UČEBNÍ TEXTY VYSOKÝCH ŠKOL

Vysoké učení technické v Brně Fakulta strojního inženýrství

Doc. RNDr. Libor Čermák, CSc.

Numerické metody

pro řešení diferenciálních rovnic

Ústav matematiky FSI VUT v Brně

Obsah

Předmluva

Tato skripta jsou určena pro studium předmětu *Numerické metody II*. Výklad navazuje na úvodní kurz *Numerické metody* a proto se předpokládá, že čtenář má základní znalosti o numerických metodách lineární algebry, řešení nelineárních rovnic, interpolaci, numerickém derivování a integrování, viz např. [4].

Skripta uvádějí celou řadu algoritmů a doprovodný text [11] k tomu přidává příklady a cvičení. Pro implementaci algoritmů a experimentování s nimi se výtečně hodí prostředí MATLABu.

První kapitola se věnuje numerickému řešení počátečních úloh pro obyčejné diferenciální rovnice. Dílčí témata jsou tradiční, tj. Rungovy-Kuttovy metody, Adamsovy metody a metody zpětného derivování.

Ve druhé kapitole jsou uvedeny čtyři základní metody řešení okrajových úloh pro obyčejné diferenciální rovnice druhého řádu, a sice metoda střelby, diferenční metoda, metoda konečných objemů a metoda konečných prvků.

Třetí kapitola je věnována numerickým metodám řešení parciálních diferenciálních rovnic. Pro eliptickou parciální diferenciální rovnici ve dvou prostorových proměnných je uvedena diskretizace diferenční metodou, metodou konečných objemů a metodou konečných prvků. Řešení parabolické a hyperbolické parciální diferenciální rovnice druhého řádu v jedné prostorové proměnné se provádí metodou přímek. Pro hyperbolickou rovnici prvního řádu v jedné prostorové proměnné je kromě metody přímek použita také metoda charakteristik.

V rámci klasických tematických okruhů jsem se snažil do skript zařadit takové numerické metody, které se v současnosti skutečně používají. Vycházel jsem přitom z osvědčených učebnic numerické matematiky, jakými jsou např. knihy [10], [18], [20], [28], a z vynikajících monografií, mezi nimi zejména [13], [23], [24], [15], [9], [2], [30]. Pokud jde o české zdroje, nejvíce podnětů jsem čerpal z knih [28], [29] a ze skript [17], [3] a [19].

Brno, květen 2015 Libor Čermák

1. Obyčejné diferenciální rovnice: počáteční úlohy

V této kapitole se budeme zabývat problematikou numerického řešení počátečních úloh pro obyčejné diferenciální rovnice.

1.1. Formulace, základní pojmy

Počáteční problém pro ODR1 spočívá v určení funkce $y(t)$, která vyhovuje diferenciální rovnici

$$y'(t) = f(t, y(t)) \tag{1.1}$$

a splňuje počáteční podmínku

$$y(a) = \eta. \tag{1.2}$$

Je-li v nějakém okolí D bodu $[a, \eta]$ funkce $f(t, y)$ spojitá a splňuje-li v tomto okolí *Lipschitzovu podmínku* s konstantou L vzhledem k proměnné y, tj. platí-li

$$|f(t, u) - f(t, v)| \leq L|u - v| \quad \forall\, [t, u]\,, [t, v] \in D\,, \tag{1.3}$$

pak bodem $[a, \eta]$ prochází jediné řešení $y(t)$ rovnice (1.1). Jestliže funkce f má v D omezenou parciální derivaci vzhledem k proměnné y, tj. když $|\partial_y f(t, y)| \leq L$, Lipschitzova podmínka (1.3) platí. Jiný standardní výsledek říká, že když je funkce f spojitá na $\langle a, b\rangle \times \mathbb{R}$ a splňuje tam Lipschitzovu podmínku, pak počáteční problém (1.1), (1.2) má jediné řešení definované v celém intervalu $\langle a, b\rangle$.

Počáteční problém pro soustavu ODR1 znamená určit funkce $y_1(t), \ldots, y_d(t)$ splňující diferenciální rovnice

$$y_j'(t) = f_j(t, y_1(t), \ldots, y_d(t)), \quad j = 1, 2, \ldots, d,$$

a počáteční podmínky

$$y_j(a) = \eta_j, \quad j = 1, 2, \ldots, d.$$

Stručnější vektorový zápis je

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \quad \mathbf{y}(a) = \boldsymbol{\eta}\,, \tag{1.4}$$

kde

$$\mathbf{y}(t) = (y_1(t), y_2(t), \ldots, y_d(t))^T, \qquad \mathbf{y}'(t) = (y_1'(t), y_2'(t), \ldots, y_d'(t))^T,$$

$$\mathbf{f}(t, \mathbf{y}(t)) = (f_1(t, \mathbf{y}(t)), f_2(t, \mathbf{y}(t)), \ldots, f_d(t, \mathbf{y}(t)))^T, \qquad \boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_d)^T.$$

Je-li v okolí D bodu $[a, \boldsymbol{\eta}]$ funkce $\mathbf{f}(t, \mathbf{y})$ spojitá a splňuje tam vzhledem k proměnné $\mathbf{y}$ Lipschitzovu podmínku s konstantou L , tj. platí-li

$$\|\mathbf{f}(t, \mathbf{u}) - \mathbf{f}(t, \mathbf{v})\| \leq L\|\mathbf{u} - \mathbf{v}\| \quad \forall\, [t, \mathbf{u}]\,, [t, \mathbf{v}] \in D\,, \tag{1.5}$$

pak bodem $[a, \boldsymbol{\eta}]$ prochází jediné řešení počáteční úlohy (1.4). Má-li $\mathbf{f}$ v D omezené parciální derivace $\{\partial f_i(t, \mathbf{y})/\partial y_j\}_{i,j=1}^d$, pak Lipschitzova podmínka (1.5) platí. Je-li $D = \langle a, b\rangle \times \mathbb{R}^d$, jediné řešení existuje v celém intervalu $\langle a, b\rangle$.

Rovnice vyššího řádu. Počáteční problém pro obyčejnou diferenciální rovnici řádu d,

$$y^{(d)}(t) = F(t, y(t), y'(t), \ldots, y^{(d-1)}(t)) \tag{1.6}$$

s počátečními podmínkami

$$y(a) = \eta_1, \quad y'(a) = \eta_2, \ldots, y^{(d-1)}(a) = \eta_d\,,$$

lze snadno převést na počáteční problém (1.4) pro d rovnic řádu prvního:

$$\begin{aligned}
y_1'(t) &= y_2(t), & y_1(a) &= \eta_1\,,\\
y_2'(t) &= y_3(t), & y_2(a) &= \eta_2\,,\\
&\vdots & &\vdots\\
y_{d-1}'(t) &= y_d(t), & y_{d-1}(a) &= \eta_{d-1}\,,\\
y_d'(t) &= F(t, y_1(t), y_2(t), \ldots, y_d(t)), & y_d(a) &= \eta_d\,,
\end{aligned}$$

kde $y_1(t) = y(t)$, $y_2(t) = y'(t)$, …, $y_d(t) = y^{(d-1)}(t)$.

V dalším budeme vždy předpokládat, že uvažovaná počáteční úloha má v intervalu $\langle a, b\rangle$ jediné řešení. Budeme také předpokládat, že funkce $\mathbf{f}(t, \mathbf{y})$ má tolik spojitých derivací, kolik jich v dané situaci bude zapotřebí.

Jedna rovnice prvního řádu s jednou neznámou funkcí je v aplikacích méně významná než soustavy rovnic. Metody přibližného řešení se však snadněji odvodí pro jednu rovnici a lze je aplikovat bezprostředně i na soustavy. Také analýza numerických metod je pro jednu rovnici podstatně snadnější. Proto se v následujícím výkladu převážně omezíme jen na jednu rovnici. Z velkého množství metod uvedeme ty, které jsou pro své dobré vlastnosti široce používány. Mezi ně bezesporu patří metody implementované do Matlabu a právě na ně se v tomto textu zaměříme.

Numerickým řešením počáteční úlohy rozumíme výpočet přibližných hodnot hledaného řešení $y(t)$ v bodech t_n dosti hustě vykrývajících interval $\langle a, b\rangle$. Nechť tedy

$$a = t_0 < t_1 < \cdots < t_Q = b$$

je *dělení* intervalu $\langle a, b\rangle$. Body t_n jsou *uzly*, vzdálenost $\tau_n = t_{n+1} - t_n$ dvou sousedních uzlů je *délka kroku* Jsou-li všechny kroky stejně dlouhé, tj. když $\tau_n = \tau = (b-a)/Q$, hovoříme o *rovnoměrném* (*ekvidistantním*) dělení intervalu $\langle a, b\rangle$. V tom případě je $t_n = a + n\tau$, $n = 0, 1, \ldots, Q$. Hodnotu přesného řešení v uzlu t_n budeme značit $y(t_n)$ a hodnotu přibližného řešení y_n. Jestliže se nám podaří najít přibližné řešení y_n, $n = 0, 1, \ldots, Q$, můžeme vypočítat přibližnou hodnotu řešení $y(t)$ v libovolném bodě $t \in \langle a, b\rangle$ interpolací.

Numerická metoda pro řešení počáteční úlohy (1.1), (1.2) je předpis pro postupný výpočet aproximací $y_1, y_2, \ldots, y_Q$, $y_0 = \eta$ z počáteční podmínky. Metoda se nazývá

k-kroková, závisí-li předpis pro výpočet aproximace y_{n+1} na předchozích aproximacích $y_n, y_{n-1}, \ldots, y_{n-k+1}$. Speciálně *jednokroková metoda* počítá přibližné řešení y_{n+1} v uzlu t_{n+1} jen pomocí znalosti přibližného řešení y_n v uzlu t_n, přibližná řešení y_{n-1}, $y_{n-2}, \ldots$ spočtená v předchozích uzlech t_{n-1}, $t_{n-2}, \ldots$ nepoužívá. Výpočet y_{n+1} nazýváme krokem metody od t_n do t_{n+1} (stručně *krokem*). Při popisu kroku budeme u délky kroku $\tau_n = t_{n+1} - t_n$ vypouštět index, tj. píšeme $\tau_n = \tau$.

1.2. Eulerovy metody

Explicitní Eulerova metoda. Nejjednodušší numerickou metodou pro řešení úlohy (1.1), (1.2) je *explicitní Eulerova metoda* (stručně EE metoda). EE metodu snadno odvodíme z Taylorovy formule

$$y(t_{n+1}) = y(t_n + \tau) = y(t_n) + \tau y'(t_n) + \tfrac{1}{2}\tau^2 y''(\xi_n), \quad \xi_n \in (t_n, t_{n+1}). \tag{1.7}$$

Uvážíme-li, že $y'(t_n) = f(t_n, y(t_n))$ a zanedbáme-li člen $\frac{1}{2}\tau^2 y''(\xi_n)$, dostaneme

$$y(t_{n+1}) \approx y(t_n) + \tau f(t_n, y(t_n)).$$

Výrazy $y(t_n)$ a $y(t_{n+1})$ nahradíme jejich přibližnými hodnotami y_n a y_{n+1}, znaménko přibližné rovnosti $\approx$ nahradíme znaménkem rovnosti a obdržíme předpis EE metody

$$y_{n+1} = y_n + \tau f(t_n, y_n)\,. \tag{1.8}$$

O explicitní metodě hovoříme proto, že pro určení y_{n+1} máme explicitní vzorec: dosazením známé hodnoty y_n do pravé strany rovnice (1.8) obdržíme hledanou hodnotu y_{n+1}. V anglicky psané literatuře se EE metoda označuje jako *Euler method* resp. *explicit Euler method* resp. *forward Euler method*.

Diskretizační chyby. Přesnost numerické metody měříme pomocí tzv. *lokální diskretizační chyby* (anglicky *local truncation error*)

$$\mathrm{lte}_n = y(t_{n+1}) - y(t_n) - \tau f(t_n, y(t_n))\,.$$

Lokální diskretizační chyba je tedy chyba, které se dopustíme v jednom kroku metody za tzv. *lokalizačního předpokladu*, že $y_n = y(t_n)$ je přesné řešení počáteční úlohy (1.1), (1.2). Z (1.7) pro EE metodu plyne

$$\mathrm{lte}_n = \tfrac{1}{2}\tau^2 y''(\xi_n) \;\text{ a tedy }\; |\mathrm{lte}_n| \le C\tau^2, \;\text{ kde }\; C = \tfrac{1}{2}\max_{t_n \le t \le t_{n+1}} |y''(t)|\,,$$

což lze stručně vyjádřit zápisem $\mathrm{lte}_n = O(\tau^2)$. Následující poznámka připomíná význam Landauova symbolu $O(\tau^p)$.

Poznámka. (*O Landauově symbolu* $O(\varphi(h))$). Nechť $\varphi(\tau)$ je funkce definovaná v intervalu $(0, \tau_0\rangle$ a p je libovolné číslo. Řekneme, že funkce $\varphi(\tau)$ je řádu $O(\tau^p)$ a píšeme $\varphi(\tau) = O(\tau^p)$, jestliže existuje kladné číslo C takové, že pro všechna $0 < \tau \le \tau_0$ platí $|\varphi(\tau)| \le C\tau^p$. □

Lokální diskretizační chyba při reálném výpočtu nevzniká, neboť obecně není splněn lokalizační předpoklad, tj. $y_n \neq y(t_n)$. Lokální diskretizační chyba se uplatní jen při

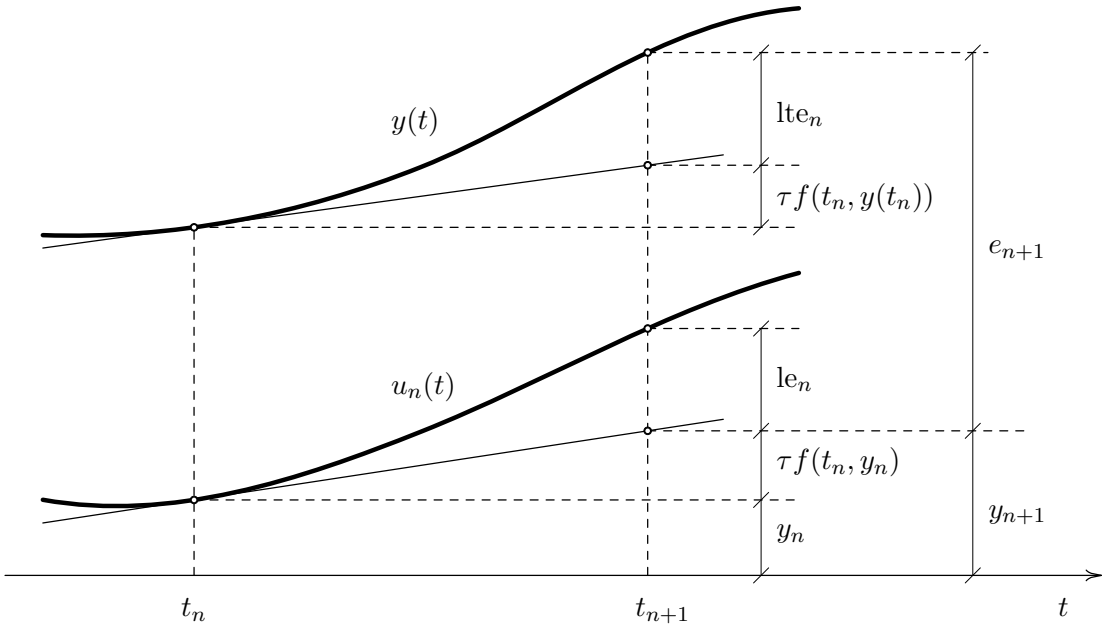
analýze vlastností numerické metody, například konvergence $y_n \to y(t_n)$. Pro praktické účely, například pro řízení délky kroku, je nutné pracovat s tzv. *lokální chybou* (anglicky local error) definovanou předpisem

$$\mathrm{le}_n = u_n(t_{n+1}) - y_{n+1},$$

kde $u_n(t)$ je tzv. *lokální řešení* počátečního problému

$$u_n'(t) = f(t, u_n(t))\,, \qquad u_n(t_n) = y_n\,.$$

Lokální chyba le_n je tedy chyba, které se skutečně dopustíme při reálném výpočtu v kroku od t_n do t_{n+1}. Dá se ukázat, že pro výpočet s dostatečně malými délkami kroků je rozdíl mezi oběma lokálními chybami prakticky zanedbatelný.



Obr. 8.1. Diskretizační chyby

Hromaděním lokálních chyb vzniká *globální diskretizační chyba*

$$e_n = y(t_n) - y_n\,.$$

V případě rovnoměrného dělení lze dokázat, že

$$|e_n| = |y(t_n) - y_n| \le C\tau, \qquad n = 0, 1, \ldots, Q, \tag{1.9}$$

kde C je konstanta nezávislá na $\tau = (b-a)/Q$. Tuto skutečnost stručně vyjádříme tvrzením, že *globální diskretizační chyba EE metody je řádu* $O(\tau)$. Říkáme také, že *EE metoda je řádu 1.* Protože $e_n \to 0$ pro $Q \to \infty$, *numerické řešení* získané EE metodou *konverguje* k řešení přesnému. Říkáme také, že *rychlost (řád) konvergence EE metody je rovna 1.* Lokální chyby lte_n, le_n a globální chyba e_{n+1} jsou zakresleny v obrázku 1.1.

Tvrzení (1.9) lze snadno ověřit v případě, že $f(t,y) = f(t)$ nezávisí na y. Pak totiž

$$y(t_{k+1}) = y(t_k) + \tau f(t_k) + \tfrac{1}{2}\tau^2 f'(\xi_k)\,,$$
$$y_{k+1} = y_k + \tau f(t_k)\,.$$

Odečtením druhé rovnice od prvé dostaneme $e_{k+1} = e_k + \frac{1}{2}\tau^2 f'(\xi_k)$, a protože $e_0 = 0$, je

$$e_n = \tfrac{1}{2}\tau^2[f'(\xi_0) + f'(\xi_1) + \cdots + f'(\xi_{n-1})]\,.$$

Označíme-li $M = \max_{a\le\xi\le b}|f'(\xi)|$, pak

$$|e_n| \le \tfrac{1}{2}\tau^2 nM \le \tfrac{1}{2}[\tau Q]M\tau = \tfrac{1}{2}(b-a)M\tau, \quad \text{neboť } \tau Q = b-a.$$

Zaokrouhlovací chyby. Dopustíme-li se v kroku od t_n do t_{n+1} zaokrouhlovací chyby, jejíž velikost $|\Delta y_{n+1}| = |\tilde{y}_{n+1} - y_{n+1}|$ nepřesáhne ε, pak lze dokázat, že po Q krocích délky τ velikost zaokrouhlovací chyby nepřesáhne $K\varepsilon\tau^{-1}$, kde K je konstanta nezávislá na ε a τ. Pro celkovou chybu EE metody pak platí

$$\max_{0\le n\le Q}|y(t_n) - \tilde{y}_n| \le C\tau + \varepsilon K\tau^{-1}\,,$$

kde C, K jsou konstanty nezávislé na τ a ε. Protože konstanta ε je malá, vliv zaokrouhlování se projeví až pro extrémně velký počet kroků Q (tj. pro velmi malé τ). Tato situace při řešení běžných úloh nenastává a tudíž vliv zaokrouhlovacích chyb bývá nepodstatný.

Stabilita. Řekneme, že počáteční problém (1.1), (1.2) je stabilní vzhledem k počáteční podmínce, jestliže malá změna počáteční hodnoty η vyvolá malou změnu řešení. Elementárním příkladem takového problému je *testovací úloha*

$$y' = \lambda y\,, \quad y(0) = 1\,, \tag{1.10}$$

kde $\lambda = \alpha + i\beta$ je komplexní číslo se zápornou reálnou složkou, tj. $\mathrm{Re}(\lambda) = \alpha < 0$. Skutečně, jestliže

$$\begin{aligned} u'(t) &= \lambda u(t)\,, \qquad u(0) = 1 \qquad \Longrightarrow \quad u(t) = e^{\lambda t} \\ v'(t) &= \lambda v(t)\,, \qquad v(0) = 1+\delta \quad \Longrightarrow \quad v(t) = (1+\delta)e^{\lambda t}, \end{aligned}$$

pak

$$|u(t) - v(t)| \le |\delta| \cdot |e^{(\alpha+i\beta)t}| = |\delta|e^{\alpha t} \cdot |\cos\beta t + i\sin\beta t| = |\delta|e^{\alpha t} \le |\delta|\,.$$

Pro řešení $y(t) = e^{\lambda t}$ testovací úlohy (1.10) rovněž platí

$$|y(t)| = |e^{\lambda t}| = e^{\alpha t}|\cos\beta t + i\sin\beta t| \to 0 \quad \text{pro} \quad t\to\infty\,.$$

Je proto přirozené požadovat, aby na rovnoměrném dělení $t_n = n\tau$, $n = 0, 1, \ldots$, numerické řešení y_n testovací úlohy (1.10) splňovalo analogickou relaci, tzv. *podmínku stability*

$$y_n \to 0 \quad \text{pro} \quad n\to\infty. \tag{1.11}$$

Aplikujeme-li EE metodu na testovací rovnici (1.10), dostaneme

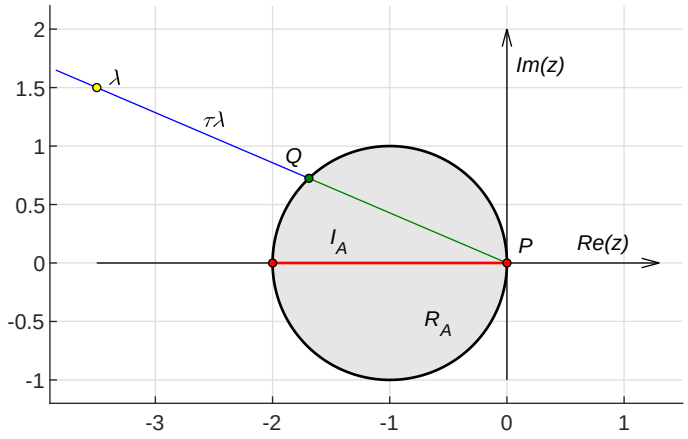
$$y_{n+1} = y_n + \tau\lambda y_n = (1+\tau\lambda)y_n = (1+\tau\lambda)^2 y_{n-1} = \cdots = (1+\tau\lambda)^{n+1} y_0\,.$$

Podmínka stability (1.11) bude splněna, právě když $|1+\tau\lambda| < 1$, neboli když $\tau\lambda$ leží v tzv. *oblasti absolutní stability* R_A:

$$\tau\lambda \in R_A = \{z \in \mathbb{C}\,|\,|z+1| < 1\}.$$

Oblast absolutní stability EE metody je tedy vnitřek jednotkového kruhu $|z+1| < 1$ komplexní roviny $\mathbb{C}$ se středem v bodě $[-1, 0]$. Průnik oblasti absolutní stability se zápornou částí reálné osy je *interval absolutní stability* I_A. Pro EE metodu $I_A = (-2, 0)$. Pro reálné $\lambda < 0$ podmínka stability (1.11) vyžaduje volit krok $\tau < 2/|\lambda|$. Z ilustračního obrázku 8.2 vyčteme, že pro zvolené λ musíme τ zvolit tak, aby $\tau\lambda$ byl vnitřní bod úsečky $\overline{PQ}$.

Tvar a velikost oblasti absolutní stability metody je spolu s řádem metody základní charakteristikou kvality numerické metody. EE metoda z tohoto pohledu příliš kvalitní není: je pouze řádu 1 a oblast její absolutní stability je malá. EE metoda se používá jen výjimečně.



Obr. 8.2. Oblast absolutní stability EE metody

Lineární stabilita. Testovací úloha (1.10) je dobrým modelem pro posouzení tzv. *lineární stability* obecné počáteční úlohy $y'(t) = f(t, y(t))$, $y(t_\alpha) = y_\alpha$. Když v Taylorově rozvoji okolo bodu (t_α, y_α),

$$f(t, y(t)) = f(t_\alpha, y_\alpha) + \frac{\partial f(t_\alpha, y_\alpha)}{\partial t}(t - t_\alpha) + \frac{\partial f(t_\alpha, y_\alpha)}{\partial y}(y(t) - y_\alpha) + O((t - t_\alpha)^2),$$

zanedbáme chybový člen, dostaneme aproximující lineární problém

$$y'(t) = \lambda_\alpha y(t) + g_\alpha(t), \quad y(t_\alpha) = y_\alpha, \tag{1.10'}$$

kde $\lambda_\alpha = \partial_y f(t_\alpha, y_\alpha)$, $g_\alpha(t) = \partial_t f(t_\alpha, y_\alpha)(t - t_\alpha) - \partial_y f(t_\alpha, y_\alpha) y_\alpha$. Jestliže

$$u' = \lambda_\alpha u + g_\alpha(t), \;\; u(t_\alpha) = y_\alpha, \qquad v' = \lambda_\alpha v + g_\alpha(t), \;\; v(t_\alpha) = y_\alpha + \delta_a,$$

pak $|u(t) - v(t)| = |\delta_\alpha| e^{\lambda_\alpha (t - t_\alpha)} \to 0$ pro $t \to \infty$, právě když $\mathrm{Re}(\lambda_\alpha) < 0$. V testovací úloze (1.10) si tedy pod λ můžeme představit hodnotu $\partial_y f(t_\alpha, y_\alpha)$ v nějakém bodě (t_α, y_α).

Implicitní Eulerova metoda. Vyjdeme opět z Taylorova rozvoje

$$y(t_n) = y(t_{n+1} - \tau) = y(t_{n+1}) - \tau y'(t_{n+1}) + \tfrac{1}{2}\tau^2 y''(\xi_n), \quad \xi_n \in (t_n, y_{n+1})\,. \tag{1.11}$$

Vypuštěním členu $\frac{1}{2} y''(\xi_n)$ a užitím rovnosti $y'(t_{n+1}) = f(t_{n+1}, y(t_{n+1}))$ obdržíme *implicitní Eulerovu metodu* (stručně IE metodu) jako předpis

$$y_{n+1} = y_n + \tau f(t_{n+1}, y_{n+1})\,. \tag{1.12}$$

V anglicky psané literatuře se IE metoda označuje jako *implicit Euler method* resp. *backward Euler method*. O implicitní metodě mluvíme proto, že neznámá y_{n+1} je rovnicí (1.12) určena implicitně. Pro dostatečně malé τ má rovnice (1.12) jediné řešení y_{n+1}, dokažte! Určit y_{n+1} znamená řešit obecně nelineární rovnici. To je ve srovnání s EE metodou problém navíc. Aby mělo použití IE metody nějaký smysl, musí mít IE metoda oproti EE metodě také nějakou přednost. Pokusme se ji najít.

Nejdříve prozkoumáme přesnost IE metody. Lokální diskretizační chyba IE metody je definována rovnicí

$$y(t_{n+1}) = y(t_n) + \tau f(t_{n+1}, y(t_{n+1})) + \mathrm{lte}_n,$$

kde $y(t)$ je řešení úlohy (1.1), (1.2). Z (1.11) plyne

$$\mathrm{lte}_n = -\tfrac{1}{2}\tau^2 y''(\xi_n) = O(\tau^2)\,.$$

IE metoda je tedy řádu 1 stejně jako EE metoda. Při podrobnějším zkoumání lze zjistit, že

$$\mathrm{lte}_n = \begin{cases} \;\;\tfrac{1}{2} y''(t_n)\tau^2 + O(\tau^3) & \text{pro EE metodu,} \\ -\tfrac{1}{2} y''(t_n)\tau^2 + O(\tau^3) & \text{pro IE metodu.} \end{cases}$$

Hlavní členy lokálních diskretizačních chyb (v případě Eulerových metod to jsou členy obsahující τ^2) jsou co do absolutní hodnoty stejné, liší se jen znaménkem. Můžeme proto oprávněně soudit, že obě metody jsou stejně přesné.

Podívejme se také na stabilitu IE metody. Pro testovací úlohu (1.10) je

$$y_{n+1} = y_n + \tau\lambda y_{n+1}\,, \quad \text{odtud} \quad y_{n+1} = \frac{1}{1 - \tau\lambda} y_n = \cdots = \left(\frac{1}{1 - \tau\lambda}\right)^{n+1} y_0$$

a tedy podmínka stability (1.11) platí, právě když $|1 - \tau\lambda| > 1$, neboli když $\tau\lambda$ leží v oblasti absolutní stability R_A:

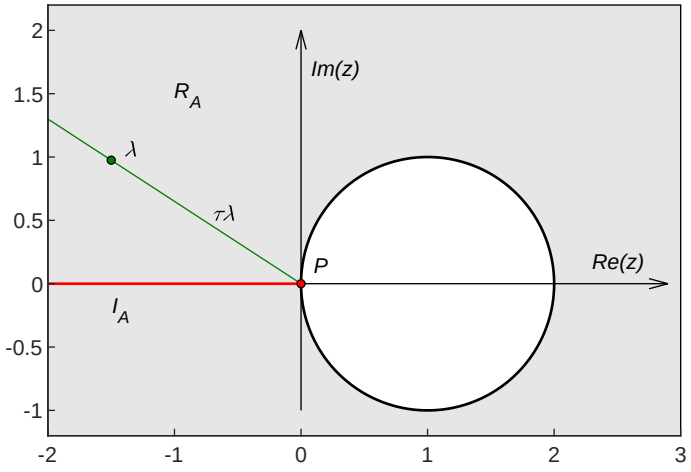
$$\tau\lambda \in R_A = \{z \in \mathbb{C} \,|\, |z - 1| > 1\}.$$

Oblast absolutní stability IE metody je tedy obrovská, je to celý vnějšek $|z-1|>1$ jednotkového kruhu komplexní roviny $\mathbb{C}$ se středem v bodě $[1,0]$. Interval I_A absolutní stability IE metody je $I_A=(-\infty,0)$. Podmínka stability (1.11) délku kroku IE metody zřejmě nijak neomezuje, viz obrázek 8.3.

Je to právě mimořádná stabilita, která je onou hledanou předností IE metody ve srovnání s EE metodou. Tento klad je však třeba vykoupit nutností řešit obecně nelineární rovnici. y_{n+1} získáme jako přibližné řešení rovnice $g(z)=0$, kde

$$g(z)=z-y_n-\tau f(t_{n+1},z)\,.$$

Protože dobrou počáteční aproximaci lze získat extrapolací z hodnot $y_n, y_{n-1},\ldots$, dá se očekávat rychlá konvergence Newtonovy metody (pro řešení nelineárních rovnic). Praktická zkušenost potvrzuje, že tomu tak skutečně je.



Obr. 8.3. Oblast absolutní stability IE metody

Lichoběžníkovou metodu dostaneme jako aritmetický průměr EE metody a IE metody:

$$y_{n+1}=y_n+\tfrac{1}{2}\tau[f(t_n,y_n)+f(t_{n+1},y_{n+1})]\,. \tag{1.13}$$

Pro lokální diskretizační chybu lte_n, definovanou rovnicí

$$y(t_{n+1})=y(t_n)+\tfrac{1}{2}\tau[f(t_n,y(t_n))+f(t_{n+1},y(t_{n+1}))]+\mathrm{lte}_n,$$

užitím Taylorovy věty odvodíme

$$\mathrm{lte}_n=-\tfrac{1}{12}\tau^3y'''(t_n)+O(\tau^4)\,.$$

Lichoběžníková metoda (stručně TR metoda podle anglického *trapezoidal rule*) je tedy implicitní metoda řádu 2. Stabilitu TR metody zjistíme řešením testovací úlohy (1.10):

$$y_{n+1}=y_n+\tfrac{1}{2}\tau\lambda[y_n+y_{n+1}],\quad\text{odtud}\quad y_{n+1}=\frac{2+\tau\lambda}{2-\tau\lambda}y_n=\cdots=\left[\frac{2+\tau\lambda}{2-\tau\lambda}\right]^{n+1}y_0\,.$$

Není těžké ověřit, že podmínka stability (1.11) platí, právě když

$$\tau\lambda \in R_A = \{z \in \mathbb{C} \,|\, \mathrm{Re}(z) < 0\}.$$

Oblast absolutní stability TR metody tedy obsahuje celou zápornou polorovinu komplexní roviny $\mathbb{C}$, interval absolutní stability $I_A = (-\infty, 0)$. TR metoda (s podporou IE metody) je v Matlabu implementována jako program `ode23t`.

1.3. Explicitní Rungovy-Kuttovy metody

Obecný tvar *s-stupňové explicitní Rungovy-Kuttovy metody* je

$$y_{n+1} = y_n + \tau(b_1 k_1 + b_2 k_2 + \cdots + b_s k_s)\,, \tag{1.14}$$

kde koeficienty k_i, $i = 1, 2, \ldots, s$, jsou určeny předpisem

$$\begin{aligned}
k_1 &= f(t_n, y_n),\\
k_2 &= f(t_n + \tau c_2, y_n + \tau a_{21} k_1),\\
k_3 &= f(t_n + \tau c_3, y_n + \tau(a_{31} k_1 + a_{32} k_2)),\\
&\vdots\\
k_s &= f(t_n + \tau c_s, y_n + \tau(a_{s1} k_1 + a_{s2} k_2 + \cdots + a_{s,s-1} k_{s-1})),
\end{aligned} \tag{1.15}$$

a kde b_i, c_i, a_{ij} jsou konstanty definující konkrétní metodu. Rungova-Kuttova metoda (1.14), (1.15) je explicitní: nejdříve spočteme k_1, pak k_2 pomocí k_1, pak k_3 pomocí k_1, k_2 a tak dále, až nakonec spočteme k_s pomocí $k_1, k_2, \ldots, k_{s-1}$. Vypočtené koeficienty k_i, $i = 1, 2, \ldots, s$, dosadíme do (1.14) a dostaneme y_{n+1}.

V dalším budeme hovořit jen o Rungových-Kuttových metodách (stručně RK metodách), tj. slůvko „explicitní“ vynecháme. Je však třeba připomenout, že existují také implicitní Rungovy-Kuttovy metody, těmi se však zabývat nebudeme.

RK metody jsou zřejmě jednokrokové: k výpočtu y_{n+1} potřebujeme znát jen y_n, předchozí hodnoty $y_{n-1}, y_{n-2}, \ldots$ v kroku od t_n do t_{n+1} nepoužijeme.

Koeficient k_i je směrnicí lokálního řešení procházejícího bodem $[t_i^*, y_i^*]$, kde

$$[t_1^*, y_1^*] = [t_n, y_n],\;\; t_i^* = t_n + \tau c_i,\;\; y_i^* = y_n + \tau(a_{i1} k_1 + a_{i2} k_2 + \cdots + a_{i,i-1} k_{i-1}),\;\; i = 2, 3, \ldots, s.$$

Do bodu $[t_{n+1}, y_{n+1}]$ se tedy dostaneme z bodu $[t_n, y_n]$ tak, že se posuneme po přímce, jejíž směrnice $k^* = b_1 k_1 + b_2 k_2 + \cdots + b_s k_s$ je váženým průměrem směrnic $k_1, k_2, \ldots, k_s$ (podle (1.16) pro všechny prakticky používané metody řádu alespoň jedna platí $\sum_{i=1}^{s} b_i = 1$).

Abychom dostali konkrétní metodu, musíme určit stupeň s a dále konstanty c_i, b_i, a_{ij}. Konstanty RK metod je zvykem zapisovat do tabulky známé jako *Butcherova tabulka*:

c_2	a_{21}				
c_3	a_{31}	a_{32}			
$\vdots$	$\vdots$				
c_s	a_{s1}	a_{s2}	$\ldots$	$a_{s,s-1}$	
	b_1	b_2	$\ldots$	b_{s-1}	b_s

Jedním z kritérií při volbě konstant RK metody je dosažení dostatečné přesnosti. Tu měříme pomocí lokální diskretizační chyby

$$\mathrm{lte}_n = y(t_{n+1}) - \left[y(t_n) + \tau \sum_{i=1}^{s} b_i k_i(t_n, y(t_n)) \right],$$

kde $k_1(t_n, y(t_n)) = f(t_n, y(t_n))$,

$$k_i(t_n, y(t_n)) = f(t_n + \tau c_i, y(t_n) + \tau \sum_{j=1}^{i-1} a_{ij} k_j(t_n, y(t_n))), \quad i = 2, 3, \ldots, s.$$

Lokální diskretizační chyba je chyba, které se dopustíme v jednom kroku za *lokalizačního předpokladu* $y_n = y(t_n)$. RK metoda je řádu p, pokud lokální diskretizační chyba je řádu $O(\tau^{p+1})$. Pro $p = 1, 2, 3$ lze odvodit následující tzv. *podmínky řádu*:

$$\begin{aligned}
&\text{řád 1:} \quad \sum_{i=1}^{s} b_i = 1\,, \\
&\text{řád 2:} \quad \sum_{i=1}^{s} b_i = 1\,, \quad \sum_{i=2}^{s} b_i c_i = \tfrac{1}{2}\,, \\
&\text{řád 3:} \quad \sum_{i=1}^{s} b_i = 1\,, \quad \sum_{i=2}^{s} b_i c_i = \tfrac{1}{2}\,, \quad \sum_{i=2}^{s} b_i c_i^2 = \tfrac{1}{3}\,, \quad \sum_{i=2}^{s} \sum_{j=2}^{i-1} b_i a_{ij} c_j = \tfrac{1}{6}\,.
\end{aligned} \tag{1.16}$$

Odvození podmínek řádu pro $p = 1, 2, 3, 4, 5$ lze najít třeba v [23]. Protože všechny prakticky používané metody splňují podmínku

$$c_i = a_{i1} + a_{i2} + \cdots + a_{i,i-1}\,, \quad i = 2, 3, \ldots, s\,, \tag{1.17}$$

budeme i my předpokládat, že podmínka (1.17) platí.

RK metoda řádu $p \geq 1$ má globální diskretizační chybu řádu $O(\tau^p)$. Předpokladem pro platnost tohoto tvrzení je dostatečná hladkost pravé strany f, konkrétně je třeba, aby funkce $f(t, y)$ měla spojité derivace až do řádu p včetně. Pokud f má spojité derivace jen do řádu $s \leq p$, pak lze pro globální chybu dokázat pouze řád $O(\tau^s)$, viz [23].

Označme $p(s)$ maximální dosažitelný řád s-stupňové RK metody. Platí

$$\begin{aligned}
&p(s) = s \quad \text{pro } s = 1, 2, 3, 4, && p(8) = 6, \\
&p(5) = 4, && p(9) = 7 \\
&p(6) = 5, && p(s) \leq s - 2 \quad \text{pro} \quad s = 10, 11, \ldots \\
&p(7) = 6, &&
\end{aligned}$$

Vidíme, že s-stupňové RK metody řádu s existují jen pro $1 \leq s \leq 4$. Například metoda řádu 5 je nejméně 6-ti stupňová. Uveďme si několik nejznámějších metod.

Metoda řádu 1. Pro $s = p = 1$ existuje jediná explicitní metoda a tou je nám již známá EE metoda $y_{n+1} = y_n + \tau f(t_n, y_n)$.

Metody řádu 2. Pro $s = p = 2$ má explicitní RK metoda Butcherovu tabulku

c_2	a_{21}	
	b_1	b_2

Podmínky (1.16) pro metodu řádu 2 stanoví

$$b_1 + b_2 = 1, \quad b_2 c_2 = \tfrac{1}{2}\,,$$

a protože ve shodě s (1.17) předpokládáme $a_{21} = c_2$, dostáváme tabulku

a	a	
	$1-b$	b

kde $ab = \frac{1}{2}$. Parametry a, b jsou tedy svázány jednou podmínkou, takže zvolíme-li $a \neq 0$, je $b = 1/(2a)$.

Pro $a = \frac{1}{2}$ je $b = 1$ a dostáváme metodu

$$y_{n+1} = y_n + \tau k_2, \quad \text{kde} \quad k_2 = f(t_n + \tfrac{1}{2}\tau, y_n + \tfrac{1}{2}\tau k_1),\ k_1 = f(t_n, y_n), \tag{1.18}$$

známou pod názvem *modifikovaná Eulerova metoda*. Budeme ji značit EM1 jako *první modifikace Eulerovy metody*. V anglicky psané literatuře je metoda (1.18) známa jako *midpoint Euler formula*.

Pro $a = 1$ je $b = \frac{1}{2}$ a dostáváme metodu

$$y_{n+1} = y_n + \tfrac{1}{2}\tau(k_1 + k_2), \quad \text{kde} \quad k_1 = f(t_n, y_n),\ k_2 = f(t_n + \tau, y_n + \tau k_1). \tag{1.19}$$

Budeme ji značit EM2 jako *druhou modifikaci Eulerovy metody*. Metoda (1.19) se také často uvádí pod názvem *Heunova metoda*.

Pro $a = \frac{2}{3}$ je $b = \frac{3}{4}$ a dostáváme metodu

$$y_{n+1} = y_n + \tfrac{1}{4}\tau(k_1 + 3k_2), \quad \text{kde} \quad k_1 = f(t_n, y_n),\ k_2 = f(t_n + \tfrac{2}{3}\tau, y_n + \tfrac{2}{3}\tau k_1),$$

známou jako *Ralstonova metoda řádu* 2 (stručně R2 metoda).

Metody řádu 3. Pro $s = p = 3$ dostáváme Butcherovu tabulku

c_2	c_2		
c_3	$c_3 - a_{32}$	a_{32}	
	b_1	b_2	b_3

a 4 podmínky pro metodu řádu 3:

$$b_1 + b_2 + b_3 = 1\,, \quad b_2 c_2 + b_3 c_3 = \tfrac{1}{2}\,,$$
$$b_2 c_2^2 + b_3 c_3^2 = \tfrac{1}{3}\,, \quad b_3 a_{32} c_2 = \tfrac{1}{6}\,.$$

Když zvolíme dva parametry $0 < c_2 < c_3$, jsou tím všechny koeficienty metody jednoznačně určeny. Volba $c_2 = \frac{1}{2}$, $c_3 = \frac{3}{4}$ vede na *Ralstonovu metodu řádu* 3 (stručně R3 metodu):

$\frac{1}{2}$	$\frac{1}{2}$		
$\frac{3}{4}$	0	$\frac{3}{4}$	
	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$

$$y_{n+1} = y_n + \tfrac{1}{9}\tau\,(2k_1 + 3k_2 + 4k_3)\,,$$
$$k_1 = f(t_n, y_n)\,, \quad k_2 = f(t_n + \tfrac{1}{2}\tau, y_n + \tfrac{1}{2}\tau k_1),$$
$$k_3 = f(t_n + \tfrac{3}{4}\tau, y_n + \tfrac{3}{4}\tau k_2).$$

Ralstonova metoda řádu 3 je základem *Runge-Kutta-Bogacki-Shampine* metody, viz [23], která je implementována do Matlabu jako funkce `ode23` a jejíž popis uvedeme v této kapitole později.

Metody řádu 4. Pro $s = p = 4$ je nejznámější *klasická Rungova-Kuttova metoda*

$$\begin{array}{c|cccc} \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

$$\begin{aligned} y_{n+1} &= y_n + \tfrac{1}{6}\tau\left(k_1 + 2k_2 + 2k_3 + k_4\right), \\ k_1 &= f(t_n, y_n)\,, \qquad\qquad k_2 = f(t_n + \tfrac{1}{2}\tau, y_n + \tfrac{1}{2}\tau k_1), \\ k_3 &= f(t_n + \tfrac{1}{2}\tau, y_n + \tfrac{1}{2}\tau k_2), \quad k_4 = f(t_n + \tau, y_n + \tau k_3). \end{aligned}$$

Klasická Rungova-Kuttova metoda (stručně cRK4) byla velmi populární v době, kdy se ještě nepoužívaly samočinné počítače a kdy proto velmi významným kritériem byla jednoduchost metody. Toto hledisko však v současné době ztratilo na významu a proto se používají jiné metody. Kvalitní dvojice metod řádu 4 a 5 jsou součástí metod *Runge-Kutta-Fehlberg* nebo *Runge-Kutta-Dormand-Prince*, viz např. [13]. Posledně jmenovaná dvojice metod je použita v matlabovské funkci `ode45`, popis uvedeme v této kapitole později.

Řízení délky kroku. V profesionálních programech uživatel zadá toleranci ε a program délku kroku vybírá tak, aby velikost odhadu est_n lokální chyby le_n nabývala pořád přibližně stejné hodnoty ε. Krok od y_n do y_{n+1} je úspěšný, když

$$|\mathrm{est}_n| \le \varepsilon\,. \tag{1.20}$$

Je-li podmínka (1.20) splněna, krok je úspěšný a pokračujeme výpočtem y_{n+2}. Pokud podmínka (1.20) splněna není, krok je neúspěšný a výpočet y_{n+1} opakujeme. Novou délku kroku τ^* určíme v případě úspěchu i neúspěchu stejným postupem, který si teď vysvětlíme.

Předpokládejme, že y_{n+1} počítáme metodou řádu p, takže

$$\mathrm{le}_n \doteq C\tau^{p+1} \doteq \mathrm{est}_n \quad \Longrightarrow \quad C \doteq \mathrm{est}_n/\tau^{p+1}\,.$$

Novou délku kroku τ^* zvolíme tak, aby velikost $|\mathrm{le}_n^*|$ lokální chyby $\mathrm{le}_n^* \doteq C(\tau^*)^{p+1}$ byla přibližně rovna zadané toleranci ε, tj.

$$|\mathrm{le}_n^*| \doteq |C|(\tau^*)^{p+1} \doteq |\mathrm{est}_n/\tau^{p+1}|(\tau^*)^{p+1} \doteq \varepsilon \quad \Longrightarrow \quad \tau^* \doteq \tau\left(\varepsilon/|\mathrm{est}_n|\right)^{1/(p+1)}\,.$$

Nová délka kroku τ^* se ještě redukuje pomocí parametru $\theta < 1$, takže

$$\tau^* = \theta\tau\left(\varepsilon/|\mathrm{est}_n|\right)^{1/(p+1)}\,. \tag{1.21}$$

V matlabovských programech `ode23` a `ode45` se bere $\theta = 0{,}8$. Současně se ještě uplatňují následující zásady.

1) Tolerance ε se uvažuje ve tvaru

$$\varepsilon = \max\{\varepsilon_r \max\{|y_n|, |y_{n+1}|\}, \varepsilon_a\},$$

kde ε_r je relativní tolerance a ε_a je tolerance absolutní. Matlabem přednastavené hodnoty jsou $\varepsilon_r = 10^{-3}$ a $\varepsilon_a = 10^{-6}$.

2) Označme τ_{min} resp. τ_{max} minimální resp. maximální povolenou délku kroku. Jestliže $\tau^* < \tau_{min}$, výpočet končí konstatováním, že danou diferenciální rovnici program neumí s požadovanou přesností vyřešit. Přitom $\tau_{min} = 16\varepsilon_m(t_n)$, kde $\varepsilon_m(t_n)$ je tzv. relativní přesnost aritmetiky pohyblivé řádové čárky, viz funkce `eps` v Matlabu. Jestliže $\tau^* > \tau_{max}$, položí se $\tau = \tau_{max}$, kde je $\tau_{max} = 0{,}1(b-a)$.

3) V případě neúspěšného kroku navrženou délku τ^* redukujeme:

$$\tau^* = \max(\tau^*, q_{min}\tau)\,,$$

kde $q_{min} = 0{,}5$ resp. 0,1 v `ode23` resp. `ode45` při prvním neúspěchu v rámci jednoho kroku a $\tau^* = 0{,}5\tau$ při opakovaném neúspěchu v témže kroku.

4) V případě úspěšného kroku délku kroku τ^* redukujeme předpisem

$$\tau^* = \min(\tau^*, q_{max}\tau),$$

kde $q_{max} = 5$.

5) V kroku bezprostředně následujícím po neúspěšném kroku se délka kroku nesmí zvětšit.

6) Počáteční délka kroku

$$\tau = \theta\varepsilon_r^{1/(p+1)} \max(|\eta|, \varepsilon_a/\varepsilon_r)/|f(a,\eta)|\,, \tag{1.22}$$

přičemž pro $\tau < \tau_{min}$ změníme τ na τ_{min} a pro $\tau > \tau_{max}$ změníme τ na τ_{max}.

7) Při programování je třeba postupovat opatrně. Příkazy programu je nutné uspořádat tak, aby nemohlo dojít k dělení nulou, viz (1.21) a (1.22).

Podrobnější informace týkající se řízení délky kroku lze najít v [23].

Odhad lokální chyby. Základní myšlenka je jednoduchá. Použijí se dvě metody, z nichž jedna je řádu p a druhá řádu $p+1$. Z výchozí hodnoty y_n spočteme y_{n+1}^{**} přesnější metodou a y_{n+1}^{*} méně přesnou metodou. Použitelný odhad lokální chyby je

$$\mathrm{est}_n = y_{n+1}^{**} - y_{n+1}^{*}.$$

Obě metody používají tutéž množinu koeficientů $\{k_j\}_{j=1}^{s}$. V tom případě je totiž získání odhadu „laciné“. Dvojice Rungových-Kuttových metod se popisují pomocí rozšířené Butcherovy tabulky,

c_2	a_{21}				
c_3	a_{31}	a_{32}			
$\vdots$	$\vdots$				
c_s	a_{s1}	a_{s2}	$\dots$	$a_{s,s-1}$	
	b_1^*	b_2^*	$\dots$	b_{s-1}^*	b_s^*
	b_1^{**}	b_2^{**}	$\dots$	b_{s-1}^{**}	b_s^{**}
	E_1	E_2	$\dots$	E_{s-1}	E_s

přičemž

$y_{n+1}^{*} = y_n + \tau\sum_{j=1}^{s} b_j^* k_j$ je metoda řádu p,

$y_{n+1}^{**} = y_n + \tau\sum_{j=1}^{s} b_j^{**} k_j$ je metoda řádu $p+1$,

$\mathrm{est}_n = \tau\sum_{j=1}^{s} E_j k_j$ je odhad lokální chyby,

takže $E_j = b_j^{**} - b_j^*$, $j = 1, 2, \dots, s$.

Pokud ve výpočtu pokračujeme přesnější metodou, tj. když $y_{n+1} = y^{**}_{n+1} = y^{*}_{n+1} + \mathrm{est}_n$, říkáme, že jsme použili dvojici metod *s lokální extrapolací*. Tento postup se v současných programech upřednostňuje. Druhou možností je pokračovat méně přesnou metodou, tj. položit $y_{n+1} = y^{*}_{n+1}$. V tom případě se přesnější metoda použije jen pro získání odhadu chyby a říkáme, že jsme dvojici metod použili *bez lokální extrapolace*. Obě níže uvedené metody BS32 a DP54 se používají jako metody s lokální extrapolací. Příkladem metody, která se obvykle používá bez lokální extrapolace, je Runge-Kutta-Fehlbergova metoda řádu 4, označovaná stručně jako RKF45, viz např. [13]. Na vysvětlenou k použitým zkratkám uveďme, že první číslo značí řád metody a druhé číslo řád pomocné metody použité pro odhad lokální chyby.

Bogacki–Shampine metoda, stručně BS32 metoda, je implementována v Matlabu jako funkce `ode23`. Rozšířená Butcherova tabulka BS32 metody je

$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{3}{4}$	0	$\frac{3}{4}$		
1	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	
	$\frac{7}{24}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{8}$
	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	0
	$-\frac{5}{72}$	$\frac{1}{12}$	$\frac{1}{9}$	$-\frac{1}{8}$

Přesnější z obou metod páru je Ralstonova R3 metoda

$$y^{**}_{n+1} = y_n + \tau\left[\tfrac{2}{9}k_1 + \tfrac{1}{3}k_2 + \tfrac{4}{9}k_3\right] = y_{n+1}$$

řádu 3. Pomocná metoda

$$y^{*}_{n+1} = y_n + \tau\left[\tfrac{7}{24}k_1 + \tfrac{1}{4}k_2 + \tfrac{1}{3}k_3 + \tfrac{1}{8}k_4\right]$$

řádu 2 používá kromě koeficientů k_1, k_2 a k_3 navíc ještě koeficient $k_4 = f(t_{n+1}, y_{n+1})$. V každém kroku se tedy počítají jen 3 nové hodnoty funkce f, neboť koeficient k_1 ve stávajícím kroku je roven koeficientu k_4 z kroku předchozího, takže nově se počítají jen koeficienty k_2, k_3 a k_4. Zařazení koeficientu k_4 do metody řádu 2 nás tedy téměř nic nestojí, umožní však zlepšit vlastnosti této metody a tím i celého páru. Metoda, ve které $k_s = f(t_{n+1}, y_{n+1})$, bývá označována jako FSAL podle anglického *First Same As Last*. Zdůrazněme, že BS32 metoda se používá jako metoda s lokální extrapolací, tj. $y_{n+1} = y^{**}_{n+1}$.

Hodnoty přibližného řešení pro $t \in \langle t_n, t_{n+1}\rangle$ spočteme dostatečně přesně pomocí kubického Hermitova polynomu $H_3(t)$ určeného podmínkami

$$\begin{aligned} H_3(t_n) &= y_n, & H_3'(t_n) &= k_1\,, \\ H_3(t_{n+1}) &= y_{n+1}, & H_3'(t_{n+1}) &= k_4\,. \end{aligned}$$

Metoda BS32 je tedy skvělá: je řádu 3, v každém kroku se pravá strana f počítá jen 3-krát, a to stačí jak na řízení délky kroku tak na výpočet řešení mezi uzly t_n a t_{n+1}.

Dormand-Prince metoda, stručně DP54 metoda, je definována rozšířenou Butcherovou tabulkou 8.1. Metoda DP54 je typu FSAL, neboť $k_7 = f(t_{n+1}, y_{n+1})$. Proto se v každém úspěšném kroku metody počítají jen koeficienty $k_2, \ldots, k_7$, koeficient k_1 byl už vypočten jako koeficient k_7 v předchozím kroku. Zdůrazněme, že DP54 metoda se používá jako metoda s lokální extrapolací, tj. $y_{n+1} = y^{**}_{n+1}$.

$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19\,372}{6\,561}$	$-\frac{25\,360}{2\,187}$	$\frac{64\,448}{6\,561}$	$-\frac{212}{729}$			
1	$\frac{9\,017}{3\,168}$	$-\frac{355}{33}$	$\frac{46\,732}{5\,247}$	$\frac{49}{176}$	$-\frac{5\,103}{18\,656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1\,113}$	$\frac{125}{192}$	$-\frac{2\,187}{6\,784}$	$\frac{11}{84}$	
	$\frac{5\,179}{57\,600}$	0	$\frac{7\,571}{16\,695}$	$\frac{393}{640}$	$-\frac{92\,097}{339\,200}$	$\frac{187}{2\,100}$	$\frac{1}{40}$
	$\frac{35}{384}$	0	$\frac{500}{1\,113}$	$\frac{125}{192}$	$-\frac{2\,187}{6\,784}$	$\frac{11}{84}$	0
	$\frac{71}{57\,600}$	0	$-\frac{71}{16\,695}$	$\frac{71}{1\,920}$	$-\frac{17\,253}{339\,200}$	$\frac{22}{525}$	$-\frac{1}{40}$

Tab 8.1. Dormand-Prince (5,4) metoda

Hodnoty přibližného řešení pro $t \in \langle t_n, t_{n+1}\rangle$ spočteme dostatečně přesně pomocí interpolačního polynomu $H_4(t) = y_n + \tau \mathbf{k}\mathbf{B}\mathbf{q}$, kde $\mathbf{k} = (k_1, k_2, k_3, k_4, k_5, k_6, k_7)$,

$$\mathbf{B} = \begin{bmatrix} 1 & -\frac{183}{64} & \frac{37}{12} & -\frac{145}{128} \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1\,500}{371} & -\frac{1\,000}{159} & \frac{1\,000}{371} \\ 0 & -\frac{125}{32} & \frac{125}{12} & -\frac{375}{64} \\ 0 & \frac{9\,477}{3\,392} & -\frac{729}{106} & \frac{25\,515}{6\,784} \\ 0 & -\frac{11}{7} & \frac{11}{3} & -\frac{55}{28} \\ 0 & \frac{3}{2} & -4 & \frac{5}{2} \end{bmatrix},$$

$\mathbf{q} = (q, q^2, q^3, q^4)^T$, přičemž $q = (t - t_n)/\tau$.

Metoda DP54 je rovněž vynikající: je řádu 5, v každém kroku se pravá strana počítá jen 6-krát, a to stačí jak pro řízení délky kroku tak pro výpočet řešení v intervalu $\langle t_n, t_{n+1}\rangle$.

Stabilita. Řešíme-li testovací rovnici (1.10) RK metodou na rovnoměrném dělení s krokem τ, dostaneme

$$y_{n+1} = P_s(\tau\lambda) y_n \,,$$

kde P_s je polynom stupně s určený pomocí konstant b_i, a_{ij} definujících RK metodu. Podmínka stability (1.11) tedy platí, právě když $|P_s(\tau\lambda)| < 1$, neboli když $\tau\lambda$ leží v oblasti absolutní stability R_A:

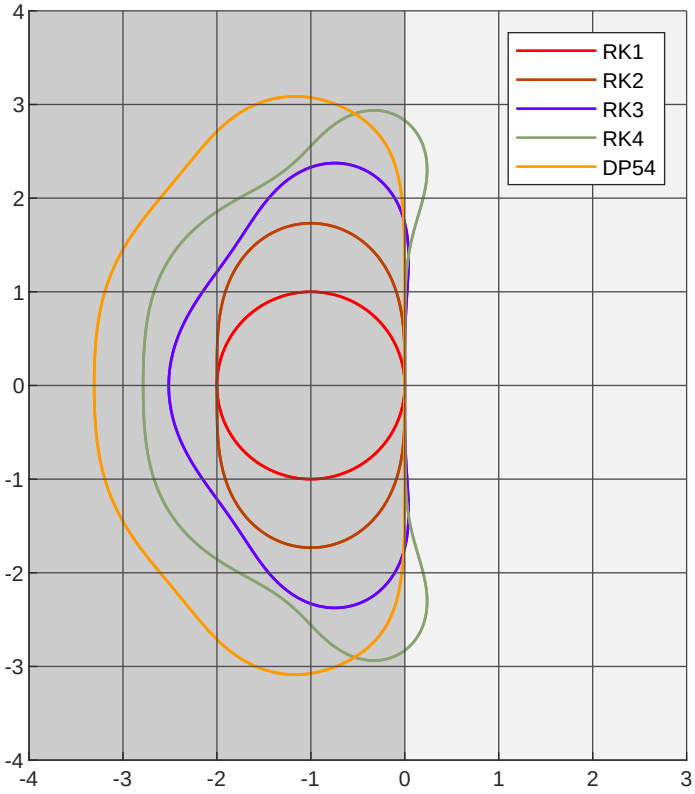
$$\tau\lambda \in R_A = \{z \in \mathbb{C} \,|\, |P_s(z)| < 1\}.$$

Explicitní RK metody mají omezenou oblast absolutní stability, neboť $|P_s(z)| \to \infty$ pro $|z| \to \infty$. Dá se ukázat, že pro $p = s \le 4$ je $P_s(z) = \sum_{i=0}^{s} z^i/i!$. Proto každá explicitní s-stupňová RK metoda řádu $p = s \le 4$ (stručně RKs metoda) má stejnou oblast absolutní

stability. Oblasti absolutní stability RKs metod pro $s = 1, 2, 3, 4$ a metody DP54 jsou uvedeny v obrázku 8.4. Intervaly absolutní stability těchto metod jsou $I_A = (\alpha, 0)$, kde

metoda	RK1	RK2	RK3	RK4	DP54
α	-2	-2	$-2{,}51$	$-2{,}79$	$-3{,}31$

Zdůrazněme, že z pohledu stability je BS32 metoda ekvivalentní s RK3 metodou, obě metody tedy maií steinou oblast a steinv interval absolutní stabilitv.



Obr. 8.4. Oblast absolutní stability RK metod

1.4. Lineární mnohokrokové metody

V této kapitole se budeme zabývat metodami, které počítají přibližné řešení y_{n+1} v uzlu t_{n+1} pomocí dříve spočtených aproximací $y_n, y_{n-1}, y_{n-2}, \ldots$ a odpovídajících hodnot $f(t_n, y_n)$, $f(t_{n-1}, y_{n-1}), f(t_{n-2}, y_{n-2}), \ldots$ pravé strany diferenciální rovnice. Tyto hodnoty jsou znovu použity tak, abychom získali y_{n+1} s vysokou přesností pomocí jen několika málo nových vyhodnocení funkce $f(t, y)$. Nejznámějšími metodami tohoto typu jsou *Adamsovy metody* a *metody zpětného derivování*. Obě skupiny metod patří do obecné třídy metod známých jako *lineární mnohokrokové metody*, stručně LMM.

1.4.1. Obecná lineární mnohokroková metoda

LMM je předpis

$$\alpha_0 y_{n+1} + \alpha_1 y_n + \cdots + \alpha_k y_{n+1-k} = \tau[\beta_0 f(t_{n+1}, y_{n+1}) + \beta_1 f_n + \cdots + \beta_k f_{n+1-k}], \quad (1.23)$$

ze kterého počítáme y_{n+1}. Přitom α_j a β_j jsou číselné koeficienty, které formuli jednoznačně určují, a f_j je zkrácený zápis pro $f(t_j, y_j)$. V dalším budeme předpokládat, že platí *normalizační podmínka* $\alpha_0 = 1$. Jestliže alespoň jeden z koeficientů α_k nebo β_k je různý od nuly, *metoda* je *k-kroková*.

Pro $\beta_0 \neq 0$ je nová hodnota y_{n+1} určena implicitně, hovoříme proto o *implicitní metodě*, pro $\beta_0 = 0$ máme *metodu explicitní*. Abychom v implicitní metodě určili y_{n+1}, musíme vyřešit obecně nelineární rovnici.

LMM lze použít, jen když jsou zadány *startovací hodnoty* $y_0, y_1, \ldots, y_{k-1}$. $y_0 = \eta$ určíme z počáteční podmínky, zbývající startovací hodnoty je však třeba získat jinou vhodnou metodou, y_r metodou nejvýše r-krokovou.

Lokální diskretizační chyba je chyba, která vznikne, když do formule (1.23) dosadíme místo přibližného řešení y_{n+1-j} přesné řešení $y(t_{n+1-j})$, tedy

$$\mathrm{lte}_n = \sum_{j=0}^{k} \alpha_j y(t_{n+1-j}) - \tau \sum_{j=0}^{k} \beta_j f(t_{n+1-j}, y(t_{n+1-j})).$$

Jestliže

$$\mathrm{lte}_n = C_{p+1}\tau^{p+1} y^{(p+1)}(t_n) + O(\tau^{p+2}),$$

řekneme, že metoda je řádu p. Člen $C_{p+1}\tau^{p+1}y^{(p+1)}(t_n)$ se nazývá *hlavní člen lokální diskretizační chyby*, konstanta C_{p+1} je tzv. *chybová konstanta*. LMM je tím přesnější, čím je vyššího řádu. Z několika metod téhož řádu je pak nejpřesnější ta metoda, pro kterou je velikost chybové konstanty $|C_{p+1}|$ nejmenší.

D-stabilita. Řekneme, že LMM je *stabilní ve smyslu Dahlquista* (stručně *D-stabilní*), jestliže všechny kořeny *prvního charakteristického polynomu*

$$\varrho(\xi) = \xi^k + \alpha_1 \xi^{k-1} + \cdots + \alpha_{k-1}\xi + \alpha_k$$

leží uvnitř jednotkového kruhu $|z| \leq 1$ komplexní roviny $\mathbb{C}$ a pokud některý kořen leží hranici $|z| = 1$, pak je jednoduchý.

Význam D-stability lze vysvětlit na rovnici $y' = 0$. Její řešení pomocí LMM vede na předpis $\sum_{j=0}^{k} \alpha_j y_{n+1-j} = 0$. Zvolíme-li startovací hodnoty $y_j = \varepsilon r^j$, $j = 0, 1, \ldots, k-1$, kde $\varrho(r) = 0$ a ε je libovolné číslo, pak $y_n = \varepsilon r^n$ pro každé n. Skutečně,

$$\sum_{j=0}^{k} \alpha_j \varepsilon r^{n+1-j} = \varepsilon r^{n+1-k} \sum_{j=0}^{k} \alpha_j r^{k-j} = \varepsilon r^{n+1-k} \varrho(r) = 0.$$

Pro $|r| > 1$ a $\varepsilon \neq 0$ je $\lim_{n\to\infty} |y_n| = \infty$, což je nepřijatelné: pro $\varepsilon = 0$ je $y_n = 0$ přesné řešení rovnice $y' = 0$, avšak pro $\varepsilon \neq 0$, $|\varepsilon| \ll 1$, tj. již pro nepatrnou poruchu startovacích

hodnot y_j, $j = 0, 1, \ldots, k-1$, dostaneme řešení zcela nevyhovující. Dá se ukázat, že vyloučit musíme také případ, kdy $|r| = 1$ je kořen $\varrho(\xi)$ násobnosti větší než 1.

Konvergence. Uvažujme D-stabilní LMM řádu $p \geq 1$. Jestliže startovací hodnoty zadáme s chybou řádu $O(\tau^p)$, pak globální diskretizační chyba je rovněž řádu $O(\tau^p)$.

Precizní formulaci a důkaz této věty lze najít např. v [29], [13]. Předpokladem její platnosti je dostatečná hladkost pravé strany f, konkrétně je třeba, aby funkce $f(t, y)$ měla spojité derivace až do řádu p včetně. Pokud f má spojité derivace jen do řádu $s \leq p$, pak lze pro globální chybu dokázat pouze řád $O(\tau^s)$.

Absolutní stabilita. Řešíme-li testovací úlohu (1.10) LMM na rovnoměrném dělení s krokem τ, dostaneme

$$\sum_{j=0}^{k} (\alpha_j - \tau\lambda\beta_j) y_{n+1-j} = 0\,. \tag{1.24}$$

Řešení hledejme ve tvaru $y_n = r^n$. Po dosazení do diferenční rovnice (1.24) obdržíme

$$\sum_{j=0}^{k} (\alpha_j - \tau\lambda\beta_j) r^{n+1-j} = r^{n+1-k} \sum_{j=0}^{k} (\alpha_j - \tau\lambda\beta_j) r^{k-j} = r^{n+1-k}\pi(r, \tau\lambda) = 0\,,$$

$$\text{kde} \quad \pi(\xi, z) = \sum_{j=0}^{k} (\alpha_j - z\beta_j)\xi^{k-j}$$

je tzv. *polynom stability* LMM. Jestliže $\pi(r, \tau\lambda) = 0$, pak $y_n = r^n$ je řešením diferenční rovnice (1.24) a podmínka stability $y_n \to 0$ pro $n \to \infty$ platí, právě když $|r| < 1$.

Oblast R_A absolutní stability LMM metody definujeme jako množinu takových bodů z komplexní roviny $\mathbb{C}$, pro které každý kořen ξ polynomu $\pi(\xi, z)$ leží uvnitř jednotkového kruhu komplexní roviny, tj. $|\xi| < 1$. Podmínka absolutní stability (1.11) tedy platí, když

$$\tau\lambda \in R_A = \{z \in \mathbb{C} \,|\, \pi(\xi, z) = 0 \;\Rightarrow\; |\xi| < 1\}.$$

1.4.2. Adamsovy metody

Integrací diferenciální rovnice (1.1) od t_n do t_{n+1} dostaneme

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t, y(t))\,\mathrm{d}t.$$

Funkci $f(t, y(t))$ aproximujeme pomocí interpolačního polynomu $P_{k-1}(t)$ stupně $k-1$, tj.

$$y(t_{n+1}) \approx y(t_n) + \int_{t_n}^{t_{n+1}} P_{k-1}(t)\,\mathrm{d}t, \quad \text{kde} \quad P_{k-1}(t_{n+1-j}) = f(t_{n+1-j}, y(t_{n+1-j})).$$

Když přibližnou rovnost nahradíme rovností a přesné řešení nahradíme řešením přibližným, dostaneme předpis

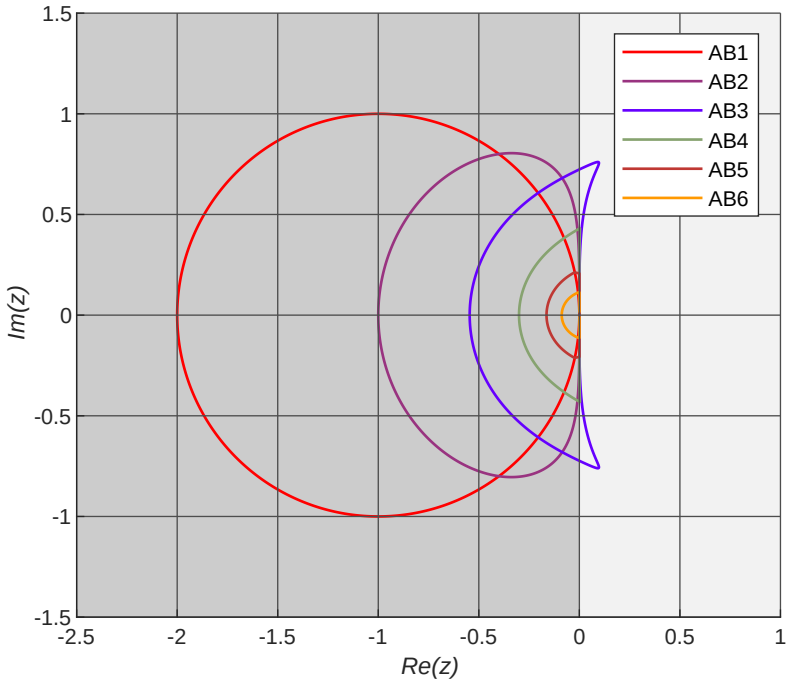
$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} P_{k-1}(t)\,\mathrm{d}t, \quad \text{kde} \quad P_{k-1}(t_{n+1-j}) = f(t_{n+1-j}, y_{n+1-j}). \tag{1.25}$$

Adams-Bashforthovy metody dostaneme, když v (1.25) zvolíme $j = 1, 2, \ldots, k$. Konstrukci polynomu $P_{k-1}(t)$ lze přehledně vyjádřit tabulkou

t	t_n	t_{n-1}	$\ldots$	t_{n+1-k}
$P_{k-1}(t)$	f_n	f_{n-1}	$\ldots$	f_{n+1-k}

Adams-Bashforthovu metodu lze zapsat ve tvaru

$$y_{n+1} = y_n + \tau \sum_{j=1}^{k} \beta^*_{k,j}\, f_{n+1-j}\,.$$



Obr. 8.5. Oblast absolutní stability ABk metod

Stručně ji budeme označovat jako ABk metodu. ABk metoda je explicitní, k-kroková, řádu k, D-stabilní. Pro konstantní délku kroku, tj. když

$$t_{n+1-j} = t_{n+1} - j\tau, \quad j = 1, 2, \ldots, k,$$

jsou koeficienty ABk metod pro $k = 1, 2, \ldots, 6$, spolu s chybovými konstantami C^*_{k+1} a dolními mezemi α^*_k intervalů absolutní stability $(\alpha^*_k, 0)$, uvedeny v následující tabulce:

k	$\beta^*_{k,1}$	$\beta^*_{k,2}$	$\beta^*_{k,3}$	$\beta^*_{k,4}$	$\beta^*_{k,5}$	$\beta^*_{k,6}$	C^*_{k+1}	α^*_k
1	1						$\frac{1}{2}$	-2
2	$\frac{3}{2}$	$-\frac{1}{2}$					$\frac{5}{12}$	-1
3	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$				$\frac{3}{8}$	$-\frac{8}{11}$
4	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$			$\frac{251}{720}$	$-\frac{3}{10}$
5	$\frac{1\,901}{720}$	$-\frac{2\,774}{720}$	$\frac{2\,616}{720}$	$-\frac{1\,274}{720}$	$\frac{251}{720}$		$\frac{95}{288}$	$-\frac{90}{551}$
6	$\frac{4\,277}{1\,440}$	$-\frac{7\,923}{1\,440}$	$\frac{9\,982}{1\,440}$	$-\frac{7\,298}{1\,440}$	$\frac{2\,877}{1\,440}$	$-\frac{475}{1\,440}$	$\frac{19\,087}{60\,480}$	$-\frac{5}{57}$

Všimněte si, že AB1 metoda je totožná s EE metodou, tj. AB1≡EE.

Adams-Moultonovy metody dostaneme, když v (1.25) zvolíme $j = 0, 1, \ldots, k-1$. Konstrukci polynomu $P_{k-1}(t)$ lze přehledně vyjádřit tabulkou

t	t_{n+1}	t_n	$\ldots$	t_{n+2-k}
$P_{k-1}(t)$	f_{n+1}	f_n	$\ldots$	f_{n+2-k}

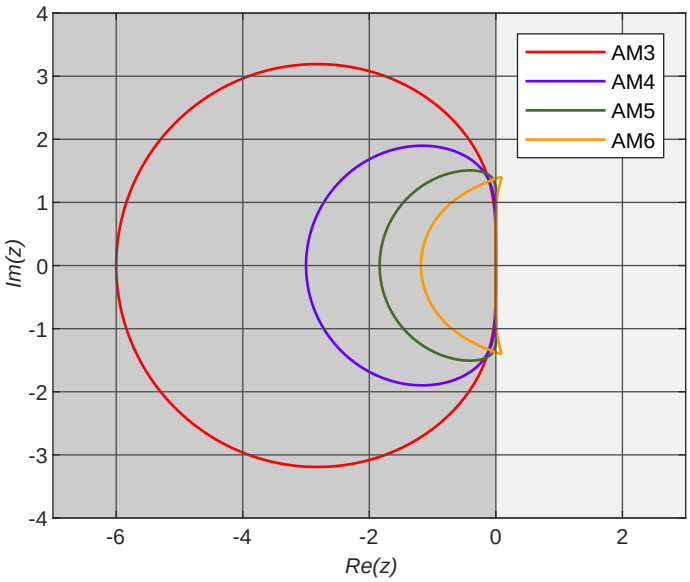
Adams-Moultonovu metodu lze zapsat ve tvaru

$$y_{n+1} = y_n + \tau\beta_{k,0}\, f(t_{n+1}, y_{n+1}) + \tau \sum_{j=1}^{k-1} \beta_{k,j}\, f_{n+1-j}\,.$$

Stručně ji budeme označovat jako AMk metodu. AMk metoda je implicitní, pro $k = 1$ je jednokroková a pro $k > 1$ je $(k-1)$-kroková, je řádu k a D-stabilní. Koeficienty AMk metod pro konstantní délku kroku a pro $k = 1, 2, \ldots, 6$, spolu s chybovými konstantami C_{k+1} a dolními mezemi α_k intervalů absolutní stability $(\alpha_k, 0)$, jsou uvedeny v následující tabulce:

k	$\beta_{k,0}$	$\beta_{k,1}$	$\beta_{k,2}$	$\beta_{k,3}$	$\beta_{k,4}$	$\beta_{k,5}$	C_{k+1}	α_k
1	1						$-\frac{1}{2}$	$-\infty$
2	$\frac{1}{2}$	$\frac{1}{2}$					$-\frac{1}{12}$	$-\infty$
3	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$				$-\frac{1}{24}$	-6
4	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$			$-\frac{19}{720}$	-3
5	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$		$-\frac{3}{160}$	$-\frac{90}{49}$
6	$\frac{475}{1\,440}$	$\frac{1\,427}{1\,440}$	$-\frac{798}{1\,440}$	$\frac{482}{1\,440}$	$-\frac{173}{1\,440}$	$\frac{27}{1\,440}$	$-\frac{863}{60\,480}$	$-\frac{45}{38}$

Všimněte si, že AM1 metoda je totožná s IE metodou, tj. AM1≡IE, a že AM2 metoda je totožná s TR metodou, tj. AM2≡TR.



Obr. 8.6. Oblast absolutní stability AMk metod

Metody prediktor-korektor. AM metody jsou přesnější a stabilnější než AB metody. Nevýhodou AM metod je jejich implicitnost. Abychom určili y_{n+1}, musíme řešit rovnici

$$y_{n+1} = \varphi(y_{n+1})\,, \quad \text{kde} \quad \varphi(z) = y_n + \tau\beta_{k,0}\, f(t_{n+1}, z) + \tau \sum_{j=1}^{k-1} \beta_{k,j}\, f_{n+1-j}\,.$$

Použít můžeme metodu prosté iterace: zvolíme počáteční aproximaci $y_{n+1}^{(0)}$ a postupně počítáme $y_{n+1}^{(s)} = \varphi(y_{n+1}^{(s-1)})$, $s = 1, 2, \ldots$. Pro dostatečně malé τ metoda konverguje. Jestliže počáteční aproximaci $y_{n+1}^{(0)}$ určíme pomocí AB metody, provedeme jen několik málo iterací

$$y_{n+1}^{(s)} = \varphi(y_{n+1}^{(s-1)})\,, \quad s = 1, 2, \ldots, S\,,$$

a nakonec položíme

$$y_{n+1} = y_{n+1}^{(S)}\,, \quad f_{n+1} = f(t_{n+1}, y_{n+1})\,,$$

dostaneme *metodu prediktor-korektor*, kterou schématicky označujeme P(EC)^{S}E. Přitom P značí předpověď počáteční aproximace AB metodou, C korekci AM metodou a E vyhodnocení pravé strany $f(t_{n+1}, y_{n+1}^{(s)})$. Zvolíme-li jako prediktor metodu ABk a jako korektor metodu AMk, dostaneme metodu, kterou značíme ABk-AMk-P(EC)^{S}E. Její chybová konstanta je rovna chybové konstantě C_{p+1} korektoru, oblast absolutní stability se blíží oblasti absolutní stability korektoru AMk až pro $S \to \infty$. Nejčastěji se používá schéma PECE, kdy se korekce provede jen jednou a pravá strana se počítá dvakrát. V dalším se omezíme právě na schéma PECE.

Abychom mohli řídit délku kroku, potřebujeme znát odhad est_n lokální chyby. Jestliže $y_{n+1}^{*} \equiv y_{n+1}^{(0)}$ spočteme prediktorem ABk a $y_{n+1}^{**} \equiv y_{n+1}^{(1)}$ korektorem AMk, pak tzv. *Milneův*

odhad lokální chyby dává

$$\mathrm{est}_n = \frac{C_{k+1}}{C^*_{k+1} - C_{k+1}}(y^{**}_{n+1} - y^*_{n+1}), \tag{1.26}$$

odvození viz [13]. Nakonec položíme

$$y_{n+1} = y^{**}_{n+1} + \mathrm{est}_n. \tag{1.27}$$

Korekce y^{**}_{n+1} pomocí odhadu lokální chyby est_n se nazývá *lokální extrapolace*. Celý krok označujeme zkratkou ABk-AMk-PECLE, přičemž písmeno L vyznačuje použití lokální extrapolace. Metoda ABk-AMk-PECLE je řádu $k+1$, oblast absolutní stability je větší než u prediktoru ABk ale menší než u korektoru AMk, viz [13], [1].

Alternativní odhad lokální chyby lze získat také tak, že y_{n+1} spočteme metodou AM(k+1) a položíme

$$\mathrm{est}_n = y_{n+1} - y^{**}_{n+1}. \tag{1.28}$$

Výslednou metodu lze označit jako ABk-AM(k+1)-PECE. Odhady (1.27) a (1.28) nejsou sice totožné, pro malé τ jsou však prakticky nerozlišitelné.

Konkrétně pro metodu AB2-AM2-PECLE organizujeme výpočet následovně:

P:	AB2:	$y^*_{n+1} = y_n + \frac{1}{2}\tau(3f_n - f_{n-1})$
E:		$f^*_{n+1} = f(t_{n+1}, y^*_{n+1})$
C:	AM2:	$y^{**}_{n+1} = y_n + \frac{1}{2}\tau(f^*_{n+1} + f_n)$
L:		$\mathrm{est}_n = -\frac{1}{6}(y^{**}_{n+1} - y^*_{n+1}), \quad y_{n+1} = y^{**}_{n+1} + \mathrm{est}_n$
E:		$f_{n+1} = f(t_{n+1}, y_{n+1}).$

V případě AB2-AM3-PECE metody nahradíme řádek L řádkem

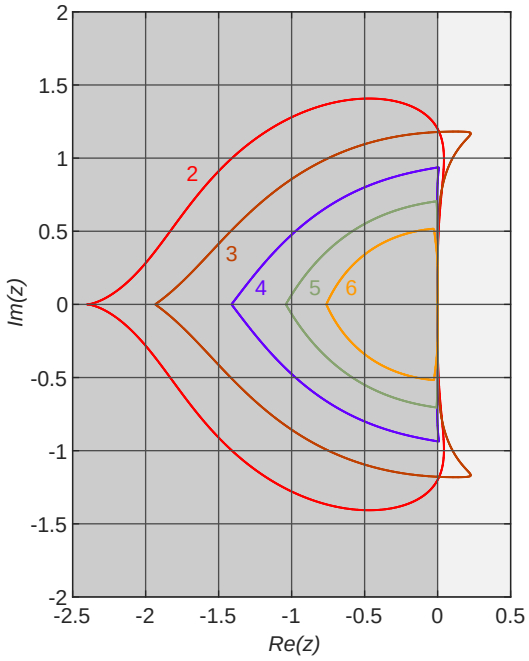
$$\text{C: AM3:} \qquad y_{n+1} = y_n + \tfrac{1}{12}\tau(5f^*_{n+1} + 8f_n - f_{n-1}), \quad \mathrm{est}_n = y_{n+1} - y^{**}_{n+1}.$$

Startovací hodnotu y_1 lze získat třeba pomocí EM2 metody.

Na obrázku 8.7 je vyznačena oblast absolutní stability metod ABk-AM(k+1)-PECE pro $k = 2, 3, \ldots, 6$: je větší než oblast absolutní stability prediktoru ABk, viz obrázek 8.5, avšak menší než oblast absolutní stability korektoru AM(k+1), viz obrázek 8.6.

Řízení délky kroku a řádu metody. Kvalitní programy založené na metodách prediktor-korektor mění jak délku kroku tak řád metody. Tak například matlabovský program `ode113` používá metody ABk-AM(k+1)-PECE pro $k = 1, 2, \ldots, 12$.

Změna řádu se provádí současně se změnou délky kroku. Algoritmy tohoto typu se označují jako VSVO (*variable step variable order*). Základní myšlenka je jednoduchá. Předpokládejme, že jsme vypočetli y_{n+1} metodou ABk-AM(k+1)-PECE s krokem délky τ. Určíme odhad est^k_n lokální chyby podle (1.27) nebo (1.28). Jestliže $|\mathrm{est}^k_n| > \varepsilon$, jde o neúspěch a výpočet y_{n+1} je třeba opakovat, v opačném případě pokračujeme výpočtem přibližného řešení y_{n+2}. V každém případě je však třeba stanovit novou délku kroku a nový řád. Řád



Obr. 8.7. Oblast absolutní stability pro metody ABk-AM(k+1)-PECE, $k = 2, 3, \ldots, 6$

se může změnit nejvýše o jednotku, tj. v metodě ABk-AM(k+1)-PECE místo k může být nově také $k-1$ nebo $k+1$. Odhady odpovídajících lokálních chyb est_n^{k-1} a est_n^{k+1} lze získat snadno, viz [23]. Nové délky kroků τ_{k-1}^*, τ_k^* a τ_{k+1}^* stanovíme podobně jako pro jednokrokovou metodu,

$$\tau_\ell^* = \theta\tau(\varepsilon/|\mathrm{est}_n^\ell|)^{1/(\ell+1)} \quad \text{pro } \ell = k-1,\ k,\ k+1\,,$$

a největší z čísel τ_{k-1}^*, τ_k^*, τ_{k+1}^* určí jak novou délku kroku tak nový řád. Konkrétně, je-li největší τ_k^*, k se nemění a jako novou délku kroku vezmeme $\tau^* = \tau_k^*$, je-li největší τ_{k+1}^*, zvětšíme k o jedničku a pokračujeme s krokem délky $\tau^* = \tau_{k+1}^*$ a je-li největší τ_{k-1}^*, k o jedničku snížíme a pokračujeme s krokem délky $\tau^* = \tau_{k-1}^*$.

Pro krok délky $\tau^* \neq \tau$ je třeba vypočítat hodnoty f_{n+1-j}^* pro $t_{n+1-j}^* = t_{n+1} - j\tau^*$. To lze snadno provést interpolací pomocí $f_{n+1}, f_n, \ldots$. Podrobný popis strategie VSVO lze najít např. v [13], [23].

Start metody není žádný problém: začneme metodou AB1-AM2-PECE, počáteční délku kroku určíme např. podle (1.22) a algoritmus VSVO se už sám rychle vyladí na správné hodnoty jak délky kroku tak řádu metody.

1.4.3. Metody zpětného derivování

Při řešení tzv. tuhých problémů, viz kapitola 1.5, je vhodné používat metody s neomezenou oblastí absolutní stability. Metody zpětného derivování (stručně BDF podle *backward differentiation formulas*) tuto vlastnost mají. Pro k-krokovou metodu zpětného derivování použijeme zkrácený zápis BDFk metoda. Dostaneme ji tak, že v rovnici

$$y'(t_{n+1}) = f(t_{n+1}, y(t_{n+1}))$$

nahradíme derivaci $y'(t_{n+1})$ pomocí derivace $P_k'(t_{n+1})$ interpolačního polynomu $P_k(t)$ stupně k procházejícího body $[t_{n+1}, y(t_{n+1})]$, $[t_n, y(t_n)]$, $\ldots$, $[t_{n+1-k}, y(t_{n+1-k})]$. Když pak nahradíme $y(t_{n+1-j})$ přibližnými hodnotami y_{n+1-j}, $j = 0, 1, \ldots, k$, dostaneme metodu

$$P_k'(t_{n+1}) = f(t_{n+1}, y_{n+1}).$$

Přehledné vyjádření aproximujícího polynomu $P_k(t)$ uvádí následující tabulka

t	t_{n+1}	t_n	$\ldots$	t_{n+1-k}
$P_k(t)$	y_{n+1}	y_n	$\ldots$	y_{n+1-k}

BDFk metodu zapíšeme ve tvaru

$$\alpha_{k,0} y_{n+1} + \sum_{j=1}^{k} \alpha_{k,j} y_{n+1-j} = \tau \beta_{k,0} f(t_{n+1}, y_{n+1}) \,.$$

Metoda BDFk je implicitní, k-kroková, řádu k a D-stabilní pro $k \leq 6$. Pro konstantní délku kroku jsou koeficienty $\alpha_{k,j}$, $\beta_{k,0}$ a chybové konstanty C_{k+1} BDFk metod uvedeny v následující tabulce:

k	$\alpha_{k,0}$	$\alpha_{k,1}$	$\alpha_{k,2}$	$\alpha_{k,3}$	$\alpha_{k,4}$	$\alpha_{k,5}$	$\alpha_{k,6}$	$\beta_{k,0}$	C_{k+1}	α_k
1	1	-1						1	$-\frac{1}{2}$	90°
2	1	$-\frac{4}{3}$	$\frac{1}{3}$					$\frac{2}{3}$	$-\frac{2}{9}$	90°
3	1	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$				$\frac{6}{11}$	$-\frac{3}{22}$	88°
4	1	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$			$\frac{12}{25}$	$-\frac{12}{125}$	73°
5	1	$-\frac{300}{137}$	$\frac{300}{137}$	$-\frac{200}{137}$	$\frac{75}{137}$	$-\frac{12}{137}$		$\frac{60}{137}$	$-\frac{10}{137}$	52°
6	1	$-\frac{360}{147}$	$\frac{450}{147}$	$-\frac{400}{147}$	$\frac{225}{147}$	$-\frac{72}{147}$	$\frac{10}{147}$	$\frac{60}{147}$	$-\frac{20}{343}$	18°

Všimněte si, že BDF1 metoda je totožná s IE metodou, tj. BDF1≡IE.

BDFk metody jsou implicitní, ve srovnání s implicitními AMk metodami ale mají značně větší chybové konstanty. Na druhé straně však metody zpětného derivování mají jednu ohromnou přednost, která plně ospravedlňuje jejich používání, a tou je neomezená oblast R_A absolutní stability. Pro metody BDF1 a BDF2 oblast R_A absolutní stability obsahuje celou zápornou polorovinu komplexní roviny $\mathbb{C}$, tj. $\mathrm{R}_A \supseteq \{z \in \mathbb{C} \,|\, \mathrm{Re}\, z < 0\}$. Takové metody se nazývají *A-stabilní*.

Abychom mohli popsat oblast absolutní stability zbývajících BDFk metod, zavedeme si jeden nový pojem. Řekneme, že numerická metoda je *A(α)-stabilní*, $\alpha \in (0, \pi/2\rangle$, jestliže její oblast absolutní stability R_A obsahuje nekonečný klín

$$W_\alpha = \{re^{i\varphi} \in \mathbb{C} \,|\, r > 0, |\varphi - \pi| < \alpha\}.$$

BDFk metody jsou (pro $k \leq 6$) A(α)-stabilní, příslušné úhly α_k (pro větší názornost ve stupních) jsou uvedeny jako poslední sloupec výše uvedené tabulky.

BDFk metody (pro $k \leq 6$) jsou dokonce L(α)-stabilní. L(α) stabilní metodu přitom definujeme jako A(α)-stabilní metodu takovou, že když $z \in R_A$, $\pi(\xi, z) = 0$, $Re(z) \to -\infty$, pak $\xi \to 0$. Pro $\alpha = \frac{\pi}{2}$ dostáváme L-stabilní metodu. BDF1 a BDF2 jsou tedy L-stabilní.

Úhel 18° metody BDF6 je příliš malý a proto se tato metoda obvykle nepoužívá. V matlabovském programu `ode15s` jsou implementovány metody zpětného derivování řádů 1 až 5. Program `ode15s` je typu VSVO, tj. volí optimální délku kroku i řád metody. Základní metodou programu `ode15s` je metoda NDF (podle *numerical differentiation formula*). NDFk metody jsou modifikace BDFk metod, mají o něco menší chybové konstanty (o 26% pro $k = 1, 2, 3$ a o 15% pro $k = 4$) a poněkud menší úhly $A(\alpha)$ stability (o 7% pro $k = 3$ a o 10% pro k=4) než odpovídající BDFk metody. Podrobnosti týkající se NDFk metod lze najít v [25].

Nelineární rovnice pro výpočet y_{n+1} se řeší pomocí několika málo kroků Newtonovy metody.

1.5. Tuhé problémy

Tuhý počáteční problém (v anglicky psané literatuře *stiff problem*) se vyznačuje několika charakteristikami, z nichž dvě si postupně uvedeme. Praktická a snadno ověřitelná je

Charakteristika 1. *Počáteční problém je tuhý, když počet kroků, který k jeho vyřešení potřebuje metoda s omezenou oblastí absolutní stability, je podstatně větší než počet kroků, který k jeho vyřešení potřebuje metoda s neomezenou oblastní absolutní stability.*

Platnost této charakteristiky ukážeme na příkladu počátečního problému

Příklad 1.1

$$\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1000 & -1001 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad x \in (0, \ell), \quad \begin{pmatrix} y_1(0) \\ y_2(0) \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{1.29}$$

Řešení je $y_1 = -e^{-t}$, $y_2 = e^{-t}$. Úlohu (1.29) jsme řešili matlabovským programem `ode45` (DP54 metoda řádu 5 s omezenou oblastí absolutní stability) a matlabovským programem `ode15s` (BDF metody VSVO řádů 1-5 s neomezenými oblastmi absolutní stability). Oba programy jsme použili se stejným požadavkem na přesnost: $\varepsilon_r = 10^{-3}$ a $\varepsilon_a = 10^{-6}$. Efektivnost obou metod lze přibližně porovnat podle počtu úspěšně provedených kroků `pk` a počtu `pf` vyhodnocení pravé strany. V následující tabulce jsou uvedeny hodnoty `pk/pf` pro několik délek ℓ intervalu integrace.

ℓ	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}
`ode45`	10/61	22/151	269/1 747	2 953/18 919	30 071/192 475
`ode15s`	10/24	10/24	12/28	42/88	71/146

Pro malé $\ell = 10^{-2}$ se na intervalu $\langle 0; 10^{-2} \rangle$ řešení poměrně rychle mění. Délku kroku zde určuje požadavek na přesnost a protože obě metody jsou téhož řádu, potřebují přibližně stejný počet kroků. Jak však ℓ vzrůstá, délku kroku stále více začíná ovlivňovat podmínka stability. □

Abychom tomuto efektu lépe porozuměli, potřebujeme několik dalších poznatků.

Stabilní problém. *Počáteční problém* $\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \mathbf{y}(a) = \boldsymbol{\eta}$, *je stabilní, když malá změna dat* $\mathbf{f}, a, \boldsymbol{\eta}$ *způsobí malou změnu řešení* $\mathbf{y}$. Zabývejme se speciálně *stabilitou vzhledem k počáteční podmínce*, konkrétně jak změna počáteční hodnoty $\boldsymbol{\eta}$ ovlivní řešení $\mathbf{y}$.

Pro ilustraci prozkoumejme stabilitu počátečního problému

$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \qquad \mathbf{y}(a) = \boldsymbol{\eta}, \tag{1.30}$$

kde $\mathbf{A}$ je číselná matice řádu d. Nechť $\mathbf{u}$ je řešení problému (1.30) a $\mathbf{v}$ je řešení téže diferenciální rovnice, avšak s porušenou počáteční podmínkou, tj.

$$\begin{aligned} \mathbf{u}' &= \mathbf{A}\mathbf{u}, \qquad \mathbf{u}(a) = \boldsymbol{\eta}, \\ \mathbf{v}' &= \mathbf{A}\mathbf{v}, \qquad \mathbf{v}(a) = \boldsymbol{\eta} + \boldsymbol{\delta}. \end{aligned}$$

Pro $\mathbf{w} = \mathbf{v} - \mathbf{u}$ dostaneme problém

$$\mathbf{w}' = \mathbf{A}\mathbf{w}, \qquad \mathbf{w}(a) = \boldsymbol{\delta},$$

který popisuje šíření počáteční poruchy $\boldsymbol{\delta}$. Pro zjednodušení výkladu předpokládejme, že matice $\mathbf{A}$ má navzájem různá vlastní čísla $\{\lambda_j\}_{j=1}^d$. Pak

$$\mathbf{w}(t) = \sum_{j=1}^{d} c_j e^{\lambda_j (t-a)} \mathbf{v}_j,$$

kde $\mathbf{v}_j$ jsou vlastní vektory příslušné vlastním číslům λ_j a c_j jsou konstanty, které určíme z počáteční podmínky:

$$\mathbf{V}\mathbf{c} = \boldsymbol{\delta}, \quad \text{kde} \quad \mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d), \quad \mathbf{c} = (c_1, c_2, \ldots, c_d)^T.$$

Odtud $\mathbf{w}(t) = \mathbf{V}\mathbf{S}(t)\mathbf{V}^{-1}\boldsymbol{\delta}$, kde $\mathbf{S} = \operatorname{diag}\{e^{\lambda_1(t-a)}, e^{\lambda_2(t-a)}, \ldots, e^{\lambda_n(t-a)}\}$. Můžeme tedy vyslovit tyto závěry:

1) Jestliže $Re(\lambda_j) < 0$ *pro všechna* j, pak $\|\mathbf{w}(t)\| \to 0$ exponenciálně pro $t \to \infty$, tj. *porucha se velmi rychle zmenšuje.*

2) Jestliže $Re(\lambda_j) \le 0$ *pro všechna* j, pak $\|\mathbf{w}(t)\| \le \|\mathbf{V}\| \cdot \|\mathbf{V}^{-1}\| \cdot \|\boldsymbol{\delta}\|$, tj. *porucha bude omezená*, přitom $\|\mathbf{w}(t)\| \to 0$ pro $\boldsymbol{\delta} \to \mathbf{o}$.

3) Jestliže *nějaké vlastní číslo* λ_j *má kladnou reálnou složku* a počáteční porucha $\boldsymbol{\delta}$ je taková, že $c_j \neq 0$, pak $\|\mathbf{w}(t)\| \to \infty$ exponenciálně pro $t \to \infty$, tj. *porucha se velmi rychle zvětšuje.*

Problém (1.30) je tedy stabilní (vzhledem k počáteční podmínce), jestliže všechna vlastní čísla matice $\mathbf{A}$ jsou navzájem různá a mají nekladnou reálnou složku.

Toto tvrzení lze rozšířit a připustit i násobná vlastní čísla se zápornou reálnou složkou:

Problém (1.30) je stabilní (vzhledem k počáteční podmínce), jestliže:

(a) všechna vlastní čísla matice $\mathbf{A}$ mají nekladnou reálnou složku,
(b) ryze imaginární vlastní čísla jsou navzájem různá.

Další rozšíření pojmu stability na obecnou počáteční úlohu zní takto:

Řekneme, že počáteční problém (1.4), tj.

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)),\ \mathbf{y}(a) = \boldsymbol{\eta},$$

je stabilní (vzhledem k počáteční podmínce), jestliže vlastní čísla $\{\lambda_j(t)\}_{j=1}^d$ Jacobiho matic

$$\mathbf{f}_{\mathbf{y}}(t, \mathbf{y}(t)) = \left\{ \frac{\partial f_i(t, \mathbf{y}(t))}{\partial y_j} \right\}_{i,j=1}^{d}, \qquad t \in \langle a, b \rangle,$$

mají nekladnou reálnou složku, tj. $\mathrm{Re}(\lambda_j(t)) \leq 0$, $j = 1, 2, \ldots, d$, $t \in \langle a, b \rangle$. Případná ryze imaginární vlastní čísla musejí být navzájem různá.

Spektrální poloměr. Označme symbolem $\varrho(\mathbf{A})$ *spektrální poloměr* matice $\mathbf{A}$ definovaný jako velikost největšího vlastního čísla $\mathbf{A}$, tj.

$$\varrho(\mathbf{A}) = \max_{j=1,2,\ldots,d} |\lambda_j|, \quad \text{kde } \lambda_j,\ j = 1, 2, \ldots, d, \text{ jsou vlastní čísla } \mathbf{A}.$$

Nyní již můžeme zformulovat další charakteristika tuhého systému.

Charakteristika 2. *Stabilní problém je tuhý, jestliže součin spektrálního poloměru Jacobiho matice* $\mathbf{f}_{\mathbf{y}}(t, \mathbf{y}) = \{\partial f_i(t, \mathbf{y})/\partial y_j\}_{i,j=1}^d$ *a délky intervalu integrace* $b - a$ *je velký, tj. když*

$$\max_{a \leq t \leq b} \varrho(\mathbf{f}_{\mathbf{y}}(t, \mathbf{y}(t)))(b - a) \gg 1. \tag{1.31}$$

Vraťme se nyní k úloze (1.30). Pro $\mathbf{f} = \mathbf{A}\mathbf{y}$ je $\mathbf{f}_{\mathbf{y}}(t, \mathbf{y}) = \mathbf{A}$. Jsou-li všechna vlastní čísla matice $\mathbf{A}$ jednoduchá a mají zápornou reálnou složku, pak pro řešení $\mathbf{y}(t)$ úlohy (1.30) platí $\mathbf{y}(t) \to \mathbf{o}$ pro $t \to \infty$. Řešíme-li takovou úlohu numericky, pak lze dokázat, že podmínka stability $\mathbf{y}_n \to \mathbf{o}$ pro $t_n = n\tau \to \infty$ platí, právě když

$$\{\lambda_1\tau, \lambda_2\tau, \ldots, \lambda_d\tau\} \subseteq R_A, \tag{1.32}$$

kde R_A je oblast absolutní stability uvažované numerické metody. Pro obecnou pravou stranu $\mathbf{f}$ podmínka stability vyžaduje volit τ tak, aby

$$\{\lambda_1^n\tau, \lambda_2^n\tau, \ldots, \lambda_d^n\tau\} \subseteq R_A, \qquad n = 0, 1, \ldots, \tag{1.32'}$$

kde $\{\lambda_i^n\}_{i=1}^d$ jsou vlastní čísla matice $\mathbf{f}_{\mathbf{y}}(t_n, \mathbf{y}_n)$.

Příklad 1.1 – pokračování. Pravá strana diferenciální rovnice je

$$\mathbf{f} = \mathbf{A}\mathbf{y}, \qquad \text{kde} \quad \mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1000 & -1001 \end{pmatrix}$$

Vlastní čísla $\mathbf{A}$ jsou $\lambda_1 = -1$, $\lambda_2 = -1000$, takže $\varrho(\mathbf{A}) = 1000$. Jde o stabilní problém (vlastní čísla $\mathbf{A}$ jsou záporná) a pro větší ℓ je $\varrho(\mathbf{A})(b - a) = 1000\,\ell \gg 1$, tj. jde o tuhý

problém. Metoda DP54 má interval absolutní stability $(-3{,}31; 0)$, takže podmínka stability (1.32) vyžaduje, aby $-3{,}31 < -1000\tau$, tj. $\tau < 0{,}00331$. Na intervalu délky ℓ je tak třeba $n_\ell > \ell/0{,}00331$ kroků délky menší než 0,00331, což je v souladu s tabulkou uvedenou v první části příkladu 1.1. Skutečně, na intervalu$\langle 10; 100\rangle$ délky 90 je třeba alespoň $90/0{,}00331 \doteq 27\,190$ kroků délky 0,00331, ve skutečnosti program `ode45` provedl přibližně stejný počet $30\,071 - 2\,953 = 27\,118$ kroků proměnné délky. Protože přesná řešení $y_1 = -e^{-t}$ a $y_2 = e^{-t}$ jsou na intervalu $\langle 10; 100\rangle$ téměř konstantní, rovná nule, je zřejmé, že délka kroku je omezena z důvodu stability a ne z důvodu přesnosti. □.

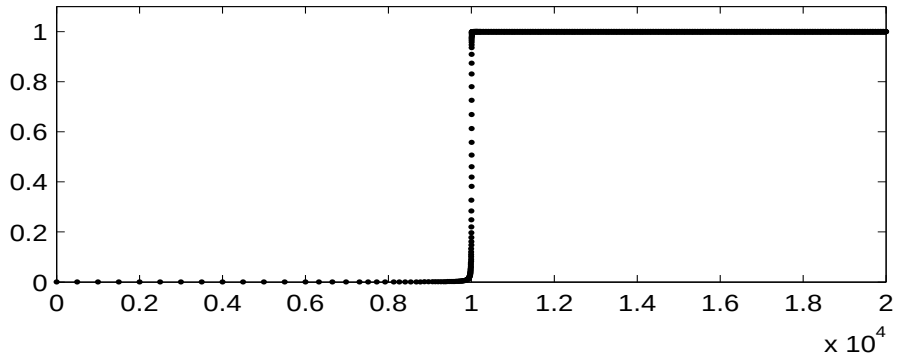
Příklad 1.2 Uvažujme počáteční problém

$$y' = y^2 - y^3\,, \qquad t \in (0, 2/\delta)\,, \qquad y(0) = \delta\,. \tag{1.33}$$

Diferenciální rovnice má dvě konstantní řešení $y = 0$ a $y = 1$: zvolíme-li počáteční podmínku $y(0) \le 0$, pak $y(t) \to 0$ pro $t \to \infty$, zatímco pro $y(0) > 0$ dostaneme $y(t) \to 1$ pro $t \to \infty$. Zvolíme-li $\delta > 0$ velmi malé, pak pravá strana $y^2 - y^3$ diferenciální rovnice nabývá malých kladných hodnot, tj. funkce $y(t)$ velmi pomalu roste a na poměrně dlouhém intervalu zůstávají její hodnoty blízké k 0. Konkrétně pro $\delta = 10^{-4}$ je $y(t) < 10^{-2}$ ještě pro $t = 9\,900$, pak $y(t)$ začíná prudce růst a pro $t > 10\,020$ je už $y(t)$ prakticky rovno 1. Spektrální poloměr $\varrho(f_y) = |2y - 3y^2|$. Pro malé $y \approx 0$ je $|2y - 3y^2| \approx 0$, pro y blízké 1 je však $|2y - 3y^2| \approx 1$. Na intervalu $(0, 9\,900)$ je výraz $9\,900|2y - 3y^2|$ charakterizující tuhost poměrně malý, nejde zde proto o tuhý problém, takže délka kroku se řídí především přesností metody. V intervalu $(9\,900; 10\,020)$ se řešení prudce mění, to mechanismus automatického řízení délky kroku zachytí a krok zkrátí. Důvodem zkrácení kroku je zde spíše prudká změna řešení než narůstající tuhost. Poté, co řešení nabude hodnotu rovnou přibližně 1, je délka kroku metody řízena stabilitou.

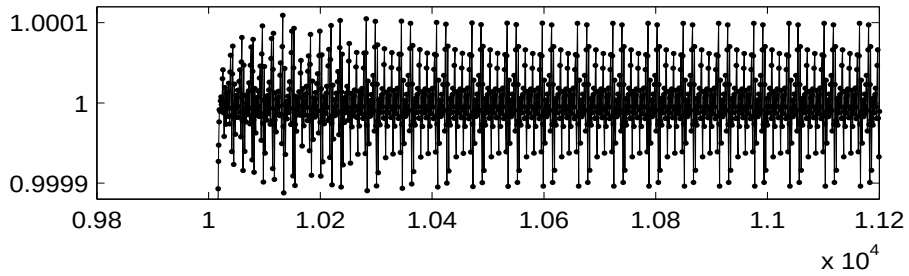
Úlohu (1.33) jsme řešili explicitní metodou DP54 (matlabovský program `ode45`) a implicitní TR metodou (matlabovský program `ode23t`). Oba programy jsme použili se stejnou přesností ($\varepsilon_r = 10^{-4}$, $\varepsilon_a = 10^{-7}$).

DP54 je explicitní Rungova-Kuttova metoda řádu 5 s intervalem absolutní stability $(-3{,}31; 0)$. Délka kroku proto musí splňovat podmínku stability $\tau|2y - 3y^2| < 3{,}31$, což pro $t > 10\,020$ znamená volit $\tau \doteq 3{,}31$. TR metoda je A-stabilní metoda řádu 2, takže tato metoda délku kroku z důvodu stability nijak neomezuje.

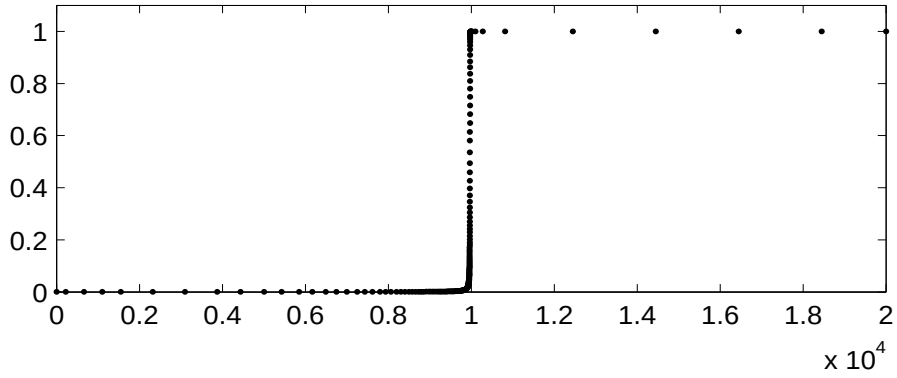


Obr. 1.8 Příklad 1.2 řešený DP54 metodou, celý výpočet

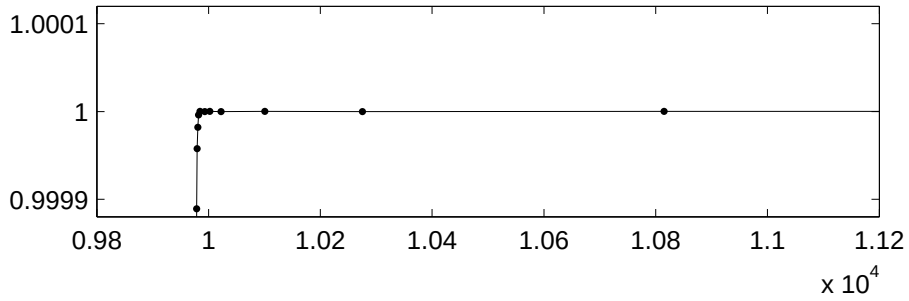
Vidíme, že v intervalu $(10\,020, 20\,000)$, kde přesné řešení je prakticky rovno 1, je TR-metoda velmi efektivní, na zdolání intervalu $(10\,020, 20\,000)$ potřebovala jen 8 kroků. Zato metoda DP54 na tomto intervalu provedla 3 005 kroků délky $\tau \doteq 3{,}31$, takže její použití rozhodně vhodné není. □



Obr. 1.9 Příklad 1.2 řešený DP54 metodou, detail



Obr. 1.10 Příklad 1.2 řešený TR metodou, celý výpočet



Obr. 1.11. Příklad 1.2 řešený TR metodou, detail

Metody pro řešení tuhých problémů. Pro řešení tuhých problémů je třeba používat metody s neomezenou oblastí absolutní stability. V Matlabu jsou to metody NDF a BDF implementované v programu `ode15s`, TR metoda implementovaná v programu `ode23t`

a dvě L-stabilní metody řádu 2: TR-BDF2 metoda (jde o kombinaci metod TR a BDF2) implementovaná v programu `ode23tb` a Rosenbrockova metoda implementovaná v programu `ode23s`, podrobnosti viz [16]. Programy `ode15s`, `ode23t` a `ode23tb` vyžadují řešení nelineárních soustav rovnic tvaru

$$\mathbf{y}_{n+1} = \tau\gamma\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) + \boldsymbol{\psi}, \tag{1.34}$$

kde parametr γ je charakteristická konstanta metody a $\boldsymbol{\psi}$ je vektor nezávislý na $\mathbf{y}_{n+1}$. Například pro TR metodu (1.13) je

$$\gamma = \tfrac{1}{2} \text{ a } \boldsymbol{\psi} = \mathbf{y}_n + \tfrac{1}{2}\tau\mathbf{f}(t_n, \mathbf{y}_n).$$

Rovnici (1.34) řešíme zjednodušenou Newtonovou metodou. Počáteční aproximaci $\mathbf{y}_{n+1}^{(0)}$ získáme dostatečně přesnou extrapolací z hodnot $\mathbf{y}_n, \mathbf{y}_{n-1}, \ldots$. Další aproximace $\mathbf{y}_{n+1}^{(s+1)}$ získáme řešením soustav lineárních rovnic

$$(\mathbf{I} - \tau\gamma\mathbf{J})\,(\mathbf{y}_{n+1}^{(s+1)} - \mathbf{y}_{n+1}^{(s)}) = \tau\gamma\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}^{(s)}) + \boldsymbol{\psi} - \mathbf{y}_{n+1}^{(s)}\,, \tag{1.35}$$

kde $\mathbf{I}$ je jednotková matice a $\mathbf{J}$ je Jacobiho matice $\mathbf{f}_{\mathbf{y}}(t_{n+1-k}, \mathbf{y}_{n+1-k})$ pro nějaké $k > 0$ (jedná se tedy o zjednodušenou Newtonovu metodu, pokud by $\mathbf{J} = \mathbf{f}_{\mathbf{y}}(t_{n+1}, \mathbf{y}_{n+1}^{(s)})$, šlo by o klasickou Newtonovu metodu). Výpočet organizujeme takto: označíme

$$\mathbf{G} = \mathbf{I} - \tau\gamma\mathbf{J}, \qquad \mathbf{d}_s = \mathbf{y}_{n+1}^{(s+1)} - \mathbf{y}_{n+1}^{(s)}, \qquad \mathbf{g}_s = \tau\gamma\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}^{(s)}) + \boldsymbol{\psi} - \mathbf{y}_{n+1}^{(s)}$$

a vypočteme nejdříve $\mathbf{d}_s$ jako řešení soustavy lineárních rovnic $\mathbf{G}\mathbf{d}_s = \mathbf{g}_s$ a pak dopočítáme $\mathbf{y}_{n+1}^{(s+1)} = \mathbf{y}_{n+1}^{(s)} + \mathbf{d}_s$. Je-li přírustek $\mathbf{d}_s$ dostatečně malý, klademe $\mathbf{y}_{n+1} = \mathbf{y}_{n+1}^{(s+1)}$.

Matice soustavy $\mathbf{G}$ obsahuje tři členy, které se mohou měnit: τ při změně délky kroku, γ při změně metody (třeba ve VSVO implementaci metod zpětného derivování) a $\mathbf{J}$ při přepočítání Jacobiho matice. Pokud se žádný z těchto členů nezmění, zůstává matice $\mathbf{G}$ stejná. Toho je třeba využít: pouze při změně $\mathbf{G}$ provedeme výpočetně náročný LU rozklad matice soustavy $\mathbf{G} = \mathbf{L}\mathbf{U}$, kde $\mathbf{L}$ je dolní trojúhelníková matice a $\mathbf{U}$ horní trojúhelníková matice. V následujících iteracích, kdy se matice soustavy $\mathbf{G}$ nemění, provádíme výpočetně nenáročná řešení dvou soustav lineárních rovnic s trojúhelníkovou maticí soustavy, tj. $\mathbf{L}\boldsymbol{\delta}_s = \mathbf{g}_s$ a pak $\mathbf{U}\mathbf{d}_s = \boldsymbol{\delta}_s$.

Program, který má pracovat efektivně, musí mít promyšlenou strategii, podle níž rozhodne, kdy změní $\mathbf{J}$, což je výpočetně nejnáročnější, a kdy jen τ nebo γ, což znamená nový LU rozklad matice $\mathbf{G}$. Používá se „konzervativní strategie“: přepočítání Jacobiho matice se provede až tehdy, když Newtonova metoda nekonverguje dostatečně rychle, délku kroku případně metodu změníme, až když očekávaný zisk takové akce převýší náklady spojené s LU rozkladem.

Jacobiho matici lze zadat přesně nebo ji lze spočítat přibližně numericky (Matlab užívá funkci `odenumjac` z privátní knihovny toolboxu `matlab\funfun`). Pokud je Jacobiho matice řídká a uživatel zadá pozice jejích nenulových prvků, výpočet Jacobiho matice lze značně urychlit. Do dalších podrobností už zacházet nebudeme, zájemce odkazujeme na skvělou monografii [23] a na matlabovskou dokumentaci [16]. Významným zdrojem poučení je studium kódů jednotlivých matlabovských programů, příliš snadné čtení to však není.

Program `ode23s`, založený na implementaci Rosenbrockovy metody, se od zbývajících programů `ode15s`, `ode23t` a `ode23tb` liší v tom, že se vyhýbá řešení nelineárních rovnic. V každém kroku Rosenbrockovy metody je třeba sestavit Jacobiho matici $\mathbf{f}_{\mathbf{y}}(t_n, \mathbf{y}_n)$ a vektor derivací $\mathbf{f}_t(t_n, \mathbf{y}_n) = \{\partial f_i(t_n, \mathbf{y}_n)/\partial t\}_{i=1}^d$, provést LU rozklad matice $\mathbf{I} - \gamma\tau\mathbf{f}_{\mathbf{y}}(t_n, \mathbf{y}_n)$ a užitím tohoto rozkladu vyřešit tři soustavy lineárních rovnic, podrobnosti viz [25].

Statistika o činnosti matlabovských programů pro řešení ODR. Kvalitní programy uživateli vždy poskytují informaci o tom, jak úspěšně si při řešení konkrétního problému počínaly. V Matlabu se dodávají tato čísla:

- `pk` počet úspěšných kroků
- `pn` počet neúspěšných kroků
- `pf` počet vyhodnocených pravých stran $\mathbf{f}$
- `pj` počet sestavených Jacobiho matic $\mathbf{J}$
- `pr` počet LU rozkladů $\mathbf{J} = \mathbf{LU}$
- `ps` počet řešených soustav lineárních rovnic $\mathbf{LU}\mathbf{d}_s = \mathbf{g}_s$

Pro ilustraci uvádíme

Příklad 1.3 Robertsonův problém

$$\begin{aligned} y_1' &= -0{,}04\, y_1 + 10^4\, y_2 y_3\,, & y_1(0) &= 1\,,\\ y_2' &= 0{,}04\, y_1 - 10^4\, y_2 y_3 - 3\cdot 10^7\, y_2^2\,, & y_2(0) &= 0\,,\\ y_3' &= 3\cdot 10^7\, y_2^2\,, & y_3(0) &= 0 \end{aligned}$$

popisuje koncentrace tří příměsí v chemické reakci, tj. $0 \le y_1, y_2, y_3 \le 1$, nezávisle proměnná t je čas, blíže viz [23]. Jacobiho matice této soustavy je

$$\mathbf{J} = \begin{pmatrix} -0{,}04 & 10^4\, y_3 & 10^4\, y_2 \\ 0{,}04 & -10^4\, y_3 - 6\cdot 10^7 y_2 & -10^4 y_2 \\ 0 & 6\cdot 10^7 y_2 & 0 \end{pmatrix}.$$

V čase $t = 0$, tj. pro $y_1 = 1$, $y_2 = y_3 = 0$, má Jacobiho matice vlastní čísla $\{-0{,}04; 0; 0\}$. Z fyzikálních úvah plyne, že $y_1, y_2 \to 0$ a $y_3 \to 1$ pro $t \to \infty$. Vlastní čísla Jacobiho matice pro $y_1 = y_2 = 0$, $y_3 = 1$, jsou $\{-10\,000{,}04; 0; 0\}$. Při řešení na intervalu $(0, 10^{10})$ je Robertsonův problém tuhý. O tom se lze ostatně snadno přesvědčit experimentálně: explicitní metody selhávají, metody pro řešení tuhých problémů zabírají. Numerickým výpočtem lze zjistit, že již pro $t > 0{,}01$ je spektrální poloměr $\varrho(\mathbf{J}) > 2\cdot 10^3$, takže Robertsonův problém lze považovat za tuhý již na nepoměrně kratším intervalu délky řádově v jednotkách.

Úlohu jsme řešili dvěma matlabovskými programy určenými pro tuhé problémy: programem `ode23t` (TR metoda) a programem `ode15s` (metody BDFk, k=1,2,...,5). Délku kroku jsme řídili pomocí tolerancí $\varepsilon_r = 10^{-3}$, $\varepsilon_a = 10^{-6}$, činnost programu `ode15s` jsme omezili tak, aby pracoval jen s BDF metodami řádů 1, 2 a 3. Do následující tabulky jsme zapsali „statistiku“ výpočtu, tj. čísla `pk`, `pn`, `pf`, `pj`, `pr`, `ps`:

	pk	pn	pf	pj	pr	ps
ode23t	238	74	794	37	188	644
ode15s	245	15	504	11	67	458

Z tabulky plyne, že oba testované programy Robertsonův problém úspěšně vyřešily, program `ode15s` se podle statistiky jeví jako efektivnější.

2. Obyčejné diferenciální rovnice: okrajové úlohy

Okrajový problém pro soustavu ODR1 spočívá v určení funkce $\mathbf{y}(x)$, která v intervalu (a, b) splňuje diferenciální rovnici

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)) \tag{2.1}$$

a v koncových bodech intervalu (a, b) vyhovuje okrajové podmínce

$$\mathbf{r}(\mathbf{y}(a), \mathbf{y}(b)) = \mathbf{o}. \tag{2.2}$$

Stejně jako v kapitole 1.1 předpokládáme, že počet rovnic jakož i počet okrajových podmínek je roven d, tedy

$$\mathbf{y}(x) = (y_1(x), y_2(x), \ldots, y_d(x))^T, \qquad \mathbf{y}'(x) = (y_1'(x), y_2'(x), \ldots, y_d'(x))^T,$$
$$\mathbf{f}(x, \mathbf{y}(x)) = (f_1(x, \mathbf{y}(x)), f_2(x, \mathbf{y}(x)), \ldots, f_d(x, \mathbf{y}(x)))^T,$$
$$\mathbf{r}(\mathbf{y}(a), \mathbf{y}(b)) = (r_1(\mathbf{y}(a), \mathbf{y}(b)), r_2(\mathbf{y}(a), \mathbf{y}(b)), \ldots, r_d(\mathbf{y}(a), \mathbf{y}(b)))^T.$$

Ve speciálním případě, když

$$\mathbf{r}(\mathbf{y}(a), \mathbf{y}(b)) = \mathbf{y}(a) - \boldsymbol{\eta} = \mathbf{o} \quad \text{nebo-li} \quad \mathbf{y}(a) = \boldsymbol{\eta}, \tag{2.3}$$

přechází problém (2.1)–(2.2) v počáteční problém (1.4). Pokud je však v podmínkách (2.2) obsažena alespoň jedna složka vektoru $\mathbf{y}(a)$ a současně také alespoň jedna složka vektoru $\mathbf{y}(b)$, jde o problém okrajový. Právě tímto případem se budeme v dalším zabývat.

Podstatný rozdíl mezi počáteční a okrajovou úlohou spočívá v tom, že zatímco řešení úlohy s počátečními podmínkami existuje a je jediné pro dosti širokou třídu diferenciálních rovnic, u okrajové úlohy s velmi jednoduchou diferenciální rovnicí je možné, že řešení neexistuje nebo naopak, že řešení je nekonečně mnoho. Tuto skutečnost si ukažme na rovnici

$$y' = y, \quad \text{jejíž obecné řešení je} \quad y = Ce^x.$$

Odtud plyne, že pro okrajovou podmínku

(a) $y(0) = y(1)$ existuje jediné řešení $y = 0$,

(b) $e\,y(0) = y(1)$ existuje nekonečně mnoho řešení $y = Ce^x$, C libovolné,

(c) $e\,y(0) = y(1) + 1$ řešení neexistuje.

V kapitole 2.1 ukážeme, jak řešit okrajový problém pro soustavu ODR1 metodou střelby. V následujících kapitolách pak popíšeme tři nejznámější metody řešení okrajového problému pro ODR2: v kapitole 2.2 diferenční metodu, v kapitole 2.3 metodu konečných objemů a v kapitole 2.4 metodu konečných prvků.

2.1. Metoda střelby

Metoda střelby je založena na numerickém řešení počátečního problému

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \qquad \mathbf{y}(a) = \boldsymbol{\eta}, \tag{2.4}$$

za omezující podmínky

$$\mathbf{r}(\boldsymbol{\eta}, \mathbf{y}(b)) = \mathbf{o}. \tag{2.5}$$

Neznámá počáteční hodnota $\boldsymbol{\eta}$ je určena implicitně rovnicí (2.5).

Lichoběžníková metoda. Nechť $a = x_0 < x_1 < \cdots < x_N = b$ je dělení intervalu $\langle a, b\rangle$, $h_i = x_{i+1} - x_i$ je délka kroku a $h = \max_i h_i$. Přibližné řešení $\mathbf{y}_i$, $i = 0, 1, \ldots, N$, dostaneme jako řešení soustavy $d(N+1)$ rovnic

$$\begin{aligned} &\mathbf{y}_{i+1} = \mathbf{y}_i + \tfrac{1}{2}h_i[\mathbf{f}(x_i, \mathbf{y}_i) + \mathbf{f}(x_{i+1}, \mathbf{y}_{i+1})], \quad i = 0, 1, \ldots, N-1\,, \\ &\mathbf{r}(\mathbf{y}_0, \mathbf{y}_N) = 0\,. \end{aligned} \tag{2.6}$$

Soustava (2.6) je obecně nelineární. Její řešení se obvykle počítá pomocí nějaké varianty Newtonovy metody. Aby nastala konvergence, je třeba dodat dosti dobrou počáteční aproximaci $\mathbf{y}_i^{(0)} \approx \mathbf{y}(x_i)$, $i = 0, 1, \ldots, N$. Lichoběžníková metoda je řádu 2, tj. je-li funkce $\mathbf{f}$ dostatečně hladká, pro chybu platí

$$\mathbf{y}(x_i) - \mathbf{y}_i = O(h^2),$$

čímž se míní, že řádu $O(h^2)$ je každá složka vektoru $\mathbf{y}(x_i) - \mathbf{y}_i$.

Ve speciálním případě, když $\mathbf{f}(x, \mathbf{y})$ je lineární v proměnné $\mathbf{y}$ a $\mathbf{r}(\mathbf{u}, \mathbf{v})$ je lineární v obou proměnných $\mathbf{u}$ i $\mathbf{v}$, bude soustava rovnic (2.6) lineární. Nechť tedy

$$\begin{aligned} &\mathbf{y}' = \mathbf{A}(x)\mathbf{y} + \mathbf{q}(x)\,, \\ &\mathbf{B}_a\mathbf{y}(a) + \mathbf{B}_b\mathbf{y}(b) = \mathbf{c}\,. \end{aligned} \tag{2.7}$$

Soustava (2.6) je pak tvaru

$$\begin{aligned} &\left[-\mathbf{I} - \tfrac{1}{2}h_i\mathbf{A}(x_i)\right]\mathbf{y}_i + \left[\mathbf{I} - \tfrac{1}{2}h_i\mathbf{A}(x_{i+1})\right]\mathbf{y}_{i+1} = \tfrac{1}{2}h_i[\mathbf{q}(x_i) + \mathbf{q}(x_{i+1})]\,, \\ &\mathbf{B}_a\mathbf{y}_0 + \mathbf{B}_b\mathbf{y}_N = \mathbf{c}\,, \end{aligned}$$

kde $\mathbf{I}$ je jednotková matice. Maticový zápis této soustavy je

$$\begin{pmatrix} \mathbf{R}_0 & \mathbf{S}_0 & & & \\ & \mathbf{R}_1 & \mathbf{S}_1 & & \\ & & \ddots & \ddots & \\ & & & \mathbf{R}_{N-1} & \mathbf{S}_{N-1} \\ \mathbf{B}_a & & & & \mathbf{B}_b \end{pmatrix} \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_{N-1} \\ \mathbf{y}_N \end{pmatrix} = \begin{pmatrix} \mathbf{v}_0 \\ \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_{N-1} \\ \mathbf{c} \end{pmatrix}, \tag{2.8}$$

kde

$$\mathbf{R}_i = -\mathbf{I} - \tfrac{1}{2}h_i\mathbf{A}(x_i)\,, \quad \mathbf{S}_i = \mathbf{I} - \tfrac{1}{2}h_i\mathbf{A}(x_{i+1})\,, \quad \mathbf{v}_i = \tfrac{1}{2}h_i[\mathbf{q}(x_i) + \mathbf{q}(x_{i+1})]\,.$$

Simpsonova metoda. Matlab nabízí pro řešení okrajového problému (2.1)–(2.2) program `bvp4c`, který je založen na použití Simpsonovy formule

$$\begin{aligned} &\mathbf{y}_{i+1} = \mathbf{y}_i + \tfrac{1}{6}h_i[\mathbf{f}(x_i,\mathbf{y}_i) + 4\mathbf{f}(x_{i+1/2},\mathbf{y}_{i+1/2}) + \mathbf{f}(x_{i+1},\mathbf{y}_{i+1})], \\ \text{kde}\quad &x_{i+1/2} = x_i + \tfrac{1}{2}h_i, \quad \mathbf{y}_{i+1/2} = \tfrac{1}{2}(\mathbf{y}_i + \mathbf{y}_{i+1}) + \tfrac{1}{8}h_i[\mathbf{f}(x_i,\mathbf{y}_i) - \mathbf{f}(x_{i+1},\mathbf{y}_{i+1})]\,, \end{aligned} \tag{2.9}$$

Simpsonova formule (2.9) je řádu 4, tj. pro chybu platí

$$\mathbf{y}(x_i) - \mathbf{y}_i = O(h^4)\,.$$

Další informace o Simpsonově formuli (2.9) lze načerpat v [24], [13] a také v [16].

V případě lineární úlohy (2.7) řešíme soustavu lineárních rovnic (2.8), matice $\mathbf{R}_i$, $\mathbf{S}_i$ a vektory $\mathbf{v}_i$ získáme z formule (2.9), v níž klademe $\mathbf{f}(x,\mathbf{y}) = \mathbf{A}(x)\mathbf{y} + \mathbf{q}(x)$. Odvození vzorců pro $\mathbf{R}_i$, $\mathbf{S}_i$ a $\mathbf{v}_i$ ponecháváme čtenáři jako cvičení.

2.2. Diferenční metoda

Ve zbytku kapitoly 2 se budeme převážně zabývat lineární ODR2

$$-\,[p(x)u']' + q(x)u = f(x), \qquad x \in (0,\ell)\,. \tag{2.10}$$

Předpokládejme, že funkce p, p', q i f jsou spojité, dále že $p(x) \geq p_0 > 0$, $q(x) \geq 0$, a uvažujme nejdříve jednoduché okrajové podmínky

$$u(0) = g_0\,, \tag{2.11a}$$

$$u(\ell) = g_\ell\,. \tag{2.11b}$$

Za uvedených předpokladů má úloha (2.10)–(2.11) jediné řešení.

Úloha (2.10)–(2.11) může popisovat například problém *tahu–tlaku prutu*, tedy prutu namáhaného pouze tahem popřípadě tlakem. V tom případě je u posunutí střednicové čáry prutu, $p = EA$, kde E je Youngův modul pružnosti a A je plocha průřezu prutu, q je měrný odpor podloží, na němž prut spočívá, f je intenzita zatížení a g_0 a g_ℓ jsou předepsaná posunutí koncových bodů prutu.

Jinou aplikací, popsanou stejnou rovnicí a stejnými okrajovými podmínkami, je například *stacionární úloha vedení tepla v tyči*. Pak u je teplota, p je koeficient tepelné vodivosti, $f - qu$ je intenzita tepelných zdrojů, g_0 a g_ℓ jsou teploty koncových bodů tyče.

Úlohu (2.10)–(2.11) lze přeformulovat do tvaru (2.1)–(2.2), stačí položit

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} u \\ pu' \end{pmatrix}, \qquad \mathbf{f} = \begin{pmatrix} y_2/p \\ qy_1 - f \end{pmatrix}, \qquad \mathbf{r} = \begin{pmatrix} y_1(0) - g_0 \\ y_1(\ell) - g_\ell \end{pmatrix}, \tag{2.12}$$

a následně řešit metodou střelby, viz kapitola 2.1. My si ale vysvětlíme jinou techniku, známou jako *diferenční metoda* (stručně FDM podle anglického *finite difference method*).

FDM je klasická diskretizační metoda, zevrubně popsaná a analyzovaná v celé řadě publikací, viz např. [15], [29].

Nechť $0 = x_0 < x_1 < \cdots < x_N = \ell$ je dělení intervalu $\langle 0, \ell \rangle$ a $h_i = x_i - x_{i-1}$ je délka kroku. Body x_i nazýváme *uzly* a množinu $\{x_i\}_{i=0}^N$ uzlů nazýváme *sítí*. Pro jednoduchost se omezíme na ekvidistantní dělení, takže $h_i = h = \ell/N$ a $x_i = ih$, $i = 0, 1, \ldots, N$.

Splnění rovnice (2.10) budeme vyžadovat pouze ve vnitřních uzlech sítě, tj.

$$-(pu')'\big|_{x=x_i} + q_i u(x_i) = f_i, \quad i = 1, 2, \ldots, N-1\,, \tag{2.13}$$

kde $q_i = q(x_i)$ a $f_i = f(x_i)$. Člen $-(pu')'\big|_{x=x_i}$ nahradíme *diferenčním podílem*. Pomocí označení

$$x_{i-1/2} = x_i - \tfrac{1}{2}h\,,\ x_{i+1/2} = x_i + \tfrac{1}{2}h\,,\ p_{i-1/2} = p(x_{i-1/2})\,,\ p_{i+1/2} = p(x_{i+1/2})$$

lze přibližně položit

$$\begin{aligned} -(pu')'\big|_{x=x_i} &\doteq -\frac{pu'\big|_{x=x_{i+1/2}} - pu'\big|_{x=x_{i-1/2}}}{h} = -\frac{p_{i+1/2}u'(x_{i+1/2}) - p_{i-1/2}u'(x_{i-1/2})}{h} \\ &\doteq -\frac{p_{i+1/2}\dfrac{u(x_{i+1}) - u(x_i)}{h} - p_{i-1/2}\dfrac{u(x_i) - u(x_{i-1})}{h}}{h}\,. \end{aligned}$$

Užitím Taylorova rozvoje snadno ověříme, že tato aproximace je řádu $O(h^2)$, tj. že platí

$$-(pu')'\big|_{x=x_i} = \frac{-p_{i-1/2}u(x_{i-1}) + (p_{i-1/2} + p_{i+1/2})u(x_i) - p_{i+1/2}u(x_{i+1})}{h^2} + O(h^2)\,. \tag{2.14}$$

Po zanedbání chyby $O(h^2)$ tak z (2.14) a (2.13) dostáváme pro přibližné řešení $u_i \approx u(x_i)$ soustavu rovnic

$$\frac{-p_{i-1/2}u_{i-1} + (p_{i-1/2} + p_{i+1/2} + h^2 q_i)u_i - p_{i+1/2}u_{i+1}}{h^2} = f_i\,, \tag{2.15}$$

pro $i = 1, 2, \ldots, N-1$. Z okrajových podmínek máme

$$u_0 = g_0\,, \tag{2.16a}$$

$$u_N = g_\ell\,. \tag{2.16b}$$

To nám umožní dosadit do první rovnice $u_0 = g_0$ a člen $-p_{1/2}g_0/h^2$ převést na pravou stranu. Podobně v poslední rovnici položíme $u_N = g_\ell$ a člen $-p_{N-1/2}g_\ell/h^2$ převedeme na pravou stranu. Matice takto upravené soustavy rovnic (2.15) je třídiagonální, pozitivně definitní, diagonálně dominantní (pro $q_i > 0$ ryze). Tyto vlastnosti zaručují, že matice soustavy je regulární a soustava rovnic má jediné řešení.

Jsou-li funkce p, q a f dostatečně hladké, pak pro chybu platí

$$u(x_i) - u_i = O(h^2)\,, \tag{2.17}$$

přesná formulace a příslušný důkaz viz [29].

Okrajové podmínky s derivací. Okrajové podmínky (2.11) se nazývají *Dirichletovy*. V aplikacích se objevují ještě další typy okrajových podmínek. Podmínky

$$p(0)u'(0) = \alpha_0 u(0) - \beta_0, \tag{2.18a}$$

$$-p(\ell)u'(\ell) = \alpha_\ell u(\ell) - \beta_\ell \tag{2.18b}$$

se nazývají *Newtonovy* nebo také *Robinovy*. Pokud $\alpha_0 = 0$ resp. $\alpha_\ell = 0$, hovoříme o *Neumannově* okrajové podmínce. Aby byla zaručena jednoznačná existence řešení, předpokládejme $\alpha_0 \geq 0$, $\alpha_\ell \geq 0$, a pokud uvažujeme okrajové podmínky (2.18a) a (2.18b), pak předpokládejme navíc buďto $\alpha_0 > 0$ nebo $\alpha_\ell > 0$ nebo $q(x) \geq q_0 > 0$ alespoň na části intervalu $(0, \ell)$. Pokud $\alpha_0 = \alpha_\ell = 0$ a $q(x) = 0$ v $(0, \ell)$, pak úloha (2.10) – (2.18) buďto nemá řešení nebo má nekonečně mnoho řešení. Skutečně, integrací (2.10) a užitím (2.18) dostaneme nutnou podmínku existence řešení, tzv. *podmínku rovnováhy*

$$\int_0^\ell [f + (pu')'] \, \mathrm{d}x = \int_0^\ell f \, \mathrm{d}x + pu'\big|_{x=0}^{x=\ell} = \int_0^\ell f \, \mathrm{d}x + \beta_0 + \beta_\ell = 0 \,.$$

Pokud podmínka rovnováhy splněna není, řešení neexistuje. Je-li však podmínka rovnováhy splněna a u je řešení, pak je řešením také funkce $u + C$, kde C je libovolná konstanta.

Interpretujeme-li Newtonovy okrajové podmínky v úloze tahu–tlaku prutu, je pu' normálová síla, α_0, α_ℓ jsou tuhosti pružin v koncových bodech prutu a β_0, β_ℓ jsou zadané síly působící na koncích prutu. V úloze stacionárního vedení tepla je pu' tepelný tok, α_0, α_ℓ jsou koeficienty přestupu tepla a β_0, β_ℓ jsou zadané tepelné toky na okrajích tyče. Tepelné toky se často uvažují ve tvaru $\beta_0 = \alpha_0 u_0^e$, $\beta_\ell = \alpha_\ell u_\ell^e$, kde u_0^e resp. u_ℓ^e je vnější teplota okolo levého resp. pravého konce tyče.

V případě zadané okrajové podmínky (2.18a) eventuálně (2.18b) musíme sestavit rovnici umožňující výpočet neznámé u_0 eventuálně u_N. Ukažme si to třeba pro případ předepsané okrajové podmínky (2.18a).

Pomůžeme si tak, že budeme požadovat, aby rovnice (2.10) platila také v levém krajním uzlu $x_0 = 0$, tj. aby rovnice (2.13) byla splněna rovněž pro $i = 0$.

Člen $-(pu')'\big|_{x=x_0}$ vyjádříme nejdříve pomocí jednostranné diference a pak pomocí centrální diference a vztahu (2.18a), tj.

$$-(pu')'\big|_{x=x_0} = -\frac{pu'\big|_{x=x_{1/2}} - pu'\big|_{x=x_0}}{\frac{1}{2}h} + O(h) =$$

$$\frac{-p_{1/2}\dfrac{u(x_1) - u(x_0)}{h} + \alpha_0 u(x_0) - \beta_0}{\frac{1}{2}h} + O(h) \,.$$

Zanedbáme-li chyby, dostaneme rovnici pro neznámou u_0

$$\frac{(p_{1/2} + h\alpha_0 + \frac{1}{2}h^2 q_0)u_0 - p_{1/2}u_1}{h^2} = \frac{1}{h}\beta_0 + \frac{1}{2}f_0 \,. \tag{2.19a}$$

V případě okrajové podmínky (2.18b) postupujeme obdobně, tj. požadujeme splnění rovnice (2.13) také pro $i = N$, a po úpravách obdržíme rovnici pro neznámou u_N

$$\frac{-p_{N-1/2}u_{N-1} + (p_{N-1/2} + h\alpha_\ell + \frac{1}{2}h^2 q_N)u_N}{h^2} = \frac{1}{h}\beta_\ell + \frac{1}{2}f_N \,. \tag{2.19b}$$

Je-li předepsána okrajová podmínka (2.18a), zapíšeme jako první rovnici (2.19a), pak následují rovnice (2.15) pro $i = 1, 2, \ldots, N-1$, a je-li předepsána okrajová podmínka (2.18b), připojíme jako poslední rovnici (2.19b). Matice výsledné soustavy rovnic je opět třídiagonální, pozitivně definitní a diagonálně dominantní. I když se při odvození rovnic (2.19) dopouštíme chyby řádu $O(h)$, pro chybu přibližného řešení platí zase vztah (2.17).

Rovnice s konvekčním členem. V aplikacích často vzniká potřeba řešit poněkud obecnější rovnici

$$-\left[p(x)u' - r(x)u\right]' + q(x)u = f(x), \qquad x \in (0, \ell)\,. \tag{2.20}$$

Tato rovnice popisuje například *transport chemické příměsi v tekutině*. V tom případě je u koncentrace příměsi v tekutině, $r = \varrho v$, kde ϱ je hustota tekutiny a v její rychlost, p je koeficient difúze a $f - qu$ je intenzita objemového zdroje příměsi. Rovnici (2.20) lze použít také k popisu *teplotního pole v tekutině*. Pak u je teplota, $r = c\varrho v$, kde c je tepelná kapacita, ϱ hustota a v rychlost tekutiny, p je tepelná vodivost a $f - qu$ je intenzita vnitřních tepelných zdrojů. S přihlédnutím k typické fyzikální interpretaci říkáme, že $-(pu')'$ je *difúzní* člen, $(ru)'$ je *konvekční člen* a $f - qu$ je *zdrojový člen*.

Pokud jde o okrajové podmínky, budeme postupovat takto: „na vtoku“, tj. bodě $x = 0$ pro $r(0) > 0$ resp. v bodě $x = \ell$ pro $r(\ell) < 0$, můžeme předpokládat, že veličinu $u(x)$ známe, a proto její hodnotu předepíšeme prostřednictvím Dirichletovy okrajové podmínky. „Na výtoku“, tj. bodě $x = 0$ pro $r(0) \le 0$ resp. v bodě $x = \ell$ pro $r(\ell) \ge 0$, můžeme zadat jak Dirichletovu tak Newtonovu okrajovou podmínku.

V [12] je dokázáno, že rovnice (2.20) doplněná o okrajové podmínky má jediné řešení třeba tehdy, když kromě dřívějších předpokladů navíc platí

$$r'(x) \ge 0, \quad \alpha_0 - \tfrac{1}{2}r(0) \ge 0, \quad \alpha_\ell + \tfrac{1}{2}r(\ell) \ge 0.$$

Jsou-li na obou okrajích $x = 0$ i $x = \ell$ předepsány Dirichletovy okrajové podmínky, stačí předpokládat

$$r'(x) \ge 0, \quad x \in \langle 0, \ell \rangle. \tag{2.21}$$

Zadáme-li Newtonovu okrajovou podmínku jen na výtoku, pak podmínka $\alpha_0 - \frac{1}{2}r(0) \ge 0$ resp. $\alpha_\ell + \frac{1}{2}r(\ell) \ge 0$ zřejmě platí, takže opět stačí předpokládat jen (2.21).

Je-li $r(0) \le 0$ a $r(\ell) \le 0$, pak levý okraj $x = 0$ je výtok a pravý okraj $x = \ell$ je vtok. Je-li $r(0) \le 0$ a $r(\ell) > 0$, pak oba okraje představují výtok. Je-li $r(0) > 0$, pak podle (2.21) také $r(\ell) > 0$, takže $x = 0$ je vtok a $x = \ell$ je výtok. Oba konce nemohou být vtokem: $r(0) > 0$ a $r(\ell) < 0$ není podle (2.21) možné.

V dynamice tekutin se ukazuje, že v nestlačitelné tekutině rychlost $\mathbf{v} = (v_1, v_2, v_3)^T$ splňuje *rovnici kontinuity* $\operatorname{div} \mathbf{v} = 0$. V jedné dimenzi tedy $v' = 0$, takže v je konstanta. Volba konstantní funkce r tedy představuje fyzikálně opodstatněnou možnost. Podmínka (2.21) je pro konstantní funkci r triviálně splněna.

Věnujme se tedy diskretizaci. Konvekční člen vyjádříme ve vnitřních uzlech pomocí centrální diference

$$(ru)'\big|_{x=x_i} = \frac{ru|_{x_{i+1/2}} - ru|_{x_{i-1/2}}}{h} + O(h^2) \tag{2.22}$$

a v koncových uzlech pomocí jednostranné diference

$$\begin{aligned}(ru)'\big|_{x=x_0} &= \frac{ru|_{x_{1/2}} - ru|_{x_0}}{\frac{1}{2}h} + O(h),\\ (ru)'\big|_{x=x_N} &= \frac{ru|_{x_N} - ru|_{x_{N-1/2}}}{\frac{1}{2}h} + O(h)\,.\end{aligned} \tag{2.23}$$

V (2.22) a (2.23) vyjádříme konvekční tok ru ve středech $x_{i+1/2}$ úseček $\langle x_i, x_{i+1}\rangle$ interpolací jako aritmetický průměr hodnot u v koncových bodech,

$$ru|_{x_{i+1/2}} = \tfrac{1}{2}r_{i+1/2}[u(x_i) + u(x_{i+1})] + O(h^2), \qquad i = 0, 1, \ldots, N-1. \tag{2.24}$$

Po zanedbání chybových členů v (2.22)–(2.24), obdržíme soustavu rovnic, jejíž tvar vyjádříme prostřednictvím dříve odvozených rovnic (2.15) a (2.19) takto: na levou stranu rovnice

$$\left.\begin{array}{l}\text{(2.19a)}\\ \text{(2.15)}\\ \text{(2.19b)}\end{array}\right\} \quad \text{přidáme člen} \quad \left\{\begin{array}{l}\left[\frac{1}{2}r_{1/2}(u_0+u_1) - r_0u_0\right]/h\,,\\ \frac{1}{2}[r_{i+1/2}(u_i+u_{i+1}) - r_{i-1/2}(u_{i-1}+u_i)]/h\,,\\ \left[r_Nu_N - \frac{1}{2}r_{N-1/2}(u_{N-1}+u_N)\right]/h.\end{array}\right. \tag{2.25}$$

Matice takto vzniklé soustavy rovnic je třídiagonální a nesymetrická. Dá se ukázat, že když

$$\tfrac{1}{2}h|r_0| < p_{1/2}, \quad \tfrac{1}{2}h|r_{i+1/2}| < p_{i+1/2}, \ \ i = 0, 1, \ldots, N-1, \quad \tfrac{1}{2}h|r_N| < p_{N-1/2}, \tag{2.26}$$

pak je matice soustavy regulární. Pro dostatečně jemné dělení intervalu $\langle 0, \ell\rangle$ bude podmínka (2.26) jistě splněna. Pro chybu opět platí (2.17).

Dominantní konvekce. Podmínka (2.26) může být značně omezující v případě, kdy konvekční koeficient výrazně převažuje nad koeficientem difúzním, tedy pro $|r| \gg p$. Pak je účelné vyjádřit konvekční tok ru ve středech $x_{i+1/2}$ úseček $\langle x_i, x_{i+1}\rangle$ takto:

$$ru|_{x_{i+1/2}} = \begin{cases} r_{i+1/2}u(x_i) + O(h) & \text{pro } r_{i+1/2} \geq 0,\\ r_{i+1/2}u(x_{i+1}) + O(h) & \text{pro } r_{i+1/2} < 0.\end{cases} \tag{2.27}$$

Aproximace (2.27) je z fyzikálního hlediska přirozená: informaci o řešení u v uzlu $x_{i+1/2}$ čerpáme ze znalosti řešení proti „proudu“, proti „větru“: pro $r_{i+1/2} > 0$ „fouká zleva“, proto použijeme hodnotu $u(x_i)$ v uzlu x_i ležícím nalevo od bodu $x_{i+1/2}$, pro $r_{i+1/2} < 0$ „fouká zprava“ a proto použijeme hodnotu $u(x_{i+1})$ v uzlu x_{i+1} ležícím napravo od bodu $x_{i+1/2}$. Jednostrannou aproximaci konvekčního toku podle (2.27) nazýváme *upwind aproximací*. Pomocí označení

$$a^+ = \max(a, 0), \qquad a^- = \min(a, 0), \quad \text{kde } a \text{ je libovolné číslo,}$$

lze (2.27) zapsat v kompaktním tvaru

$$ru|_{x_{i+1/2}} = r^+_{i+1/2}u(x_i) + r^-_{i+1/2}u(x_{i+1}) + O(h), \quad i = 0, 1, \ldots, N-1. \tag{2.28}$$

Po zanedbání chybových členů v (2.22), (2.23) a (2.28) obdržíme výslednou soustavu rovnic, jejíž tvar vyjádříme prostřednictvím dříve odvozených rovnic (2.15) a (2.19) takto: na levou stranu rovnice

$$\left.\begin{array}{l}(2.19a)\\(2.15)\\(2.19b)\end{array}\right\}\ \text{přidáme člen}\ \left\{\begin{array}{l}[(r^+_{1/2}-r_0)u_0+r^-_{1/2}u_1]/h,\\ [r^+_{i+1/2}u_i+r^-_{i+1/2}u_{i+1}]/h-[r^+_{i-1/2}u_{i-1}+r^-_{i-1/2}u_i]/h,\\ [-r^+_{N-1/2}u_{N-1}+(r_N-r^-_{N-1/2})u_N]/h\,.\end{array}\right. \tag{2.29}$$

Matice takto vzniklé soustavy je regulární nezávisle na jemnosti dělení intervalu $\langle 0,\ell\rangle$. Pro chybu však platí jen

$$u(x_i)-u_i=O(h)\,. \tag{2.30}$$

Pro dosažení přesnosti řádu $O(h^2)$ je třeba používat přesnější upwind aproximaci konvekčního toku, třeba

$$ru|_{x_{i+1/2}}=r^+_{i+1/2}\left[\tfrac{3}{2}u(x_i)-\tfrac{1}{2}u(x_{i-1})\right]+r^-_{i+1/2}\left[\tfrac{3}{2}u(x_{i+1})-\tfrac{1}{2}u(x_{i+2})\right]+O(h^2), \tag{2.28'}$$

$i=0,1,\ldots,N-1$, kde $u(x_{-1}):=2u(x_0)-u(x_1)$, $u(x_{N+1}):=2u(x_N)-u(x_{N-1})$.

Příklad 2.1. Zabývejme se řešením modelové úlohy

$$-\varepsilon u''+u'=0\quad\text{pro}\quad x\in(0,1),\quad u(0)=0,\ u(1)=1,$$

kde $\varepsilon>0$ je konstanta. Přesné řešení

$$u(x)=\frac{1-e^{x/\varepsilon}}{1-e^{1/\varepsilon}}$$

je rostoucí funkce.

Diskretizací konvekčního členu pomocí centrální diference (2.25_2), tj. pomocí druhého vztahu v (2.25), dostaneme rovnici

$$-\varepsilon\,\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}+\frac{u_{i+1}-u_{i-1}}{2h}=0\,,$$

kterou lze při označení $\kappa:=h/\varepsilon$ zapsat ekvivalentně ve tvaru

$$-(1+\tfrac{1}{2}\kappa)u_{i-1}+2u_i-(1-\tfrac{1}{2}\kappa)u_{i+1}=0\,.$$

Pro $\kappa=2$ dostaneme $u_i=u_{i-1}$, takže řešení je $u_i=0$ pro $i<N$ a $u_N=1$. Pro $\kappa\neq 2$ snadným výpočtem ověříme, že diferenční rovnici vyhovuje řešení

$$u_i=C_1+C_2\left[\frac{2+\kappa}{2-\kappa}\right]^i,$$

kde C_1 a C_2 jsou konstanty, které určíme z okrajových podmínek. Pro $\kappa>2$ přibližné řešení u_i osciluje okolo C_1, což je v rozporu s chováním přesného řešení u, rostoucí posloupnost $\{u_i\}_{i=0}^N$ dostaneme jen pro $|\kappa|<2$, což je podmínka (2.26) pro $p=\varepsilon$, $r=1$.

Naši modelovou úlohu vyřešíme také upwind technikou. Konvekční člen nahradíme levostrannou diferencí podle (2.29_2) a dostaneme rovnici

$$-\varepsilon\,\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}+\frac{u_i-u_{i-1}}{h}=0.$$

Diferenční rovnici vyhovuje řešení

$$u_i=C_1+C_2(1+\kappa)^i,$$

které je rostoucí nezávisle na velikosti κ. □

2.3. Metoda konečných objemů

(stručně FVM podle anglického *finite volume method*) se používá především pro řešení problémů proudění ve více prostorových proměnných, např. viz [26], [8]. Princip této metody však lze objasnit i pro jednodimenzionální úlohu popsanou diferenciální rovnicí (2.20) a okrajovými podmínkami (2.11) a (2.0).

Ke každému uzlu x_i přiřadíme *konečný objem* B_i (stručně *buňku*) takto: pro vnitřní uzly $B_i=\langle x_{i-1/2},x_{i+1/2}\rangle$, $i=1,2,\dots,N-1$, a pro koncové uzly $B_0=\langle 0,x_{1/2}\rangle$, $B_N=\langle x_{N-1/2},\ell\rangle$. Zřejmě $\langle 0,\ell\rangle=\bigcup_{i=0}^{N}B_i$. Integrací rovnice (2.20) přes buňku B_i dostaneme *bilanční rovnici*

$$\int_{B_i}-[pu'-ru]'\,\mathrm{d}x+\int_{B_i}qu\,\mathrm{d}x=\int_{B_i}f\,\mathrm{d}x\,. \tag{2.31}$$

Předpokládejme nejdříve, že B_i přísluší vnitřnímu uzlu. Pak dostaneme

$$-[pu'-ru]_{x=x_{i-1/2}}^{x=x_{i+1/2}}+\int_{x_{i-1/2}}^{x_{i+1/2}}qu\,\mathrm{d}x=\int_{x_{i-1/2}}^{x_{i+1/2}}f\,\mathrm{d}x\,. \tag{2.32}$$

Difúzní tok aproximujeme pomocí centrální diference,

$$p_{i-1/2}u'(x_{i-1/2})=p_{i-1/2}\frac{u(x_i)-u(x_{i-1})}{h}+O(h^2),\quad i=1,2,\dots,N-1, \tag{2.33}$$

a konvekční tok aproximujeme pomocí centrální aproximace (2.24) resp. upwind aproximace (2.28). Integrály v rovnici (2.32) spočteme obdélníkovou formulí,

$$\int_{x_{i-1/2}}^{x_{i+1/2}}qu\,\mathrm{d}x=hq_iu(x_i)+O(h^3)\,,\qquad\int_{x_{i-1/2}}^{x_{i+1/2}}f\,\mathrm{d}x=hf_i+O(h^3).$$

Zanedbáme-li chyby, dostaneme aproximaci bilanční rovnice (2.32) ve tvaru (2.25_2) resp. (2.29_2).

Je-li v koncovém bodě $x=0$ resp. $x=\ell$ předepsána Newtonova okrajová podmínka (2.18a) resp. (2.18b), je třeba uvážit bilanční rovnici (2.31) také pro buňku B_0 resp. B_N. Tak například pro buňku B_0 máme

$$-(pu'-ru)\big|_{x=x_0}^{x=x_{1/2}}+\int_{x_0}^{x_{1/2}}qu\,\mathrm{d}x=\int_{x_0}^{x_{1/2}}f\,\mathrm{d}x\,.$$

Difúzní tok v bodě $x_{1/2}$ aproximujeme centrální diferencí (2.33), konvekční tok v bodě $x_{1/2}$ aproximujeme pomocí (2.24) resp. (2.28), člen $p(0)u'(0)$ vyjádříme pomocí okrajové podmínky (2.18a) a integrály spočteme levostrannou obdélníkovou formulí,

$$\int_{x_0}^{x_{1/2}} qu \,\mathrm{d}x = \tfrac{1}{2}hq_0u(x_0) + O(h^2)\,, \qquad \int_{x_0}^{x_{1/2}} f \,\mathrm{d}x = \tfrac{1}{2}hf_0 + O(h^2)\,.$$

Zanedbáme-li chyby, dostaneme rovnici (2.25_1) resp. (2.29_1). Rovnici (2.25_3) resp. (2.29_3) odvodíme obdobně z bilanční rovnice (2.31) zapsané pro buňku B_N.

Metodou konečných objemů jsme tedy dostali stejné rovnice jako rovnice odvozené metodou diferenční. Přednosti metody konečných objemů ve srovnání s metodou diferenční vyniknou až při řešení parciálních diferenciálních rovnic ve více prostorových proměnných.

2.4. Metoda konečných prvků

Nejuniverzálnější metodou diskretizace okrajových úloh je *metoda konečných prvků* (stručně FEM podle anglického *finite element method*, česky MKP). V úlohách mechaniky pevné fáze je to jednoznačně nejpoužívanější metoda. I když přednosti MKP lze plně ocenit teprve u úloh ve dvou a třech prostorových proměnných, podstatu metody lze objasnit i na jednodimenzionální úloze. Z řady publikací věnovaných MKP zmiňme např. [9], [30], [2].

Východiskem pro diskretizaci metodou konečných prvků je slabá formulace okrajové úlohy. Proto si teď ukážeme, jak z *klasické formulace* (2.20), (2.11), (2.18) formulaci slabou dostaneme. Nejdříve zavedeme následující

Označení. Symbolem $C\langle 0,\ell\rangle$ budeme značit prostor všech funkcí, které jsou v intervalu $\langle 0,\ell\rangle$ spojité, a symbolem $C^k\langle 0,\ell\rangle$ pak prostor všech funkcí, které jsou v intervalu $\langle 0,\ell\rangle$ spojité spolu se svými derivacemi až do řádu k včetně (C podle anglického *continuous*).

Říkáme, že bod a je pro funkci $f(x)$ *bodem nespojitosti prvního druhu*, existuje-li v a konečná limita zprava i zleva [označme tyto limity $f(a+0)$ resp. $f(a-0)$] a je-li $f(a+0) \neq f(a-0)$.

Funkce $f(x)$ definovaná v intervalu $\langle 0,\ell\rangle$ se nazývá *po částech spojitá v intervalu* $\langle 0,\ell\rangle$, je-li v $\langle 0,\ell\rangle$ spojitá s výjimkou konečného počtu bodů, v nichž má nespojitost prvního druhu.

Prostor funkcí po částech spojitých v intervalu $\langle 0,\ell\rangle$ označíme $PC\langle 0,\ell\rangle$ (PC podle anglického *piecewise continuous*). Symbolem $PC^k\langle 0,\ell\rangle$ značíme prostor funkcí, které jsou v intervalu $\langle 0,\ell\rangle$ spojité spolu se svými derivacemi až do řádu $k-1$ včetně, a jejichž k-tá derivace je v intervalu $\langle 0,\ell\rangle$ po částech spojitá.

V dalším budeme používat zejména prostor $C^1\langle 0,\ell\rangle$ funkcí, které jsou v intervalu $\langle 0,\ell\rangle$ spojité spolu se svou první derivací, a prostor $PC^1\langle 0,\ell\rangle$ funkcí, které jsou v intervalu $\langle 0,\ell\rangle$ spojité a mají v něm po částech spojitou první derivaci.

Slabá formulace. Začneme tím, že zavedeme pojem *testovací funkce*: funkci $v \in C^1\langle 0,\ell\rangle$ nazveme testovací, jestliže $v = 0$ v tom krajním bodě intervalu $\langle 0,\ell\rangle$, v němž je předepsána Dirichletova okrajová podmínka. Pro konkrétnost se omezíme na okrajové podmínky (2.11a) a (2.18b), takže testovací funkce v splňuje $v(0) = 0$. Násobme rovnici (2.10)

testovací funkcí v a integrujme přes $\langle 0, \ell\rangle$. Integrací per-partes členu $\int_0^\ell [-(pu')']v \,\mathrm{d}x$ a následným užitím okrajové podmínky (2.18b) a vztahu $v(0) = 0$ obdržíme

$$\int_0^\ell fv \,\mathrm{d}x = \int_0^\ell [-(pu' - ru)' + qu]\, v \,\mathrm{d}x = -(pu' - ru)v\big|_{x=0}^{x=\ell} +$$

$$+\int_0^\ell [(pu' - ru)v' + quv] \,\mathrm{d}x = [\alpha_\ell u(\ell) - \beta_\ell + r(\ell)u(\ell)\,]v(\ell) + \int_0^\ell [(pu' - ru)v' + quv] \,\mathrm{d}x.$$

Odvodili jsme tedy, že řešení u úlohy (2.10), (2.11a) a (2.18b) musí splňovat kromě Dirichletovy okrajové podmínky $u(0) = g_0$ také rovnost

$$\int_0^\ell [(pu' - ru)v' + quv] \,\mathrm{d}x + [\alpha_\ell + r(\ell)]u(\ell)v(\ell) = \int_0^\ell fv \,\mathrm{d}x + \beta_\ell v(\ell) \tag{2.34}$$

pro každou funkci $v \in C^1\langle 0, \ell\rangle$, $v(0) = 0$. Okrajová podmínka (2.18b) Newtonova typu, která se stala součástí integrální rovnice (2.34) a je tak automaticky splněna, se nazývá *přirozenou okrajovou podmínkou*. Dirichletovu okrajovou podmínku (2.11a), která součástí rovnice (2.34) není a jejíž explicitní splnění proto musíme vyžadovat, nazýváme *podstatnou* nebo také *hlavní okrajovou podmínkou*. Rovnice (2.34) je dobře definována i v případě, kdy funkce u a v jsou z prostoru $X \equiv PC^1\langle 0, \ell\rangle$. Testovací funkce pak volíme z *prostoru* $V = \{v \in X |\, v(0) = 0\}$ *testovacích funkcí* a řešení u z *množiny* $W = \{v \in X |\, v(0) = g_0\}$ *přípustných řešení*. Dále označíme

$$\begin{aligned} a(u, v) &= \int_0^\ell [(pu' - ru)v' + quv] \,\mathrm{d}x + [\alpha_\ell + r(\ell)]u(\ell)v(\ell), \\ L(v) &= \int_0^\ell fv \,\mathrm{d}x + \beta_\ell v(\ell). \end{aligned} \tag{2.35}$$

Pak úlohu

$$\text{najít } u \in W \text{ splňující } a(u, v) = L(v) \quad \forall v \in V \tag{2.36}$$

nazýváme *slabou formulací* problému (2.10), (2.11a), (2.18b). Řešení úlohy (2.36) nazveme *slabým řešením*. Slabá formulace je obecnější než formulace klasická, neboť klade nižší nároky na hladkost dat:

$$\text{klasická formulace:} \quad p, r \in C^1\langle 0, \ell\rangle, q, f \in C\langle 0, \ell\rangle\,, \tag{2.37a}$$

$$\text{slabá formulace:} \quad p, r, q, f \in PC\langle 0, \ell\rangle\,. \tag{2.37b}$$

Jestliže $p, r', q, f \in PC\langle 0, \ell\rangle$, $p \geq p_0 > 0$, $q + \frac{1}{2}r' \geq 0$, $\alpha_\ell + \frac{1}{2}r(\ell) \geq 0$, pak úloha (2.36) má jediné slabé řešení, viz [12].

Ukázali jsme si, že klasické řešení je vždy také řešení slabé, viz odvození rovnice (2.34). Opak obecně neplatí, tj. slabé řešení nemusí být řešení klasické. Jsou-li však funkce p, r, q a f dostatečně hladké, konkrétně platí-li podmínky (2.37a), pak lze dokázat, že slabé řešení $u \in C^2\langle 0, \ell\rangle$ je řešení klasické.

Slabá formulace má v úloze tahu–tlaku prutu (kdy $r = 0$) význam *principu virtuálních posunutí* a samotné testovací funkce $v \in V$ mají význam virtuálních posunutí δu přípustných řešení $u \in W$. Slabá formulace je tedy zcela přirozená, neboť konkrétně pro úlohu tahu–tlaku prutu popisuje základní fyzikální zákon mechaniky kontinua.

Slabá formulace pro všechny kombinace okrajových podmínek. Uveďme si tvar V, W, $a(u,v)$ a $L(v)$ pro všechny možné kombinace okrajových podmínek.

(DD) Okrajové podmínky (2.11a), (2.11b)

$$V = \{v \in X \,|\, v(0) = v(\ell) = 0\}, \quad W = \{v \in X \,|\, v(0) = g_0,\, v(\ell) = g_\ell\}, \tag{2.38-DD}$$
$$a(u,v) = \int_0^\ell [(pu' - ru)v' + quv]\,\mathrm{d}x, \quad L(v) = \int_0^\ell fv\,\mathrm{d}x\,.$$

(DN) Okrajové podmínky (2.11a), (2.18b)

$$V = \{v \in X \,|\, v(0) = 0\}, \quad W = \{v \in X \,|\, v(0) = g_0\}, \tag{2.38-DN}$$
$$a(u,v) = \int_0^\ell [(pu' - ru)v' + quv]\,\mathrm{d}x + [\alpha_\ell + r(\ell)]u(\ell)v(\ell), \quad L(v) = \int_0^\ell fv\,\mathrm{d}x + \beta_\ell v(\ell)\,.$$

(ND) Okrajové podmínky (2.18a), (2.11b)

$$V = \{v \in X \,|\, v(\ell) = 0\}, \quad W = \{v \in X \,|\, v(\ell) = g_\ell\}, \tag{2.38-ND}$$
$$a(u,v) = \int_0^\ell [(pu' - ru)v' + quv]\,\mathrm{d}x + [\alpha_0 - r(0)]u(0)v(0), \quad L(v) = \int_0^\ell fv\,\mathrm{d}x + \beta_0 v(0)\,.$$

(NN) Okrajové podmínky (2.18a), (2.18b)

$$V = X, \quad W = X, \tag{2.38-NN}$$
$$a(u,v) = \int_0^\ell [(pu' - ru)v' + quv]\,\mathrm{d}x + [\alpha_0 - r(0)]u(0)v(0) + [\alpha_\ell + r(\ell)]u(\ell)v(\ell),$$
$$L(v) = \int_0^\ell fv\,\mathrm{d}x + \beta_0 v(0) + \beta_\ell v(\ell)\,.$$

Ve všech čtyřech případech je zaručena jednoznačná existence slabého řešení úlohy (2.36), pokud $p, r', q, f \in PC\langle 0,\ell\rangle$, $p \geq p_0 > 0$, $q + \frac{1}{2}r' \geq 0$, $\alpha_0 - \frac{1}{2}r(0) \geq 0$, $\alpha_\ell + \frac{1}{2}r(\ell) \geq 0$. V případě úlohy (2.38-NN) je třeba navíc předpokládat $\alpha_0 - \frac{1}{2}r(0) > 0$ nebo $\alpha_\ell + \frac{1}{2}r(\ell) > 0$ nebo $q + \frac{1}{2}r' \geq q_0 > 0$ alespoň na části $\langle 0,\ell\rangle$, viz [12].

Diskretizace užitím lineárního prvku. Na intervalu $\langle 0,\ell\rangle$ zvolíme dělení $0 = x_0 < x_1 < \cdots < x_N = \ell$ a na každé úsečce $\langle x_{i-1}, x_i\rangle$ délky $h_i = x_i - x_{i-1}$ hledáme přibližné řešení $U(x)$ ve tvaru lineárního polynomu procházejícího body $[x_{i-1}, u_{i-1}]$ a $[x_i, u_i]$, takže

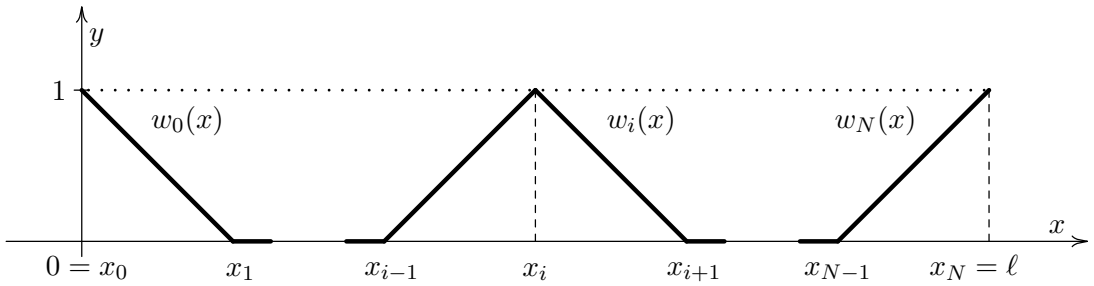
$$U(x) = U_{i-1}w_{i-1}(x) + U_i w_i(x), \quad \text{kde } w_{i-1}(x) = \frac{x_i - x}{h_i}, \quad w_i(x) = \frac{x - x_{i-1}}{h_i}.$$

Funkce $U(x)$ je tedy na celém intervalu $\langle 0, \ell\rangle$ *spojitou po částech lineární funkcí* určenou předpisem

$$U(x) = \sum_{i=0}^{N} U_i w_i(x), \tag{2.39}$$

kde $w_i(x)$ se jsou tzv. *bázové funkce*, lineární na každé úsečce $\langle x_{k-1}, x_k\rangle$ a takové, že

$$w_i(x_j) = \begin{cases} 1 \text{ pro } i = j, \\ 0 \text{ pro } i \neq j. \end{cases}$$



Obr. 5.1. Lineární Lagrangeovy bázové funkce

Úsečku $\langle x_{i-1}, x_i\rangle$, na které je definována lineární funkce určená svými hodnotami U_{i-1} resp. U_i v uzlech x_{i-1} resp. x_i, nazýváme *Lagrangeovým lineárním prvkem* nebo také *Lagrangeovým lineárním elementem*. Délku největšího dílku dělení $\{x_i\}_{i=0}^N$ označíme jako h, tj. $h = \max_{1\le i\le N} h_i$. Nechť X_h je prostor všech spojitých po částech lineárních funkcí tvaru $\sum_{i=0}^N \Theta_i w_i(x)$, kde $\{\Theta_i\}_{i=0}^N$ jsou libovolná reálná čísla. Zřejmě $X_h \subset X$. Nechť $V_h = V \cap X_h$ a $W_h = W \cap X_h$. Pak přibližné řešení U, tzv. *MKP řešení*, obdržíme z *diskrétní slabé formulace*

$$\text{najít } U \in W_h \text{ splňující } a_h(U, v) = L_h(v) \quad \forall v \in V_h. \tag{2.40}$$

Přitom index h v $a_h(U, v)$ resp. $L_h(v)$ značí, že integrál $\int_0^\ell [(pU' - r)v' + qUv]\,\mathrm{d}x$ v $a(U, v)$ resp. $\int_0^\ell fv\,\mathrm{d}x$ v $L(v)$ počítáme numericky kvadraturní formulí řádu alespoň jedna.

V dalším budeme pro konkrétnost uvažovat slabou formulaci (2.38-NN). Označíme-li $Q^i(\varphi)$ přibližně spočtenou hodnotu $\int_{x_{i-1}}^{x_i} \varphi\,\mathrm{d}x$, je

$$a_h(U, v) = \sum_{i=1}^{N} Q^i([pU' - rU]v' + qUv) + [\alpha_0 - r(0)]U(x_0)v(x_0) + [\alpha_\ell + r(\ell)]U(x_\ell)v(x_\ell),$$

$$L_h(v) = \sum_{i=1}^{N} Q^i(fv) + \beta_0 v(x_0) + \beta_\ell v(x_N)\,. \tag{2.41}$$

Nechť $v(x) = \sum_{i=0}^{N} \Theta_i w_i(x) \in V_h$ je libovolná testovací funkce (tj. $\Theta_i = v(x_i)$ je libovolné číslo) a $U(x) = \sum_{j=0}^{N} \Delta_j w_j(x)$ je MKP řešení (tj. $\Delta_j = U(x_j)$). Pak z (2.40) pro úlohu (2.38-NN) plyne

$$\begin{aligned} 0 = a_h(U,v) - L_h(v) = a_h\left(\sum_{j=0}^{N}\Delta_j w_j, \sum_{i=0}^{N}\Theta_i w_i\right) - L_h\left(\sum_{i=0}^{N}\Theta_i w_i\right) = \\ = \sum_{i=0}^{N}\Theta_i\left[\sum_{j=0}^{N} a_h(w_j,w_i)\Delta_j - L_h(w_i)\right] = \boldsymbol{\theta}^T\left[\mathbf{K}\boldsymbol{\Delta} - \mathbf{F}\right], \end{aligned} \tag{2.42}$$

kde $\boldsymbol{\theta} = (\Theta_0, \Theta_1, \ldots, \Theta_N)^T$, $\mathbf{K} = \{k_{ij}\}_{i,j=0}^{N}$ pro $k_{ij} = a_h(w_j, w_i)$, $\boldsymbol{\Delta} = (\Delta_0, \Delta_1, \ldots, \Delta_N)^T$ a $\mathbf{F} = (F_0, F_1, \ldots, F_N)^T$ pro $F_i = L_h(w_i)$. Protože $\boldsymbol{\theta}$ je libovolný vektor, musí platit

$$\mathbf{K}\boldsymbol{\Delta} = \mathbf{F}. \tag{2.43}$$

Matice $\mathbf{K}$ bývá označována jako *matice tuhosti* a vektor $\mathbf{F}$ jako *vektor zatížení*. Toto pojmenování pochází z prvních aplikací MKP v pružnosti a stalo se univerzálním označením pro matici soustavy a pro vektor pravé strany v soustavě rovnic vzniklé diskretizací jakékoliv úlohy MKP.

Soustavu rovnic (2.43) sestavíme pomocí tzv. *elementárních matic tuhosti* $\mathbf{K}^i$ a *elementárních vektorů zatížení* $\mathbf{F}^i$ příslušných elementům $\langle x_{i-1}, x_i\rangle$, $i = 1, 2, \ldots, N$. Pomocí obdélníkové formule vyjádříme

$$Q^i(p\,U'v') = h_i p_{i-1/2}\frac{\Delta_i - \Delta_{i-1}}{h_i}\frac{\Theta_i - \Theta_{i-1}}{h_i} = [\boldsymbol{\theta}^i]^T\mathbf{K}^{i1}\boldsymbol{\Delta}^i,$$

kde

$$\boldsymbol{\theta}^i = \begin{pmatrix}\Theta_{i-1}\\ \Theta_i\end{pmatrix}, \qquad \mathbf{K}^{i1} = \frac{p_{i-1/2}}{h_i}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix} \quad \text{a} \quad \boldsymbol{\Delta}^i = \begin{pmatrix}\Delta_{i-1}\\ \Delta_i\end{pmatrix}.$$

Znovu pomocí obdélníkové formule dostaneme

$$Q^i(-r\,Uv') = -\,h_i r_{i-1/2}\frac{\Delta_i + \Delta_{i-1}}{2}\frac{\Theta_i - \Theta_{i-1}}{h_i} = [\boldsymbol{\theta}^i]^T\mathbf{K}^{i2}\boldsymbol{\Delta}^i,$$

kde

$$\mathbf{K}^{i2} = \tfrac{1}{2}r_{i-1/2}\begin{pmatrix}1 & 1\\ -1 & -1\end{pmatrix}.$$

Dále pomocí lichoběžníkové formule obdržíme

$$Q^i(qUv) = \tfrac{1}{2}h_i(q_{i-1}\Theta_{i-1}\Delta_{i-1} + q_i\Theta_i\Delta_i) = [\boldsymbol{\theta}^i]^T\mathbf{K}^{i3}\boldsymbol{\Delta}^i,$$

kde

$$\mathbf{K}^{i3} = \tfrac{1}{2}h_i\begin{pmatrix}q_{i-1} & 0\\ 0 & q_i\end{pmatrix},$$

a položíme $\mathbf{K}^i = \mathbf{K}^{i1}+\mathbf{K}^{i2}+\mathbf{K}^{i3}$. Nakonec, opět pomocí lichoběžníkové formule, dostaneme

$$Q^i(fv) = \tfrac{1}{2}h_i(f_{i-1}\Theta_{i-1} + f_i\Theta_i) = [\boldsymbol{\theta}^i]^T\mathbf{F}^i, \quad \text{kde} \quad \mathbf{F}^i = \tfrac{1}{2}h_i\begin{pmatrix} f_{i-1} \\ f_i \end{pmatrix}.$$

Z rovnice

$$\begin{aligned}
0 = {} & a_h(U,v) - L_h(v) = \\
= {} & \left[\sum_{i=1}^{N} Q^i([pU' - rU]v' + qUv) + [\alpha_0 - r(x_0)]U(x_0)v(x_0) + [\alpha_\ell + r(x_N)]U(x_N)v(x_N)\right] - \\
& \left[\sum_{i=1}^{N} Q^i(fv) + \beta_0 v(x_0) + \beta_\ell v(x_N)\right] = \\
= {} & \sum_{i=1}^{N}[\boldsymbol{\theta}^i]^T[\mathbf{K}^i\boldsymbol{\Delta}^i - \mathbf{F}^i] + \Theta_0[(\alpha_0 - r_0]\Delta_0 - \beta_0] + \Theta_N[(\alpha_\ell + r_N)\Delta_N - \beta_\ell]
\end{aligned}$$

a z rovnice (2.42) tak dostaneme rovnost

$$\begin{aligned}
\boldsymbol{\theta}^T[\mathbf{K}\boldsymbol{\Delta} - \mathbf{F}] = {} & \sum_{i=1}^{N}[\boldsymbol{\theta}^i]^T[\mathbf{K}^i\boldsymbol{\Delta}^i - \mathbf{F}^i] + \\
& \Theta_0[(\alpha_0 - r_0)\Delta_0 - \beta_0] + \Theta_N[(\alpha_\ell + r_N)\Delta_N - \beta_\ell],
\end{aligned} \tag{2.44}$$

z níž plyne postup, jak pomocí elementárních matic $\mathbf{K}^i = \{k^i_{rs}\}_{r,s=1}^2$, elementárních vektorů $\mathbf{F}^i = (F^i_1, F^i_2)^T$ a čísel α_0, r_0, β_0, α_ℓ, r_N, β_ℓ sestavit globální matici $\mathbf{K}$ a globální vektor $\mathbf{F}$:

$k^1_{11}+\alpha_0-r_0$	k^1_{12}		$\dots$			$\beta_0+F^1_1$
k^1_{21}	$k^1_{22}+k^2_{11}$	k^2_{12}	$\dots$			$F^1_2+F^2_1$
	k^2_{21}	$k^2_{22}+k^3_{11}$	$\dots$			$F^2_2+F^3_1$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
			$\dots$	$k^{N-1}_{22}+k^N_{11}$	k^N_{12}	$F^{N-1}_2+F^N_1$
			$\dots$	k^N_{21}	$k^N_{22}+\alpha_\ell+r_N$	$F^N_2+\beta_\ell$

Tab 2.1: Matice soustavy $\mathbf{K}$ a vektor pravé strany $\mathbf{F}$.

stačí srovnat členy se stejnými indexy u parametrů Θ a Δ (pro určení prvků matice $\mathbf{K}$) nebo jen u parametru Θ (pro určení prvků vektoru $\mathbf{F}$) na levé a na pravé straně rovnice (2.44). Struktura matice soustavy $\mathbf{K}$ a vektoru pravé strany $\mathbf{F}$ je patrná z tabulky 2.1.

Pro ekvidistantní dělení je výsledná soustava rovnic (2.43) stejná jako soustava rovnic (2.25), kterou jsme odvodili diferenční metodou.

Výpočet probíhá podle následujícího algoritmu:

1) Matici $\mathbf{K}$ a vektor $\mathbf{F}$ vynulujeme.

2) Postupně procházíme jednotlivé prvky $\langle x_{i-1}, x_i\rangle$, $i = 1, 2, \ldots, N$, na každém z nich vypočteme elementární matici $\mathbf{K}^i = \{k^i_{rs}\}^2_{r,s=1}$ a elementární vektor $\mathbf{F}^i = \{F^i_r\}^2_{r=1}$ a koeficienty k^i_{rs} resp. F^i_r přičteme k odpovídajícím prvkům matice $\mathbf{K}$ resp. vektoru $\mathbf{F}$ v souladu s tabulkou 2.1.

3) Matici $\mathbf{K}$ a vektor $\mathbf{F}$ modifikujeme podle uvažovaných okrajových podmínek:

   a) Je-li v levém krajním bodě předepsána Dirichletova okrajová podmínka (2.11a), odstraníme první řádek matice $\mathbf{K}$ a první řádek vektoru $\mathbf{F}$, pak od pravé strany odečteme první sloupec matice $\mathbf{K}$ násobený předepsanou hodnotou g_0 a nakonec vynecháme také první sloupec matice $\mathbf{K}$.

   b) Je-li v levém krajním bodě předepsána Newtonova okrajová podmínka (2.18a), přičteme k prvku v levém horním rohu matice $\mathbf{K}$ číslo $\alpha_0 - r_0$ a k prvnímu prvku vektoru $\mathbf{F}$ přičteme číslo β_0.

   c) Je-li v pravém krajním bodě předepsána Dirichletova okrajová podmínka (2.11b), odstraníme poslední řádek matice $\mathbf{K}$ a poslední řádek vektoru $\mathbf{F}$, pak od pravé strany odečteme poslední sloupec matice $\mathbf{K}$ násobený předepsanou hodnotou g_ℓ a nakonec vynecháme také poslední sloupec matice $\mathbf{K}$.

   d) Je-li v pravém krajním bodě předepsána Newtonova okrajová podmínka (2.18b), přičteme k prvku v pravém dolním rohu matice $\mathbf{K}$ číslo $\alpha_\ell + r_N$ a k poslednímu prvku vektoru $\mathbf{F}$ přičteme číslo β_ℓ.

4) Vyřešíme soustavu lineárních rovnic $\mathbf{Ku} = \mathbf{F}$. Podle zvolených okrajových podmínek tak získáme

$$\begin{aligned}
&\mathbf{u} = (u_1, u_2, \ldots, u_{N-1})^T && \text{v případě okrajových podmínek (2.11a), (2.11b),}\\
&\mathbf{u} = (u_1, u_2, \ldots, u_N)^T && \text{v případě okrajových podmínek (2.11a), (2.18b),}\\
&\mathbf{u} = (u_0, u_1, \ldots, u_{N-1})^T && \text{v případě okrajových podmínek (2.18a), (2.11b),}\\
&\mathbf{u} = (u_0, u_1, \ldots, u_N)^T && \text{v případě okrajových podmínek (2.18a), (2.18b).}
\end{aligned}$$

Pro chybu $u - U$ a její derivaci platí za předpokladu $u \in C^2\langle 0, \ell\rangle$ odhad

$$u - U = O(h^2)\,, \qquad u' - U' = O(h)\,. \tag{2.45}$$

Dominantní konvekce. Upwind aproximaci konvekčního členu dostaneme z modifikované rovnosti (2.40), a sice

$$\text{najít } U \in W_h \text{ splňující } a_h(U, v) + c_h(U, v) = L_h(v) \quad \forall v \in V_h, \tag{2.40'}$$

kde člen

$$c_h(U, v) = \sum_{i=1}^{N} Q^i \left(\delta_i U' v'\right), \qquad \delta_i = \tfrac{1}{2} h_i |\sigma_i r|, \tag{2.46}$$

reprezentuje tzv. *umělou difúzi*. Vhodnou volbou koeficientů $\{\sigma_i\}_{i=1}^N$ ovlivňujeme velikost dodané umělé difúze. Rovnost (2.40') dostaneme z rovnosti (2.40) tak, že v ní nahradíme „difúzní člen“ $\sum_{i=1}^N Q^i(pU'v')$ členem $\sum_{i=1}^N Q^i([p+\delta_i]U'v')$, tj. na prvku $\langle x_{i-1}, x_i\rangle$ přidáme k difúzi p umělou difúzi δ_i. Pomocí obdélníkové formule dostaneme

$$Q^i\left(\tfrac{1}{2}h_i|\sigma_i r|U'v'\right) = [\boldsymbol{\theta}^i]^T\mathbf{K}^{i4}\boldsymbol{\Delta}^i, \tag{2.47}$$

kde

$$\mathbf{K}^{i4} = \tfrac{1}{2}|\sigma_i r_{i-1/2}|\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{2.48}$$

Položíme $\mathbf{K}^i = \mathbf{K}^{i1} + \mathbf{K}^{i2} + \mathbf{K}^{i3} + \mathbf{K}^{i4}$ a soustavu rovnic sestavíme podle tabulky 2.1. Pro ekvidistantní dělení a $\sigma_i = 1$, $i = 1, 2, \ldots, N$, dostaneme rovnice (2.29). Ověřte! Pro chybu v tom případě platí pouze

$$u - U = O(h). \tag{2.49}$$

Když p, r jsou konstanty, $q = f = 0$, okrajové podmínky jsou Dirichletovy, viz (2.11), a dělení je ekvidistantní, pak pro

$$\sigma_i \equiv \sigma = \coth\kappa - \frac{1}{\kappa}, \quad \text{kde} \quad \kappa = \frac{rh}{2p}, \tag{2.50}$$

dostaneme přesné řešení, viz [5]. Číslo κ je známo jako *lokální Pecletovo číslo*. Snadno ověříme, že funkce $\sigma(\kappa)$ je rostoucí, $\lim\limits_{\kappa\to-\infty}\sigma(\kappa) = -1$, $\lim\limits_{\kappa\to 0}\sigma(\kappa) = 0$, $\lim\limits_{\kappa\to\infty}\sigma(\kappa) = 1$. To je žádoucí chování: pro dominantní konvekci $|\sigma| \to 1$, tj. dostaneme čistý upwind, a pro dominantní difúzi $|\sigma| \to 0$, tj. dostaneme standardní schéma bez umělé difúze.

Proto lze doporučit univerzální volbu

$$\sigma_i = \coth\kappa_i - \frac{1}{\kappa_i}, \quad \text{kde} \quad \kappa_i = \frac{r_{i-1/2}h_i}{2p_{i-1/2}}, \quad i = 1, 2, \ldots, N, \tag{2.51}$$

která je vhodná pro každou konvekčně-difúzní úlohu, nezávisle na tom, zda je konvekce dominantní či nikoliv.

3. Parciální diferenciální rovnice

Parciální diferenciální rovnice (stručně PDR) vyjadřuje vztah mezi funkcí několika proměnných a jejími parciálními derivacemi. Parciální rovnice a jejich soustavy jsou matematickým modelem mnoha technických úloh. Četné modely byly vytvořeny již v minulých staletích, jejich praktické řešení však umožnily teprve výkonné počítače. Prostředky klasické matematické analýzy se zkoumá existence, jednoznačnost, hladkost a další vlastnosti řešení v závislosti na koeficientech rovnice, okrajových a počátečních podmínkách a na oblasti, ve které má být rovnice splněna. Tuto oblast budeme standardně značit symbolem Ω. Pro některé jednodušší úlohy lze těmito prostředky najít i přesné řešení, často ve tvaru nekonečné řady. Převážnou většinu těchto úloh však dovedeme řešit pouze přibližně, numericky.

Označení *eliptické, parabolické* nebo *hyperbolické* získaly PDR na základě formální podobnosti s rovnicemi kuželoseček. *Stacionární* úlohy jsou na čase nezávislé, úlohy *nestacionární* na čase závisejí. K jednoznačnému určení řešení nestačí samotná diferenciální rovnice, je nutno zadat ještě okrajové podmínky a u nestacionárních úloh také počáteční podmínky. Ve třech následujících kapitolách uvedeme nejjednodušší rovnice druhého řádu. Omezíme se přitom na PDR ve dvou proměnných, tj. ve stacionárním případě jde o úlohu ve dvou prostorových proměnných x, y a v nestacionárním případě se uvažují úlohy s jedinou prostorovou proměnnou x, druhou proměnnou je čas t.

Diskretizaci v prostorových proměnných provedeme pomocí diferenční metody, metody konečných objemů a metody konečných prvků. Metoda konečných prvků jednoznačně dominuje při řešení problémů mechaniky pevné fáze, zatímco pro proudění tekutin se více používají programy pracující na bázi metody konečných objemů.

Několik pojmů. *Uzávěr* množiny $M \in \mathbb{R}^d$ je sjednocením bodů množiny M a bodů ležících na její hranici ∂M. Uzávěr M značíme $\overline{M}$, tj. $\overline{M} = M \cup \partial M$.

Oblastí rozumíme otevřenou souvislou množinu v $\mathbb{R}^d$.

Nechť Ω je oblast. Prostor funkcí, které jsou v $\bar{\Omega}$ spojité spolu se svými derivacemi až do řádu k, značíme $C^k(\overline{\Omega})$. Prostor $C^0(\overline{\Omega})$ funkcí spojitých v $\overline{\Omega}$ značíme stručně $C(\overline{\Omega})$.

Řekneme, že funkce u je v Ω po částech spojitá, jestliže $\overline{\Omega}$ je sjednocením uzávěrů konečného počtu navzájem disjunktních podoblastí, tj. $\overline{\Omega} = \bigcup \overline{\Omega}_i$, $\Omega_i \cap \Omega_j = \emptyset$ pro $i \neq j$, a jestliže u je na každé z podoblastí Ω_i spojitá a spojitě prodloužitelná až do hranice, tj. existuje spojitá funkce $\overline{u}_i \in C(\overline{\Omega}_i)$ s vlastností $\overline{u}_i = u_i$ v Ω_i. Prostor po částech spojitých funkcí značíme $PC(\Omega)$. Symbolem $PC^k(\Omega)$ značíme prostor funkcí, které jsou v oblasti $\overline{\Omega}$ spojité spolu se všemi svými derivacemi až do řádu $k-1$ včetně a jejichž k-té derivace jsou po částech spojité. Tak třeba $PC^1(\Omega)$ je prostor funkcí, které jsou v $\overline{\Omega}$ spojité a jejichž první derivace jsou v Ω po částech spojité.

Definici funkce spojité a funkce po částech spojité na hranici lze prostřednictvím parametrického vyjádření hranice převést na definici funkce spojité a funkce po částech spojité na úsečce, viz kapitola 2.4. Pokud jde o značení, tak třeba $PC(\Gamma_\ell)$ je prostor po částech spojitých funkcí na části $\Gamma_\ell \subset \partial\Omega$ hranice oblasti Ω.

3.1. Úloha eliptického typu

3.1.1. Formulace úlohy

Buď Ω omezená *oblast* v $\mathbb{R}^2$. O *hranici* $\Gamma = \partial\Omega$ oblasti Ω předpokládejme, že je sjednocením uzávěrů dvou navzájem disjunktních částí Γ_1 a Γ_2, tj. $\Gamma = \overline{\Gamma}_1 \cup \overline{\Gamma}_2$, $\Gamma_1 \cap \Gamma_2 = \emptyset$. Dále $\mathbf{n} = (n_1, n_2)^T$ nechť je *jednotkový vektor vnější normály* hranice a

$$\frac{\partial u}{\partial n} = \mathbf{n} \cdot \nabla u = n_1 \frac{\partial u}{\partial x} + n_2 \frac{\partial u}{\partial y}$$

je *derivace ve směru vnější normály.*

Nechť $p(x,y) \geq p_0 > 0$, $q(x,y) \geq 0$, $f(x,y)$, $g(x,y)$, $\alpha(x,y) \geq 0$ a $\beta(x,y)$ jsou dané funkce. Naším úkolem je určit funkci $u(x,y)$, která uvnitř Ω vyhovuje diferenciální rovnici

$$-\frac{\partial}{\partial x}\Big(p(x,y)\frac{\partial u}{\partial x}\Big) - \frac{\partial}{\partial y}\Big(p(x,y)\frac{\partial u}{\partial y}\Big) + q(x,y)u = f(x,y) \qquad \text{v } \Omega\,, \tag{3.1}$$

na hranici Γ_1 splňuje Dirichletovu okrajovou podmínku

$$u = g(x,y) \qquad \text{na } \Gamma_1\,, \tag{3.2}$$

a na hranici Γ_2 splňuje Newtonovu okrajovou podmínku

$$-p(x,y)\frac{\partial u}{\partial n} = \alpha(x,y)u - \beta(x,y) \qquad \text{na } \Gamma_2\,. \tag{3.3}$$

Je-li v (3.3) $\alpha = 0$, dostaneme *Neumannovu* okrajovou podmínku.

V případě, že p je konstanta a $q = 0$, dělíme rovnici (3.1) číslem p a vznikne *Poissonova rovnice*

$$-\Delta u = f(x,y) \quad \text{v } \Omega, \tag{3.4}$$

kde *Laplaceův operátor* Δ aplikovaný na funkci u je

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\,.$$

Rovnice $\Delta u = 0$ se nazývá *Laplaceva rovnice.*

Fyzikální význam. Úlohu (3.1)–(3.3) můžeme interpretovat jako *stacionární vedení tepla* v nekonečném hranolu o průřezu Ω. Pak u je teplota, p tepelná vodivost, $-p\nabla u$ je tepelný tok, q je měrný tepelný odpor, f je intenzita vnitřních tepelných zdrojů, okrajová podmínka (3.2) předepisuje teplotu na povrchu a v okrajové podmínce (3.3) je $-p\,\partial u/\partial n$ tepelný tok ve směru vnější normály, α je koeficient přestupu tepla a β je předepsaný tepelný tok.

Poissonova rovnice s homogenní Dirichletovou okrajovou podmínkou $u = 0$ vyjadřuje např. průhyb membrány upevněné na okraji a zatížené tlakem úměrným funkci f.

Laplaceova rovnice s Neumannovou okrajovou podmínkou popisuje např. potenciální proudění: u je potenciál vektoru rychlosti $\mathbf{v} = (v_1, v_2)^T$, kde $v_1 = \partial u/\partial x$, $v_2 = \partial u/\partial y$.

Rovnice $0 = \Delta u = \mathrm{div}\mathbf{v} = 0$ je známa jako podmínka nestlačitelnosti. Neumannova okrajová podmínka $\partial u/\partial n = \mathbf{v}\cdot\mathbf{n} = \beta$ předepisuje normálovou složku rychlosti $v_n = \mathbf{v}\cdot\mathbf{n}$. Pomocí Gauss-Ostrogradského věty, viz např. [21],

$$\int_\Omega \mathrm{div}\,\mathbf{v}\,\mathrm{d}x\,\mathrm{d}y = \int_{\partial\Omega} \mathbf{v}\cdot\mathbf{n}\,\mathrm{d}S = \int_{\partial\Omega} \frac{\partial u}{\partial n}\,\mathrm{d}S = \int_{\partial\Omega} \beta\,\mathrm{d}S = 0,$$

dostáváme nutnou podmínku existence řešení. Potenciál u není určen jednoznačně: je-li u řešení, je také $u+C$ řešení, kde C je libovolná konstanta. Rychlost $\mathbf{v}$ však už jednoznačně určena je.

Existence a jednoznačnost řešení. Podmínky zaručující existenci a jednoznačnost klasického řešení $u \in C^2(\overline{\Omega})$ problému (3.1)–(3.3) jsou komplikované a proto je zde uvádět nebudeme. V praktických aplikacích vystačíme s existencí tzv. *slabého řešení*, které existuje za podmínek v aplikacích běžně splněných.

V dalším budeme předpokládat, že Ω je mnohoúhelník, $p \geq p_0 > 0$, $q \geq 0$, f jsou po částech spojité v Ω, g je spojitá na Γ_1, $\alpha \geq 0$, β jsou po částech spojité na Γ_2, a pokud $\Gamma = \Gamma_2$, pak buďto $q \geq q_0 > 0$ na části Ω nebo $\alpha \geq \alpha_0 > 0$ na části Γ_2. Za těchto předpokladů existuje jediné slabé řešení $u \in PC^1(\Omega)$ problému (3.1)–(3.3), viz např. [12]. Je-li $\Gamma = \Gamma_2$, $q = 0$, $\alpha = 0$ a pokud $\int_\Omega f\,\mathrm{d}x\,\mathrm{d}y + \int_\Gamma \beta\,\mathrm{d}s = 0$, pak má úloha (3.1)–(3.3) nekonečně mnoho řešení: je-li u řešení, pak také $u + C$ je řešení, kde C je libovolná konstanta.

Zesílením uvedených předpokladů lze docílit toho, že slabé řešení je také řešení klasické. Tyto zesílené předpoklady však obvykle odporují požadavkům praktických aplikací.

3.1.2. Diferenční metoda

Diskretizace okrajové úlohy ve dvou dimenzích je analogická diskretizaci jednodimenzionální úlohy, viz kapitola 2.2.

Princip metody. Abychom výklad nezatěžovali detaily nepodstatnými z hlediska numerické metody, začneme řešením Dirichletovy úlohy pro Poissonovu rovnici na čtverci, tj. řešíme rovnici (3.4) s okrajovou podmínkou (3.2) pro $\Gamma_1 = \Gamma$, když $\Omega = (0,\ell)\times(0,\ell)$ je čtverec se stranou délky ℓ.

Na Ω vytvoříme pravidelnou čtvercovou síť. Diferenční metoda se proto také často nazývá *metoda sítí*. Zvolme tedy $N > 1$ celé a definujme *krok* $h = \ell/N$. Označme $x_i = ih$, $i = 0, 1, \ldots, N$, $y_j = jh$, $j = 0, 1, \ldots, N$. Body $[x_i, y_j]$, $i, j = 0, 1, \ldots, N$, nazveme *uzly sítě*. Rovnice (3.4) má být splněna ve všech bodech $[x, y]$ uvnitř Ω, musí tedy být také splněna ve všech vnitřních uzlech, tj.

$$-\frac{\partial^2 u(x_i,y_j)}{\partial x^2} - \frac{\partial^2 u(x_i,y_j)}{\partial y^2} = f(x_i,y_j), \quad i,j = 1,2,\ldots,N-1\,. \tag{3.5}$$

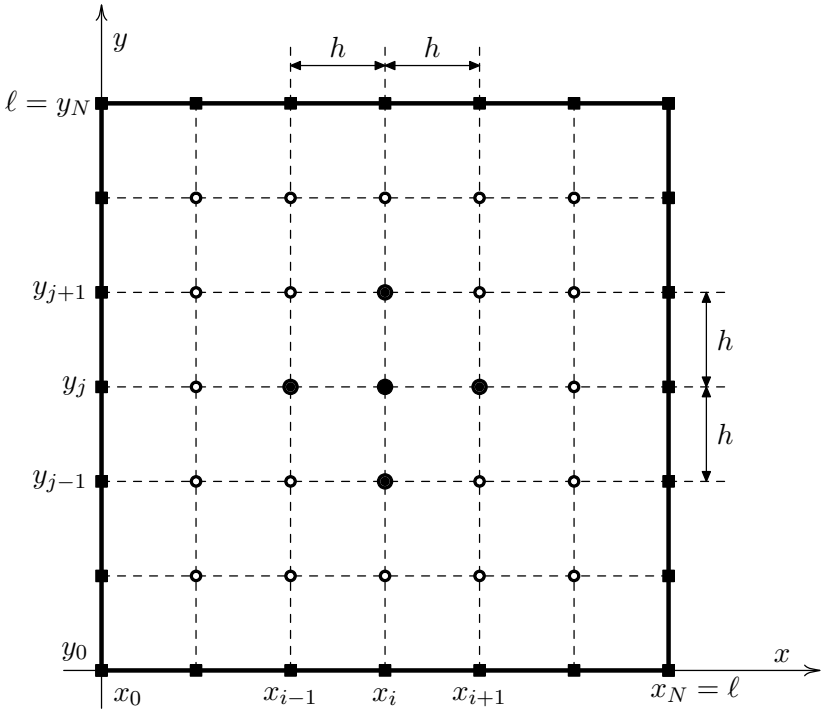
Parciální derivace vyjádříme pomocí centrálních diferencí

$$-\frac{\partial^2 u(x_i,y_j)}{\partial x^2} = \frac{-u(x_{i-1},y_j) + 2u(x_i,y_j) - u(x_{i+1},y_j)}{h^2} + O(h^2)\,, \tag{3.6a}$$

$$-\frac{\partial^2 u(x_i,y_j)}{\partial y^2} = \frac{-u(x_i,y_{j-1}) + 2u(x_i,y_j) - u(x_i,y_{j+1})}{h^2} + O(h^2)\,, \tag{3.6b}$$

dosadíme do rovnice (3.5) a chybové členy $O(h^2)$ zanedbáme. Po vynásobení h^2 dostaneme soustavu *síťových rovnic*

$$-u_{i-1,j} - u_{i,j-1} + 4u_{ij} - u_{i,j+1} - u_{i+1,j} = h^2 f_{ij}\,, \quad i,j = 1,2,\ldots,N-1\,, \tag{3.7}$$



Obr. 3.1. Síť

kde u_{ij} je aproximace $u(x_i, y_j)$ a $f_{ij} = f(x_i, y_j)$. Z okrajové podmínky (3.2) dostaneme

$$u_{ij} = g_{ij} \quad \text{pro } i = 0 \text{ nebo } i = N \text{ nebo } j = 0 \text{ nebo } j = N\,, \tag{3.8}$$

přičemž $g_{ij} = g(x_i, y_j)$. Když z (3.8) dosadíme do (3.7) a na levé straně ponecháme pouze členy s neznámými u_{ij}, dostaneme soustavu $(N-1)^2$ lineárních algebraických rovnic, kterou můžeme zapsat maticově ve tvaru

$$\mathbf{Ku} = \mathbf{F}\,. \tag{3.9}$$

Pro $N = 4$ má soustava rovnic (3.9) následující tvar:

$$\left(\begin{array}{rrr|rrr|rrr}
4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
-1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\
\hline
-1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\
0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\
0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\
\hline
0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4
\end{array}\right)
\left(\begin{array}{c} u_{11} \\ u_{12} \\ u_{13} \\ \hline u_{21} \\ u_{22} \\ u_{23} \\ \hline u_{31} \\ u_{32} \\ u_{33} \end{array}\right)
=
\left(\begin{array}{l} h^2 f_{11} + g_{01} + g_{10} \\ h^2 f_{12} + g_{02} \\ h^2 f_{13} + g_{03} + g_{14} \\ \hline h^2 f_{21} + \qquad g_{20} \\ h^2 f_{22} \\ h^2 f_{23} + \qquad g_{24} \\ \hline h^2 f_{31} + g_{41} + g_{30} \\ h^2 f_{32} + g_{42} \\ h^2 f_{33} + g_{43} + g_{34} \end{array}\right)$$

Pravá strana rovnice odpovídající uzlu, který není sousedem hranice, obsahuje pouze člen $h^2 f_{ij}$, pro uzly nejbližší vrcholům čtverce přibudou dva členy s funkcí g a pro ostatní sousedy hranice jeden člen s funkcí g.

Soustavu (3.9) lze napsat v blokovém tvaru

$$\begin{pmatrix} \mathbf{B} & -\mathbf{I} & \mathbf{O} & \dots & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ -\mathbf{I} & \mathbf{B} & -\mathbf{I} & \dots & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & -\mathbf{I} & \mathbf{B} & \dots & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \dots & -\mathbf{I} & \mathbf{B} & -\mathbf{I} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \dots & \mathbf{O} & -\mathbf{I} & \mathbf{B} \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \\ \vdots \\ \vdots \\ \mathbf{u}_{N-2} \\ \mathbf{u}_{N-1} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \\ \mathbf{F}_3 \\ \vdots \\ \vdots \\ \mathbf{F}_{N-2} \\ \mathbf{F}_{N-1} \end{pmatrix} \tag{3.10}$$

kde $\mathbf{I}$ je jednotková matice řádu $N-1$, $\mathbf{O}$ je nulová matice řádu $N-1$ a $\mathbf{B}$ je třídiagonální matice řádu $N-1$ tvaru

$$\mathbf{B} = \begin{pmatrix} 4 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 4 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 4 & \dots & 0 & 0 & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 4 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 4 \end{pmatrix}.$$

Vektory $\mathbf{u}_i = (u_{i1}, u_{i2}, \dots, u_{i,N-1})^T$, $i = 1, 2, \dots, N-1$, jsou části celkového vektoru $\mathbf{u}$ neznámých, vektory $\mathbf{F}_i = (F_{i1}, F_{i2}, \dots, F_{i,N-1})^T$, $i = 1, 2, \dots, N-1$, jsou části celkového vektoru $\mathbf{F}$ pravých stran.

Pro homogenní okrajovou podmínku $g(x, y) = 0$ je $F_{ij} = h^2 f_{ij}$, pro nehomogenní okrajovou podmínku je v některých rovnicích příspěvek z hranice, přesně

$$\begin{aligned} F_{ij} &= h^2 f_{ij}\,, & & 2 \le i \le N-2, \quad 2 \le j \le N-2\,, \\ F_{1j} &= h^2 f_{1j} + g_{0j}\,, & F_{N-1,j} &= h^2 f_{N-1,j} + g_{Nj}\,, \quad 2 \le j \le N-2\,, \\ F_{i1} &= h^2 f_{i1} + g_{i0}\,, & F_{i,N-1} &= h^2 f_{i,N-1} + g_{iN}\,, \quad 2 \le i \le N-2\,, \\ F_{11} &= h^2 f_{11} + g_{01} + g_{10}\,, & F_{1,N-1} &= h^2 f_{1,N-1} + g_{0,N-1} + g_{1N}\,, \\ F_{N-1,1} &= h^2 f_{N-1,1} + g_{N1} + g_{N-1,0}\,, & F_{N-1,N-1} &= h^2 f_{N-1,N-1} + g_{N,N-1} + g_{N-1,N}\,. \end{aligned}$$

Matice $\mathbf{K}$ soustavy (3.9) má řadu vynikajících vlastností. Některé z nich jsou patrné okamžitě: $\mathbf{K}$ je symetrická, diagonálně dominantní, pásová a pětidiagonální. Matice $\mathbf{K}$ je pro velké N řídká, neboť počet jejích nenulových prvků je malý: z celkového počtu $(N-1)^4$ je jich nenulových jen $5(N-1)^2 - 4(N-1)$. Dá se dokázat, že $\mathbf{K}$ je pozitivně definitní a tedy regulární.

Jsou-li funkce f a g dostatečně hladké, pak pro chybu metody platí

$$u(x_i, y_j) - u_{ij} = O(h^2)\,. \tag{3.11}$$

Obdélníková oblast. Je-li $\Omega = (0,a)\times(0,b)$ obdélník a na něm chceme zvolit pravidelnou síť, musíme ve směru osy y vybrat obecně jiný krok než ve směru osy x. Nechť tedy $N \geq 1$ je počet dílků ve směru osy x a $M \geq 1$ je počet dílků ve směru osy y. Položíme $h = a/N$, $k = b/M$ a definujeme $x_i = ih$, $i = 0,1,\ldots,N$, a $y_j = jk$, $j = 0,1,\ldots,M$. Síť je tvořena uzly $[x_i, y_j]$ pro $i = 0,1,\ldots,N$, $j = 0,1,\ldots,M$. Při diskretizaci Poissonovy rovnice (3.4) postupuje analogicky jako dříve, jen místo (3.6b) použijeme

$$-\frac{\partial^2 u(x_i,y_j)}{\partial y^2} = \frac{-u(x_i,y_{j-1}) + 2u(x_i,y_j) - u(x_i,y_{j+1})}{k^2} + O(k^2)\,, \tag{3.6b'}$$

a tak místo rovnic (3.7) dostaneme pro $i = 1,2,\ldots,N-1$ a $j = 1,2,\ldots,M-1$ rovnice

$$\frac{-u_{i-1,j} + 2u_{ij} - u_{i+1,j}}{h^2} + \frac{-u_{i,j-1} + 2u_{ij} - u_{i,j+1}}{k^2} = f_{ij}\,. \tag{3.7'}$$

Je-li řešení u dostatečně hladké, pro chybu platí

$$u(x_i,y_j) - u_{ij} = O(h^2 + k^2)\,. \tag{3.11'}$$

Obecnější rovnice. Při diskretizaci rovnice (3.1) aproximujeme členy $-[pu_x]_x$ a $-[pu_y]_y$ pomocí vzorce (2.14). Tak ve vnitřních uzlech dostaneme místo (3.7') rovnici

$$\begin{aligned}&\frac{-p_{i-1/2,j}u_{i-1,j} + (p_{i-1/2,j} + p_{i+1/2,j})u_{ij} - p_{i+1/2,j}u_{i+1,j}}{h^2} + \\ &\frac{-p_{i,j-1/2}u_{i,j-1} + (p_{i,j-1/2} + p_{i,j+1/2})u_{ij} - p_{i,j+1/2}u_{i,j+1}}{k^2} + q_{ij}u_{ij} = f_{ij}\,.\end{aligned} \tag{3.7"}$$

Přitom index $_{i\pm1/2}$ znamená, že za argument x dosadíme $x_i \pm \frac{1}{2}h$, a podobně index $_{j\pm1/2}$ znamená, že za y dosadíme $y_j \pm \frac{1}{2}k$.

Je-li řešení u dostatečně hladké, pro chybu opět platí (3.11').

Newtonova okrajová podmínka. Při diskretizaci postupujeme obdobně jako v jednodimenzionálním případě, viz odvození vztahů (2.19). Ukážeme si to na příkladu podmínky

$$-p(a,y)\frac{\partial u(x,y)}{\partial x}\Big|_{x=a} = \alpha^e(y)u(a,y) - \beta^e(y)\,, \qquad 0 < y < b\,, \tag{3.12e}$$

na východní (east) straně obdélníka Ω. Splnění rovnice (3.5) požadujeme i v uzlech $[x_N, y_j]$ na straně $x = a$. Při diskretizaci $-[pu_x]_x(x_N,y_j)$, $1 \leq j \leq M-1$, postupujeme zcela analogicky jako v jedné dimenzi, tj.

$$\begin{aligned}-[pu_x]_x(x_N,y_j) = &-\frac{p_{Nj}u_x(x_N,y_j) - p_{N-1/2,j}u_x(x_{N-1/2},y_j)}{\frac{1}{2}h} + O(h) = \\ &\frac{\alpha_j^e u(x_N,y_j) - \beta_j^e + p_{N-1/2,j}\dfrac{u(x_N,y_j) - u(x_{N-1},y_j)}{h}}{\frac{1}{2}h} + O(h)\,.\end{aligned} \tag{3.13e}$$

Pomocí (3.13e) a užitím standardní aproximace členu $-[pu_y]_y(x_N, y_j)$ dostaneme ve vnitřních uzlech $[x_N, y_j]$ východní strany $x = a$, tj. pro $j = 1, 2, \ldots, M-1$, rovnici

$$\begin{aligned}
&\frac{-p_{N-1/2,j}u_{N-1,j} + (p_{N-1/2,j} + h\alpha_j^e)u_{Nj}}{h^2} + \\
&\frac{-p_{N,j-1/2}u_{N,j-1} + (p_{N,j-1/2} + p_{N,j+1/2})u_{Nj} - p_{N,j+1/2}u_{N,j+1}}{2k^2} + \\
&\frac{1}{2}q_{Nj}u_{Nj} = \frac{1}{2}f_{Nj} + \frac{1}{h}\beta_j^e\,.
\end{aligned} \tag{3.14e}$$

Předpokládejme, že Newtonova okrajová podmínka je předepsána také na severní (north) straně straně Ω, tj. že platí

$$-p(x,b)\frac{\partial u(x,y)}{\partial y}\Big|_{y=b} = \alpha^n(x)u(x,b) - \beta^n(x)\,, \qquad 0 < x < a\,. \tag{3.12n}$$

Podobně jako při odvození (3.13e) dostaneme

$$-[pu_y]_y(x_i, y_M) = \frac{\alpha_i^n u(x_i, y_M) - \beta_i^n + p_{i,M-1/2}\dfrac{u(x_i,y_M) - u(x_i,y_{M-1})}{k}}{\frac{1}{2}k} + O(k)\,. \tag{3.13n}$$

Odtud a užitím standardní aproximace členu $-[pu_x]_x(x_i, y_M)$ dostaneme pro vnitřní uzly horní strany $y = b$, tj. pro $i = 1, 2, \ldots, N-1$, rovnici

$$\begin{aligned}
&\frac{-p_{i-1/2,M}u_{i-1,M} + (p_{i-1/2,M} + p_{i+1/2,M})u_{iM} - p_{i+1/2,M}u_{i+1,M}}{2h^2} + \\
&\frac{-p_{i,M-1/2}u_{i,M-1} + (p_{i,M-1/2} + k\alpha_i^n)u_{iM}}{k^2} + \frac{1}{2}q_{iM}u_{iM} = \frac{1}{2}f_{iM} + \frac{1}{k}\beta_i^n\,.
\end{aligned} \tag{3.14n}$$

Pokud je Newtonova okrajová podmínka předepsána současně na východní i severní straně, dostaneme v severovýchodním rohu $[x_N, y_M]$ (pomocí (3.13e) pro $j = M$ a (3.13n) pro $i = N$) rovnici

$$\begin{aligned}
&\frac{-p_{N-1/2,M}u_{N-1,M} + (p_{N-1/2,M} + h\alpha_M^e)u_{NM}}{2h^2} + \\
&\frac{-p_{N,M-1/2}u_{N,M-1} + (p_{N,M-1/2} + k\alpha_N^n)u_{NM}}{2k^2} + \\
&\frac{1}{4}q_{NM}u_{NM} = \frac{1}{4}f_{NM} + \frac{1}{2h}\beta_M^e + \frac{1}{2k}\beta_N^n\,.
\end{aligned} \tag{3.14ne}$$

Při diskretizaci Newtonovy okrajové podmínky na ostatních stranách a v rozích postupujeme podobně. Je-li řešení u dostatečně hladké, pro chybu opět platí (3.11').

Rovnice s konvekčním členem je tvaru

$$-\frac{\partial}{\partial x}\Big(p(x,y)\frac{\partial u}{\partial x} - r_1(x,y)u\Big) - \frac{\partial}{\partial y}\Big(p(x,y)\frac{\partial u}{\partial y} - r_2(x,y)u\Big) + q(x,y)u = f(x,y). \tag{3.15}$$

Nechť $\mathbf{r} = (r_1, r_2)^T$. Na části $\Gamma_1 = \{\mathbf{x} \in \partial\Omega \mid \mathbf{r} \cdot \mathbf{n} \leq 0\}$ hranice předepíšeme Dirichletovu okrajovou podmínku a na zbývající části $\Gamma_2 = \{\mathbf{x} \in \partial\Omega \mid \mathbf{r} \cdot \mathbf{n} > 0\}$ hranice předepíšeme Newtonovu okrajovou podmínku. V dynamice tekutin Γ_1 může být vtok, to když $\mathbf{r} \cdot \mathbf{n} < 0$, nebo neprostupná stěna, to když $\mathbf{r} \cdot \mathbf{n} = 0$. Část Γ_2 reprezentuje výtok. Abychom zajistili jednoznačnou existenci řešení, předpokládejme, že

$$\operatorname{div} \mathbf{r}(x, y) \equiv \frac{\partial r_1(x, y)}{\partial x} + \frac{\partial r_2(x, y)}{\partial y} \geq 0, \qquad (x, y) \in \Omega. \tag{3.16}$$

V dynamice tekutin $\operatorname{div} \mathbf{r}(x, y) = 0$. Konvekční členy $(r_1 u)_x$ a $(r_2 u)_y$ aproximujeme podobně jako v jednorozměrné úloze, viz. kapitola 2.2. Ukažme si to pro konvekční člen $(r_1 u)_x$. Omezíme se přitom jen na vnitřní uzel $[x_i, y_j]$. Pomocí centrální diference dostaneme

$$(r_1 u)_x(x_i, y_j) \approx \left([r_1 u](x_{i+1/2}, y_j) - [r_1 u](x_{i-1/2}, y_j) \right) / h.$$

Dalším krokem je aproximace konvekčních toků $[r_1 u](x_{i+1/2}, y_j)$ a $[r_1 u](x_{i-1/2}, y_j)$. Protože aproximace obou těchto toků je založena na stejných pravidlech, věnujme se podrobně jen aproximaci $[r_1 u](x_{i+1/2}, y_j)$. Ta závisí na dvourozměrné analogii podmínky (2.26): pokud platí

$$\tfrac{1}{2}h|[r_1]_{0j}| < p_{1/2,j}, \quad \tfrac{1}{2}h|[r_1]_{i+1/2,j}| < p_{i+1/2,j},\ i = 0, 1, \ldots, N-1, \quad \tfrac{1}{2}h|[r_1]_{Nj}| < p_{N-1/2,j}, \tag{3.17}$$

určíme $u(x_{i+1/2}, y_j)$ interpolací z hodnot $u(x_i, y_j)$ a $u(x_{i+1}, y_j)$, tj.

$$[r_1 u](x_{i+1/2}, y_j) \approx \tfrac{1}{2}[r_1]_{i+1/2,j} \left(u(x_i, y_j) + u(x_{i+1}, y_j) \right), \tag{3.18}$$

v opačném případě použijeme upwind aproximaci

$$[r_1 u](x_{i+1/2}, y_j) \approx [r_1]^+_{i+1/2,j} u(x_i, y_j) + [r_1]^-_{i+1/2,j} u(x_{i+1}, y_j). \tag{3.19}$$

Konvekční člen $(r_2 u)_y(x_i, y_j)$ aproximujeme obdobně jako člen $(r_1 u)_x(x_i, y_j)$, při rozhodování o typu aproximace konvekčního toku $[r_2 u](x_i, y_{j+1/2})$ se řídíme podmínkou

$$\tfrac{1}{2}k|[r_2]_{i0}| < p_{i,1/2}, \quad \tfrac{1}{2}k|[r_2]_{i,j+1/2}| < p_{i,j+1/2},\ j = 0, 1, \ldots, M-1, \quad \tfrac{1}{2}k|[r_2]_{iM}| < p_{i,M-1/2}. \tag{3.20}$$

Aproximujeme-li konvekční členy interpolací, pak za předpokladu dostatečné hladkosti řešení u pro chybu platí (3.11'). Jestliže však konvekční členy aproximujeme užitím upwind aproximace, řád chyby se o jednotku sníží, pro chybu platí jen

$$u(x_i, y_j) - u_{ij} = O(h + k).$$

Pro dosažení přesnosti řádu $O(h^2 + k^2)$ je třeba používat přesnější upwind aproximaci konvekčního toku, viz (2.28').

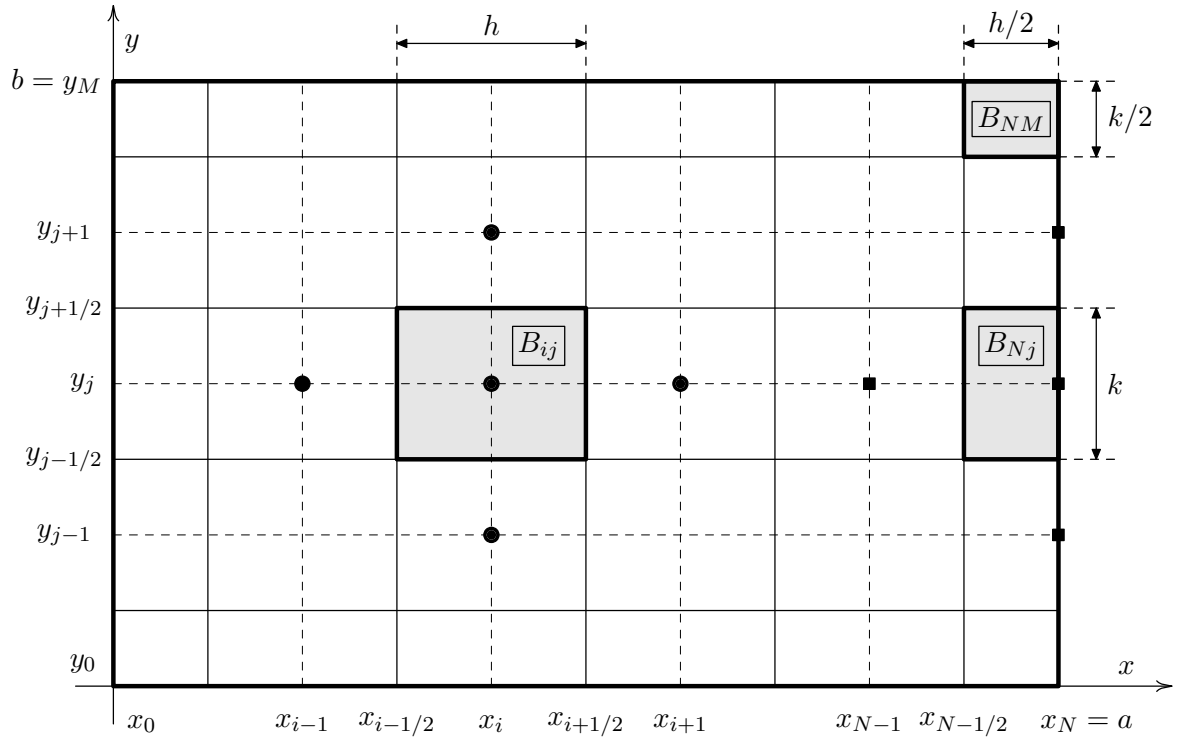
Dirichletova okrajová podmínka. Standardní postup je tento: je-li v uzlu $[x_i, y_j]$ předepsána hodnota řešení $u(x_i, y_j) = g_{ij}$, rovnici pro tento uzel vůbec nesestavujeme a v rovnicích obsahujících u_{ij} položíme $u_{ij} = g_{ij}$.

Podmínky $u_{ij} = g_{ij}$ lze vynutit i jinak. Nejdříve sestavíme soustavu rovnic jako kdyby na části Γ_1 hranice byla předepsána Newtonova okrajová podmínka (3.3) s $\alpha = \beta = 0$. Následně modifikujeme rovnice příslušné uzlům ležícím na Γ_1. Předpokládejme, že uzlu $[x_i, y_j]$ s předepsanou hodnotou $u(x_i, y_j) = g_{ij}$ přísluší r-tá rovnice soustavy $\mathbf{Ku} = \mathbf{F}$. Pak provedeme $k_{rr} := \kappa k_{rr}, F_r := \kappa k_{rr} g_{ij}$, kde κ je velké číslo, např. $\kappa = 10^{20}$. To způsobí, že mimodiagonální koeficienty v r-té rovnici budou oproti velkému diagonálnímu koeficientu prakticky zanedbatelné, takže r-tá rovnice nabude přibližně tvaru $\kappa k_{rr} u_r \doteq \kappa k_{rr} g_{ij}$ nebo-li $u_r \doteq g_{ij}$, což jsme potřebovali zajistit.

3.1.3. Metoda konečných objemů

Metodu vysvětlíme pro konvekčně-difúzní úlohu (3.15), (3.2), (3.3).

Pravidelná síť. Uvažujme nejdříve případ, kdy $\Omega = (0, a) \times (0, b)$ je obdélník. Na Ω zvolme uzly $[x_i, y_j] = [ih, jk]$, $i = 0, 1, \ldots, N$, $j = 0, 1, \ldots, M$, kde $h = a/N$, $k = b/M$.



Obr. 3.2. Vnitřní buňka B_{ij}, hraniční buňka B_{Nj} a rohová buňka B_{NM}

Obdélník

$$B_{ij} = \left[\langle x_{i-1/2}, x_{i+1/2}\rangle \times \langle y_{j-1/2}, y_{j+1/2}\rangle\right] \cap \overline{\Omega}\,, \qquad i = 0, 1, \ldots, N\,, \quad j = 0, 1, \ldots, M,$$

nazveme *konečným objemem* (stručně *buňkou*), viz Obr. 3.2. Integrací diferenciální rovnice (3.15) přes buňku B_{ij} dostaneme bilanční rovnici

$$\int_{B_{ij}} \left[-(pu_x - r_1 u)_x - (pu_y - r_2 u)_y + qu\right] \mathrm{d}x\,\mathrm{d}y = \int_{B_{ij}} f\,\mathrm{d}x\,\mathrm{d}y\,. \tag{3.21}$$

Uvažujme nejdříve případ, kdy B_{ij} je vnitřní buňka, tj. když $0 < i < N$ a $0 < j < M$. Pak $B_{ij} = \langle x_{i-1/2}, x_{i+1/2}\rangle \times \langle y_{j-1/2}, y_{j+1/2}\rangle$, viz Obr.3.2. Integrací per-partes obdržíme

$$\begin{aligned}
&-\int_{y_{j-1/2}}^{y_{j+1/2}} [pu_x - r_1 u]\,(x_{i+1/2}, y)\,\mathrm{d}y + \int_{y_{j-1/2}}^{y_{j+1/2}} [pu_x - r_1 u]\,(x_{i-1/2}, y)\,\mathrm{d}y - \\
&-\int_{x_{i-1/2}}^{x_{i+1/2}} [pu_y - r_2 u]\,(x, y_{j+1/2})\,\mathrm{d}x + \int_{x_{i-1/2}}^{x_{i+1/2}} [pu_y - r_2 u]\,(x, y_{j-1/2})\,\mathrm{d}x + \qquad (3.22)\\
&+\int_{B_{ij}} qu\,\mathrm{d}x\,\mathrm{d}y = \int_{B_{ij}} f\,\mathrm{d}x\,\mathrm{d}y\,.
\end{aligned}$$

Ukážeme si, jak provést aproximaci prvního členu v (3.22), zbývající jednoduché integrály se aproximují podobně. Pomocí obdélníkové formule dostaneme

$$-\int_{y_{j-1/2}}^{y_{j+1/2}} [pu_x - r_1 u]\,(x_{i+1/2}, y)\,\mathrm{d}y \approx k[-p_{i+1/2,j} u_x(x_{i+1/2}, y_j) + [r_1 u](x_{i+1/2}, y_j)]\,,$$

derivaci $u_x(x_{i+1/2}, y_j)$ aproximujeme pomocí centrální diference

$$u_x(x_{i+1/2}, y_j) \approx [u(x_{i+1}, y_j) - u(x_i, y_j)]/h$$

a $[r_1 u](x_{i+1/2}, y_j)$ aproximujeme stejně jako v diferenční metodě, viz (3.17)–(3.19).

Dvojné integrály v (3.22) aproximujeme pomocí součinové obdélníkové formule:

$$\int_{B_{ij}} qu\,\mathrm{d}x\,\mathrm{d}y \approx hkq_{ij}u(x_i, y_j)\,, \qquad \int_{B_{ij}} f\,\mathrm{d}x\,\mathrm{d}y \approx hkf_{ij}\,.$$

Po dosazení do (3.22) dospějeme ke stejné rovnici, jakou bychom dostali pomocí diferenční metody.

Věnujme se nyní případu, kdy uzel $[x_i, y_j]$ leží na hranici $\partial\Omega$, tj. když $i = 0$ nebo $i = N$ nebo $j = 0$ nebo $j = M$. Jestliže uzel $[x_i, y_j]$ leží na části $\overline{\Gamma}_1$ hranice a je v něm tedy předepsána hodnota $u(x_i, y_j) = g(x_i, y_j)$, pak bilanční rovnici (3.21) pro takový uzel vůbec nesestavujeme. Uvažujme tedy případ, kdy uzel $[x_i, y_j]$ je vnitřním bodem hranice Γ_2. Pro konkrétnost budeme uvažovat vnitřní uzel východní strany obdélníka Ω, tj. uzel $[x_N, y_j]$, $0 < j < M$, pro který $B_{Nj} = \langle x_{N-1/2}, x_N\rangle \times \langle y_{j-1/2}, y_{j+1/2}\rangle$, viz Obr. 3.2. Bilanční rovnici (3.21) upravíme integrací per-partes,

$$\begin{aligned}
&-\int_{y_{j-1/2}}^{y_{j+1/2}} [pu_x - r_1 u]\,(x_N, y)\,\mathrm{d}y + \int_{y_{j-1/2}}^{y_{j+1/2}} [pu_x - r_1 u]\,(x_{N-1/2}, y)\,\mathrm{d}y - \\
&-\int_{x_{N-1/2}}^{x_N} [pu_y - r_2 u]\,(x, y_{j+1/2})\,\mathrm{d}x + \int_{x_{N-1/2}}^{x_N} [pu_y - r_2 u]\,(x, y_{j-1/2})\,\mathrm{d}x + \qquad (3.23)\\
&+\int_{B_{Nj}} qu\,\mathrm{d}x\,\mathrm{d}y = \int_{B_{Nj}} f\,\mathrm{d}x\,\mathrm{d}y\,.
\end{aligned}$$

První integrál v (3.23) upravíme užitím Newtonovy okrajové podmínky (3.12e),

$$-\int_{y_{j-1/2}}^{y_{j+1/2}} [pu_x - r_1 u]\,(x_N, y)\,\mathrm{d}y = \int_{y_{j-1/2}}^{y_{j+1/2}} [\alpha^e(y) u(x_N, y) - \beta^e(y) + [r_1 u](x_N, y)]\,\mathrm{d}y\,,$$

a pak použijeme obdélníkovou formuli, takže

$$-\int_{y_{j-1/2}}^{y_{j+1/2}} [pu_x - r_1 u]\,(x_N, y)\,\mathrm{d}y \approx k[(\alpha_j^e + [r_1]_{Nj})u(x_N, y_j) - \beta_j^e]\,.$$

Druhý integrál v (3.23) aproximujeme stejně jako obdobný integrál ve vnitřní buňce. Třetí integrál v (3.23) aproximujeme užitím jednostranné obdélníkové formule

$$-\int_{x_{N-1/2}}^{x_N} [pu_y - r_2 u]\,(x, y_{j+1/2})\,\mathrm{d}x \approx -\tfrac{1}{2}h[pu_y - r_2 u](x_N, y_{j+1/2})\,,$$

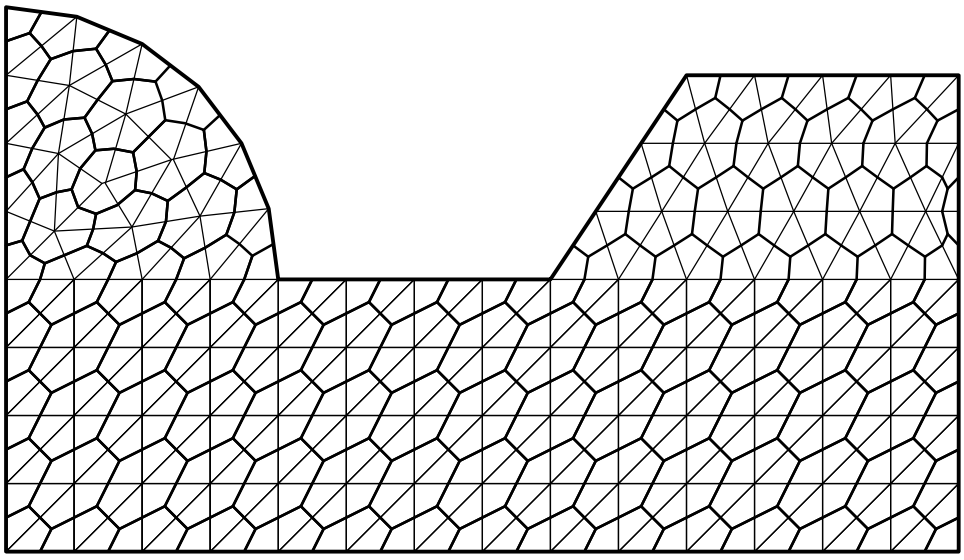
zbývající aproximace se provedou stejně jako ve vnitřní buňce. Čtvrtý integrál v (3.23) aproximujeme obdobně jako třetí integrál. Poslední dva integrály v (3.23) aproximujeme užitím jednostranné součinové obdélníkové formule

$$\int_{B_{Nj}} qu\,\mathrm{d}x\,\mathrm{d}y \approx \tfrac{1}{2}hkq_{Nj}u(x_N, y_j), \qquad \int_{B_{Nj}} f\,\mathrm{d}x\,\mathrm{d}y \approx \tfrac{1}{2}hkf_{Nj}\,.$$

Po dosazení do (3.22) opět dojdeme ke stejné rovnici, jakou bychom dostali pomocí diferenční metody.

Diskretizaci bilanční rovnice pro rohové konečné objemy lze provést pomocí dříve uvedených obratů a proto zde již tuto diskretizaci neuvádíme.

Obecnější síť buněk. Mnohoúhelník $\overline{\Omega}$ vyjádříme jako sjednocení konečného počtu uzavřených trojúhelníků, z nichž každé dva jsou buďto disjunktní nebo mají společný vrchol nebo stranu. Vrcholy trojúhelníků nazveme uzly. Soubor všech trojúhelníků vytváří tzv. *triangulaci* oblasti Ω, v metodě konečných objemů označovanou jako *primární síť*, viz Obr. 3.3. Ke každému uzlu P_i přiřadíme buňku B_i. Sestavíme ji z přilehlých částí C_{ijk} všech trojúhelníků s vrcholem P_i. Objasněme si to podrobněji.



Obr. 3.3. Duální síť

Nechť T_{ijk} je trojúhelník s vrcholy P_i, P_j a P_k. Označme P_{ij} střed strany $\overline{P_iP_j}$, P_{ik} střed strany $\overline{P_iP_k}$ a P_{ijk} těžiště trojúhelníka T_{ijk}. Pak do buňky B_i zahrneme čtyřúhelník C_{ijk} s vrcholy P_i, P_{ij}, P_{ijk} a P_{ik}, viz Obr. 3.4. Tento postup opakujeme pro všechny trojúhelníky T_{ijk}, které obsahují uzel P_i, a sjednocením přilehlých částí C_{ijk} dostaneme buňku B_i. Vnitřní buňka B_i příslušná vnitřnímu uzlu $P_i \notin \partial\Omega$ je zakreslena na Obr. 3.5, hraniční buňka pro $P_i \in \partial\Omega$ pak na Obr. 3.6. Množina všech buněk se nazývá *duální síť*, viz Obr. 3.3.

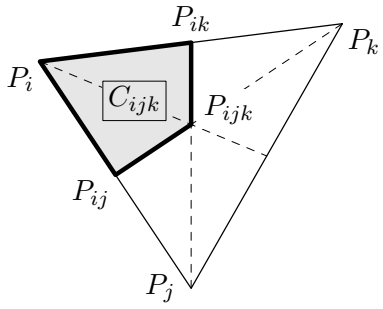
Diskretizace vychází opět z bilanční rovnice (3.21). Pomocí Gauss-Ostrogradského věty

$$\int_{B_i} \operatorname{div} \mathbf{w}\,\mathrm{d}x\,\mathrm{d}y = \int_{\partial B_i} (\mathbf{w}\cdot\mathbf{n})\,\mathrm{d}s, \quad \text{kde} \quad \mathbf{w} = \begin{pmatrix} pu_x - r_1 u \\ pu_y - r_2 u \end{pmatrix},$$

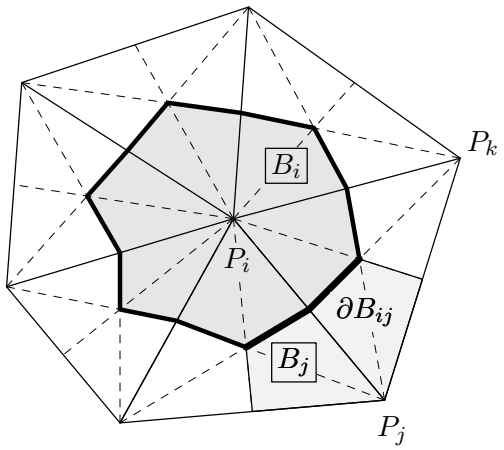
dostaneme analog rovnice (3.22),

$$\int_{\partial B_i} \left[-p\frac{\partial u}{\partial n_i} + (\mathbf{r}\cdot\mathbf{n}_i)u \right] \mathrm{d}s + \int_{B_i} qu\,\mathrm{d}x\,\mathrm{d}y = \int_{B_i} f\,\mathrm{d}x\,\mathrm{d}y\,, \tag{3.22'}$$

kde $\mathbf{r}\cdot\mathbf{n}_i = r_1 n_1 + r_2 n_2$ a $\mathbf{n}_i = (n_1^i, n_2^i)^T$ je jednotkový vektor vnější normály hranice ∂B_i buňky B_i.



Obr. 3.4. $C_{ijk} = B_i \cap T_{ijk}$



Obr. 3.5. Vnitřní buňka B_i

Klíčová je aproximace křivkového integrálu. Věnujme se nejdříve případu, kdy B_i je vnitřní buňka. Hranici ∂B_i vyjádříme jako sjednocení společných částí $\partial B_{ij} = \partial B_i \cap \partial B_j$ buňky B_i a sousedních buněk B_j, tj. $\partial B_i = \bigcup_j \partial B_{ij}$. Protože $\int_{\partial B_i} = \sum_j \int_{\partial B_{ij}}$, stačí popsat aproximaci jen na ∂B_{ij}. To lze provést třeba takto:

$$\int_{\partial B_{ij}} \left[-p\frac{\partial u}{\partial n_i} + (\mathbf{r}\cdot\mathbf{n}_i)u \right] \mathrm{d}s \approx$$

$$|\partial B_{ij}| \left[p(P_{ij}) \frac{u(P_i) - u(P_j)}{|\overline{P_iP_j}|} + (\mathbf{r}\cdot\mathbf{n})_{ij} u(P_{ij}) \right],$$

kde $|\partial B_{ij}|$ je délka ∂B_{ij}, $|\overline{P_iP_j}|$ je délka úsečky $\overline{P_iP_j}$,

$$(\mathbf{r}\cdot\mathbf{n})_{ij} = \mathbf{r}(P_{ij})\cdot\mathbf{n}_{ij} = r_1(P_{ij})n_1^{ij} + r_2(P_{ij})n_2^{ij} \quad \text{a} \quad \mathbf{n}_{ij} = (n_1^{ij}, n_2^{ij})^T = (P_j - P_i)/|\overline{P_iP_j}|$$

je jednotkový vektor ve směru vektoru $\overrightarrow{P_iP_j}$. Zbývá aproximovat hodnotu $u(P_{ij})$ ve středu úsečky P_iP_j. Jestliže je dominantní difúze, třeba když

$$\tfrac{1}{2}|\overline{P_iP_j}||(\mathbf{r}\cdot\mathbf{n})_{ij}| < p(P_{ij})\,,$$

zvolíme aritmetický průměr

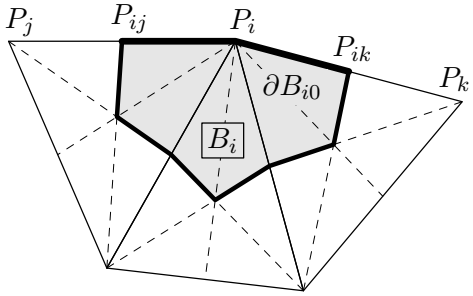
$$u(P_{ij}) \approx \tfrac{1}{2}(u(P_i)+u(P_j))\,,$$

v opačném případě užijeme upwind aproximaci,

$$u(P_{ij}) \approx \begin{cases} u(P_i), \\ u(P_j), \end{cases} \quad \text{pokud} \quad (\mathbf{r}\cdot\mathbf{n})_{ij} \quad \begin{cases} \geq 0, \\ < 0. \end{cases}$$

Dvojné integrály aproximujeme jako součin obsahu $|B_{ij}|$ buňky B_{ij} a hodnoty integrandu v bodu P_i, tj.

$$\int_{B_i} qu\,\mathrm{d}x\,\mathrm{d}y \approx |B_i|q(P_i)u(P_i)\,, \qquad \int_{B_i} f\,\mathrm{d}x\,\mathrm{d}y \approx |B_i|f(P_i)\,.$$



Obr. 3.6. Hraniční buňka B_i

Pro uzel P_i ležící na hranici Γ_1 buňku B_i nepotřebujeme a klademe $u(P_i) = g(P_i)$. Pro vnitřní uzel P_i hranice Γ_2 sestrojíme hraniční buňku, viz Obr. 3.6. Hranice $\partial B_i = \bigcup_j \partial B_{ij} \cup \partial B_{i0}$ je sjednocením společných částí $\partial B_{ij} = \partial B_i \cap \partial B_j$ buňky B_i a sousedních buněk B_j a dále části $\partial B_{i0} = \partial B_i \cap \Gamma_2$ hranice buňky B_i ležící na hranici Γ_2, $\partial B_{i0} = \overline{P_iP_{ij}} \cup \overline{P_iP_{ik}}$ na Obr. 3.6.

Při aproximaci $\int_{\partial B_{i0}}$ využijeme předepsanou Newtonovu okrajovou podmínku (3.3),

$$\int_{\partial B_{i0}} \left[-p\frac{\partial u}{\partial n_i} + (\mathbf{r}\cdot\mathbf{n}_i)u\right]\mathrm{d}s = \int_{\partial B_{i0}} [\alpha u - \beta + (\mathbf{r}\cdot\mathbf{n}_i)u]\,\mathrm{d}s \approx$$
$$|\partial B_{i0}|\,[\alpha(P_i)u(P_i) - \beta(P_i)] + \mathbf{r}(P_i)\cdot\left[\,|\overline{P_iP_{ij}}|\mathbf{n}^i_{ij} + |\overline{P_iP_{ik}}|\mathbf{n}^i_{ik}\,\right]u(P_i),$$

kde $\mathbf{n}^i_{ij}$ resp. $\mathbf{n}^i_{ik}$ je vektor vnější normály buňky B_i na straně $\overline{P_iP_j}$ resp. $\overline{P_iP_k}$. Zbývající aproximace se provedou stejně jako u vnitřní buňky.

Více podrobností o metodě konečných objemů, včetně přesnější varianty upwind aproximace konvekčního členu, lze najít např. v [12].

3.1.4. Metoda konečných prvků

Metodu vysvětlíme pro konvekčně-difúzní úlohu (3.15), (3.2), (3.3).

Slabé řešení. Stejně jako v kapitole 2.4 úlohu převedeme na tvar vhodný pro nasazení MKP, tj. odvodíme slabou formulaci naší úlohy. K tomu účelu násobíme rovnici (3.1)

testovací funkcí $v \in C^1(\overline{\Omega})$ s vlastností $v = 0$ na Γ_1 a integrujeme přes oblast Ω, tj. provedeme

$$-\int_\Omega [(pu_x - r_1 u)_x + (pu_y - r_2 u)_y] v \,\mathrm{d}x\,\mathrm{d}y + \int_\Omega quv \,\mathrm{d}x\,\mathrm{d}y = \int_\Omega fv \,\mathrm{d}x\,\mathrm{d}y\,. \tag{3.24}$$

První člen na levé straně upravíme pomocí Gauss-Ostrogradského věty

$$\int_\Omega \operatorname{div} \mathbf{w} \,\mathrm{d}x\,\mathrm{d}y = \int_{\partial\Omega} (\mathbf{w}\cdot\mathbf{n}) \,\mathrm{d}s, \quad \text{kde} \quad \mathbf{w} = \begin{pmatrix} (pu_x - r_1 u)v \\ (pu_y - r_2 u)v \end{pmatrix},$$

na tvar

$$-\int_\Omega [(pu_x - r_1 u)_x + (pu_y - r_2 u)_y] v \,\mathrm{d}x\,\mathrm{d}y = -\int_{\partial\Omega} [(pu_x - r_1 u)n_1 + (pu_y - r_2 u)n_2] v \,\mathrm{d}s +$$
$$\int_\Omega [(pu_x - r_1 u)v_x + (pu_y - r_2 u)v_y] \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}x\,\mathrm{d}y\,.$$

Křivkový integrál dále upravíme: využijeme toho, že $v = 0$ na Γ_1 a že na Γ_2 platí okrajová podmínka (3.3). Tak dostaneme

$$-\int_{\partial\Omega} [(pu_x - r_1 u)n_1 + (pu_y - r_2 u)n_2] v \,\mathrm{d}s = \int_{\Gamma_2} \left[-p\frac{\partial u}{\partial n} + (r_1 n_1 + r_2 n_2)u \right] v \,\mathrm{d}s =$$
$$\int_{\Gamma_2} [\alpha u - \beta + (\mathbf{r}\cdot\mathbf{n})u] v \,\mathrm{d}s\,.$$

Dosadíme-li z posledních dvou rovností do (3.24) vidíme, že řešení u úlohy (3.15), (3.2), (3.3) splňuje rovnici

$$\int_\Omega [p(u_x v_x + u_y v_y) - u(r_1 v_x + r_2 v_y) + quv] \,\mathrm{d}x\,\mathrm{d}y + \int_{\Gamma_2} (\alpha + \mathbf{r}\cdot\mathbf{n})uv] \,\mathrm{d}s =$$
$$\int_\Omega fv \,\mathrm{d}x\,\mathrm{d}y + \int_{\Gamma_2} \beta v \,\mathrm{d}s \tag{3.25}$$

pro každou funkci $v \in C^1(\overline{\Omega})$ s vlastností $v = 0$ na Γ_1. Rovnice (3.25) má zřejmě smysl i v případě, když funkce u a v jsou jen z prostoru $X \equiv PC^1(\overline{\Omega})$. Testovací funkce tedy volíme z *prostoru* $V = \{v \in X \,|\, v = 0 \text{ na } \overline{\Gamma}_1\}$ *testovacích funkcí* a řešení u z *množiny* $W = \{v \in X \,|\, v = g \text{ na } \overline{\Gamma}_1\}$ *přípustných řešení*. Označíme-li

$$\begin{aligned} a(u,v) &= \int_\Omega [\, p\,(\nabla u\cdot\nabla v) - u\,(\mathbf{r}\cdot\nabla v) + quv] \,\mathrm{d}x\,\mathrm{d}y + \int_{\Gamma_2} (\alpha + \mathbf{r}\cdot\mathbf{n})uv \,\mathrm{d}s\,, \\ L(v) &= \int_\Omega fv \,\mathrm{d}x\,\mathrm{d}y + \int_{\Gamma_2} \beta v \,\mathrm{d}s, \end{aligned} \tag{3.26}$$

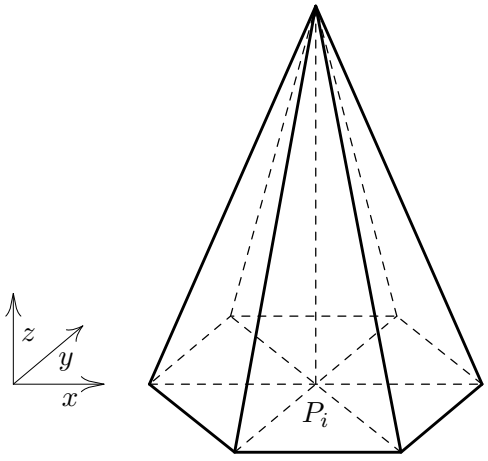
pak úlohu

$$\text{najít } u \in W \text{ splňující } a(u,v) = L(v) \quad \forall v \in V \tag{3.27}$$

nazveme *slabou formulací* úlohy (3.15), (3.2), (3.3) a funkci u nazveme *slabým řešením*.

Diskretizace. Omezíme se na případ, že Ω je mnohoúhelník. $\overline{\Omega}$ vyjádříme jako sjednocení konečného počtu uzavřených trojúhelníků T_e, z nichž každé dva jsou buďto disjunktní nebo mají společný vrchol nebo společnou stranu, viz Obr. 3.3. Množinu $\mathcal{T}$ všech trojúhelníků nazveme triangulací oblasti Ω. Trojúhelníky budeme značit $T_1, T_2, \ldots, T_{N_T}$. Vrcholy trojúhelníků budeme nazývat uzly, značíme je $P_1, P_2, \ldots, P_M$. Předpokládejme, že společné body částí $\overline{\Gamma}_1$ a $\overline{\Gamma}_2$ hranice Γ jsou uzly triangulace $\mathcal{T}$. Množinu stran trojúhelníků $T \in \mathcal{T}$, jejichž sjednocení je $\overline{\Gamma}_2$, označíme jako $\mathcal{S}$. Strany budeme značit $S_1, S_2, \ldots, S_{N_S}$. Nejdelší stranu trojúhelníků triangulace $\mathcal{T}$ označíme jako h.

Funkci, která je v $\overline{\Omega}$ spojitá a která je na každém trojúhelníku $T \in \mathcal{T}$ lineární, nazveme *spojitou po částech lineární funkcí*. Každá taková funkce $v(x, y)$ je jednoznačně určena svými hodnotami $v(x_i, y_i)$ ve vrcholech $P_i \equiv [x_i, y_i]$ triangulace $\mathcal{T}$. Prostor všech spojitých po částech lineárních funkcí označíme X_h. Speciálním případem funkcí z X_h jsou bázové funkce $w_i(x, y)$, které jsou v P_i rovny jedné a v ostatních uzlech jsou rovny nule, tj.



Obr. 3.7. Bázová funkce $w_i(x, y)$

$$w_i(P_j) = \begin{cases} 1 & \text{pro } \begin{cases} i = j, \\ i \neq j. \end{cases} \\ 0 \end{cases}$$

Každá funkce $v \in X_h$ může být pomocí svých hodnot v uzlech a pomocí bázových funkcí vyjádřena ve tvaru

$$v(x, y) = \sum_{i=1}^{M} v(x_i, y_i) w_i(x, y)\,.$$

Dále definujme prostor testovacích funkcí

$$V_h = \{v \in X_h \mid v(x_i, y_i) = 0 \ \forall P_i \in \overline{\Gamma}_1\}$$

a množinu přípustných řešení

$$W_h = \{v \in X_h \mid v(x_i, y_i) = g(x_i, y_i) \ \forall P_i \in \overline{\Gamma}_1\}.$$

Trojúhelník, na němž je definována lineární funkce jednoznačně určená svými hodnotami ve vrcholech, se nazývá *Lagrangeův lineární trojúhelníkový prvek*.

Nyní už máme vše potřebné k dispozici: přibližné řešení U, tzv. *MKP řešení*, obdržíme z *diskrétní slabé formulace*

$$\text{najít } U \in W_h \text{ splňující } a_h(U, v) = L_h(v) \quad \forall v \in V_h\,, \tag{3.28}$$

kde

$$a_h(U, v) = \sum_{T_e \in \mathcal{T}} \left[Q^{T_e}(p\,(\nabla U \cdot \nabla v)) + Q^{T_e}(-U\,(\mathbf{r} \cdot \nabla v)) + Q^{T_e}(qUv) \right] + \sum_{S_e \in \mathcal{S}} Q^{S_e}((\alpha + \mathbf{r} \cdot \mathbf{n})\,Uv),$$

$$L_h(v) = \sum_{T_e \in \mathcal{T}} Q^{T_e}(fv) + \sum_{S_e \in \mathcal{S}} Q^{S_e}(\beta v). \tag{3.29}$$

Přitom symbolem $Q^{T_e}(\varphi)$ jsme označili kvadraturní formuli pro výpočet $\int_{T_e} \varphi \,\mathrm{d}x\,\mathrm{d}y$ na

trojúhelníku T_e a symbolem $Q^{S_e}(\varphi)$ formuli pro výpočet $\int_{S_e} \varphi \, \mathrm{d}s$ na straně S_e. Jako vhodné kvadraturní formule lze doporučit:

1) členy $p(\nabla U \cdot \nabla v)$ a $U(\mathbf{r} \cdot \nabla v)$ integrujeme na trojúhelníku T formulí

$$Q^T(\varphi) = |T| \varphi(P_0) \,, \tag{3.30}$$

kde $|T|$ je plocha trojúhelníka T, $P_0 = \frac{1}{3}(P_1 + P_2 + P_3)$ je těžiště T a P_1, P_2, P_3 jsou vrcholy T;

2) členy qUv a fv integrujeme na trojúhelníku T formulí

$$Q^T(\varphi) = \tfrac{1}{3}|T| \left[\varphi(P_1) + \varphi(P_2) + \varphi(P_3)\right] ; \tag{3.31}$$

3) členy $(\alpha + \mathbf{r} \cdot \mathbf{n})Uv$ a βv integrujeme na straně S lichoběžníkovou formulí

$$Q^S(\varphi) = \tfrac{1}{2}|S| [\varphi(P_1) + \varphi(P_2)] \,, \tag{3.32}$$

kde $|S|$ je délka strany S a P_1, P_2 jsou koncové body S.

Formule (3.30), (3.31) a (3.32) jsou řádu 1, tj. jsou přesné, když φ je polynom stupně 1.

Předpokládejme, že uzly jsou očíslovány tak, že $P_1, P_2, \ldots, P_N$ leží buďto uvnitř oblasti Ω nebo uvnitř hranice Γ_2, a že $P_{N+1}, P_{N+2}, \ldots, P_M$ leží na hranici $\overline{\Gamma}_1$. To není žádné omezení, neboť uzly lze vždy přečíslovat tak, aby byl tento předpoklad splněn. Pak

$$U(x,y) = \sum_{j=1}^{N} \Delta_j w_j(x,y) + \sum_{j=N+1}^{M} g_j w_j(x,y) \,, \qquad v(x,y) = \sum_{i=1}^{N} \Theta_i w_i(x,y) \,, \tag{3.33}$$

kde $\Delta_j = U(P_j)$, $g_j = g(P_j)$ a $\Theta_i = v(P_i)$. Dosazením do (3.28) obdržíme

$$\begin{aligned} 0 = a_h(U,v) - L_h(v) &= a_h\left(\sum_{j=1}^{N} \Delta_j w_j + \sum_{j=N+1}^{M} g_j w_j, \sum_{i=1}^{N} \Theta_i w_i\right) - L_h\left(\sum_{i=1}^{N} \Theta_i w_i\right) = \\ &= \sum_{i=1}^{N} \Theta_i \sum_{j=1}^{N} a_h(w_j, w_i)\Delta_j - \sum_{i=1}^{N} \Theta_i \left[L_h(w_i) - \sum_{j=N+1}^{M} a_h(w_j, w_i) g_j\right] = \\ &= \boldsymbol{\theta}^T \left[\mathbf{K}\boldsymbol{\Delta} - \mathbf{F}\right] , \end{aligned} \tag{3.34}$$

kde $\boldsymbol{\theta} = (\Theta_1, \Theta_2, \ldots, \Theta_N)^T$, $\mathbf{K} = \{k_{ij}\}_{i,j=1}^{N}$ pro $k_{ij} = a_h(w_j, w_i)$, $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, \ldots, \Delta_N)^T$ a $\mathbf{F} = (F_1, F_2, \ldots, F_N)^T$ pro $F_i = L_h(w_i) - \sum_{j=N+1}^{M} a_h(w_j, w_i) g_j$. Protože $\boldsymbol{\theta}$ je libovolný vektor, musí platit

$$\mathbf{K}\boldsymbol{\Delta} = \mathbf{F}. \tag{3.35}$$

Matice $\mathbf{K}$ je řídká a při vhodném očíslování uzlů je pásová. Pro $\mathbf{r} = \mathbf{o}$ je $\mathbf{K}$ pozitivně definitní a tedy regulární. Pro $\mathbf{r} \neq \mathbf{o}$ je $\mathbf{K}$ nesymetrická, postačující podmínkou regularity matice $\mathbf{K}$ je dostatečně malé h neboli dostatečně jemná triangulace. Vyřešením soustavy lineárních rovnic (3.35) získáme $\Delta_j = U(P_j)$, $j = 1, 2, \ldots, N$.

Za předpokladu dostatečné hladkosti slabého řešení u pro chybu platí

$$u - U = O(h^2)\,, \qquad u_x - U_x = O(h)\,, \quad u_y - U_y = O(h). \tag{3.36}$$

Algoritmus. Matici $\mathbf{K}$ a vektor $\mathbf{F}$ sestavíme pomocí elementárních matic $\mathbf{K}^{T_e}$ a elementárních vektorů $\mathbf{F}^{T_e}$ pro $T_e \in \mathcal{T}$ a elementárních matic $\mathbf{K}^{S_e}$ a elementárních vektorů $\mathbf{F}^{S_e}$ pro $S_e \in \mathcal{S}$. Matici $\mathbf{K}$ budeme nazývat také globální maticí tuhosti a vektor $\mathbf{F}$ globálním vektorem zatížení.

Elementární matice a elementární vektor na trojúhelníku. Nechť `ELEM` je tabulka typu $N_T \times 3$, která v řádku e obsahuje čísla vrcholů trojúhelníka T_e. Uvažme jeden konkrétní trojúhelník T_e triangulace $\mathcal{T}$ s vrcholy $P_1^e = [x_1^e, y_1^e]$, $P_2^e = [x_2^e, y_2^e]$ a $P_3^e = [x_3^e, y_3^e]$. Pro uzel P_r^e, $r = 1, 2, 3$, je r lokálním číslem uzlu na trojúhelníku T_e. Globálním číslem uzlu P_r^e je číslo $i = \texttt{ELEM}(e, r)$. P_r^e a P_i jsou tedy jen různá označení téhož uzlu.

Řešení U a testovací funkce v je na trojúhelníku T_e tvaru

$$\begin{aligned} U(x,y)\Big|_{T_e} &\equiv U^e(x,y) = \Delta_1^e w_1^e(x,y) + \Delta_2^e w_2^e(x,y) + \Delta_3^e w_3^e(x,y)\,, \\ v(x,y)\Big|_{T_e} &\equiv v^e(x,y) = \Theta_1^e w_1^e(x,y) + \Theta_2^e w_2^e(x,y) + \Theta_3^e w_3^e(x,y)\,, \end{aligned} \tag{3.37}$$

kde $\Delta_r^e = U(P_r^e)$, $\Theta_r^e = v(P_r^e)$ a kde

$$w_r^e(x,y) = a_r^e x + b_r^e y + c_r^e \tag{3.38}$$

je bázová funkce příslušná k uzlu P_r^e, $r = 1, 2, 3$. Z (3.37) a (3.38) dostaneme

$$\nabla U^e = \begin{pmatrix} \partial U^e/\partial x \\ \partial U^e/\partial y \end{pmatrix} = \mathbf{B}^e \boldsymbol{\Delta}^{T_e}, \quad \nabla v^e = \begin{pmatrix} \partial v^e/\partial x \\ \partial v^e/\partial y \end{pmatrix} = \mathbf{B}^e \boldsymbol{\theta}^{T_e},$$

$$\text{kde} \quad \mathbf{B}^e = \begin{pmatrix} a_1^e & a_2^e & a_3^e \\ b_1^e & b_2^e & b_3^e \end{pmatrix} \quad \text{a} \quad \boldsymbol{\Delta}^{T_e} = \begin{pmatrix} \Delta_1^e \\ \Delta_2^e \\ \Delta_3^e \end{pmatrix}, \quad \boldsymbol{\theta}^{T_e} = \begin{pmatrix} \Theta_1^e \\ \Theta_2^e \\ \Theta_3^e \end{pmatrix}$$

jsou vektory parametrů na trojúhelníku T_e. Koeficienty a_r^e a b_r^e lze snadno spočítat z rovnic

$$w_r^e(P_s^e) = \begin{cases} 1 \\ 0 \end{cases} \quad \text{pro} \quad \begin{cases} r = s\,, \\ r \neq s\,, \end{cases} \qquad r, s = 1, 2, 3,$$

nebo-li maticově

$$\begin{pmatrix} x_1^e & y_1^e & 1 \\ x_2^e & y_2^e & 1 \\ x_3^e & y_3^e & 1 \end{pmatrix} \begin{pmatrix} a_1^e & a_2^e & a_3^e \\ b_1^e & b_2^e & b_3^e \\ c_1^e & c_2^e & c_3^e \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Označíme-li

$$\mathbf{D}^e = \begin{pmatrix} x_1^e & y_1^e & 1 \\ x_2^e & y_2^e & 1 \\ x_3^e & y_3^e & 1 \end{pmatrix}, \quad \text{pak} \quad \mathbf{B}^e = [\mathbf{D}^e]^{-1}_{[1:2,1:3]}\,,$$

tj. matice $\mathbf{B}^e$ je rovna prvním dvěma řádkům matice inverzní k matici $\mathbf{D}^e$.

Nechť $P_0^e = \frac{1}{3}(P_1^e + P_2^e + P_3^e)$ je těžiště trojúhelníka T_e. Užitím kvadraturní formule (3.30) pak obdržíme

$$Q^{T_e}(\,p\,(\nabla U \cdot \nabla v)) = |T_e|\,p(P_0^e)\left(\mathbf{B}^e\boldsymbol{\theta}^{T_e}\right)^T\mathbf{B}^e\boldsymbol{\Delta}^{T_e} = [\boldsymbol{\theta}^{T_e}]^T\mathbf{K}^{T_e,1}\boldsymbol{\Delta}^{T_e}, \tag{3.39}$$

kde elementární matice

$$\mathbf{K}^{T_e,1} = |T_e|\,p_0^e\,[\mathbf{B}^e]^T\mathbf{B}^e \tag{3.40}$$

a

$$|T_e| = \tfrac{1}{2}|\det(\mathbf{D}^e)|$$

je plocha trojúhelníka T_e.

Označíme-li $\mathbf{w}^e = (w_1^e, w_2^e, w_3^e)^T$, pak $U^e = [\mathbf{w}^e]^T\boldsymbol{\Delta}^{T_e}$ podle (3.37). Protože

$$-U^e(\mathbf{r}\cdot\nabla v^e) = -[\nabla v^e]^T\mathbf{r}\,U^e = -[\mathbf{B}^e\boldsymbol{\theta}^{T_e}]^T\mathbf{r}[\mathbf{w}^e]^T\boldsymbol{\Delta}^{T_e} = -[\boldsymbol{\theta}^{T_e}]^T[\mathbf{B}^e]^T\mathbf{r}[\mathbf{w}^e]^T\boldsymbol{\Delta}^{T_e},$$

pomocí kvadraturní formule (3.30) dostaneme

$$Q^{T_e}(-U\,(\mathbf{r}\cdot\nabla v)) = [\boldsymbol{\theta}^{T_e}]^T\left[-|T_e|[\mathbf{B}^e]^T\mathbf{r}(P_0^e)(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3})\right]\boldsymbol{\Delta}^{T_e} = [\boldsymbol{\theta}^{T_e}]^T\mathbf{K}^{T_e,2}\boldsymbol{\Delta}^{T_e}, \tag{3.41}$$

kde elementární matice

$$\mathbf{K}^{T_e,2} = -\tfrac{1}{3}|T_e|[\mathbf{B}^e]^T\mathbf{r}_0^e(1,1,1) \equiv (\mathbf{s}^e,\mathbf{s}^e,\mathbf{s}^e) \quad \text{pro} \quad \mathbf{s}^e = -\tfrac{1}{3}|T_e|[\mathbf{B}^e]^T\mathbf{r}_0^e. \tag{3.42}$$

Užitím kvadraturní formule (3.31) dostaneme

$$\begin{aligned} Q^{T_e}(qUv) =& \tfrac{1}{3}|T_e|[q(P_1^e)U(P_1^e)v(P_1^e) + q(P_2^e)U(P_2^e)v(P_2^e) + q(P_3^e)U(P_3^e)v(P_1^e)] = \\ & \tfrac{1}{3}|T_e|[q(P_1^e)\Theta_1^e\Delta_1^e + q(P_2^e)\Theta_2^e\Delta_2^e + q(P_3^e)\Theta_3^e\Delta_3^e] = [\boldsymbol{\theta}^{T_e}]^T\mathbf{K}^{T_e,3}\boldsymbol{\Delta}^{T_e}, \end{aligned} \tag{3.43}$$

kde elementární matice

$$\mathbf{K}^{T_e,3} = \tfrac{1}{3}|T_e|\begin{pmatrix} q_1^e & 0 & 0 \\ 0 & q_2^e & 0 \\ 0 & 0 & q_3^e \end{pmatrix}. \tag{3.44}$$

Celkem tak

$$\sum_{T_e\in\mathcal{T}}\left\{Q^{T_e}(\,p\,(\nabla U\cdot\nabla v)) + Q^{T_e}(-U\,(\mathbf{r}\cdot\nabla v)) + Q^{T_e}(qUv)\right\} = [\boldsymbol{\theta}^{T_e}]^T\mathbf{K}^{T_e}\boldsymbol{\Delta}^{T_e}, \tag{3.45}$$

kde

$$\mathbf{K}^{T_e} = \mathbf{K}^{T_e,1} + \mathbf{K}^{T_e,2} + \mathbf{K}^{T_e,3} \tag{3.46}$$

je elementární matice na trojúhelníku T_e. Užitím kvadraturní formule (3.31) obdržíme

$$\begin{aligned} Q^{T_e}(fv) =& \tfrac{1}{3}|T_e|[f(P_1^e)v(P_1^e) + f(P_2^e)v(P_2^e) + f(P_3^e)v(P_3^e)] = \\ & \tfrac{1}{3}|T_e|[f(P_1^e)\Theta_1^e + f(P_2^e)\Theta_2^e + f(P_3^e)\Theta_3^e] = [\boldsymbol{\theta}^{T_e}]^T\mathbf{F}^{T_e}, \end{aligned} \tag{3.47}$$

kde

$$\mathbf{F}^{T_e} = \tfrac{1}{3}|T_e| \begin{pmatrix} f_1^e \\ f_2^e \\ f_3^e \end{pmatrix} \tag{3.48}$$

je elementární vektor na trojúhelníku T_e.

Elementární matice a elementární vektor na straně. Nechť `SIDE` je tabulka typu $N_S \times 2$, která v řádku e obsahuje čísla krajních bodů strany S_e. Uvažme konkrétní stranu S_e hranice Γ_2 s koncovými body $P_1^e = [x_1^e, y_1^e]$ a $P_2^e = [x_2^e, y_2^e]$. Pro uzel P_r^e, $r = 1, 2$, je r lokálním číslem uzlu na straně S_e. Globálním číslem uzlu P_r^e je číslo $i = \texttt{SIDE}(e, r)$. P_r^e a P_i jsou tedy jen různá označení téhož uzlu.

Řešení U a testovací funkce v je na straně S_e tvaru

$$U(x,y)\Big|_{S_e} \equiv U^e(x,y) = \Delta_1^e w_1^e(x,y) + \Delta_2^e w_2^e(x,y)\,,$$

$$v(x,y)\Big|_{S_e} \equiv v^e(x,y) = \Theta_1^e w_1^e(x,y) + \Theta_2^e w_2^e(x,y)\,,$$

kde $\Delta_r^e = U(P_r^e)$, $\Theta_r^e = v(P_r^e)$ a kde $w_r^e(x,y)$ je bázová funkce příslušná k uzlu P_r^e, $r = 1, 2$. Nechť $\mathbf{n}_e = (n_1^e, n_2^e)^T$ je jednotkový vektor vnější normály na straně S_e. Užitím formule (3.32) obdržíme

$$Q^{S_e}((\alpha + \mathbf{r}\cdot\mathbf{n})\,Uv) = \tfrac{1}{2}|S_e|\{[(\alpha + \mathbf{r}\cdot\mathbf{n}_e)Uv](P_1^e) + [(\alpha + \mathbf{r}\cdot\mathbf{n}_e)Uv](P_2^e)\} = \tag{3.49}$$

$$\tfrac{1}{2}|S_e|[\Theta_1^e(\alpha(P_1^e) + \mathbf{r}(P_1^e)\cdot\mathbf{n}_e]\Delta_1^e + \Theta_2^e(\alpha(P_2^e) + \mathbf{r}(P_2^e)\cdot\mathbf{n}_e)\Delta_2^e] = [\boldsymbol{\theta}^{S_e}]^T\mathbf{K}^{S_e}\boldsymbol{\Delta}^{S_e},$$

kde

$$\mathbf{K}^{S_e} = \tfrac{1}{2}|S_e| \begin{pmatrix} \alpha_1^e + \mathbf{r}_1^e\cdot\mathbf{n}_e & 0 \\ 0 & \alpha_2^e + \mathbf{r}_2^e\cdot\mathbf{n}_e \end{pmatrix} \tag{3.50}$$

je elementární matice na straně S_e a

$$\boldsymbol{\Delta}^{S_e} = \begin{pmatrix} \Delta_1^e \\ \Delta_2^e \end{pmatrix}, \qquad \boldsymbol{\theta}^{S_e} = \begin{pmatrix} \Theta_1^e \\ \Theta_2^e \end{pmatrix}$$

jsou vektory parametrů na straně S_e. Podobně odvodíme

$$Q^{S_e}(\beta v) = \tfrac{1}{2}|S_e|[\beta(P_1^e)v(P_1^e) + \beta(P_2^e)v(P_2^e)] =$$

$$\tfrac{1}{2}|S_e|[\beta(P_1^e)\Theta_1^e + \beta(P_2^e)\Theta_2^e] = [\boldsymbol{\theta}^{S_e}]^T\mathbf{F}^{S_e}, \tag{3.51}$$

kde

$$\mathbf{F}^{S_e} = \tfrac{1}{2}|S_e| \begin{pmatrix} \beta_1^e \\ \beta_2^e \end{pmatrix} \tag{3.52}$$

je elementární vektor na straně S_e.

Sestavení soustavy rovnic. Zkombinujeme-li (3.34), (3.29), (3.45), (3.47), (3.49) a (3.51) vidíme, že pro MKP řešení U a libovolnou testovací funkci $v \in V_h$ platí

$$0 = a_h(U,v) - L_h(v) = \boldsymbol{\theta}^T \left[\mathbf{K}\boldsymbol{\Delta} - \mathbf{F}\right] =$$

$$= \sum_{T_e \in \mathcal{T}} [\boldsymbol{\theta}^{T_e}]^T \left[\mathbf{K}^{T_e}\boldsymbol{\Delta}^{T_e} - \mathbf{F}^{T_e}\right] + \sum_{S_e \in \mathcal{S}} [\boldsymbol{\theta}^{S_e}]^T \left[\mathbf{K}^{S_e}\boldsymbol{\Delta}^{S_e} - \mathbf{F}^{S_e}\right].$$

Z této rovnosti lze odvodit pravidla pro sestavení globální matice $\mathbf{K}$ a globálního vektoru $\mathbf{F}$ z lokálních matic $\mathbf{K}^{T_e}$, $\mathbf{K}^{S_e}$ a lokálních vektorů $\mathbf{F}^{T_e}$, $\mathbf{F}^{S_e}$. Postupujeme podle následujícího algoritmu:

1) Matici $\mathbf{K}$ řádu N a sloupcový vektor $\mathbf{F}$ délky N naplníme nulami.

2) Pro každý trojúhelník $T_e \in \mathcal{T}$ sestavíme elementární matici $\mathbf{K}^{T_e} = \{k_{rs}^{T_e}\}_{r,s=1}^3$, viz (3.46), a elementární vektor $\mathbf{F}^{T_e} = \{F_r^{T_e}\}_{r=1}^3$, viz (3.48).

   Pro $r,s = 1,2,3$ položíme $i = \texttt{ELEM}(e,r)$, $j = \texttt{ELEM}(e,s)$ a

$$\text{pokud} \quad \begin{cases} i \le N \text{ a } j \le N, \\ i \le N \text{ a } j > N, \\ i \le N, \end{cases} \quad \text{provedeme} \quad \begin{cases} k_{ij} \leftarrow k_{ij} + k_{rs}^{T_e}, \\ F_i \leftarrow F_i - k_{rs}^{T_e} g(P_j), \\ F_i \leftarrow F_i + F_r^{T_e}. \end{cases}$$

3) Pro každou stranu $S_e \in \mathcal{S}$ sestavíme elementární matici $\mathbf{K}^{S_e} = \{k_{rs}^{S_e}\}_{r,s=1}^2$, viz (3.50), a elementární vektor $\mathbf{F}^{S_e} = \{F_r^{S_e}\}_{r=1}^2$, viz (3.52).

   Pro $r = 1,2$ položíme $i = \texttt{SIDE}(e,r)$ a

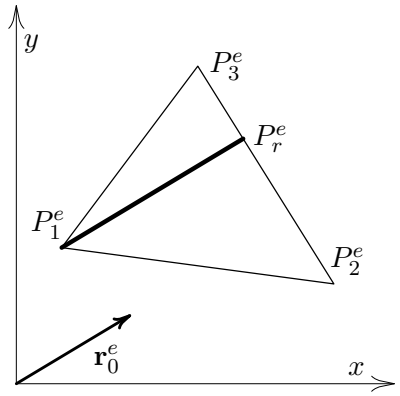
   pokud $i \le N$, provedeme $k_{ii} \leftarrow k_{ii} + k_{rr}^{S_e}$, $F_i \leftarrow F_i + F_r^{S_e}$.

Dominantní konvekce. Jednou z možností, jak se vypořádat s dominantní konvekcí, je postup známý jako *metoda umělé směrové difúze*, zkratka *SD* (podle anglického *Streamline Diffusion*), viz [5]. Vyjdeme z modifikované diskrétní slabé formulace

$$\text{najít } U \in W_h \text{ splňující } a_h(U,v) + c_h(U,v) = L_h(v), \quad \forall v \in V_h, \tag{3.28'}$$

kde

$$c_h(U,v) = \sum_{i=1}^{N} Q^{T_e} \left(\delta_e \left(\mathbf{r}\cdot\nabla U\right)\left(\mathbf{r}\cdot\nabla v\right)\right).$$



Obr. 3.8. $h_r^e = |\overline{P_1^e P_r^e}|$

Nechť $|\mathbf{r}| = \sqrt{r_1^2 + r_2^2}$ je délka vektoru $\mathbf{r}$ a nechť h_r^e je průměr trojúhelníka T_e ve směru vektoru $\mathbf{r}_0^e$, viz Obr. 3.8. Pak

$$\delta_e = \frac{h_r^e}{2|\mathbf{r}|}|\sigma_e|.$$

Parametr σ_e lze zvolit následovně:

$$\sigma_e = \coth\kappa_e - \frac{1}{\kappa_e}, \quad \text{kde} \quad \kappa_e = \frac{r_0^e h_r^e}{2p_0^e}$$

je lokální Pecletovo číslo. Pomocí obdélníkové

formule dostaneme

$$Q^{T_e}\left(\delta_e\,(\mathbf{r}\cdot\nabla U)\,(\mathbf{r}\cdot\nabla v)\right) = [\boldsymbol{\theta}^{T_e}]^T\mathbf{K}^{e4}\boldsymbol{\Delta}^{T_e}, \quad \text{kde} \quad \mathbf{K}^{e4} = \delta_e|T_e|[\mathbf{B}^e]^{T_e}\mathbf{r}_0^e[\mathbf{r}_0^e]^T\mathbf{B}^e.$$

Položíme $\mathbf{K}^e = \mathbf{K}^{e1} + \mathbf{K}^{e2} + \mathbf{K}^{e3} + \mathbf{K}^{e4}$ a sestavíme soustavu rovnic podle výše uvedeného algoritmu.

Rovnost (3.28') dostaneme z rovnosti (3.28) tak, že v ní nahradíme „difúzní člen" $Q^{T_e}(p\nabla U\cdot\nabla v)$ členem $Q^{T_e}\big([(p\mathbf{I}+\delta_e\mathbf{r}\mathbf{r}^T)\nabla U]\cdot\nabla v\big)$, kde $\mathbf{I}$ je jednotková matice řádu 2. Matice $\delta_e\mathbf{r}\mathbf{r}^T$ reprezentuje umělou směrovou difúzi.

Pro řešení konvekčně-difúzních úloh na bázi metody konečných prvků se používají důmyslnější postupy, zejména metoda SUPG (podle anglického *Streamline Upwind Petrov-Galerkin*), viz [5], nebo metoda DG-FEM (podle anglického *Discontinuous Galerkin Finite Element Method*), viz [7].

3.2. Úloha parabolického typu

Formulace úlohy. Hledáme funkci $u(x,t)$ definovanou pro $x\in\langle 0,\ell\rangle$, $t\in\langle 0,T\rangle$, která vyhovuje diferenciální rovnici

$$c(x)\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(p(x)\frac{\partial u}{\partial x}\right) + q(x)u = f(x,t)\,, \qquad x\in(0,\ell),\ t\in(0,T)\,, \tag{3.53}$$

Dirichletovým okrajovým podmínkám

$$u(0,t) = g_0(t)\,, \qquad u(\ell,t) = g_\ell(t), \qquad t\in(0,T), \tag{3.54}$$

nebo Newtonovým okrajovým podmínkám

$$\begin{aligned} p(0)\frac{\partial u(0,t)}{\partial x} &= \alpha_0 u(0,t) - \beta_0(t), \\ -p(\ell)\frac{\partial u(\ell,t)}{\partial x} &= \alpha_\ell u(\ell,t) - \beta_\ell(t), \end{aligned} \qquad t\in(0,T), \tag{3.55}$$

a počáteční podmínce

$$u(x,0) = \varphi(x), \qquad x\in(0,\ell)\,. \tag{3.56}$$

Možný je také případ, kdy je na jednom okraji předepsána Dirichletova podmínka a na druhém podmínka Newtonova. Úloha (3.53) – (3.56) popisuje *nestacionární vedení tepla v tyči* délky ℓ, T je doba trvání děje. Proměnná x je prostorová, t má význam času, $u(x,t)$ je teplota v bodě x a v čase t, funkce p, q, f, g_0, g_ℓ, β_0, β_ℓ a konstanty α_0, α_ℓ mají stejný význam jako v kapitole 2, c je objemová tepelná kapacita (tj. tepelná kapacita vztažená na jednotku objemu), φ je teplota tyče v čase $t=0$. Tepelné toky se často uvažují ve tvaru $\beta_0(t) = \alpha_0 u_0^e(t)$, $\beta_\ell(t) = \alpha_\ell u_\ell^e(t)$, kde u_0^e resp. u_ℓ^e je teplota okolí levého resp. pravého konce tyče.

Předpokládejme že funkce c, p, q, f, g_0, g_ℓ, β_0, β_ℓ a φ jsou dostatečně hladké a že jsou splněny podmínky nezápornosti $c\ge c_0>0$, $p\ge p_0>0$, $q\ge 0$, $\alpha_0\ge 0$, $\alpha_\ell\ge 0$.

Aby existovalo klasické řešení úlohy (3.53) – (3.56), je třeba pro Dirichletovy okrajové podmínky připojit ještě tzv. *podmínky souhlasu*

$$\varphi(0) = g_0(0)\,, \quad \varphi(\ell) = g_\ell(0)\,.$$

Tyto vztahy, vyjadřující soulad počáteční podmínky a Dirichletovy okrajové podmínky, nebývají v aplikacích obvykle splněny. Také funkce c, p, q, f, g_0, g_ℓ, β_0, β_ℓ bývají často jen po částech spojité. V tom případě má úloha (3.53) – (3.56) pouze tzv. slabé řešení (jehož přesnou definici zde uvádět nebudeme). Pro praktické účely je slabé řešení zcela vyhovující a numerické metody, které si uvedeme, budou poskytovat přibližné hodnoty takového slabého řešení.

Úloha (3.53) – (3.56) je okrajovou úlohou druhého řádu vzhledem k proměnné x a počáteční úlohou prvního řádu vzhledem k proměnné t. Při její diskretizaci proto zkombinujeme postupy z prvních dvou kapitol.

Prostorová diskretizace se provádí metodami kapitoly 2. Pro jednoduchost budeme uvažovat úlohu s Dirichletovými okrajovými podmínkami, diferenční metodu a rovnoměrné dělení intervalu $\langle 0, \ell\rangle$ s krokem $h = \ell/N$, tj. $x_i = ih$, $i = 0, 1, \ldots, N$. Budeme požadovat, aby rovnice (3.53) byla splněna ve vnitřních uzlech x_i, $i = 1, 2, \ldots, N-1$, tj.

$$c(x_i)\frac{\partial u(x_i,t)}{\partial t} - \left[\frac{\partial}{\partial x}\left(p\frac{\partial u}{\partial x}\right)\right](x_i,t) + q(x_i)u(x_i,t) = f(x_i,t)\,.$$

Derivaci podle x vyjádříme pomocí diference analogicky jako v (2.14), tj.

$$-\left[\frac{\partial}{\partial x}\left(p\,\frac{\partial u}{\partial x}\right)\right](x_i,t) = \frac{1}{h^2}\Big[-p_{i-1/2}u(x_{i-1},t)+ \\ + \big[p_{i-1/2}+p_{i+1/2}\big]u(x_i,t) - p_{i+1/2}u(x_{i+1},t)\Big] + O(h^2)\,,$$

Zanedbáme-li chybu, dostaneme soustavu rovnic

$$c_i\dot{u}_i(t) + \frac{1}{h^2}\Big[-p_{i-1/2}u_{i-1}(t) + (p_{i-1/2}+p_{i+1/2}+h^2q_i)u_i(t) - \\ - p_{i+1/2}u_{i+1}(t)\Big] = f_i(t)\,, \qquad i = 1,2,\ldots,N-1, \quad t\in(0,T)\,, \tag{3.57}$$

v níž $u_i(t)$ je aproximace $u(x_i,t)$, $\dot{u}_i(t) = \dfrac{\mathrm{d}u_i(t)}{\mathrm{d}t}$ je aproximace $\partial_t u(x_i,t)$ a $f_i(t) = f(x_i,t)$. Z okrajových podmínek (3.54) máme

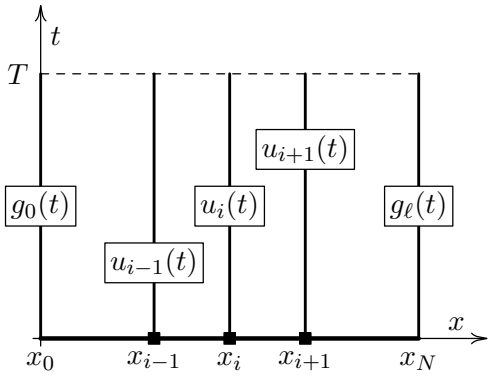
$$u_0(t) = g_0(t)\,, \qquad u_N(t) = g_\ell(t)\,, \qquad t\in(0,T), \tag{3.58}$$

což dosadíme do (3.57) pro $i = 1$ a $i = N-1$. Z počáteční podmínky obdržíme

$$u_i(0) = \varphi(x_i), \qquad i = 1,2,\ldots,N-1\,. \tag{3.59}$$

(3.57) je soustava obyčejných diferenciálních rovnic prvního řádu pro $N-1$ hledaných funkcí $u_i(t)$, $i = 1,2,\ldots,N-1$, s počátečními podmínkami (3.59). Původní úlohu,

tj. určení jedné funkce $u(x,t)$, která na obdélníku $\langle 0,\ell\rangle \times \langle 0,T\rangle$vyhovuje (3.53), (3.54) a (3.56), jsme nahradili počáteční úlohou pro soustavu $N-1$ obyčejných diferenciálních rovnic s neznámými funkcemi $u_i(t)$ definovanými na úsečkách $x = x_i$, $i = 1,2,\ldots,N-1$, $t \in \langle 0,T\rangle$. Tento postup se nazývá *metoda přímek* (anglicky *method of lines*, stručně MOL). Postup, při němž se diskretizace provádí jen vzhledem k některým (nezávisle) proměnným, se nazývá *semidiskretizace*.



Obr. 3.9. Metoda přímek

Počáteční problém (3.57)–(3.59) jsme tedy získali semidiskretizací úlohy (3.53), (3.54) a (3.56) vzhledem k prostorové proměnné x. Úlohu (3.57) – (3.59) můžeme zapsat maticově ve tvaru

$$\mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{F}(t),\ \ t \in (0,T),\ \ \mathbf{u}(0) = \boldsymbol{\varphi}\,, \quad (3.60)$$

kde $\mathbf{C}$ je diagonální matice s kladnými prvky na diagonále, tzv. *matice tepelné kapacity*,

$\mathbf{K}$ je třídiagonální pozitivně definitní matice známá jako *matice tepelné vodivosti*, $\mathbf{F}(t)$ je tzv. *vektor tepelných zdrojů*, $\mathbf{u}(t) = (u_1(t), u_2(t), \ldots, u_{N-1}(t))^T$ je vektor neznámých funkcí a $\boldsymbol{\varphi} = (\varphi(x_1), \varphi(x_2), \ldots, \varphi(x_{N-1}))^T$ je vektor počátečních teplot.

Newtonova okrajová podmínka. Protože Newtonovy okrajové podmínky se v úloze vedení tepla používají nejčastěji, uvedeme si rovnice reprezentující diskterizaci Newtonovy okrajové podmínky vzhledem k proměnné x v levém resp. pravém koncovém bodu intervalu $\langle 0,\ell\rangle$. Bude-li tedy Newtonova okrajová podmínka předepsána v bodě $x = 0$, zařadíme před rovnice (3.57) jako první rovnici

$$\frac{1}{2}c_0\dot{u}_0(t) + \frac{1}{h^2}\left[(p_{1/2} + h\alpha_0 + \tfrac{1}{2}h^2q_0)u_0(t) - p_{1/2}u_1(t)\right] = \frac{1}{h}\beta_0(t) + \frac{1}{2}f_0(t), \quad (3.57a)$$

a bude-li Newtonova okrajová podmínka předepsána v bodě $x = \ell$, přidáme za poslední z rovnic (3.57) rovnici

$$\frac{1}{2}c_N\dot{u}_N(t) + \frac{1}{h^2}\left[-p_{N-1/2}u_{N-1}(t) + (p_N + h\alpha_\ell + \tfrac{1}{2}h^2q_N)u_N(t)\right] = \frac{1}{h}\beta_\ell(t) + \frac{1}{2}f_N(t). \quad (3.57b)$$

Výsledná matice $\mathbf{K}$ je opět třídiagonální a symetrická. Jestliže $\alpha_0 = \alpha_\ell = 0$ a $q_i = 0$, $i = 0,1,\ldots,N$, pak je matice $\mathbf{K}$ pozitivně semidefinitní, ve všech ostatních případech je $\mathbf{K}$ pozitivně definitní. Připomeňme: matice $\mathbf{A}$ je pozitivně semidefinitní, jestliže je symetrická a $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ pro každý vektor $\mathbf{x}$.

Časová diskretizace. Prozkoumejme nejdříve charakter počáteční úlohy (3.60) za zjednodušujících předpokladů. Pro $c = p = 1$, $q = f = 0$, $u(0,t) = u(\ell,t) = 0$, lze úlohu (3.60) zapsat ve tvaru

$$\dot{\mathbf{u}} = \mathbf{A}\mathbf{u}\,, \quad t \in (0,T)\,, \qquad \mathbf{u}(0) = \boldsymbol{\varphi}\,, \quad (3.61)$$

kde

$$\mathbf{A} = -\frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 \end{pmatrix}. \tag{3.62}$$

Vlastní čísla matice $\mathbf{A}$

$$\lambda_i = -\frac{4N^2}{\ell^2}\sin^2\frac{\pi i}{2N}, \qquad i = 1, 2, \dots, N-1,$$

jsou reálná záporná a $\max_i |\lambda_i| \to \infty$ pro $N \to \infty$. Podle (1.31) je tedy problém (3.61) pro velké N tuhý. Totéž platí také pro úlohu (3.60), tj. jde o tuhý problém. Vzhledem k (1.32) je proto vhodné řešit počáteční problém (3.60) metodou, jejíž oblast absolutní stability obsahuje celou zápornou reálnou poloosu komplexní roviny. Metody s touto vlastností se nazývají *A_0-stabilní*. Patří mezi ně všechny metody doporučené v kapitole 1.5 pro řešení tuhých problémů, zejména tedy implicitní Eulerova metoda, lichoběžníková metoda nebo metody zpětného derivování. V Matlabu lze použít některý z programů `ode23s`, `ode23t`, `ode23tb` a `ode15s`. Při řešení úloh vedení tepla se často používá metoda časové diskretizace známá jako

Theta metoda, kterou v následujícím textu stručně popíšeme. Nechť tedy $0 = t_0 < t_1 < \cdots < t_Q = T$ je dělení intervalu $\langle 0, T\rangle$, $\tau_n = t_{n+1} - t_n$ je délka kroku a $\mathbf{u}^n$ je přibližné řešení v čase t_n, tj. $u_i^n \approx u_i(t_n) \approx u(x_i, t_n)$. Nechť θ je pevně zvolené číslo z intervalu $\langle 0, 1\rangle$. Označme $t_{n+\theta} = t_n + \theta\tau_n$. Rovnici (3.60) zapíšeme pro $t = t_{n+\theta}$ a dostaneme

$$\mathbf{C}\dot{\mathbf{u}}^{n+\theta} + \mathbf{K}\mathbf{u}^{n+\theta} = \mathbf{F}^{n+\theta}, \tag{3.63}$$

kde $\mathbf{F}^{n+\theta} = \mathbf{F}(t_{n+\theta})$, $\mathbf{u}^{n+\theta} = \mathbf{u}(t_{n+\theta})$, $\dot{\mathbf{u}}^{n+\theta} = \dot{\mathbf{u}}(t_{n+\theta})$. Předpokládejme, že $\mathbf{u}(t)$ je na intervalu $\langle t_n, t_{n+1}\rangle$ lineární funkce, tj.

$$\mathbf{u}(t) = \mathbf{u}^n + \frac{t - t_n}{\tau_n}(\mathbf{u}^{n+1} - \mathbf{u}^n).$$

Pak $\dot{\mathbf{u}}^{n+\theta} = \dfrac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau_n}$ a $\mathbf{u}^{n+\theta} = (1-\theta)\mathbf{u}^n + \theta\mathbf{u}^{n+1}$. Po dosazením do (3.63) a malé úpravě dostaneme *θ-metodu*

$$(\mathbf{C} + \tau_n\theta\mathbf{K})\mathbf{u}^{n+1} = (\mathbf{C} - \tau_n(1-\theta)\mathbf{K})\mathbf{u}^n + \tau_n\mathbf{F}^{n+\theta}. \tag{3.64}$$

Snadno ověříme, že pro $\frac{1}{2} \le \theta \le 1$ je θ-metoda A-stabilní a tedy také A_0-stabilní, a že pro $0 \le \theta < \frac{1}{2}$ je oblast absolutní stability θ-metody omezená a interval absolutní stability je interval $(-2/(1-2\theta), 0)$. Pokud jde o přesnost, θ-metoda je řádu 1 pro $\theta \ne \frac{1}{2}$ a řádu 2 pro $\theta = \frac{1}{2}$. Všimněte si, že z θ-metody dostaneme pro $\theta = 0$ EE metodu, pro $\theta = 1$ IE

metodu a pro $\theta = \frac{1}{2}$ TR metodu. Metoda (3.64) pro $\theta = \frac{1}{2}$ je známa jako *Crankova-Nicolsonova metoda*. Často se zapisuje v analogickém tvaru

$$(\mathbf{C} + \tfrac{1}{2}\tau_n \mathbf{K})\mathbf{u}^{n+1} = (\mathbf{C} - \tfrac{1}{2}\tau_n \mathbf{K})\mathbf{u}^n + \tfrac{1}{2}\tau_n(\mathbf{F}^n + \mathbf{F}^{n+1})\,. \tag{3.65}$$

Přesnost i stabilita metody (3.64) pro $\theta = \frac{1}{2}$ a metody (3.65) je stejná.

Matice $\mathbf{C} + \tau_n \theta \mathbf{K}$ soustavy (3.64) nezávisí na čase a je pozitivně definitní. Soustavu (3.64) je proto účelné řešit pomocí Choleského rozkladu, který stačí provést jen jednou.

Nelineární úloha vedení tepla vznikne tehdy, když některé z funkcí $c, p, q, f, \alpha, \beta$ závisejí na teplotě u. Odpovídající počáteční úlohu

$$\mathbf{C}(\mathbf{u})\dot{\mathbf{u}} = \mathbf{F}(t, \mathbf{u}) - \mathbf{K}(\mathbf{u})\mathbf{u}, \quad t \in (0, T)\,, \qquad \mathbf{u}(0) = \boldsymbol{\varphi}, \tag{3.60'}$$

lze v Matlabu vyřešit programem `ode23t` nebo `ode15s`. Řešení jednodimenzionálního parabolického problému, a to i nelineárního, umožňuje matlabovský program `pdepe`.

Rovnice s konvekčním členem

$$c(x)\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(p(x)\frac{\partial u}{\partial x} - r(x)u\right) + q(x)u = f(x,t), \quad x \in (0,\ell),\ t \in (0,T), \tag{3.53'}$$

popisuje šíření tepla v tekutině, $r = cv$, kde $v(x)$ je rychlost proudění. Diskretizaci konvekčního členu provedeme stejně jako v kapitole 2.2, viz (2.24), (2.28), (2.28'). Pro časovou diskretizaci úlohy s dominantní konvekcí lze doporučit metody IE, TR nebo BDF2.

3.3. Úloha hyperbolického typu

Formulace úlohy. Hledáme funkci $u(x,t)$ definovanou pro $x \in \langle 0, \ell\rangle$, $t \in \langle 0, T\rangle$, která vyhovuje diferenciální rovnici

$$\rho(x)\frac{\partial^2 u}{\partial t^2} + c(x)\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(p(x)\frac{\partial u}{\partial x}\right) + q(x)u = f(x,t), \quad x \in (0,\ell),\ t \in (0,T), \tag{3.66}$$

Dirichletovým okrajovým podmínkám

$$u(0,t) = g_0(t), \qquad u(\ell,t) = g_\ell(t), \qquad t \in (0,T), \tag{3.67}$$

nebo Newtonovým okrajovým podmínkám

$$\begin{aligned} p(0)\frac{\partial u(0,t)}{\partial x} &= \alpha_0 u(0,t) - \beta_0(t), \\ -p(\ell)\frac{\partial u(\ell,t)}{\partial x} &= \alpha_\ell u(\ell,t) - \beta_\ell(t), \end{aligned} \qquad t \in (0,T), \tag{3.68}$$

a počátečním podmínkám

$$u(x,0) = \varphi(x), \quad \frac{\partial u(x,0)}{\partial t} = \psi(x), \qquad x \in (0,\ell). \tag{3.69}$$

Možný je také případ, kdy na jednom okraji je předepsána Dirichletova podmínka a na druhém Newtonova. Tato úloha vyjadřuje *příčné kmitání tenké struny* (nebo *podélné kmitání prutu*) délky ℓ. Proměnná x je prostorová, t má význam času, $u(x,t)$ je deformace (tj. příčná výchylka pro strunu nebo podélné posunutí v případě prutu) v bodě x a v čase t, ρ je hustota, c udává tlumení (pro $c > 0$ jsou kmity tlumené, pro $c = 0$ netlumené), p charakterizuje tuhost, q odpor okolního prostředí a f vnější síly. Dirichletovy okrajové podmínky předepisují deformaci, Newtonovy okrajové podmínky vnitřní síly: α_0 resp. α_ℓ reprezentuje tuhost pružné podpory a β_0 resp. β_ℓ bodovou vnější sílu. Počáteční podmínky určují počáteční deformaci φ a její počáteční rychlost ψ.

Předpokládejme, že funkce ρ, c, p, q, f, g_0, g_ℓ, β_0, β_ℓ, φ a ψ jsou dostatečně hladké, že jsou splněny podmínky nezápornosti $\rho \geq \rho_0 > 0$, $c \geq 0$, $p \geq p_0 > 0$, $q \geq 0$, $\alpha_0 \geq 0$, $\alpha_\ell \geq 0$ a že platí podmínky souhlasu

$$\varphi(0) = g_0(0)\,, \quad \psi(0) = g_0'(0)\,, \qquad\qquad \varphi(\ell) = g_\ell(0)\,, \quad \psi(\ell) = g_\ell'(0)\,,$$

vyjadřující soulad počátečních podmínek a Dirichletových okrajových podmínek. Pak má úloha (3.66) – (3.69) jediné klasické řešení.

Podmínky souhlasu v aplikacích někdy nebývají splněny. Také funkce ρ, c, p, q, f, g_0, g_ℓ, β_0, β_ℓ bývají často jen po částech spojité. V tom případě má úloha (3.53) – (3.56) pouze tzv. slabé řešení (jehož přesnou definici zde ovšem uvádět nebudeme). Pro praktické účely je slabé řešení zcela vyhovující a numerické metody, které si uvedeme, budou poskytovat přibližné hodnoty takového slabého řešení.

Úloha (3.66) – (3.69) je okrajovou úlohou druhého řádu vzhledem k proměnné x a počáteční úlohou druhého řádu vzhledem k proměnné t. Při její diskretizaci proto opět použijeme postupy z prvních dvou kapitol.

Prostorová diskretizace se provede stejně jako u parabolického problému metodou přímek, viz kapitola 3.2. Opět předpokládáme rovnoměrné dělení a Dirichletovu okrajovou podmínku. Pro $u_i(t)$ aproximující $u(x_i,t)$ při označení $\dot{u}_i(t) = \dfrac{\mathrm{d}u_i(t)}{\mathrm{d}t}$, $\ddot{u}_i(t) = \dfrac{\mathrm{d}^2u_i(t)}{\mathrm{d}t^2}$ dostaneme soustavu rovnic

$$\rho_i\ddot{u}_i(t) + c_i\dot{u}_i(t) + \frac{1}{h^2}\Big(-p_{i-1/2}u_{i-1}(t) + [p_{i-1/2} + p_{i+1/2} + h^2q_i]u_i(t) - \tag{3.70}$$

$$- p_{i+1/2}u_{i+1}(t)\Big) = f_i(t)\,, \qquad i = 1,2,\ldots,N-1, \quad t \in (0,T)\,.$$

Z okrajových podmínek (3.67) máme

$$u_0(t) = g_0(t), \qquad u_N(t) = g_\ell(t), \qquad t \in (0,T)\,, \tag{3.71}$$

což dosadíme do (3.70) pro $i = 1$ a $i = N-1$. Z počátečních podmínek (3.69) dostaneme

$$u_i(0) = \varphi(x_i), \qquad \dot{u}_i(0) = \psi(x_i), \qquad i = 1,2,\ldots,N-1\,. \tag{3.72}$$

(3.70) je soustava obyčejných diferenciálních rovnic druhého řádu pro $N-1$ hledaných funkcí $u_i(t)$, $i = 1,2,\ldots,N-1$, s počátečními podmínkami (3.72).

Počáteční úlohu (3.70)–(3.72) můžeme zapsat maticově ve tvaru

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{F}(t)\,, \quad t \in (0,T), \qquad \mathbf{u}(0) = \boldsymbol{\varphi}, \quad \dot{\mathbf{u}}(0) = \boldsymbol{\psi}\,, \tag{3.73}$$

kde $\mathbf{M}$ je diagonální matice s kladnými prvky na diagonále, tzv. *matice hmotnosti*, $\mathbf{C}$ je diagonální matice s nezápornými prvky na diagonále, tzv. *matice útlumu*, $\mathbf{K}$ je třídiagonální pozitivně definitní matice, tzv. *matice tuhosti*, $\mathbf{F}(t)$ je tzv. *vektor (vnějšího) zatížení*, $\boldsymbol{\varphi} = (\varphi(x_1), \varphi(x_2), \ldots, \varphi(x_{N-1}))^T$, $\boldsymbol{\psi} = (\psi(x_1), \psi(x_2), \ldots, \psi(x_{N-1}))^T$ jsou vektory počátečních hodnot a $\mathbf{u}(t) = (u_1(t), u_2(t), \ldots, u_{N-1}(t))^T$ je vektor neznámých funkcí.

Časová diskretizace. Abychom mohli posoudit tuhost problému (3.73), zapíšeme ho jako počáteční problém prvního řádu,

$$\mathbf{R}\dot{\mathbf{w}} + \mathbf{S}\mathbf{w} = \mathbf{G}(t)\,, \quad t \in (0,T)\,, \qquad \mathbf{w}(0) = \boldsymbol{\kappa}\,, \tag{3.74}$$

kde

$$\mathbf{R} = \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{M} \end{pmatrix},\ \mathbf{w} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix},\ \mathbf{S} = \begin{pmatrix} \mathbf{O} & -\mathbf{I} \\ \mathbf{K} & \mathbf{C} \end{pmatrix},\ \mathbf{G}(t) = \begin{pmatrix} \mathbf{o} \\ \mathbf{F}(t) \end{pmatrix},\ \boldsymbol{\kappa} = \begin{pmatrix} \boldsymbol{\varphi} \\ \boldsymbol{\psi} \end{pmatrix},$$

přičemž $\mathbf{v} = \dot{\mathbf{u}}$, $\mathbf{I}$ je jednotková matice, $\mathbf{O}$ je nulová matice a $\mathbf{o}$ je nulový vektor.

Prozkoumejme charakter počáteční úlohy (3.74) za zjednodušujících předpokladů. Pro $\rho = p = 1$, $c = konst$, $q = f = 0$, $u(0,t) = u(\ell,t) = 0$, je $\mathbf{M} = \mathbf{I}$, $\mathbf{K} = -\mathbf{A}$ a $\mathbf{C} = c\mathbf{I}$, kde $\mathbf{A}$ je definována předpisem (3.62). Rovnice (3.74) se tak zjednoduší, dostaneme

$$\dot{\mathbf{w}} = \mathbf{P}\mathbf{w}, \quad \mathbf{w}(0) = \boldsymbol{\kappa}, \qquad \text{kde} \quad \mathbf{P} = \begin{pmatrix} \mathbf{O} & \mathbf{I} \\ \mathbf{A} & -c\mathbf{I} \end{pmatrix}. \tag{3.75}$$

Dá se ukázat, že vlastní čísla matice $\mathbf{P}$

$$\lambda_i = -\tfrac{1}{2}c \pm \sqrt{\tfrac{1}{4}c^2 - \omega_i^2}, \quad \text{kde} \quad \omega_i = \frac{2N}{\ell}\sin\frac{\pi i}{2N}\,, \quad i = 1, 2, \ldots, N-1.$$

Zřejmě $\max_i |\lambda_i| \to \infty$ pro $N \to \infty$. Pro $c = 0$ jsou vlastní čísla ryze imaginární a pro $c > 0$ mají zápornou reálnou složku. Podle (1.31) je tedy problém (3.75) pro velké N tuhý. Protože reálné složky vlastních čísel jsou zdola ohraničené, $\mathrm{Re}(\lambda_i) \geq -c$, a velikost imaginárních složek je neomezená, říkáme, že problém (3.75) je *oscilatoricky tuhý*. Problém (3.74) se chová obdobně. Vzhledem k (1.32) je proto vhodné řešit úlohu (3.75) pro $c > 0$ metodou, jejíž oblast absolutní stability obsahuje pás $R_c = \{z \in \mathbb{C} \mid -c < \mathrm{Re}(z) < 0\}$.

Pro $c = 0$ lze odvodit rovnost

$$[\mathbf{u}(t)]^T\mathbf{u}(t) + [\mathbf{v}(t)]^T\mathbf{K}^{-1}\mathbf{v}(t) = \boldsymbol{\varphi}^T\boldsymbol{\varphi} + \boldsymbol{\psi}^T\mathbf{K}^{-1}\boldsymbol{\psi}, \qquad t > 0,$$

vyjadřující stabilitu počátečního problému (3.75) vzhledem k počáteční podmínce. Pro přibližné řešení $\mathbf{w}_n \approx \mathbf{w}(n\tau)$ spočtené lichoběžníkovou metodou lze odvodit analogickou rovnost

$$\mathbf{u}_n^T\mathbf{u}_n + \mathbf{v}_n^T\mathbf{K}^{-1}\mathbf{v}_n = \boldsymbol{\varphi}^T\boldsymbol{\varphi} + \boldsymbol{\psi}^T\mathbf{K}^{-1}\boldsymbol{\psi}, \qquad n = 1, 2, \ldots$$

To je skvělé, lichoběžníková metoda aplikovaná na řešení úlohy (3.75) pro $c = 0$ vykazuje stejnou stabilitu jako přesné řešení.

Protože lichoběžníková metoda je A-stabilní, je vhodnou metodou pro každé $c \geq 0$. Z matlabovských programů lze doporučit programy `ode23t` a `ode23tb`, které jsou na lichoběžníkové metodě založeny.

Lichoběžníková metoda aplikovaná na úlohu (3.74) vede na předpis

$$(\mathbf{R} + \tfrac{1}{2}\tau_n \mathbf{S})\mathbf{w}^{n+1} = (\mathbf{R} - \tfrac{1}{2}\tau_n \mathbf{S})\mathbf{w}^n + \tfrac{1}{2}\tau_n(\mathbf{G}^n + \mathbf{G}^{n+1}), \tag{3.76}$$

viz (3.65). Pro efektivní výpočet složek $\mathbf{u}^{n+1}$ a $\mathbf{v}^{n+1}$ vektoru $\mathbf{w}^{n+1}$ je vhodné soustavu rovnic (3.76) zapsat po složkách, tj.

$$\mathbf{u}^{n+1} - \tfrac{1}{2}\tau_n \mathbf{v}^{n+1} = \mathbf{u}^n + \tfrac{1}{2}\tau_n \mathbf{v}^n,$$
$$(\mathbf{M} + \tfrac{1}{2}\tau_n \mathbf{C})\mathbf{v}^{n+1} + \tfrac{1}{2}\tau_n \mathbf{K}\mathbf{u}^{n+1} = (\mathbf{M} - \tfrac{1}{2}\tau_n \mathbf{C})\mathbf{v}^n - \tfrac{1}{2}\tau_n \mathbf{K}\mathbf{u}^n + \tfrac{1}{2}\tau_n(\mathbf{F}^{n+1} + \mathbf{F}^n).$$

Z prvé rovnice vyjádříme $\mathbf{v}^{n+1}$, viz (3.78), dosadíme do druhé rovnice a obdržíme rovnici pro výpočet $\mathbf{u}^{n+1}$:

$$\hat{\mathbf{K}}\mathbf{u}^{n+1} = \widehat{\mathbf{F}}, \qquad \text{kde} \tag{3.77}$$
$$\widehat{\mathbf{K}} = \frac{4}{\tau_n^2}\mathbf{M} + \frac{2}{\tau_n}\mathbf{C} + \mathbf{K}\,, \qquad \widehat{\mathbf{F}} = \left(\frac{4}{\tau_n^2}\mathbf{M} + \frac{2}{\tau_n}\mathbf{C} - \mathbf{K}\right)\mathbf{u}^n + \frac{4}{\tau_n}\mathbf{M}\mathbf{v}^n + \mathbf{F}^n + \mathbf{F}^{n+1}.$$

Až vypočítáme $\mathbf{u}^{n+1}$, dopočítáme

$$\mathbf{v}^{n+1} = \frac{2}{\tau_n}(\mathbf{u}^{n+1} - \mathbf{u}^n) - \mathbf{v}^n. \tag{3.78}$$

Startujeme přitom z počátečních hodnot

$$\mathbf{u}^0 = \boldsymbol{\varphi}, \quad \mathbf{v}^0 = \boldsymbol{\psi}. \tag{3.79}$$

Matice $\hat{\mathbf{K}}$ soustavy (3.77) nezávisí na čase a je pozitivně definitní. Soustavu (3.77) je proto účelné řešit pomocí Choleského rozkladu, který stačí provést jen jednou.

Metoda (3.77)–(3.79) je speciálním případem *Newmarkovy metody* hojně používané v inženýrské praxi, viz [2].

3.4. Hyperbolická rovnice prvního řádu

Formulace úlohy. Hledáme funkci $u(x,t)$ definovanou pro $x \in \langle 0, \ell\rangle$, $t \in \langle 0, T\rangle$, která vyhovuje diferenciální rovnici

$$\frac{\partial u}{\partial t} + a(x,t)\frac{\partial u}{\partial x} = f(x,t)\,, \quad x \in (0,\ell)\,,\ t \in (0,T)\,, \tag{3.80}$$

okrajovým podmínkám

$$\begin{aligned} u(0,t) &= g_0(t)\,, \\ u(\ell,t) &= g_\ell(t)\,, \end{aligned} \quad \text{pokud} \quad \begin{aligned} a(0,t) &> 0\,, \\ a(\ell,t) &< 0\,, \end{aligned} \qquad t \in (0,T)\,, \tag{3.81}$$

a počáteční podmínce

$$u(x,0) = \varphi(x)\,. \tag{3.82}$$

Nechť $Q = \langle 0, \ell \rangle \times \langle 0, T \rangle$ je obdélník, v němž hledáme řešení, a ∂Q^+ ta je část hranice ∂Q obdélníka Q, na níž je předepsána počáteční nebo okrajová podmínka, tj.

$$\partial Q^+ = \{[x,t] \in \partial Q \,|\, t=0 \text{ nebo } x=0 \text{ (pokud } a(0,t)>0) \text{ nebo } x=\ell \text{ (pokud } a(\ell,t)<0)\}\,.$$

Pro každý bod $P_0 = [x_0, t_0] \in Q \setminus \partial Q^+$ určeme řešení $x(t)$ počátečního problému

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = a(x(t),t)\,, \qquad t < t_0\,, \quad x(t_0) = x_0. \tag{3.83}$$

Křivka $x(t)$ se nazývá *charakteristika* příslušná bodu $[x_0, t_0]$. Předpokládejme, že charakteristika protíná ∂Q^+ v jediném bodě $P_0^* = [x_0^*, t_0^*]$. Tomuto bodu říkáme *pata charakteristiky*. Podle pravidla o derivování složené funkce

$$\frac{\mathrm{d}u(x(t),t)}{\mathrm{d}t} = \frac{\partial u(x(t),t)}{\partial x}\frac{\mathrm{d}x(t)}{\mathrm{d}t} + \frac{\partial u(x(t),t)}{\partial t} = \frac{\partial u(x(t),t)}{\partial t} + a(x(t),t)\frac{\partial u(x(t),t)}{\partial x}\,.$$

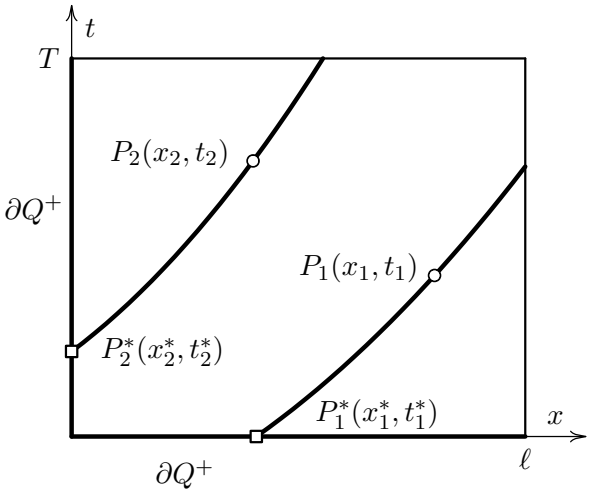
Úlohu (3.80)–(3.82) proto můžeme na charakteristice $x(t)$ zapsat ve tvaru

$$\frac{\mathrm{d}u(x(t),t)}{\mathrm{d}t} = f(x(t),t), \;\; t \in (t_0^*, t_0)\,, \quad u(x(t_0^*), t_0^*) = \begin{cases} \varphi(x_0^*) & \text{pro} \;\; t_0^* = 0\,, \\ g_0(t_0^*) & \text{pro} \;\; x_0^* = 0\,, \\ g_\ell(t_0^*) & \text{pro} \;\; x_0^* = \ell\,. \end{cases} \tag{3.84}$$

Předpokládejme, že funkce a, f, g_0, g_ℓ a φ jsou spojité, že funkce $a(0,t)$ i $a(\ell,t)$ nemění znaménko a že okrajová podmínka je kompatibilní s počáteční podmínkou, tj. platí

$$\varphi(0) = g_0(0) \text{ (pokud } a(0,0) > 0)\,, \quad \varphi(\ell) = g_\ell(0) \text{ (pokud } a(\ell,0) < 0)\,.$$

Pak úloha (3.83)–(3.84), a tedy rovněž úloha (3.80)–(3.82), má jediné klasické řešení.



Obr. 3.10. Charakteristiky

Rovnice (3.80) bývá označována jako *advekční* rovnice nebo také *transportní* rovnice. Úloha (3.80)–(3.82) popisuje třeba šíření příměsi prouděním, tj. bez vlivu difúze: $u(x,t)$ je koncentrace příměsi v tekutině v bodu x a v čase t, $a = \varrho v$ je součin hustoty ϱ tekutiny a její rychlosti v, f je zdrojový člen. Charakteristika $x(t)$ je trajektorie, po které se částečka příměsi pohybuje. Pokud bychom připustili také difúzní šíření příměsi, dostali bychom *konvekčně-difúzní rovnici*

$$c(x)\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(r(x,t)u) - \frac{\partial}{\partial x}\left(p(x)\frac{\partial u}{\partial x}\right) = f(x,t), \quad x \in (0,\ell)\,,\; t \in (0,T),$$

v níž p je koeficient difúze, viz (3.53'). Na rovnici advekce lze proto nahlížet jako na limitní případ konvekčně-difúzní rovnice, v níž zanedbáme účinky difúze. Jiná forma advekční rovnice tedy je

$$c(x)\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(r(x,t)u) = f(x,t)\,, \quad x \in (0,\ell)\,,\ t \in (0,T). \tag{3.80'}$$

Jestliže rovnice (3.80') popisuje přenos tepla advekcí, pak u je teplota, c je objemová tepelná kapacita (tj. tepelná kapacita vztažená na jednotku objemu), $r = cv$, kde v je rychlost proudění a f je tepelný zdroj.

Metoda přímek. Uvažujme rovnoměrné dělení intervalu $\langle 0,\ell\rangle$, tj. $x_i = ih$, $i = 0,1,\ldots,N$, kde $h = \ell/N$. Předpokládejme, že funkce $a(x,t)$ v krajních bodech intervalu $\langle 0,\ell\rangle$ nemění znaménko. Splnění rovnice (3.80) budeme požadovat ve vnitřních uzlech $x_1, x_2, \ldots, x_{N-1}$ a pro $a(0,t) < 0$ také v uzlu x_0 a pro $a(\ell,t) > 0$ také v uzlu x_N. V rovnici

$$\frac{\partial u(x_i,t)}{\partial t} + a(x_i,t)\frac{\partial u(x_i,t)}{\partial x} = f(x_i,t)$$

aproximujeme člen $u_x(x_i,t)$ ve vnitřních uzlech centrální diferencí a v krajních uzlech jednostrannou diferencí. Pokud třeba $a(0,t) > 0$, $a(\ell,t) > 0$ pro všechna $t \in \langle 0,T\rangle$, dostaneme

$$\begin{aligned}
&\dot u_1(t) + a_1(t)[u_2(t) - g_0(t)]/(2h) = f_1(t)\,,\\
&\dot u_i(t) + a_i(t)[u_{i+1}(t) - u_{i-1}(t)]/(2h) = f_i(t)\,, \quad i = 2,3,\ldots,N-1\,,\\
&\dot u_N(t) + a_N(t)[3u_N(t) - 4u_{N-1}(t) + u_{N-2}(t)]/(2h) = f_N(t)\,,
\end{aligned} \tag{3.85}$$

kde $a_i(t) = a(x_i,t)$, $f_i(t) = f(x_i,t)$ a $u_i(t)$ je aproximace $u(x_i,t)$. Z (3.82) dostaneme počáteční podmínky

$$u_i(0) = \varphi_i\,, \quad i = 1,2,\ldots,N\,, \tag{3.86}$$

kde $\varphi_i = \varphi(x_i)$. Počáteční problém (3.85)–(3.86) řešíme vhodnou numerickou metodou. K jejímu výběru nám poslouží zjednodušený model. Předpokládejme, že $a = 1$, $f = 0$ a uvažme periodické okrajové podmínky $u(0,t) = u(\ell,t)$. Pomocí centrálních diferencí odvodíme

$$\begin{aligned}
&\dot u_1(t) + [u_2(t) - u_N(t)]/(2h) = 0\,,\\
&\dot u_i(t) + [u_{i+1}(t) - u_{i-1}(t)]/(2h) = 0\,, \quad i = 2,3,\ldots,N-1\,,\\
&\dot u_N(t) + [u_1(t) - u_{N-1}(t)]/(2h) = 0\,.
\end{aligned} \tag{3.85'}$$

Počáteční problém (3.85')–(3.86) zapíšeme maticově, tj.

$$\dot{\mathbf{u}} = \mathbf{A}\mathbf{u}\,, \quad t \in (0,T)\,, \qquad \mathbf{u}(0) = \boldsymbol{\varphi}\,, \tag{3.87}$$

kde $\mathbf{u}(t) = (u_1(t), u_2(t), \ldots, u_N(t))^T$, $\boldsymbol{\varphi}(t) = (\varphi_1(t), \varphi_2(t), \ldots, \varphi_N(t))^T$ a

$$\mathbf{A} = -\frac{1}{2h}\begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & -1 \\ -1 & 0 & 1 & \ldots & 0 & 0 \\ 0 & -1 & 0 & \ldots & 0 & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & & -1 & 0 \end{pmatrix}. \tag{3.88}$$

Dá se ukázat, že vlastní čísla matice $\mathbf{A}$

$$\lambda_i = I\frac{N}{\ell}\sin\frac{2\pi i}{N}, \quad I = \sqrt{-1}, \qquad i = 1, 2, \ldots, N,$$

leží na imaginární ose, $\max_i |\lambda_i| \to \infty$ pro $N \to \infty$, tj. pro velké N je počáteční problém (3.87) oscilatoricky tuhý.

Antisymetrická matice $\mathbf{A}$ je diagonizovatelná pomocí unitární matice[1] vlastních vektorů, tj. platí $\mathbf{V}^H\mathbf{A}\mathbf{V} = \mathbf{D}$, kde $\mathbf{D} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$, viz [27]. Řešení počáteční úlohy (3.87) lze zapsat ve tvaru $\mathbf{u}(t) = \mathbf{V}\mathbf{S}(t)\mathbf{V}^H\boldsymbol{\varphi}$, kde $\mathbf{S}(t) = \mathrm{diag}\{e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_N t}\}$. Ověřte! Odtud již snadno obdržíme $\|\mathbf{u}(t)\|_2 = \|\boldsymbol{\varphi}\|_2$ pro každé $t > 0$. Ověřte! K numerickému řešení se proto skvěle hodí lichoběžníková metoda, pro kterou $\|\mathbf{u}_n\|_2 = \|\boldsymbol{\varphi}\|_2$ pro každé $n = 1, 2, \ldots$ Ověřte!

Problém (3.85)–(3.86) se chová podobně: vlastní čísla odpovídající matice $\mathbf{A}$ leží v záporné komplexní polorovině v blízkosti imaginární osy a $|\lambda_{max} - IN/\ell| \to 0$ pro $N \to \infty$. Pro řešení počátečního problému (3.85)–(3.86) lze kromě lichoběžníkové metody doporučit také metody, jejichž oblast absolutní stability obsahuje obdélník

$$R_{\alpha\beta} = \{z \in \mathbb{C} \,|\, -\alpha \le Re(z) \le 0, -\beta \le Im(z) \le \beta\}.$$

Pro metodu BS32 je $(\alpha; \beta) \doteq (1{,}64; 1{,}73)$ a pro metodu DP54 je $(\alpha; \beta) \doteq (3{,}19; 0{,}99)$, viz např. [13]. Délka kroku těchto dvou metod je sice z důvodu stability omezena, toto omezení však není nijak dramatické, délka kroku bude řádu $O(h)$. Z matlabovských programů lze tedy doporučit programy `ode23t`, `ode23` a `ode45`.

Metoda charakteristik je další vhodnou technikou pro řešení úlohy (3.80)–(3.82). Její podstatu vysvětlíme nejdříve pro zjednodušenou úlohu

$$\begin{aligned} &\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0\,, \quad x \in (0, \ell)\,,\ t \in (0, T)\,, \\ &u(0, t) = g_0(t), \quad t \in (0, T)\,, \\ &u(x, 0) = \varphi(x), \quad x \in (0, \ell)\,, \end{aligned} \tag{3.89}$$

[1] Čtvercová *matice* $\mathbf{V}$ *je unitární*, jestliže $\mathbf{V}^H\mathbf{V} = \mathbf{V}\mathbf{V}^H = \mathbf{I}$. Přitom $\mathbf{V}^H$ je *matice Hermitovsky sdružená*, tj. transponovaná a komplexně sdružená: když $\mathbf{V} = \{v_{ij}\}_{i,j=1}^N$ a $\mathbf{V}^H = \{v_{ij}^H\}_{i,j=1}^N$, pak $v_{ij}^H = \bar{v}_{ji}$ je číslo komplexně sdružené s číslem v_{ji}.

kde $a > 0$ je konstanta. Diferenciální rovnici přepíšeme pomocí charakteristik na tvar

$$\frac{\mathrm{d}u(x(t),t)}{\mathrm{d}t} = 0. \tag{3.90}$$

Zvolme rovnoměrné dělení intervalu $\langle 0, \ell\rangle$ s krokem $h = \ell/N$, tj. $x_i = ih$, $i = 0, 1, \ldots, N$, a rovnoměrné dělení intervalu $\langle 0, T\rangle$ s krokem $\tau = T/Q$, tj. $t_n = n\tau$, $n = 0, 1, \ldots, Q$. Charakteristika vycházející z bodu $[x_i, t_{n+1}]$ je přímka $x_i(t) = x_i + a(t - t_{n+1})$. V čase $t = t_n$ je

$$x_i^n := x_i(t_n) = x_i - a\tau\,.$$

Předpokládejme, že $a\tau \le h$. Pak pro $i = 1, 2, \ldots, N$ bod $x_i^n \in \langle x_{i-1}, x_i\rangle$, zejména tedy $x_i^n \in \langle 0, \ell\rangle$, takže $u(x_i^n, t_n)$ má smysl. Integrací rovnice (3.90) od t_n do t_{n+1} obdržíme

$$\int_{t_n}^{t_{n+1}} \frac{\mathrm{d}u(x_i(t),t)}{\mathrm{d}t}\,\mathrm{d}t = u(x_i, t_{n+1}) - u(x_i^n, t_n) = 0\,, \text{ a odtud } u(x_i, t_{n+1}) = u(x_i^n, t_n)\,.$$

Numerickou metodu dostaneme tak, že $u(x_i^n, t_n)$ určíme přibližně interpolací, tj. počítáme

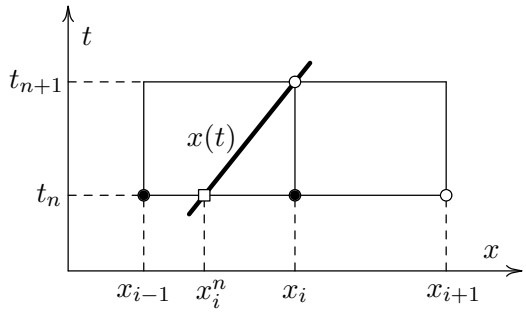
$$u_i^{n+1} = P_k(x_i^n)\,, \tag{3.91}$$

kde $P_k(x)$ je interpolační polynom stupně $k \ge 1$ splňující

$$P_k(x_{i-1}) = u_{i-1}^n\,, \quad P_k(x_i) = u_i^n\,. \tag{3.92}$$

Pro lineární polynom P_1 z podmínek (3.92) odvodíme

$$P_1(x) = u_i^n + \frac{u_i^n - u_{i-1}^n}{h}(x - x_i) \quad \text{a odtud} \quad P_1(x_i^n) = u_i^n - \frac{a\tau}{h}(u_i^n - u_{i-1}^n)\,.$$



Obr. 3.11. Upwind metoda

Dostali jsme tak *upwind* metodu

$$u_i^{n+1} = u_i^n - \frac{a\tau}{h}(u_i^n - u_{i-1}^n)\,. \tag{3.93}$$

Název metody nám připomíná, že veličina u v uzlu x_i závisí jen hodnotách této veličiny v uzlech ležících „proti proudu, proti větru“, pro $a > 0$ tedy nalevo od x_i. V bodu $x_0 = 0$ užijeme okrajovou podmínku, takže $u_0^{n+1} = g_0(t_{n+1})$.

Dá se ukázat, že když

$$\nu := \frac{a\tau}{h} \le 1\,, \tag{3.94}$$

pak za předpokladu dostatečné hladkosti přesného řešení pro chybu platí

$$u(x_i, t_n) - u_i^n = O(\tau)\,, \tag{3.95}$$

viz [15]. Říkáme, že metoda (3.93) je řádu jedna. Pro $\nu = 1$ dokonce $u_i^n = u(x_i, t_n)$ je přesné! Číslo $\nu = a\tau/h$ se nazývá *Courantovo číslo* a podmínka (3.94) se nazývá *Courantova-Friedrichsova-Lewyova podmínka*, stručně *CFL podmínka*.

Metodu řádu dva dostaneme tak, že v (3.91) použijeme interpolační polynom druhého stupně. Když k podmínkám (3.92) přidáme ještě podmínku

$$P_2(x_{i+1}) = u_{i+1}^n\,, \tag{3.96}$$

vypočteme $P_2(x_i^n)$ a dosadíme do (3.91), dostaneme *Laxovu-Wendroffovu* metodu

$$u_i^{n+1} = u_i^n - \frac{a\tau}{2h}(u_{i+1}^n - u_{i-1}^n) + \frac{a^2\tau^2}{2h^2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n)\,. \tag{3.97}$$

V bodu x_0 užijeme okrajovou podmínku, takže $u_0^{n+1} = g_0(t_{n+1})$. Pro aproximaci v uzlu x_N můžeme použít interpolační polynom $P_2(x)$, který kromě podmínek (3.92) vyžaduje navíc splnění podmínky $P_2(x_{N-2}) = u_{N-2}^n$.

Jestliže pro obecný uzel x_i přidáme k podmínkám (3.92) podmínku

$$P_2(x_{i-2}) = u_{i-2}^n\,, \tag{3.98}$$

vypočteme $P_2(x_i^n)$ a dosadíme do (3.91), obdržíme *Beamovu-Warmingovu* metodu

$$u_i^{n+1} = u_i^n - \frac{a\tau}{2h}(3u_i^n - 4u_{i-1}^n + u_{i-2}^n) + \frac{a^2\tau^2}{2h^2}(u_i^n - 2u_{i-1}^n + u_{i-2}^n)\,. \tag{3.99}$$

Jestliže $u_0^{n+1} = g_0(t_{n+1})$ a počítáme-li u_i^{n+1}, $i = 1, 2, \ldots, N-1$, podle (3.97) a u_N^{n+1} podle (3.99), je-li splněna CFL podmínka (3.94) a je-li přesné řešení u dostatečně hladké, pro chybu platí

$$u(x_i, t_n) - u_i^n = O(\tau^2)\,. \tag{3.100}$$

Jestliže $u_0^{n+1} = g_0(t_{n+1})$ a počítáme-li u_1^{n+1} podle (3.97) a u_i^{n+1}, $i = 2, 3, \ldots, N$, podle (3.99), je-li splněna CFL podmínka (3.94) a je-li přesné řešení u dostatečně hladké, pro chybu opět platí (3.100).

Věnujme se dále případu, kdy konstanta $a < 0$. Pak okrajová podmínka bude předepsána vpravo, tj. podmínku (3.89_2) nahradí podmínka

$$u(\ell, t) = g_\ell(t)\,, \qquad t \in (0, T)\,. \tag{$3.89_2'$}$$

Charakteristika vycházející z bodu $[x_i, t_{n+1}]$ bude směřovat doprava, takže za předpokladu platnosti CFL podmínky

$$\nu := \frac{|a|\tau}{h} \le 1 \tag{3.101}$$

bude $x_i(t_n) \in \langle x_i, x_{i+1}\rangle$ pro $i = 0, 1, \ldots, N-1$. Podobně jako dříve odvodíme upwind metodu

$$u_i^{n+1} = u_i^n - \frac{|a|\tau}{h}(u_i^n - u_{i+1}^n). \tag{3.93'}$$

Laxova-Wendroffova metoda na znaménku a nezávisí, takže opět platí (3.97). Pro Beamovu-Warmingovu metodu dostaneme

$$u_i^{n+1} = u_i^n - \frac{|a|\tau}{2h}(3u_i^n - 4u_{i+1}^n + u_{i+2}^n) + \frac{a^2\tau^2}{2h^2}(u_i^n - 2u_{i+1}^n + u_{i+2}^n). \tag{3.99'}$$

V obecném případě $a = a(x,t)$ určíme x_i^n numericky: $x_i^n \approx x_i(t_n)$, kde

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = a(x_i(t),t)\,, \qquad x_i(t_{n+1}) = x_i\,. \tag{3.101}$$

Pro upwind metodu použijeme EE metodu, takže

$$x_i^n = x_i - a_i^n\tau\,, \quad \text{kde} \quad a_i^n = a(x_i, t_{n+1})\,.$$

Pro Laxovu-Wendroffovu metodu a Beamovu-Warmingovu použijeme EM2 metodu, tj.

$$x_i^n = x_i - a_i^n\tau, \quad \text{kde} \quad a_i^n = \tfrac{1}{2}(k_1+k_2), \quad k_1 = a(x_i,t_{n+1})\,,\ k_2 = a(x_i - \tau k_1, t_n)\,.$$

Vzorce (3.93), (3.93'), (3.97), (3.99) a (3.99') se změní jen v tom, že v nich místo a píšeme a_i^n. Délka kroku musí splňovat CFL podmínku

$$\frac{\max_i |a_i^n|\tau}{h} \le 1. \tag{3.101'}$$

Jestliže v rovnici (3.89_1) uvažujeme nenulovou pravou stranu $f(x,t)$, tj. řešíme-li místo rovnice (3.90) rovnici

$$\frac{\mathrm{d}u(x(t),t)}{\mathrm{d}t} = f(x(t),t), \tag{3.90'}$$

přičteme k pravým stranám formulí aproximaci $Q_i^n(f)$ integrálu $\int_{t_n}^{t_{n+1}} f(x_i(t),t)\,\mathrm{d}t$. Pro upwind metodu stačí použít jednostrannou obdélníkovou formuli

$$Q_i^n(f) = \tau f(x_i, t_{n+1}).$$

Pro Laxovu-Wendroffovu metodu a Beamovu-Warmingovu metodu užijeme lichoběžníkovou formuli

$$Q_i^n(f) = \tfrac{1}{2}\tau[f(x_i,t_{n+1}) + f(x_i^n, t_n)].$$

Metoda konečných objemů. Pro diskretizaci rovnice (3.80') se výborně hodí metoda konečných objemů na časoprostorových objemech $B_i = \langle x_{i-1/2}, x_{i+1/2}\rangle \times \langle t_n, t_{n+1}\rangle$, více o tom viz [14].

Literatura

[1] R. Ashino, M. Nagase, R. Vaillancourt: *A survey of the MATLAB ODE suite,* Technical Report CRM-2651, Centre de recherches mathematiques, University of Ottawa, 2000.

[2] K. J. Bathe: *Finite Elements Procedures,* Prentice-Hall, Upper Saddle River, NJ, 1996.

[3] M. Brandner, J. Egermaier, H. Kopincová: *Numerické metody pro řešení evolučních parciálních diferenciálních rovnic,* učební text ZČU a VŠB, Plzeň, 2012.

[4] L. Čermák, R. Hlavička: *Numerické metody,* CERM, učební text FSI VUT Brno, 2017.

[5] J. Donea, A. Huerta: *Finite Element Methods for Flow Problems,* John Wiley & Sons Ltd, Chichester, 2003.

[6] D. R. Durran: *Numerical Methods for Fluid Dynamics: with Application to Geophysics (2nd edition),* Springer, New York, 2010.

[7] M. Feistauer, J. Felcman, I. Straškraba: *Mathematical and Computational Methods for Compressible Flow,* Oxford Science Publications, New York, 2003.

[8] J. H. Ferziger, M. Perić: *Computational Methods for Fluid Dynamics (3rd edition),* Springer, Berlin, 2002.

[9] J. Fish, T. Belytschko: *A First Course in Finite Elements,* John Willey & Sons Ltd, Chichester, 2007.

[10] M. T. Heath: *Scientific Computing. An Introductory Survey,* McGraw-Hill, New York, 2002.

[11] R. Hlavička: *Numerické metody pro řešení diferenciálních rovnic: Průvodce softwarem a počítačová cvičení v prostředí MATLABu,* učební text FSI VUT Brno, [on-line], dostupné z `http://mathonline.fme.vutbr.cz/Numericke-metody-II/sc-1246-sr-1-a-263/default.aspx`.

[12] P. Knabner, L. Angermann: *Numerical Methods for Elliptic and Parabolic Partial Differential Equations,* Springer, New York, 2003.

[13] J. D. Lambert: *Numerical Methods in Ordinary Differential Systems. The Initial Value Problem,* J. Willey & Sons, Chichester, 1993.

[14] R. J. LeVecque: *Finite Volume Methods for Hyperbolic Problems,* Cambridge University Press, Cambridge, 2002.

[15] R. J. LeVecque: *Finite Difference Methods for Ordinary and Partial Differential Equations,* Siam, Philadelphia, 2007.

[16] The MathWorks, Inc.: *MATLAB® Mathematics*, R2012b, [on-line], dostupné z `http://www.mathworks.com/help/pdf_doc/matlab/math.pdf`.

[17] S. Míka, P. Přikryl, M. Brandner: *Speciální numerické metody: Numerické metody řešení okrajových úloh pro diferenciální rovnice,* Vydavatelský servis, Plzeň, 2006.

[18] C. B. Moler: *Numerical Computing with MATLAB,* Siam, Philadelphia, 2004. Electronic edition: The MathWorks, Inc., Natick, MA, 2004, [on-line], dostupné z `http://www.mathworks.com/moler`.

[19] S. Míka, P. Přikryl: *Numerické metody řešení obyčejných diferenciálních rovnic, okrajové úlohy,* učební text ZČU Plzeň, 1994.

[20] A. Quarteroni, R. Sacco, F. Saleri: *Numerical Mathematics,* Springer, Berlin, 2000.

[21] K. Rektorys: *Přehled užité matematiky I,II,* Prometheus, Praha, 1995.

[22] L. F. Shampine: *Some practical Runge-Kutta formulas,* Math. Comp., Vol. 46, Num. 173 (1986), str. 135-150.

[23] L. F. Shampine: *Numerical Solution of ordinary differential equations,* Chapman & Hall, New York, 1994.

[24] L. F. Shampine, I. Gladwell, S. Thompson: *Solving ODEs with MATLAB,* Cambridge University Press, Cambridge, 2003.

[25] L. F. Shampine, L. W. Reichelt: *The MATLAB ODE suite,* SIAM J. Sci. Comput., Vol. 18 (1997), No. 1, str. 1-22.

[26] H. K. Versteeg, W. Malalasekera: *An Introduction to Computational Fluid Dynamics: The finite volume method (2nd edition),* Prentice Hall, Harlow, 2007.

[27] J. Stoer, R. Bulirsch: *Introduction to Numerical Analysis,* 2. vyd., Springer, New York, 1993.

[28] E. Vitásek: *Numerické metody,* SNTL, Praha, 1987.

[29] E. Vitásek: *Základy teorie numerických metod pro řešení diferenciálních rovnic,* Academia, Praha, 1994.

[30] O. C. Zienkiewicz, R. L. Taylor: *The Finite Element Method, Volume I: The Basis,* Butterworth-Heinemann, Oxford, 2000.