Je-li hranice oblasti lipschitzovská, je prostor $\mathscr{D}$ roven podprostoru $\mathscr{H}_0^2(\Omega)$ takových funkcí ze Sobolevova prostoru $\mathscr{H}^2(\Omega)$, které splňují okrajové podmínky (3.16) ve smyslu stop. Extrém funkcionálu je tedy třeba hledat v prostoru $\mathscr{H}_0^2(\Omega)$.

Řešení dané okrajové úlohy lze i v tomto případě charakterizovat jako funkci u, pro niž platí rovnice

$$[u,v] = (f,v) \tag{3.21}$$

pro každou funkci $v \in \mathscr{D}_L$.

Uvědomíme-li si, že bilineární forma (3.19) vznikla za skalárního součinu (Lu,v) užitím Greenova vzorce, je snadné formulovat variačně i úlohy s obecnějším operátorem čtvrtého řádu.

3.1.3 Jiné typy okrajových podmínek, nehomogenní okrajové podmínky

Až dosud jsme se v příkladech variační formulace omezili na nejjednodušší typy okrajových podmínek. Popsaný postup lze však snadno přenést i na podstatně obecnější okrajové úlohy. Tak např. řešení diferenciální rovnice (2.1) s homogenní Newtonovou okrajovou podmínkou

$$\frac{\partial u}{\partial n} = -ku \text{ na } \Gamma \tag{3.22}$$

je ekvivalentní s úlohou nalézt minimum funkcionálu $F(u) = [u,u] - 2(f,u)$, kde

$$[u,v] = \int_\Omega \Big[p(x,y)\Big(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\Big) + q(x,y)uv\Big]\,\mathrm{d}x\,\mathrm{d}y + \int_\Gamma k(s)uv\,\mathrm{d}x, \tag{3.23}$$

na celém prostoru $\mathscr{H}^1(\Omega)$. Protože integrály na pravé straně identity (3.23) vznikly úpravou vzorce (Lu,v) pomocí Greenova vzorce při využití okrajových podmínek (3.22), je naznačeno, jak postupovat i v případě jiných okrajových podmínek.

Úlohy s nehomogenními okrajovými podmínkami jsme převedli na homogenní případ pomocí přípustné funkce. Lze však také postupovat přímo. Tak např. řešení diferenciální rovnice (2.1) s nehomogenní Dirichletovou okrajovou podmínku lze

nalézt minimalizací funkcionálu (3.5) ve třídě funkcí, které patří do Sobolevova prostoru $\mathscr{H}^1(\Omega)$ a které splňují tuto nehomogenní podmínku.

K tomu, abychom dostali řešení rovnice (2.1) s nehomogenní Newtonovou podmínkou (2.68), je třeba minimalizovat funkcionál

$$F(u) = \int_\Omega \Big\{p(x,y)\Big[\Big(\frac{\partial u}{\partial x}\Big)^2 + \Big(\frac{\partial u}{\partial y}\Big)^2\Big] + q(x,y)u^2\Big\}\,\mathrm{d}x\,\mathrm{d}y + \int_\Gamma k(s)[u^2 - 2u\gamma(s)]\,\mathrm{d}s - 2\int_\Omega fu\,\mathrm{d}x\,\mathrm{d}y \tag{3.24}$$

na celém prostoru $\mathscr{H}^1(\Omega)$.

Z předchozích příkladů je vidět, že Dirichletovy okrajové podmínky a Newtonovy okrajové podmínky se chovají podstatně odlišně. Dirichletovy podmínky je třeba při minimalizaci příslušného funkcionálu předem splnit. Naproti tomu Newtonovy podmínky vedou sice k nutnosti modifikace funkcionálu, který minimalizujeme, členem obsahujícím integraci po hranici dané oblasti, při volbě množiny funkcí, na níž hledáme minimum, k nim však nemusíme přihlížet a nalezené minimum už je splní automaticky. Toto odlišné chování souvisí s tím, jaký Hilbertův prostor vznikne při zúplnění množiny $\mathscr{D}_L$ v energetické normě. Zatímco v případě (homogenní) Dirichletovy podmínky prvky výsledného prostoru tuto podmínku splňují, v případě Newtonovy podmínky tomu tak není. Proto se také, a to i v případě obecnějších eliptických diferenciálních operátorů, podmínky, které se chovají jako Dirichletovy okrajové podmínky v našem příkladě, nazývají *stabilní okrajové podmínky*, a ty je tedy třeba při hledání minima splnit. Okrajové podmínky, které se chovají jako Newtonovy podmínky, se nazývají *nestabilní* nebo *přirozené okrajové podmínky*. Ty mají vliv na tvar příslušného funkcionálu, při hledání minima však k nim není třeba přihlížet. Do dalších podrobností nebudeme zacházet a odkazujeme čtenáře opět na příslušnou speciální literaturu.

3.2 Základní přibližné metody

V tomto odstavci popíšeme velmi stručně Ritzovu a Galerkinovu metodu řešení eliptických okrajových úloh, neboť jde o takřka doslovné opakování toho, co bylo o těchto metodách řečeno v odst. 4.2. kap. II.

3.2.1 Ritzova metoda

V odst. 3.1 jsme ukázali, že celá řada okrajových úloh pro parciální diferenciální rovnice eliptického typu je ekvivalentní úloze na hledání minima funkcionálu F typu $F(u) = [u,u] - 2(f,u)$ ve vhodném prostoru $\mathscr{D}$ (např. v Sobolevově prostoru $\mathscr{H}^k(\Omega)$ nebo jeho podprostoru $\mathscr{H}_0^k(\Omega)$). Výraz $[.,.]$ přitom znamená skalární součin, který je ekvivalentní se skalárním součinem ve zmíněném Sobolevově prostoru a který souvisí s daným diferenciálním operátorem a eventuálně i s okrajovými podmínkami způsobem, který byl naznačen v předchozím textu.

Základní myšlenka Ritzovy metody spočívá v tom (srv. odst. 4.1.1 z kap. II), že se zvolí konečnědimenzionální podprostor $\mathscr{D}_N$ prostoru $\mathscr{D}$ a za aproximaci řešení dané okrajové úlohy se pokládá funkce u_N, pro kterou nabývá funkcionál F svého minima v tomto konečnědimenzionálním prostoru. Zvolíme-li za bázi v prostoru D_N funkce $\Phi_1, \ldots, \Phi_N$, je hledaná aproximace u_N dána vzorcem

$$u_N = \sum_{k=1}^{N} c_k \Phi_k, \tag{3.25}$$

přičemž vektor koeficientů $c = (c_1, \ldots, c_N)^T$ je řešením soustavy lineárních rovnic

$$Ac = g, \tag{3.26}$$

kde A je čtvercová matice řádu N, jejíž prvky a_{ij} se určí ve vzorců

$$a_{ij} = [\Phi_i, \Phi_j], \tag{3.27}$$

a $g = (g_1, \ldots, g_N)^T$ je N-dimenzionální vektor o složkách daných rovnicemi

$$g_i = (f, \Phi_i). \tag{3.28}$$

Matice A se stejně jako v jednodimenzionálním případě nazývá *Gramova matice* nebo *matice tuhosti* příslušná k bázi $\Phi_1, \ldots, \Phi_n$ a je v námi uvedených příkladech symetrická a pozitivně definitní. Vektor g se nazývá *vektor zatížení.*

3.2.2 Galerkinova metoda

Základní myšlenka Galerkinovy metody vychází z toho, jak už bylo ostatně řečeno v odst. 4.2.2 v kap. II, že řešení dané okrajové úlohy vyhovuje rovnici $[u, v] = (f, v)$ pro každou funkci $v \in \mathscr{D}$. Je-li $\mathscr{D}_N$ konečnědimenzionální podprostor prostoru $\mathscr{D}$, je galerkinovské přibližné řešení taková funkce $u_N \in \mathscr{D}_N$, pro niž platí

$$[u_N, v_N] = (f, v_N) \tag{3.29}$$

pro každou funkci $v_N \in \mathscr{D}_N$. Tvoří-li funkce $\Phi_1, \ldots, \Phi_N$ bázi v prostoru $\mathscr{D}_N$, je toto přibližné řešení dáno stejnými vzorci jako u Ritzovy metody.

Galerkinova metoda je tedy totožná s Ritzovou metodou, jsou-li obě realizovatelné, a stejně jako v jednodimenzionálním případě je Galerkinova metoda obecnější, neboť se dá užít i v těch případech, kdy danou okrajovou úlohu nelze formulovat jako minimalizační úlohu.

Kvalita aproximace získané jak Galerkinovou, tak Ritzovou metodou je dána volbou konečnědimenzionálního prostoru $\mathscr{D}_N$, tj. volbou bázových funkcí $\Phi_1, \ldots, \Phi_N$. Zejména v předpočítačové éře numerického počítání byla navržena řada postupů, jak tyto bázové funkce volit hlavně s ohledem na to, aby v součtu (3.25) stačilo k přijatelné představě o chování řešení vzít jen několik málo sčítanců. Tato snaha byla diktována zejména tím, že na mechanických stolních kalkulátorech nebylo možné

a únosné provádět příliš velký počet početních operací. Tak např. se doporučovalo vzít za bázové funkce vlastní funkce „podobného" problému, přičemž podobný problém znamenal problém, v němž se např. místo proměnných koeficientů vzaly konstantní koeficienty, daná oblast se vnořila do nějaké kanonické oblasti apod. Je zřejmé, že univerzálnost těchto postupů byla značně omezená a že problémy, které bylo možno takto uspokojivě řešit, byly jen velice speciální.

V současné době se Ritzova a Galerkinova metoda užívá takřka výhradně v souvislosti s metodou konečných prvků. V této metodě byl nalezen značně univerzální způsob konstrukce konečnědimenzionálních podprostorů prostoru, v němž hledáme řešení. Metoda konečných prvků tak umožňuje řešit s velkou přesností i značně obecné problémy. Pro její důležitost jí budeme věnovat samostatný článek i přesto, že spadá pod obecné schéma variačních metod.

4 Metoda konečných prvků

Při popisu Ritzovy a Galerkinovy metody v odst. 3.2 jsme uvedli, že tyto metody mají v současné době největší význam ve spojitosti s metodou konečných prvků jako systematického způsobu konstrukce konečnědimenzionálních prostorů, v nichž hledáme příslušné aproximace. V tomto článku popíšeme důležité základní typy prostorů konečných prvků vhodných k řešení dvoudimenzionálních problémů a ukážeme teoretické problémy, které je třeba při jejich užití řešit. Všimneme si také, velice stručně, některých aspektů početní realizace metody konečných prvků.

Začneme rekapitulací základních hledisek, která je záhodno respektovat při konstrukci konečnědimenzionálních prostorů nutných k užití Ritzovy-Galerkinovy metody. Tato hlediska jsou do značné míry podobná těm, která jsme uvedli už v jednodimenzionálním případě, a proto je připomeneme jen velice krátce. Pro určitost a jednoduchost budeme mít zde na mysli úlohy druhého řádu. Pro rozdíl přibližného řešení u_N získaného Ritzovou-Galerkinovou metodou a přesného řešení u dané úlohy platí i zde nerovnost typu

$$\|u - u_N\|_{\mathscr{H}^1(\Omega)} \leqq M \inf_{\tilde{u} \in \mathscr{D}_N} \|u - \tilde{u}\|_{\mathscr{H}^1(\Omega)} \tag{4.1}$$

Jde-li o Ritzovu metodu, tj. je-li daná okrajové úloha definovaná jako minimalizační úloha a je-li příslušná energetická norma ekvivalentní s normou v Sobolevově prostoru $\mathscr{H}^1$, plyne tato nerovnost z faktu, že i ve dvoudimenzionálním případě je přibližné řešení ortogonální projekcí (v energetickém skalárním součinu) přesného řešení do příslušného podprostoru. Nerovnost (4.1) však platí pro podstatně širší množinu problémů čítaje v to i okrajové úlohy, které nelze formulovat jako úlohy na hledání minima funkcionálu F a u nichž lze hovořit pouze o slabém řešení. V této souvislosti se nerovnost (4.1) spojuje se jménem *Céovým* (viz např. Ciarlet (1978)). Otázka přesnosti Ritzovy-Galerkinovy aproximace je tedy opět jako v jedné dimenzi otázkou, jak přesně lze aproximovat funkci z prostoru $\mathscr{D}$ (což je většinou prostor

$\mathscr{H}^1(\Omega)$ nebo jeho podprostor) funkcí z prostoru $\mathscr{D}_N$. První požadavek, který je tedy při konstrukci podprostoru $\mathscr{D}_N$ žádoucí respektovat, je, aby měl dobré aproximační vlastnosti.

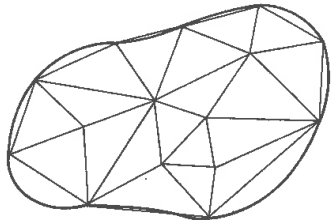
Druhé, stejně důležité hledisko při konstrukci prostoru $\mathscr{D}_N$ je požadavek maximální efektivnosti vzniklého algoritmu. Z rovnic (3.25) až (3.28) je vidět, že při realizaci Ritzovy-Galerkinovy metody je třeba vypočítat prvky matice A a složky vektoru g, řešit soustavu lineárních rovnic (3.26) a konečně provést součet (3.25). Zvolený prostor $\mathscr{D}_N$ musí tedy umožňovat takovou volbu báze, aby se prvky matice A a složky vektoru g počítaly co nejsnáze a aby matice soustavy (3.26) byla pokud možno řídká nebo ještě lépe pásová. Posledního požadavku se dosáhne tehdy, mají-li bázové funkce malé nosiče ve srovnání s oblastí Ω. Skalární součin ve vzorci (3.27) je totiž dán pomocí integrálů přes oblast Ω a zmíněná vlastnost má pak za následek, že většina prvků matice A je rovna nule. Je tedy žádoucí volit bázové funkce tak, aby měly i tuto vlastnost. Konečně zmenšení počtu potřebných operací se dosáhne také tehdy, mají-li některé (nebo všechny) koeficienty c_k přímo význam hodnot přibližného řešení v některých bodech dané oblasti, neboť pak odpadne nutnost počítat hodnoty součtu (3.25).

Metoda konečných prvků tyto požadavky do značné míry splňuje. Její základní myšlenka je založena na interpolaci a představuje přímé, nikoliv však triviální rozšíření myšlenek, z nichž jsme vycházeli v jednodimenzionálním případě. Při konstrukci prostoru $\mathscr{D}_N$ je tedy třeba rozdělit danou oblast Ω na konečný počet geometricky jednoduchých podoblastí a interpolaci provést pomocí funkcí, které jsou po částech rovny polynomům v těchto jednotlivých podoblastech. Každý polynom definovaný na určité podoblasti budeme charakterizovat funkčními hodnotami, eventuálně hodnotami některých derivací v nějakých pevně zvolených bodech této podoblasti tak, aby byl těmito podmínkami jednoznačně určen. Tyto body nazveme stejně jako v jednodimenzionálním případě *uzly* a příslušné veličiny, které v nich zadáváme, *uzlovými parametry*. První problém, kterým se tedy musíme zabývat, je volba rozkladu oblasti Ω. Tento rozklad by měl být na jedné straně tak obecný, aby dovoloval modelovat nepravidelné oblasti, na druhé straně tak jednoduchý, aby příliš nenarůstala výpočetní složitost. Praktické zkušenosti ukazují, že těmto do jisté míry protichůdným požadavkům velmi uspokojivě vyhovují jednoduché trojúhelníky nebo čtyřúhelníky. Při konstrukci konečnědimenzionálních prostorů metodou konečných prvků budeme tedy v dalším výkladu předpokládat, že daná oblast je sjednocením konečně mnoha trojúhelníků nebo čtyřúhelníků, přičemž kombinace obou útvarů současně nebudeme připouštět. Dále pak budeme předpokládat, že pro libovolné dva útvary rozkladu nastává právě jeden z těchto případů: (a) útvary jsou disjunktní; (b) útvary mají společnou celou jednu stranu; (c) útvary mají společný jeden vrchol (viz obr. 4.1 a 4.2).

O útvarech, pro něž na stává případ (b) nebo (c) budeme mluvit jako o sousedních útvarech. Za uzly, v nichž zadáváme uzlové parametry, tj. funkční hodnoty a hodnoty derivací, připouštíme kromě vrcholů ještě i některé další body ležící v daném

Obr. 4.1

Rozklad oblasti na trojúhelníky



Obr. 4.2

Rozklad oblasti na čtyřúhelníky

trojúhelníku nebo čtyřúhelníku (např. středy stran, těžiště apod.). Protože některé uzly jsou společné pro více útvarů rozkladu, užijí se hodnoty odpovídajících uzlových parametrů pro interpolaci na všech těchto sousedních útvarech. Útvar rozkladu, na něm definované uzlové parametry a příslušný interpolační polynom nazveme *konečným prvkem*. Často také, zejména tam, kde nebude hrozit nedorozumění, budeme termín prvek užívat pro samotný útvar rozkladu. Souhrn všech funkcí, které jsou polynomiální na jednotlivých prvcích rozkladu, mají v sousedních prvcích stejné uzlové parametry a jsou dostatečně hladké, tvoří při všech možných hodnotách uzlových parametrů konečnědimenzionální podprostor $\mathscr{D}_h^k$ Sobolevova prostoru $\mathscr{H}^k(\Omega)$. Prostor $\mathscr{D}_h^k$ budeme nazývat *prostorem konečných prvků* a jeho dimenze je zřejmě rovna počtu všech uzlových parametrů. Index h, kterým jsme opatřili jeho označení, udává závislost tohoto prostoru na velikosti jednotlivých prvků rozkladu použitých k jeho konstrukci a můžeme jej pokládat za rovný délce největší strany užitých prvků.

Při numerické realizaci Ritzovy-Galerkinovy metody je třeba sestrojit v prostoru $\mathscr{P}_h^k$ vhodnou bázi. Ve všech konkrétních případech prostorů konečných prvků, které v tomto článku zavedeme, je možno za bázi vzít množinu funkcí Φ_j, které sestrojíme tak, že j-tý uzlový parametr položíme rovný jedné a všechny ostatní rovny nule. Takto sestrojená bázová funkce bude nenulová jen na těch prvcích rozkladu, které obsahují j-tý uzlový parametr. Bude mít tedy malý nosič, takže jeden ze zformulovaných požadavků na prostor, v němž hledáme řešení, je splněn. Protože některé nebo všechny uzlové parametry budou mít, jak uvidíme později, význam hodnot aproximované funkce, bude při popsané konstrukci báze splněn i další požadavek, totiž, že řešením soustavy lineárních rovnic (3.26) už dostaneme přímo veličiny, které nás zajímají.

Tím, že jsme se omezili na trojúhelníkové a čtyřúhelníkové prvky, dopouštíme se při obecné křivočaré hranici vždy jakési chyby, neboť sjednocení všech prvků rozkladu tvoří polygon, který se nemůže přesně rovnat nepolygonální oblasti. Zjemněním rozkladu lze však danou oblast aproximovat s libovolnou přesností. Kromě toho se k lepšímu vystižení hranice oblasti užívají také křivočaré prvky (např. trojúhelníky, jejichž některé strany jsou parabolické oblouky apod.). Příslušné prostory konečných prvků se přitom konstruují na zcela stejných principech, jak bylo naznačeno.

V dalších několika odstavcích si všimneme některých nejdůležitějších speciálních případů konstrukce prostorů konečných prvků.

4.1 Trojúhelníkové prvky

V tomto odstavci popíšeme konstrukci takových prostorů konečných prvků, kdy základním prvkem rozkladu je trojúhelník.

4.1.1 Lineární Lagrangeův prvek

Tento konečný prvek je založen na myšlence Lagrangeovy interpolace a jeho konstrukce se opírá o následující větu:

Věta 4.1. *Buď K trojúhelník o vrcholech $P_j = (x_j, y_j)$, $j = 1,2,3$, a buďte u_j, $j = 1,2,3$ libovolná reálná čísla. Pak existuje právě jeden polynom $\Pi_1(x,y)$ prvního stupně, pro který platí*

$$\Pi_1(x_j, y_j) = u_j, \quad j = 1,2,3. \tag{4.2}$$

D ů k a z . Geometricky je tvrzení věty zřejmé, neboť každé tři body, které neleží v přímce, určují v trojrozměrném euklidovském prostoru právě jednu rovinu. Proveďme nicméně příslušný důkaz podrobně, abychom ukázali, jak postupovat později v komplikovanějších případech.

Hledaný polynom je tvaru

$$\Pi_1(x,y) = \alpha + \beta x + \gamma y, \tag{4.3}$$

kde α, β a γ jsou konstanty, které je třeba zvolit tak, aby platilo

$$\alpha + \beta x_j + \gamma y_j = u_j, \quad j = 1,2,3. \tag{4.4}$$

Tato soustava tří rovnic o třech neznámých má při libovolné pravé straně právě jedno řešení tehdy a jen tehdy, má-li příslušná homogenní soustava rovnic pouze triviální řešení, neboli, jinými slovy, je-li každý polynom prvního stupně, který se anuluje ve vrcholech trojúhelníku K, identicky roven nule. Uvažovaný polynom je polynomem prvního stupně na každé úsečce ležící v trojúhelníku K. Odtud především plyne, že se anuluje na každé straně tohoto trojúhelníku, neboť je podle předpokladu roven nule v jejích koncových bodech. Odtud však už bezprostředně plyne, že se anuluje i v každém dalším bodě v K. Abychom to dokázali, stačí vést takovým bodem libovolnou přímku a uvažovat zkoumaný polynom na ní. Věta je dokázána.

Často je výhodné psát polynom Π_1 ve tvaru

$$\Pi_1(x,y) = \sum_{j=1}^{3} u_j p_j(x,y), \tag{4.5}$$

kde p_j, $j = 1,2,3$ jsou polynomy, pro něž platí

$$p_i(x_j, y_j) = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \quad i,j = 1,2,3. \tag{4.6}$$

Z Cramerova pravidla ihned dostáváme, že je

$$\begin{aligned} p_1(x,y) &= \frac{1}{\det T}[(x_2y_3 - x_3y_2) - (y_3 - y_2)x + (x_3 - x_2)y], \\ p_2(x,y) &= \frac{1}{\det T}[(x_3y_1 - x_1y_3) - (y_1 - y_3)x + (x_1 - x_3)y], \\ p_3(x,y) &= \frac{1}{\det T}[(x_1y_2 - x_2y_1) - (y_2 - y_1)x + (x_2 - x_1)y], \end{aligned} \tag{4.7}$$

kde T je matice soustavy (4.4), tj.

$$T = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}. \tag{4.8}$$

Poznámka 4.1. Rovnice

$$p_j(x,y) = 0 \tag{4.9}$$

je rovnice přímky, v níž leží strana trojúhelníku K protější k vrcholu (x_j, y_j).

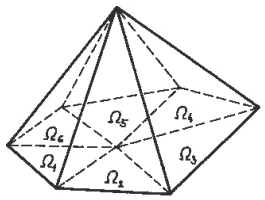
Tato poznámka je sice úplně elementární, bude však důležitá pro vyšetřování prostorů konečných prvků založených na polynomech vyšších stupňů než jedna.

Věta 4.1 dovoluje zavést Lagrangeův trojúhelníkový prvek jako polynom prvního stupně na příslušném trojúhelníku, který interpoluje funkční hodnoty v jeho

vrcholech. Uzlové parametry jsou tedy v tomto případě funkční hodnoty a uzly vrcholy trojúhelníků rozkladu. Prostor konečných prvků sestrojený pomocí popsaného prvku označíme symbolem $\mathscr{T}_h^1$. Každá funkce z tohoto prostoru je na každém trojúhelníku rozkladu lineární a protože v společných uzlech jsou uzlové parametry stejné, je také spojitá. Její parciální derivace jsou tedy po částech konstantní, z čehož plyne, že prostor $\mathscr{T}_h^1$ je podprostorem Sobolevova prostoru $\mathscr{H}^1$. Je proto vhodný k řešení problémů druhého řádu. Bázi v prostoru $\mathscr{T}_h^1$ vytvoříme postupem zmíněným výše, tj. za j-tou bázovou funkci bereme funkci Φ_j, která je v j-tém uzlu rovna jedné, ve všech ostatních uzlech rovna nule a na každém trojúhelníku je lineární. Každá bázová funkce je tedy v každém trojúhelníku rozkladu tvaru r-bokého jehlanu, kde číslo r udává počet trojúhelníků, které obsahují tento uzel (viz obr. 4.3).

Obr. 4.3

Bázové funkce v prostoru $\mathscr{T}_h^1$



Bázová funkce příslušná k uzlu, který leží na hranici oblasti, je přirozeně tvořena pouze částí tohoto jehlanu. Chceme-li sestrojit prostor konečných prvků, který je podprostorem prostoru $\mathscr{H}_0^1$ (a který je tedy vhodný k řešení problémů druhého řádu s homogenní Dirichletovou okrajovou podmínkou), vypustíme z popsané soustavy bázových funkcí ty funkce, které jsou rovny jedné v uzlu ležícím na hranici.

Z nerovnosti (4.1) plyne, že rychlost konvergence Ritzovy-Galerkinovy metody je dána přesností, se kterou je možno aproximovat přesné řešení prvkem užitého konečnědimenzionálního prostoru. Abychom byli schopni učinit si představu o rychlosti konvergence, musíme mít dispozici nějaké poznatky o aproximačních vlastnostech prostoru konečných prvků $\mathscr{T}_h^1$. V jednodimenzionálním případě jsme mohli každou funkci z prostoru $\mathscr{H}^1(a,b)$ jednoduše aproximovat po částech lineární funkcí pomocí interpolace. Zde je situace složitější, neboť obecná funkce z prostoru $\mathscr{H}^1(\Omega)$ nemusí být nejen spojitá, ale dokonce ani omezená (přesněji řečeno, není možné dosáhnout této vlastnosti změnou její definice na množině míry nula, neboť prvek prostoru $\mathscr{H}^1(\Omega)$ není jedna funkce, ale třída funkcí lišících se na množině míry nula), takže o její hodnotě nemá obecně smysl mluvit. Aproximaci dané funkce $f \in \mathscr{H}^1(\Omega)$ funkcí $f_h \in \mathscr{T}_h^1$ nelze tedy sestrojit přímo pomocí interpolace. Je však

možné postupovat např. tak, že se nejprve sestrojí funkce $\tilde{f} \in \mathscr{H}^2(\Omega)$, která funkci f aproximuje ve smyslu normy prostoru $\mathscr{H}^1$. To je možné, neboť množina $\mathscr{H}^2(\Omega)$ je hustá v prostoru $\mathscr{H}^1$. K funkci $\tilde{f}$ se pak sestrojí funkce f_h jako její interpolace, což má smysl, neboť funkce $\tilde{f}$ je už spojitá (tato skutečnost plyne z tzv. Sobolevových vět o vnoření, viz např. Nečas (1967)). Pokud o aproximované funkci nemáme žádné další informace než jen ty, že leží v prostoru $\mathscr{H}^1(\Omega)$, nelze o řádu chyby takto i jakkoliv jinak sestrojené aproximace nic říci. Aby bylo možné rozdíl $f - f_h$ nějakým použitelným způsobem odhadnout, je třeba zpřísnit požadavky na hladkost aproximované funkce. Jako příklad tvrzení tohoto typu uveďme následující větu.

Věta 4.2. *Nechť K je uzavřený trojúhelník o vrcholech $P_j = (x_j, y_j)$, $j = 1,2,3$. Nechť $u \in \mathscr{C}^2(K)$ a nechť Πu je polynom prvního stupně určený podmínkami $(\Pi u)(P_j) = u(P_j)$, $j = 1,2,3$. Pak platí*

$$\|u - \Pi u\|_{\mathscr{L}_\infty(K)} \leqq 2h_K^2 \max_{\alpha_1+\alpha_2=2} \left\| \frac{\partial^{\alpha_1+\alpha_2} u}{\partial x^{\alpha_1} \partial y^{\alpha_2}} \right\|_{\mathscr{L}_\infty(K)} \tag{4.10}$$

a

$$\max_{\alpha_1+\alpha_2=1} \left\| \frac{\partial^{\alpha_1+\alpha_2}(u - \Pi u)}{\partial x^{\alpha_1} \partial y^{\alpha_2}} \right\|_{\mathscr{L}_\infty(K)} \leqq 6 \frac{h_K^2}{\varrho_K} \max_{\alpha_1+\alpha_2=2} \left\| \frac{\partial^{\alpha_1+\alpha_2} u}{\partial x^{\alpha_1} \partial y^{\alpha_2}} \right\|_{\mathscr{L}_\infty(K)} \tag{4.11}$$

kde

$$\|u\|_{\mathscr{L}_\infty(K)} = \max_{(x,y)\in K} |u(x,y)| \tag{4.12}$$

a h_K je nejdelší strana a ϱ_K průměr kružnice vepsané trojúhelníku K.

D ů k a z . Zřejmě je

$$(\Pi u)(x,y) = \sum_{j=1}^{3} u(x_j, y_j) p_j(x,y), \tag{4.13}$$

kde p_j jsou polynomy (4.7). Podle Taylorova vzorce platí

$$u(\xi,\eta) = u(x,y) + \frac{\partial u}{\partial x}(x,y)(\xi - x) + \frac{\partial u}{\partial y}(x,y)(\eta - y) + R(x,y;\xi,\eta), \tag{4.14}$$

kde

$$\begin{aligned} R(x,y;\xi,\eta) = {} & \tfrac{1}{2}\frac{\partial^2 u}{\partial x^2}(\tilde{\xi},\tilde{\eta})(\xi - \eta)^2 + \\ & + \frac{\partial^2 u}{\partial x \partial y}(\tilde{\xi},\tilde{\eta})(\xi - x)(\eta - y) + \tfrac{1}{2}\frac{\partial^2 u}{\partial y^2}(\tilde{\xi},\tilde{\eta})(\eta - y)^2 \end{aligned} \tag{4.15}$$

a bod $(\tilde{\xi},\tilde{\eta})$ leží na úsečce spojující body (ξ,η) a (x,y). Položíme-li speciálně $(\xi,\eta) = (x_j, y_j)$, máme pro $j = 1,2,3$

$$u(x_j, y_j) = u(x,y) + q_j(x,y) + R_j(x,y), \tag{4.16}$$

kde

$$q_j(x,y) = \frac{\partial u}{\partial x}(x,y)(x_j - x) + \frac{\partial u}{\partial y}(x,y)(y_j - y) \tag{4.17}$$

a

$$R_j(x,y) = R(x,y;x_j,y_j). \tag{4.18}$$

Protože je $|x_j - x| \leqq h_K$ a $|y_j - y| \leqq h_K$, dostáváme pro funkci R_j odhad

$$|R_j(x,y)| \leqq 2h_K^2 \left\| \frac{\partial^{\alpha_1+\alpha_2} u}{\partial x^{\alpha_1} \partial y^{\alpha_2}} \right\|_{\mathscr{L}_\infty(K)} \tag{4.19}$$

Dosadíme-li do (4.13) ze (4.16), dostaneme

$$(\Pi u)(x,y) = u(x,y)\sum_{j=1}^{3} p_j(x,y) + \\ + \sum_{j=1}^{3} q_j(x,y)p_j(x,y) + \sum_{j=1}^{3} R_j(x,y)p_j(x,y). \tag{4.20}$$

Dokažme nyní, že je

$$\sum_{j=1}^{3} p_j(x,y) = 1 \tag{4.21}$$

a

$$\sum_{j=1}^{3} q_j(x,y)p_j(x,y) = 0. \tag{4.22}$$

Důkaz obou těchto rovností je založen na skutečnosti plynoucí ihned z věty 4.1, totiž, že je

$$u = \Pi u \tag{4.23}$$

pro každou funkci u, která je na K lineární. Dosadíme-li do (4.20) funkci $u(x,y) \equiv \equiv 1$, pro niž zřejmě platí (4.23), dostáváme ihned (4.21), neboť v tomto speciálním případě je $q_j(x,y) \equiv R_j(x,y) \equiv 0$. Abychom dokázali (4.22), položme ve (4.20) $u(x,y) = \gamma x + \delta y$, kde γ a δ jsou libovolná reálná čísla. Protože i v tomto případě platí (4.23) a protože je $R_j(x,y) = 0$, máme

$$u(x,y) = u(x,y) + \sum_{j=1}^{3} [\gamma(x_j - x) + \delta(y_j - y)]p_j(x,y). \tag{4.24}$$

Pro libovolná reálná γ a δ tedy platí

$$\sum_{j=1}^{3} [\gamma(x_j - x) + \delta(y_j - y)]p_j(x,y) = 0. \tag{4.25}$$

Položíme-li speciálně $\gamma = \partial u/\partial x$ a $\delta = \partial u/\partial y$, dostáváme odtud platnost (4.22).

Použijeme-li v (4.20) vztahy (4.21) a (4.22), je

$$(\Pi u)(x,y) = u(x,y) + \sum_{j=1}^{3} R_j(x,y)p_j(x,y) \tag{4.26}$$

neboli

$$u(x,y) - (\Pi u)(x,y) = -\sum_{j=1}^{3} R_j(x,y)p_j(x,y). \tag{4.27}$$

Protože pro $(x,y) \in K$ platí $0 \leqq p_j(x,y) \leqq 1$, je

$$|u(x,y) - (\Pi u)(x,y)| \leqq \sum_{j=1}^{3} |R_j(x,y)|p_j(x,y) \leqq \\ \leqq \max_j |R_j(x,y)| \sum_{j=1}^{3} p_j(x,y) = \max_j |R_j(x,y)|. \tag{4.28}$$

Odtud a z nerovnosti (4.19) už bezprostředně plyne nerovnost (4.10).

Abychom dokázali nerovnost (4.11), derivujme rovnici (4.13) podle x; dostaneme

$$\frac{\partial \Pi u}{\partial x} = \sum_{j=1}^{3} u(x_j,y_j)\frac{\partial p_j}{\partial x}(x,y). \tag{4.29}$$

Dosadíme-li do této rovnice z rovnice (4.16), máme

$$\frac{\partial \Pi u}{\partial x} = u(x,y)\sum_{j=1}^{3} \frac{\partial p_j}{\partial x}(x,y) + \sum_{j=1}^{3} q_j(x,y)\frac{\partial p_j}{\partial x}(x,y) + \\ + \sum_{j=1}^{3} R_j(x,y)\frac{\partial p_j}{\partial x}(x,y). \tag{4.30}$$

Protože však je

$$\sum_{j=1}^{3} \frac{\partial p_j}{\partial x}(x,y) = \frac{\partial}{\partial x}\sum_{j=1}^{3} p_j(x,y) = 0, \tag{4.31}$$

jak plyne z (4.21) a

$$\sum_{j=1}^{3} q_j(x,y)\frac{\partial p_j}{\partial x}(x,y) = \frac{\partial u}{\partial x}(x,y), \tag{4.32}$$

jak plyne z rovnice (4.30), položíme-li v ní $u(x,y) = \gamma x + \delta y$, plyne z (4.30), že

platí

$$\frac{\partial \Pi u}{\partial x}(x,y) = \frac{\partial u}{\partial x}(x,y) + \sum_{j=1}^{3} R_j(x,y)\frac{\partial p_j}{\partial x}(x,y) \tag{4.33}$$

neboli

$$\frac{\partial u}{\partial x}(x,y) - \frac{\partial \Pi u}{\partial x}(x,y) = -\sum_{j=1}^{3} R_j(x,y)\frac{\partial p_j}{\partial x}(x,y). \tag{4.34}$$

Protože platí

$$\max_{(x,y)\in K}\left|\frac{\partial p_j}{\partial x}(x,y)\right| \leqq \frac{1}{\varrho_K}, \tag{4.35}$$

jak se snadno zjistí, plyne z (4.34) a (4.19) odhad

$$\left|\frac{\partial u}{\partial x} - \frac{\partial \Pi u}{\partial x}(x,y)\right| \leqq 6\frac{h_K^2}{\varrho_K}\max_{\alpha_1+\alpha_2=2}\left\|\frac{\partial^{\alpha_1+\alpha_2}u}{\partial x^{\alpha_1}\partial y^{\alpha_2}}\right\|_{\mathscr{L}_\infty(K)} \tag{4.36}$$

Úplně stejně odhadneme rozdíl $\partial u/\partial y - \partial(\Pi u)/\partial y$. Platí tedy (4.11) a důkaz věty 4.2 je zakončen.

Ve větě 4.2 jsme tedy nalezli tvrzení, které nám dává představu o velikosti chyby interpolace v prostoru $\mathscr{T}_h^1$. Protože však v nerovnostech (4.10) a (4.11) vystupuje $\mathscr{L}_\infty$-norma, nejsou tyto odhady ideální ke konstrukci odhadu chyby Ritzovy-Galerkinovy metody na základě nerovnosti (4.1), neboť v ní se užívá norma typu $\mathscr{L}_2$-normy. Zformulujme proto větu, která je paralelní k větě 4.2, v níž však jsou příslušné veličiny měřeny integrálními normami.

Věta 4.3. *Nechť jsou splněny předpoklady věty 4.2. Pak existuje absolutní konstanta C taková, že platí*

$$\|u - \Pi u\|_{\mathscr{L}_2(K)} \leqq Ch_K^2|v|_{\mathscr{H}^2(K)} \tag{4.37}$$

a

$$|u - \Pi u|_{\mathscr{H}^1(K)} \leqq C\frac{h_K^2}{\varrho_K}|v|_{\mathscr{H}^2(K)}, \tag{4.38}$$

kde

$$|u|_{\mathscr{H}^r(K)} = \sum_{\alpha_1+\alpha_2=r}\left(\int_K\left|\frac{\partial^{\alpha_1+\alpha_2}u}{\partial x^{\alpha_1}\partial y^{\alpha_2}}\right|^2 \mathrm{d}x\,\mathrm{d}y\right)^{1/2}, \tag{4.39}$$

a je to tedy r-tá seminorma funkce u.

Důkaz této věty, který už není tak elementární, jako tomu bylo u věty 4.2, nalezne čtenář v článku Dupontově a Scottově (1980).

Jak už jsme řekli, ke konstrukci odhadů chyby interpolace a tedy také chyby Ritzovy-Galerkinovy aproximace řešení okrajové úlohy se častěji užívá věty 4.3 než věty 4.2. Abychom se však vyhnuli značným technickým komplikacím, uvedli jsme zde pouze důkaz věty 4.2, neboť se domníváme, že je dostatečně charakteristický k ilustraci postupů v obdobných situacích.

Užijme ještě větu 4.3 k odhadu chyby interpolace na celé oblasti Ω. Předpokládejme, že je dána posloupnost rozkladů $\{Z_h\}$ dané oblasti charakterizovaných parametrem

$$h = \max_{K\in Z_h} h_K \tag{4.40}$$

a předpokládejme dále, že pro každý trojúhelník K rozkladu platí

$$\frac{\varrho_K}{h_K} \geqq \beta, \tag{4.41}$$

kde β je kladná konstanta nezávislá na h. Takovou posloupnost rozkladů obvykle nazýváme *regulární posloupností rozkladů*. Požadavek značí, že v rozkladu Z_h nejsou přípustné trojúhelníky, které mají některý úhel příliš malý.

Buď nyní $f \in \mathscr{H}^2(\Omega)$ a buď f_h její interpolace. Sečtení nerovností (4.37) pro všechny trojúhelníky rozkladu dává

$$\begin{aligned}\|f - f_h\|_{\mathscr{L}_2(\Omega)}^2 &= \sum_{K\in Z_h}\|f - f_h\|_{\mathscr{L}_2(K)}^2 \leqq \\ &\leqq \sum_{K\in Z_h} C^2h_K^4|f|_{\mathscr{H}^2(K)}^2 \leqq C^2h^4\sum_{K\in Z_h}|f|_{\mathscr{H}^2(K)}^2 = \\ &= C^2h^4|f|_{\mathscr{H}^2(\Omega)}^2.\end{aligned} \tag{4.42}$$

Analogicky plyne z nerovnosti (4.38), použijeme-li zároveň (4.41), že je

$$\begin{aligned}|f - f_h|_{\mathscr{H}^1(\Omega)}^2 &\leqq \sum_{K\in Z_h} C^2\frac{h_K^4}{\varrho_K^2}|f|_{\mathscr{H}^2(K)}^2 \leqq \\ &\leqq \sum_{K\in Z_h}\frac{C^2h_K^2}{\beta^2}|f|_{\mathscr{H}^2(K)}^2 \leqq \frac{C^2h^2}{\beta^2}|f|_{\mathscr{H}^2(\Omega)}^2\end{aligned} \tag{4.43}$$

Spojením nerovností (4.42) a (4.43) dostáváme, že pro $h \leqq h_0$ platí

$$\|f - f_h\|_{\mathscr{H}^1(\Omega)} \leqq Mh|f|_{\mathscr{H}^2(\Omega)}, \tag{4.44}$$

kde M je konstanta daná vzorcem

$$M = \left(C^2h_0^2 + \frac{C^2}{\beta^2}\right)^{1/2}, \tag{4.45}$$

takže nezávisí na h. Pro $\mathscr{L}_2$-normu rozdílu $f - f_h$ samozřejmě platí lepší odhad

$$\|f - f_h\|_{\mathscr{L}_2(\Omega)} \leqq Ch^2|f|_{\mathscr{H}^2(\Omega)}, \tag{4.46}$$

jak plyne ihned ze vzorce (4.42).

Rychlost konvergence Ritzovy-Galerkinovy metody je zde tedy $O(h^2)$ měřeno normou prostoru $\mathscr{L}_2$ a $O(h)$ měřeno normou prostoru $\mathscr{H}^1$.

Poznamenejme konečně, že provedeme-li triangulaci dané oblasti pravidelně (je-li např. Ω čtverec, sestrojíme v něm nejprve pravidelnou čtvercovou síť a vzniklé čtverce rozdělíme navzájem rovnoběžnými úhlopříčkami na trojúhelníky) a k přibližnému výpočtu integrálů, které je třeba počítat při sestavování Gramovy matice a vektoru pravé strany, užijeme vhodné kvadraturní vzorce, dostaneme metodu sítí.

4.1.2 Kvadratický Lagrangeův prvek

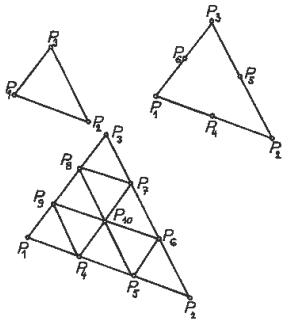
Sestrojme nyní prostor konečných prvků založený na interpolaci polynomy druhého stupně. Dokážeme proto nejprve následující větu.

Věta 4.4. *Buď K trojúhelník o vrcholech $P_j = (x_j, y_j)$, $j = 1, 2, 3$, označme $P_4 = (x_4, y_4)$, $P_5 = (x_5, y_5)$ a $P_6 = (x_6, y_6)$ středy jeho stran (viz obr. 4.4) a buďte u_j, $j = 1, \dots, 6$, libovolná reálná čísla. Pak existuje právě jeden polynom Π_2 druhého stupně, pro který platí*

$$\Pi_2(x_j, y_j) = u_j, \quad j = 1, \dots, 6. \tag{4.47}$$

Obr. 4.4

Lineární, kvadratický a kubický Lagrangeův prvek



D ů k a z . Zřejmě stačí dokázat, že jediný polynom druhého stupně, který se anuluje v bodech P_j, je nulový polynom. Buď tedy u polynom druhého stupně, pro který platí

$$u(x_j, y_j) = 0, \quad j = 1, \dots, 6, \tag{4.48}$$

a dokažme, že tento polynom je nulový. Uvažujme k tomu cíli stranu P_2P_3 daného trojúhelníka. Podél této strany je funkce u zřejmě polynom druhého stupně v jedné proměnné, který je roven nule ve třech navzájem různých bodech; odtud ihned plyne, že polynom u je na úsečce P_2P_3 identicky roven nule. Podle poznámky 4.1 je však rovnice $p_1(x, y) = 0$, kde funkce p_1 je definovaná první rovnicí (4.7), rovnicí přímky P_2P_3. Musí tedy platit

$$u(x, y) = p_1(x, y) u_1(x, y), \tag{4.49}$$

kde u_1 je polynom stupně nejvýše jedna. Stejným způsobem dokážeme, že funkce u se anuluje na úsečce P_1P_3. Pro každý bod $(x, y) \in K$ tedy platí

$$u(x, y) = p_1(x, y) p_2(x, y) u_0, \tag{4.50}$$

kde u_0 je nyní polynom stupně nula, tj. konstanta. Dosadíme-li však do rovnice (4.50) za (x, y) bod P_4 ležící mezi body P_1 a P_2, dostaneme

$$0 = \tfrac{1}{2}\tfrac{1}{2} u_0, \tag{4.51}$$

což dokazuje, že je $u_0 = 0$. Nulovost polynomu u pak už plyne z (4.50). Tím je zakončen důkaz věty.

Na základě věty 4.4 lze zvolit vrcholy trojúhelníka a středy jeho stran za uzly a funkční hodnoty za uzlové parametry a sestrojit tak jejich interpolací kvadratický Lagrangeův prvek.

Polynom Π_2 splňující podmínky (4.47) lze psát ve tvaru

$$\Pi_2(x, y) = \sum_{j=1}^{6} u_j p_j^{(2)}(x, y), \tag{4.52}$$

kde elementární polynomy $p_j^{(2)}$ jsou podobně jako v případě lineárních prvků definovány rovnicemi.

$$p_j^{(2)}(x_k, y_k) = \begin{cases} 1, & j = k, \\ 0, & j \neq k, \end{cases} \qquad j, k = 1, \dots, 6. \tag{4.53}$$

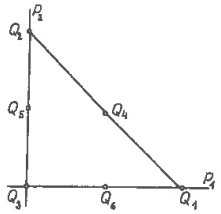
Abychom odvodili konkrétní tvar těchto polynomů, užijeme pojmu tzv. *referenčního trojúhelníku.* Zavedení tohoto trojúhelníku vychází z poznámky 4.1. Protože pro funkce p_1, p_2 a p_3 vystupující v této poznámce platí rovnice (4.21), převádí transformace

$$\begin{aligned} p_1 &= p_1(x, y), \\ p_2 &= p_2(x, y) \end{aligned} \tag{4.54}$$

trojúhelník $P_1P_2P_3$ roviny (x, y) na trojúhelník $Q_1Q_2Q_3$ roviny (p_1, p_2), přičemž je $Q_1 = (1, 0)$, $Q_2 = (0, 1)$ a $Q_3 = (0, 0)$ (viz obr. 4.5). Trojúhelník $Q_1Q_2Q_3$ se nazývá jednotkový referenční trojúhelník nebo stručně referenční trojúhelník.

Obr. 4.5

Referenční trojúhelník



Protože transformace (4.54) je lineární, je snadné vypočítat inverzní transformaci z roviny (p_1, p_2) na rovinu (x, y). Tato transformace je samozřejmě opět lineární a je tvaru

$$x = x(p_1, p_2), \qquad y = y(p_1, p_2), \tag{4.55}$$

kde

$$\begin{aligned} x(p_1, p_2) &= x_3 + (x_1 - x_3)p_1 + (x_2 - x_3)p_2, \\ y(p_1, p_2) &= y_3 + (y_1 - y_3)p_1 + (y_2 - y_3)p_2. \end{aligned} \tag{4.56}$$

Transformace (4.54) a (4.55) umožňují podle potřeby přecházet snadno od referenčního trojúhelníka k obecnému a naopak.

Protože transformace (4.55) je lineární, je každý polynom druhého stupně v proměnných x, y polynomem druhého stupně v proměnných p_1, p_2. Elementární polynom $p_1^{(2)}$ nabývá v uzlu P_1 hodnotu jedna a v ostatních uzlech hodnotu 0. Protože přímka $p_1 = 0$ prochází uzly P_2, P_3 a P_5, a přímka $p_1 = 1/2$ uzly P_4 a P_6, nesmí polynom $p_1^{(2)}$ záviset na proměnné p_2 a musí být roven nule pro $p_1 = 0$ a $p_1 = 1/2$ (a samozřejmě jedné pro $p_1 = 1$). Jediný polynom druhého stupně, který splňuje tyto podmínky, je polynom $p_1(2p_1 - 1)$. Polynom $p_1^{(2)}$ je tedy roven výrazu $p_2(2p_1 - 1)$, kde za p_1 je třeba dosadit ze vzorců (4.7) (nebo (4.54)). Polynomy $p_2^{(2)}$ a $p_3^{(2)}$ sestrojíme na základě úplně stejných úvah. Elementární polynom $p_4^{(2)}$, který je roven jedné v uzlu P_4 a v ostatních uzlech je roven nule, sestrojíme pomocí přímek $p_1 = 0$ a $p_2 = 0$, které procházejí body P_2, P_3 a P_5 a body P_1, P_3 a P_6. Polynom $p_4^{(2)}$ musí být tedy tvaru $\alpha p_1 p_2$, přičemž konstanta α se určí z podmínky $p_4^{(2)}(x_4, y_4) = 1$, která je splněna pro $\alpha = 4$. Elementární polynomy $p_j^{(2)}$ jsou tedy dány vzorci

$$\begin{aligned} p_1^{(2)} &= p_1(2p_1 - 1), \\ p_2^{(2)} &= p_2(2p_2 - 1), \end{aligned} \tag{4.57}$$

$$\begin{aligned} p_3^{(2)} &= p_3(2p_3 - 1), \\ p_4^{(2)} &= 4p_1p_2, \\ p_5^{(2)} &= 4p_2p_3, \\ p_6^{(2)} &= 4p_1p_3, \end{aligned}$$

přičemž p_1, p_2 a p_3 jsou lineární polynomy definované rovnicemi (4.7).

Postup, pomocí něhož jsme odvodili tvar polynomů majících vlastnosti (4.53), demonstruje, že užití referenčního trojúhelníku může podstatně zjednodušit příslušné úvahy. Možnost práce s referenčním trojúhelníkem, eventuálně s jiným referenčním obrazcem, jsou-li užity jiné základní obrazce než trojúhelníky, patří rovněž k charakteristickým rysům metody konečných prvků.

Prostor konečných prvků sestrojený pomocí právě popsaného kvadratického prvku označíme symbolem $\mathscr{T}_h^2$. Vytvoříme-li v tomto prostoru bázi jako v předchozím odstavci, tj. za j-tou bázovou funkci vezmeme funkci, která je po částech kvadratická, která je v j-tém uzlu rovna jedné a v ostatních uzlech rovna nule, je tato bázová funkce různá od nuly pouze v trojúhelnících obsahujících zmíněný uzel, a má tedy malý nosič. Na každém trojúhelníku, kde je nenulová, je přitom rovna některé elementární funkci (4.57). Funkce z prostoru $\mathscr{T}_h^2$ jsou zřejmě spojité, mají tedy po částech spojité první parciální derivace a vytvářejí tak podprostor Sobolevova prostoru $\mathscr{H}^1$. Jsou tedy stejně jako lineární prvky vhodné k řešení úloh druhého řádu. O konstrukci aproximace funkce z $\mathscr{H}^1$ funkcí z $\mathscr{T}_h^2$ platí totéž, co bylo řečeno v odst. 4.1.1 o aproximaci funkcemi z $\mathscr{T}_h^1$. Uveďme ještě pro úplnost bez důkazu, že je-li posloupnost triangulací regulární, je-li $f \in \mathscr{H}^r(\Omega)$ pro $r = 2$ nebo 3 a je-li f_h interpolace funkce f, platí pro rozdíl $f - f_h$ odhad

$$\|f - f_h\|_{\mathscr{H}^p(\Omega)} \leqq M h^{r-p} |f|_{\mathscr{H}^r(\Omega)}, \quad p = 0, 1, \tag{4.58}$$

kde $\mathscr{H}^0(\Omega) = \mathscr{L}_2(\Omega)$. Přitom předpoklad, že je $f \in \mathscr{H}^r(\Omega)$ s r větším než 3, už nevede k zlepšení uvedeného odhadu.

Užití kvadratických Lagrangeových prvků vede tedy při řešení rovnice druhého řádu k chybě řádu $O(h^2)$ měřeno normou Sobolevova prostoru $\mathscr{H}^1(\Omega)$, je-li hledané řešení dostatečně hladké.

4.1.3 Kubický Lagrangeův prvek

Abychom sestrojili tento prvek, vezmeme za uzly vrcholy trojúhelníka, body, které dělí jeho strany na tři stejné díly, a těžiště (viz obr. 4.4), za uzlové parametry funkční hodnoty v těchto uzlech a interpolujeme je polynomem třetího stupně. Označme užité uzly $P_1 = (x_1, y_1), \ldots, P_{10} = (x_{10}, y_{10})$. Následující věta ospravedlňuje popsaný postup.

Věta 4.5. *Buď K trojúhelník $P_1P_2P_3$ a buďte u_j, $j = 1, \ldots, 10$, libovolná reálná*

čísla. Pak existuje právě jeden polynom Π_3 třetího stupně, pro který platí

$$\Pi_3(x_j, y_j) = u_j, \quad j = 1, \ldots, 10. \tag{4.59}$$

D ů k a z . Protože podmínek (4.59) je právě tolik, kolik má koeficientů polynom třetího stupně, stačí opět jako ve větě 4.1 a 4.4 dokázat, že polynom u třetího stupně, pro který platí

$$u(x_j, y_j) = 0, \quad j = 1, \ldots, 10, \tag{4.60}$$

je nutně nulový polynom. Polynom u je identicky roven nule podél každé strany trojúhelníku K, neboť je na těchto stranách polynomem třetího stupně v jedné proměnné, který se anuluje ve čtyřech navzájem různých bodech. Polynom u lze tedy psát ve tvaru

$$u(x, y) = \gamma p_1(x, y) p_2(x, y) p_3(x, y), \tag{4.61}$$

kde polynomy p_1, p_2 a p_3 jsou opět definovány rovnicemi (4.7) a γ je konstanta. Dosadíme-li nyní do (4.61) za (x, y) těžiště trojúhelníka, dostáváme z (4.60), že je

$$0 = \gamma \tfrac{1}{3}\tfrac{1}{3}\tfrac{1}{3}. \tag{4.62}$$

Je tedy $\gamma = 0$ a u je podle (4.61) nulový polynom, což dokazuje větu.

Polynom Π_3 lze podobně jako v případě Lagrangeova kvadratického prvku psát ve tvaru

$$\Pi_3(x, y) = \sum_{j=1}^{10} u_j p_j^{(3)}(x, y), \tag{4.63}$$

kde polynomy $p_j^{(3)}$ jsou definovány rovnicemi

$$p_j^{(3)}(x_k, y_k) = \begin{cases} 1, & j = k, \\ & j, k = 1, \ldots, 10. \\ 0, & j \neq k, \end{cases} \tag{4.64}$$

Elementární polynomy $p_j^{(3)}$ se přitom dají vyjádřit pomocí elementárních polynomů lineární interpolace.

Prostor konečných prvků vytvořený pomocí polynomů Π_3 označíme $\mathscr{T}_h^3$. Je opět podprostorem Sobolevova prostoru $\mathscr{H}^1$. Vytvoříme-li v něm bázi standardním postupem popsaným výše, je složena z funkcí, které jsou na jednotlivých trojúhelnících rovny některé elementární funkci $p_j^{(3)}$ a které mají malý nosič. Označíme-li opět f_h interpolaci funkce $f \in \mathscr{H}^r$, kde za číslo r připouštíme nyní hodnoty 2, 3 a 4, platí pro chybu nerovnost (4.58). Tento odhad už nelze dále zlepšit zvětšováním čísla r, tj. dalším zvyšováním hladkosti interpolované funkce. Měříme-li tedy velikost chyby normou Sobolevova prostoru $\mathscr{H}^1(\Omega)$, vede Ritzova-Galerkinova metoda při použití prostoru $\mathscr{T}_h^3$ k rychlosti konvergence řádu $O(h^3)$, samozřejmě stejně jako dříve za předpokladu dostatečné hladkosti hledaného řešení.

4.1.4 Obecný Lagrangeův prvek

V tomto odstavci popíšeme stručně bez jakýchkoliv důkazů, jak lze sestrojit obecný Lagrangeův prvek založený na interpolaci obecným polynomem stupně m. Polynom Π_m stupně m má $s_m = (m+1)(m+2)/2$ koeficientů, a lze jej tedy užít k interpolaci v s_m uzlech umístěných v uvažovaném trojúhelníku. Uspořádáme-li jednotlivé sčítance v obecném polynomu stupně m do schématu znázorněného v tab. 4.1, máme v něm návod, jak potřebné uzly rozmístit symetricky: Každou stranu trojúhelníku $P_1P_2P_3$ rozdělíme na m stejných dílů a takto vzniklé dělicí body spojíme přímkami rovnoběžnými se stranami trojúhelníku. Tímto postupem rozdělíme daný trojúhelník na m^2 podobných trojúhelníků, jejichž vrcholy, kterých je právě s_m, vezmeme za uzly. Popsaným návodem jsme se ostatně už řídili ve speciálních případech uvedených v odst. 4.1.1 až 4.1.3.

Tabulka 4.1

Schéma obecného polynomu ve dvou proměnných

				1						
				x		y				
			x^2		xy		y^2			
		x^3		x^2y		xy^2		y^3		
	x^4		x^3y		x^2y^2		xy^3		y^4	
...		...		...		...		...		...

Označíme-li u_j, $j = 1, \ldots, s_m$, hodnoty interpolované funkce v právě sestrojených uzlech, lze polynom Π_m psát ve tvaru

$$\Pi_m(x, y) = \sum_{j=1}^{s_m} u_j p_j^{(m)}(x, y), \tag{4.65}$$

kde každá funkce $p_j^{(m)}$ je polynom stupně m, který nabývá v právě jednom uzlu hodnoty jedna a v ostatních uzlech hodnot nula. Každý z těchto elementárních polynomů $p_j^{(m)}$ lze opět složit z lineárních polynomů p_1, p_2 a p_3.

Pro příslušný prostor konečných prvků $\mathscr{T}_h^m$ platí $\mathscr{T}_h^m \subset \mathscr{H}^1$ a jeho aproximační vlastnosti jsou popsány opět nerovností (4.58), v níž předpokládáme $f \in \mathscr{H}^r$ a za r připouštíme hodnoty $2, \ldots, m+1$. Bázové funkce v prostoru $\mathscr{T}_h^m$ mají opět malé nosiče a na jednotlivých trojúhelnících rozkladu jsou rovny elementárním polynomům $p_j^{(m)}$. Prostor $\mathscr{T}_h^m$ je možno užít opět pouze k řešení problémů druhého řádu, neboť není podprostorem žádného Sobolevova prostoru $\mathscr{H}^k$ s $k > 1$. Maximální dosažitelná rychlost konvergence je $O(h^m)$.

4.1.5 Hermitův prvek

Jako alternativu k interpolaci funkčních hodnot v mnoha uzlech je možné uvažovat interpolaci hodnot dané funkce a hodnot některých jejích derivací, tj. interpolaci Hermitova typu, v menším počtu uzlů. Jedna z možností tohoto postupu v případě polynomů třetího stupně je založena na následující větě.

Věta 4.6. *Polynom třetího stupně je jednoznačně určen svými funkčními hodnotami a hodnotami prvních parciálních derivací podle x a y ve vrcholech trojúhelníku $P_1P_2P_3$ a funkční hodnotou v jeho těžišti.*

D ů k a z . Protože i v tomto případě je počet podmínek roven počtu koeficientů uvažovaného polynomu, stačí dokázat, že polynom u, pro který platí

$$u(x_j, y_j) = \frac{\partial u}{\partial x}(x_j, y_j) = \frac{\partial u}{\partial y}(x_j, y_j) = 0, \quad j = 1, 2, 3, \tag{4.66}$$
$$u(x_4, y_4) = 0,$$

kde (x_j, y_j), $j = 1, 2, 3$, jsou souřadnice vrcholů trojúhelníku $P_1P_2P_3$ a $P_4 = (x_4, y_4)$ je jeho těžiště, je nulový polynom. Z rovnic (4.66) plyne, že je

$$\frac{\partial u}{\partial s}(x_j, y_j) = \frac{\partial u}{\partial x}(x_j, y_j)s_1 + \frac{\partial u}{\partial y}(x_j, y_j)s_2 = 0, \tag{4.67}$$

kde $\partial/\partial s$ je derivace ve směru $s = (s_1, s_2)$. Speciálně je tedy

$$\frac{\partial u}{\partial s}(x_2, y_2) = \frac{\partial u}{\partial s}(x_3, y_3) = 0, \tag{4.68}$$

kde s je směr od bodu P_2 do bodu P_3. Protože je také $u(x_2, y_2) = u(x_3, y_3) = 0$ a protože funkce u je podél úsečky P_2P_3 polynomem třetího stupně, je funkce u rovna na této úsečce identicky nule. Podobně se dokáže, že u se anuluje i na zbývajících dvou stranách trojúhelníka. Důkaz věty se nyní už zakončí za užití podmínky $u(x_4, y_4) = 0$ stejně jako důkaz věty 4.5.

Kubický Hermitův prvek dostaneme tak, že za uzly vezmeme vrcholy P_1, P_2 a P_3 daného trojúhelníka a jeho těžiště P_4 a za uzlové parametry funkční hodnoty v uzlech P_1 až P_4 a hodnoty prvních parciálních derivací podle x a y v uzlech P_1 až P_3.

Prostor konečných prvků vytvořený právě popsaným kubickým prvkem označíme $\mathscr{S}_h^3$. Vytvoříme-li jeho bázi naším standardním postupem, tj. položíme-li vždy právě jeden uzlový parametr rovný jedné a ostatní parametry rovny nule, mají bázové funkce opět malé nosiče. Prostor $\mathscr{S}_h^3$ je stejně jako prostor $\mathscr{T}_h^3$ pouze podprostorem Sobolevova prostoru $\mathscr{H}^1$, takže je vhodný pouze k řešení problémů druhého řádu. Jeho aproximační vlastnosti jsou popsány nerovností (4.58) s $r = 3$ a 4, a jsou tedy stejné jako aproximační vlastnosti prostoru $\mathscr{T}_h^3$. Za číslo r v nerovnosti (4.58) zde však nepřipouštíme hodnotu 2, jako tomu bylo v případě prostoru $\mathscr{T}_h^3$. Důvod je ten, že ke konstrukci interpolace v prostoru $\mathscr{S}_h^3$ nestačí užít pouze

hodnoty interpolované funkce, ale jsou zapotřebí hodnoty jejích derivací, a ty pro funkci z prostoru $\mathscr{H}^2$ nemusí mít smysl.

Podobně jako jsme vytvořili kubický Hermitův prvek a prostor konečných prvků $\mathscr{S}_h^3$, dají se vytvořit prostory $\mathscr{S}_h^{2\nu+1}$ pro libovolné $\nu = 1, 2, \ldots$. Stačí k tomu vzít za uzly stejně jako v kubickém případě vrcholy daného trojúhelníku a jeho těžiště a za uzlové parametry hodnoty interpolované funkce a jejích všech parciálních derivací až do řádu ν včetně ve vrcholech trojúhelníku a hodnoty interpolované funkce a jejích parciálních derivací až do řádu $\nu - 1$ včetně v těžišti. Prostor $\mathscr{S}_h^{2\nu+1}$ je stále vhodný pouze k řešení rovnic druhého řádu, neboť je podprostorem pouze prostoru $\mathscr{H}^1$ a není podprostorem žádného Sobolevova prostoru $\mathscr{H}^k$ pro $k > 1$. Lze však pomocí něj, stejně jako pomocí obecného Lagrangeova prvku, docílit rychlosti konvergence Ritzovy-Galerkinovy metody libovolně vysokého řádu, neboť jeho aproximační vlastnosti jsou dány nerovností (4.58) s $r = \nu + 2, \ldots, 2\nu + 2$.

4.1.6 Prostory konečných prvků pro řešení diferenciálních rovnic čtvrtého řádu

Prostory konečných prvků, které jsme až dosud sestrojili, obsahují funkce, jejichž první parciální derivace mohou být podél stran trojúhelníků rozkladu nespojité. Takové funkce proto patří pouze do Sobolevova prostoru $\mathscr{H}^1(\Omega)$. Abychom sestrojili prostory, které jsou podprostory Sobolevova prostoru $\mathscr{H}^2(\Omega)$ a které jsou tak vhodné k řešení problémů čtvrtého řádu, je třeba, aby byly spojité nejen příslušné funkce samotné, ale i jejich první parciální derivace. Nejjednodušší prvek, pomocí něhož to lze dosáhnout, je založen na interpolaci polynomy pátého stupně. Při volbě uzlů a uzlových parametrů, kterých je zapotřebí 21, je však třeba postupovat jinak, než jsme postupovali při konstrukci prostoru $\mathscr{S}_h^5$ založeného také na polynomech pátého stupně. Požadované spojitosti prvních parciálních derivací se dosáhne, jak hned uvidíme, tak, že za uzly se zvolí vrcholy P_1, P_2 a P_3 daného trojúhelníku a středy jeho stran P_4, P_5 a P_6 a za uzlové parametry se berou funkční hodnoty prvních a druhých parciálních derivací ve vrcholech a hodnoty derivace podle normály ve středu stran. Nejprve však ukážeme, že polynom pátého stupně je zvolenými uzlovými parametry skutečně jednoznačně určen.

Věta 4.7. *Polynom Π_5 pátého stupně je jednoznačně určen svými hodnotami a hodnotami parciálních derivací až do řádu dvě ve vrcholech trojúhelníka a hodnotami derivací podle normály ve středech jeho stran.*

D ů k a z . Zřejmě stačí podobně jako při důkazu např. věty 4.6 dokázat, že polynom u pátého stupně, pro který platí

$$\frac{\partial^{\alpha_1+\alpha_2} u}{\partial x^{\alpha_1} \partial y^{\alpha_2}}(x_j, y_j) = 0, \quad \alpha_1 + \alpha_2 \leqq 2, \ \alpha_1 \geqq 0, \ \alpha_2 \geqq 0, \ j = 1, 2, 3, \tag{4.69}$$
$$\frac{\partial u}{\partial n}(x_j, y_j) = 0, \quad j = 4, 5, 6,$$

je nutně nulový polynom (je samozřejmě $P_j = (x_j, y_j)$, $j = 1, \dots, 6$, a body P_j jsou rozmístěny jako na obr. 4.4). Abychom to nahlédli, všimneme si především, že pro $j = 2, 3$ platí

$$u(x_j, y_j) = \frac{\partial u}{\partial s}(x_j, y_j) = \frac{\partial^2 u}{\partial s^2}(x_j, y_j) = 0, \tag{4.70}$$

kde s značí směr úsečky P_2P_3. Protože funkce u je na úsečce P_2P_3 polynom stupně nejvýše pět, plyne z podmínek (4.70), že je na ní identicky roven nule. Protože funkce $\partial u/\partial n$ je na této úsečce polynom stupně nejvýše čtyři a protože pro $j = 2, 3$ platí

$$\frac{\partial u}{\partial n}(x_j, y_j) = \frac{\partial}{\partial s}\left(\frac{\partial u}{\partial n}\right)(x_j, y_j) = 0 \tag{4.71}$$

a samozřejmě také

$$\frac{\partial u}{\partial n}(x_5, y_5) = 0, \tag{4.72}$$

je funkce $\partial u/\partial n$ identicky rovna nule na úsečce P_2P_3. Protože tedy obě funkce u a $\partial u/\partial n$ se anulují na P_2P_3, dá se polynom u psát ve tvaru

$$u(x, y) = [p_1(x, y)]^2 u_3(x, y), \tag{4.73}$$

kde u_3 je polynom nejvýše třetího stupně. Poslední tvrzení platí ovšem i pro další dvě strany trojúhelníku $P_1P_2P_3$; platí tedy

$$u(x, y) = \gamma[p_1(x, y)]^2[p_2(x, y)]^2[p_3(x, y)]^2. \tag{4.74}$$

Protože však u je polynomem stupně nejvýše pět, je nutně $\gamma = 0$. Je tedy $u \equiv 0$ a věta je dokázána.

Buďte nyní K_1 a K_2 dva trojúhelníky se společnou stranou S, buďte u_1 a u_2 polynomy pátého stupně definované na K_1, resp. K_2 a předpokládejme, že platí

$$\begin{aligned} &\frac{\partial^{\alpha_1+\alpha_2} u_1}{\partial x^{\alpha_1} \partial y^{\alpha_2}} = \frac{\partial^{\alpha_1+\alpha_2} u_2}{\partial x^{\alpha_1} \partial y^{\alpha_2}}, \text{ v koncových bodech } S, \\ &\frac{\partial u_1}{\partial n} = \frac{\partial u_2}{\partial n} \text{ ve středu úsečky } S, \end{aligned} \tag{4.75}$$

kde n je směr normály k úsečce S. Pro rozdíl $w = u_1 - u_2$ tedy platí (4.70), (4.71) a (4.72), a je tedy

$$w = \frac{\partial w}{\partial n} = 0 \text{ na } S. \tag{4.76}$$

Je-li však $w = 0$ na S, je také

$$\frac{\partial w}{\partial s} = 0 \text{ na } S \tag{4.77}$$

(s je směr úsečky S). Relace (4.76) a (4.77) však ukazují, že funkce u, která je definovaná na $K_1 \cup K_2$ tak, že je rovna funkci u_1 na K_1 a funkci u_2 na K_2, je spojitá spolu se svými prvními derivacemi podél úsečky S.

Prostor konečných prvků vytvořený pomocí právě sestrojeného prvku, označme jej $\mathscr{R}_h^5$, je tedy skutečně podprostorem Sobolevova prostoru $\mathscr{H}^2(\Omega)$. Je-li f_h interpolace funkce $f \in \mathscr{H}^r(\Omega)$, platí pro rozdíl $f - f_h$ opět nerovnost (4.58), kde za p připouštíme nyní hodnoty 0, 1 a 2.

Na podobném principu lze sestrojit prostor konečných prvků $\mathscr{R}_h^{4\nu+1}$, který je podprostorem Sobolevova prostoru $\mathscr{H}^{\nu+1}(\Omega)$, je tvořen po částech polynomy stupně $4\nu + 1$ a je vhodný k řešení úloh řádu $2\nu + 2$. Za uzlové parametry je třeba vzít hodnoty interpolované funkce a jejích derivací až do řádu 2ν ve vrcholech trojúhelníku, funkční hodnoty a hodnoty derivací až do řádu $\nu - 2$ v těžišti a hodnoty první derivace podle normály ve středu každé strany, hodnoty druhé derivace podle normály v bodech dělících každou stranu na tři stejné díly atd., až hodnoty ν-té derivace podle normály v bodech dělících každou stranu na $\nu + 1$ stejných dílů. Za předpokladu, že pro aproximovanou funkci platí $f \in \mathscr{H}^r(\Omega)$ pro $r = 2\nu + 2, \dots, 4\nu + 2$, je chyba aproximace popsána nerovností (4.58) s $p = 0, \dots, \nu + 1$. Prostor $\mathscr{R}_h^{4\nu+1}$ lze samozřejmě užít i k řešení rovnic nižších řádů.

4.2. Čtyřúhelníkové prvky

V současné době jsou čtyřúhelníkové prvky méně populární než trojúhelníkové prvky. Všimneme si jich proto pouze velice stručně.

4.2.1. Obdélníkové Lagrangeovy prvky

Tyto prvky podobně jako obdélníkové Hermitovy prvky, které popíšeme v odst. 4.2.2, jsou vhodné k řešení úloh na oblastech, jejichž hranice je tvořena rovnoběžkami se souřadnicovými osami. Jsou založeny na myšlence Lagrangeovy interpolace a vzniknou tak, že za lokální aproximaci na každém obdélníku vezmeme funkci, která vznikne formálním vynásobením polynomů stupně m v proměnné x a y. Obecný tvar polynomu tohoto typu je

$$U(x, y) = \sum_{i=0}^{m} \sum_{j=0}^{m} \alpha_{ij} x^i y^j, \tag{4.78}$$

takže k jednoznačnému určení jeho koeficientů je třeba $(m + 1)^2$ uzlových parametrů. Obvykle se za ně berou funkční hodnoty ve vrcholech navzájem podobných obdélníků, které vzniknou tak, že každou stranu obdélníku rozdělíme na m stejných dílů a dělicí body spojíme rovnoběžkami se stranami. Tak např. v případě $m = 1$ jsou uzlovými parametry hodnoty ve vrcholech obdélníku a mluvíme o bilineární aproximaci, v případě $m = 2$ jsou uzlovými parametry funkční hodnoty ve

vrcholech obdélníku, ve středech jeho stran a v těžišti a mluvíme o bikvadratické aproximaci apod.

Vzniklé prostory končených prvků jsou tvořeny podobně jako v případě trojúhelníkových Lagrangeových prvků funkcemi, které mají po částech spojité derivace, a jsou proto podprostory Sobolevova prostoru $\mathscr{H}^1(\Omega)$. Jsou tedy vhodné k řešení úloh druhého řádu. Ve spojení s Ritzovou-Galerkinovou metodou vedou k rychlosti konvergence, která je stejného řádu jako při užití příslušných trojúhelníkových Lagrangeových prvků. Bázové funkce se v těchto prostorech konstruují opět zcela stejně jako v případě trojúhelníkových prvků. Za bázovou funkci se bere funkce, která má právě jeden uzlový parametr rovný jedné a ostatní rovny nule. Je zřejmé, že takto vzniklé funkce budou mít malé nosiče. Označíme-li u_j, $j = 1, \dots, (m+1)^2$, hodnoty interpolované funkce v uzlech, lze psát na daném obdélníku interpolační polynom U ve tvaru

$$U(x,y) = \sum_{j=1}^{(m+1)^2} u_j \psi_j^{(m)}(x,y), \tag{4.79}$$

kde pro elementární funkce $\psi_j^{(m)}$ platí

$$\psi_j^{(m)}(x_i, y_i) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{4.80}$$

Každá bázová funkce je tedy na tom obdélníku, kde je nenulová, rovna některé elementární funkci $\psi_j^{(m)}$.

Protože obdélníkové prvky se užívají hlavně v těch případech, kdy hranice dané oblasti je tvořena úsečkami rovnoběžnými se souřadnými osami, omezíme se v dalším na obdélník, jehož strany jsou rovnoběžné se souřadnicovými osami. Každý takový obdélník lze převést jednoduchou transformací typu

$$\begin{aligned} p_1 &= \frac{1}{h_1}(x - \xi), \\ p_2 &= \frac{1}{h_2}(y - \eta) \end{aligned} \tag{4.81}$$

na jednotkový referenční čtverec v rovině (p_1, p_2). Každou elementární funkci $\psi_j^{(m)}$ tedy stačí zadat pouze v referenční rovině a pak užít transformaci inverzní k transformaci (4.81).

Tak např. pro $m = 1$ jsou funkce $\psi_j^{(1)}$ dány vzorci

$$\begin{aligned} \psi_1^{(1)} &= p_1 p_2, \\ \psi_2^{(1)} &= (1 - p_1) p_2, \\ \psi_3^{(1)} &= (1 - p_1)(1 - p_2), \\ \psi_4^{(1)} &= p_1 (1 - p_2), \end{aligned} \tag{4.82}$$

kde indexy 1 až 4 se vztahují k vrcholům referenčního čtverce $(1,1)$, $(0,1)$, $(0,0)$ a $(1,0)$. Podobně lze také vyjádřit funkce $\psi_j^{(2)}$.

4.2.2 Obdélníkové Hermitovy prvky

Tyto prvky jsou založeny na Hermitově interpolaci a formálně je sestrojíme jako součin jednorozměrných Hermitových spline-funkcí (viz odst. 4.3.2 z kap. II). Vzniklý prostor konečných prvků označíme $\mathscr{Q}_h^{2\nu-1}$. Tento prostor je tvořen funkcemi, které se na každém obdélníku rovnají funkci U, která je součinem polynomů v x a y stupňů $2\nu - 1$, a je tedy tvaru

$$U(x,y) = \sum_{i=0}^{2\nu-1} \sum_{j=0}^{2\nu-1} \alpha_{ij} x^i y^j. \tag{4.83}$$

Polynom (4.83) má $4\nu^2$ koeficientů, a je tedy určen $4\nu^2$ uzlovými parametry. Za ty se volí hodnoty derivací $\partial^{\alpha_1+\alpha_2}/(\partial x^{\alpha_1} \partial y^{\alpha_2})$ pro $\alpha_1, \alpha_2 = 0, \dots, \nu - 1$ ve vrcholech obdélníku. Pro prostor $\mathscr{Q}_h^{2\nu-1}$ platí $\mathscr{Q}_h^{2\nu-1} \subset \mathscr{H}^\nu(\Omega)$, takže je vhodný k řešení problémů řádu 2ν. Báze v něm se sestrojí opět standardním postupem, přičemž elementární polynomy, pomocí nichž jsou definovány bázové funkce na jednotlivých obdélnících, se sestrojí obdobně jako v případě Lagrangeových prvků.

4.3 Algoritmické otázky spojené s metodou konečných prvků

V předchozích odstavcích jsme popsali základní myšlenky, na nichž je založena metoda konečných prvků, takže je více méně jasné, jak při řešení konkrétní okrajové úlohy postupovat. Je třeba (i) zvolit prostor konečných prvků a jeho bázi; (ii) sestavit matici tuhosti a vektor zatížení; (iii) řešit vzniklou soustavu lineárních rovnic; (iv) vypočítat hodnoty přibližného řešení. Body, které jsme uvedli, jsou samozřejmě stejné jako v jednodimenzionálním případě, neboť princip Ritzovy-Galerkinovy metody nezávisí na počtu dimenzí. Dát však odpověď na problémy, které jsou v těchto bodech obsaženy tak, aby vzniklý algoritmus byl co nejefektivnější, je zde však podstatně komplikovanější. Tak např. při konstrukci prostoru konečných prvků a jeho báze jsme se v jednodimenzionálním případě snažili dosáhnout především toho, aby příslušná matice tuhosti byla pásová o co možná nejmenší šíři pásu. To zde platí v podstatě také, ale jen potud, pokud počet neznámých dovolí řešit uvažovanou soustavu eliminací. Je-li třeba užít některou iterační metodu (zejména z důvodu potřeby paměti), je pouhá řídkost příslušné matice většinou stejně užitečná jako její pásovost. Volbou prostoru končených prvků lze však velmi podstatně ovlivnit další vlastnosti této matice, které umožňují užít zvolenou iterační metodu co nejefektivněji (srv. odst. 2.3.2). Proto při vytváření programů užívajících k řešení okrajových úloh metodu konečných prvků se setkáme s netriviálními problémy už hned při návrhu konkrétního prostoru konečných prvků.

Přístup k problémům týkajících se bodu (ii) je v podstatě stejný jako jsme uvedli v odst. 4.3.3 z kap. II. Opět jednotlivé prvky matice tuhosti skládáme z příspěvků na jednotlivých útvarech rozkladu a integrály, které je třeba přitom vypočítat, transformujeme na integrály na referenčních útvarech. Technické komplikace při provádění tohoto postupu jsou však značné. Už jen problém, jak ekonomicky popsat rozklad dané oblasti na elementární útvary, není zdaleka triviální. K tomuto problému přistupují problémy spojené s přibližným výpočtem dvourozměrných integrálů, kdy příslušné kvadraturní vzorce jsou rovněž komplikovanější než v jedné dimenzi, nutnost eventuálního výpočtu křivkových integrálů apod. Pokus podat jen hrubý návod jak se vypořádat s uvedenými problémy, by vlastně znamenal vypracovat projekt implementace metody konečných prvků pro řešení určité skupiny problémů, což přesahuje rámec této elementární učebnice. Proto jsme se omezili na těchto několik poznámek.

CVIČENÍ

1. Sestavte metodu přímek pro řešení Dirichletova problémů pro Laplaceovu rovnici $\Delta u = 0$ na jednotkovém čtverci tak, že nahradíte diferenčním podílem $[u(x, y-h) - 2u(x,y) + u(x, y+h)]/h^2$ derivaci $\partial^2 u/\partial y^2$.
2. Dokažte princip maxima pro operátor $L_h^{(1)}$ definovaný rovnicí (2.2).
3. Odvoďte chybu aproximace derivace $\partial^2 u/(\partial x \partial y)$ dané vzorcem (2.13).
4. Dokažte platnost vzorce (2.22).
5. Proveďte podrobně výpočty potřebné k důkazu věty 2.3 na str. 251.
6. Dokažte, že pro operátor Δ_h definovaný rovnicí (2.21) platí princip maxima.
7. Dokažte, že metoda sítí $(\Delta_h u)_{ks} = 0$ pro řešení Dirichletova problému pro Laplaceovu rovnici v takové oblasti Ω, že všechny hraniční uzly leží na její hranici, vede k rychlosti konvergence $O(h^6)$. Návod: Použijte pomocnou funkci jako v důkazu věty 2.4 na str. 254 a výsledek cvič. 6.)
8. Dokažte, že metoda sítí (2.27), (2.61) má chybu řádu $O(h^2)$.
9. Dokažte konvergenci metody sítí (2.27), (2.66) pro řešení Neumannovy úlohy.
10. Dokažte, že vektory $v^{(\nu,\mu)}$, jejichž složky jsou dány vzorci (2.91), tvoří ortonormovanou soustavu vlastních vektorů matice A_h metody (2.86).
11. Prověřte, že číslování uzlů tak, jak je popsáno na str. 275, skutečně vede k matici tvaru (2.120).
12. Prověřte platnost vzorců (2.121) až (2.125).
13. Dokažte správnost algoritmu cyklické redukce.

14. Rozmyslete si, proč oblast $\Omega = \{(x,y);\ (x^2+y^2 < 1) \wedge (x \notin \langle 0,1\rangle)\}$ nemá lipschitzovskou hranici.
15. Prověřte, že funkce $[-\frac{1}{2}\ln(x^2+y^2)]^{1/4}$ patří do $\mathscr{H}^1(\Omega)$, kde $\Omega = \{(x,y);\ x^2 + y^2 < 1\}$, a přitom není spojitá.
16. Dokažte, že pro úlohu (2.1), (3.1) platí nerovnost (4.1).
17. Dokažte nerovnost (4.35).
18. Nalezněte užitím referenčního trojúhelníku vyjádření elementárních polynomů $p_j^{(3)}$ kubické Lagrangeovy interpolace.
19. Odvoďte tvar elementárních funkcí $\psi_j^{(2)}$ bikvadratického obdélníkového prvku.

POZNÁMKY K LITERATUŘE

Čl. 1. Teorie okrajových úloh pro parciální diferenciální rovnice eliptického typu je rozpracována do značných podrobností a je jí věnována rozsáhlá literatura. Podrobný výklad založený na funkcionální analýze nalezne čtenář např. u Nečase (1967), Lionse a Magenese (1968) a v knize Rektorysově (1976). Okrajové úlohy pro rovnice vznikající v matematické teorii pružnosti jsou velmi detailně rozebrány v knize Nečasově a Hlaváčkově (1983). Klasického pojetí se přidržují např. knihy Petrovského (1950), Mirandy (1955) a Berse, Johna a Schechtera (1964). O numerické problematice uvažované v této kapitole nalezne čtenář mnoho cenného materiálu např. v knihách Collatze (1951), Forsytha a Wasowa (1960), Babušky, Prágera a Vitáska (1964), Berezina a Židkova (1966), Babušky, Prágera a Vitáska (1966), Mitchella (1969), Marčuka (1980) a Meise a Mareowitze (1981). Poslední ze zmíněných knih obsahuje kromě toho některé užitečné programy a kniha Marčukova rozsáhlou bibliografii. Výběr látky zpracované v této kapitole se do značné míry kryje s obsahem příslušné kapitoly jiné autorovy knihy (Vitásek (1987)), která však jako přehled základních numerických metod neobsahuje žádné důkazy.

Čl. 2. O metodě sítí bylo napsáno velmi mnoho. Obsáhlou bibliografii nalezne čtenář např. v už zmíněné knize Marčukově (1980) a ve sborníku redigovaném Jacobsem (1977). Základní principy této metody jsou vyloženy např. ve všech knihách, o nichž jsme se zmínili, kniha Samarského a Andrejeva (1967) je monografií, která je celá věnována metodě sítí pro řešení eliptických problémů. O užití integrálních identit k sestavení diferenčních rovnic viz Babuška, Práger a Vitásek (1966). Souvislost konvergence se stabilitou diferenčních schémat studuje např. Rjabeňkij a Filippov (1966) a Samarskij a Gulin (1973). Rovněž existuje velmi rozsáhlá literatura zabývající se řešením soustav lineárních rovnic, které vznikají v metodě sítí. Řadu informací i v tomto směru nalezne čtenář v obecných pramenech, které jsme uvedli. Sborník Jacobsův (1977) podává velmi zasvěcený přehled o nejnovějších modifikacích algoritmů založených hlavně na přímých metodách a obsáhlou

bibliografii. Varga (1962), Wachspress (1966), Young (1971) a Samarskij, Nikolajev (1978) představují jedny z nejúplnějších pramenů ke studiu iteračních metod. Upozorněme také na rozsáhlý soubor programů pro řešení eliptických okrajových úloh, v němž se kromě metody sítí užívá i metoda konečných prvků a který je podrobně popsán v knize Riceově a Boisvertově (1985).

Čl. 3. Teoretický základ, na němž spočívají variační metody řešení eliptických úloh, je v podstatě stejný jako v případě obyčejných diferenciálních rovnic. Proto i zde odkazujeme na Rektoryse (1974) a na knihy Michlinovy (1966 a 1970). Kromě toho většina knih zabývajících se metodou konečných prvků, např. Strang, Fix (1973), Mitchell a Wait (1977), Ciarlet (1978), Johnson (1988), Křížek a Neittaanmäki (1990), obsahuje aspoň přehled základních teoretických výsledků.

Čl. 4. Metoda konečných prvků patří v současné době k jedné z nejužívanějších metod pro řešení eliptických okrajových úloh. Přestože její bouřlivý rozvoj začíná teprve na sklonku šedesátých let, je jí už věnována obrovská literatura. Např. už v r. 1976 vyšla kniha Norrieova a de Vriesova, která je cele věnována pouze bibliografii této metody. Velmi obsáhlou bibliografii nalezne čtenář také v další knize Norrieově a de Vriesově (1978), která představuje velmi užitečný pramen pro všechny aspekty spojené s metodou konečných prvků. Kniha Ciarletova (1978), která má rovněž obsáhlou komentovanou bibliografii, je sice pro studium dosti náročná, jako jeden z nejúplnějších pramenů je však třeba ji maximálně doporučit. Velmi přístupně je napsána kniha Mitchellova a Waitova (1977), kniha Axelssonova a Barkerova (1984) a zejména kniha Johnsonova (1988). I v nich nalezne čtenář mnoho cenných informací. Řada dalších pramenů ke studiu metody končených prvků je uvedena v seznamu literatury, aniž bychom o nich zde konkrétně hovořili, upozorněme jen ještě na zcela elementární knihu Beckerovu, Careyovu a Odenovu (1981). Otázky spojené s problematikou řešení vzniklých soustav lineárních rovnic a s technikou programování jsou rovněž v poslední době intenzívně studovány. Relevantním pramenem je např. Bathe a Wilson (1976) a příslušné kapitoly z knihy Norrieovy a de Vriesovy (1978). Softwarové realizaci je výhradně věnován např. sborník redigovaný Kardestuncerem a Norriem (1987).

LITERATURA

AXELSSON, O. – BARKER, V.A.: Finite Element Solution of Boundary Value Problems, Theory and Computation. New York–San Francisco–London, Academic Press 1984.

BABUŠKA, I. – PRÁGER, M. – VITÁSEK, E.: Numerické řešení diferenciálních rovnic. Praha, SNTL 1964.

BABUŠKA, I. – PRÁGER, M. – VITÁSEK, E.: Numerical Processes in Differential Equations. London–New York–Sydney, Interscience Publishers 1966.

BATHE, K.J. – WILSON, E.L.: Numerical Methods in Finite Element Analysis.

Englewood Cliffs, N.J., Prentice–Hall 1976.

BECKER, E.B. – CAREY, G.F. – ODEN, J.T.: Finite Elements. An Introduction, Vol. 1. Englewood Cliffs, N.J., Prentice–Hall 1981.

BEREZIN, I.S. – ŽIDKOV, N.P.: Metody vyčislenij. 3. vyd. Moskva, Nauka 1966, 2 sv.

BERS, L. – JOHN, F. – SCHECHTER, M.: Partial Differential Equations. New York–London–Sydney, Interscience Publishers 1964. (Překlad do ruštiny: Moskva, Mir 1966.)

CIARLET, P.G.: The Finite Element Method fo Elliptic Problems. Amsterdam, North Holland 1978.

COLLATZ, L.: Numerische Behandlung von Differentialgleichungen. Berlin–Göttingen–Heidelberg, Springer–Verlag 1951. (Překlad do ruštiny: Moskva, IL 1953.)

DUPONT, T. – SCOTT, R.: Polynomial Approximations in Sobolev Spaces. Math. Comp. *34*, 1980, s. 441 – 463.

FORSYTHE, G.E. – WASOW, W.R.: Finite Difference Methods for Partial Differential Equations. New York–London, J. Wiley and Sons 1960. (Překlad do ruštiny: Moskva, IL 1963.)

GALLAGHER, R.G.: Finite Element Analysis. Englewood Cliffs, N.J., Prentice–Hall 1975.

JACOBS, D.A.G. (ed.): The State of the Art in Numerical Analysis. London–New York–San Francisco, Academic Press 1977.

JOHNSON, C.: Numerical Solutions of Partial Differential Equations by the Finite Element Method. Cambridge, Cambridge University Press 1988.

KARDESTUNCER, H. – NORRIE, D.H. (eds.): Finite Element Handbook. New York, McGraw-Hill 1987.

KŘÍŽEK, M. – NEITTAANMÄKI, P.: Finite Element Approximation of Variational Problems and Applications. Essex, Longman 1990.

LIONS, J.L. – MAGENES, E.: Problèmes aux limites nonhomogènes et applications. Paris, Dunod 1968–1970, 3 sv. (Překlad do ruštiny: Moskva, Mir 1971.)

MARČUK, G.I.: Metody vyčislitel'noj matematiky. 2. vyd. Moskva, Nauka 1980. (Překlad do češtiny: Praha, Academia 1987.)

MARTIN, H.C. – CAREY, G.F.: Introduction to Finite Element Analysis. New York, McGraw-Hill 1973.

MEIS, T. – MARCOWITZ, V.: Numerical Solution of Partial Differential Equations. New York–Heidelberg–Berlin, Springer–Verlag 1981.

MICHLIN, S.G.: Čislennaja realizacija variacionnych metodov. Moskva, Nauka 1966.

MICHLIN, S.G.: Variacionnyje metody v matematičeskoj fizike. 2. vyd. Moskva, Nauka 1970.

MIRANDA, C.: Equazioni alle derivate parziali di tipo ellittico. Berlin–Göttingen–Heidelberg, Springer–Verlag 1955. (Překlad do ruštiny: Moskva, IL 1957.)

MITCHELL, A.R.: Computational Methods in Partial Differential Equations. London–New York–Sydney–Toronto, J. Wiley and Sons 1969.
MITCHELL, A.R. – WAIT, R.: The Finite Element Method in Partial Eifferential Equations. Chichester–New York–Brisbane–Toronto, J. Wiley and Sons 1977. (Překlad do ruštiny: Moskva, Mir 1981.)
NEČAS, J.: Les méthodes directes en théorie des équations elliptiques. Praha, Academia 1967.
NEČAS, J. – HLAVÁČEK, I.: Úvod do matematické teorie pružných a pružně plastických těles. Praha SNTL 1983.
NORRIE, D.H. – VRIES, G.A., DE: A Finite Element Bibliography. New York, Plenum Press 1976.
NORRIE, D.H. – VRIES, G.A., DE: An Introduction to Finite Element Analysis. New York–San Francisco–London, Academic Press 1978. (Překlad do ruštiny: Moskva, Mir 1981.)
ODEN, J.T.: Finite Element of Nonlinear Continua. New York, McGraw-Hill 1972. (Překlad do ruštiny: Moskva, Mir 1975.)
PETROVSKIJ, I.G.: Lekcii ob uravnenijach s častnymi proizvodnymi. Moskva–Leningrad, Gostechizdat 1950. (Překlad do češtiny: Praha, Přírodovědecké vydavatelství 1952.)
PROSKUROWSKI, W. – WIDLUND, O.: On the Numerical Solution of Helmholtz's Equaiton by the Capacitance Matrix Method. Math. Comp. *30*, 1976, s. 433 – 468.
REKTORYS, K.: Variační metody v inženýrských problémech a v problémech matematické fyziky. Praha, SNTL 1974.
RICE, J.R. – BOISVERT, R.G.: Solving Elliptic Problems Using ELLPACK. New York–Berlin–Geidelberg, Springer–Verlag 1985.
RJABEŇKIJ, V.S. – FILLIPPOV, A.F.: Ob ustojčivosti raznostnych uravněnij. Moskva, Gostechizdat 1956.
SAMARSKIJ, A.A. – GULIN, A.V.: Ustojčivosť raznostnych schem. Moskva, Nauka 1973.
SAMARSKIJ, A.A. – ANDREJEV, V.D.: Raznostnyje metody dlja elliptičeskich uravnenij. Moskva, Nauka 1976.
SAMARSKIJ, A.A. – NIKOLAJEV, J.S.: Metody rešenija setočnych uravnenij. Moskva, Nauka 1978. (Překlad do češtiny: Praha, Academia 1984.)
SCHWARTZ, L.: Théorie des distributions I et II. 2. vyd. Paris, Hermann 1957.
STONE, H.L.: Iterative Solutions of Implicit Approximations of Multidimensional Partial Differential Equaitons. SIAM J. Numer. Anal. *5*, 1968, s. 530 – 558.
STRANG, G. – FIX, G.J.: An Analysis of the Finite Element Method. Englewood Cliffs, N.J., Prentice–Hall 1973. (Překlad do ruštiny: Moskva, Mir 1977.)
VARGA, R.: Matrix Iterative Analysis. Englewood Cliffs, N.J., Prentice–Hall 1962.
WACHSPRESS, E.L.: Iterative Solution of Elliptic Systems. Englewood Cliffs, N.J., Prentice–Hall 1966.

YOUNG, D.H.: Iterative Soluytion of Large Linear Systems. New York–London, Academic Press 1971.
ZIENKIEWICZ, O.C.: The Finite Elements Method in Engineering Science. 2. vyd. New York, McGraw-Hill 1971. (Překlad do ruštiny: Moskva, Mir 1975.)

Kapitola IV.

Parciální diferenciální rovnice parabolického typu

1 Úvod

Dostatečně obecným příkladem parciální diferenciální rovnice parabolického typu v $m+1$ proměnných je rovnice

$$Lu = f(x,t), \quad (x,t) \in R, \tag{1.1}$$

kde $R = \Omega \times (0,T)$, $\Omega \subset \mathbf{E}^m$, je omezená oblast, T je kladná konstanta a

$$Lu = c\frac{\partial u}{\partial t} - \sum_{i,j=1}^{m} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + qu. \tag{1.2}$$

Koeficienty $c(x,t)$, $a_{ij}(x,t)$, $q(x,t)$ a pravá strana $f(x,t)$ jsou dané funkce $m+1$ proměnných, o nichž se obvykle předpokládá, že pro ně platí

$$c(x,t) \geqq c_0, \quad a_{ij}(x,t) = a_{ji}(x,t), \tag{1.3}$$
$$\sum_{i,j=1}^{m} a_{ij}(x,t)\xi_i\xi_j \geqq p_0 \sum_{i=1}^{m} \xi_i^2,$$

pro každé $(x,t) \in R$ a $\xi_i \in \mathbf{E}^1$, kde c_0 a p_0 jsou kladné konstanty. Parabolický diferenciální operátor L lze tedy psát také ve tvaru $Lu = c\partial u/\partial t + L_e u$, kde L_e je eliptický diferenciální operátor.

Typická úloha pro rovnici (1.1) je úloha nalézt funkci u, která splňuje v R rovnici (1.1) a pro niž platí

$$u(x,0) = g(x), \quad x \in \Omega, \tag{1.4}$$

a

$$\alpha\frac{\partial u}{\partial n_c} + \beta u = \gamma, \quad x \in \Gamma,\ t \in (0,T), \tag{1.5}$$

kde funkce α, β a γ jsou definovány na $\Gamma \times (0,T)$, $\alpha(x,t) + \beta(x,t) > 0$ a $\partial u/\partial n_c$ je derivace ve směru konormály.

Podmínka (1.4) je počáteční podmínka a podmínka (1.5) okrajová podmínka. Pro okrajovou podmínku (1.5) se také užívá při $\alpha \equiv 0$ (podobně jako v případě okrajové úlohy pro eliptickou parciální diferenciální rovnici) název Dirichletova podmínka, při $\beta \equiv 0$ Neumannova podmínka a v obecném případě Newtonova podmínka.

Otázkami řešitelnosti právě zformulované úlohy se nebudeme zabývat a odkážeme čtenáře na obsáhlou specializovanou literaturu (viz např. Petrovskij (1952) nebo Friedman (1964) pokud jde o klasickou teorii, nebo Lions (1961) pokud jde o moderní funkcionálně analytickou teorii.) Důvody pro to jsou podobné, jako tomu bylo v kap. III, a nebudeme je proto znovu opakovat. Všimneme si však jiné okolnosti patrné ostatně na první podhled. Z rovnice (1.1) a z doplňujících podmínek (1.4) a (1.5) je vidět, že proměnná t, jejíž fyzikální význam je obvykle čas, má mezi všemi nezávislými proměnnými výjimečné postavení a že vzhledem k této proměnné má daná úloha charakter úlohy s počátečními podmínkami, zatímco vzhledem k ostatním proměnným, nazvěme je prostorovými proměnnými, jde o okrajovou úlohu. Dá se proto očekávat, že algoritmy pro řešení dané úlohy budou mít vzhledem k proměnné t rekurentní charakter, a že tedy metody pro numerické řešení parabolických problémů budou vykazovat současně rysy metod pro řešení úloh s počátečními podmínkami i okrajových úloh. Tato skutečnost bude v dalším výkladu zřetelně patrná a přinese také některé specifické zvláštnosti.

Z metod pro přibližné řešení parabolických parciálních diferenciálních rovnic popíšeme podrobně zejména metodu sítí čítaje v to důležitý speciální případ metody střídavých směrů. Dále pak si všimneme tzv. *semidiskrétních metod*. Tyto metody jsou obdobou metody přímek, o níž jsme se zmínili v kap. III, a spočívají v tom, že diskretizace se provede pouze vzhledem k prostorovým proměnným $x_1, \ldots, x_m$ a proměnná t (čas) se ponechá spojitá. U těchto metod se tedy řešení původní úlohy aproximuje řešením soustavy obyčejných diferenciálních rovnic. Pokud se při zmíněné diskretizaci vychází z myšlenky metody sítí, vznikne klasická metoda přímek, užijeme-li metodu konečných prvků, dostaneme semidiskrétní metodu Galerkinova typu. Podobné semidiskrétním metodám jsou metody Rotheovy, kdy se diskretizace provádí pouze vzhledem k proměnné t. Náhradní úloha je pak okrajová úloha pro eliptickou soustavu diferenciálních rovnic. I o těchto metodách se stručně zmíníme.

2 Metoda sítí

Metoda sítí vychází i zde z úplně stejných principů, jak jsme o nich hovořili v kap. II a III, a je to jedna z nejpopulárnějších metod pro řešení parabolických diferenciálních rovnic. Užití metody sítí k řešení okrajových úloh pro obyčejné diferenciální rovnice a pro eliptické parciální diferenciální rovnice vede, jak víme, na soustavy lineárních algebraických rovnic. Zde je tomu v principu také tak, v důsledku zvláštního postavení proměnné t však budou tyto soustavy speciálního charakteru, což přinese některé další jevy, zejména problém stability, s nimiž jsme se až dosud ne-

setkali. Abychom tyto problémy co nejjednodušeji objasnili, vyšetříme metodu sítí nejprve v jednoduchém speciálním případě.

2.1 Rovnice pro vedení tepla v jedné prostorové proměnné

V tomto odstavci se budeme zabývat rovnicí

$$Lu \equiv \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \quad x \in (0,1),\ t \in (0,T) \tag{2.1}$$

s počáteční podmínkou

$$u(x,0) = g(x), \quad x \in (0,1) \tag{2.2}$$

a okrajovými podmínkami

$$u(0,t) = u(1,t) = 0, \quad t \in (0,T), \tag{2.3}$$

která je zřejmě speciálním případem rovnice (1.1) a která se vzhledem ke své nejběžnější fyzikální interpretaci nazývá rovnicí *pro vedení tepla*. Začneme dvěma nejjednoduššími variantami metody sítí pro řešení této rovnice. Předtím však ještě poznamenejme, že existence a jednoznačnost řešení úlohy (2.1) až (2.3) je zaručena např. spojitostí a omezeností funkce g.

2.1.1 Explicitní a implicitní metoda

V případě úlohy (2.1) až (2.3) je množina R obdélník $(0,1) \times (0,T)$. Sestrojme v něm síť složenou z přímek $x = x_k$ a $t = t_l$, kde $x_k = kh$, $k = 0,\dots,n$, $h = 1/n$, $t_l = l\tau$, $l = 0,\dots,r$, $\tau = T/r$ a n a r jsou přirozená čísla. Skutečnost, že proměnné x a t nemají v rovnici (2.1) stejnou roli tedy respektujeme tím, že uvažujeme síť obdélníkovou nikoliv čtvercovou. Znakem $\bar{R}_{h,\tau}$ označme množinu všech uzlů (x_k,t_l) ležících v $\bar{R}$, znakem $R_{h,\tau}$ označme ty uzly (x_k,t_l), pro něž je $k = 1,\dots,n-1$ a $l = 1,\dots,r$, a $\Gamma_{h,\tau}$ uzly, které leží na úsečkách $t = 0$, $0 \leqq x \leqq 1$, $x = 0$, $0 < t \leqq T$ a $x = 1$, $0 < t \leqq T$. Je-li nyní $L_{h,\tau}^{(0)}$ operátor, který funkci $u = u_k^{(l)}$ definované na množině uzlů $\bar{R}_{h,\tau}$ přiřazuje funkci $L_{h,\tau}^{(0)}u$ definovanou na množině $R_{h,\tau}$ a který je dán vzorcem

$$(L_{h,\tau}^{(0)}u)_k^{(l)} = \frac{1}{\tau}[u_k^{(l)} - u_k^{(l-1)}] - \frac{1}{h^2}[u_{k-1}^{(l-1)} - 2u_k^{(l-1)} + u_{k+1}^{(l-1)}], \quad (x_k,t_l) \in R_{h,\tau}, \tag{2.4}$$

platí pro každou dostatečně hladkou funkci u

$$(L_{h,\tau}^{(0)}u^{(\mathrm{pr})})_k^{(l)} = (Lu)(x_k,t_l) + O(\tau + h^2), \quad (x_k,t_l) \in R_{h,\tau}, \tag{2.5}$$

kde $u^{(\mathrm{pr})}$ je funkce definovaná na množině $\bar{R}_{h,\tau}$ předpisem $(u^{(\mathrm{pr})})_k^{(l)} = u(x_k,t_l)$, jak se snadno zjistí pomocí Taylorova vzorce.

Vzhledem k této rovnici je přirozené hledat přibližné řešení $u_k^{(l)}$ ze soustavy rovnic

$$(L_{h,\tau}^{(0)}u)_k^{(l)} = 0, \quad k = 1,\dots,n-1,\ l = 1,\dots,r. \tag{2.6}$$

Protože těchto rovnici je zřejmě méně než neznámých, je k nim třeba ještě připojit rovnice

$$u_k^{(0)} = g(x_k), \quad k = 0,\dots,n, \tag{2.7}$$

a

$$u_0^{(l)} = u_n^{(l)} = 0, \quad l = 1,\dots,r, \tag{2.8}$$

získané z počáteční podmínky a z okrajových podmínek.

Všimněme si, že jsme se zde přiklonili k možnosti předepsat hodnoty přibližného řešení v rohových uzlech $(0,0)$ a $(1,0)$ z počáteční podmínky. Stejně oprávněně jsme mohli také užít okrajové podmínky. Žádáme-li totiž, aby řešení úlohy (2.1) až (2.3) bylo alespoň spojité v $\bar{R}$, je nutné, aby pro funkci g platily podmínky $g(0) = g(1) = 0$. Tyto podmínky se nazývají *podmínkami souhlasu*, a nejsou-li splněny, je nutně řešení dané okrajové úlohy v bodech $(0,0)$ a $(1,0)$ nespojité. V tomto případě je pak jedno, kterou z uvedených možností zvolíme.

Soustava (2.6) až (2.8) představuje soustavu $(n+1)(r+1)$ rovnic pro $(n+1)(r+1)$ hodnot hledaného řešení v uzlech sítě ležících v $\bar{R}_{h,\tau}$. Tato soustava je však velice speciální. Každá z rovnic (2.6) obsahuje pouze jednu hodnotu přibližného řešení v čase $t = t_l$ a rovnice (2.7) a (2.8) pak přímo udávají hodnoty některých neznámých. Vynásobíme-li rovnice (2.6) číslem τ, dostaneme po elementární úpravě

$$u_k^{(l)} = \beta u_{k-1}^{(l-1)} + (1-2\beta)u_k^{(l-1)} + \beta u_{k+1}^{(l-1)}, \quad k = 1,\dots,n-1,\ l = 1,\dots,r, \tag{2.9}$$

kde

$$\beta = \frac{\tau}{h^2}. \tag{2.10}$$

Protože přibližné řešení $u_k^{(0)}$ pro $k = 0,\dots,n$ známe, můžeme použitím rovnic (2.9) s $l = 1$ vypočítat hodnoty $u_k^{(1)}$ pro $k = 1,\dots,n-1$; hodnoty $u_k^{(0)}$, $u_k^{(1)}$ opět známe (v našem případě jsou to nuly), můžeme tedy vypočítat $u_k^{(2)}$ pro $k = 1,\dots,n-1$ atd. Přibližné řešení počítáme tedy rekurentně postupně pro přibývající časové řádky sítě. Žádný problém s řešením odvozené soustavy vlastně nevzniká a rovnice (2.9), (2.7) a (2.8) popisující diferenční schéma udávají přímo algoritmus přibližného řešení uvažované úlohy. Z tohoto důvodu se právě popsaná varianta metody sítí nazývá *explicitní metoda* nebo *explicitní schéma*.

Lokální chyba, tj. chyba, která vznikne náhradou derivací v původní diferenciální rovnici diferenčními podíly, je u explicitní metody za předpokladu dostatečné hladkosti přesného řešení velikosti $O(\tau + h^2)$ (srv. rovnici (2.5)). Je tedy při dostatečně jemné síti libovolně malá. V eliptickém případě měla tato skutečnost většinou za

následek konvergenci, neboť diferenční rovnice vzniklé na základě zmíněného jednoduchého principu náhrady derivací diferenčními podíly byly už v podstatě vždy korektní, tj. jejich řešení spojitě záviselo na vstupních datech úlohy. Abychom posoudili platnost tohoto závěru v případě parabolické rovnice, začneme jednoduchým příkladem.

Příklad 2.1. Řešme rovnici (2.1) s okrajovými podmínkami (2.3) a s počáteční podmínkou $g(x) = \sin \pi x$. V tomto jednoduchém příkladě je přesné řešení dáno vzorcem

$$u(x,t) = e^{-\pi^2 t} \sin \pi x, \tag{2.11}$$

takže lze snadno posoudit chybu vypočteného přibližného řešení. Protože lokální chyba je řádu $O(\tau + h^2)$, zvolili jsme časový integrační krok τ řádově rovný druhé mocnině prostorového integračního kroku h. Přibližné řešení a jeho chyba v bodě $x = 1/2$ je pro dvě alternativy sítě uvedeno v tab. 2.1 (symbol ∞ v ní značí, že došlo k přeplnění).

Tabulka 2.1

Řešení rovnice (2.1) explicitní metodou

T	$\tau = 10^{-4}$, $h = 10^{-2}$		$\tau = 10^{-4}$, $h = 2.10^{-2}$	
.10 000	přibl. řeš.	chyba	přibl. řeš.	chyba
1	0,999 013 2	0,000 000 4	0,999 012 9	0,000 000 6
2	0,998 027 1	0,000 000 9	0,998 027 3	0,000 000 7
3	0,997 042 5	0,000 001 0	0,997 042 4	0,000 001 1
4	0,996 057 8	0,000 002 1	0,996 058 2	0,000 001 7
5	0,995 076 7	0,000 000 7	0,995 075 0	0,000 002 4
6	0,994 089 9	0,000 005 8	0,994 093 2	0,000 002 6
7	0,993 121 3	−0,000 006 2	0,993 112 1	0,000 002 9
8	0,992 107 9	0,000 027 5	0,992 131 9	0,000 003 5
9	0,991 220 1	−0,000 063 4	0,991 153 0	0,000 003 7
10	0,989 986 4	0,000 192 6	0,990 174 8	0,000 004 4
⋮				
96	$-2{,}83.10^{37}$	$2{,}83.10^{37}$	0,909 562 3	0,000 039 6
97	$8{,}52.10^{37}$	$-8{,}52.10^{37}$	0,908 664 7	0,000 040 0
98	$-2{,}56.10^{38}$	$2{,}56.10^{38}$	0,907 768 2	0,000 040 1
99	$7{,}72.10^{38}$	$-7{,}72.10^{38}$	0,906 872 3	0,000 040 5
100	$-2{,}32.10^{38}$	$2{,}32.10^{38}$	0,905 977 5	0,000 040 6
⋮				
10 000	∞	∞	0,000 051 5	0,000 000 2

Z této tabulky vidíme, že v případě sítě $\tau = 10^{-4}$, $h = 10^{-2}$ dostáváme úplně nesmyslné výsledky, zatímco v druhém případě, i když užitá síť je hrubší, jsou výsledky přijatelné.

Pokusme se vysvětlit, proč tento na první pohled jistě překvapivý jev nastal. Položíme-li $u^{(l)} = (u_1^{(l)}, \ldots, u_{n-1}^{(l)})^T$, můžeme rovnici (2.9) s příslušnými okrajovými podmínkami (2.8) psát ve tvaru

$$u^{(l)} = A_{\mathrm{E}} u^{(l-1)}, \quad l = 1, \ldots, r, \tag{2.12}$$

kde

$$A_{\mathrm{E}} = \begin{bmatrix} 1-2\beta & \beta & 0 & \ldots & 0 \\ \beta & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta \\ 0 & \ldots & 0 & \beta & 1-2\beta \end{bmatrix} \tag{2.13}$$

je matice řádu $n-1$ a $u^{(0)}$ je známý vektor určený počátečními podmínkami. Matice A_{E} charakterizuje operaci přechodu od $(l-1)$-ního časového řádku sítě k l-tému řádku, a nazývá se proto *matice přechodu*.

Z rovnice (2.12) ihned plyne, že platí

$$u^{(l)} = A_{\mathrm{E}}^l u^{(0)}, \quad l = 0, \ldots, r. \tag{2.14}$$

Chování složek vektoru $u^{(l)}$ v závislosti na l je tedy určeno chováním prvků matice A_{E}^l. Vlastní čísla λ_ν matice A_{E} a příslušné vlastní vektory $v^{(\nu)} = (v_1^{(\nu)}, \ldots, v_{n-1}^{(\nu)})^T$ jsou však dány jednoduchými vzorci

$$\lambda_\nu = 1 - 4\beta \sin^2 \frac{\nu\pi}{2n}, \quad \nu = 1, \ldots, n-1, \tag{2.15}$$

a

$$v_k^{(\nu)} = \sin \frac{\nu\pi k}{n}, \quad k = 1, \ldots, n-1, \tag{2.16}$$

jak se snadno zjistí přímým výpočtem. Ze vzorce (2.15) plyne, že vlastní číslo λ_{n-1} je pro velká n přibližně rovno číslu $1-4\beta$. Některé prvky matice A_{E}^l se tedy chovají jako $(1-4\beta)^l$. Je-li však $\beta > 1/2$, je $1-4\beta < -1$, a tedy alespoň jeden prvek matice A_{E}^l musí při rostoucím l exponenciálně růst. Tato situace právě nastala pro první síť užitou v příkladě 2.1, neboť pro ni je $\beta = 1$.

Platí-li však, že je

$$\beta \leqq \frac{1}{2}, \tag{2.17}$$

plyne ihned ze vzorce (2.15), že všechna vlastní čísla matice A_{E} jsou v absolutní hodnotě menší než jedna. Prvky matice A_{E}^l jsou tedy omezené konstantou nezávis-

lou na r a n. Tento případ nastal pro druhou síť v uvedeném příkladě, neboť pro ni je $\beta = 1/4$.

Na základě uvedeného rozboru je zřejmé, že podmínka (2.17) je v případě uvažované diferenciální rovnice nutnou podmínkou konvergence. Následující věta ukazuje, že jde i o podmínku postačující.

Věta 2.1. *Nechť řešení problému* (2.1) až (2.3) *existuje a má v* $\bar{R}$ *dvě spojité derivace podle* t *a čtyři spojité derivace podle* x. *Nechť dále platí* (2.17). *Pak existují konstanty* M *a* $h_0 > 0$ *takové, že pro* $h \leqq h_0$ *a pro každý uzel* $(x_k, t_l) \in \bar{R}_{h,\tau}$ *platí*

$$|u_k^{(l)} - u(x_k, t_l)| \leqq M(\tau + h^2). \tag{2.18}$$

D ů k a z . Pro přesné řešení platí za uvedených předpokladů

$$(L_{h,\tau}^{(0)} u^{(\mathrm{pr})})_k^{(l)} = -\varepsilon_k^{(l)}, \quad (x_k, t_l) \in R_{h,\tau}, \tag{2.19}$$

kde funkce $u^{(\mathrm{pr})}$ je definována stejně jako ve vzorci (2.5) a kde

$$|\varepsilon_k^{(l)}| \leqq K(\tau + h^2). \tag{2.20}$$

Pro chybu $\eta_k^{(l)} = u_k^{(l)} - u(x_k, t_l)$ tedy platí

$$\begin{aligned} &(L_{h,\tau}^{(0)} \eta)_k^{(l)} = \varepsilon_k^{(l)}, \quad k = 1, \ldots, n-1, \ l = 1, \ldots, r, \\ &\eta_k^{(0)} = 0, \quad k = 0, \ldots, n, \\ &\eta_0^{(l)} = \eta_n^{(l)} = 0, \quad l = 1, \ldots, r, \end{aligned} \tag{2.21}$$

neboli

$$\begin{aligned} &\eta_k^{(l)} = \beta \eta_{k-1}^{(l-1)} + (1 - 2\beta)\eta_k^{(l-1)} + \beta \eta_{k+1}^{(l-1)} + \tau \varepsilon_k^{(l)}, \\ &k = 1, \ldots, n-1, \ l = 1, \ldots, r, \end{aligned} \tag{2.22}$$

s příslušnými počátečními a okrajovými podmínkami. Z rovnic (2.22) a z nerovností (2.20) však snadno indukcí dokážeme, že platí

$$|\eta_k^{(l)}| \leqq l\tau K(\tau + h^2) \tag{2.23}$$

pro $l = 1, \ldots, r$, neboť koeficienty v lineární kombinaci na pravé straně rovnice (2.22) jsou v důsledku splnění nerovnosti (2.17) nezáporné. Je tedy

$$|\eta_k^{(l)}| \leqq r\tau K(\tau + h^2) = TK(\tau + h^2), \tag{2.24}$$

což dokazuje větu.

Užití operátoru $L_{h,\tau}^{(0)}$ vede tedy ke konvergentní metodě, při limitním přechodu jsou však parametry sítě vázány nerovností (2.17). Z právě dokončeného důkazu věty 2.1 však plyne, že podmínka (2.17) zaručuje nejen konvergenci uvažované metody při $h \to 0$, $\tau \to 0$, ale i její uspokojivé chování vzhledem k zaokrouhlovacím

chybám. Při splnění této podmínky je tedy explicitní metoda v tomto smyslu stabilní. Naopak rozumné chování metody, u níž je podmínka (2.17) porušena, se nedá očekávat ani ve speciálním případě. Tak např. při užití první sítě v příkladě 2.1 jsou přesné hodnoty přibližného řešení $u_k^{(l)}$ dány vzorcem

$$u_k^{(l)} = \left(1 - 4\sin^2 \frac{\pi}{200}\right)^l \sin \frac{k\pi}{100}. \tag{2.25}$$

Z tohoto vzorce je vidět, že v případě speciální počáteční podmínky $u(x, 0) = \sin \pi x$ přesné hodnoty přibližného řešení nejen že nejsou nesmyslné, ale dokonce uspokojivě aproximují přesné řešení. Jelikož však mocniny matice přechodu užitého schématu nejsou omezené, rekurentní výpočet přibližného řešení se v důsledku zaokrouhlování zhroutí, a to, jak je vidět z tab. 2.1, velmi rychle.

Explicitní schéma, které jsme právě popsali, je tedy *stabilní* ve výše zmíněném smyslu, je-li splněna podmínka (2.17); v opačném případě je *nestabilní*. Protože jeho stabilita závisí na splnění podmínky, která podřizuje časové dělení sítě prostorovému dělení, mluvíme o *relativně stabilním schématu*. Podmínku stability (2.17) musíme samozřejmě respektovat i při zjemňování sítě. V tomto případě se tato podmínka může ukázat značně restriktivní, neboť rozpůlíme-li např. prostorový integrační krok, je třeba za časový krok brát čtvrtinu původního; počet potřebných časových řádků tak může neúnosně růst. Proto vzniká přirozená otázka, zda neexistuje diferenční schéma pro řešení úlohy (2.1) až (2.3), které by ke své stabilitě nevyžadovalo splnění žádné podmínky typu (2.17). Pokusme se takové schéma, které nazveme *absolutně stabilní*, nalézt.

K explicitnímu schématu jsme došli tak, že jsme derivaci $\partial u/\partial t$ v uzlu (x_k, t_{l-1}) aproximovali podílem $[u(x_k, t_l) - u(x_k, t_{l-1})]/\tau$. Tento podíl však aproximuje stejně dobře uvedenou derivaci i v uzlu (x_k, t_l). Stejně oprávněně můžeme tedy dojít místo k operátoru $L_{h,\tau}^{(0)}$ k operátoru $L_{h,\tau}^{(1)}$ danému rovnicí

$$(L_{h,\tau}^{(1)} u)_k^{(l)} = \frac{1}{\tau}[u_k^{(l)} - u_k^{(l-1)}] - \frac{1}{h^2}[u_{k-1}^{(l)} - 2u_k^{(l)} + u_{k+1}^{(l)}], \tag{2.26}$$

neboť opět pro každou dostatečně hladkou funkci u platí

$$(L_{h,\tau}^{(1)} u^{(\mathrm{pr})})_k^{(l)} = (Lu)(x_k, t_l) + O(\tau + h^2) \tag{2.27}$$

(jak ihned plyne z Taylorova vzorce). Funkce $u^{(\mathrm{pr})}$ je přitom definována stejně jako ve vzorci (2.5).

Přibližné řešení $u_k^{(l)}$ se tedy můžeme pokusit hledat z rovnic

$$(L_{h,\tau}^{(1)} u)_k^{(l)} = 0, \quad k = 1, \ldots, n-1, \ l = 1, \ldots, r, \tag{2.28}$$

s počáteční podmínkou (2.7) a okrajovými podmínkami (2.8). Přepíšeme-li rovnice (2.28) podobně jako v případě explicitní metody, dostaneme

$$\begin{aligned} &-\beta u_{k-1}^{(l)} + (1 + 2\beta)u_k^{(l)} - \beta u_{k+1}^{(l)} = u_k^{(l-1)}, \\ &k = 1, \ldots, n-1, \ l = 1, \ldots, r, \end{aligned} \tag{2.29}$$

s příslušnými počátečními a okrajovými podmínkami. Použijeme-li znovu dříve zavedeného vektorového označení $u^{(l)} = (u_1^{(l)}, \dots, u_{n-1}^{(l)})^T$, lze rovnici (2.29) spolu s okrajovými podmínkami (2.8) přepsat maticově do tvaru

$$A_{\mathrm{I}} u^{(l)} = u^{(l-1)}, \quad l = 1, \dots, r, \tag{2.30}$$

kde

$$A_{\mathrm{I}} = \begin{bmatrix} 1+2\beta & -\beta & 0 & \dots & 0 \\ -\beta & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\beta \\ 0 & \dots & 0 & -\beta & 1+2\beta \end{bmatrix} \tag{2.31}$$

je opět třídiagonální matice řádu $n-1$ a $u^{(0)}$ je známý vektor určený počáteční podmínkou. Řešení probíhá opět rekurentně od jednoho časového řádku k druhému, k získání přibližného řešení v l-tém časovém řádku za předpokladu, že v $(l-1)$-ním časovém řádku je už známe, je však třeba řešit soustavu $n-1$ lineárních rovnic o $n-1$ neznámých s třídiagonální maticí (2.31). Z tohoto důvodu se metoda (2.28) nazývá *implicitní metoda* nebo *implicitní schéma*. Výpočet je proveditelný při libovolné hodnotě β, neboť matice A_{I} je zřejmě regulární. Dá se to zjistit např. pomocí lemmatu 3.5 z kap. II (viz str. 166), neboť matice A_{I} je ireducibilně diagonálně dominantní při libovolném kladném β, má kladné diagonální prvky a nekladné nediagonální prvky.

Roli matice přechodu implicitní metody hraje matice A_{I}^{-1}. Vlastní čísla μ_ν této matice jsou dána vzorcem

$$\mu_\nu = \frac{1}{1 + 4\beta \sin^2 \frac{\nu\pi}{2n}}, \quad \nu = 1, \dots, n-1, \tag{2.32}$$

takže jsou při libovolném $\beta > 0$ v absolutní hodnotě menší než 1. Přechod od $(l-1)$-ního časového řádku k l-tému časovému řádku je tedy stabilní při libovolném poměru prostorového a časového dělení sítě a v implicitní metodě tak máme příklad absolutně stabilní metody.

Obraťme se nyní ke studiu její konvergence. Stejně jako v případě explicitní metody, platí pro chybu $\eta_k^{(l)} = u_k^{(l)} - u(x_k, t_l)$ této metody rovnice

$$\begin{aligned} &(L_{h,\tau}^{(1)} \eta)_k^{(l)} = \varepsilon_k^{(l)}, \quad k = 1, \dots, n-1, \ l = 1, \dots, r, \\ &\eta_k^{(0)} = 0, \quad k = 0, \dots, n, \\ &\eta_0^{(l)} = \eta_n^{(l)} = 0, \quad l = 1, \dots, r, \end{aligned} \tag{2.33}$$

a existuje konstanta K taková, že je

$$|\varepsilon_k^{(l)}| \leqq K(\tau + h^2). \tag{2.34}$$

Rovnice pro chybu implicitní metody jsou tedy také implicitní, takže už nejsme schopni tak snadno jako v případě explicitní metody odhadnout na jejich základě velikost chyby. Dokážeme proto nejprve několik pomocných tvrzení.

Lemma 2.1. *Nechť pro funkci $\eta_k^{(l)}$ definovanou na množině $\bar{R}_{h,\tau}$ platí*

$$(L_{h,\tau}^{(1)} \eta)_k^{(l)} \leqq 0, \quad (x_k, t_l) \in R_{h,\tau}. \tag{2.35}$$

Pak pro každý uzel $(x_k, t_l) \in \bar{R}_{h,\tau}$ platí

$$\eta_k^{(l)} \leqq \max_{(x_k,t_l) \in \Gamma_{h,\tau}} \eta_k^{(l)}. \tag{2.36}$$

Důkaz. Buď $M = \max_{(x_k,t_l) \in \bar{R}_{h,\tau}} \eta_k^{(l)}$ a nechť existuje uzel $(x_{k_0}, t_{l_0}) \in R_{h,\tau}$, pro který je $\eta_{k_0}^{(l_0)} = M$. V tomto uzlu platí nerovnost (2.35), kterou můžeme psát ve tvaru

$$(1+2\beta)\eta_{k_0}^{(l_0)} - \beta\eta_{k_0-1}^{(l_0)} - \beta\eta_{k_0+1}^{(l_0)} - \eta_{k_0}^{(l_0-1)} \leqq 0. \tag{2.37}$$

Předpokládáme-li nyní, že aspoň v jednom z uzlů (x_{k_0}, t_{l_0-1}), (x_{k_0-1}, t_{l_0}) a (x_{k_0+1}, t_{l_0}) je $\eta < M$, plyne z nerovnosti (2.37) ($\beta > 0$), že je

$$\begin{aligned} 0 &= (1+2\beta)M - \beta M - \beta M - M < \\ &< (1+2\beta)\eta_{k_0}^{(l_0)} - \beta\eta_{k_0-1}^{(l_0)} - \beta\eta_{k_0+1}^{(l_0)} - \eta_{k_0}^{(l_0-1)} \leqq 0, \end{aligned} \tag{2.38}$$

a to je spor. Je tedy $\eta_{k_0-1}^{(l_0)} = \eta_{k_0+1}^{(l_0)} = \eta_{k_0}^{(l_0-1)} = M$. Pokračujeme-li tímto způsobem dále, dospějeme nutně k uzlu z $\Gamma_{h,\tau}$, ve kterém funkce η nabývá hodnoty M. Lemma je dokázáno.

Právě dokázané lemma ukazuje, že operátor $L_{h,\tau}^{(1)}$ splňuje princip maxima.

Lemma 2.2. *Nechť pro funkci $\eta_k^{(l)}$ definovanou v $\bar{R}_{h,\tau}$ platí rovnice (2.33) a buď*

$$\varepsilon = \max_{(x_k,t_l) \in R_{h,\tau}} |\varepsilon_k^{(l)}|. \tag{2.39}$$

Pak existuje konstanta M (nezávislá na h a τ) taková, že platí

$$|\eta_k^{(l)}| \leqq M\varepsilon, \quad k = 0, \dots, n, \ l = 0, \dots, r. \tag{2.40}$$

Důkaz. Položme $r_k^{(l)} = 1 - \mathrm{e}^{x_k - 1}$ pro $k = 0, \dots, n$ a $l = 0, \dots, r$. Podle Taylorova vzorce je

$$\begin{aligned} (L_{h,\tau}^{(1)} r)_k^{(l)} &= \frac{1}{h^2}(\mathrm{e}^{x_{k-1}-1} - 2\mathrm{e}^{x_k - 1} + \mathrm{e}^{x_{k+1}-1}) = \\ &= \frac{\mathrm{e}^{-1}}{h^2}(\mathrm{e}^{x_{k-1}} - 2\mathrm{e}^{x_k} + \mathrm{e}^{x_{k+1}}) = \\ &= \mathrm{e}^{-1}(\mathrm{e}^{x_k} + \tfrac{1}{12}h^2\mathrm{e}^{\xi_k}) \geqq \mathrm{e}^{-1}. \end{aligned} \tag{2.41}$$

Položme dále

$$\xi_k^{(l)} = -e\varepsilon r_k^{(l)} \pm \eta_k^{(l)}, \quad k = 0, \dots, n, \; l = 0, \dots, r, \tag{2.42}$$

takže je

$$(L_{h,\tau}^{(1)}\xi)_k^{(l)} = -e\varepsilon(L_{h,\tau}^{(1)}r)_k^{(l)} \pm \varepsilon_k^{(l)}. \tag{2.43}$$

Protože podle (2.41) je $e(L_{h,\tau}^{(1)}r)_k^{(l)} \geqq 1$, plyne z rovnice (2.39), že platí

$$-e\varepsilon(L_{h,\tau}^{(1)}r)_k^{(l)} \leqq \varepsilon_k^{(l)} \leqq e\varepsilon(L_{h,\tau}^{(1)}r)_k^{(l)}, \tag{2.44}$$

neboli

$$-e\varepsilon(L_{h,\tau}^{(1)}r)_k^{(l)} \pm \varepsilon_k^{(l)} \leqq 0. \tag{2.45}$$

Platí tedy

$$(L_{h,\tau}^{(1)}\xi)_k^{(l)} \leqq 0 \tag{2.46}$$

a z lemmatu 2.1 dostáváme

$$\xi_k^{(l)} \leqq \max_{(x_k,t_l)\in\Gamma_{h,\tau}} \xi_k^{(l)}. \tag{2.47}$$

Funkce $\xi_k^{(l)}$ je však na množině $\Gamma_{h,\tau}$ zřejmě nekladná, a je tedy nekladná v celé množině $\bar{R}_{h,\tau}$. Protože funkce $r_k^{(l)}$ je v $\bar{R}_{h,\tau}$ omezená, plyne tvrzení lemmatu z rovnice (2.42).

Důkaz následující konvergenční věty pro implicitní metodu už nyní nepředstavuje žádný problém.

Věta 2.2. *Nechť řešení problému* (2.1) až (2.3) *má v* $\bar{R}$ *dvě spojité derivace podle* t *a čtyři spojité derivace podle* x *a nechť* $u_k^{(l)}$ *je přibližné řešení vypočtené implicitní metodou* (2.29). *Pak existuje konstanta* M *taková, že pro dostatečně malé* h *a* τ *platí*

$$|u_k^{(l)} - u(x_k, t_l)| \leqq M(\tau + h^2). \tag{2.48}$$

D ů k a z . Tvrzení věty plyne ihned z rovnice (2.27) a z lemmatu 2.2.

Konvergenci implicitní metody se nám teď podařilo dokázat bez jakýchkoliv doplňujících předpokladů o poměru τ/h^2. Při implicitní metodě proto můžeme volit časový integrační krok τ bez jakéhokoliv ohledu na zvolený prostorový integrační krok, aniž bychom narušili konvergenci. Tuto skutečnost však nemůžeme obecně plně využít. Nechceme-li totiž, aby převládla chyba, která vznikne v důsledku diskretizace proměnné t, je třeba brát časové oko sítě τ řádově tak velké, jako je kvadrát prostorového oka sítě h (viz odhad (2.48)). Splnění této podmínky má však za následek, že při zjemňování sítě musíme stejně jako u explicitní metody při rozpůlení prostorového integračního kroku rozdělit časový integrační krok na

čtyři díly. Vzniká proto přirozená otázka, není-li možné sestrojit takové diferenční schéma pro řešení úlohy (2.1) až (2.3), které by vedlo k chybě řádu $O(\tau^2 + h^2)$ a při kterém by stejně jako při implicitní metodě nebylo třeba brát žádný ohled na h při volbě τ. Konstrukci takového schématu provedeme v následujícím odstavci.

2.1.2 Crankovo-Nicolsonovo schéma

Uvědomíme-li si, že úloha (2.1) až (2.3) má vzhledem k proměnné t charakter úlohy s počáteční podmínkou (na tuto skutečnost jsme už ostatně upozornili v úvodu) a že při řešení obyčejné diferenciální rovnice $y' = f(t, y)$ metodami typu $y_{l+1} = y_l + \tau[\alpha f_{l+1} + (1-\alpha)f_l]$, kde α je číselný parametr, je chyba obecně řádu $O(\tau)$, avšak při $\alpha = 1/2$ řádu $O(\tau^2)$, přirozeně se objeví otázka, zda mezi aproximacemi daného parabolického operátoru tvaru

$$\begin{aligned}(L_{h,\tau}^{(\alpha)}u)_k^{(l)} = {} & \frac{1}{\tau}[u_k^{(l)} - u_k^{(l-1)}] - \frac{\alpha}{h^2}[u_{k-1}^{(l)} - 2u_k^{(l)} + u_{k+1}^{(l)}] - \\ & - \frac{1-\alpha}{h^2}[u_{k-1}^{(l-1)} - 2u_k^{(l-1)} + u_{k+1}^{(l-1)}], \quad 0 \leqq \alpha \leqq 1,\end{aligned} \tag{2.49}$$

nemá hodnota $1/2$ parametru α opět výlučné postavení. Následující věta ukazuje, že tomu tak skutečně je.

Věta 2.3. *Nechť funkce* u *má v* $\bar{R}$ *tři spojité derivace podle* t *a čtyři spojité derivace podle* x *a nechť* $u^{(\mathrm{pr})}$ *je funkce definovaná na síti* $\bar{R}_{h,\tau}$ *předpisem* $(u^{(\mathrm{pr})})_k^{(l)} = u(x_k, t_l)$. *Pak pro každý uzel* $(x_k, t_l) \in R_{h,\tau}$ *platí*

$$(L_{h,\tau}^{(\alpha)}u^{(\mathrm{pr})})_k^{(l)} = \alpha(Lu)(x_k, t_l) + (1-\alpha)(Lu)(x_k, t_{l-1}) + O(\tau + h^2) \tag{2.50}$$

při obecném α *a*

$$(L_{h,\tau}^{1/2}u^{(\mathrm{pr})})_k^{(l)} = \tfrac{1}{2}(Lu)(x_k, t_l) + \tfrac{1}{2}(Lu)(x_k, t_{l-1}) + O(\tau^2 + h^2). \tag{2.51}$$

D ů k a z . Pomocí Taylorova vzorce snadno vypočteme, že platí

$$\begin{aligned}&(L_{h,\tau}^{(\alpha)}u^{(\mathrm{pr})})_k^{(l)} = \\ &= \alpha\Big[\frac{\partial u}{\partial t}(x_k, t_l) - \frac{\partial^2 u}{\partial x^2}(x_k, t_l) - \tfrac{1}{2}\tau\frac{\partial^2 u}{\partial t^2}(x_k, t_l) + O(\tau^2 + h^2)\Big] + \\ &\quad + (1-\alpha)\Big[\frac{\partial u}{\partial t}(x_k, t_{l-1}) - \frac{\partial^2 u}{\partial x^2}(x_k, t_{l-1}) + \tfrac{1}{2}\tau\frac{\partial^2 u}{\partial t^2}(x_k, t_{l-1}) + \\ &\quad + O(\tau^2 + h^2)\Big] = \\ &= \alpha(Lu)(x_k, t_l) + (1-\alpha)(Lu)(x_k, t_{l-1}) - \\ &\quad - \tfrac{1}{2}\tau\Big[\alpha\frac{\partial^2 u}{\partial t^2}(x_k, t_l) - (1-\alpha)\frac{\partial^2 u}{\partial t^2}(x_k, t_{l-1})\Big] + O(\tau^2 + h^2).\end{aligned} \tag{2.52}$$

Z rovnice (2.52) však už tvrzení věty plyne snadno.

Z důkazu věty 2.3 je zřejmé, že v případě $\alpha \neq 1/2$ není třeba předpokládat spojitost třetí derivace podle t, ale stačí pouze spojitost druhé derivace, stejně jako tomu bylo u explicitní a implicitní metody.

Přibližné řešení $u_k^{(l)}$ budeme nyní hledat z rovnice

$$(L_{h,\tau}^{(\alpha)} u)_k^{(l)} = 0, \quad k = 1, \dots, n-1, \ l = 1, \dots, r, \tag{2.53}$$

s počáteční podmínkou (2.7) a okrajovými podmínkami (2.8). Použijeme-li už vícekrát užitou vektorovou symboliku, můžeme rovnice (2.53) spolu s příslušnými okrajovými podmínkami zapsat ve tvaru

$$A_{\mathrm{U}} u^{(l)} = A_{\mathrm{L}} u^{(l-1)}, \quad l = 1, \dots, r, \tag{2.54}$$

kde $u^{(0)}$ je daný vektor a matice A_{U} a A_{L} jsou třídiagonální matice dané vzorci

$$A_{\mathrm{U}} = \begin{bmatrix} 1+2\alpha\beta & -\alpha\beta & 0 & \dots & 0 \\ -\alpha\beta & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\alpha\beta \\ 0 & \dots & 0 & -\alpha\beta & 1+2\alpha\beta \end{bmatrix} \tag{2.55}$$

a

$$A_{\mathrm{L}} = \begin{bmatrix} 1-2(1-\alpha)\beta & (1-\alpha)\beta & 0 & \dots & 0 \\ (1-\alpha)\beta & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & (1-\alpha)\beta \\ 0 & \dots & 0 & (1-\alpha)\beta & 1-2(1-\alpha)\beta \end{bmatrix}. \tag{2.56}$$

Matice A_{U} je zřejmě regulární, neboť je podle Collatzova lemmatu (viz lemma 3.5 z kap. II) monotónní, takže rovnice (2.53) s příslušnými počátečními a okrajovými podmínkami definují skutečně metodu. Při $\alpha \neq 0$ je tato metoda implicitní v tom smyslu, že k tomu, abychom vypočetli přibližné řešení na nějakém časovém řádku, je třeba řešit soustavu lineárních algebraických rovnic.

Jak je to s konvergencí zavedené třídy metod? V případě $\alpha = 0$ byla situace velice jednoduchá a konvergence této metody byla takřka zřejmá (ovšem za splnění podmínky (2.17)). V případě $\alpha = 1$ jsme rovněž bez podstatných obtíží dokázali konvergenci užitím lemmatu 2.2, tj. pomocí principu maxima. Tento princip ostatně platil i v případě $\alpha = 0$. V obecném případě lze sice také dosáhnout platnosti principu maxima, je však k tomu třeba klást na poměr $\beta = \tau/h^2$ nepřirozené požadavky. Proto budeme postupovat poněkud jinak, abychom se vyhnuli nutnosti užití principu maxima.

Pro chybu $\eta_k^{(l)}$ všech uvažovaných metod platí zřejmě

$$(L_{h,\tau}^{(\alpha)} \eta)_k^{(l)} = \varepsilon_k^{(l)}, \quad k = 1, \dots, n-1, \ l = 1, \dots, r, \tag{2.57}$$

s počáteční podmínkou

$$\eta_k^{(0)} = 0, \quad k = 0, \dots, n, \tag{2.58}$$

a okrajovými podmínkami

$$\eta_0^{(l)} = \eta_n^{(l)} = 0, \quad l = 1, \dots, r. \tag{2.59}$$

Přepíšeme-li tyto rovnice maticově, dostaneme

$$\begin{aligned} A_{\mathrm{U}} \eta^{(l)} &= A_{\mathrm{L}} \eta^{(l-1)} + \tau \epsilon^{(l)}, \quad l = 1, \dots, r, \\ \eta^{(0)} &= 0, \end{aligned} \tag{2.60}$$

kde

$$\eta^{(l)} = (\eta_1^{(l)}, \dots, \eta_{n-1}^{(l)})^T \tag{2.61}$$

a

$$\epsilon^{(l)} = (\varepsilon_1^{(l)}, \dots, \varepsilon_{n-1}^{(l)})^T \tag{2.62}$$

a matice A_{U} a A_{L} jsou definované rovnicemi (2.55) a (2.56). Z věty 2.3 víme, že pravé strany rovnic (2.57) jsou (za předpokladu dostatečné hladkosti přesného řešení) malé. Rádi bychom usoudili, že tento fakt má za následek i malost příslušného řešení. Je tedy třeba dokázat, že řešení uvažovaných diferenčních rovnic spojitě závisí na pravých stranách, neboli, užijeme-li terminologie z odst. 2.4 kap. III, že uvažované diferenční rovnice jsou stabilní vzhledem k pravým stranám. Pro pohodlí čtenáře vyslovíme zde nezbytné definice specializované na náš případ ještě jednou.

Definice 2.1. Řekneme, že diferenční schéma dané operátorem $L_{h,\tau}^{(\alpha)}$ je *stabilní vzhledem k pravé straně*, existuje-li konstanta M (nezávislá na h a τ) taková, že pro každé řešení $\eta^{(l)}$ rovnice (2.60), pro které je $\|\epsilon^{(l)}\| \leqq \varepsilon$ pro $l = 1, \dots, r$, platí

$$\|\eta^{(l)}\| \leqq M\varepsilon. \tag{2.63}$$

Zde se rozumí, že symbol $\|.\|$ značí nějakou vhodnou pevně zvolenou normu v $(n-1)$-dimenzionálním vektorovém prostoru. Např. při vyšetřování implicitní metody jsme vlastně konstruovali odhad typu (2.63) na základě principu maxima a za vektorovou normu jsme přitom brali $\mathscr{L}_\infty$ normu (tj. $\|x\| = \max_i |x_i|$).

Z definice 2.1 je vidět, že je-li dané schéma stabilní vzhledem k pravé straně a jsme-li schopni odhadnout lokální chyby — to není většinou velký problém — je metoda konvergentní. Stabilita vzhledem k pravé straně se však neověřuje vždy pohodlně, a proto zavedeme ještě jeden pojem a zformulujeme další pomocné tvrzení, které její studium podstatně usnadní.

Definice 2.2. Řekneme, že diferenční schéma dané operátorem $L_{h,\tau}^{(\alpha)}$ je *stejnoměrně stabilní vzhledem k počátečním podmínkám*, existuje-li konstanta M (nezávislá na h a τ) taková, že pro každé celé s z intervalu $\langle 0, r\rangle$ a pro každou soustavu vektorů $\boldsymbol{\eta}^{(l)}$, $l = s, s+1, \ldots, r$, která je řešením rovnic

$$A_{\mathrm{U}}\boldsymbol{\eta}^{(l)} = A_{\mathrm{L}}\boldsymbol{\eta}^{(l-1)} \tag{2.64}$$

pro $l = s+1, \ldots, r$, platí

$$\|\boldsymbol{\eta}^{(l)}\| \leqq M\|\boldsymbol{\eta}^{(s)}\|, \quad l = s, \ldots, r. \tag{2.65}$$

Lemma 2.3. *Nechť pro diferenční schéma dané operátorem $L_{h,\tau}^{(\alpha)}$ platí*
(i) *existuje konstanta M (nezávislá na h a τ) taková, že je $\|A_{\mathrm{U}}^{-1}\| \leqq M$;*
(ii) *schéma je stejnoměrně stabilní vzhledem k počátečním podmínkám.*
Pak schéma je stabilní vzhledem k pravé straně.

D ů k a z . Buď $\boldsymbol{\eta}^{(l)}$ řešením rovnic (2.60) a buď $\|\boldsymbol{\epsilon}^{(l)}\| \leqq \varepsilon$ pro $l = 1, \ldots, r$. Přiřaďme každému celému číslu m, $0 \leqq m \leqq r$, soustavu vektorů $\boldsymbol{\xi}^{(l)}(m)$, $l = 0, \ldots, r$, takto: $\boldsymbol{\xi}^{(l)}(m) = \boldsymbol{0}$ pro $l = 0, \ldots, m$ a pro $l = m+1, \ldots, r$ jsou vektory $\boldsymbol{\xi}^{(l)}(m)$ řešením rovnic

$$A_{\mathrm{U}}\boldsymbol{\xi}^{(l)}(m) = A_{\mathrm{L}}\boldsymbol{\xi}^{(l-1)}(m) + \tau\boldsymbol{\epsilon}^{(l)}. \tag{2.66}$$

Zvolme nyní pevně m, $1 \leqq m \leqq r$ a položme

$$\boldsymbol{v}^{(l)} = \boldsymbol{\xi}^{(l)}(m-1) - \boldsymbol{\xi}^{(l)}(m). \tag{2.67}$$

Předně je $\boldsymbol{v}^{(l)} = \boldsymbol{0}$ pro $l = 0, \ldots, m-1$ a $\boldsymbol{v}^{(m)} = \boldsymbol{\xi}^{(m)}(m-1)$. Dále je

$$A_{\mathrm{U}}\boldsymbol{v}^{(l)} = A_{\mathrm{L}}\boldsymbol{v}^{(l-1)} \tag{2.68}$$

pro $l = m+1, \ldots, r$. V důsledku stejnoměrné stability vzhledem k počátečním podmínkám existuje konstanta M_1, taková, že platí

$$\|\boldsymbol{v}^{(l)}\| \leqq M_1\|\boldsymbol{v}^{(m)}\| \tag{2.69}$$

pro $l = m \ldots, r$. Protože však pro $l < m$ je $\boldsymbol{v}^{(l)} = \boldsymbol{0}$, platí vztah (2.69) pro $l = 0, \ldots, r$. Platí tedy

$$\|\boldsymbol{\xi}^{(l)}(m-1) - \boldsymbol{\xi}^{(l)}(m)\| \leqq M_1\|\boldsymbol{\xi}^{(m)}(m-1)\| \tag{2.70}$$

pro $l = 0, \ldots, r$. Odhadněme nyní normu vektoru $\boldsymbol{\xi}^{(m)}(m-1)$. Tento vektor je podle své definice řešením rovnice

$$A_{\mathrm{U}}\boldsymbol{\xi}^{(m)}(m-1) = A_{\mathrm{L}}\boldsymbol{\xi}^{(m-1)}(m-1) + \tau\boldsymbol{\epsilon}^{(m)}. \tag{2.71}$$

Protože však vektor $\boldsymbol{\xi}^{(m-1)}(m-1)$ je nulový, je podle předpokladu (i)

$$\|\boldsymbol{\xi}^{(m)}(m-1)\| \leqq \tau M\varepsilon. \tag{2.72}$$

Odtud a z nerovnosti (2.70) tedy dostáváme, že platí

$$\|\boldsymbol{\xi}^{(l)}(m-1) - \boldsymbol{\xi}^{(l)}(m)\| \leqq \tau M M_1\varepsilon \tag{2.73}$$

pro $l = 0, \ldots, r$. Pišme nerovnost (2.73) postupně pro $m = 1, \ldots, r$; dostaneme

$$\begin{aligned} &\|\boldsymbol{\xi}^{(l)}(0) - \boldsymbol{\xi}^{(l)}(1)\| \leqq \tau M M_1\varepsilon, \\ &\|\boldsymbol{\xi}^{(l)}(1) - \boldsymbol{\xi}^{(l)}(2)\| \leqq \tau M M_1\varepsilon, \\ &\qquad\vdots \\ &\|\boldsymbol{\xi}^{(l)}(r-1) - \boldsymbol{\xi}^{(l)}(r)\| \leqq \tau M M_1\varepsilon. \end{aligned} \tag{2.74}$$

Spojením těchto nerovností máme

$$\|\boldsymbol{\xi}^{(l)}(0) - \boldsymbol{\xi}^{(l)}(r)\| \leqq r\tau M M_1\varepsilon = T M M_1\varepsilon \tag{2.75}$$

pro $l = 0, \ldots, r$. Vektor $\boldsymbol{\xi}^{(l)}(r)$ je však pro $l = 0, \ldots, r$ nulový vektor a $\boldsymbol{\xi}^{(l)}(0)$ splňuje tytéž rovnice jako vektor $\boldsymbol{\eta}^{(l)}$. Je tedy $\boldsymbol{\xi}^{(l)}(0) = \boldsymbol{\eta}^{(l)}$ pro $l = 0, \ldots, r$ a nerovnosti (2.75) dokazují lemma.

Upozorněme, že v definicích 2.1 a 2.2 a v právě dokázaném lemmatu není nikterak podstatné, že se konkrétně jednalo o operátor $L_{h,\tau}^{(\alpha)}$ daný rovnicemi (2.49). Příslušné pojmy lze zřejmě zavést pro libovolný diferenční operátor, jen vede-li na soustavu rovnic, kterou je možné psát ve tvaru (2.54).

Chceme-li dokázat konvergenci studované metody, stačí v důsledku lemmatu 2.3 prověřit jeho předpoklady. Protože jsme se rozhodli, že se chceme vyhnout principu maxima, užijeme tzv. *metodu separace proměnných*, zvanou také *Fourierova metoda*. Začneme zavedením vhodné normy $(n-1)$-dimenzionálních vektorů, které v našich úvahách vystupují. Vzhledem k postupu, který chceme užít, je přirozené požadovat, jak uvidíme, aby uvažovaný $(n-1)$-dimenzionální vektorový prostor byl Hilbertův prostor. Normu budeme proto definovat pomocí skalárního součinu. Obyčejný euklidovský skalární součin zde není vhodný, neboť složky vektorů, s nimiž pracujeme, aproximují hodnoty funkcí definovaných ve všech bodech intervalu $(0,1)$, a je tedy žádoucí, aby námi zavedený skalární součin konvergoval pro $h \to 0$ k $\mathscr{L}_2(0,1)$-skalárnímu součinu (srv. požadavky (2.175) z kap. III.). Tuto vlastnost má jednoduchá modifikace euklidovského skalárního součinu, kdy za skalární součin vektorů $\boldsymbol{v}$, $\boldsymbol{w}$ bereme číslo $(\boldsymbol{v}, \boldsymbol{w})_h$ dané vzorcem

$$(\boldsymbol{v}, \boldsymbol{w})_h = h\sum_{k=1}^{n-1} v_k w_k. \tag{2.76}$$

Norma vektoru $\boldsymbol{v}$ je tedy dána výrazem

$$\|\boldsymbol{v}\|_h = \left(h\sum_{k=1}^{n-1} v_k^2\right)^{1/2}. \tag{2.77}$$

Označme ještě vektorový prostor opatřený normou (2.77) symbolem $\mathbf{E}_h^{(n-1)}$.

Stabilita diferenčních schémat $L_{h,\tau}^{(\alpha)}$ v právě zavedené normě je popsána v následující větě.

Věta 2.4. *Buď dán diferenční operátor* $L_{h,\tau}^{(\alpha)}$ *s* $0 \leqq \alpha \leqq 1$ *a nechť v případě, že je* $0 \leqq \alpha < 1/2$ *platí*

$$\beta = \frac{\tau}{h^2} \leqq \frac{1}{2(1-2\alpha)}. \tag{2.78}$$

Pak diferenční schéma dané tímto operátorem je stejnoměrně stabilní vzhledem k počátečním podmínkám.

D ů k a z . Pišme matice A_U a A_L dané vzorci (2.55) a (2.56) ve tvaru

$$A_U = I + \alpha\beta P_0 \tag{2.79}$$

a

$$A_L = I - (1-\alpha)\beta P_0, \tag{2.80}$$

kde I je jednotková matice a třídiagonální matice P_0 řádu $n-1$ je dána vzorcem

$$P_0 = \begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 2 \end{bmatrix}. \tag{2.81}$$

Snadno zjistíme, že vektory $v^{(\nu)}$, $\nu = 1, \dots, n-1$, jejichž složky jsou dány rovnicemi (2.16), tvoří ortogonální soustavu vlastních vektorů matice (2.81), a to jak v obyčejném euklidovském, prostoru, tak v prostoru $\mathsf{E}_h^{(n-1)}$, a že příslušná vlastní čísla jsou

$$\lambda_\nu = 4\sin^2\frac{\nu\pi}{2n}, \quad \nu = 1, \dots, n-1. \tag{2.82}$$

Vzhledem k rovnicím (2.79) a (2.80) jsou vektory $v^{(\nu)}$ také vlastními vektory matice A_U, resp. A_L s odpovídajícími vlastními čísly

$$\lambda_\nu^{(U)} = 1 + \alpha\beta\lambda_\nu, \quad \nu = 1, \dots, n-1, \tag{2.83}$$

resp.

$$\lambda_\nu^{(L)} = 1 - (1-\alpha)\beta\lambda_\nu, \quad \nu = 1, \dots, n-1. \tag{2.84}$$

Abychom dokázali tvrzení věty, je třeba odhadnout řešení soustavy (2.64) pomocí počáteční podmínky. Vzhledem k právě zjištěné skutečnosti, že matice A_U a A_L mají tytéž vlastní vektory, se pokusíme splnit rovnici (2.64) soustavou vektorů $z^{(l)}$ tvaru

$$z^{(l)} = c(l)v^{(\nu)}, \tag{2.85}$$

kde $c(l)$ je skalární veličina závisející pouze na indexu l. Funkci $\eta_k^{(l)}$, která řeší rovnici (2.64), se tedy snažíme nalézt ve tvaru součinu dvou funkcí, z nichž jedna je funkcí pouze proměnné l a druhá funkcí pouze proměnné k. Odtud také pochází název metoda separace proměnných. Má tedy platit rovnice

$$c(l)A_U v^{(\nu)} = c(l-1)A_L v^{(\nu)}. \tag{2.86}$$

Odtud plyne, že funkce $c(l)$ musí splňovat rovnici

$$c(l) = \frac{\lambda_\nu^{(L)}}{\lambda_\nu^{(U)}} c(l-1). \tag{2.87}$$

Každá soustava vektorů $z^{(l)}$ tvaru

$$z^{(l)} = c\left[\frac{\lambda_\nu^{(L)}}{\lambda_\nu^{(U)}}\right]^{l+\delta} v^{(\nu)}, \tag{2.88}$$

kde c je libovolná konstanta a δ libovolné celé číslo, tedy řeší rovnici (2.64).

Vzhledem k tomu, že vektory $v^{(\nu)}$ tvoří ortogonální bázi v prostoru $\mathsf{E}_h^{(n-1)}$, existují k libovolnému vektoru $\eta = \eta^{(s)} \in \mathsf{E}_h^{(n-1)}$ čísla $c_1, \dots, c_{n-1}$ taková, že platí

$$\eta^{(s)} = \sum_{\nu=1}^{n-1} c_\nu v^{(\nu)}. \tag{2.89}$$

Z ortogonality vektorů $v^{(\nu)}$ přitom plyne, že je

$$\|\eta^{(s)}\|_h^2 = \sum_{\nu=1}^{n-1} c_\nu^2 \|v^{(\nu)}\|_h^2. \tag{2.90}$$

Z rovnic (2.89) a (2.88) tedy dostáváme, že řešení soustavy (2.64) určené počáteční podmínkou $\eta^{(s)}$ lze psát ve tvaru

$$\eta^{(l)} = \sum_{\nu=1}^{n-1} c_\nu \left[\frac{\lambda_\nu^{(L)}}{\lambda_\nu^{(U)}}\right]^{l-s} v^{(\nu)}. \tag{2.91}$$

Přitom platí

$$\|\eta^{(l)}\|_h^2 = \sum_{\nu=1}^{n-1} c_\nu^2 \left[\frac{\lambda_\nu^{(L)}}{\lambda_\nu^{(U)}}\right]^{2l-2s} \|v^{(\nu)}\|_h^2, \tag{2.92}$$

jak plyne opět z ortogonality vektorů $v^{(\nu)}$. Podle rovnic (2.83) a (2.84) je

$$\frac{\lambda_\nu^{(L)}}{\lambda_\nu^{(U)}} = \frac{1-(1-\alpha)\beta\lambda_\nu}{1+\alpha\beta\lambda_\nu}. \tag{2.93}$$

Platí-li

$$-1 \leqq \frac{1-(1-\alpha)\beta\lambda_\nu}{1+\alpha\beta\lambda_\nu} \leqq 1, \tag{2.94}$$

je

$$\|\eta^{(l)}\|_h \leqq \|\eta^{(s)}\|_h, \tag{2.95}$$

jak plyne ihned ze vzorců (2.92) a (2.90). Splnění nerovností (2.94) tedy stačí ke stabilitě. Pravá z těchto nerovností je splněna vždy, neboť je $0 \leqq \alpha \leqq 1$, $\beta > 0$ a $\lambda_\nu > 0$. Je-li $\alpha \geqq 1/2$, je splněna i levá z těchto nerovností bez dalších doplňujících podmínek, neboť v tomto případě je $(1-2\alpha)\beta\lambda_\nu \leqq 0$, a tedy tím spíše platí nerovnost

$$(1-2\alpha)\beta\lambda_\nu < 2 \tag{2.96}$$

neboli

$$-1 - \alpha\beta\lambda_\nu < 1 - (1-\alpha)\beta\lambda_\nu. \tag{2.97}$$

Je-li však $1 \leqq \alpha < 1/2$ a platí-li navíc nerovnost (2.78), platí opět (2.96), a tedy také (2.97), neboť je $\lambda_\nu < 4$, jak plyne ihned z rovnic (2.82). Tím je důkaz věty zakončen.

Diferenční schémata $L_{h,\tau}^{(\alpha)}$ jsou tedy při $1/2 \leqq \alpha \leqq 1$ absolutně stabilní. Vzhledem k tomu, že pro velká n je jedno z vlastních čísel matice přechodu $A_U^{-1}A_L$ velice blízké číslu $[1-4(1-\alpha)\beta]/(1+4\alpha\beta)$ a vzhledem k tomu, že toto číslo je v případě $0 \leqq \alpha < 1/2$ a při porušení podmínky (2.78) menší než -1, je tato podmínka i nutnou podmínkou stability. Uvažovaná diferenční schémata jsou proto při $0 \leqq \alpha < 1/2$ pouze relativně stabilní a nerovnost (2.78) je jejich podmínkou stability.

Věta 2.4 spolu s lemmatem 2.3 dovolují už snadno dokázat základní konvergenční větu pro uvažovanou třídu schémat.

Věta 2.5. *Nechť řešení okrajové úlohy* (2.1) – (2.3) *má v* $\bar{R}$ *dvě spojité derivace podle* t *a čtyři spojité derivace podle* x. *Nechť v případě* $\alpha = 1/2$ *má řešení navíc spojitou třetí derivaci podle* t. *Nechť konečně při* $0 \leqq \alpha < 1/2$ *je splněna nerovnost* (2.78). *Pak existuje konstanta* M *taková že pro chybu* $\eta_k^{(l)} = u_k^{(l)} - u(x_k, t_l)$ *přibližného řešení spočítaného schématem* $L_{h,\tau}^{(\alpha)}$ *platí*

$$\|\eta^{(l)}\|_h \leqq M(\tau + h^2) \tag{2.98}$$

v obecném případě a

$$\|\eta^{(l)}\|_h \leqq M(\tau^2 + h^2) \tag{2.99}$$

při $\alpha = 1/2$.

Důkaz. Za uvedených předpokladů splňuje chyba $\eta_k^{(l)}$ rovnice (2.60) a přitom je

$$|\varepsilon_k^{(l)}| \leqq M(\tau + h^2) \tag{2.100}$$

při $\alpha \neq 1/2$ a

$$|\varepsilon_k^{(l)}| \leqq M(\tau^2 + h^2) \tag{2.101}$$

při $\alpha = 1/2$, jak plyne ihned z věty 2.3. Matice A_U je symetrická a její vlastní čísla jsou dána vzorcem (2.83). Vzhledem k tomu, že $\|A_U^{-1}\|_h$ je zřejmě obyčejná spektrální norma, platí

$$\|A_U^{-1}\|_h = \frac{1}{\lambda_{\min}^{(U)}} \leqq 1. \tag{2.102}$$

Protože schéma $L_{h,\tau}^{(\alpha)}$ je podle věty 2.4 zároveň stejnoměrně stabilní vzhledem k počátečním podmínkám, je podle lemmatu 2.3 stabilní vzhledem k pravé straně, a zbývá tedy už jen odhadnout velikost normy vektorů $\boldsymbol{\epsilon}^{(l)}$. Z nerovností (2.100) a (2.101) však ihned plyne, že je

$$\|\epsilon^{(l)}\|_h^2 = h\sum_{k=1}^{n-1}[\varepsilon_k^{(l)}]^2 \leqq \tag{2.103}$$

$$\leqq \begin{cases} M^2(\tau+h^2)^2 h\sum_{k=1}^{n-1} 1 \leqq M^2(\tau+h^2)^2, & \alpha \neq 1/2, \\ M^2(\tau^2+h^2)^2 h\sum_{k=1}^{n-1} 1 \leqq M^2(\tau^2+h^2)^2, & \alpha = 1/2. \end{cases}$$

Důkaz je hotov.

Výsledek, který jsme dostali, je tedy poněkud slabší, než tomu bylo v případě explicitní a čistě implicitní metody, a to v tom smyslu, že jsme užili horší normu než normu užitou v odst. 2.1.1. Toto zhoršení není příliš podstatné a navíc je lze důmyslnějším postupem, jak uvidíme, odstranit. Hlavního výsledku, tj. sestrojení absolutně stabilního diferenčních schématu s celkovou chybou řádu $O(\tau^2 + h^2)$ jsme však dosáhli. Takové schéma je dáno operátorem $L_{h,\tau}^{(1/2)}$ a příslušná metoda se nazývá *Crankova-Nicolsonova metoda*. Tato metoda je velmi populární a v praxi se často užívá. Důvod je ten, že její užití je podstatně ekonomičtější, než je tomu např. u explicitní metody. Čtenáře možná toto tvrzení na první pohled poněkud překvapí, neboť při užití Crankova-Nicolsonova schématu je třeba navíc ve srovnání s explicitní metodou řešit v každém kroku soustavu lineárních algebraických rovnic. Tato soustava má však třídiagonální matici, takže počet potřebných operací je řádově roven počtu neznámých, tj. veličině $O(1/h)$. Protože celková diskretizační chyba je $O(\tau^2 + h^2)$, je rozumné volit $\tau = O(h)$. Počet potřebných časových řádku je tedy $O(1/h)$ a celkový počet operací je řádově roven $O(1/h^2)$, což je pro malá h podstatně méně než u explicitní metody, kde je toto číslo rovno $O(1/h^3)$.

Zakončeme tento odstavec ještě jedním postřehem. Z toho, co jsme až dosud uvedli, by mohl vzniknout dojem, že všechna schémata pro řešení rovnice pro vedení tepla, která jsou sestavena „přirozeným" způsobem, jsou buď absolutně stabilní, nebo v nejhorším případě relativně stabilní. Tak tomu však není. Abychom to uká-

zali, uvažujme schéma

$$\frac{u_k^{(l+1)} - u_k^{(l-1)}}{2\tau} - \frac{u_{k-1}^{(l)} - 2u_k^{(l)} + u_{k+1}^{(l)}}{h^2} = 0, \tag{2.104}$$
$$k = 1, \dots, n-1, \; l = 1, \dots, r-1,$$

které vznikne tak, že derivaci $\partial u/\partial t$ v uzlu (x_k, t_l) nahradíme podílem $[u(x_k, t_{l+1}) - u(x_k, t_{l-1})]/(2\tau)$, který ji aproximuje s přesností $O(\tau^2)$. Schéma (2.104) je tedy odvozeno zcela v duchu metody sítí a jeho lokální chyba je jako u Crankova-Nicolsonova schématu řádu $O(\tau^2 + h^2)$. S Crankovým-Nicolsonovým schématem dokonce souvisí ještě těsněji, neboť nahradíme-li v něm veličiny $u_k^{(l)}$ průměry $[u_k^{(l+1)} + u_k^{(l-1)}]/2$ a užijeme-li je jen v sudých časových řádcích, dostaneme přímo Crankovo-Nicolsonovo schéma. V podobě, jak je schéma (2.104) zapsáno, je to však schéma třívrstvové, takže je nutné k němu připojit kromě okrajových podmínek (2.8) a počáteční podmínky (2.7) ještě další počáteční podmínku v čase $t = \tau$, kterou je třeba vypočítat jiným vhodným způsobem. Rovnice (2.104) s příslušnými okrajovými podmínkami lze zapsat maticově ve tvaru

$$u^{(l+1)} = u^{(l-1)} - 2\beta P_0 u^{(l)}, \tag{2.105}$$

kde P_0 je matice daná vzorcem (2.81). Zavedeme-li $(2n-2)$-dimenzionální vektory $v^{(l)}$ rovnicí

$$v^{(l)} = \begin{bmatrix} u^{(l)} \\ u^{(l+1)} \end{bmatrix}, \tag{2.106}$$

lze třívrstvové schéma (2.105) psát jako dvouvrstvové schéma

$$v^{(l)} = Rv^{(l-1)}, \quad l = 1, \dots, r-1, \tag{2.107}$$

kde matice R je řádu $2n-2$ a je daná vzorcem

$$R = \begin{bmatrix} 0 & I \\ I & -2\beta P_0 \end{bmatrix}. \tag{2.108}$$

Přímým výpočtem se snadno zjistí, že všechna vlastní čísla matice R se určí řešením $n-1$ kvadratických rovnic $\xi^2 + 2\beta\lambda_\nu \xi - 1 = 0$, $\nu = 1, \dots, n-1$, kde čísla λ_ν jsou vlastní čísla matice P_0, a jsou tedy dána vzorci (2.82). Z uvedených rovnic je však vidět, že ať zvolíme síť jakkoliv, vždy existují vlastní čísla matice R, která jsou v absolutní hodnotě větší než 1. Schéma (2.104) je tedy absolutně nestabilní.

2.2 Obecná parabolická rovnice v jedné prostorové proměnné

V tomto odstavci se budeme zabývat metodou sítí pro řešení diferenciální rovnice

$$Lu = f(x,t) \quad \text{v } R, \tag{2.109}$$

kde R je opět množina $\{(x,t);\, 0 < x < 1,\, 0 < t < T\}$ a

$$Lu = c(x,t)\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(p(x,t)\frac{\partial u}{\partial x}\right) + q(x,t)u, \tag{2.110}$$

s počáteční podmínkou

$$u(x,0) = g(x), \quad 0 < x < 1, \tag{2.111}$$

a okrajovými podmínkami

$$u(0,t) = \gamma^{(0)}(t), \quad u(1,t) = \gamma^{(1)}(t), \quad 0 < t < T, \tag{2.112}$$

(Dirichletova nebo první okrajová úloha) nebo

$$(l^{(0)}u)(t) = \gamma^{(0)}(t), \quad (l^{(1)}u)(t) = \gamma^{(1)}(t), \quad 0 < t < T, \tag{2.113}$$

kde

$$(l^{(0)}u)(t) = -p(0,t)\frac{\partial u}{\partial x}(0,t) + \beta^{(0)}(t)u(0,t), \tag{2.114}$$
$$(l^{(1)}u)(t) = p(1,t)\frac{\partial u}{\partial x}(1,t) + \beta^{(1)}(t)u(1,t).$$

Funkce a, p, q, f, $\beta^{(0)}$, $\beta^{(1)}$, $\gamma^{(0)}$, $\gamma^{(1)}$ a g jsou dány a předpokládáme, že jsou spojité a že funkce p má navíc spojitou derivaci v $\bar{R}$. Kromě toho předpokládáme, že existují konstanty p_0, p_1, c_0, c_1, q_1, β_0, β_1 takové, že platí

$$0 < p_0 \leqq p(x,t) \leqq p_1, \quad 0 < c_0 \leqq c(x,t) \leqq c_1, \quad 0 \leqq q(x,t) \leqq q_1 \tag{2.115}$$

a že je buď $\beta^{(i)}(t) \equiv 0$ pro $i = 1, 2$ (v tom případě mluvíme o Neumannově nebo druhé okrajové úloze) nebo

$$0 < \beta_0 \leqq \beta^{(i)}(t) \leqq \beta_1 \tag{2.116}$$

(v tomto případě hovoříme o Newtonově nebo třetí okrajové úloze). Uvedené předpoklady jsou příkladem podmínek, které zaručují jednoznačnou řešitelnost výše zmíněných úloh.

Následující odstavec věnujeme sestavení příslušných diferenčních operátorů.

2.2.1 Odvození metody

Na základě výsledků, k nimž jsme dospěli při studiu metody sítí pro okrajové úlohy pro obyčejné diferenciální rovnice v druhé kapitole a na základě myšlenky studovat ne jednu izolovanou metodu, ale celou třídu metod závislých na parametru $\alpha \in \langle 0, 1\rangle$, je přirozené aproximovat diferenciální operátory L a $l^{(i)}$ operátory $L_{h,\tau}^{(\alpha)}$ a $l_{h,\tau}^{(i,\alpha)}$, které funkci u definované na množině $\bar{R}_{h,\tau}$ (jejíž význam je stejný jako

v odst. 2.1.1) přiřazují funkci definovanou na $R_{h,\tau}$, resp. na množině $\{l;\ l = 1, \dots, r\}$ a které jsou dány vzorci

$$
\begin{aligned}
(2.117)\quad (L_{h,\tau}^{(\alpha)}u)_k^{(l)} =& [\alpha c(x_k,t_l) + (1-\alpha)c(x_k,t_{l-1})]\frac{u_k^{(l)} - u_k^{(l-1)}}{\tau} - \\
&- \frac{\alpha}{h^2}\{p(x_k - h/2, t_l)u_{k-1}^{(l)} - [p(x_k - h/2, t_l) + \\
&+ p(x_k + h/2, t_l)]u_k^{(l)} + p(x_k + h/2, t_l)u_{k+1}^{(l)}\} - \\
&- \frac{1-\alpha}{h^2}\{p(x_k - h/2, t_{l-1})u_{k-1}^{(l-1)} - [p(x_k - h/2, t_{l-1}) + \\
&+ p(x_k + h/2, t_{l-1})]u_k^{(l-1)} + p(x_k + h/2, t_{l-1})u_{k+1}^{(l-1)}\} + \\
&+ \alpha q(x_k, t_l)u_k^{(l)} + (1-\alpha)q(x_k, t_{l-1})u_k^{(l-1)}, \quad (x_k, t_l) \in R_{h,\tau}
\end{aligned}
$$

a

$$
\begin{aligned}
(2.118)\quad (l_{h,\tau}^{(0,\alpha)}u)^{(l)} =& \tfrac{1}{2}h[\alpha c(x_0,t_l) + (1-\alpha)c(x_0,t_{l-1})]\frac{u_0^{(l)} - u_0^{(l-1)}}{\tau} - \\
&- \Big[p(x_0 + h/2, t_l)\frac{u_1^{(l)} - u_0^{(l)}}{h} - \beta^{(0)}(t_l)u_0^{(l)}\Big] - \\
&- (1-\alpha)\Big[p(x_0 + h/2, t_{l-1})\frac{u_1^{(l-1)} - u_0^{(l-1)}}{h} - \\
&- \beta^{(0)}(t_{l-1})u_0^{(l-1)}\Big] + \\
&+ \tfrac{1}{2}h\alpha q(x_0,t_l)u_0^{(l)} + \tfrac{1}{2}h(1-\alpha)q(x_0,t_{l-1})u_0^{(l-1)},
\end{aligned}
$$

$$
\begin{aligned}
(l_{h,\tau}^{(1,\alpha)}u)^{(l)} =& \tfrac{1}{2}h[\alpha c(x_n,t_l) + (1-\alpha)c(x_n,t_{l-1})]\frac{u_n^{(l)} - u_n^{(l-1)}}{\tau} + \\
&+ \alpha\Big[p(x_n - h/2, t_l)\frac{u_n^{(l)} - u_{n-1}^{(l)}}{h} + \beta^{(1)}(t_l)u_n^{(l)}\Big] + \\
&+ (1-\alpha)\Big[p(x_n - h/2, t_{l-1})\frac{u_n^{(l-1)} - u_{n-1}^{(l-1)}}{h} + \beta^{(1)}(t_{l-1})u_n^{(l-1)}\Big] + \\
&+ \tfrac{1}{2}h\alpha q(x_n,t_l)u_n^{(l)} + \tfrac{1}{2}h(1-\alpha)q(x_n,t_{l-1})u_n^{(l-1)}.
\end{aligned}
$$

K těmto rovnicím se dojde nejsnáze tak, že v diferenciální rovnici (1.12) z kap. II píšeme místo pravé strany f funkci $-c\partial u/\partial t + f$, výraz $-c\partial u/\partial t$ v místě (x_k, t_l), resp. (x_k, t_{l-1}) aproximujeme výrazem $c(x_k,t_l)[u(x_k,t_l) - u(x_k,t_{l-1})]/\tau$, resp. $c(x_k,t_{l-1})[u(x_k,t_l) - u(x_k,t_{l-1})]/\tau$, napíšeme rovnice typu (3.40) a (3.58) z kap. II pro bod (x_k,t_l) a (x_k,t_{l-1}) a zkombinujeme je s vahami α a $(1-\alpha)$.

Opět nás bude zajímat, jako dříve při vyšetřování metody sítí, jaký bude výsledek aplikace právě zavedených operátorů na funkce definované na $\bar{R}_{h,\tau}$, které vzniknou z hodnot (dostatečně hladké) funkce u v uzlových bodech. To popíšeme v následujících dvou větách, v nichž klademe jako dříve $(u^{(\mathrm{pr})})_k^{(l)} = u(x_k,t_l)$.

Věta 2.6. *Nechť koeficienty diferenciálních operátorů* (2.110) a (2.114) *jsou dostatečně hladké a nechť u je dostatečně hladká funkce definovaná v $\bar{R}$. Pak existují funkce v_0 a v_1 definované a spojitě diferencovatelné v intervalu $\langle 0, T\rangle$ a takové, že platí*

$$
\begin{aligned}
(2.119)\quad & (L_{h,\tau}^{(\alpha)}u^{(\mathrm{pr})})_k^{(l)} = \alpha(Lu)(x_k,t_l) + (1-\alpha)(Lu)(x_k,t_{l-1}) + O(\tau + h^2), \\
& k = 1, \dots, n-1,\ l = 1, \dots, r,
\end{aligned}
$$

a

$$
\begin{aligned}
(2.120)\quad (l_{h,\tau}^{(i,\alpha)}u^{(\mathrm{pr})})^{(l)} =& \tfrac{1}{2}h[(Lu)(i,t_l) + (1-\alpha)(Lu)(i,t_{l-1})] + \\
&+ \alpha(l^{(i)}u)(t_l) + (1-\alpha)(l^{(i)}u)(t_{l-1}) - \\
&- [\alpha v_i(t_l) + (1-\alpha)v_i(t_{l-1})]h^2 + O(\tau h) + O(h^3), \\
& l = 1, \dots, r,\ i = 0, 1.
\end{aligned}
$$

Věta 2.7. *Nechť koeficienty diferenciálních operátorů* (2.110) a (2.114) *jsou dostatečně hladké a nechť u je dostatečně hladká funkce definovaná v $\bar{R}$. Pak existují funkce v_0 a v_1 definované a spojitě diferencovatelné v intervalu $\langle 0, T\rangle$ a takové, že platí*

$$
\begin{aligned}
(2.121)\quad & (L_{h,\tau}^{(1/2)}u^{(\mathrm{pr})})_k^{(l)} = \tfrac{1}{2}(Lu)(x_k,t_l) + \tfrac{1}{2}(Lu)(x_k,t_{l-1}) + O(\tau^2 + h^2), \\
& k = 1, \dots, n-1,\ l = 1, \dots, r,
\end{aligned}
$$

a

$$
\begin{aligned}
(2.122)\quad (l_{h,\tau}^{(i,1/2)}u^{(\mathrm{pr})})^{(l)} =& \tfrac{1}{2}h[\tfrac{1}{2}(Lu)(i,t_l) + \tfrac{1}{2}(Lu)(i,t_{l-1})] + \\
&+ \tfrac{1}{2}(l^{(i)}u)(t_l) + \tfrac{1}{2}(l^{(i)}u)(t_{l-1}) - \\
&- \tfrac{1}{2}[v_i(t_l) + v_i(t_{l-1})]h^2 + O(\tau^2 h) + O(h^3), \\
& l = 1, \dots, r,\ i = 0, 1.
\end{aligned}
$$

Důkazy obou těchto vět se provedou přímočaře pomocí Tayolorova vzorce. Vyžadují však dosti zdlouhavé psaní a protože jsme už řadu podobných důkazů podrobně prováděli, přenecháme je čtenáři.

V právě uvedených větách jsme nespecifikovali přesně hladkostní předpoklady, které je třeba klást na koeficienty daných diferenciálních operátorů a na funkci u. Zformulovat tyto požadavky není samozřejmě žádný problém; že jsme tak neučinili, je motivováno zejména tím, že jsme nechtěli těmito pro další úvahy nepodstatnými podrobnostmi snižovat přehlednost citovaných vět.

Na základě vět 2.6 a 2.7 se zdá rozumné nahradit danou okrajovou úlohu rovnicemi

$$
\begin{aligned}
(2.123)\quad & (L_{h,\tau}^{(\alpha)}u)_k^{(l)} = \alpha f(x_k,t_l) + (1-\alpha)f(x_k,t_{l-1}), \\
& k = 1, \dots, n-1,\ l = 1, \dots, r,
\end{aligned}
$$

s počátečními podmínkami

$$u_k^{(0)} = g(x_k), \quad k = 0, \dots, n, \tag{2.124}$$

a okrajovými podmínkami

$$u_0^{(l)} = \gamma^{(0)}(t_l), \quad u_n^{(l)} = \gamma^{(1)}(t_l), \quad l = 1, \dots, r, \tag{2.125}$$

v případě Dirichletovy okrajové úlohy a podmínkami

$$\begin{aligned}(l_{h,\tau}^{(0,\alpha)}u)^{(l)} &= \tfrac{1}{2}h\alpha f(x_0,t_l) + \tfrac{1}{2}h(1-\alpha)f(x_0,t_{l-1}) + \\ &+ \alpha\gamma^{(0)}(t_l) + (1-\alpha)\gamma^{(0)}(t_{l-1}), \quad l = 1, \dots, r,\end{aligned} \tag{2.126}$$

$$\begin{aligned}(l_{h,\tau}^{(1,\alpha)}u)^{(l)} &= \tfrac{1}{2}h\alpha f(x_n,t_l) + \tfrac{1}{2}h(1-\alpha)f(x_n,t_{l-1}) + \\ &+ \alpha\gamma^{(1)}(t_l) + (1-\alpha)\gamma^{(1)}(t_{l-1}), \quad l = 1, \dots, r,\end{aligned}$$

v případě druhé a třetí okrajové úlohy, neboť se přitom dopustíme lokálně malé chyby. Především je třeba se přesvědčit o tom, že rovnice (2.123) spolu s příslušnými počátečními a okrajovými podmínkami definují skutečně algoritmus, neboli jinými slovy jednoznačně určují veličinu $u_k^{(l)}$. Provedeme to analogicky jako dříve. Buď opět $u^{(l)}$ vektor, jehož složky jsou hodnoty přibližného řešení v l-tém časovém řádku. Vynásobíme-li rovnice (2.123) integračním krokem τ a rovnice (2.126) číslem τ/h, dostaneme

$$A_{\mathrm{U}}^{(l)}u^{(l)} = A_{\mathrm{L}}^{(l)}u^{(l-1)} + \frac{\tau}{h^2}[\alpha f^{(l)} + (1-\alpha)f^{(l-1)}], \quad l = 1, \dots, r, \tag{2.127}$$

kde

$$\begin{aligned}A_{\mathrm{U}}^{(l)} &= \alpha C^{(l)} + (1-\alpha)C^{(l-1)} + \alpha\frac{\tau}{h^2}P^{(l)}, \\ A_{\mathrm{L}}^{(l)} &= \alpha C^{(l)} + (1-\alpha)C^{(l-1)} - (1-\alpha)\frac{\tau}{h^2}P^{(l-1)}, \quad l = 1, \dots, r,\end{aligned} \tag{2.128}$$

$C^{(l)}$ je diagonální matice, $P^{(l)}$ je symetrická třídiagonální matice, obě řádu $n-1$ nebo $n+1$ podle druhu okrajových podmínek a $f^{(l)}$ je $(n-1)$-dimenzionální, resp. $(n+1)$-dimenzionální vektor,

$$C^{(l)} = \{c_{km}^{(l)}\}, \quad c_{kk}^{(l)} = c(x_k,t_l), \quad k = 1, \dots, n-1, \tag{2.129}$$

resp.

$$c_{00}^{(l)} = \tfrac{1}{2}c(x_0,t_l), \quad c_{kk}^{(l)} = c(x_k,t_l), \quad k = 1, \dots, n-1, \quad c_{nn}^{(l)} = \tfrac{1}{2}c(x_n,t_l),$$

$$\begin{aligned}P^{(l)} &= \{p_{km}^{(l)}\}, \\ p_{kk}^{(l)} &= p(x_k - h/2, t_l) + p(x_k + h/2, t_l) + h^2 q(x_k,t_l), \quad k = 1, \dots, n-1, \\ p_{k,k+1}^{(l)} &= p_{k+1,k}^{(l)} = -p(x_k + h/2, t_l), \quad k = 1, \dots, n-2,\end{aligned} \tag{2.130}$$

resp.

$$\begin{aligned}p_{00}^{(l)} &= p(x_0 + h/2, t_l) + h\beta^{(0)}(t_l) + \tfrac{1}{2}h^2 q(x_0,t_l), \\ p_{kk}^{(l)} &= p(x_k - h/2, t_l) + p(x_k + h/2, t_l) + h^2 q(x_k,t_l), \quad k = 1, \dots, n-1, \\ p_{nn}^{(l)} &= p(x_n - h/2, t_l) + h\beta^{(1)}(t_l) + \tfrac{1}{2}h^2 q(x_n,t_l), \\ p_{k,k+1}^{(l)} &= p_{k+1,k}^{(l)} = -p(x_k + h/2, t_l), \quad k = 1, \dots, n-2,\end{aligned}$$

$$\begin{aligned}f^{(l)} &= \{f_k^{(l)}\}, \\ f_1^{(l)} &= h^2 f(x_1,t_l) + p(x_0 + h/2, t_l)\gamma^{(0)}(t_l), \\ f_k^{(l)} &= h^2 f(x_k,t_l), \quad k = 2, \dots, n-2, \\ f_{n-1}^{(l)} &= h^2 f(x_{n-1},t_l) + p(x_n - h/2, t_l)\gamma^{(1)}(t_l),\end{aligned} \tag{2.131}$$

resp.

$$\begin{aligned}f_0^{(l)} &= \tfrac{1}{2}h^2 f(x_0,t_l) + h\gamma^{(0)}(t_l), \\ f_k^{(l)} &= h^2 f(x_k,t_l), \quad k = 1, \dots, n-1, \\ f_n^{(l)} &= \tfrac{1}{2}h^2 f(x_n,t_l) + h\gamma^{(1)}(t_l).\end{aligned}$$

Matice A_{U} je podle lemmatu 3.5 z kap. 2 monotónní. Je proto regulární a rovnice (2.123) s příslušnými počátečními a okrajovými podmínkami tvoří tedy skutečně metodu. Je-li $u_k^{(l)}$ přibližné řešení dané úlohy vypočtené z rovnic (2.123) s počáteční podmínkou (2.124) a okrajovými podmínkami (2.125), resp. (2.126) a jsou-li koeficienty a přesné řešení dané úlohy dostatečně hladké, platí pro chybu $\eta_k^{(l)} = u_k^{(l)} - u(x_k,t_l)$ podle vět 2.6 a 2.7 rovnice

$$(l_{h,\tau}^{(\alpha)}\eta)_k^{(l)} = \varepsilon_k^{(l)}, \quad k = 1, \dots, n-1, \; l = 1, \dots, r, \tag{2.132}$$

s počáteční podmínkou

$$\eta_k^{(0)} = 0, \quad k = 0, \dots, n, \tag{2.133}$$

a okrajovými podmínkami

$$\eta_0^{(l)} = \eta_n^{(l)} = 0, \quad l = 1, \dots, r, \tag{2.134}$$

resp.

$$\begin{aligned}(l_{h,\tau}^{(0,\alpha)}\eta)^{(l)} &= \varepsilon_0^{(l)} + \alpha\delta_0^{(l)} + (1-\alpha)\delta_0^{(l-1)}, \\ (l_{h,\tau}^{(1,\alpha)}\eta)^{(l)} &= \varepsilon_n^{(l)} + \alpha\delta_1^{(l)} + (1-\alpha)\delta_1^{(l-1)}, \quad l = 1, \dots, r,\end{aligned} \tag{2.135}$$

kde

$$\varepsilon_k^{(l)} = O(\tau + h^2), \quad k = 1, \dots, n-1, \; l = 1, \dots, r, \tag{2.136}$$

a

$$\varepsilon_{in}^{(l)} = O(\tau h) + O(h^3),$$
$$\delta_i^{(l)} = v_i(t_l)h^2, \quad i = 0, 1, \ l = 1, \dots, r, \tag{2.137}$$

v obecném případě a

$$\varepsilon_k^{(l)} = O(\tau^2 + h^2), \quad k = 1, \dots, n-1, \ l = 1, \dots, r, \tag{2.138}$$

a

$$\varepsilon_{in}^{(l)} = O(\tau^2 h) + O(h^3),$$
$$\delta_i^{(l)} = v_i(t_l)h^2, \quad i = 0, 1, \ l = 1, \dots, r, \tag{2.139}$$

v případě, že je $\alpha = 1/2$.

Pravé strany rovnic pro chybu jsou tedy malé a abychom dokázali konvergenci, je třeba dokázat, že z toho plyne i malost příslušných řešení. Tyto otázky vyšetříme v následujících dvou odstavcích.

2.2.2 Konvergence, speciální případy

Vzpomeňme si, jak jsme postupovali při vyšetřování konvergence u rovnice pro vedení tepla. V případě explicitní metody ($\alpha = 0$) byl postup přímočarý vzhledem k jejímu rekurentnímu charakteru. V případě čistě implicitní metody ($\alpha = 1$) jsme užili principu maxima. Konečně v případě obecného α jsme užili metody separace proměnných. Je přirozené položit si otázku, zda lze tyto postupy přenést na zde vyšetřovaný obecný případ. Pro explicitní a implicitní metodu to zřejmě možné je. Nám však jde zejména o obecnou metodu, protože při $\alpha = 1/2$ je její lokální chyba minimální. Pokusíme se tedy užít metodu separace proměnných. Při jejím užití bude patrně vadit proměnnost koeficientů a hlavně pak jejich závislost na t. Předpokládejme tedy, že koeficienty dané diferenciální rovnice a okrajových podmínek jsou nezávislé na čase a uvažujme pro jednoduchost Dirichletovu úlohu. Jiné okrajové podmínky by totiž působily rovněž určité obtíže vzhledem k tomu, že rovnice (2.135) nejsou homogenní. Pro chybu $\eta_k^{(l)}$ přibližného řešení tedy platí rovnice (2.132) s podmínkami (2.133) a (2.134). Přepíšeme-li tyto rovnice pomocí vektorové symboliky, dostaneme

$$A_{\mathrm{U}}^{(l)}\boldsymbol{\eta}^{(l)} = A_{\mathrm{L}}^{(l)}\boldsymbol{\eta}^{(l-1)} + \tau\boldsymbol{\epsilon}^{(l)}, \tag{2.140}$$

kde matice $A_{\mathrm{U}}^{(l)}$ a $A_{\mathrm{L}}^{(l)}$ nezávisí na indexu l. Horní index budeme tedy v dalším vypouštět. Vzhledem k předpokladu (2.115) o koeficientu c můžeme bez újmy na obecnosti předpokládat, že je $c(x) \equiv 1$, a že matice A_{U} a A_{L} lze tedy psát takto:

$$A_{\mathrm{U}} = I + \alpha\beta P,$$
$$A_{\mathrm{L}} = I - (1-\alpha)\beta P, \tag{2.141}$$

kde P je třídiagonální matice s prvky

$$p_{kk} = p(x_k - h/2) + p(x_k + h/2) + h^2 q(x_k), \quad k = 1, \dots, n-1,$$
$$p_{k,k+1} = p_{k+1,k} = -p(x_k + h/2), \quad k = 1, \dots, n-2, \tag{2.142}$$

a kde jsme opět položili $\tau/h^2 = \beta$. Dostali jsme se tedy do úplně stejné situace, jako když jsme vyšetřovali konvergenci obecné metody pro rovnici pro vedení tepla. Budeme proto postupovat stejně jako v odst. 2.1.2. Konvergenci dokážeme, dokážeme-li, že dané schéma je stabilní vzhledem k pravé straně. Tuto poslední skutečnost pak dokážeme použitím lemmatu 2.3. Ověříme proto jeho předpoklady.

Odhadněme nejprve velikost normy matice A_{U}^{-1} v prostoru $\mathsf{E}_h^{(n-1)}$. Protože je to obyčejná spektrální norma, jak už víme, a protože matice A_{U}^{-1} je zřejmě symetrická, stačí k tomu nalézt její vlastní čísla. Matice P je samozřejmě také symetrická. Označme její navzájem ortogonální vlastní vektory $\boldsymbol{w}^{(\nu)}$, $\nu = 1, \dots, n-1$, a jim odpovídající vlastní čísla λ_ν. Kromě toho, že matice P je symetrická, je navíc diagonálně dominantní. Odtud a z Geršgorinovy věty o lokalizaci vlastních čísel plyne, že její vlastní čísla jsou nezáporná. Z Collatzova lemmatu však plyne, že matice P je monotónní, a tedy regulární. To vše dohromady dává, že matice P je pozitivně definitní. Z první rovnice (2.141) plyne, že vektory $\boldsymbol{w}^{(\nu)}$ jsou také vlastními vektory matice A_{U} a že příslušná vlastní čísla $\lambda_\nu^{(\mathrm{U})}$ jsou dána vzorci

$$\lambda_\nu^{(\mathrm{U})} = 1 + \alpha\beta\lambda_\nu, \quad \nu = 1, \dots, n-1. \tag{2.143}$$

Odtud máme, že je $\lambda_\nu^{(\mathrm{U})} \geqq 1$ pro $\nu = 1, \dots, n-1$, což dává

$$\|A_{\mathrm{U}}^{-1}\|_h = \frac{1}{\lambda_{\min}^{(\mathrm{U})}} \leqq 1. \tag{2.144}$$

První podmínka lemmatu 2.3 je tedy splněna.

Zkoumejme dále stejnoměrnou stabilitu daného schématu vzhledem k počátečním podmínkám. To znamená odhadnout normu vektorů $\boldsymbol{\eta}^{(l)}$, které jsou řešením soustavy

$$A_{\mathrm{U}}\boldsymbol{\eta}^{(l)} = A_{\mathrm{L}}\boldsymbol{\eta}^{(l-1)}, \quad l = s+1, \dots, r, \tag{2.145}$$

normou vektoru $\boldsymbol{\eta}^{(s)}$. Abychom toho dosáhli, stačí opět opakovat úvahy, které jsme už prováděli v odst. 2.1.2. Vektor $\boldsymbol{w}^{(\nu)}$ je totiž nejen vlastním vektorem matice A_{U} s odpovídajícím vlastním číslem daným rovnicí (2.143), ale je samozřejmě i vlastním vektorem matice A_{L} s vlastním číslem $\lambda_\nu^{(\mathrm{L})}$ daným vzorcem

$$\lambda_\nu^{(\mathrm{L})} = 1 - (1-\alpha)\beta\lambda_\nu. \tag{2.146}$$

Řešení rovnice (2.145) je tedy možné opět psát ve tvaru

$$\boldsymbol{\eta}^{(l)} = \sum_{\nu=1}^{n-1} c_\nu \left[\frac{\lambda_\nu^{(\mathrm{L})}}{\lambda_\nu^{(\mathrm{U})}}\right]^{l-s} \boldsymbol{w}^{(\nu)}, \tag{2.147}$$

kde

$$\eta^{(s)} = \sum_{\nu=1}^{n-1} c_\nu w^{(\nu)}. \tag{2.148}$$

Odtud už stejnoměrná stabilita vzhledem k počátečním podmínkám snadno plyne, platí-li, že je

$$\left|\frac{\lambda_\nu^{(L)}}{\lambda_\nu^{(U)}}\right| \leqq 1, \quad \nu = 1, \ldots, n-1. \tag{2.149}$$

Tato podmínka je však stejně jako v odst. 2.1.2 splněna v případě, že je $1/2 \leqq \alpha \leqq 1$, bez jakýchkoliv doplňujících podmínek, neboť matice P je pozitivně definitní. V případě, že je $0 \leqq \alpha < 1/2$, k jejímu splnění je nutné a stačí, aby platila nerovnost

$$\beta = \frac{\tau}{h^2} \leqq \frac{2}{(1-2\alpha)\lambda_{\max}}, \tag{2.150}$$

kde $\lambda_{\max}$ je největší vlastní číslo matice P. Dokázali jsme tedy následující větu.

Věta 2.8. *Nechť řešení Dirichletovy úlohy pro rovnici* (2.109), *jejíž koeficienty nezávisí na* t, *je dostatečně hladké a nechť při* $0 \leqq \alpha < 1/2$ *platí navíc nerovnost* (2.150). *Pak existuje konstanta* M *taková, že pro chybu* $\eta_k^{(l)}$ *přibližného řešení vypočteného pomocí operátoru* $L_{h,\tau}^{(\alpha)}$ *platí*

$$\|\eta^{(l)}\|_h \leqq \begin{cases} M(\tau + h^2) & \text{pro } \alpha \neq 1/2, \\ M(\tau^2 + h^2) & \text{pro } \alpha = 1/2. \end{cases} \tag{2.151}$$

Metoda daná operátorem $L_{h,\tau}^{(1/2)}$ se opět nazývá Crankova-Nicolsonova metoda a vede mezi všemi metodami danými operátory $L_{h,\tau}^{(\alpha)}$ k nejekonomičtějšímu algoritmu.

Pokusme se ještě zkonkretizovat nerovnost (2.150), tj. odhadnout maximální vlastní číslo matice P. Použijeme k tomu, že pro symetrickou pozitivně definitní matici platí

$$\lambda_{\max} = \sup_{v \neq 0} \frac{(Pv, v)_h}{(v, v)_h}. \tag{2.152}$$

Buď tedy $v = (v_1, \ldots, v_{n-1})^T$ a vypočtěme veličinu $(Pv, v)_h$. Položíme-li ještě $v_0 = v_n = 0$, je zřejmě

$$(Pv, v)_h = h \sum_{k=1}^{n-1} \{-p(x_k - h/2)v_{k-1} + [p(x_k - h/2) + p(x_k + h/2) + h^2 q(x_k)]v_k - p(x_k + h/2)v_{k+1}\}v_k. \tag{2.153}$$

Odtud však sumací per partes (srv. lemma 3.10 z kap. II) snadno dostaneme, že je

$$(Pv, v)_h = h \sum_{k=1}^{n} p(x_k - h/2)(v_k - v_{k-1})^2 + h \sum_{k=1}^{n-1} h^2 q(x_k) v_k^2. \tag{2.154}$$

Použijeme-li nyní nerovnosti (2.115), dostáváme odtud, že platí

$$(Pv, v)_h \leqq (\tilde{P}v, v)_h, \tag{2.155}$$

kde

$$\tilde{P} = p_1 P_0 + h^2 q_1 I \tag{2.156}$$

a matice P_0 je dána vzorcem (2.81). Je tedy $\lambda_{\max}^{(P)} \leqq \lambda_{\max}^{(\tilde{P})} = p_1 \lambda_{\max}^{(P_0)} + h^2 q_1$. V odst. 2.1.2 jsme však zjistili, že je $\lambda_{\max}^{(P_0)} \leqq 4$. K splnění nerovnosti (2.150) tedy stačí, aby platila nerovnost

$$\beta \leqq \frac{1}{(1-2\alpha)(2p_1 + h^2 q_1/2)}. \tag{2.157}$$

2.2.3 Konvergence, obecný případ

Začneme obecnou rovnicí (2.109) a explicitní metodou a omezíme se jako na nejsložitější na třetí okrajovou úlohu. U tohoto omezení zůstaneme i v dalším vyšetřování. Případy ostatních okrajových úloh by se studovaly analogicky; navíc by většinou došlo ještě k dalšímu zjednodušení.

V tomto případě tedy pro chybu $\eta_k^{(l)}$ platí (operátory $L_{h,\tau}^{(0)}$, $l_{h,\tau}^{(0,0)}$ a $l_{h,\tau}^{(1,0)}$ jsou dány vzorci (2.117) a (2.118))

$$\begin{aligned} (L_{h,\tau}^{(0)}\eta)_k^{(l)} &= \varepsilon_k^{(l)}, \quad k = 1, \ldots, n-1, \ l = 1, \ldots, r, \\ (l_{h,\tau}^{(0,0)}\eta)^{(l)} &= \varepsilon_0^{(l)} + \delta_0^{(l-1)}, \quad l = 1, \ldots, r, \\ (l_{h,\tau}^{(1,0)}\eta)^{(l)} &= \varepsilon_n^{(l)} + \delta_1^{(l-1)}, \quad l = 1, \ldots, r, \end{aligned} \tag{2.158}$$

a za předpokladu dostatečné hladkosti je

$$\begin{aligned} \varepsilon_k^{(l)} &= O(\tau + h^2), \quad k = 1, \ldots, n-1, \ l = 1, \ldots, r, \\ \varepsilon_{in}^{(l)} &= O(\tau h) + O(h^3), \\ \delta_i^{(l-1)} &= v_i(t_{l-1})h^2, \quad i = 0, 1, \ l = 1, \ldots, r, \end{aligned} \tag{2.159}$$

kde funkce v_i mají spojitou derivaci v intervalu $\langle 0, T\rangle$.

Abychom dokázali konvergenci, je třeba odhadnout řešení soustavy (2.158). Přepišme je proto tak, že první rovnici vynásobíme číslem $\tau/c(x_k, t_{l-1})$ a druhé dvě

čísly $2\tau/[hc(x_0,t_{l-1})]$ a $2\tau/[hc(x_n,t_{l-1})]$. Dostaneme

$$\eta_k^{(l)} = \beta\frac{p(x_k-h/2,t_{l-1})}{c(x_k,t_{l-1})}\eta_{k-1}^{(l-1)} + \Big[1-\beta\frac{p(x_k-h/2,t_{l-1})+p(x_k+h/2,t_{l-1})+h^2q(x_k,t_{l-1})}{c(x_k,t_{l-1})}\Big]\eta_k^{(l-1)} + \beta\frac{p(x_k+h/2,t_{l-1})}{c(x_k,t_{l-1})}\eta_{k+1}^{(l-1)} + \frac{\tau}{c(x_k,t_{l-1})}\varepsilon_k^{(l)}, \quad k=1,\dots,n-1, \tag{2.160}$$

$$\eta_0^{(l)} = \Big[1-\beta\frac{2p(x_0+h/2,t_{l-1})+2h\beta^{(0)}(t_{l-1})+h^2q(x_0,t_{l-1})}{c(x_0,t_{l-1})}\Big]\eta_0^{(l-1)} + 2\beta\frac{p(x_0+h/2,t_{l-1})}{c(x_0,t_{l-1})}\eta_1^{(l-1)} + \frac{2\tau}{h}\frac{1}{c(x_0,t_{l-1})}[\varepsilon_0^{(l)}+\delta_0^{(l-1)}]$$

a analogickou rovnici pro druhý kraj. (Položili jsme opět $\beta = \tau/h^2$.) Na základě tohoto vyjádření snadno dokážeme následující větu.

Věta 2.9. *Nechť koeficienty dané diferenciální rovnice a okrajových podmínek splňují nerovnosti* (2.115) a (2.116). *Nechť dále platí nerovnost*

$$\beta \leqq \frac{c_0}{2(p_1+\beta_1 h)+q_1h^2}. \tag{2.161}$$

Pak existuje konstanta M taková, že pro řešení rovnic (2.158) *s* $\delta_0^{(l)} = \delta_1^{(l)} = 0$ *platí*

$$\max_{k=0,\dots,n}|\eta_k^{(l)}| \leqq M\Big[\max_{\nu=1,\dots,l}\max(h^{-1}|\varepsilon_0^{(\nu)}|, \max_{k=1,\dots,n-1}|\varepsilon_k^{(\nu)}|, h^{-1}|\varepsilon_n^{(\nu)}|) + \max_{k=0,\dots,n}|\eta_k^{(0)}|\Big], \quad l=1,\dots,r. \tag{2.162}$$

D ů k a z tohoto tvrzení plyne snadno úplnou indukcí, neboť vzhledem k podmínce (2.161) jsou koeficienty v hranatých závorkách na pravé straně rovnice (2.160) nezáporné.

Z této věty a z odhadů (2.159) ihned plyne, že explicitní schéma konverguje jako $O(\tau+h)$. Rychlost konvergence $O(\tau+h^2)$, kterou na základě analogie s výše vyšetřovanými případy očekáváme, bychom dostali jen v tom případě, kdyby náhodou bylo $\delta_i^{(l)} = 0$. Tak tomu samozřejmě obecně není, a k vyspravení odhadu chyby tedy potřebujeme jemněji odhadnout řešení rovnic (2.158), ve kterých je $\varepsilon_k^{(l)} = 0$. K takovému odhadu použijeme dvě skoro zřejmá pomocná tvrzení.

Lemma 2.4. *Nechť platí nerovnosti* (2.115), (2.116) a (2.161) a *nechť* $\chi_k^{(l)}$ a $\eta_k^{(l)}$ *jsou dvě funkce definované na síti* $\bar{R}_{h,\tau}$, *pro něž platí*

$$|(L_{h,\tau}^{(0)}\eta)_k^{(l)}| \leqq (L_{h,\tau}^{(0)}\chi)_k^{(l)}, \quad k=1,\dots,n-1,\ l=1,\dots,r,$$
$$|(l_{h,\tau}^{(i,0)}\eta)^{(l)}| \leqq (l_{h,\tau}^{(i,0)}\chi)^{(l)}, \quad i=0,1,\ l=1,\dots,r,$$
$$|\eta_k^{(0)}| \leqq \chi_k^{(0)}, \quad k=0,\dots,n. \tag{2.163}$$

Pak platí

$$|\eta_k^{(l)}| \leqq \chi_k^{(l)}, \quad k=0,\dots,n,\ l=0,\dots,r. \tag{2.164}$$

D ů k a z tohoto tvrzení plyne ihned z analogického rozpisu jako v rovnici (2.160). Poznamenejme také, že tvrzení lemmatu 2.4 je v podstatě ekvivalentní principu maxima pro operátor $L_{h,\tau}^{(0)}$.

Lemma 2.5. *Nechť pro koeficienty dané diferenciální rovnice a okrajových podmínek platí nerovnosti* (2.115) a (2.116) a *nechť při pevném* l *je* $\varphi_k^{(l)}$ *řešením soustavy*

$$p(x_k-h/2,t_l)\varphi_{k-1}^{(l)} - [p(x_k-h/2,t_l)+p(x_k+h/2,t_l)]\varphi_k^{(l)} + p(x_k+h/2,t_l)\varphi_{k+1}^{(l)} = 0, \quad k=1,\dots,n-1,$$
$$-p(x_0+h/2,t_l)\frac{\varphi_1^{(l)}-\varphi_0^{(l)}}{h} + \beta^{(0)}(t_l)\varphi_0^{(l)} = \delta_0^{(l)},$$
$$p(x_n-h/2,t_l)\frac{\varphi_n^{(l)}-\varphi_{n-1}^{(l)}}{h} + \beta^{(1)}(t_l)\varphi_n^{(l)} = \delta_1^{(l)}. \tag{2.165}$$

Pak existuje konstanta M (nezávislá na k a l) taková, že platí

$$\max_{k=0,\dots,n}|\varphi_k^{(l)}| \leqq M\max(|\delta_0^{(l)}|,|\delta_1^{(l)}|),$$
$$\max_{k=1,\dots,n}\Big|\frac{\varphi_k^{(l)}-\varphi_{k-1}^{(l)}}{h}\Big| \leqq M\max(|\delta_0^{(l)}|,|\delta_1^{(l)}|),$$
$$\max_{k=0,\dots,n}\Big|\frac{\varphi_k^{(l)}-\varphi_k^{(l-1)}}{\tau}\Big| \leqq M\max\Big(|\delta_0^{(l)}|,|\delta_1^{(l)}|,|\delta_0^{(l-1)}|,|\delta_1^{(l-1)}|, \frac{\delta_0^{(l)}-\delta_0^{(l-1)}}{\tau}, \frac{\delta_1^{(l)}-\delta_1^{(l-1)}}{\tau}\Big). \tag{2.166}$$

D ů k a z . Tvrzení lemmatu plyne snadno ze skutečnosti, že řešení soustavy (2.165) se dá psát vzorcem

$$\varphi_k^{(l)} = \frac{1}{\Psi^{(l)}}\Big\{\delta_0^{(l)}\Big[1+\beta^{(1)}(t_l)\sum_{\nu=k+1}^{n}\frac{h}{p(x_\nu-h/2,t_l)}\Big] + \delta_1^{(l)}\Big[1+\beta^{(0)}(t_l)\sum_{\nu=1}^{k}\frac{h}{p(x_\nu-h/2,t_l)}\Big]\Big\}, \tag{2.167}$$

kde

$$\Psi^{(l)} = \beta^{(0)}(t_l)+\beta^{(1)}(t_l)+\beta^{(0)}(t_l)\beta^{(1)}(t_l)\sum_{\nu=1}^{n}\frac{h}{p(x_\nu-h/2,t_l)}, \tag{2.168}$$

jak se snadno ověří dosazením do rovnic (2.165).

Pomocí těchto dvou lemmat už snadno dokážeme následující větu.

Věta 2.10. *Nechť koeficienty dané diferenciální rovnice a okrajových podmínek splňují nerovnosti* (2.115) a (2.116) *a nechť platí nerovnost* (2.161). *Nechť dále je* $\eta_k^{(l)}$ *řešením soustavy* (2.158) *s nulovou počáteční podmínkou a s* $\varepsilon_k^{(l)} = 0$ *pro* $k = 0, \dots, n$ *a* $l = 1, \dots, r$. *Pak existuje konstanta* M *taková, že pro* $l = 1, \dots, r$ *platí*

$$\max_{k=0,\dots,n} |\eta_k^{(l)}| \leqq M \max_{\nu=0,\dots,l-1} (|\delta_0^{(\nu)}|, |\delta_1^{(\nu)}|). \tag{2.169}$$

D ů k a z . Zvolme pevně l, $1 \leqq l \leqq r$, položme

$$\theta_0^{(l)} = \max_{\nu=0,\dots,l-1} |\delta_0^{(\nu)}|, \quad \theta_1^{(l)} = \max_{\nu=0,\dots,l-1} |\delta_1^{(\nu)}| \tag{2.170}$$

a buď $\chi_k^{(\nu)}$ řešením soustavy

$$\begin{aligned} (L_{h,\tau}^{(0)}\chi)_k^{(\nu)} &= 0, \quad k = 1, \dots, n-1, \ \nu = 1, \dots, l, \\ (l_{h,\tau}^{(i,0)}\chi)^{(\nu)} &= \theta_i^{(l)}, \quad i = 0, 1, \ \nu = 1, \dots, l, \\ \chi_k^{(0)} &= 0, \quad k = 0, \dots, n. \end{aligned} \tag{2.171}$$

Protože v důsledku lemmatu 2.4 je $|\eta_k^{(\nu)}| \leqq \chi_k^{(\nu)}$ pro $\nu = 0, \dots, l$, stačí odhadnout řešení soustavy (2.171). Odhadujme $\chi_k^{(\nu)}$ ve tvaru součtu

$$\chi_k^{(\nu)} = \varphi_k^{(\nu)} + \Psi_k^{(\nu)}, \tag{2.172}$$

kde $\varphi_k^{(\nu)}$ řeší při pevném ν, $0 \leqq \nu \leqq l$, rovnice

$$\begin{aligned} &p(x_k - h/2, t_\nu)\varphi_{k-1}^{(\nu)} - [p(x_k - h/2, t_\nu) + p(x_k + h/2, t_\nu)]\varphi_k^{(\nu)} + \\ &+ p(x_k + h/2, t_\nu)\varphi_{k+1}^{(\nu)} = 0, \quad k = 1, \dots, n-1, \end{aligned} \tag{2.173}$$

s okrajovými podmínkami

$$\begin{aligned} -p(x_0 + h/2, t_\nu)\frac{\varphi_1^{(\nu)} - \varphi_0^{(\nu)}}{h} + \beta^{(0)}(t_\nu)\varphi_0^{(\nu)} &= \theta_0^{(l)}, \\ p(x_n + h/2, t_\nu)\frac{\varphi_n^{(\nu)} - \varphi_{n-1}^{(\nu)}}{h} + \beta^{(1)}(t_\nu)\varphi_n^{(\nu)} &= \theta_1^{(l)}. \end{aligned} \tag{2.174}$$

Funkci $\varphi_k^{(\nu)}$ umíme odhadnout pomocí lemmatu 2.5 a funkce $\Psi_k^{(\nu)}$ řeší rovnice

$$\begin{aligned} (L_{h,\tau}^{(0)}\Psi)_k^{(\nu)} &= -c(x_k, t_{\nu-1})\frac{\varphi_k^{(\nu)} - \varphi_k^{(\nu-1)}}{\tau} - q(x_k, t_{\nu-1})\varphi_k^{(\nu-1)}, \\ &\quad k = 1, \dots, n-1, \ \nu = 1, \dots, l, \\ (l_{h,\tau}^{(i,0)}\Psi)^{(\nu)} &= -\tfrac{1}{2}hc(x_{in}, t_{\nu-1})\frac{\varphi_{in}^{(\nu)} - \varphi_{in}^{(\nu-1)}}{\tau} - \\ &\quad - \tfrac{1}{2}hq(x_{in}, t_{\nu-1})\varphi_{in}^{(\nu-1)}, \quad i = 0, 1, \ \nu = 1, \dots, l, \\ \Psi_k^{(0)} &= -\varphi_k^{(0)}, \quad k = 0, \dots, n. \end{aligned} \tag{2.175}$$

Řešení posledně napsané soustavy umíme odhadnout pomocí pravých stran užitím věty 2.9, přičemž pravé strany rovnic (2.175) umíme odhadnout užitím lemmatu 2.5. Protože však pravé strany rovnic (2.174) nezávisí na běžném indexu ν, redukuje se pravá strana poslední nerovnosti v (2.166) na výraz $M \max(|\theta_0^{(l)}|, |\theta_1^{(l)}|)$. Odtud však už tvrzení věty plyne.

Z vět 2.9 a 2.10 dostáváme už ihned nejen konvergenci explicitní metody v případě obecné parabolické rovnice, ale i odhad chyby $O(\tau + h^2)$.

Obraťme se nyní ke studiu schématu $L_{h,\tau}^{(\alpha)}$ s příslušnými okrajovými podmínkami v případě, že je $1/2 \leqq \alpha \leqq 1$. Případ $0 < \alpha < 1/2$ vyžaduje, jak jsme viděli v příslušném odstavci ve speciálním případě, omezení na poměr τ/h^2, a proto jej jako méně zajímavý vyšetřovat nebudeme, i když by to bylo následující metodikou dobře možné. Pokusíme se postupovat paralelně k explicitnímu případu, tj. zkusíme odhadnout řešení rovnice (2.132) s okrajovými podmínkami (2.135) zvlášť pro případ $\delta_i^{(l)} \equiv 0$ a $\varepsilon_k^{(l)} \equiv 0$.

K tomu cíli použijeme postup, který, pokud je nám známo, užil k odhadu řešení parabolické rovnice v závislosti na její pravé straně jako první Lees (1960). Základní myšlenku tohoto postupu ukážeme na příkladě diferenciální rovnice

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = \varepsilon(x, t) \tag{2.176}$$

s okrajovými podmínkami

$$u(0, t) = u(1, t) = 0. \tag{2.177}$$

Budeme přitom předpokládat, že všechny operace, které budou následovat, jsou oprávněné, aniž bychom podrobně zkoumali potřebné předpoklady, neboť nám nejde v daný okamžik o získání přesného tvrzení, ale pouze o vysvětlení základní myšlenky. Vynásobme rovnici (2.176) výrazem $\partial u/\partial t$ a integrujme podle x od 0 do 1. Dostaneme

$$\int_0^1 \left(\frac{\partial u}{\partial t}\right)^2 \mathrm{d}x - \int_0^1 \frac{\partial^2 u}{\partial x^2}\frac{\partial u}{\partial t}\,\mathrm{d}x = \int_0^1 \varepsilon(x, t)\frac{\partial u}{\partial t}\,\mathrm{d}x. \tag{2.178}$$

Je však

$$\begin{aligned} -\int_0^1 \frac{\partial^2 u}{\partial x^2}\frac{\partial u}{\partial t}\,\mathrm{d}x &= -\Big[\frac{\partial u}{\partial x}\frac{\partial u}{\partial t}\Big]_0^1 + \int_0^1 \frac{\partial u}{\partial x}\frac{\partial^2 u}{\partial x \partial t}\,\mathrm{d}x = \\ &= \tfrac{1}{2}\int_0^1 \frac{\partial}{\partial t}\Big[\Big(\frac{\partial u}{\partial x}\Big)^2\Big]\,\mathrm{d}x = \tfrac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 \Big(\frac{\partial u}{\partial x}\Big)^2\,\mathrm{d}x, \end{aligned} \tag{2.179}$$

neboť platí (2.177). Platí tedy

$$\int_0^1 \left(\frac{\partial u}{\partial t}\right)^2 \mathrm{d}x + \tfrac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 \left(\frac{\partial u}{\partial x}\right)^2 \mathrm{d}x = \int_0^1 \varepsilon(x, t)\frac{\partial u}{\partial t}\,\mathrm{d}x. \tag{2.180}$$

Vzhledem k zřejmé nerovnosti $|ab| \leqq \frac{1}{2}a^2 + \frac{1}{2}b^2$, platné pro libovolná reálná a, b, kterou jsme už vícekrát užili, platí

$$\left|\varepsilon(x,t)\frac{\partial u}{\partial t}\right| \leqq \tfrac{1}{2}\varepsilon^2(x,t) + \tfrac{1}{2}\left(\frac{\partial u}{\partial t}\right)^2, \tag{2.181}$$

a tedy

$$\int_0^1 \left(\frac{\partial u}{\partial t}\right)^2 \mathrm{d}x + \tfrac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 \left(\frac{\partial u}{\partial x}\right)^2 \mathrm{d}x \leqq \tag{2.182}$$
$$\leqq \tfrac{1}{2}\int_0^1 \varepsilon^2(x,t)\,\mathrm{d}x + \tfrac{1}{2}\int_0^1 \left(\frac{\partial u}{\partial t}\right)^2 \mathrm{d}x.$$

Odtud však plyne

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 \left(\frac{\partial u}{\partial x}\right)^2 \mathrm{d}x \leqq \int_0^1 \varepsilon^2(x,t)\,\mathrm{d}x, \tag{2.183}$$

a tedy, položíme-li

$$D(t) = \int_0^1 \left[\frac{\partial u}{\partial x}(x,t)\right]^2 \mathrm{d}x, \tag{2.184}$$

$$D(t) \leqq D(0) + \int_0^t \left[\int_0^1 \varepsilon^2(x,\tau)\,\mathrm{d}x\right]\mathrm{d}\tau. \tag{2.185}$$

Je však

$$u(x,t) = \int_0^x \frac{\partial u}{\partial x}(\xi,t)\,\mathrm{d}\xi; \tag{2.186}$$

odtud pomocí Schwarzovy nerovnosti dostáváme

$$u^2(x,t) \leqq x\int_0^x \left[\frac{\partial u}{\partial x}(\xi,t)\right]^2 \mathrm{d}\xi \leqq D(t). \tag{2.187}$$

Z (2.185) tedy máme

$$|u(x,t)| \leqq \left\{D(0) + \int_0^t \left[\int_0^1 \varepsilon^2(x,\tau)\,\mathrm{d}x\right]\mathrm{d}\tau\right\}^{1/2}. \tag{2.188}$$

Odhad tohoto typu bychom rádi získali pro řešení soustavy (2.132), (2.135). Jak už jsme řekli, budeme postupovat opět dvoufázově, tj. odhadneme zvlášť případ $\delta_i^{(l)} = 0$ a případ $\varepsilon_k^{(l)} = 0$. Dříve než zformulujeme tvrzení, která budou paralelní k větám 2.9 a 2.10, dokážeme ještě tři pomocná tvrzení, se dvěma z nichž jsme se už v podstatě setkali v kap. II.

Lemma 2.6. *Buď p libovolná funkce definovaná v intervalu $\langle 0,1\rangle$ a buďte y_k*

a z_k libovolné funkce definované pro $k = 0,\ldots,n$. Pak platí

$$h\sum_{k=1}^{n-1}\frac{1}{h^2}\{p(x_k - h/2)y_{k-1} - \tag{2.189}$$
$$- [p(x_k - h/2) + p(x_k + h/2)]y_k + p(x_k + h/2)y_{k+1}\}z_k =$$
$$= -h\sum_{k=1}^{n} p(x_k - h/2)\frac{y_k - y_{k-1}}{h}\frac{z_k - z_{k-1}}{h} +$$
$$+ p(x_n - h/2)\frac{y_n - y_{n-1}}{h}z_n - p(x_0 + h/2)\frac{y_1 - y_0}{h}z_0.$$

D ů k a z tohoto lemmatu je analogický jako důkaz lemmatu 3.11 z kap. II (viz str. 184), a přenecháme jej proto čtenáři.

Lemma 2.7. *Nechť koeficienty p a $\beta^{(i)}$ dané diferenciální rovnice a okrajových podmínek splňují nerovnosti* (2.115) a (2.116). *Pak existuje konstanta M taková, že pro libovolnou funkci $\chi_k^{(l)}$ definovanou na síti $\bar{R}_{h,j\tau}$ platí nerovnost*

$$[\chi_k^{(l)}]^2 \leqq M D(\chi_k^{(l)}), \tag{2.190}$$

kde

$$D(\chi_k^{(l)}) = h\sum_{k=1}^{n} p(x_k - h/2, t_l)\left[\frac{\chi_k^{(l)} - \chi_{k-1}^{(l)}}{h}\right]^2 + \tag{2.191}$$
$$+ \beta^{(0)}(t_l)[\chi_0^{(l)}]^2 + \beta^{(1)}(t_l)[\chi_n^{(l)}]^2.$$

D ů k a z . Tvrzení tohoto lemmatu je v podstatě identické s tvrzením lemmatu 3.12 z kap. II (viz str. 185).

Lemma 2.8. *Buď g funkce definovaná a spojitě diferencovatelná v intervalu $\langle 0,T\rangle$ a nechť pro ni v tomto intervalu platí $g(t) \geqq g_0 > 0$, $|g'(t)| \leqq g_1$. Pak pro libovolnou funkci $v^{(l)}$ definovanou pro $l = 0,\ldots,r$ a pro libovolné α, $1/2 \leqq \alpha \leqq 1$, platí*

$$|[\alpha g(t_l) - (1-\alpha)g(t_{l-1})]v^{(l)}v^{(l-1)}| \leqq \tag{2.192}$$
$$\leqq \tfrac{1}{2}\left(2\alpha - 1 + \alpha\frac{g_1}{g_0}\tau\right)\{g(t_l)[v^{(l)}]^2 + g(t_{l-1})[v^{(l-1)}]^2\}.$$

D ů k a z . Z učiněných předpokladů plyne především, že platí

$$\left|\frac{g(t_l) - g(t_{l-1})}{\tau}\right| \leqq g_1. \tag{2.193}$$

Protože je $g(t_l) \geqq g_0$ a $g(t_{l-1}) \geqq g_0$, je

$$[g(t_l)]^{1/2}[g(t_{l-1})]^{1/2} \geqq g_0, \tag{2.194}$$

a tedy

$$\left|\frac{g(t_l)-g(t_{l-1})}{\tau}\right| \leqq \frac{g_1}{g_0}[g(t_l)]^{1/2}[g(t_{l-1})]^{1/2}. \tag{2.195}$$

Platí-li $g(t_l) < g(t_{l-1})$, je také

$$[g(t_l)]^{1/2} < [g(t_{l-1})]^{1/2} \leqq \left(1+\tau\frac{g_1}{g_0}\right)[g(t_{l-1})]^{1/2}. \tag{2.196}$$

Je-li však $g(t_l) \geqq g(t_{l-1})$, plyne z nerovnosti (2.195), že platí

$$g(t_l) \leqq g(t_{l-1}) + \tau\frac{g_1}{g_0}[g(t_l)]^{1/2}[g(t_{l-1})]^{1/2}. \tag{2.197}$$

Platí tedy

$$[g(t_l)]^{1/2} \leqq \frac{g(t_{l-1})}{[g(t_l)]^{1/2}} + \tau\frac{g_1}{g_0}[g(t_{l-1})]^{1/2}. \tag{2.198}$$

Za výše uvedeného předpokladu je však $[g(t_{l-1})/g(t_l)]^{1/2} \leqq 1$, a tedy $g(t_{l-1})/[g(t_l)]^{1/2} \leqq [g(t_{l-1})]^{1/2}$. Dosadíme-li z této nerovnosti do nerovnosti (2.198), dostaneme, že platí

$$[g(t_l)]^{1/2} \leqq \left(1+\tau\frac{g_1}{g_0}\right)[g(t_{l-1})]^{1/2}. \tag{2.199}$$

Nerovnost (2.199) tedy platí bez ohledu na relaci mezi čísly $g(t_l)$ a $g(t_{l-1})$. Z nerovností (2.195) a (2.199) plyne, že je

$$\begin{aligned}
&|[\alpha g(t_l) - (1-\alpha)g(t_{l-1})]v^{(l)}v^{(l-1)}| = \\
&= |(2\alpha-1)g(t_l) + (1-\alpha)[g(t_l)-g(t_{l-1})]|\ |v^{(l)}v^{(l-1)}| \leqq \\
&\leqq \Big\{(2\alpha-1)[g(t_l)]^{1/2}\Big(1+\tau\frac{g_1}{g_0}\Big)[g(t_{l-1})]^{1/2} + \\
&\quad + (1-\alpha)\tau\frac{g_1}{g_0}[g(t_l)]^{1/2}[g(t_{l-1})]^{1/2}\Big\}|v^{(l)}v^{(l-1)}| = \\
&= \Big[(2\alpha-1)\Big(1+\tau\frac{g_1}{g_0}\Big) + (1-\alpha)\tau\frac{g_1}{g_0}\Big][g(t_l)]^{1/2}[g(t_{l-1})]^{1/2}|v^{(l)}v^{(l-1)}|.
\end{aligned} \tag{2.200}$$

Odhadneme-li pravou část poslední nerovnosti podle známého vzorce $|ab| \leqq \frac{1}{2}a^2 + \frac{1}{2}b^2$, dostaneme požadovanou nerovnost. Lemma je dokázáno.

Nyní už můžeme zformulovat a dokázat větu, která je analogická k větě 2.9.

Věta 2.11. *Nechť koeficienty dané diferenciální rovnice a okrajových podmínek jsou dostatečně hladké a nechť splňují nerovnosti* (2.115) *a* (2.116). *Nechť dále* $\eta_k^{(l)}$ *je řešení soustavy* (2.132) *s okrajovými podmínkami* (2.135) *s* $\delta_i^{(l)} \equiv 0$ *a nechť je*

$1/2 \leqq \alpha \leqq 1$. *Pak existují konstanty* M *a* $\tau_0 > 0$ *takové, že pro* $\tau \leqq \tau_0$ *platí*

$$\begin{aligned}
&\max_{k=0,\dots,n}|\eta_k^{(l)}| \leqq \\
&\leqq M\Big\{D(\eta_k^{(0)}) + \sum_{\nu=1}^{l}\tau\Big[\frac{2}{h}|\varepsilon_0^{(\nu)}|^2 + h\sum_{k=1}^{n-1}|\varepsilon_k^{(\nu)}|^2 + \frac{2}{h}|\varepsilon_n^{(\nu)}|^2\Big]\Big\}^{1/2}.
\end{aligned} \tag{2.201}$$

D ů k a z . Vynásobme každou z rovnic (2.132) při pevném l výrazem $h[\eta_k^{(l)} - \eta_k^{(l-1)}]$ a vzniklé rovnice sečtěme od $k=1$ do $k=n-1$. Dostaneme

$$\begin{aligned}
&\tau h\sum_{k=1}^{n-1}[\alpha c(x_k,t_l) + (1-\alpha)c(x_k,t_{l-1})]\Big[\frac{\eta_k^{(l)}-\eta_k^{(l-1)}}{\tau}\Big]^2 = \\
&- \alpha h\sum_{k=1}^{n-1}\frac{1}{h^2}\{p(x_k-h/2,t_l)\eta_{k-1}^{(l)} - [p(x_k-h/2,t_l)+p(x_k+h/2,t_l)]\eta_k^{(l)} + \\
&+ p(x_k+h/2,t_l)\eta_{k+1}^{(l)}\}[\eta_k^{(l)}-\eta_k^{(l-1)}] - \\
&- (1-\alpha)h\sum_{k=1}^{n-1}\frac{1}{h^2}\{p(x_k-h/2,t_{l-1})\eta_{k-1}^{(l-1)} - [p(x_k-h/2,t_{l-1})+ \\
&+ p(x_k+h/2,t_{l-1})]\eta_k^{(l-1)} + p(x_k+h/2,t_{l-1})\eta_{k+1}^{(l-1)}\}[\eta_k^{(l)}-\eta_k^{(l-1)}] + \\
&+ \alpha h\sum_{k=1}^{n-1}q(x_k,t_l)\eta_k^{(l)}[\eta_k^{(l)}-\eta_k^{(l-1)}] + \\
&+ (1-\alpha)h\sum_{k=1}^{n-1}q(x_k,t_{l-1})\eta_k^{(l-1)}[\eta_k^{(l)}-\eta_k^{(l-1)}] = \\
&= h\sum_{k=1}^{n-1}\varepsilon_k^{(l)}[\eta_k^{(l)}-\eta_k^{(l-1)}].
\end{aligned} \tag{2.202}$$

Upravíme-li tuto rovnost pomocí lemmatu 2.6, máme

$$\begin{aligned}
&\tau h\sum_{k=1}^{n-1}[\alpha c(x_k,t_l) + (1-\alpha)c(x_k,t_{l-1})]\Big[\frac{\eta_k^{(l)}-\eta_k^{(l-1)}}{\tau}\Big]^2 + \\
&+ \alpha h\sum_{k=1}^{n}p(x_k-h/2,t_l)\Big[\frac{\eta_k^{(l)}-\eta_{k-1}^{(l)}}{h}\Big]^2 - \\
&- \alpha h\sum_{k=1}^{n}p(x_k-h/2,t_l)\frac{\eta_k^{(l)}-\eta_{k-1}^{(l)}}{h}\frac{\eta_k^{(l-1)}-\eta_{k-1}^{(l-1)}}{h} - \\
&- \alpha p(x_n-h/2,t_l)\frac{\eta_n^{(l)}-\eta_{n-1}^{(l)}}{h}[\eta_n^{(l)}-\eta_n^{(l-1)}] + \\
&+ \alpha p(x_0+h/2,t_l)\frac{\eta_1^{(l)}-\eta_0^{(l)}}{h}[\eta_0^{(l)}-\eta_0^{(l-1)}] -
\end{aligned} \tag{2.203}$$

$$
\begin{aligned}
&- (1-\alpha)h\sum_{k=1}^{n} p(x_k - h/2, t_{l-1})\Big[\frac{\eta_k^{(l-1)} - \eta_{k-1}^{(l-1)}}{h}\Big]^2 + \\
&+ (1-\alpha)h\sum_{k=1}^{n} p(x_k - h/2, t_{l-1})\frac{\eta_k^{(l)} - \eta_{k-1}^{(l)}}{h}\frac{\eta_k^{(l-1)} - \eta_{k-1}^{(l-1)}}{h} - \\
&- (1-\alpha)p(x_n - h/2, t_{l-1})\frac{\eta_n^{(l-1)} - \eta_{n-1}^{(l-1)}}{h}[\eta_n^{(l)} - \eta_n^{(l-1)}] + \\
&+ (1-\alpha)p(x_0 + h/2, t_{l-1})\frac{\eta_1^{(l-1)} - \eta_0^{(l-1)}}{h}[\eta_0^{(l)} - \eta_0^{(l-1)}] + \\
&+ \tau h\sum_{k=1}^{n-1}[\alpha q(x_k, t_l)\eta_k^{(l)} + (1-\alpha)q(x_k, t_{l-1})\eta_k^{(l-1)} - \varepsilon_k^{(l)}]\cdot \\
&\cdot\frac{\eta_k^{(l)} - \eta_k^{(l-1)}}{\tau} = 0.
\end{aligned}
$$

V důsledku okrajových podmínek (2.135) je ($\delta_i^{(l)} \equiv 0$!)

$$
\begin{aligned}
(2.204)\ &\alpha p(x_0 + h/2, t_l)\frac{\eta_1^{(l)} - \eta_0^{(l)}}{h} + (1-\alpha)p(x_0 + h/2, t_{l-1})\frac{\eta_1^{(l-1)} - \eta_0^{(l-1)}}{h} = \\
&= \tfrac{1}{2}h[\alpha c(x_0, t_l) + (1-\alpha)c(x_0, t_{l-1})]\frac{\eta_0^{(l)} - \eta_0^{(l-1)}}{\tau} + \\
&+ \alpha\beta^{(0)}(t_l)\eta_0^{(l)} + (1-\alpha)\beta^{(0)}(t_{l-1})\eta_0^{(l-1)} + \\
&+ \tfrac{1}{2}h\alpha q(x_0, t_l)\eta_0^{(l)} + \tfrac{1}{2}h(1-\alpha)q(x_0, t_{l-1})\eta_0^{(l-1)} - \varepsilon_0^{(l)}
\end{aligned}
$$

a

$$
\begin{aligned}
(2.205)\quad &- \Big[\alpha p(x_n - h/2, t_l)\frac{\eta_n^{(l)} - \eta_{n-1}^{(l)}}{h} + \\
&+ (1-\alpha)p(x_n - h/2, t_{l-1})\frac{\eta_n^{(l-1)} - \eta_{n-1}^{(l-1)}}{h}\Big] = \\
&= \tfrac{1}{2}h[\alpha c(x_n, t_l) + (1-\alpha)c(x_n, t_{l-1})]\frac{\eta_n^{(l)} - \eta_n^{(l-1)}}{\tau} + \\
&+ \alpha\beta^{(1)}(t_l)\eta_n^{(l)} + (1-\alpha)\beta^{(1)}(t_{l-1})\eta_n^{(l-1)} + \\
&+ \tfrac{1}{2}h\alpha q(x_n, t_l)\eta_n^{(l)} + \tfrac{1}{2}h(1-\alpha)q(x_n, t_{l-1})\eta_n^{(l-1)} - \varepsilon_n^{(l)}.
\end{aligned}
$$

Dosadíme-li do identity (2.203) ze vztahů (2.204) a (2.205), dostaneme

$$
\begin{aligned}
(2.206)\quad &\tau h\sum_{k=1}^{n-1}[\alpha c(x_k, t_l) + (1-\alpha)c(x_k, t_{l-1})]\Big[\frac{\eta_k^{(l)} - \eta_k^{(l-1)}}{\tau}\Big]^2 + \\
&+ \alpha h\sum_{k=1}^{n} p(x_k - h/2, t_l)\Big[\frac{\eta_k^{(l)} - \eta_{k-1}^{(l)}}{h}\Big]^2 -
\end{aligned}
$$

$$
\begin{aligned}
&- (1-\alpha)h\sum_{k=1}^{n} p(x_k - h/2, t_{l-1})\Big[\frac{\eta_k^{(l-1)} - \eta_{k-1}^{(l-1)}}{h}\Big]^2 - \\
&- h\sum_{k=1}^{n}[\alpha p(x_k - h/2, t_l) - (1-\alpha)p(x_k - h/2, t_{l-1})]\cdot \\
&\cdot\frac{\eta_k^{(l)} - \eta_{k-1}^{(l)}}{h}\frac{\eta_k^{(l-1)} - \eta_{k-1}^{(l-1)}}{h} + \\
&+ \tfrac{1}{2}\tau h[\alpha c(x_n, t_l) + (1-\alpha)c(x_n, t_{l-1})]\Big[\frac{\eta_n^{(l)} - \eta_n^{(l-1)}}{\tau}\Big]^2 + \\
&+ \alpha\beta^{(1)}(t_l)[\eta_n^{(l)}]^2 + (1-\alpha)\beta^{(1)}(t_{l-1})\eta_n^{(l)}\eta_n^{(l-1)} - \\
&- \alpha\beta^{(1)}(t_l)\eta_n^{(l)}\eta_n^{(l-1)} - (1-\alpha)\beta^{(1)}(t_{l-1})[\eta_n^{(l-1)}]^2 + \\
&+ \tfrac{1}{2}\tau h[\alpha q(x_n, t_l)\eta_n^{(l)} + (1-\alpha)q(x_n, t_{l-1})\eta_n^{(l-1)}]\frac{\eta_n^{(l)} - \eta_n^{(l-1)}}{\tau} - \\
&- \tau\varepsilon_n^{(l)}\frac{\eta_n^{(l)} - \eta_n^{(l-1)}}{\tau} + \\
&+ \tfrac{1}{2}\tau h[\alpha c(x_0, t_l) + (1-\alpha)c(x_0, t_{l-1})]\Big[\frac{\eta_0^{(l)} - \eta_0^{(l-1)}}{\tau}\Big]^2 + \\
&+ \alpha\beta^{(0)}(t_l)[\eta_0^{(l)}]^2 + (1-\alpha)\beta^{(0)}(t_{l-1})\eta_0^{(l)}\eta_0^{(l-1)} - \\
&- \alpha\beta^{(0)}(t_l)\eta_0^{(l)}\eta_0^{(l-1)} - (1-\alpha)\beta^{(0)}(t_{l-1})[\eta_0^{(l-1)}]^2 + \\
&+ \tfrac{1}{2}\tau h[\alpha q(x_0, t_l)\eta_0^{(l)} + (1-\alpha)q(x_0, t_{l-1})\eta_0^{(l-1)}]\frac{\eta_0^{(l)} - \eta_0^{(l-1)}}{\tau} - \\
&- \tau\varepsilon_0^{(l)}\frac{\eta_0^{(l)} - \eta_0^{(l-1)}}{\tau} + \\
&+ \tau h\sum_{k=1}^{n}[\alpha q(x_k, t_l)\eta_k^{(l)} + (1-\alpha)q(x_k, t_{l-1})\eta_k^{(l-1)} - \\
&- \varepsilon_k^{(l)}]\frac{\eta_k^{(l)} - \eta_k^{(l-1)}}{\tau} = 0.
\end{aligned}
$$

Poslední rovnost pišme ve tvaru

$$
(2.207)\qquad S^{(l)} + \alpha D(\eta_k^{(l)}) = (1-\alpha)D(\eta_k^{(l-1)}) + Q^{(l)} + R^{(l)},
$$

kde $D(\eta_k^{(l)})$ je definováno rovnicí (2.191),

$$
\begin{aligned}
(2.208)\quad S^{(l)} = \tau\Big\{&\tfrac{1}{2}h[\alpha c(x_0, t_l) + (1-\alpha)c(x_0, t_{l-1})]\Big[\frac{\eta_0^{(l)} - \eta_0^{(l-1)}}{\tau}\Big]^2 + \\
&+ h\sum_{k=1}^{n-1}[\alpha c(x_k, t_l) + (1-\alpha)c(x_k, t_{l-1})]\Big[\frac{\eta_k^{(l)} - \eta_k^{(l-1)}}{\tau}\Big]^2 +
\end{aligned}
$$

$$+\tfrac{1}{2}h[\alpha c(x_n,t_l)+(1-\alpha)c(x_n,t_{l-1})]\Big[\frac{\eta_n^{(l)}-\eta_n^{(l-1)}}{\tau}\Big]^2\Big\},$$

(2.209) $$Q^{(l)} = -\tau\Big\{\tfrac{1}{2}h[\alpha q(x_0,t_l)\eta_0^{(l)}+(1-\alpha)q(x_0,t_{l-1})\eta_0^{(l-1)}]\frac{\eta_0^{(l)}-\eta_0^{(l-1)}}{\tau}+$$
$$+h\sum_{k=1}^{n-1}[\alpha q(x_k,t_l)\eta_k^{(l)}+(1-\alpha)q(x_k,t_{l-1})\eta_k^{(l-1)}]\frac{\eta_k^{(l)}-\eta_k^{(l-1)}}{\tau}+$$
$$+\tfrac{1}{2}h[\alpha q(x_n,t_l)\eta_n^{(l)}+(1-\alpha)q(x_n,t_{l-1})\eta_n^{(l-1)}]\frac{\eta_n^{(l)}-\eta_n^{(l-1)}}{\tau}\Big\}+$$
$$+\tau\Big[\varepsilon_0^{(l)}\frac{\eta_0^{(l)}-\eta_0^{(l-1)}}{\tau}+h\sum_{k=1}^{n-1}\varepsilon_k^{(l)}\frac{\eta_k^{(l)}-\eta_k^{(l-1)}}{\tau}+\varepsilon_n^{(l)}\frac{\eta_n^{(l)}-\eta_n^{(l-1)}}{\tau}\Big]$$

a

(2.210) $$R^{(l)} = h\sum_{k=1}^{n}[\alpha p(x_k-h/2,t_l)-$$
$$-(1-\alpha)p(x_k-h/2,t_{l-1})]\frac{\eta_k^{(l)}-\eta_{k-1}^{(l)}}{h}\frac{\eta_k^{(l-1)}-\eta_{k-1}^{(l-1)}}{h}+$$
$$+[\alpha\beta^{(0)}(t_l)-(1-\alpha)\beta^{(0)}(t_{l-1})]\eta_0^{(l)}\eta_0^{(l-1)}+$$
$$+[\alpha\beta^{(1)}(t_l)-(1-\alpha)\beta^{(1)}(t_{l-1})]\eta_n^{(l)}\eta_n^{(l-1)}.$$

Odhadněme nyní shora pravou stranu výrazu (2.207). Začněme s $R^{(l)}$. Vzhledem k předpokladu o dostatečné hladkosti koeficientů a vzhledem k nerovnostem (2.115) a (2.116) splňují funkce p a $\beta^{(i)}$ předpoklady lemmatu 2.8. Platí tedy

(2.211) $$\Big|[\alpha p(x_k-h/2,t_l)-$$
$$-(1-\alpha)p(x_k-h/2,t_{l-1})]\frac{\eta_k^{(l)}-\eta_{k-1}^{(l)}}{h}\frac{\eta_k^{(l-1)}-\eta_{k-1}^{(l-1)}}{h}\Big|\leqq$$
$$\leqq\tfrac{1}{2}(2\alpha-1+\alpha M_1\tau)\Big\{p(x_k-h/2,t_l)\Big[\frac{\eta_k^{(l)}-\eta_{k-1}^{(l)}}{h}\Big]^2+$$
$$+p(x_k-h/2,t_{l-1})\Big[\frac{\eta_k^{(l-1)}-\eta_{k-1}^{(l-1)}}{h}\Big]^2\Big\}$$

a

(2.212) $$|[\alpha\beta^{(i)}(t_l)-(1-\alpha)\beta^{(i)}(t_{l-1})]\eta_{in}^{(l)}\eta_{in}^{(l-1)}|\leqq$$
$$\leqq\tfrac{1}{2}(2\alpha-1+\alpha M_1\tau)\{\beta^{(i)}(t_l)[\eta_{in}^{(l)}]^2+\beta^{(i)}(t_{l-1})[\eta_{in}^{(l-1)}]^2\}$$

pro $i=0,1$, kde M_1 je kladná konstanta. Z nerovností (2.211) a (2.212) však už plyne, že platí

(2.213) $$|R^{(l)}|\leqq\tfrac{1}{2}(2\alpha-1+\alpha M_1\tau)[D^{(l)}+D^{(l-1)}].$$

Zde jsme položili pro stručnost $D^{(l)}=D(\eta_k^{(l)})$.

Obraťme se dále k odhadu veličiny $Q^{(l)}$. Abychom to provedli, uvědomíme se především, že nerovnost

(2.214) $$|ab|\leqq\frac{1}{\gamma}a^2+\tfrac{1}{4}\gamma b^2$$

platí pro libovolná reálná a a b a pro libovolné $\gamma>0$. Skutečně, tato nerovnost plyne ihned ze zřejmé nerovnosti

(2.215) $$\Big(\frac{a}{\gamma^{1/2}}\pm\tfrac{1}{2}\gamma^{1/2}b\Big)^2\geqq 0.$$

Užitím nerovnosti (2.214) snadno zjistíme, že platí

(2.116) $$\Big|q(x_k,t_l)\eta_k^{(l)}\frac{\eta_k^{(l)}-\eta_k^{(l-1)}}{\tau}\Big|\leqq\frac{q_1^2}{c_0}[\eta_k^{(l)}]^2+\tfrac{1}{4}c(x_k,t_l)\Big[\frac{\eta_k^{(l)}-\eta_k^{(l-1)}}{\tau}\Big]^2,$$
$$k=0,\dots,n,$$

(2.217) $$\Big|q(x_k,t_{l-1})\eta_k^{(l-1)}\frac{\eta_k^{(l)}-\eta_k^{(l-1)}}{\tau}\Big|\leqq$$
$$\leqq\frac{q_1^2}{c_0}[\eta_k^{(l-1)}]^2+\tfrac{1}{4}c(x_k,t_{l-1})\Big[\frac{\eta_k^{(l)}-\eta_k^{(l-1)}}{\tau}\Big]^2,\quad k=0,\dots,n,$$

(2.218) $$\Big|\varepsilon_k^{(l)}\frac{\eta_k^{(l)}-\eta_k^{(l-1)}}{\tau}\Big|\leqq\frac{1}{c_0}[\varepsilon_k^{(l)}]^2+$$
$$+\tfrac{1}{4}[\alpha c(x_k,t_l)+(1-\alpha)c(x_k,t_{l-1})]\Big[\frac{\eta_k^{(l)}-\eta_k^{(l-1)}}{\tau}\Big]^2,$$
$$k=1,\dots,n-1,$$

(2.219) $$\Big|\varepsilon_0^{(l)}\frac{\eta_0^{(l)}-\eta_0^{(l-1)}}{\tau}\Big|\leqq\frac{1}{c_0}\frac{2}{h}[\varepsilon_0^{(l)}]^2+$$
$$+\tfrac{1}{8}h[\alpha c(x_0,t_l)+(1-\alpha)c(x_0,t_{l-1})]\Big[\frac{\eta_0^{(l)}-\eta_0^{(l-1)}}{\tau}\Big]^2$$

a

(2.220) $$\Big|\varepsilon_n^{(l)}\frac{\eta_n^{(l)}-\eta_n^{(l-1)}}{\tau}\Big|\leqq\frac{1}{c_0}\frac{2}{h}[\varepsilon_n^{(l)}]^2+$$
$$+\tfrac{1}{8}h[\alpha c(x_n,t_l)+(1-\alpha)c(x_nk,t_{l-1})]\Big[\frac{\eta_n^{(l)}-\eta_n^{(l-1)}}{\tau}\Big]^2.$$

Použitím nerovností (2.216) až (2.220) dostáváme

$$
\begin{aligned}
(2.221)\quad |Q^{(l)}| \leqq{}& \tau h \frac{q_1^2}{c_0} \sum_{k=1}^{n-1} \{\alpha [\eta_k^{(l)}]^2 + (1-\alpha)[\eta_k^{(l-1)}]^2\} + \\
&+ \tfrac{1}{4}\tau h \sum_{k=1}^{n-1} [\alpha c(x_k, t_l) + (1-\alpha) c(x_k, t_{l-1})] \Big[\frac{\eta_k^{(l)} - \eta_k^{(l-1)}}{\tau}\Big]^2 + \\
&+ \tfrac{1}{2}\tau h \frac{q_1^2}{c_0} \{\alpha [\eta_0^{(l)}]^2 + (1-\alpha)[\eta_0^{(l-1)}]^2\} + \\
&+ \tfrac{1}{2}\tau h \tfrac{1}{4} [\alpha c(x_0, t_l) + (1-\alpha) c(x_0, t_{l-1})] \Big[\frac{\eta_0^{(l)} - \eta_0^{(l-1)}}{\tau}\Big]^2 + \\
&+ \tfrac{1}{2}\tau h \frac{q_1^2}{c_0} \{\alpha [\eta_n^{(l)}]^2 + (1-\alpha)[\eta_n^{(l-1)}]^2\} + \\
&+ \tfrac{1}{2}\tau h \tfrac{1}{4} [\alpha c(x_n, t_l) + (1-\alpha) c(x_n, t_{l-1})] \Big[\frac{\eta_n^{(l)} - \eta_n^{(l-1)}}{\tau}\Big]^2 + \\
&+ \frac{\tau}{c_0} \frac{2}{h} [\varepsilon_0^{(l)}]^2 + \frac{\tau h}{c_0} \sum_{k=1}^{n-1} [\varepsilon_k^{(l)}]^2 + \frac{\tau}{c_0} \frac{2}{h} [\varepsilon_n^{(l)}]^2 + \\
&+ \tfrac{1}{8}\tau h [\alpha c(x_0, t_l) + (1-\alpha) c(x_0, t_{l-1})] \Big[\frac{\eta_0^{(l)} - \eta_0^{(l-1)}}{\tau}\Big]^2 + \\
&+ \tfrac{1}{4}\tau h \sum_{k=1}^{n-1} [\alpha c(x_k, t_l) + (1-\alpha) c(x_k, t_{l-1})] \Big[\frac{\eta_k^{(l)} - \eta_k^{(l-1)}}{\tau}\Big]^2 + \\
&+ \tfrac{1}{8}\tau h [\alpha c(x_n, t_l) + (1-\alpha) c(x_n, t_{l-1})] \Big[\frac{\eta_n^{(l)} - \eta_n^{(l-1)}}{\tau}\Big]^2 .
\end{aligned}
$$

Tuto nerovnost však můžeme přepsat do tvaru

$$
\begin{aligned}
(2.222)\quad |Q^{(l)}| \leqq{}& \tfrac{1}{2} S^{(l)} + \tau \frac{q_1^2}{c_0} \alpha \Big\{ \tfrac{1}{2} h [\eta_0^{(l)}]^2 + h \sum_{k=1}^{n-1} [\eta_k^{(l)}]^2 + \tfrac{1}{2} h [\eta_n^{(l)}]^2 \Big\} + \\
&+ \tau \frac{q_1^2}{c_0} (1-\alpha) \Big\{ \tfrac{1}{2} h [\eta_0^{(l-1)}]^2 + h \sum_{k=1}^{n-1} [\eta_k^{(l-1)}]^2 + \tfrac{1}{2} h [\eta_n^{(l-1)}]^2 \Big\} + \\
&+ \frac{\tau}{c_0} E^{(l)},
\end{aligned}
$$

kde jsme položili

$$
(2.223)\qquad E^{(l)} = \frac{2}{h} [\varepsilon_0^{(l)}]^2 + h \sum_{k=1}^{n-1} [\varepsilon_k^{(l)}]^2 + \frac{2}{h} [\varepsilon_n^{(l)}]^2 .
$$

Podle lemmatu 2.7 je

$$
(2.224)\qquad \tfrac{1}{2} h [\eta_0^{(l)}]^2 + h \sum_{k=1}^{n-1} [\eta_k^{(l)}]^2 + \tfrac{1}{2} h [\eta_n^{(l)}]^2 \leqq M D^{(l)} .
$$

Dosadíme-li z této nerovnosti do nerovnosti (2.222), máme

$$
\begin{aligned}
(2.225)\qquad |Q^{(l)}| \leqq{}& \tfrac{1}{2} S^{(l)} + \tau \alpha \frac{q_1^2}{c_0} M D^{(l)} + \\
&+ \tau (1-\alpha) \frac{q_1^2}{c_0} M D^{(l-1)} + \frac{\tau}{c_0} E^{(l)} .
\end{aligned}
$$

Použijeme-li nyní v rovnici (2.207) nerovnosti (2.213) a (2.225) a vezmeme-li v úvahu, že je $S^{(l)} \geqq 0$ pro každé l, dostaneme

$$
\begin{aligned}
(2.226)\qquad & \tfrac{1}{2} \Big\{ 1 - \Big[\alpha M_1 + \frac{2 q_1^2}{c_0} \alpha M \Big] \tau \Big\} D^{(l)} \leqq \\
& \leqq \tfrac{1}{2} \Big\{ 1 + \Big[\alpha M_1 + \frac{2 q_1^2}{c_0} (1-\alpha) M \Big] \tau \Big\} D^{(l-1)} + \frac{\tau}{c_0} E^{(l)} ,
\end{aligned}
$$

neboli, protože je $\alpha \geqq 1/2$,

$$
(2.227)\qquad (1 - \tau \alpha M_2) D^{(l)} \leqq (1 + \tau \alpha M_2) D^{(l-1)} + \frac{2\tau}{c_0} E^{(l)} ,
$$

kde

$$
(2.228)\qquad M_2 = M_1 + \frac{2 q_1^2}{c_0} M .
$$

Pro $\tau \leqq \tau_0 < 1/(\alpha M_2)$ plyne z nerovnosti (2.227) nerovnost

$$
(2.229)\qquad D^{(l)} \leqq \frac{1 + \tau \alpha M_2}{1 - \tau \alpha M_2} D^{(l-1)} + \tau M_3 E^{(l)} ,
$$

kde

$$
(2.230)\qquad M_3 = \frac{2}{c_0 (1 - \alpha M_2 \tau_0)} .
$$

Z nerovnosti (2.229) pak už dostáváme, že je

$$
(2.231)\qquad D^{(l)} \leqq \Big(\frac{1 + \tau \alpha M_2}{1 - \tau \alpha M_2} \Big)^l D^{(0)} + \tau M_3 \sum_{\nu=1}^{l} \Big(\frac{1 + \tau \alpha M_2}{1 - \tau \alpha M_2} \Big)^{l-\nu} E^{(\nu)} ,
$$

jak se snadno přesvědčíme úplnou indukcí. Pro $\tau \leqq \tau_0$ je však $(1 + \tau \alpha M_2)/(1 - \tau \alpha M_2) > 1$, a tedy platí

$$
(2.232)\qquad \Big(\frac{1 + \tau \alpha M_2}{1 - \tau \alpha M_2} \Big)^s \leqq \Big(\frac{1 + \tau \alpha M_2}{1 - \tau \alpha M_2} \Big)^r
$$

pro $s = 0, \ldots, r$. Konečně pro dostatečně malá τ platí

$$
(2.233)\qquad \Big(\frac{1 + \tau \alpha M_2}{1 - \tau \alpha M_2} \Big)^r = \Big(\frac{1 + \tau \alpha M_2}{1 - \tau \alpha M_2} \Big)^{T/\tau} \leqq M_3 \mathrm{e}^{2 \alpha M_2 T} ,
$$

neboť je

$$
(2.234)\qquad \lim_{\tau \to 0} \Big(\frac{1 + \tau \alpha M_2}{1 - \tau \alpha M_2} \Big)^{T/\tau} = \mathrm{e}^{2 \alpha M_2 T} .
$$

Dosadíme-li konečně odhad (2.232) a (2.233) do nerovnosti (2.231) a použijeme-li lemmatu 2.7, dostaneme nerovnost (2.201). Věta je dokázána.

Dále je třeba dokázat ještě větu, která je paralelní k větě 2.10.

Věta 2.12. *Nechť koeficienty dané diferenciální rovnice a okrajových podmínek jsou dostatečně hladké a nechť jsou splněny nerovnosti* (2.115) a (2.116). *Nechť dále* $\eta_k^{(l)}$ *je řešením soustavy* (2.132) *s nulovou počáteční podmínkou a s okrajovými podmínkami* (2.135) *s* $\varepsilon_k^{(l)} \equiv 0$. *Pak existuje konstanta* M *(nezávislá na* h *a* τ*) taková, že pro každé dostatečně malé* τ *platí*

$$\max_{k=0,\dots,n} |\eta_k^{(l)}| \leqq M \left\{ \max \left[|\delta_0^{(l)}|^2, |\delta_1^{(l)}|^2, \max(|\delta_0^{(0)}|^2, |\delta_1^{(0)}|^2) + \right.\right. \tag{2.235}$$
$$\left.\left. + \tau \sum_{\nu=1}^{l} \max \left(|\delta_0^{(\nu)}|^2, |\delta_1^{(\nu)}|^2, |\delta_0^{(\nu-1)}|^2, |\delta_1^{(\nu-1)}|^2, \right.\right.\right.$$
$$\left.\left.\left. \left| \frac{\delta_0^{(\nu)} - \delta_0^{(\nu-1)}}{\tau} \right|^2, \left| \frac{\delta_1^{(\nu)} - \delta_1^{(\nu-1)}}{\tau} \right|^2 \right) \right] \right\}^{1/2}.$$

D ů k a z je podobný důkazu věty 2.10. Veličinu $\eta_k^{(l)}$, kterou máme odhadnout, pišme ve tvaru

$$\eta_k^{(l)} = \varphi_k^{(l)} + \psi_k^{(l)}, \tag{2.236}$$

kde $\varphi_k^{(l)}$ je řešením soustavy (2.165) z lemmatu 2.5. Podle tohoto lemmatu platí

$$\max_{k=0,\dots,n} |\varphi_k^{(l)}| \leqq M \{\max[|\delta_0^{(l)}|^2, |\delta_1^{(l)}|^2]\}^{1/2}. \tag{2.237}$$

Funkce $\psi_k^{(l)}$ splňuje soustavu

$$\begin{aligned} (L_{h,\tau}^{(\alpha)} \psi)_k^{(l)} &= \varepsilon_k^{(l)}, \\ \psi_k^{(0)} &= -\varphi_k^{(0)}, \\ (l_{h,\tau}^{(0,\alpha)} \psi)^{(l)} &= \varepsilon_0^{(l)}, \\ (l_{h,\tau}^{(1,\alpha)} \psi)^{(l)} &= \varepsilon_n^{(l)}, \end{aligned} \tag{2.238}$$

kde je

$$\begin{aligned} \varepsilon_k^{(l)} &= (L_{h,\tau}^{(\alpha)} \eta)_k^{(l)} - (L_{h,\tau}^{(\alpha)} \varphi)_k^{(l)} = \\ &= -[\alpha c(x_k, t_l) + (1-\alpha) c(x_k, t_{l-1})] \frac{\varphi_k^{(l)} - \varphi_k^{(l-1)}}{\tau} - \\ &\quad - \alpha q(x_k, t_l) \varphi_k^{(l)} - (1-\alpha) q(x_k, t_{l-1}) \varphi_k^{(l-1)}, \\ &\quad k = 1, \dots, n-1, \end{aligned} \tag{2.239}$$

$$\begin{aligned} \varepsilon_0^{(l)} &= (l_{h,\tau}^{(0,\alpha)} \eta)^{(l)} - (l_{h,\tau}^{(0,\alpha)} \varphi)^{(l)} = \\ &= -\tfrac{1}{2} h [\alpha c(x_0, t_l) + (1-\alpha) c(x_0, t_{l-1})] \frac{\varphi_0^{(l)} - \varphi_0^{(l-1)}}{\tau} - \\ &\quad - \tfrac{1}{2} h \alpha q(x_0, t_l) \varphi_0^{(l)} - \tfrac{1}{2} h (1-\alpha) q(x_0, t_{l-1}) \varphi_0^{(l-1)} \end{aligned} \tag{2.240}$$

a analogická rovnice platí pro $\varepsilon_n^{(l)}$. Položíme-li tedy

$$E^{(l)} = \max \left[|\delta_0^{(l)}|^2, |\delta_1^{(l)}|^2, |\delta_0^{(l-1)}|^2, |\delta_1^{(l-1)}|^2, \right. \tag{2.241}$$
$$\left. \left| \frac{\delta_0^{(l)} - \delta_0^{(l-1)}}{\tau} \right|^2, \left| \frac{\delta_1^{(l)} - \delta_1^{(l-1)}}{\tau} \right|^2 \right],$$

je znovu podle lemmatu 2.5

$$|\varepsilon_k^{(l)}| \leqq M [E^{(l)}]^{1/2}, \quad k = 1, \dots, n-1, \; l = 1, \dots, r, \tag{2.242}$$

a

$$|\varepsilon_0^{(l)}| \leqq h M [E^{(l)}]^{1/2}, \quad |\varepsilon_n^{(l)} \leqq h M [E^{(l)}]^{1/2}, \tag{2.243}$$

Použijeme-li k odhadu funkce $\psi_k^{(l)}$ větu 2.11, máme

$$\max_{k=0,\dots,n} |\psi_k^{(l)}| \leqq \tag{2.244}$$
$$\leqq M \left[D(\varphi_k^{(0)}) + \tau \sum_{\nu=1}^{l} \left(\frac{2}{h} h^2 E^{(\nu)} + h \sum_{k=1}^{n-1} E^{(\nu)} + \frac{2}{h} h^2 E^{(\nu)} \right) \right]^{1/2} \leqq$$
$$\leqq M_1 \left[D(\varphi_k^{(0)}) + \tau \sum_{\nu=1}^{l} E^{(\nu)} \right]^{1/2}$$

Na druhé straně podle definice čísla $D(\varphi_k^{(0)})$ a podle lemmatu 2.5 je

$$D(\varphi_k^{(0)}) \leqq M \max[|\delta_0^{(0)}|^2, |\delta_1^{(0)}|^2]. \tag{2.245}$$

Nerovnosti (2.237), (2.244) a (2.245) však už dokazují větu.

Závěrem tohoto odstavce zformulujme ještě konvergenční větu pro obecnou parabolickou rovnici a pro $1/2 \leqq \alpha \leqq 1$.

Věta 2.13. *Nechť přesné řešení* u *okrajové úlohy* (2.109), (2.111), (2.114) *je dostatečně hladké, nechť jsou splněny nerovnosti* (2.115), (2.116) *a nechť* $u_k^{(l)}$ *je přibližné řešení vypočtené pomocí operátorů* $L_{h,\tau}^{(\alpha)}$, $l_{h,\tau}^{(0,\alpha)}$, $l_{h,\tau}^{(1,\alpha)}$ *s* $1/2 \leqq \alpha \leqq 1$. *Pak existuje konstanta* M *taková, že platí*

$$|u_k^{(l)} - u(x_k, t_l)| \leqq M(\tau + h^2) \tag{2.246}$$

v obecném případě a

$$|u_k^{(l)} - u(x_k, t_l)| \leqq M(\tau^2 + h^2) \tag{2.247}$$

v případě, že je $\alpha = 1/2$.

D ů k a z plyne bezprostředně z vět 2.11 a 2.12. Stačí si pouze uvědomit, že chyba $u_k^{(l)} = u(x_k, t_l)$ splňuje rovnici (2.132) s okrajovými podmínkami (2.135), pro jejichž pravé strany platí odhady (2.136) a (2.137), resp. (2.138) a (2.139) a že v důsledku spojité diferencovatelnosti funkcí v_i jsou výrazy $[\delta_i^{(l)} - \delta_i^{(l-1)}]/\tau$ řádu $O(h^2)$.

Crankova-Nicolsonova metoda vede tedy i v obecném případě k nejefektivnějšímu algoritmu, neboť její celková diskretizační chyba je řádu $O(\tau^2 + h^2)$, takže lze opět volit $\tau = O(h)$. Upozorněme také na to, že věta 2.13 je obecnější než věta 2.8 nejen proto, že se v ní vyšetřuje obecná parabolická rovnice, ale také proto, že chyba se v ním měří $\mathscr{L}_\infty$-normou.

2.3 Dvou- a vícedimenzionální parabolické rovnice

V předchozích odstavcích jsme viděli, že problematika metody sítí při řešení parabolických parciálních diferenciálních rovnic v jedné prostorové proměnné je dosti úzce svázána s problematikou řešení okrajových úloh pro obyčejné diferenciální rovnice. Tak např. otázky spojené s přepisem okrajových podmínek se řešily zcela stejně jako v případě obyčejných diferenciálních rovnic. Nové byly pouze problémy vyvolané přítomností význačné proměnné, času. Dá se očekávat, že analogická situace nastane i v případě řešení parabolických diferenciálních rovnic ve dvou (a více) prostorových proměnných. Komplikace proti prostorově jednodimenzionálnímu případu budou souviset tedy zejména s přepisem okrajových podmínek, což je problematika, jíž jsme se zabývali ve třetí kapitole, zatímco parabolický charakter úlohy (tj. přítomnost proměnné t) přinese patrně již jen takové problémy, se kterými jsme se setkali v předchozích odstavcích. Proto budeme postupovat stručně a omezíme se vlastně jen na jednoduché ilustrativní příklady.

2.3.1 Základní metody

Způsob, jakým lze řešit metodu sítí parabolickou parciální rovnici ve dvou a více prostorových proměnných, si ukážeme na Dirichletově úloze pro dvoudimenzionální rovnici pro vedení tepla. Budeme tedy řešit rovnici

$$Lu \equiv \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = 0 \tag{2.248}$$

v $\Omega \times (0,T)$, kde Ω je omezená oblast v $\mathbf{E}^2$, s počáteční podmínkou

$$u(x,y,0) = g(x,y), \quad (x,y) \in \Omega, \tag{2.249}$$

a s okrajovými podmínkami

$$u(x,y,t) = \gamma(x,y,t), \quad (x,y) \in \Gamma,\ t \in (0,T). \tag{2.250}$$

Vzhledem k úvahám, které jsme prováděli v kap. III a vzhledem k postupu v předchozích odstavcích budeme přibližné řešení v bodech (x_k, y_s, t_l), kde $(x_k, y_s) \in \bar{\Omega}^{(h)}$ a $t_l = l\tau$, $l = 0, \dots, r$, $\tau = T/r$, hledat ze soustavy

$$(L_{h,\tau}^{(\alpha)} u)_{ks}^{(l)} = 0, \quad (x_k, y_s) \in \Omega^{(h)},\ l = 1, \dots, r, \tag{2.251}$$

s počátečními podmínkami

$$u_{ks}^{(0)} = g(x_k, y_s), \quad (x_k, y_s) \in \bar{\Omega}^{(h)}, \tag{2.252}$$

s okrajovými podmínkami

$$(l_{h,\tau} u)_{ks}^{(l)} = \Lambda_{h,\tau}(\gamma; x_k, y_s, t_l), \quad (x_k, y_s) \in \Gamma^{(h)},\ l = 1, \dots, r. \tag{2.253}$$

Množiny $\Omega^{(h)}$, $\bar{\Omega}^{(h)}$, $\Gamma^{(h)}$ zde mají samozřejmě stejný význam jako v odst. 2.1 kap. III a $L_{h,\tau}^{(\alpha)}$ je operátor, který funkci $u_{ks}^{(l)}$ definované pro $(x_k, y_s) \in \bar{\Omega}^{(h)}$ a pro $l = 0, \dots, r$ přiřazuje funkci $(L_{h,\tau}^{(\alpha)} u)_k^{(l)}$ definovanou pro $(x_k, y_s) \in \Omega^{(h)}$, $l = 1, \dots, r$ a danou předpisem

$$\begin{aligned}(L_{h,\tau}^{(\alpha)} u)_k^{(l)} = {} & \frac{u_{ks}^{(l)} - u_{ks}^{(l-1)}}{\tau} - \\ & - \frac{\alpha}{h^2}[u_{k+1,s}^{(l)} + u_{k-1,s}^{(l)} + u_{k,s+1}^{(l)} + u_{k,s-1}^{(l)} - 4u_{ks}^{(l)}] - \\ & - \frac{1-\alpha}{h^2}[u_{k+1,s}^{(l-1)} + u_{k-1,s}^{(l-1)} + u_{k,s+1}^{(l-1)} + u_{k,s-1}^{(l-1)} - 4u_{ks}^{(l-1)}].\end{aligned} \tag{2.254}$$

Symbol $l_{h,\tau}$ v okrajové podmínce (2.253) je pak lineární operátor tvaru

$$\begin{aligned}(l_{h,\tau} u)_{ks}^{(l)} = {} & u_{ks}^{(l)} - \chi_1 u_{k_1 s_1}^{(l)} - \chi_2 u_{k_2 s_2}^{(l)} - \\ & - \chi_3 u_{k_3 s_3}^{(l)} - \chi_4 u_{k_4 s_4}^{(l)}, \quad (x_k, y_s) \in \Gamma^{(h)},\end{aligned} \tag{2.255}$$

kde uzly (x_{k_i}, y_{s_i}) jsou sousední k uzlu (x_k, y_s) a platí-li $(x_{k_i}, y_{s_i}) \notin \bar{\Omega}^{(h)}$, je příslušné χ_i rovno nule (vzhledem k definici množiny $\Gamma^{(h)}$ existuje aspoň jedno takové i) a ostatní koeficienty jsou (stejná) nezáporná čísla, přičemž platí

$$\sum_{i=1}^{4} \chi_i \leqq \chi_0 < 1, \tag{2.256}$$

kde χ_0 je konstanta nezávislá na h a τ. Konečně operátory $\Lambda_{h,\tau}$ jsou dány vzorcem

$$\begin{aligned}\Lambda_{h,\tau}(\gamma; x_k, y_s, t_l) = {} & \xi_1 \gamma(x_{k_1'}, y_{s_1'}, t_l) + \xi_2 \gamma(x_{k_2'}, y_{s_2'}, t_l) + \\ & + \xi_3 \gamma(x_{k_3'}, y_{s_3'}, t_l) + \xi_4 \gamma(x_{k_4'}, y_{s_4'}, t_l), \\ & (x_k, y_s) \in \Gamma^{(h)},\ l = 1, \dots, r,\end{aligned} \tag{2.257}$$

kde ξ_i jsou jistá čísla, body $(x_{k_i'}, y_{s_i'})$ jsou průsečíky přímek sítě procházející uzlem (x_k, y_s) s hranicí Γ, jejichž vzdálenost od bodu (x_k, y_s) je menší než h a $\xi_i = 0$, jestliže takový průsečík neexistuje. Upozorněme, že to jsou právě ta i, pro něž je koeficient χ_i v (2.255) od nuly různý (srv. odst. 2.1.2 z kap. III.).

Pro $\alpha = 0$ dostaneme *explicitní metodu*, ve které se hodnoty přibližného řešení v l-tém časovém řádku vypočítávají přímo jako lineární kombinace hodnot přibližného řešení v $(l-1)$-ním časovém řádku. To je sice z numerického hlediska příjemné, hned však uvidíme, že za to musíme zaplatit dosti vážným omezením na velikost časového integračního kroku. *Implicitní metoda*, která vznikne při $\alpha = 1$, touto nevýhodou netrpí, abychom však při její užití dostali přibližné řešení v l-tém časovém řádku, je třeba řešit soustavu $O(1/h^2)$ rovnic o $O(1/h^2)$ neznámých. Podobně je tomu i u ostatních schémat s $\alpha \neq 0$.

Vyšetřeme nyní konvergenci uvedených metod. Začneme explicitní a implicitní metodou, které pojednáme najednou.

Užitím Taylorova vzorce se snadno zjistí, že pro každou dostatečně hladkou funkci u platí

$$(2.258)\qquad (L_{h,\tau}^{(\alpha)} u^{(\mathrm{pr})})_{ks}^{(l)} = \alpha(Lu)(x_k, y_s, t_l) + (1-\alpha)(Lu)(x_k, y_s, t_{l-1}) + \\ + O(\tau + h^2), \quad (x_k, y_s) \in \Omega^{(h)},\ l = 1, \dots, r,$$

a

$$(2.259)\qquad (L_{h,\tau}^{(\alpha)} u^{(\mathrm{pr})})_{ks}^{(l)} = \Lambda_{h,\tau}(u; u_k, y_s, t_l) + O(h^2), \\ (x_k, y_s) \in \Gamma^{(h)},\ l = 1, \dots, r.$$

Zde jsme položili jako už vícekrát v této i předchozích kapitolách $(u^{(\mathrm{pr})})_{ks}^{(l)} = u(x_k, y_s, t_l)$ pro $(x_k, y_s) \in \bar{\Omega}^{(h)}$ a $l = 0, \dots, r$.

Je-li tedy přesné řešení okrajové úlohy (2.248) až (2.250) dostatečně hladké, platí pro celkovou diskretizační chybu $\eta_{ks}^{(l)} = u_{ks}^{(l)} - u(x_k, y_s, t_l)$ rovnice

$$(2.260)\qquad (L_{h,\tau}\eta)_{ks}^{(l)} = O(\tau + h^2), \quad (x_k, y_s) \in \Omega^{(h)},\ l = 1, \dots, r, \\ \eta_{ks}^{(0)} = 0, \quad (x_k, y_s) \in \bar{\Omega}^{(h)} \\ (l_{h,\tau}\eta)_{ks}^{(l)} = O(h^2), \quad (x_k, y_s) \in \Gamma^{(h)},\ l = 1, \dots, r.$$

Ukažme dále, že operátor $L_{h,\tau}^{(\alpha)}$ splňuje při $\alpha = 0$ a $\alpha = 1$ princip maxima.

Lemma 2.9. *Buď dána libovolná funkce $\eta_{ks}^{(l)}$ definovaná pro $(x_k, y_s) \in \bar{\Omega}^{(h)}$ a pro $l = 0, \dots, r$. Buď dále $\alpha = 0$ nebo $\alpha = 1$ a nechť platí*

$$(2.261)\qquad (L_{h,\tau}^{(\alpha)}\eta)_{ks}^{(l)} \leqq 0, \quad (x_k, y_s) \in \Omega^{(h)},\ l = 1, \dots, r.$$

Nechť konečně v případě, že je $\alpha = 0$ platí

$$(2.262)\qquad \beta \equiv \frac{\tau}{h^2} \leqq \tfrac{1}{4}.$$

Pak platí

$$(2.263)\qquad \eta_{ks}^{(l)} \leqq \max_{(x_k, y_s, t_l) \in \Gamma_{\mathrm{p}}^{(h)}} \eta_{ks}^{(l)},$$

kde $\Gamma_{\mathrm{p}}^{(h)}$ je množina bodů (x_k, y_s, t_l) takových, že je $(x_k, y_s) \in \Gamma^{(h)}$ a $l = 0, \dots, r$ nebo $(x_k, y_s) \in \Omega^{(h)}$ a $l = 0$.

Důkaz je zcela analogický důkazu lemmatu 2.1, a proto jej přenecháme čtenáři.

Lemma 2.10. *Nechť při $\alpha = 0$ nebo $\alpha = 1$ platí*

$$(2.264)\qquad |(L_{h,\tau}^{(\alpha)}\eta)_{ks}^{(l)}| \leqq (L_{h,\tau}^{(\alpha)}\psi)_{ks}^{(l)}, \quad (x_k, y_s) \in \Omega^{(h)},\ l = 1, \dots, r, \\ |\eta_{ks}^{(0)}| \leqq \psi_{ks}^{(0)}, \quad (x_k, y_s) \in \bar{\Omega}^{(h)}, \\ |(l_{h,\tau}\eta)_{ks}^{(l)}| \leqq (l_{h,\tau}\psi)_{ks}^{(l)}, (x_k, y_s) \in \Gamma^{(h)},\ l = 1, \dots, r,$$

a nechť v případě, že je $\alpha = 0$, platí navíc nerovnost (2.262). Pak platí

$$(2.265)\qquad |\eta_{ks}^{(l)}| \leqq \psi_{ks}^{(l)}, \quad (x_k, y_s) \in \bar{\Omega}^{(h)}, \quad l = 0, \dots, r.$$

D ů k a z . Lemma je bezprostředním důsledkem lemmmatu 2.9.

Na základě právě zformulovaných pomocných tvrzení je už snadné dokázat konvergenční větu pro případ $\alpha = 0$ a $\alpha = 1$.

Věta 2.14. *Nechť řešení $u(x, y, t)$ diferenciální rovnice (2.248) s počáteční podmínkou (2.249) a okrajovými podmínkami (2.250) je dostatečně hladké. Dále nechť existuje dostatečně hladké řešení $w(x, y, t)$ diferenciální rovnice*

$$(2.266)\qquad (Lw)(x, y, t) = 1, \quad (x, y) \in \Omega,\ t \in (0, T),$$

s okrajovou podmínkou

$$(2.267)\qquad w(x, y, t) = 1, \quad (x, y) \in \Gamma,\ t \in (0, T),$$

a s nezápornou počáteční podmínkou. Buď konečně $u_{ks}^{(l)}$ řešení soustavy (2.251) s počáteční podmínkou (2.252) a okrajovými podmínkami (2.253) při $\alpha = 0$ nebo $\alpha = 1$ a nechť platí (2.262), je-li $\alpha = 0$. Pak existuje konstanta M taková, že je

$$(2.268)\qquad |u_{ks}^{(l)} - u(x_k, y_s, t_l)| \leqq M(\tau + h^2)$$

pro $(x_k, y_s) \in \bar{\Omega}^{(h)}$ a $l = 0, \dots, r$.

D ů k a z . Položme $\eta_{ks}^{(l)} = u_{ks}^{(l)} - u(x_k, y_s, t_l)$. Na základě úvah úplně analogických jako v důkazu věty 3.7 z kap. II nebo věty 2.4 z kap. III (viz str. 179 nebo 254) snadno zjistíme, že lze nalézt konstantu M tak, aby pro funkci $\psi_{ks}^{(l)} = Mw(x_k, y_s, t_l)(\tau + h^2)$ platily nerovnosti (2.264). Tvrzení věty pak plyne ihned z lemmatu 2.10 a omezenosti funkce w.

Analogická tvrzení se dokáží i pro ostatní okrajové podmínky, pokud jsou formulovány tak, aby platilo lemma 2.10.

Obraťme se konečně k vyšetření konvergence metod daných operátorem $L_{h,\tau}^{(\alpha)}$ při obecném α. Zde je situace komplikovanější než u explicitní nebo čistě implicitní

metody, neboť při $1/2 \leqq \alpha < 1$ je princip maxima splněn jen za dodatečných omezení na poměr časového a prostorového integračních kroku, což je v jistém smyslu nepřirozené. Ani postup užitý v odst. 2.2.3 sem nelze bezprostředně přenést, neboť ve dvou (a více) dimenzích neplatí nerovnost (2.190).

Pokud se spokojíme při vyšetřování operátorů $L_{h,\tau}^{(\alpha)}$ s normou analogickou normě dané vzorcem (2.77) a omezíme se na případ, že okrajové podmínky jsou splněny přesně, lze se opřít o lemma 2.3 a postupovat analogicky jako v odst. 2.1.2. K důkazu konvergence v tomto případě bude tedy stačit zjištění, že užité schéma je stejnoměrně stabilní vzhledem k počátečními podmínkám.

Zformulujeme a dokažme příslušné tvrzení pro případ okrajové úlohy (2.248) až (2.250), kde oblast Ω je čtverec $(0,1) \times (0,1)$. Položíme-li v tomto případě $h = 1/n$, kde n je přirozené číslo, tvoří funkce definované na množině $\Omega^{(h)}$ $(n-1)^2$-dimenzionální vektorový prostor. Zaveďme v tomto prostoru normu analogickou normě (2.77), tj. za normu vektoru v o složkách v_{ks}, $(x_k, y_s) \in \Omega^{(h)}$ berme číslo

$$\|v\|_{h^2} = \Big[h^2 \sum_{(x_k,y_s)\in\Omega^{(h)}} v_{ks}^2\Big]^{1/2} \tag{2.269}$$

a tak vzniklý vektorový prostor označme $\mathbf{E}_{h^2}^{(n-1)^2}$.

Věta 2.15. *Buď $\Omega = (0,1) \times (0,1)$, buď $h = 1/n$ a nechť v případě, že je $0 \leqq \alpha < 1/2$, platí nerovnost*

$$\beta \equiv \frac{\tau}{h^2} \leqq \frac{1}{4(1-2\alpha)}. \tag{2.270}$$

Pak diferenční schéma dané operátorem $L_{h,\tau}^{(\alpha)}$ je v normě prostoru $\mathbf{E}_{h^2}^{(n-1)^2}$ stejnoměrně stabilní vzhledem k počátečním podmínkám.

D ů k a z . Ve shodě s definicí 2.2 máme dokázat, že existuje konstanta M taková, že pro každé řešení soustavy

$$(L_{h,\tau}^{(\alpha)}\eta)_{ks}^{(l)} = 0, \quad (x_k, y_s) \in \Omega^{(h)}, \ l = m+1, \dots, r, \tag{2.271}$$

s okrajovými podmínkami

$$\eta_{ks}^{(l)} = 0, \quad (x_k, y_s) \in \Gamma^{(h)}, \ l = m, \dots, r, \tag{2.272}$$

platí

$$\|\eta^{(l)}\|_{h^2} \leqq M \|\eta^{(m)}\|_{h^2}, \tag{2.273}$$

pro každé $m = 0, \dots, r-1$ a pro $l = m, \dots, r$, kde $\eta^{(l)}$ je vektor o složkách $\eta_{ks}^{(l)}$, $(x_k, y_s) \in \Omega^{(h)}$. Abychom toho dosáhli, užijeme stejně jako v odst. 2.1.2 metodu separace proměnných, která nám dovolí psát řešení soustavy (2.271), (2.272) v uzavřeném tvaru.

Buď tedy m libovolné pevně zvolené celé číslo, pro něž je $0 \leqq m < r$, a buď

$\eta^{(m)}$ libovolný vektor z prostoru $\mathbf{E}_{h^2}^{(n-1)^2}$. Protože soustava vektorů $v^{(\nu,\mu)}$, $\nu, \mu = 1, \dots, n-1$, o složkách $v_{ks}^{(\nu,\mu)}$ daných vzorci

$$v_{ks}^{(\nu,\mu)} = \sin\frac{\nu\pi k}{n} \sin\frac{\mu\pi s}{n}, \quad k, s = 1, \dots, n-1, \tag{2.274}$$

tvoří zřejmě ortogonální bázi v prostoru $\mathbf{E}_{h^2}^{(n-1)^2}$ (srv. také vzorec (2.91) z odst. 2.2 kap. III), dá se vektor $\eta^{(m)}$ psát ve tvaru

$$\eta^{(m)} = \sum_{(x_\nu,y_\mu)\in\Omega^{(h)}} \alpha_{\nu\mu}^{(m)} v^{(\nu,\mu)}, \tag{2.275}$$

kde $\alpha_{\nu\mu}^{(m)}$ jsou vhodné konstanty, a přitom platí

$$\|\eta^{(m)}\|_{h^2} = \sum_{(x_\nu,y_\mu)\in\Omega^{(h)}} [\alpha_{\nu\mu}^{(m)}]^2 \|v^{(\nu,\mu)}\|_{h^2}^2. \tag{2.276}$$

Hledejme řešení soustavy (2.271), (2.272) ve tvaru

$$\eta_{ks}^{(l)} = \sum_{(x_\nu,y_\mu)\in\Omega^{(h)}} c(l)\alpha_{\nu\mu}^{(m)} v_{ks}^{(\nu,\mu)}. \tag{2.277}$$

Dosazením do rovnic (2.271), (2.272) snadno zjistíme, že k jejich splnění je třeba, aby pro funkci $c(l)$ platilo

$$\begin{aligned} &c(m) = 1, \\ &(1+\alpha\beta\lambda_{\nu\mu})c(l) = \big(1-(1-\alpha)\beta\lambda_{\nu\mu}\big)c(l-1), \\ &l = m+1, \dots, r, \end{aligned} \tag{2.278}$$

kde

$$\lambda_{\nu\mu} = 4\Big(\sin^2\frac{\nu\pi}{2n} + \sin^2\frac{\mu\pi}{2n}\Big), \quad \nu, \mu = 1, \dots, n-1. \tag{2.279}$$

Je tedy

$$c(l) = \Big[\frac{1-(1-\alpha)\beta\lambda_{\nu\mu}}{1+\alpha\beta\lambda_{\nu\mu}}\Big]^{l-m} \tag{2.280}$$

a z rovnic (2.277) plyne, že platí

$$\|\eta^{(l)}\|_{h^2}^2 = \sum_{(x_\nu,y_\mu)\in\Omega^{(h)}} \Big|\frac{1-(1-\alpha)\beta\lambda_{\nu\mu}}{1+\alpha\beta\lambda_{\nu\mu}}\Big|^{2l-2m} [\alpha_{\nu\mu}^{(m)}]^2 \|v^{(\nu,\mu)}\|_{h^2}^2. \tag{2.281}$$

Vyšetřované schéma bude tedy stejnoměrně stabilní vzhledem k počátečním podmínkám právě tehdy, budou-li platit nerovnosti

$$-1 \leqq \frac{1-(1-\alpha)\beta\lambda_{\nu\mu}}{1+\alpha\beta\lambda_{\nu\mu}} \leqq 1. \tag{2.282}$$

Pravá z těchto nerovností je splněna vždy, neboť podle vzorce (2.279) je $\lambda_{\nu\mu} \geqq 0$. Je-li $1/2 \leqq \alpha \leqq 1$, platí i levá nerovnost (2.282) bez jakýchkoliv doplňujících

podmínek, neboť v tomto případě je $(1-2\alpha)\beta\lambda_{\nu\mu} \leqq 0$, a tedy tím spíše platí nerovnost

$$(1-2\alpha)\beta\lambda_{\nu\mu} < 2, \tag{2.283}$$

neboli

$$-1-\alpha\beta\lambda_{\nu\mu} < 1-(1-\alpha)\beta\lambda_{\nu\mu}. \tag{2.284}$$

Je-li naopak $0 \leqq \alpha < 1/2$ a platí-li navíc nerovnost (2.270), platí opět (2.283), neboť je $\lambda_{\nu\mu} < 8$. Tím je věta dokázána.

Schéma dané operátorem $L_{h,\tau}^{(\alpha)}$ při obecném α vede tedy v normě (2.269) k rychlosti konvergence $O(\tau + h^2)$. Protože ve vzorci (2.258) lze zřejmě při $\alpha = 1/2$ zaměnit sčítanec $O(\tau + h^2)$ sčítancem $O(\tau^2 + h^2)$, má metoda daná operátorem $L_{h,\tau}^{(\alpha)}$, zvaná i zde Crankova-Nicolsonova metoda, celkovou diskretizační chybu řádu $O(\tau^2 + h^2)$. V prostorově jednodimenzionální případě vedlo Crankovo-Nicolsonovo schéma k algoritmu s minimálním počtem operací mezi všemi metodami vycházejícími z operátorů $L_{h,\tau}^{(\alpha)}$. Zde tomu už tak není. Při užití Crankova-Nicolsonova schématu musíme totiž stejně jako u čistě implicitní metody řešit v každém časovém řádku soustavu $O(1/h^2)$ rovnic o $O(1/h^2)$ neznámých. Matice této soustavy je sice pásová (srv. odst. 2.1 z kap. III), šíře pásu je však $O(1/h)$, takže k řešení zmíněné soustavy např. Gaussovou eliminační metodou je třeba řádově $O(1/h^4)$ operací. Položíme-li $\tau = O(h)$, je celkový počet operací $O(1/h^5)$ a toto číslo je dokonce větší než u explicitní metody. To je také důvod, proč se zavádějí tzv. metody střídavých směrů a lokálně jednorozměrné metody. Tyto metody jsou co do velikosti celkové chyby v podstatě ekvivalentní Crankově-Nicolsonově metodě, vedou však k podstatně ekonomičtějším algoritmům. Některé z nich si nyní stručně popíšeme.

2.3.2 Metody střídavých směrů

Hlavním důvodem zavedení těchto metod je, jak už jsme uvedli, snaha po dosažení maximální výpočetní ekonomie. Základní myšlenka, která se sledovala při jejich vzniku, byla spojit výhody implicitních metod spočívající v nezávislosti jejich stability na poměru časového a prostorového integračního kroku, s tou skutečností, že soustavy lineárních algebraických rovnic s třídiagonální maticí lze velmi ekonomicky řešit Gaussovou eliminační metodou (neboť v této situaci je počet potřebných operací úměrný počtu rovnic). Princip těchto metod ukážeme na příkladě okrajové úlohy (2.248) až (2.250), kde Ω je čtverec $(0,1)\times(0,1)$. Kromě přibližného řešení $u_{ks}^{(l+1)}$ v čase $t = t_{l+1}$ se v těchto metodách počítá ještě pomocná hodnota $u_{ks}^{(l+1/2)}$ v mezivrstvě $t = t_{l+1/2} = t_l + \tau/2$, a to tak, že k výpočtu hodnot $u_{ks}^{(l+1/2)}$ se za předpokladu, že hodnoty $u_{ks}^{(l)}$ jsou už známé, použijí jiné vzorce než k výpočtu hodnot $u_{ks}^{(l+1)}$. Jedna z historicky prvních metod tohoto typu pochází od *Peacemana*

a *Rachforda* a je dána dvojicí vzorců

$$\begin{aligned}
\frac{u_{ks}^{(l+1/2)} - u_{ks}^{(l)}}{\tau} &= \frac{1}{2h^2}[u_{k-1,s}^{(l+1/2)} - 2u_{ks}^{(l+1/2)} + u_{k+1,s}^{(l+1/2)}] + \\
&\quad + \frac{1}{2h^2}[u_{k,s-1}^{(l)} - 2u_{ks}^{(l)} + u_{k,s+1}^{(l)}], \\
\frac{u_{ks}^{(l+1)} - u_{ks}^{(l+1/2)}}{\tau} &= \frac{1}{2h^2}[u_{k-1,s}^{(l+1/2)} - 2u_{ks}^{(l+1/2)} + u_{k+1,s}^{(l+1/2)}] + \\
&\quad + \frac{1}{2h^2}[u_{k,s-1}^{(l+1)} - 2u_{ks}^{(l+1)} + u_{k,s+1}^{(l+1)}], \\
&\quad k,s = 1,\dots,n-1,\ l = 0,\dots,r-1.
\end{aligned} \tag{2.285}$$

První z rovnic (2.285) tedy vznikla tak, že v intervalu $\langle t_l, t_{l+1/2}\rangle$ jsme aproximovali derivaci $\partial^2 u/\partial x^2$ příslušným diferenčním podílem v horní (tj. počítané) časové vrstvě a derivaci $\partial^2 u/\partial y^2$ v dolní (tj. známé) vrstvě. V druhém vzorci (2.285) jsme tento postup prohodili.

Přepíšeme-li rovnice (2.285) ve tvaru

$$\begin{aligned}
&-\tfrac{1}{2}\beta u_{k-1,s}^{(l+1/2)} + (1+\beta)u_{ks}^{(l+1/2)} - \tfrac{1}{2}\beta u_{k+1,s}^{(l+1/2)} = \\
&= \tfrac{1}{2}\beta u_{k-1,s}^{(l)} + (1-\beta)u_{ks}^{(l)} + \tfrac{1}{2}\beta u_{k+1,s}^{(l)}, \\
&-\tfrac{1}{2}\beta u_{k-1,s}^{(l+1)} + (1+\beta)u_{ks}^{(l+1)} - \tfrac{1}{2}\beta u_{k+1,s}^{(l+1)} = \\
&= \tfrac{1}{2}\beta u_{k-1,s}^{(l+1/2)} + (1-\beta)u_{ks}^{(l+1/2)} + \tfrac{1}{2}\beta u_{k+1,s}^{(l+1/2)}, \\
&k,s = 1,\dots,n-1,\ l = 0,\dots,r-1
\end{aligned} \tag{2.286}$$

(položili jsme opět jako obvykle $\beta = \tau/h^2$), je ihned vidět, že hodnoty přibližného řešení $u_{1s}^{(l+1/2)}, u_{2s}^{(l+1/2)}, \dots, u_{n-1,s}^{(l+1/2)}$ v čase $t = t_{l+1/2}$ dostaneme při pevném s řešením soustavy $(n-1)$ lineárních rovnic $(n-1)$ neznámých s třídiagonální regulární maticí a hodnoty přibližného řešení $u_{k1}^{(l+1)}, u_{k2}^{(l+1)}, \dots, u_{k,n-1}^{(l+1)}$ v čase $t = t_{l+1}$ při pevném k rovněž řešením $(n-1)$ lineárních rovnic o $(n-1)$ neznámých s třídiagonální maticí (v našem speciálním případě budou dokonce všechny zmíněné třídiagonální matice stejné). Příslušný algoritmus tedy probíhá tak, že hodnoty přibližného řešení v mezivrstvě $t = t_{l+1/2}$ počítáme postupně na přímkách $y = y_s$ rovnoběžných s osou x a hodnoty přibližného řešení ve vrstvě $t = t_{l+1}$ postupně na přímkách $x = x_k$ rovnoběžných s osou y. Z tohoto charakteru algoritmu také pochází název metoda střídavých směrů. K tomu, abychom vypočítali přibližného řešení $(l+1)$-ní časové vrstvě za předpokladu, že přibližné řešení v l-té časové vrstvě je už vypočteno, je tedy třeba řešit dvakrát $(n-1)$ soustav lineárních rovnic o $(n-1)$ neznámých s třídiagonálními maticemi. Celkový počet operací potřebných k získání přibližného řešení v jedné časové vrstvě je tedy řádově roven číslu $O(1/h^2)$.

Při dalším vyšetřování Peacemanovy-Rachfordovy metody budeme už velice struční. Její lokální chyba (tj. chyba, se kterou se vypočte přibližné řešení v $(l+1)$-ní časové vrstvě za předpokladu, že řešení v l-té vrstvě je přesné) je řádu

$O(\tau^2 + h^2)$, jak se snadno zjistí standardním užitím Taylorova vzorce. Že i celková diskretizační chyba je téhož řádu (v normě prostoru $\mathsf{E}_{h^2}^{(n-1)^2}$) plyne pak už v podstatě z lemmatu 2.3 a z následující věty.

Věta 2.16. *Peacemanova-Rachfordova metoda střídavých směrů je v normě* (2.269) *stejnoměrně stabilní vzhledem k počátečním podmínkám.*

D ů k a z . Podobně jako v důkazu věty 2.15 máme dokázat, že pro každé řešení soustavy (2.285) s nulovými okrajovými podmínkami platí odhad

$$\|u^{(l)}\|_{h^2} \leqq M\|u^{(m)}\|_{h^2} \tag{2.287}$$

pro $m = 0, \ldots, r-1$ a pro $l = m, \ldots, r$. Funkci $u_{ks}^{(l)}$ hledejme opět ve tvaru

$$u_{ks}^{(l)} = \sum_{(x_\nu, y_\mu)\in\Omega^{(h)}} c(l)\alpha_{\nu\mu}^{(m)} v_{ks}^{(\nu,\mu)}, \tag{2.288}$$

kde funkce $v^{(\nu,\mu)}$ jsou dány vzorcem (2.274) a čísla $\alpha_{\nu\mu}^{(m)}$ jsou Fourierovy koeficienty počátečního vektoru, tj. platí

$$u_{ks}^{(l)} = \sum_{(x_\nu, y_\mu)\in\Omega^{(h)}} \alpha_{\nu\mu}^{(m)} v_{ks}^{(\nu,\mu)}, \tag{2.289}$$

Dosadíme-li (2.288) do (2.285), snadno zjistíme, že musí platit

$$\begin{aligned} \frac{c(l+1/2)-c(l)}{\tau} &= -\frac{1}{2h^2}\varrho_\nu c(l+1/2) - \frac{1}{2h^2}\varrho_\mu c(l), \\ \frac{c(l+1)-c(l+1/2)}{\tau} &= -\frac{1}{2h^2}\varrho_\mu c(l+1) - \frac{1}{2h^2}\varrho_\nu c(l+1/2), \end{aligned} \tag{2.290}$$

kde

$$\varrho_\nu = 4\sin^2\frac{\nu\pi}{2n}, \quad \nu = 1, \ldots, n-1. \tag{2.291}$$

Položíme-li znovu $\beta = \tau/h^2$ a vyloučíme-li $c(l+1/2)$ z rovnic (2.290), dostaneme

$$c(l+1) = \frac{(1-\frac{1}{2}\beta\varrho_\nu)(1-\frac{1}{2}\beta\varrho_\mu)}{(1+\frac{1}{2}\beta\varrho_\nu)(1+\frac{1}{2}\beta\varrho_\mu)}c(l). \tag{2.292}$$

Zopakujeme-li nyní doslova úvahy z důkazu věty 2.15, dostaneme, že podmínka

$$\left|\frac{(1-\frac{1}{2}\beta\varrho_\nu)(1-\frac{1}{2}\beta\varrho_\mu)}{(1+\frac{1}{2}\beta\varrho_\nu)(1+\frac{1}{2}\beta\varrho_\mu)}\right| \leqq 1 \tag{2.293}$$

je nutná a postačující podmínka požadované stability. Tato podmínka je však vzhledem k (2.291) splněna pro libovolné $\beta > 0$. Věta je dokázána.

Peacemanova-Rachfordova metoda je tedy absolutně stabilní a je při ní rozumné klást $\tau = O(h)$. Výpočet pak vyžaduje $O(1/h^3)$ operací, což je výsledek podstatně příznivější než u kterékoliv z metod popsaných v odst. 2.3.1.

Metod střídavých směrů byla navržena celá řada. Uveďme z nich ještě velmi známou metodu *D'jakonovovu*, která je dána dvojicí vzorců

$$\begin{aligned} (I - \tfrac{1}{2}\beta\delta_x^2)u^{(l+1/2)} &= (I + \tfrac{1}{2}\beta\delta_x^2)(I + \tfrac{1}{2}\beta\delta_y^2)u^{(l)}, \\ (I - \tfrac{1}{2}\beta\delta^2 y)u^{(l+1)} &= u^{(l+1/2)}. \end{aligned} \tag{2.294}$$

Zde jsme použili stručný zápis diferenčních schémat pomocí identického operátoru I a operátoru *centrální diference*. Centrální diferenci přitom definujeme, jak je obvyklé, jako operátor, který funkci f přiřazuje funkci δf předpisem $(\delta f)(x) = f(x + h/2) - f(x - h/2)$. Mocninou operátoru δ pak rozumíme operátor δ^k definovaný rekurentně vztahem $(\delta^k f)(x) = (\delta^{k-1} f)(x + h/2) - (\delta^{k-1} f)(x - h/2)$. Pokud užijeme tuto operaci na funkci více proměnných, pak připojený index označuje proměnnou, které se operace týká. I pro D'Jakonovovu metodu platí věta 2.16, jak se snadno zjistí, a její lokální chyba je opět $O(\tau^2 + h^2)$.

Poznamenejme, že zapíšeme-li Peacemanovu-Rachfordovu metodu analogickým způsobem jako metodu D'jakonovovu, dostaneme dvojici vzorců

$$\begin{aligned} (I - \tfrac{1}{2}\beta\delta_x^2)u^{(l+1/2)} &= (I + \tfrac{1}{2}\beta\delta_y^2)u^{(l)}, \\ (I - \tfrac{1}{2}\beta\delta_y^2)u^{(l+1)} &= (I + \tfrac{1}{2}\beta\delta_x^2)u^{(l+1/2)}. \end{aligned} \tag{2.295}$$

V souvislosti s metodami střídavých směrů je ještě třeba upozornit na jednu velmi závažnou skutečnost. Údajům o velikosti chyb, které jsme uvedli, je nutno rozumět tak, že se týkají přibližného řešení $u^{(l)}$ pro *celá* l a že tato přibližná řešení jsou dána jak v případě Peacemanovy-Rachfordovy metody, tak v případě D'jakonovovy metody vzorcem

$$(I - \tfrac{1}{2}\beta\delta_x^2)(I - \tfrac{1}{2}\beta\delta_y^2)u^{(l+1)} = (I + \tfrac{1}{2}\beta\delta_x^2)(I + \tfrac{1}{2}\beta\delta_y^2)u^{(l)} \tag{2.296}$$

s počátečními podmínkami $u_{ks}^{(0)} = g(x_k, y_s)$ a okrajovými podmínkami $u_{ks}^{(l)} = \gamma(x_k, y_s, t_l)$, $(x_k, y_s) \in \Gamma^{(h)}$, $l = 0, \ldots, r$. Veličiny $u^{(l+1/2)}$ s lomenými indexy přitom pokládáme za pomocné. Tyto pomocné veličiny *nemusí* aproximovat řešení v žádné časové vrstvě a zavádějí se jen proto, abychom řešení soustav rovnic, které je nutno při užití schématu (2.296) řešit, získali ekonomicky. Veličiny $u^{(l+1/2)}$ jsou tedy vlastně definovány vzorci

$$u^{(l+1/2)} = \tfrac{1}{2}(I - \tfrac{1}{2}\beta\delta_y^2)u^{(l+1)} + \tfrac{1}{2}(I + \tfrac{1}{2}\beta\delta_y^2)u^{(l)} \tag{2.297}$$

(Peacemanova-Rachfordova metoda) a

$$u^{(l+1/2)} = (I - \tfrac{1}{2}\beta\delta_y^2)u^{(l+1)} \tag{2.298}$$

(D'jakonovova metoda). Odtud však plyne, že i okrajové podmínky, které pro tyto veličiny potřebujeme při počítání podle vzorců (2.294), (2.295), je třeba volit

v souladu s rovnicemi (2.297) a (2.298), tj. je třeba je počítat ze vzorců

$$u_{ks}^{(l+1/2)} = \tfrac{1}{2}(I - \tfrac{1}{2}\beta\delta_y^2)\gamma(x_k, y_s, t_{l+1}) + \\ + \tfrac{1}{2}(I + \tfrac{1}{2}\beta\delta_y^2)\gamma(x_k, y_s, t_l), \quad (x_k, y_s) \in \Gamma^{(h)} \tag{2.299}$$

(Peacemanova-Rachfordova metoda) a

$$u_{ks}^{(l+1/2)} = (I - \tfrac{1}{2}\beta\delta_y^2)\gamma(x_k, y_s, t_{l+1}) \tag{2.300}$$

(D'jakonovova metoda). Kdybychom kladli zdánlivě přirozeně $u_{ks}^{(l+1/2)} = \gamma(x_k, y_s, t_{l+1/2})$, narušili bychom odhady chyb, které jsme uvedli.

Závěrem tohoto odstavce upozorněme ještě na jednu zajímavou a důležitou souvislost. V odst. 2.3.2 z kap. III jsme zavedli Peacemanovu-Rachfordovu metodu střídavých směrů jako iterační metodu pro řešení soustav lineárních rovnic, které vznikají při řešení okrajových úloh pro eliptické parciální diferenciální rovnice metodou sítí. Tato metoda vznikla ve skutečnosti tak, že řešení eliptické rovnice se pokládá za ustálené řešení parabolické rovnice a tato parabolická rovnice se pak řeší metodou (2.285).

2.3.3 Lokálně jednorozměrné metody

Tyto metody jsou velmi podobné metodám střídavých směrů. I zde se přibližné řešení $u^{(l+1)}$ v čase $t = t_{l+1}$ za předpokladu, že přibližné řešení $u^{(l)}$ v čase $t = t_l$ je už vypočteno, počítá tak, že se nejprve vypočte pomocná hodnota $u^{(l+1/2)}$ pomocí jednoho vzorce a pak teprve konečná hodnota pomocí jiného vzorce. Nejběžnější metoda tohoto typu pochází od Janenka a je dána dvojicí vzorců

$$(I - \tfrac{1}{2}\beta\delta_x^2)u^{(l+1/2)} = (I + \tfrac{1}{2}\beta\delta_x^2)u^{(l)}, \\ (I - \tfrac{1}{2}\beta\delta_y^2)u^{(l+1)} = (I + \tfrac{1}{2}\beta\delta_y^2)u^{(l+1/2)}. \tag{2.301}$$

Přibližné řešení se tedy počítá tak, jako kdyby v časovém úseku mezi časy $t = t_l$ a $t = t_{l+1/2}$ vedl materiál teplo jen ve směru osy x a v časovém úseku $\langle t_{l+1/2}, t_{l+1}\rangle$ pouze ve směru osy y. Odtud také pochází název metody. Je zřejmé, že jeden dvoukrok uvedené metody se realizuje stejně jako u metod střídavých směrů řešením soustav s třídiagonálními maticemi a že je řádově stejně pracný. Rovněž je zřejmé, jak lze tuto metodu přenést na případ více prostorových proměnných. Je k tomu pouze třeba rozdělit časový úsek $\langle t_l, t_{l+1}\rangle$ na m mezivrstev, kde m je počet prostorových proměnných, a v každé mezivrstvě aproximovat pouze derivaci podle jedné prostorové proměnné.

Metoda (2.301) je v normě (2.269) opět stejnoměrně stabilní vzhledem k počátečním podmínkám bez jakéhokoliv dalšího omezení na parametry τ a h a v případě homogenních okrajových podmínek vede k rychlosti konvergence $O(\tau^2 + h^2)$. V případě nehomogenních Dirichletových okrajových podmínek se obvykle doporučuje brát za okrajové hodnoty pomocné veličiny $u^{(l+1/2)}$ hodnoty okrajové funkce, tj.

klást $u_{ks}^{(l+1/2)} = \gamma(x_k, y_s, t_{l+1/2})$. Tato volba však vede podobně jako u metod střídavých směrů ke ztrátě přesnosti. Na rozdíl od tamní situace je však zde obtížné udat korekční vzorce typu (2.299) a (2.300), které by tuto obtíž odstraňovaly. To je také důvod, proč jsou lokálně jednorozměrné metody podstatně méně populární než metody střídavých směrů.

3 Semidiskrétní metody

Již vícekrát jsme v této kapitole upozornili na to, že základní úloha pro parciální diferenciální rovnice parabolického typu má vzhledem k proměnné t charakter úlohy s počátečními podmínkami pro obyčejnou diferenciální rovnici a vzhledem k prostorovým proměnným charakter okrajové úlohy pro obyčejnou diferenciální rovnici, resp. pro parciální diferenciální rovnici eliptického typu. Metody, které v tomto článku popíšeme, si kladou za cíl této skutečnosti využít, a to tak, že diskretizace se provádí pouze vzhledem k prostorovým proměnným nebo pouze vzhledem k proměnné t.

Nejdůležitějšími reprezentanty metod první skupiny jsou metoda přímek a semidiskrétní metody Galerkinova typu. V metodě přímek se diskretizace prostorových proměnných provádí na základě metody sítí, u semidiskrétních metod Galerkinova typu se vychází z Ritzovy-Galerkinovy metody pro řešení eliptických úloh a postup se většinou spojuje s metodou končených prvků.

Metody druhé skupiny, kdy se diskretizace provádí pouze vzhledem k proměnné t, se nazývají Rotheovy metody.

Je třeba upozornit, že metody obou zmíněných skupin jsou odlišného charakteru než metoda sítí, jejíž různé varianty jsme popsali v předešlém článku. Zatímco výsledkem algoritmů založených na metodě sítí jsou přibližné hodnoty hledaného řešení v nějaké konečné množině bodů zadané oblasti, tedy veličiny, které nás už bezprostředně zajímají, za aproximaci přesného řešení v případě semidiskrétních metod slouží stále ještě nekonečněrozměrné problémy, i když jednodušší než původní. Konkrétně jde v případě metod první skupiny o soustavy obyčejných diferenciálních rovnic s počátečními podmínkami, v případě Rotheovy metody o okrajové úlohy pro obyčejné nebo parciální eliptické diferenciální rovnice. K řešení těchto úloh se pak dají užít metody, které byly popsány v kap. I, II a III.

Abychom se mohli v dalším výkladu soustředit na základní myšlenku vyšetřovaných metod a nezastřeli ji množstvím technických podrobností, omezíme se v tomto článku prakticky výhradně na prostorově jednodimenzionální rovnici pro vedení tepla (2.1) s počáteční podmínkou (2.2) a s homogenními Dirichletovými okrajovými podmínkami (2.3) a možnosti rozšíření na obecnější úlohy jen stručně okomentujeme.

3.1 Metoda přímek

Tato metoda vychází, jak už jsme uvedli, z myšlenky metody sítí. Na rozdíl od ní se však síť sestrojí jen v oblasti prostorových proměnných a diferenčními podíly se nahradí pouze derivace podle těchto proměnných. Přesné řešení daného problému je pak aproximováno řešením soustavy obyčejných diferenciálních rovnic s počátečními podmínkami. V odst. 3.1.1 popíšeme klasickou variantu této metody a v odst. 3.1.2 si stručně všimneme Numerovovy metody, v níž se uvedená základní myšlenka realizuje dosti důmyslně, a která má proto jistou popularitu.

3.1.1 Klasická metoda přímek

Buď u řešením problému (2.1) až (2.3), buď n přirozené číslo a rozdělme interval $\langle 0, 1\rangle$ na n dílků délky $h = 1/n$ dělicími body $x_k = kh$, $k = 0, \dots, n$. Nahradíme-li druhou derivaci $\partial^2 u/\partial x^2$ v bodě (x_k, t) podílem $[u(x_{k-1}, t) - 2u(x_k, t) + u(x_{k+1}, t)]/h^2$ a chybu, které se přitom dopustíme, zanedbáme, dostaneme pro přibližné řešení $u_k(t)$ na přímce $x = x_k$, $0 \leqq t \leqq T$, soustavu obyčejných diferenciálních rovnic

$$\dot{u}_k(t) = \frac{1}{h^2}[u_{k-1}(t) - 2u_k(t) + u_{k+1}(t)], \quad k = 1, \dots, n-1, \ t \in (0, T), \tag{3.1}$$

kde

$$u_0(t) = u_n(t) = 0, \tag{3.2}$$

$$u_k(0) = g(x_k), \quad k = 1, \dots, n-1, \tag{3.3}$$

a tečku užíváme jako stručné označení derivace podle proměnné t.

Je zřejmé, že stejně jednoduše se dá postupovat i v případě složitějších parabolických rovnic.

Chceme-li provést úplnou diskretizaci daného problému, je třeba řešit soustavu (3.1) vhodnou numerickou metodou. Užijeme-li např. nejprostší Eulerovu metodu, dostaneme explicitní metodu sítí. Podobně řešení této soustavy lichoběžníkovým pravidlem vede ke Crankově-Nicolsonově schématu. Souvislost mezi klasickou metodou přímek a metodou sítí je tedy skutečně velice úzká.

V této souvislosti je také třeba upozornit na jednu důležitou a pro metodu přímek a pro semidiskrétní metody, v nichž se ponechává čas spojitý, vůbec typickou okolnost. Matice soustavy diferenciálních rovnic (3.1) je matice $-(1/h^2)P_0$, kde P_0 je matice (2.81). Její vlastní čísla μ_ν jsou tedy dána vzorcem

$$\mu_\nu = -\frac{4}{h^2}\sin^2\frac{\nu\pi}{2n}, \quad \nu = 1, \dots, n-1 \tag{3.4}$$

(srv. vzorec (2.82)). Z tohoto vzorce však plyne, že absolutní hodnota vlastního čísla μ_1 konverguje pro $h \to 0$ k $-\pi^2$, zatímco absolutní hodnota vlastního čísla μ_{n-1} roste při $h \to 0$ nade všechny meze. Soustava (3.1) tvoří tedy soustavu diferenciálních rovnic se silným tlumením (srv. odst. 6.3 z kap. I) a tento její charakter

je tím výraznější, čím je integrační krok h menší. Prakticky to má za následek, že řešení této soustavy obecnou metodou pro řešení úloh s počátečními podmínkami pro obyčejné diferenciální rovnice vede obecně pouze k relativně stabilním diferenčním schématům a že absolutní stabilitu lze očekávat pouze při užití A-stabilních nebo aspoň $A(0)$-stabilních metod.

Konvergenční vlastnosti klasické metody přímek jsou uvedeny v následující větě.

Věta 3.1. *Nechť řešení problému* (2.1) až (2.3) *existuje — označme je* u — *a nechť má v množině* $R = \{(x, t);\ 0 \leqq x \leqq 1,\ 0 \leqq t \leqq T\}$ *spojitou čtvrtou derivaci podle* x. *Nechť dále* u_k *je přibližné řešení vypočtené z rovnic* (3.1) až (3.3). *Pak existují konstanty* M *a* $h_0 > 0$ *takové, že pro každé* $h \leqq h_0$ *a* $t \in \langle 0, T\rangle$ *platí*

$$\|u^{(h)}(t) - u(t)\|_h \leqq Mh^2, \tag{3.5}$$

kde

$$u^{(h)}(t) = [u_1(t), \dots, u_{n-1}(t)]^T \tag{3.6}$$

a

$$u(t) = [u(x_1, t), \dots, u(x_{n-1}, t)]^T. \tag{3.7}$$

D ů k a z . Připomeňme především, že norma užitá ve vzorci (3.5) je norma v prostoru $\mathbf{E}_h^{(n-1)}$, tj. je to obyčejná euklidovská norma vynásobená činitelem $h^{1/2}$ (viz vzorec (2.77)). Užijeme-li označení (3.6), dá se soustava (3.1) psát maticově ve tvaru

$$\dot{u}^{(h)}(t) = -\frac{1}{h^2}P_0 u^{(h)}(t), \tag{3.8}$$

kde P_0 je už výše zmíněná matice (2.81). Položme dále

$$e(t) = u^{(h)}(t) - u(t) \tag{3.9}$$

a

$$r(t) = \left[\frac{\partial u}{\partial t}(x_1, t), \dots, \frac{\partial u}{\partial t}(x_{n-1}, t)\right]^T, \tag{3.10}$$

takže je

$$\dot{e}(t) = -\frac{1}{h^2}P_0 u^{(h)}(t) - r(t). \tag{3.11}$$

Z předpokladů věty plyne, že platí

$$\frac{\partial u}{\partial t}(x_k, t) = \frac{\partial^2 u}{\partial x^2}(x_k, t) = \frac{1}{h^2}[u(x_{k-1}, t) - 2u(x_k, t) + u(x_{k+1}, t)] + \varepsilon_k(t), \tag{3.12}$$

přičemž je pro $h < h_0$, $t \in \langle 0, T\rangle$ a $k = 1, \dots, n-1$

$$|\varepsilon_k(t)| \leqq Mh^2 \tag{3.13}$$

a M a $h_0 > 0$ jsou vhodné konstanty. Položíme-li

$$\varepsilon(t) = [\varepsilon_1(t), \dots, \varepsilon_{n-1}(t)]^T, \tag{3.14}$$

plyne ze vzorců (3.10) a (3.12), že pro vektor r platí rovnice

$$r(t) = -\frac{1}{h^2} P_0 u + \varepsilon(t). \tag{3.15}$$

Celkem tedy dostáváme, že vektor $e(t)$ splňuje soustavu diferenciálních rovnic

$$\dot{e}(t) = -\frac{1}{h^2} P_0 e(t) - \varepsilon(t) \tag{3.16}$$

s počáteční podmínkou

$$e(0) = 0. \tag{3.17}$$

Je tedy

$$e(t) = -\int_0^t \mathrm{e}^{-\frac{1}{h^2}(t-\tau)P_0} \varepsilon(\tau)\, \mathrm{d}\tau. \tag{3.18}$$

Odtud však plyne, že je

$$\|e(t)\|_h \leqq \int_0^t \|\mathrm{e}^{-\frac{1}{h^2}(t-\tau)P_0}\|_h \|\varepsilon(\tau)\|_h \, \mathrm{d}\tau. \tag{3.19}$$

Matice $\exp\left[-(1/h^2)(t-\tau)P_0\right]$ je symetrická a všechna její vlastní čísla jsou dána vzorcem $\exp\left[-(1/h^2)(t-\tau)\lambda_\nu\right]$, $\nu = 1, \dots, n-1$, kde λ_ν jsou vlastní čísla čísla matice P_0 (viz vzorec (2.82)). Všechna vlastní čísla matice $\exp\left[-(1/h^2)(t-\tau)P_0\right]$ jsou tedy pro $\tau \in \langle 0, t\rangle$ menší než jedna. Protože maticová norma indukovaná vektorovou normou $\|.\|_h$ je zřejmě obyčejná spektrální norma, plyne odtud, že platí

$$\|\mathrm{e}^{-\frac{1}{h^2}(t-\tau)P_0}\|_h \leqq 1, \quad \tau \in \langle 0, t\rangle. \tag{3.20}$$

Z nerovnosti (3.13) plyne, že pro $t \in \langle 0, T\rangle$ platí

$$\|\varepsilon(t)\|_h = \left[h \sum_{i=1}^{n-1} |\varepsilon_i(t)|^2\right]^{1/2} \leqq (hnM^2h^4)^{1/2} = Mh^2. \tag{3.21}$$

Z nerovností (3.19), (3.20) a (3.21) však už snadno dostáváme, že je

$$\|e(t)\|_h \leqq MTh^2, \tag{3.22}$$

což dokazuje větu.

3.1.2 Numerovova metoda

Tato metoda vychází z pozorování, že pro každou dostatečně hladkou funkci y platí

$$\begin{aligned} &y(x-h) - 2y(x) + y(x+h) = \\ &= \tfrac{1}{12}h^2[y''(x-h) + 10y''(x) + y''(x+h)] + O(h^6). \end{aligned} \tag{3.23}$$

Máme-li na mysli opět úlohu (2.1) až (2.3) a píšeme-li v této rovnici $\partial u(x,t)/\partial t$ místo $y''(x)$, dostaneme po zanedbání chybového členu pro přibližné řešení $u_k(t)$ soustavu rovnic

$$\begin{aligned} &\tfrac{1}{12}[\dot{u}_{k-1}(t) + 10\dot{u}_k(t) + \dot{u}_{k+1}(t)] = \frac{1}{h^2}[u_{k-1}(t) - 2u_k(t) + u_{k+1}(t)], \\ &k = 1, \dots, n-1, \; u_0(t) = u_n(t) = 0 \end{aligned} \tag{3.24}$$

s počátečními podmínkami (3.3).

Právě popsaná metoda se nazývá *Numerovova metoda*. O její konvergenci platí následující věta.

Věta 3.2. *Nechť řešení problému* (2.1) *až* (2.3) *existuje a má v R šest spojitých derivací podle x. Nechť $u^{(h)}(t)$ je vektor přibližného řešení vypočteného Numerovovou metodou a nechť $u(t)$ je vektor přesného řešení. Pak platí*

$$\|u^{(h)}(t) - u(t)\|_h = O(h^4). \tag{3.25}$$

D ů k a z je skoro totožný s důkazem věty 3.1, a proto budeme postupovat velmi rychle. Označíme-li matici na levé straně soustavy (3.24) symbolem B, zjistíme snadno, že pro vektor chyby e definovaný opět rovnicí (3.9) platí

$$B\,\dot{e}(t) = -\frac{1}{h^2} P_0 e(t) - \varepsilon(t), \tag{3.26}$$

přičemž složky vektoru ε jsou řádu $O(h^4)$. Pro matici B zřejmě platí $B = (1/12)(12I - P_0)$. Z tohoto vyjádření a ze znalosti vlastních čísel matice P_0 (viz znovu vzorec (2.82)) však ihned plyne, že matice B je pozitivně definitní a že pro ni platí $\|B\|_h = O(1)$ a $\|B^{-1}\|_h = O(1)$. Vynásobíme-li tedy rovnici (3.16) zleva maticí B^{-1}, dostaneme soustavu rovnic, jejíž řešení odhadneme úplně stejným způsobem jako řešení soustavy (3.16) v důkazu věty 3.1. Důkaz je hotov.

Numerovova metoda tedy konverguje s rychlostí řádu $O(h^4)$. Protože matice B je třídiagonální, vede numerické řešení soustavy (3.24) k řešení soustavy lineárních rovnic s třídiagonální maticí pro každý časový řádek. Numerovova metoda není tedy o nic pracnější než implicitní metoda sítí, která má ovšem řádově větší diskretizační chybu. To jej jistě pozoruhodné a pramení odtud i určitá popularita Numerovovy metody.

3.2 Semidiskrétní metody Galerkinova typu

Základní myšlenka tohoto postupu je obdobná jako u Galerkinovy metody pro řešení okrajových úloh pro obyčejné diferenciální rovnice, resp. pro parciální rovnice eliptického typu.

Protože Galerkinova metoda v eliptickém případě vychází z pojmu slabého řešení, musí náš první krok spočívat v zavedení podobného pojmu pro parabolický případ. Postup je zcela analogický jako v eliptickém případě, a popíšeme jej proto jen velmi stručně, a to na příkladě modelové úlohy (2.1) až (2.3).

Buď $\mathscr{H}_0^1$ podprostor těch funkcí u ze Sobolevova prostoru $\mathscr{H}^1$, které splňují okrajovou podmínku $u(0) = u(1) = 0$. Řekneme, že funkce $u\colon \langle 0, T\rangle \to \mathscr{H}_0^1$ je *slabým řešením* úlohy (2.1) až (2.3), platí-li pro každé $t \in (0, T\rangle$ a pro každé $v \in \mathscr{H}_0^1$ identita

$$(\dot{u}(t), v) + [u(t), v] = 0 \tag{3.27}$$

a pro $t = 0$ a pro každé $v \in \mathscr{H}_0^1$ identita

$$(u(0), v) = (g, v) \tag{3.28}$$

(počáteční podmínka). Zde kulaté závorky značí skalární součin v prostoru $\mathscr{L}_2(0,1)$, $[u, v]$ je bilineární forma definovaná vzorcem

$$[u, v] = \int_0^1 u'(x)v'(x)\,\mathrm{d}x \tag{3.29}$$

(a je to tedy skalární součin v prostoru $\mathscr{H}_0^1$) a tečka označuje stejně jako výše derivaci funkce u podle t.

Buď dále $\mathscr{D}_h$ konečnědimenzionální podprostor prostoru $\mathscr{H}_0^1$. *Semidiskrétní Galerkinovu aproximaci* řešení úlohy (3.27), (3.28) pak definujeme jako funkci $u_h\colon \langle 0, T\rangle \to \mathscr{D}_h$, pro niž platí

$$(\dot{u}_h(t), v) + [u_h(t), v] = 0, \quad t \in (0, T\rangle \tag{3.30}$$

a

$$(u_h(0), v) = (g, v) \tag{3.31}$$

pro každé $v \in \mathscr{D}_h$.

Je zřejmé, že pro zavedení uvedených pojmů není vůbec podstatné, že bilineární forma $[u, v]$ je definovaná právě vzorcem (3.29) a že na pravé straně rovnic (3.27), resp. (3.30) je nula. Úplně stejně by se postupovalo i v případě rovnice $\partial u/\partial t + Lu = f$, kde L je lineární eliptický diferenciální operátor druhého řádu v libovolném počtu prostorových proměnných.

Tvoří-li funkce $\Phi_1, \ldots, \Phi_N$ bázi v prostoru $\mathscr{D}_h$, je hledaná aproximace u_h tvaru

$$u_h(t, x) = \sum_{i=1}^{N} \eta_i(t)\Phi_i(x), \tag{3.32}$$

kde η_i jsou reálné funkce proměnné t. Dosadíme-li do rovnice (3.30) za u_h podle (3.32) a za v postupně funkce Φ_j, $j = 1, \ldots, N$, dostaneme

$$\begin{aligned} &\sum_{i=1}^{N} \dot{\eta}_i(t)(\Phi_i, \Phi_j) + \sum_{i=1}^{N} \eta_i(t)[\Phi_i, \Phi_j] = 0, \quad t \in (0, T\rangle, \\ &\sum_{i=1}^{N} \eta_i(0)(\Phi_i, \Phi_j) = (g, \Phi_j). \end{aligned} \tag{3.33}$$

Hledané koeficienty $\boldsymbol{\eta} = [\eta_1(t), \ldots, \eta_{n-1}(t)]^T$ tedy splňují soustavu diferenciálních rovnic

$$B\dot{\boldsymbol{\eta}}(t) + A\boldsymbol{\eta}(t) = 0, \quad t \in (0, T\rangle, \tag{3.34}$$

s počáteční podmínkou

$$B\boldsymbol{\eta}(0) = \boldsymbol{\eta}^{(0)}, \tag{3.35}$$

kde A a B jsou symetrické matice řádu N dané vzorci

$$\begin{aligned} A &= \{[\Phi_i, \Phi_j]\} = \left\{ \int_0^1 \Phi_i'(x)\Phi_j'(x)\,\mathrm{d}x \right\}, \\ B &= \{(\Phi_i, \Phi_j)\} = \left\{ \int_0^1 \Phi_i(x)\Phi_j(x)\,\mathrm{d}x \right\} \end{aligned} \tag{3.36}$$

a $\boldsymbol{\eta}^{(0)}$ je N-dimenzionální vektor o složkách (g, Φ_i). Matice A a B jsou zřejmě symetrické a pozitivně definitní, soustavu diferenciálních rovnic (3.34) lze tedy psát ve tvaru

$$\dot{\boldsymbol{\eta}}(t) + \tilde{A}\boldsymbol{\eta}(t) = 0, \tag{3.37}$$

kde

$$\tilde{A} = B^{-1}A. \tag{3.38}$$

Za prostor $\mathscr{D}_h$ se většinou volí prostor konečných prvků, neboť pak je zaručeno, že matice A i B jsou pásové, což podstatně usnadňuje řešení soustavy (3.34). Poznamenejme v této souvislosti, že v důsledku přítomnosti matice B v soustavě (3.34) nemá smysl řešit tuto soustavu explicitní metodou, neboť tak jak tak musíme v každém kroku řešit soustavu lineárních rovnic. V tomto smyslu je zde tedy situace podobná jako u Numerovovy metody vyšetřované v odst. 3.1.2.

V dalším textu se omezíme na případ, že za prostor $\mathscr{D}_h$ bereme prostor $\mathscr{D}_{h1}^{(0)}$ po částech lineárních funkcí, které splňují nulové Dirichletovy okrajové podmínky a že příslušné dělení intervalu $\langle 0, 1\rangle$ je ekvidistantní s podintervaly délky h. Na tomto speciálním případě pak ukážeme jeden z možných způsobů vyšetřování konvergence studované metody. Uvidíme, že už i v této jednoduché situaci nebude ani zdaleka vše tak elementární, jako tomu bylo u metody sítí nebo u klasické metody přímek. Než zformulujeme a dokážeme příslušné tvrzení, uvedeme několik lemmat, která při

jeho důkazu použijeme. Abychom však nemuseli psát příliš mnoho indexů, umluvíme se ještě předtím, že pod symbolem absolutní hodnoty budeme v dalším rozumět euklidovskou normu vektoru nebo jí indukovanou normu matice a že obyčejný symbol normy bez jakéhokoliv indexu bude značit normu v prostoru $\mathscr{L}_2(0,1)$.

Lemma 3.1. *Buď $\mathscr{D}_h \subset \mathscr{H}_0^1$ prostor po částech lineárních funkcí s ekvidistantním dělením o délce dílku $h = 1/(N+1)$ a buď $\Phi_1, \dots, \Phi_N$ jeho standardní báze. Pak pro matice A a B definované vzorci* (3.36) *platí*

$$A = \frac{1}{h} P_0, \quad B = \tfrac{1}{6} h(6I - P_0), \tag{3.39}$$

kde P_0 je matice (2.81) *a*

$$|A| = O\Big(\frac{1}{h}\Big), \quad |B| = O(h), \quad |B^{-1}| = O\Big(\frac{1}{h}\Big). \tag{3.40}$$

D ů k a z . Vzorce (3.39) odvodíme přímým výpočtem za použití konkrétního tvaru bázových funkcí (viz vzorec (4.83) z kap. II). Z těchto vzorců a ze znalosti vlastních čísel matice P_0 (viz vzorec (2.82)) však plyne, že vlastní čísla $\lambda_\nu^{(A)}$ a $\lambda_\nu^{(B)}$ matice A a B jsou dána vzorci

$$\lambda_\nu^{(A)} = \frac{4}{h} \sin^2 \frac{\nu\pi}{2(N+1)}, \quad \nu = 1, \dots, N, \tag{3.41}$$

a

$$\lambda_\nu^{(B)} = \tfrac{1}{3} h\Big(1 + 2\cos^2 \frac{\nu\pi}{2(N+1)}\Big), \quad \nu = 1, \dots, N. \tag{3.42}$$

Odhady (3.40) pak plynou odtud a z toho, že je $|A| = \lambda_{\max}^{(A)}$, $|B| = \lambda_{\max}^{(B)}$ a $|B^{-1}| = 1/\lambda_{\min}^{(B)}$.

Lemma 3.2. *Existují kladné konstanty c a C takové, že pro každou funkci*

$$\eta = \sum_{i=1}^{N} \eta_i \Phi_i \in \mathscr{D}_h \tag{3.43}$$

platí

$$ch|\eta|^2 \leqq \|\eta\|^2 \leqq Ch|\eta|^2. \tag{3.44}$$

D ů k a z . Je

$$\|\eta\|^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} \eta_i \eta_j (\Phi_i, \Phi_j) = \boldsymbol{\eta}^T B \boldsymbol{\eta}. \tag{3.45}$$

Protože pro symetrickou a pozitivně definitní matici B platí

$$\lambda_{\min}^{(B)} |\boldsymbol{\eta}|^2 \leqq \boldsymbol{\eta}^T B \boldsymbol{\eta} \leqq \lambda_{\max}^{(B)} |\boldsymbol{\eta}|^2, \tag{3.46}$$

plyne tvrzení ze vzorce (3.42).

Lemma 3.3. *Nechť je dána množina matic C_t, kde t probíhá libovolnou množinu indexů I, a nechť pro každé $t \in I$ je matice C_t symetrická a pozitivně semidefinitní. Pak existuje konstanta M taková, že pro každé $t \in I$ platí*

$$|C_t \mathrm{e}^{-C_t}| \leqq M. \tag{3.47}$$

D ů k a z . Vlastní čísla symetrické matice $C_t \mathrm{e}^{-C_t}$ jsou tvaru $\lambda_t \mathrm{e}^{-\lambda_t}$, kde λ_t je vlastní číslo matice C_t. Protože matice C_t je pozitivně semidefinitní, je $\lambda_t \in \langle 0, \infty)$. Reálná funkce $\lambda \mathrm{e}^{-\lambda}$ reálné proměnné $\lambda \in \langle 0, \infty)$ však nabývá zřejmě v bodě $\lambda = 1$ svého maxima. Odtud už tvrzení plyne.

Lemma 3.4. *Slabé řešení úlohy $-y'' = \theta$, $y(0) = y(1) = 0$, kde $\theta \in \mathscr{L}_2(0,1)$ existuje a leží dokonce v prostoru $\mathscr{H}^2$. Navíc existuje konstanta M taková, že pro každé $\theta \in \mathscr{L}_2(0,1)$ platí odhad*

$$\|y\|_{\mathscr{H}^2} \leqq M\|\theta\|. \tag{3.48}$$

D ů k a z . Za uvedených předpokladů je výraz (θ, v) zřejmě lineární funkcionál nad prostorem $\mathscr{H}_0^1$. Podle Rieszovy věty tedy existuje právě jeden prvek $y \in \mathscr{H}_0^1$ takový, že platí

$$[y, v] = (\theta, v) \tag{3.49}$$

pro každé $v \in \mathscr{H}_0^1$. Slabé řešení uvažované okrajové úlohy tedy existuje a je určeno jednoznačně. Snadno se ověří, že je lze psát ve tvaru

$$y(x) = (1-x) \int_0^x \xi\theta(\xi)\,\mathrm{d}\xi + x \int_x^1 (1-\xi)\theta(\xi)\,\mathrm{d}\xi. \tag{3.50}$$

Z tohoto vyjádření však plyne, že je

$$y'(x) = -\int_0^x \xi\theta(\xi)\,\mathrm{d}\xi + \int_x^1 (1-\xi)\theta(\xi)\,\mathrm{d}\xi. \tag{3.51}$$

Protože obě funkce $\xi\theta(\xi)$ i $(1-\xi)\theta(\xi)$ jsou lebesgueovsky integrovatelné, je funkce y' absolutně spojitá a platí pro ni $y''(x) = -\theta(x)$ skoro všude na intervalu $\langle 0,1\rangle$. Odtud však už nerovnost (3.48) plyne bezprostředně.

Věta 3.3. *Nechť u je slabé řešení problému* (2.1) *a*ž (2.3) *a nechť pro ně platí $u(t) \in \mathscr{H}^2$ pro každé $t \in \langle 0, T\rangle$. Nechť $\mathscr{D}_h$ je podprostor prostoru $\mathscr{H}_0^1$ z lemmatu 3.1 a buď u_h příslušné semidiskrétní řešení. Pak existují konstanty M a $h_0 > 0$ takové, že platí*

$$\max_{t\in\langle 0,T\rangle} \|u(t) - u_h(t)\| \leqq h^2 M\Big(1 + \ln\frac{T}{h^2}\Big) \max_{t\in\langle 0,T\rangle} \|u(t)\|_{\mathscr{H}^2} \tag{3.52}$$

pro každé $h \leqq h_0$.

D ů k a z . Buď t pevně zvolené a buď $\varphi_h : \langle 0, t\rangle \to \mathscr{D}_h$ pomocná funkce, která je řešením úlohy

$$-(\dot{\varphi}_h(s), v) + [\varphi_h(s), v], \quad s \in \langle 0, t\rangle , \quad v \in \mathscr{D}_h, \tag{3.53}$$

s „koncovou" podmínkou

$$\varphi_h(t) = e_h(t). \tag{3.54}$$

Přitom jsme položili

$$e_h(s) = u_h(s) - \tilde{u}_h(s), \tag{3.55}$$

kde funkce $\tilde{u}_h(s)$ je ortogonální projekce v prostoru $\mathscr{H}_0^1$ přesného řešení $u(s)$ na podprostor $\mathscr{D}_h$. Pro každou funkci $v \in \mathscr{D}_h$ a pro každé $s \in \langle 0, T\rangle$ tedy platí

$$[u(s) - \tilde{u}_h(s), v] = 0. \tag{3.56}$$

Úloha (3.53), (3.54) má právě jedno řešení, neboť položíme-li

$$\varphi_h(s) = \sum_{i=1}^{N} \xi_i(s)\Phi_i(x), \tag{3.57}$$

a

$$e_h(t) = \sum_{i=1}^{N} \xi_i^{(0)}\Phi_i(x), \tag{3.58}$$

je ekvivalentní s úlohou

$$-B\dot{\boldsymbol{\xi}}(s) + A\boldsymbol{\xi}(s) = 0, \quad s \in (0, t), \qquad \boldsymbol{\xi}(t) = \boldsymbol{\xi}^{(0)}. \tag{3.59}$$

Položme v (3.53) $v = e_h(s)$. Dostaneme

$$-\big(\dot{\varphi}_h(s), e_h(s)\big) + [\varphi_h(s), e_h(s)] = 0. \tag{3.60}$$

Integrací v mezích od 0 do t odtud máme

$$\int_0^t \big(\dot{\varphi}_h(s), e_h(s)\big)\,\mathrm{d}\,s = \int_0^t [\varphi_h(s), e_h(s)]\,\mathrm{d}\,s. \tag{3.61}$$

Úprava per partes dá (je třeba si uvědomit, že skalární součin v prostoru $\mathscr{L}_2$ je integrál)

$$\big(\varphi_h(t), e_h(t)\big) - \big(\varphi_h(0), e_h(0)\big) - \int_0^t \big(\varphi_h(s), \dot{e}_h(s)\big)\,\mathrm{d}\,s = \tag{3.62}$$
$$= \int_0^t [\varphi_h(s), e_h(s)]\,\mathrm{d}\,s$$

neboli, protože je $\varphi_h(t) = e_h(t)$,

$$\|e_h(t)\|^2 = \big(\varphi_h(0), e_h(0)\big) + \int_0^t \big\{\big(\varphi_h(s), \dot{e}_h(s)\big) + [\varphi_h(s), e_h(s)]\big\}\,\mathrm{d}\,s. \tag{3.63}$$

Položme dále

$$\theta(s) = u(s) - \tilde{u}_h(s). \tag{3.64}$$

Je

$$\begin{aligned}
&\int_0^t \big\{\big(\dot{\theta}(s), \varphi_h(s)\big) + [\theta(s), \varphi_h(s)]\big\}\,\mathrm{d}\,s = \\
&= \int_0^t \big\{\big(\dot{u}(s) - \dot{\tilde{u}}_h(s), \varphi_h(s)\big) + [u(s) - \tilde{u}_h(s), \varphi_h(s)]\big\}\,\mathrm{d}\,s = \\
&= \int_0^t \big\{\big(\dot{u}(s), \varphi_h(s)\big) + [u(s), \varphi_h(s)]\big\}\,\mathrm{d}\,s - \\
&\quad - \int_0^t \big\{\big(\dot{\tilde{u}}(s), \varphi_h(s)\big) + [\tilde{u}_h(s), \varphi_h(s)]\big\}\,\mathrm{d}\,s = \\
&= \int_0^t \big\{\big(\dot{u}_h(s), \varphi_h(s)\big) + [u_h(s), \varphi_h(s)]\big\}\,\mathrm{d}\,s - \\
&\quad - \int_0^t \big\{\big(\dot{\tilde{u}}(s), \varphi_h(s)\big) + [\tilde{u}_h(s), \varphi_h(s)]\big\}\,\mathrm{d}\,s = \\
&= \int_0^t \big\{\big(\dot{e}_h(s), \varphi_h(s)\big) + [e_h(s), \varphi_h(s)]\big\}\,\mathrm{d}\,s.
\end{aligned} \tag{3.65}$$

Dále je

$$\begin{aligned}
\big(\theta(0), \varphi_h(0)\big) &= \big(u(0) - \tilde{u}_h(0), \varphi_h(0)\big) = \\
&= \big(u_h(0) - \tilde{u}_h(0) + u(0) - u_h(0), \varphi_h(0)\big) = \\
&= \big(e_h(0), \varphi_h(0)\big) + \big(u(0) - u_h(0), \varphi_h(0)\big) = \big(e_h(0), \varphi_h(0)\big),
\end{aligned} \tag{3.66}$$

přičemž poslední rovnost platí v důsledku počátečních podmínek, které splňují funkce u a u_h. Dosadíme-li poslední dva vztahy do rovnice (3.63), máme

$$\|e_h(t)\|^2 = \big(\theta(0), \varphi_h(0)\big) + \int_0^t \big\{\big(\dot{\theta}(s), \varphi_h(s)\big) + [\theta(s), \varphi_h(s)]\big\}\,\mathrm{d}\,s. \tag{3.67}$$

Provedeme-li v prvním integrálu na pravé straně rovnice (3.67) integraci per partes a užijeme-li toho, že podle definice funkce $\tilde{u}_h$ je $[\theta(s), \varphi_h(s)] = [u(s) - \tilde{u}_h(s), \varphi_h(s)] = 0$, dostáváme konečně

$$\|e_h(t)\|^2 = \big(\theta(t), \varphi_h(t)\big) - \int_0^t \big(\theta(s), \dot{\varphi}_h(s)\big)\,\mathrm{d}\,s. \tag{3.68}$$

Abychom mohli právě odvozenou reprezentaci podstatné části chyby dál zužitkovat, je třeba odhadnout funkci φ_h a její derivaci. Použijeme k tomu vyjádření

(3.57). Řešení soustavy diferenciálních rovnic (3.59) lze psát ve tvaru

$$\xi(s) = \mathrm{e}^{-(t-s)\tilde{A}}\xi(t), \tag{3.69}$$

kde

$$\tilde{A} = B^{-1}A. \tag{3.70}$$

Protože matice $\tilde{A}$ je pozitivně definitní, platí

$$|\mathrm{e}^{-(t-s)\tilde{A}}| \leqq 1 \tag{3.71}$$

pro každé $s \in \langle 0, t\rangle$ (srv. důkaz věty 3.1). Z rovnice (3.69) proto plyne, že je

$$|\xi(s)|^2 \leqq |\xi(t)|^2, \quad s \in \langle 0, t\rangle. \tag{3.72}$$

Derivujeme-li rovnici (3.69) podle s, máme

$$\dot{\xi}(s) = \tilde{A}\mathrm{e}^{-(t-s)\tilde{A}}\xi(t) = \frac{1}{t-s}(t-s)\tilde{A}\mathrm{e}^{-(t-s)\tilde{A}}\xi(t). \tag{3.73}$$

Odtud použitím lemmatu 3.3 dostáváme

$$|\dot{\xi}(s)| \leqq \frac{M}{t-s}|\xi(t)|. \tag{3.74}$$

Pišme dále

$$\int_0^t |\dot{\xi}(s)|\,\mathrm{d}\,s = \int_0^{t-h^2} |\dot{\xi}(s)|\,\mathrm{d}\,s + \int_{t-h^2}^t |\dot{\xi}(s)|\,\mathrm{d}\,s. \tag{3.75}$$

Podle (3.74) je

$$\int_0^{t-h^2} |\dot{\xi}(s)|\,\mathrm{d}\,s \leqq M|\xi(t)| \int_0^{t-h^2} \frac{\mathrm{d}\,s}{t-s} = M|\xi(t)| \ln \frac{t}{h^2}. \tag{3.76}$$

Dosadíme-li do druhého integrálu na pravé straně rovnice (3.75) za $\dot{\xi}(s)$ z rovnice (3.59) a normu vektoru $\xi(s)$ odhadneme podle (3.72), dostáváme

$$\int_{t-h^2}^t |\dot{\xi}(s)|\,\mathrm{d}\,s \leqq |\xi(t)|\,|\tilde{A}|h^2. \tag{3.77}$$

Podle lemmatu 3.1 je však $|\tilde{A}| = O(1/h^2)$, takže celkem dostáváme z rovnice (3.75) a nerovností (3.76) a (3.77)

$$\int_0^t |\dot{\xi}(s)|\,\mathrm{d}\,s \leqq M|\xi(t)|\Big(1 + \ln \frac{t}{h^2}\Big). \tag{3.78}$$

Z lemmatu 3.2 a z nerovnosti (3.72) plyne, že je

$$\|\varphi_h(s)\|^2 \leqq Ch|\xi(s)|^2 \leqq Ch|\xi(t)|^2 \leqq \frac{C}{c}\|\varphi_h(t)\|^2. \tag{3.79}$$

Platí tedy (je $\varphi_h(t) = e_h(t)$)

$$\|\varphi_h(s)\| \leqq M\|e_h(t)\|. \tag{3.80}$$

(Konstanta M je zde i dále samozřejmě generická.) Zcela analogicky odvodíme z nerovnosti (3.78) nerovnost

$$\int_0^t \|\dot{\varphi}_h(s)\|\,\mathrm{d}\,s \leqq M\|e_h(t)\|\Big(1 + \ln \frac{t}{h^2}\Big). \tag{3.81}$$

Nyní jsme už skoro hotovi. Z vyjádření (3.68) plyne

$$\|e_h(t)\|^2 \leqq \|\theta(t)\|\,\|e_h(t)\| + \int_0^t \|\theta(s)\|\,\|\dot{\varphi}_h(s)\|\,\mathrm{d}\,s. \tag{3.82}$$

Dosadíme-li do této nerovnosti odhad (3.81), dostáváme

$$\|e_h(t)\|^2 \leqq (\max_{s\in(0,t)} \|\theta(s)\|)\|e_h(t)\|\Big[1 + M\Big(1 + \ln \frac{t}{h^2}\Big)\Big], \tag{3.83}$$

a tedy celkem

$$\|e_h(t)\| \leqq M\Big(1 + \ln \frac{t}{h^2}\Big) \max_{s\in(0,t)} \|\theta(s)\|. \tag{3.84}$$

Píšeme-li nyní celkovou diskretizační chybu $u - u_h$ ve tvaru

$$u - u_h = u - \tilde{u}_h + \tilde{u}_h - u_h = \theta - e_h, \tag{3.85}$$

dostáváme pro ni z odhadu (3.84) nerovnost

$$\|u - u_h\| \leqq M\Big(1 + \ln \frac{t}{h^2}\Big) \max_{s\in(0,t)} \|\theta(s)\|. \tag{3.86}$$

Abychom dokázali nerovnost (3.52), stačí už jen odhadnout normu funkce θ. Řešme k tomu cíli pro tuto funkci okrajovou úlohu zformulovanou v lemmatu 3.4. Funkce θ patří podle (3.64) nejen do prostoru $\mathscr{L}_2$, ale dokonce do $\mathscr{H}_0^1$, takže ji lze dosadit do identity (3.49). Dostaneme tak rovnici

$$\|\theta\|^2 = [y, \theta] = [y - y_h, \theta], \tag{3.87}$$

kde y_h je interpolace funkce θ v prostoru $\mathscr{D}_h$. Je tomu tak proto, že podle definice funkce θ je $[\theta, v] = 0$ pro každou funkci $v \in \mathscr{D}_h$. Z rovnice (3.87), z věty 4.5 z kap. II (viz str. 227) a z lemmatu 3.4 postupně dokážeme, že platí

$$\|\theta\|^2 \leqq \|y - y_h\|_{\mathscr{H}^1}\,\|\theta\|_{\mathscr{H}^1} \leqq Mh\|y\|_{\mathscr{H}^2}\,\|\theta\|_{\mathscr{H}^1} \leqq M_1 h\|\theta\|\,\|\theta\|_{\mathscr{H}^1}. \tag{3.88}$$

Odtud však plyne, že je

$$\|\theta\| \leqq Mh\|\theta\|_{\mathscr{H}^1}. \tag{3.89}$$

Na druhé straně však je

$$\|\theta\|_{\mathscr{H}^1} = \|u - \tilde{u}_h\|_{\mathscr{H}^1} \leqq \|u - v\|_{\mathscr{H}^1} \tag{3.90}$$

pro každou funkci $v \in \mathscr{D}_h$, neboť $\tilde{u}_h$ je ortogonální projekce na podprostor $\mathscr{D}_h$. Tedy speciálně, opět podle věty 4.5 z kap. II, máme

$$\|\theta\|_{\mathscr{H}^2} \leqq Mh\|u\|_{\mathscr{H}^2} \tag{3.91}$$

Celkem tedy platí

$$\|\theta(s)\| \leqq Mh^2\|u(s)\|_{\mathscr{H}^2} \tag{3.92}$$

Odtud a z nerovnosti (3.86) však už plyne nerovnost (3.52) bezprostředně. Věta je dokázána.

3.3 Metody Rotheova typu

V předchozích odstavcích jsme popsali ty semidiskrétní metody pro numerické řešení parciálních diferenciálních rovnic parabolického typu, které vznikly tak, že diskretizace se prováděla pouze vzhledem k prostorovým proměnným. Řešení původního problému pak bylo aproximováno úlohou s počátečními podmínkami pro soustavu obyčejných diferenciálních rovnic. Obrácení tohoto postupu, tj. provedení diskretizace pouze vzhledem k proměnné t, tvoří základní myšlenku *metod Rotheova typu*. Podrobněji se budeme zabývat jen nejjednodušší z nich a omezíme se opět na modelovou úlohu (2.1) až (2.3).

Položíme-li $t_l = l\tau$, kde $\tau = T/r$, nahradíme-li derivaci $\partial u/\partial t$ v bodě (x, t_l) podílem $[u(x,t_l) - u(x,t_{l-1})]/\tau$ a chybu, které se přitom dopustíme, zanedbáme, je přibližné řešení $u^{(l)}(x)$, $l = 1, \ldots, r$, v čase $t = t_l$ dáno řešením diferenciální rovnice

$$-\frac{\mathrm{d}^2 u^{(l)}}{\mathrm{d}x^2} + \frac{1}{\tau}u^{(l)} = \frac{1}{\tau}u^{(l-1)}, \quad l = 1, \ldots, r, \tag{3.93}$$

s okrajovými podmínkami

$$u^{(l)}(0) = u^{(l)}(1) = 0. \tag{3.94}$$

Popsaná metoda tedy vznikne tak, že se na danou diferenciální rovnici díváme jako na obyčejnou diferenciální rovnici vzhledem k proměnné t a tu řešíme implicitní Eulerovou metodou. Náhradní problém je tedy okrajová úloha pro obyčejnou diferenciální rovnici, k jejímuž řešení lze většinou bez podstatných obtíží užít kteroukoliv z metod popsaných v kap. II. Často přitom vznikne některá z variant metody sítí.

Při vyšetřování konvergence Rotheovy metody vyjdeme ze dvou elementárních lemmat, která vyjadřují některé jednoduché vlastnosti diferenciální rovnice typu (3.93).

Lemma 3.5. *Buď y funkce spojitá v intervalu $\langle 0,1\rangle$, která je navíc dvakrát spojitě diferencovatelná v $(0,1)$. Buď dále q kladná konstanta a nechť platí*

$$-y''(x) + qy(x) \leqq 0 \tag{3.95}$$

pro každé $x \in (0,1)$. Položme konečně

$$M = \max_{x\in\langle 0,1\rangle} y(x) \tag{3.96}$$

a předpokládejme, že je $M > 0$. Pak platí

$$M = \max[y(0), y(1)]. \tag{3.97}$$

D ů k a z . Buď $M = y(x_0)$ a předpokládejme, že je $x_0 \in (0,1)$. Protože funkce y nabývá v bodě x_0 svého maxima a je v něm dvakrát spojitě diferencovatelná, platí $y'(x_0) = 0$ a $y''(x_0) \leqq 0$. Podle (3.95) v tomto bodě platí

$$0 \geqq -y''(x_0) + qy(x_0) \geqq qy(x_0) = qM > 0, \tag{3.98}$$

což není možné. Proto je $x_0 = 0$ nebo $x_0 = 1$ a důkaz lemmatu je hotov.

Diferenciální operátor na levé straně nerovnosti (3.95) splňuje tedy princip maxima.

Lemma 3.6. *Buďte u a η dvě funkce spojité v intervalu $\langle 0,1\rangle$ a dvakrát spojitě diferencovatelné v otevřeném intervalu $(0,1)$. Nechť dále pro každé $x \in (0,1)$ platí*

$$|-u''(x) + qu(x)| \leqq -\eta''(x) + q\eta(x), \tag{3.99}$$

kde q je kladná konstanta. Nechť konečně je $|u(0)| \leqq \eta(0)$ a $|u(1)| \leqq \eta(1)$. Pak platí

$$|u(x)| \leqq \eta(x) \tag{3.100}$$

pro každé $x \in \langle 0,1\rangle$.

D ů k a z . Položme $\varphi = \pm u - \eta$. Z nerovnosti (3.99) ihned plyne, že pro každé $x \in (0,1)$ platí

$$-\varphi''(x) + q\varphi(x) = \mp u''(x) \pm qu(x) + \eta''(x) - q\eta(x) \leqq 0. \tag{3.101}$$

Buď nyní

$$M = \max_{x\in\langle 0,1\rangle} [\pm u(x) - \eta(x)] \tag{3.102}$$

a předpokládejme, že je $M > 0$. Pak je podle lemmatu 3.5

$$M = \max[\pm u(0) - \eta(0), \pm u(1) - \eta(1)] \leqq 0. \tag{3.103}$$

To je však spor dokazující, že je $M \leqq 0$. Odtud už tvrzení lemmatu plyne bezprostředně.

Věta 3.4. *Nechť řešení úlohy (2.1) až (2.3) existuje a má v množině $R = \{(x,t);\ 0 \leqq x \leqq 1, 0 \leqq t \leqq T\}$ dvě spojité derivace podle proměnné t. Nechť dále $u^{(l)}(x)$ je přibližné řešení vypočtené Rotheovou metodou (3.93), (3.94). Pak existují konstanty M a $\tau_0 > 0$ takové, že pro $\tau \leqq \tau_0$ platí*

$$|u^{(l)}(x) - u(x,t_l)| \leqq M\tau, \quad l = 1, \ldots, r. \tag{3.104}$$

D ů k a z . Chyba $\eta^{(l)}(x) = u^{(l)}(x) - u(x, t_l)$ Rotheovy metody splňuje rovnice

$$-\frac{\mathrm{d}^2\eta^{(l)}}{\mathrm{d}x^2} + \frac{1}{\tau}\eta^{(l)} = \frac{1}{\tau}\eta^{(l-1)} + \varepsilon^{(l)}, \tag{3.105}$$
$$\eta^{(l)}(0) = \eta^{(l)}(1) = 0, \quad l = 1, \ldots, r,$$

a

$$\eta^{(0)} = 0, \tag{3.106}$$

přičemž existuje konstanta M taková, že pro každé $x \in \langle 0, 1\rangle$ a pro každé dostatečně malé τ platí

$$|\varepsilon^{(l)}(x)| \leqq M\tau. \tag{3.107}$$

Položme v rovnici (3.105) nejprve $l = 1$. Pak pro chybu v prvním časovém řádku platí

$$-\frac{\mathrm{d}^2\eta^{(1)}}{\mathrm{d}x^2} + \frac{1}{\tau}\eta^{(1)} = \varepsilon^{(1)}, \tag{3.108}$$
$$\eta^{(1)}(0) = \eta^{(1)}(1) = 0.$$

Podle lemmatu 3.6 je však řešení této okrajové úlohy majorizováno funkcí

$$\varphi(x) = M\tau\psi(x), \tag{3.109}$$

kde funkce ψ je řešením pomocné okrajové úlohy

$$-\psi'' + \frac{1}{\tau}\psi = 1, \tag{3.110}$$
$$\psi(0) = \psi(1) = 0.$$

Pro funkci φ totiž platí vzhledem k nerovnosti (3.107)

$$-\varphi''(x) + \frac{1}{\tau}\varphi(x) = M\tau \geqq \left| -\frac{\mathrm{d}^2\eta^{(1)}(x)}{\mathrm{d}x^2} + \frac{1}{\tau}\eta^{(1)}(x)\right|. \tag{3.111}$$

Funkce ψ, která řeší okrajovou úlohu (3.110), je však dána vzorcem

$$\psi(x) = \tau - \tau\frac{\cosh\left(\frac{1}{\tau^{1/2}}x - \frac{1}{2\tau^{1/2}}\right)}{\cosh\frac{1}{2\tau^{1/2}}}, \tag{3.112}$$

jak se snadno zjistí přímým výpočtem. Odtud však plyne, že pro každé $x \in \langle 0, 1\rangle$ platí

$$0 \leqq \psi(x) \leqq \tau. \tag{3.113}$$

(Abychom to nahlédli, stačí vypočíst maximum funkce (3.112) v intervalu $\langle 0, 1\rangle$.) Je tedy

$$|\eta^{(1)}(x)| \leqq M\tau\psi(x) \leqq M\tau^2. \tag{3.114}$$

Nyní již snadno dokážeme úplnou indukcí, že platí

$$|\eta^{(l)}(x)| \leqq Ml\tau^2. \tag{3.115}$$

Skutečně, předpokládáme-li, že nerovnost (3.115) platí pro $l - 1$, dostaneme z (3.105), (3.107) a z indukčního předpokladu, že je

$$\left| -\frac{\mathrm{d}^2\eta^{(l)}(x)}{\mathrm{d}x^2} + \frac{1}{\tau}\eta^{(l)}(x)\right| \leqq M(l-1)\tau + M\tau = Ml\tau. \tag{3.116}$$

Odtud však úplně stejnou úvahou, jako jsme dospěli k nerovnosti (3.114), dokážeme, že platí (3.115). Věta je dokázána.

CVIČENÍ

1. Formulujte přesně hladkostní předpoklady, které stačí k platnosti vzorce (2.5).
2. Odvoďte vzorec (2.25).
3. Proveďte příslušné numerické výpočty dokazující komentář pod vzorcem (2.25).
4. Prověřte platnost vzorce (2.32).
5. Dokažte platnost principu maxima pro operátor $L_{h,\tau}^{(0)}$ daný vzorcem (2.4) (tj. pro explicitní schéma).
6. Dokažte, že maticová norma indukovaná normou prostoru $\mathbf{E}_h^{(m-1)}$ je obyčejná spektrální norma.
7. Na základě rovnic (3.40) a (3.58) z kap. II odvoďte tvar operátorů $L_{h,\tau}^{(\alpha)}$ a $I_{h,\tau}^{(i,\alpha)}$ daných vzorci (2.117) a (2.118).
8. Formulujte přesně hladkostní předpoklady ve větách 2.6 a 2.7 (viz str. 353) a dokažte jejich tvrzení.
9. Dokažte konvergenci explicitní metody pro rovnici (2.109) s Dirichletovou okrajovou podmínkou. Nalezněte podmínku stability.
10. Proveďte podrobně výpočty potřebné k důkazu lemmatu 2.4 ze str. 360.
11. Dokažte, že splnění podmínky (2.161) implikuje platnost principu maxima pro operátor $L_{h,\tau}^{(0)}$ daný vzorcem (2.117).
12. Formulujte a dokažte pro D'jakonovovu metodu střídavých směrů větu analogickou k větě 2.16 (viz str. 384).
13. Dokažte stejnoměrnou stabilitu vzhledem k počátečním podmínkám Janenkovy metody (2.301).
14. Dokažte platnost odhadu (3.23).

POZNÁMKY K LITERATUŘE

Čl. 1. Teorii parciálních diferenciálních rovnic parabolického typu je věnována skoro stejně rozsáhlá literatura jako teorii eliptických rovnic, viz např. Petrovskij (1960), Friedman (1964), Ladyženskaja, Solonnikov a Uralceva (1967), pokud jde o klasickou teorii, Lions (1961) a Lions-Magenes (1968), pokud jde o funkcionálně analytické metody. Dobrými obecnými prameny k problematice této kapitoly jsou např. Collatz (1951), Forsythe, Wasow (1960), Babuška, Práger, Vitásek (1964), Berezin, Židkov (1966), Babuška, Práger, Vitásek (1966), Mitchell (1969), Ansorge (1978) a Marčuk (1980).

Čl. 2. Základy metody sítí jsou vyloženy ve všech obecných pramenech uvedených k čl. 1; kniha Sauljevova (1960) je věnována speciálně řešení parabolických rovnic metodou sítí. Ke studiu problémů stability jsou velmi dobrými prameny Richtmyer (1957), Richtmyer, Morton (1967) a Samarskij (1971). Obsáhlou bibliografii metody sítí jakož i ostatních metod pro řešení parabolických rovnic nalezne čtenář ve sborníku Jacobsově (1977). Speciálně metodami střídavých směrů a lokálně jednorozměrnými metodami se zabývá Janenko (1966).

Čl. 3. Klasická metoda přímek je populární zejména v sovětské literatuře a zabývá se jí např. Berezin, Židkov (1966). Základní poučení o ní lze nalézt také ve sborníku vydaném Hallem a Wattem (1976). Semidiskrétní metody Galerkinova typu se zejména v poslední době dosti intenzívně studují a poučení o nich lze nalézt už v celé řadě knižních publikací. Uveďme z nich jmenovitě knihu Strangovu a Fixovu (1973), Mitchellovu a Waitovu (1977), Thoméovu (1984), Goeringovu, Roosovu a Tobiskovu (1988) a zejména pak knihu Johnsonovu (1988), z níž jsme také vydatně čerpali. Rotheova metoda se už dlouho užívá k teoretickým úvahám, viz např. Ladyženskaja, Solonnikov, Uralceva (1967). Kniha Rektorysova (1982) je monografií, která je v podstatě celá věnována této metodě.

LITERATURA

ANSORGE, R.: Differenzenapproximationen partieller Anfangsweraufgaben. Stuttgart, Teubner 1978.

BABUŠKA, I. – PRÁGER, M. – VITÁSEK, E.: Numerické řešení diferenciálních rovnic. Praha, SNTL 1964.

BABUŠKA, I. – PRÁGER, M. – VITÁSEK, E.: Numerical Processes in Differential Equations. London–New York–Sydney, Interscience Publishers 1966.

BEREZIN, I.S. – ŽIDKOV, N.P.: Metody vyčislenij. 3. vyd. Moskva, Nauka 1966, 2 sv.

COLLATZ, L.: Numerische Behandlung von Differentialgleichungen. Berlin–Göttingen–Heidelberg, Springer–Verlag 1951. (Překlad do ruštiny: Moskva, IL 1953.)

D'JAKONOV, E.G.: Raznostnyje schemy s rasščepljajuščimsja operatorom. Ž. Vyčisl. Mat. i Mat. Fiz. *2*, 1962, s. 549 – 568.

FORSYTHE, K.E. – WASOV, W.R.: Finite Difference Methods for Partial Differential Equations. New York–London, J. Wiley and Sons 1960. (Překlad do ruštiny: Moskva, IL 1963.)

FRIEDMAN, A.: Partial Differential Equations of Parabolic Type. Englewood Cliffs, N.J., Prentice–Hall 1964.

GOERING, H. – ROOS, H.G. – TOBISKA, L.: Finite-Element-Methode. Berlin, Akademie–Verlag 1988.

HALL, G. – WATT, J.M. (eds.): Modern Numerical Methods for Ordinary Differential Equations. Oxford, Clarendon-Press 1976. (Překlad do ruštiny: Moskva, Mir 1979.)

JACOBS, D.A.H. (ed.): The State of the Art in Numerical Analysis. London–New York–San Francisco, Academic Press 1977.

JANENKO, N.N.: Metod drobnych šagov. Novosibirsk, Iz. NGU 1966. (Překlad do angličitny: New York–Geidelberg–Berlin, Springer–Verlag 1971.)

JOHNSON, C.: Numerical Solutions of Partial Differential Equations by the Finite Element Method. Cambridge, Cambridge University Press 1988.

LADYŽENSKAJA, O.A. – SOLONNIKOV, V.A. – URALCEVA, N.N.: Linejnyje i kvazilinejnyje uravnenija paraboličeskoho tipa. Moskva, Nauka 1967.

LEES, M.: Apriori Estimates for the Solution of Difference Approximations to Parabolic Equations. Duke Math. J. *27*, 1960, s. 297 – 311.

LIONS, J.L.: Equations differentielles operationelles et problèmes aux limites. New York–Heidelberg–Berlin, Springer–Verlag 1961.

LIONS, J.L. – MAGENES, E.: Problèmes aux limites non homogènes et applications. Vol. 2. Paris, Dunod 1968.

MARČUK, G.I.: Metody vyčislitel'noj matematiki. 2. vyd. Moskva, Nauka 1980. (Překlad do češtiny: Praha, Academia 1987.)

MITCHELL, A.R.: Computational Methods in Partial Differential Equations. London–New York–Sydney–Toronto, J. Wiley and Sons 1969.

MITCHELL, A.R. – WAIT, R.: The Finite Element Method in Partial Differential Equations. Chichester–New York–Brisbane–Toronto, J. Wiley and Sons 1977. (Překlad do ruštiny: Moskva, Mir 1981.)

PEACEMAN, D.W. – RACHFORD, H.H.: The Numerical Solution of Parabolic and Elliptic Differential Equations. J. Soc. Indust. Appl. Mat. *3*, 1955, s. 28 – 41.

PETROVSKIJ, I.G.: Lekcii ob uravnenijach s častnymi proizvodnymi. Moskva–Leningrad, Gostechizdat 1950. (Překlad do češtiny: Praha, Přírodovědecké vydavatelství 1952.)

REKTORYS, K.: The Method of Discretization in Time and Partial Differential Equations. Dordrecht, J. Reidel 1982. (České vydání: Praha, SNTL 1985.)

RICHTMYER, R.D.: Difference Methods for Initial Value Problems. New York–London, Interscience Publishers 1957. (Překlad do ruštiny: Moskva, IL 1960.)
RYCHTMYER, R.D. – MORTON, K.N.: Difference Methods for Initial Value Problems. New York–London–Sydney, J. Wiley and Sons 1967.
SAMARSKIJ, A.A.: Vvedenije v teoriju raznostnych schem. Moskva, Nauka 1970.
SAULJEV, V.K.: Integrirovanije uravnenij paraboličeskoho tipa metodom setok. Moskva, Fizmatgiz 1960.
STRANG, G. – FIX, K.J.: An Analysis of the Finite Element Method. Englewood Cliffs, N.J., Prentice–Hall 1973. (Překlad do ruštiny: Moskva, Mir 1977.)
THOMÉE, V.: Galerkin Finite Element Methods for Parabolic Problems. Berlin–New York–Heidelberg, Springer–Verlag 1984.

REJSTŘÍK