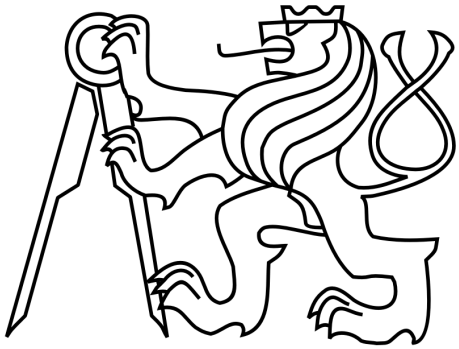
České vysoké učení technické v Praze
Fakulta stavební
Katedra mechaniky



Rozšíření předmětů
‘Numerická analýza konstrukcí 1 a 2’

Tento výukový text byl vytvořen na základě podpory grantu
FRVŠ 1249/2013

Ing. Filip Kolařík
Ing. Martin Horák
Prof. Dr. Ing. Bořek Patzák
Ing. Jan Stránský
Ing. Petr Havlásek

Contents

Úvod

Jednou z nejdůležitějších činností inženýrů je modelování a řešení reálných, přirozených jevů. Ty mohou pocházet z nejrůznějších oblastí od technických věd, přes moderní fyziku a biologii až po společenské vědy a ekonomii. Pomocí fyzikálních zákonů (případně jiných znalostí) musí inženýr sestavit odpovídající matematický model, který zkoumaný problém vystihuje. Matematický model lze chápat jako množinu vztahů mezi proměnnými, vystupujícími ve zkoumaném problému, která ho analyticky popisuje. Modely fyzikálních problémů, na které se zde omezíme, jsou obvykle založeny na fundamentálních principech, jako jsou zákony zachování hmoty, hybnosti a energie. Ty jsou potom popsány pomocí algebraických, diferenciálních, nebo integrálních rovnic.

S řadou těchto problémů se studenti již setkali v nejrůznějších předmětech nejen na katedře mechaniky. Jako příklady lze uvést diferenciální rovnice pro tažený-tlačený prut či ohybovou čáru nosníku odvozenou v předmětu Pružnost a pevnost

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(EA\frac{\mathrm{d}u}{\mathrm{d}x}\right)+f_x=0, \tag{0.1}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(EI\frac{\mathrm{d}^2w}{\mathrm{d}x^2}\right)+f_z=0, \tag{0.2}$$

nebo diferenciální rovnice popisující stacionární vedení tepla, která byla řešená v Numerické analýze konstrukcí 1

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\lambda\frac{\mathrm{d}T}{\mathrm{d}x}\right)+Q=0, \tag{0.3}$$

Jedním z cílů tohoto textu je ukázat, že uvedené problémy je možné roztřídit do skupin, ve kterých mají příslušné rovnice společné jisté vlastnosti a znaky. Z matematického hlediska jde pak v rámci dané skupiny často o jedinou rovnici, kde různou fyzikální interpretací jejích koeficientů získáme modely pro různé fyzikální jevy. Například rovnice pro prut namáhaný tahem (1.1), stacionární (nezávislá na čase) rovnice vedení tepla (0.3), patří mezi takzvaně eliptické problémy. Nestacionární (časově závislý) problém vedení tepla

$$\rho c_v\frac{\partial T}{\partial t}-\frac{\partial}{\partial x}\left(\lambda\frac{\partial T}{\partial x}\right)+Q=0, \tag{0.4}$$

je úlohou parabolickou a úlohy probírané v předmětu Dynamika stavebních konstrukcí, například podélné kmitání prutu, nebo harmonický oscilátor

$$\rho\frac{\partial^2u}{\partial t^2}-\frac{\partial}{\partial x}\left(EA\frac{\partial u}{\partial x}\right)-f_x=0, \tag{0.5}$$

patří mezi hyperbolické problémy. Formální vymezení těchto pojmů je uvedeno v Dodatku 2.8. Uvedené rozdělení diferenciálních rovnic umožňuje vyzdvihnout a shrnout jejich společné znaky, což hraje zásadní roli při vyšetřování vlastností těchto rovnic, zjišťování existence a jednoznačnosti řešení až po volbu metody řešení a s tím spojenou analýzu zvolené metody.

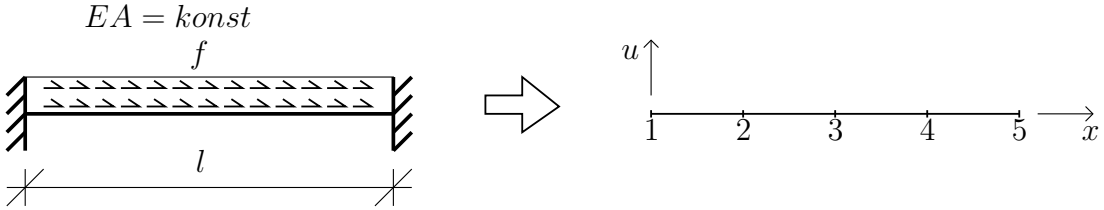
Prakticky významné problémy jsou však často dvourozměrné, nebo i třírozměrné, přičemž geometrie úlohy může být značně komplexní. Nalezení analytického řešení proto obecně není možné. Dostupné analytické metody (jako například Fourierova metoda separace proměnných) mají navíc většinou pouze omezenou použitelnost. I v případech, kde je obecné řešení úlohy známé, nemusí se podařit nalézt příslušné partikulární řešení, neboť to závisí na okrajových podmínkách úlohy. Na rozdíl od obyčejných diferenciálních rovnic (ODR), kde partikulární řešení závisí jen na libovolných konstantách, v případě parciálních diferenciálních rovnic (PDR) partikulární řešení závisí na libovolných funkcích. Problém určení těchto funkcí, které závisí na okrajových podmínkách, je pak často analyticky neřešitelný. Z teoretického hlediska se proto analýza problému často omezuje na vyšetřování existence a jednoznačnosti řešení, případně jeho regularitu, nikoliv však na zkonstruování tohoto řešení. Z uvedeného plyne nutnost mít k dispozici numerické metody, které hledané řešení umožňují získat alespoň přibližně. Jak však uvidíme dále, při správném použití lze docílit teoreticky libovolné přesnosti, a navíc je možno řešit i velmi složité úlohy. Nejvýznamnější a nejpoužívanější numerické metody řešení diferenciálních rovnic jsou bezesporu metoda konečných prvků a metoda konečných diferencí, někdy také zvaná metoda sítí. S nimi se studenti již setkali v některých předchozích předmětech. S metodou sítí v Analýze konstrukcí a s metodou konečných prvků v Numerické analýze konstrukcí 1. Cílem tohoto textu je rozšířit a zejména prohloubit tyto znalosti, ukázat matematické základy jednotlivých metod, předvést analýzu konkrétních numerických schémat, ukázat jejich výhody a omezení a rozsah použitelnosti.

1 Metoda konečných diferencí

1.1 Úvod

Metoda konečných diferencí (MKD), někdy zvaná také metoda sítí, je přibližnou metodou pro řešení diferenciálních rovnic (ať už obyčejných nebo parciálních). Idea v podstatě všech numerických metod pro řešení diferenciálních rovnic je stejná. Zhruba řečeno, daný spojitý problém, ve kterém na zkoumané oblasti hledáme takovou funkci, která v každém bodě oblasti vyhovuje předepsaným rovnicím, chceme převést na diskrétní problém, kdy hledaná funkce bude přiblížením řešení původního problému a bude vyhovovat diskretizovaným rovnicím pouze v konečně mnoha bodech řešené oblasti, případně bude tyto rovnice splňovat pouze ve vhodném zobecněném smyslu. Touto oblastí může být například nosník, na němž hledáme rozložení vnitřních sil, nebo stěna, ve které nás zajímá rozložení teploty. Od spojitého a tedy nekonečně dimenzionálního problému tak přecházíme k diskrétnímu, konečně dimenzionálnímu problému. Původní diferenciální rovnice (nebo soustava diferenciálních rovnic) se převede na soustavu algebraických rovnic. Pokud původní rovnice byla lineární, pak i soustava algebraických rovnic bude lineární. V případě nelineárních diferenciálních rovnic bude vzniklá diskretizovaná soustava algebraických rovnic také nelineární a i tu bude posléze třeba řešit některou z přibližných metod. Stručnou informaci o numerických metodách obecně podává Dodatek 2.9.

V metodě konečných diferencí se řešená oblast nahradí množinou diskrétních bodů, které tvoří síť (odtud název metody). Tyto diskrétní body se nazývají uzly sítě. V uzlech sítě se potom diferenciální operátory (nebo jednoduše derivace) vystupující v diferenciální rovnici nahradí operátory diferenčními (podíly diferencí). Vše si pro snadnější pochopení nejprve ukážeme na jednoduchém příkladu. Mějme oboustranně vetknutý sloup délky l, zatížený rovnoměrným



Obr. 1.1: *Příklad 1 - Zadání*

spojitým zatížením dle Obr. 1.3. Tahová tuhost prutu EA nechť je po délce prutu konstantní. Známý princip říká, že aby byla celá konstrukce v rovnováze, musí být v rovnováze každá její část. Převedeno do řeči matematiky musí v každém bodě intervalu $(0, l)$ splněna rovnice rovnováhy. Navíc v krajních bodech intervalu musí být splněny okrajové podmínky, simulující způsob podepření/zatížení nosníku na jeho koncích. Z teorie pružnosti je známo, že příslušná rovnice rovnováhy zní

$$EAu_{,xx} + f_x = 0, \tag{1.1}$$

kde f_x je podélné rovnoměrné spojité zatížení prutu a čárkou jsme označili derivaci podle proměnné, která je uvedena spodním indexem za touto čárkou, tj. $u_{,xx}$ značí druhou derivaci u podle x. Toto označení budeme užívat v případech, kdy zpřehlední zápis. K této rovnici přísluší

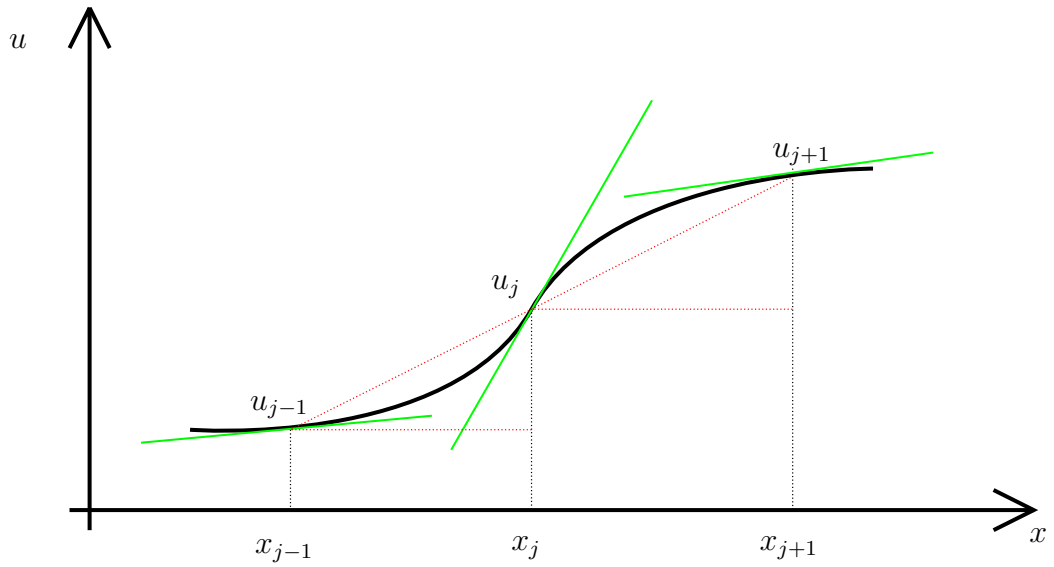
okrajové podmínky ve tvaru

$$u(0) = 0, \tag{1.2}$$
$$u(l) = 0. \tag{1.3}$$

Pro úplnost připomeňme, že zcela analogická rovnice popisuje i časově ustálený problém vedení tepla v jedné dimenzi (pouze normálová tuhost EA je nahrazena tepelnou vodivostí λ). Tuto úlohu je pochopitelně možné vyřešit analyticky [1]. Znalost analytického řešení nám poslouží k porovnání a ověření řešení získaného numericky metodou sítí. Řešení této úlohy lze zapsat ve tvaru

$$u(x) = -\frac{f_x}{2EA}x^2 + \frac{f_x l}{2EA}x. \tag{1.4}$$

Ověření, že (1.4) je řešením (1.1) je snadné na základě dosazení. Nyní úlohu vyřešíme metodou



Obr. 1.2: *Příklad 1. - náhrada derivace*

sítí. K tomu účelu je nejprve třeba celou oblast (v našem případě interval $(0, l)$) diskretizovat, to znamená vytvořit síť uzlů. Jak hustá síť bude, záleží čistě na nás. Zvolme nejprve pro jednoduchost síť dle obrázku. Dělení intervalu je pro jednoduchost ekvidistantní. Obecně to však není nutné. Vzdálenost dvou sousedních uzlů x_{j+1} a x_j označíme h, tzv. prostorový krok sítě. Místo požadování platnosti rovnice (1.1) v každém bodě $(0, l)$, budeme tuto rovnici řešit pouze v uzlech vytvořené sítě. Navíc v těchto uzlech nahradíme derivaci pomocí podílu diferencí. Náhrada derivace je založena na geometrické představě pojmu derivace jakožto směrnice tečny ke grafu funkce v daném bodě. Při použití označení z Obr. 1.2 lze pro derivaci v bodě x_j psát přibližný vztah:

$$\frac{\mathrm{d}u(x_j)}{\mathrm{d}x} \approx \frac{u(x_{j+1}) - u(x_j)}{h} = \frac{\Delta_{+x}u(x_j)}{h} \tag{1.5}$$

Přitom výrazu

$$\Delta_{+x}u(x) = u(x + h) - u(x) \tag{1.6}$$

[1] Ohledně řešení lineárních ODR odkazujeme čtenáře na přednášky z Matematiky 2.

budeme říkat dopředná diference. Podobně můžeme nahradit derivaci pomocí zpětné diference vztahem

$$\frac{\mathrm{d}u(x_j)}{\mathrm{d}x} \approx \frac{u(x_j) - u(x_{j-1})}{h} = \frac{\Delta_{-x}u(x_i)}{h} \tag{1.7}$$

kde výraz

$$\Delta_{-x}u(x) = u(x) - u(x-h) \tag{1.8}$$

nazýváme zpětnou diferencí. Je výhodné definovat také centrální diferenci:

$$\delta_x u(x) = u(x + h/2) - u(x - h/2) \tag{1.9}$$

Znovu a detailněji se náhradami derivací pomocí podílů diferencí a přesností těchto náhrad budeme zabývat v dalších oddílech. Uvedené diference jsou diferencemi prvního řádu, neboť nahrazují první derivace. V rovnici (1.1) však vystupuje druhá derivace, kterou je nutné vhodně aproximovat. Vyjdeme z faktu, že druhá derivace funkce je derivací její první derivace, a definujeme analogicky diferenci druhého řádu vztahem

$$\begin{gathered}\delta_x^2 u(x) = \delta_x(\delta_x u(x)) = \delta_x(u(x + h/2) - u(x - h/2)) = \\ \delta_x u(x + h/2) - \delta_x u(x - h/2) = u(x + h) - 2u(x) + u(x - h).\end{gathered} \tag{1.10}$$

Druhou derivaci funkce v bodě x_j je pak možno nahradit pomocí následujícího vztahu

$$\frac{\mathrm{d}^2u(x_j)}{\mathrm{d}x^2} \approx \frac{\frac{\delta_x u(x_j+h/2)}{h} - \frac{\delta_x u(x_j-h/2)}{h}}{h} = \frac{u(x_j + h) - 2u(x_j) + u(x_j - h)}{h^2} = \frac{\delta_x^2 u(x_j)}{h^2}. \tag{1.11}$$

Pokud nyní zavedeme následující označení pro aproximaci hledané funkce u v bodě x_j

$$u(x_j) \approx U_j, \tag{1.12}$$

je možné původní formulaci problému (1.1) převést na diskretizovaný tvar

$$EA\frac{\delta_x^2 U_j}{h^2} + f_j = EA\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2} + f_j = 0, \tag{1.13}$$

který je platný ve všech vnitřních uzlech sítě (v našem případě dělení intervalu $(0, l)$ jde o uzly $j = 2, 3, 4$). V rovnici (1.13) jsme označili f_j hodnotu spojitého zatížení v bodě x_j, tj. $f(x_j)$ (protože je $f(x)$ konstantní funkce, je $f_j = f$ ve všech uzlech. Pokud (1.13) rozepíšeme pro jednotlivé vnitřní uzly, dostaneme následující sadu rovnic

$$\begin{aligned} j = 2: &\quad \frac{EA}{h^2}(U_3 - 2U_2 + U_1) + f_2 = 0 \\ j = 3: &\quad \frac{EA}{h^2}(U_4 - 2U_3 + U_2) + f_3 = 0 \\ j = 4: &\quad \frac{EA}{h^2}(U_5 - 2U_4 + U_3) + f_4 = 0. \end{aligned} \tag{1.14}$$

K nim náleží ještě okrajové podmínky zohledňující uložení prutu, které v diskretizované podobě nabudou tvaru

$$\begin{aligned} j = 1: &\quad U_1 = 0 \\ j = 5: &\quad U_5 = 0. \end{aligned} \tag{1.15}$$

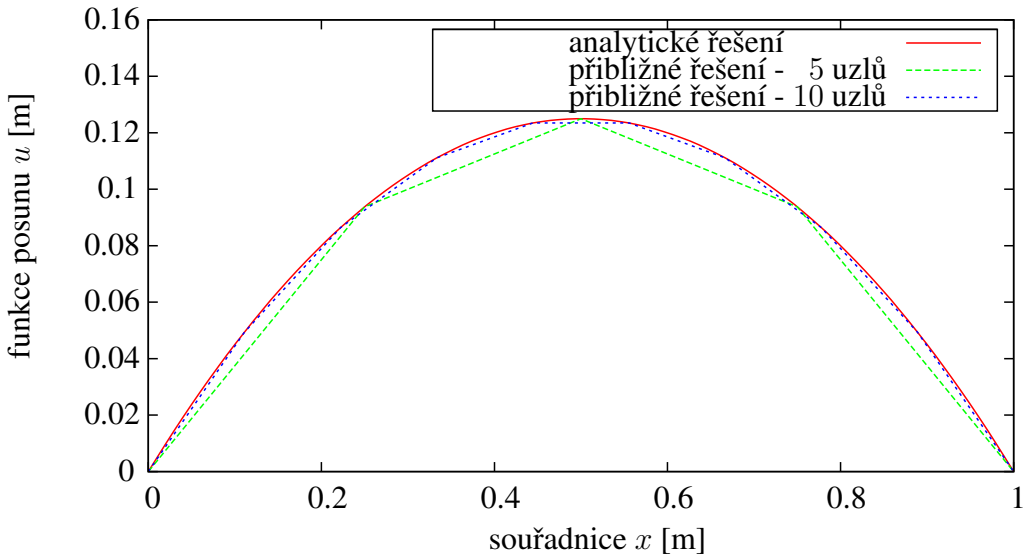
Spojením (1.14) a (1.15) dostaneme výslednou soustavu rovnic, kterou lze v maticové podobě zapsat jako

$$\frac{EA}{h^2}\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}\begin{pmatrix} U_2 \\ U_3 \\ U_4 \end{pmatrix} = \begin{pmatrix} f_2 \\ f_3 \\ f_4 \end{pmatrix}. \tag{1.16}$$

Řešením soustavy rovnic (2.123) získáme vektor uzlových hodnot obsahující aproximaci hledané funkce posunutí u ve vnitřních uzlech sítě. Pro dané dělení intervalu vychází

$$\begin{pmatrix} U_2 \\ U_3 \\ U_4 \end{pmatrix} = \frac{fh^2}{EA}\begin{pmatrix} 3/2 \\ 2 \\ 3/2 \end{pmatrix}. \tag{1.17}$$

Na Obr. **??** je uvedeno porovnání analytického řešení a přibližných řešení získaných pomocí



Obr. 1.3: *Příklad 1 - Zadání*

MKD při dělení na 5 a 10 uzlů. Z obrázku je patrné, že pro hustější dělení intervalu získáváme přesnější řešení. Přirozeně přitom vznikají otázky, jaké se při daném dělení dopouštíme chyby, jak rychle se chyba zmenšuje, pokud zjemňujeme dělení sítě, případně zda je možné chybu teoreticky zmenšit libovolně (a tedy se dostat libovolně blízko k přesnému řešení). Abychom na tyto otázky mohli odpovědět, zavedeme nejprve následující pojem chyby diskretizace, kterou označíme $\varepsilon_h(x_j)$, definované jako reziduum[2] diskretizovaného problému (1.13), v němž nahradíme přibližné řešení U_j přesným řešením $u(x_j)$, tj.

$$\varepsilon_h(x_j) = EA\frac{\delta_x^2 u(x_j)}{h^2} + f_j = EA\frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} + f_j.$$

[2]Reziduum rovnice je chyba vzniklá tím, že jsme do ní nedosadili funkci, která ji identicky splňuje. Mějme například soustavu algebraických rovnic $Ax = b$, kde matice A je regulární. Pokud je vektor y jejím řešením, pak $Ay - b = 0$. Pro libovolný vektor $z \neq y$ je potom $Ay - b = r \neq 0$. Vektor r se nazývá reziduum této rovnice.

V libovolném vnitřním bodě intervalu $(0, l)$ tedy pro libovolné dělení platí

$$\varepsilon_h(x) = EA\frac{\delta_x^2 u(x)}{h^2} + f(x) = EA\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + f(x). \tag{1.18}$$

Použitím nekonečného Taylorova rozvoje následovně upravíme jednotlivé členy (1.18)

$$u(x+h) = u(x) + u_{,x}(x)h + \frac{1}{2}u_{,xx}(x)h^2 + \frac{1}{6}u_{,xxx}(x)h^3 + \frac{1}{24}u_{,xxxx}(x)h^4 + ... \tag{1.19}$$

$$u(x-h) = u(x) - u_{,x}(x)h + \frac{1}{2}u_{,xx}(x)h^2 - \frac{1}{6}u_{,xxx}(x)h^3 + \frac{1}{24}u_{,xxxx}(x)h^4 + ... \tag{1.20}$$

Dosazením těchto rozvojů do (1.18) dostáváme

$$\varepsilon_h(x) = EA\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + f(x) = EA\frac{u_{,xx}(x)h^2 + \frac{1}{12}u_{,xxxx}(x)h^4 + ...}{h^2} + f(x) =$$
$$EA\left(u_{,xx}(x) + \frac{1}{12}u_{,xxxx}(x)h^2 + ...\right) + f(x) = \frac{EA}{12}u_{,xxxx}h^2 + ... \tag{1.21}$$

Poslední rovnost (1.21) plyne z faktu, že $EAu_{xx}(x) + f(x) = 0$ (což není nic jiného, než původně řešená rovnice). Třemi tečkami jsou přitom označeny členy vyššího řádu. Pro další analýzu je však výhodnější uvažovat pouze konečné Taylorova rozvoje ve tvaru

$$u(x+h) = u(x) + u_{,x}(x)h + \frac{1}{2}u_{,xx}(x)h^2 + \frac{1}{6}u_{,xxx}(x)h^3 + \frac{1}{24}u_{,xxxx}(\xi)h^2, \quad \xi \in (x-h, h+h). \tag{1.22}$$

Analogicky rozvineme i $u(x-h)$. Za předpokladu dostatečné hladkosti vstupních dat (tzn. funkce $f(x)$) existuje konstanta M taková, že $|u_{,xxxx}(x)| \leq M$ všude na $(0, l)$. Potom dosazením (1.22) do (1.21) platí následující nerovnost

$$|\varepsilon_h(x)| \leq EA\frac{M}{12}h^2 = \mathcal{O}(h^2). \tag{1.23}$$

Nerovnost (1.23) říká, že se chyba diskretizace $\varepsilon_h(x)$ pro $h \to 0$ chová jako $\mathcal{O}(h^2)$ a je tedy druhého řádu přesnosti.[3] Jinými slovy, při zmenšení prostorového kroku h na polovinu se chyba způsobená nahrazením derivace podílem diferencí zmenší čtyřikrát. Vidíme, že pro $h \to 0$ jde velikost chyby diskretizace k nule. Vyvstává nyní důležitá otázka, zda metoda konverguje k přesnému řešení. Abychom mohli tuto otázku zodpovědět, definujme tzv. chybu aproximace jako rozdíl přibližné a přesné hodnoty hledané funkce u v daném uzlu sítě, tj.

$$e_j = U_j - u(x_j). \tag{1.24}$$

Zajímá nás, zda pro $h \to 0$ jde také $e_j \to 0$, jinými slovy zda se přibližná hodnota hledané funkce u blíží při zjemňování sítě k přesné hodnotě. Dosaďme nejprve z (1.24) za U_j do

[3] Symbol $\mathcal{O}$ značí následující skutečnost. Pro libovolné $x_0 \in [-\infty, \infty]$ řekneme, že $f(x) = \mathcal{O}(g(x))$ pro $x \to x_0$, pokud existuje konstanta $K > 0$ tak, že $|f(x)| \leq K|g(x)|$ na okolí bodu x_0. Lze ukázat, že ekvivalentně platí $\lim_{x\to x_0} \frac{f(x)}{g(x)} = K < \infty$ a tedy funkce $f(x)$ a $g(x)$ jsou na okolí bodu x_0 "stejného řádu". V našem případě to znamená, že funkce $\varepsilon_h(x)$ a h^2 se blízko nule chovají zhruba stejně.

diskretizované rovnice (1.13). Tím získáme

$$\underbrace{EA\frac{u(x_{j+1})-2u(x_j)+u(x_{j-1})}{h^2}+f_j}_{\varepsilon_h(x_j)}+EA\frac{e_{j+1}-2e_j+e_{j-1}}{h^2}=0, \tag{1.25}$$

neboli

$$-EA\frac{\delta_x^2 e(x_j)}{h^2}=\varepsilon_h(x_j). \tag{1.26}$$

Pro další postup budeme předpokládat, že diskrétní operátor $-EA\delta_x^2/h^2$ splňuje tzv. "diskrétní princip maxima". Jeho platnost dokážeme později pro obecnější Laplaceův operátor. Diskrétní princip maxima pro náš případ říká, že pokud $-EA\delta_x^2 U_j/h^2 \leq 0$ pro všechny vnitřní uzly, pak se (nezáporného) maxima nabývá na hranici, tj. v krajních bodech intervalu. Pokud označíme množinu vnitřních uzlů $J_\Omega = \{x_j; j \in \{1,2,...,J-1\}\}$ a množinu hraničních uzlů $\partial J_\Omega = \{x_j; j \in \{0,J\}\}$, pak lze diskrétní princip maxima matematicky zapsat následovně

$$\left(-EA\frac{\delta_x^2 U_P}{h^2}\leq 0;\ \ \forall P\in J_\Omega\right)\ \ \Rightarrow\ \ \max_{P\in J_\Omega} U_P \leq \max\{0, \max_{Q\in\partial J_\Omega} U_Q\}. \tag{1.27}$$

Přitom U_P jsme označili hodnotu funkce U v bodě P. K vlastnímu odvození chyby aproximace definujme pomocnou srovnávací funkci

$$\Phi(x)=\left(x-\frac{1}{2}\right)^2. \tag{1.28}$$

Volba funkce $\Phi(x)$ závisí na řešeném problému a není jednoznačně určena. Obecně lze říci, že se ji snažíme zvolit tak, aby konstanta u odhadu chyby vyšla co nejmenší. Jelikož $\Phi(x)$ má nulové čtvrté derivace, je po dosazení $\Phi(x)$ do odhadu (1.23) konstanta M nulová a po dosazení do (1.26) vychází

$$-EA\frac{\delta_x^2\Phi_P}{h^2}=-2,\quad \forall P\in J_\Omega. \tag{1.29}$$

Označme nyní

$$\Psi_P = e_P + \frac{1}{2}\frac{h^2}{12}M\Phi_P. \tag{1.30}$$

Potom platí

$$-EA\frac{\delta_x^2\Psi_P}{h^2}=-EA\frac{\delta_x^2 e_P}{h^2}-EA\frac{h^2}{12}M=\varepsilon_P-EA\frac{h^2}{12}M\leq 0,\quad \forall P\in J_\Omega, \tag{1.31}$$

kde poslední nerovnost plyne z (1.23). Funkci Ψ_P v libovolném bodě P jsme zvolili právě tak, aby byla pro ni splněna podmínka diskrétního principu maxima. Tudíž platí

$$\Psi_P \leq \frac{1}{2}\frac{h^2}{12}M\max_{Q\in\partial\Omega}\{\Phi_Q\}=\frac{1}{8}\frac{h^2}{12}M,\quad \forall P\in J_\Omega, \tag{1.32}$$

protože v krajních bodech intervalu je srovnávací funkce $\Phi(x)$ rovna $1/4$. Jelikož dále z definice platí $e_P \leq \Psi_P$, dostáváme

$$U_P - u_P \leq \frac{1}{96}Mh^2. \tag{1.33}$$

Budeme-li definovat $\Psi_P = -e_P + \frac{1}{2}\frac{h^2}{12}M\Phi_P$, dostaneme analogický odhad pro $-(U_P - u_P)$, z čehož plyne požadovaný odhad chyby aproximace

$$|U_P - u_P| = |e_P| \leq \frac{1}{96}Mh^2, \tag{1.34}$$

a vidíme, že se zmenšujícím se prostorovým krokem h konverguje schéma kvadraticky k přesnému řešení.

1.2 Parabolické problémy - nestacionární vedení tepla v 1D

Nyní rozvineme, zobecníme a rozšíříme výsledky a metody naznačené v úvodu. Budeme přitom uvažovat následující problém časově závislého vedení tepla v jedné prostorové proměnné

$$v_{,t} = \kappa v_{,xx}, \quad \forall t > 0, x \in (0,1) \tag{1.35}$$

$$v(0,t) = \alpha, \quad v(1,t) = \beta, \quad \forall t > 0, \tag{1.36}$$

$$v(x,0) = v^0(x) = u^0(x) + \alpha(1-x) + \beta x, \quad \forall x \in [0,1]. \tag{1.37}$$

Studenti se s tímto problémem již setkali v předmětu NAK 1. Touto úlohou lze popsat například šíření tepla zdí domu (osa x je kolmá na zeď), nebo v izolované homogenní tyči konečné délky. Parametr κ značí difuzivitu (materiálový parametr popisující vodivost, tj. schopnost materiálu vést teplo), α a β jsou okrajové podmínky, tzn. předepsané teploty na na obou koncích řešené oblasti. Pro zjednodušení analýzy, kterou zde budeme provádět, lze substitucí $u(x,t) = v(x,t/\kappa) - \alpha(1-x) - \beta x$ úlohu převést na tvar

$$u_{,t} = u_{,xx}, \quad \forall t > 0, x \in (0,1) \tag{1.38}$$

$$u(0,t) = u(1,t) = 0, \quad \forall t > 0, \tag{1.39}$$

$$u(x,0) = u^0(x) \tag{1.40}$$

s jednotkovou difuzivitou a homogenními okrajovými podmínkami. Ověření čtenář provede snadno sám porovnáním dvojnásobné derivace v prostorové proměnné a jednoduché derivace v čase. Opačnou substitucí lze naopak rovnice (1.38) - (1.40) převést na rovnice (1.35) - (1.37) s reálným fyzikálním významem. Analytické řešení rovnice (1.38) - (1.40) už není tak snadné, jako v případě stacionárního vedení tepla popsaného rovnicí analogickou (1.1). Za určitých předpokladů ovšem řešení nalézt lze. Postup k jeho nalezení se nazývá Fourierova metoda a základním předpokladem je, že hledáme řešení v separovaném tvaru:

$$u(x,t) = X(x)T(t). \tag{1.41}$$

Uvedený tvar (1.41) má však velmi dobrý smysl a jeho předpoklad je velmi přirozený. Lze se něj totiž dívat jako na rozložení teploty $X(x)$ napříč řešenou oblastí (například stěnou), které má v každém čase jinou amplitudu $T(t)$. Rozložení počáteční teploty $X(x)$ pro $T(0)$ se tedy vyvíjí v čase podle toho, jak se mění $T(t)$. Za tohoto předpokladu prostým dosazením (1.41) do (1.38) dostáváme

$$X(x)T'(t) = X''(x)T(t), \qquad \forall t > 0, \;\; x \in (0,1), \tag{1.42}$$

kde čárkou značíme obyčejnou derivaci. Pokud vydělíme rovnici (1.42) $T(t)$ a označíme $\lambda = -T'(t)/T(t)$, musí funkce $X(x)$ splňovat

$$-X''(x) = \lambda X(x), \qquad x \in (0,1), \;\; X(0) = X(1) = 0. \tag{1.43}$$

Přitom předpokládáme, že $\exists t_0 > 0$ a $x_0 \in (0,1)$ tak, že $u(x_0,t_0) \neq 0$. Vynásobením (1.43) funkcí $X(x)$, jejím integrováním na intervalu od 0 do 1 a postupnou dvojnásobnou aplikací

per-partes postupně dostáváme:

$$\lambda \int_0^1 X^2(x)\mathrm{d}x = -\int_0^1 X''(x)X(x)\mathrm{d}x = -\underbrace{[X'(x)X(x)]_0^1}_{=0} + \int_0^1 [X'(x)]^2\mathrm{d}x > 0, \quad (1.44)$$

z čehož plyne $\lambda > 0$, jelikož integrál z druhé mocniny jakékoliv funkce je kladné číslo. Obecné řešení rovnice (1.43) lze proto psát ve tvaru

$$X(x) = a\,\sin\sqrt{\lambda}x + b\,\cos\sqrt{\lambda}x. \quad (1.45)$$

Z okrajové podmínky $X(0) = 0$ plyne $b = 0$, z $X(1) = 0$ pak $\sqrt{\lambda} = m\pi, m \in \mathbb{N}$. Existuje tedy (spočetně mnoho) vlastních čísel λ_m a odpovídajících vlastních vektorů X_m[4]:

$$\lambda_m = (m\pi)^2, \qquad X_m(x) = \sin(m\pi x). \quad (1.46)$$

Analogicky vyloučením $X(x)$ z (1.42) dostáváme rovnici

$$T'(t) = -\lambda T(t), \qquad \forall t > 0, \quad (1.47)$$

která má (až na multiplikativní konstantu) řešení

$$T(t) = \mathrm{e}^{-\lambda_m t}. \quad (1.48)$$

Vezmeme-li v úvahu původní předpoklad (1.41), můžeme celkové řešení (1.38) - (1.40) psát ve tvaru

$$u(x,t) = a_m \mathrm{e}^{-\lambda_m t}\sin(m\pi x), \quad (1.49)$$

kde a_m je konstanta, kterou určíme níže. Vzhledem k linearitě řešené rovnice je i libovolná lineární kombinace funkcí (1.49) řešením (1.38) - (1.40). Nabízí se nyní uvažovat řešení ve tvaru nekonečné řady

$$u(x,t) = \sum_{m=1}^{\infty} a_m \mathrm{e}^{-\lambda_m t}\sin(m\pi x). \quad (1.50)$$

Z počáteční podmínky pro čas $t = 0$ plyne

$$u(x,0) = u^0(x) = \sum_{m=1}^{\infty} a_m \sin(m\pi x), \quad (1.51)$$

což znamená, že koeficienty a_m jsou Fourierovy koeficienty funkce $u^0(x)$ v rozvoji do "sinové" řady[5]. Proto dále platí

$$a_m = 2\int_0^1 u^0(x)\sin(m\pi x). \quad (1.52)$$

Z teorie Fourierových řad je známo, že řada (1.52) konverguje v prostoru $L^2(0,1)$ pro libovolnou funkci $u^0(x) \in L^2(0,1)$, a tedy (1.50) řeší (1.38) - (1.40).

[4]Zde opět odkazujeme čtenáře na přednášky z Matematiky 3 a 4

[5]Ohledně základních pojmů z teorie Fourierových řad se může čtenář informovat například v Dodatku **??**

Ačkoliv by se mohlo zdát, že získáním analytického řešení "máme vyhráno", je třeba poznamenat, že pro praktické využití se tento výsledek blíží spíše řešení numerickému. Jednak jsme totiž obecně schopni získat koeficienty a_m pouze přibližně (integrál (1.52) nemusíme být obecně schopni vyčíslit přesně), a navíc jsme v každém případě schopni sečíst pouze konečný počet členů řady (1.50). Z hlediska Fourierovy metody samotné je však největším omezením fakt, že není snadno zobecnitelná na komplikovanější úlohy. Je proto třeba se uchýlit k jiným metodám, například metodě konečných diferencí, jako v této kapitole.

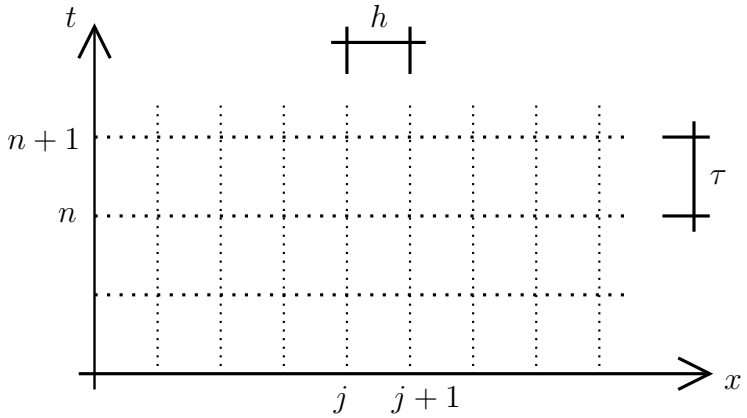
Podobně jako v případě taženého-tlačeného prutu, řešeného v úvodní kapitole, musíme nejprve vytvořit síť, na které budeme diskretizovat rovnici (1.38) - (1.40). Rozdíl však bude v tom, že nyní hledaná funkce teploty $u(x,t)$ závisí kromě prostorové proměnné ještě na čase a musíme proto provést diskretizaci vzhledem k oběma těmto proměnným. Výsledna síť tak bude dvourozměrná, jak je vidět na Obr. 1.4. Zvolme nyní pevné $J \in \mathbb{N}$ a položme $h = 1/J$, přičemž h nazveme prostorový krok sítě a J počet uzlů sítě v prostorové proměnné. Podobně zvolme časový krok sítě $\tau > 0$ a definujme body

$$x_j = jh, \quad j = 0, 1, ..., J, \qquad t_n = n\tau, \quad n = 0, 1, 2, ... \tag{1.53}$$

Dvojice (x_j, t_n) nazýváme uzly sítě (jsou to průsečíky přímek rovnoběžných s osami a procházejícími body x_j a t_n - viz Obr. 1.4). Přitom budeme pro jednoduchost uvažovat konstantní prostorový i časový krok, ačkoliv to obecně není nutné. Podobně jako v úvodním příkladu zavedeme také aproximaci hledané funkce u v bodě (x_j, t_n):

$$u(x_j, t_n) \approx U_j^n. \tag{1.54}$$

Tyto přibližné hodnoty získáme opět tak, že derivace v rovnici (1.38) nahradíme diferencemi[6] a následně řešíme získané diferenční rovnice[7] v čase počínaje $n = 0$.



Obr. 1.4: *Schéma výpočetní sítě pro nestacionární úlohu tepla v 1D*

[6]Správněji bychom měli říkat podílem diferencí, například $\Delta_{+x}u(x)/h$ je podíl diferencí ve funkční hodnotě $u(x)$ a proměnné x. Jelikož ale nehrozí nedorozumění, budeme i pro tento podíl užívat pojmu diference.

[7]Pro stručnou informaci o diferenčních rovnicích, viz například Dodatek **??**.

Druhou derivaci v prostorové proměnné nahradíme (s drobnou úpravou, neboť nyní jde o funkci dvou proměnných) stejně, jako v případě taženého-tlačeného prutu, tedy diferenčním schématem

$$u_{,xx}(x_j,t_n) \approx \frac{\delta_x^2 u(x_j,t_n)}{h^2} = \frac{u(x_j+h,t_n)-2u(x_j,t_n)+u(x_j-h,t_n)}{h^2}. \tag{1.55}$$

V případě časové derivace však máme na výběr z více možností, viz (1.56). Podle toho, zda zvolíme dopřednou či zpětnou diferenci, získáme takzvaně explicitní či implicitní schéma. V dalších odstavcích uvidíme, že tato schémata mají kvalitativně různé vlastnosti.

$$u_{,t}(x_j,t_n) \approx \frac{u(x_j,t_{n+1})-u(x_j,t_n)}{\tau}, \qquad u_{,t}(x_j,t_n) \approx \frac{u(x_j,t_n)-u(x_j,t_{n-1})}{\tau} \tag{1.56}$$

V další kapitole se blíže seznámíme s explicitním schématem pro řešení úlohy (1.38) - (1.40).

1.2.1 Explicitní schéma

Explicitní schéma pro jednorozměrnou úlohu nestacionárního vedení tepla (1.38) - (1.40) je nejjednodušší postup numerického řešení dané úlohy. Pro náhradu derivací v bodě (x_j,t_n) se použijí diference

$$u_{,t}(x_j,t_n) \approx \frac{u(x_j,t_{n+1})-u(x_j,t_n)}{h}, \qquad u_{,xx}(x_j,t_n) \approx \frac{u(x_j+h,t_n)-2u(x_j,t_n)+u(x_j-h,t_n)}{h^2}.$$

Pokud právě uvedené vztahy dosadíme do rovnice (1.38) a místo přesných hodnot funkce $u(x,t)$ dosadíme její aproximace pomocí (1.54), dostaneme diferenční rovnici

$$\frac{U_j^{n+1}-U_j^n}{\tau} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2}, \tag{1.57}$$

kterou po označení $\mu = \tau/h^2$ můžeme přepsat do následujícího tvaru, vhodného pro výpočet

$$U_j^{n+1} = U_j^n + \mu(U_{j+1}^n - 2U_j^n + U_{j-1}^n). \tag{1.58}$$

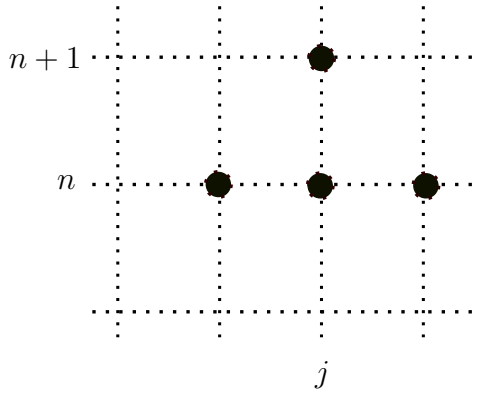
Ze vztahu (1.58) je vidět, že libovolnou (přibližnou) hodnotu hledané funkce U_j^{n+1} na časové hladině t_{n+1} lze spočítat pouze z hodnot na časové hladině t_n - viz Obr. 1.5. Proto se schématu (1.58) říká explicitní diferenční schéma. K rovnici (1.58) ještě musíme připojit počáteční a okrajové podmínky. Ty je třeba rovněž vyjádřit v uzlech (x_j,t_n):

$$U_j^0 = u^0(x_j), \qquad j = 1,2,...,J-1, \tag{1.59}$$

$$U_0^n = U_J^n = 0, \qquad n = 0,1,2,... \tag{1.60}$$

S podmínkami (1.59) a (1.60) lze postupně určit hodnoty pro všechny vnitřní hodnoty intervalu pro všechna $n > 0$. Přitom zatím předpokládáme, že počáteční a okrajové podmínky jsou konzistentní, tj.

$$u^0(0) = u^0(1) = 1. \tag{1.61}$$



Obr. 1.5: *Ilustrace explicitního schématu. Hodnota v čase t_{n+1} se vypočte pouze z hodnot v čase t_n.*

Ukážeme si nyní jednoduchý příklad, abychom motivovali další výklad vyšetřování vlastností numerických schémat. Řešme proto úlohu (1.38) - (1.40) s následující počáteční podmínkou:

$$u^0(x) = \begin{cases} 2x, & x \in \left[0, \frac{1}{2}\right] \\ 2 - 2x, & x \in \left[\frac{1}{2}, 1\right]. \end{cases} \tag{1.62}$$

Při takto definované počáteční podmínce je možné Fourierovou metodou získat libovolně přesné řešení, jelikož integrály (1.52) lze přesně spočítat. Přesné řešení má tvar

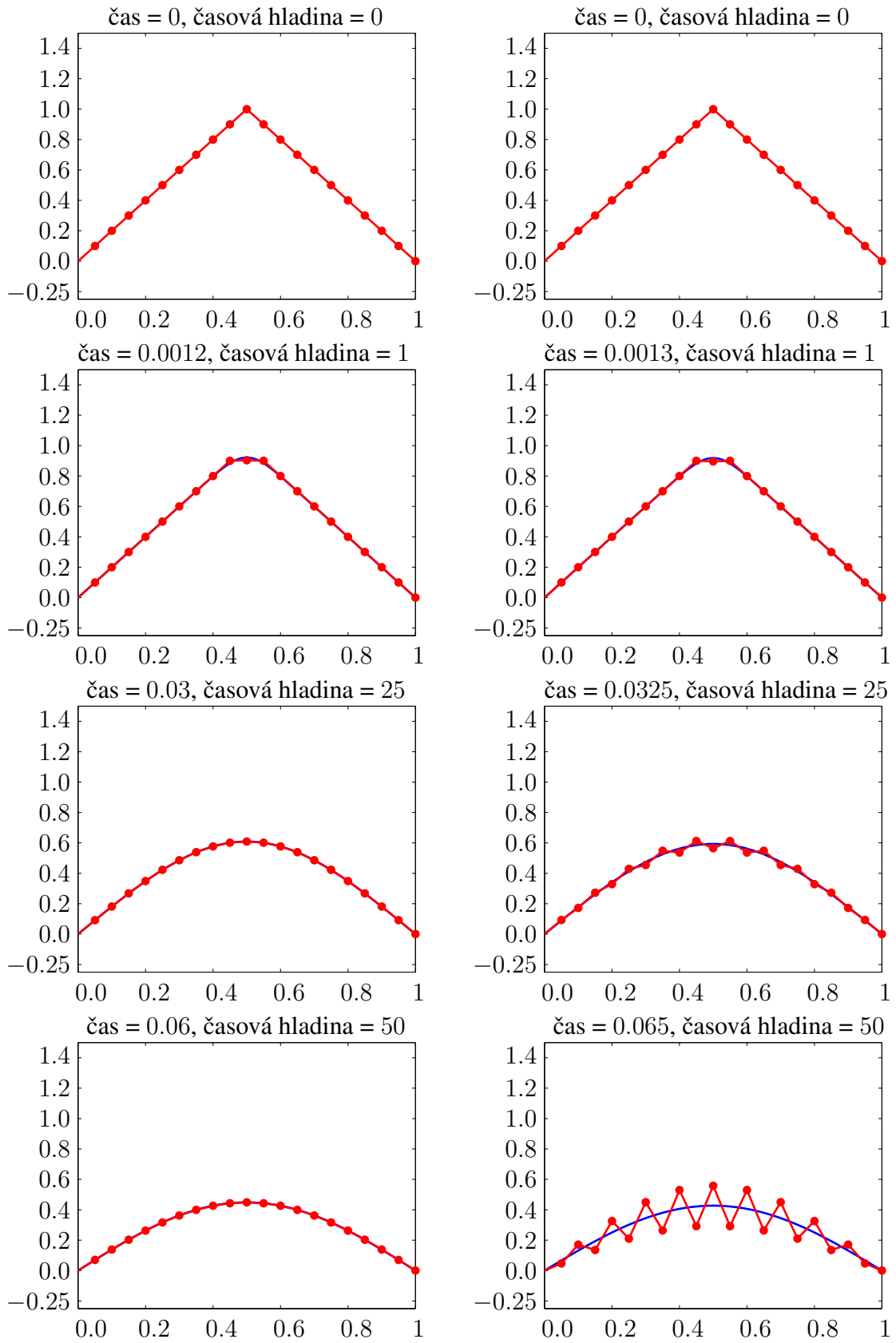
$$u_{exact} = \sum_{m=1}^{\infty} \frac{8}{(m\pi)^2} \sin\left(\frac{m\pi}{2}\right) \sin\left(m\pi x\right) \mathrm{e}^{-(m\pi)^2 t}. \tag{1.63}$$

Řada (1.63) konverguje velice rychle a už pro malá m získáváme dostatečně přesné řešení. To použijeme k porovnání s řešením získaným metodou sítí. Zvolme počet uzlů $J = 20$, čemuž odpovídá velikost prostorového kroku $h = 0.05$. Na Obr. 1.6 je porovnání přesného řešení (modrá čára) a přibližného řešení získaného metodou sítí (červená čára) pro různé hodnoty časového kroku τ. V levém sloupci jsou výsledky pro $\tau = 0.0012$, vpravo pro $\tau = 0.0013$. Zatímco pro první hodnotu časového kroku dostáváme velmi dobrou shodu s přesným řešením, při zdánlivě nepatrné změně o 0.0001 dojde postupně k naprosté destrukci přibližného řešení. Jak uvidíme v další kapitole, tento jev není náhodný, nýbrž jde o principiální záležitost. Jedná se o stabilitu či nestabilitu numerického schématu, v níž hraje klíčovou roli hodnota poměru časového a prostorového kroku $\mu = \tau/h^2$, zavedeného výše. Než se pustíme do analýzy explicitního schématu, zavedeme následující označení pro diference.

Značení pro diference

Dopředné diference:

$$\Delta_{+t} u(x,t) = u(x, t+\tau) - u(x,t), \qquad \Delta_{+x} u(x,t) = u(x+h, t) - u(x,t)$$



Obr. 1.6: Porovnání přesného (modrá čára) a přibližného (červená čára) řešení v čase t = 0, po 1., 25. a 50. kroku. Vlevo je časový krok $\tau = 0.0012$, vpravo $\tau = 0.0013$

Zpětné diference:

$$\Delta_{-t}u(x,t) = u(x,t) - u(x,t-\tau), \qquad \Delta_{-x}u(x,t) = u(x,t) - u(x-h,t)$$

Centrální diference:

$$\delta_t u(x,t) = u\left(x, t+\frac{1}{2}\tau\right) - u\left(x, t-\frac{1}{2}\tau\right), \qquad \delta_x u(x,t) = u\left(x+\frac{1}{2}h, t\right) - u\left(x-\frac{1}{2}h, t\right)$$

Centrální diference s dvojnásobnou délkou intervalu:

$$\Delta_{0x} = \frac{1}{2}(\Delta_{+x} + \Delta_{-x})u(x,t) = \frac{1}{2}\left[u(x+h,t) - u(x-h,t)\right]$$
$$\Delta_{0t} = \frac{1}{2}(\Delta_{+t} + \Delta_{-t})u(x,t) = \frac{1}{2}\left[u(x,t+\tau) - u(x,t-\tau)\right]$$

Centrální diference druhého řádu:

$$\delta_x^2 u(x,t) = \delta_x\left(u\left(x+\frac{1}{2}h,t\right) - u\left(x-\frac{1}{2}h,t\right)\right) = u(x+h,t) - 2u(x,t) + u(x-h,t).$$

Čtenář snadno ověří, že pro centrální diferenci druhého řádu platí $\delta_x^2 = \Delta_{+x}\Delta_{-x}$. Analogicky by se zavedla i diference druhého řádu v časové proměnné.

Nyní se vrátíme k explicitnímu schématu (1.58), které budeme psát ve tvaru

$$\underbrace{\frac{U_j^{n+1} - U_j^n}{\tau}}_{\approx u_{,t}} = \underbrace{\frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2}}_{\approx u_{,xx}}.$$

Podobně jako v úvodní kapitole nyní budeme definovat chybu diskretizace explicitního schématu a vyšetříme, jakého řádu přesnosti toto schéma je. Chybou diskretizace nazveme rozdíl levé a pravé strany (1.64), kam dosadíme místo aproximovaných hodnot U_j^n přesné hodnoty $u(x_j,t_n)$

$$\varepsilon_{h,\tau}(x_j,t_n) = \frac{u(x_j,t_n+\tau) - u(x_j,t_n)}{\tau} - \frac{u(x_j+h,t_n) - 2u(x_j,t_n) + u(x_j-h,t_n)}{h^2} = \frac{\Delta_{+t}u(x_j,t_n)}{\tau} - \frac{\delta_x^2 u(x_j,t_n)}{h^2}. \tag{1.64}$$

Odhad chyby diskretizace provedeme pomocí Taylorova rozvoje, jehož použitím dostaneme následující vztahy

$$\Delta_{+t}u(x,t) = u(x,t+\tau) - u(x,t) = u_{,t}(x,t)\tau + \frac{1}{2}u_{,tt}(x,t)\tau^2 + \frac{1}{6}u_{,ttt}(x,t)\tau^3 + ...$$
$$\delta_x^2 u(x,t) = u(x+h,t) - 2u(x,t) + u(x-h,t) =$$
$$(u(x,t) + u_{,x}(x,t)h + \frac{1}{2}u_{,xx}(x,t)h^2 + \frac{1}{6}u_{,xxx}(x,t)h^3 + ...) - 2u(x,t)$$
$$(u(x,t) - u_{,x}(x,t)h + \frac{1}{2}u_{,xx}(x,t)h^2 - \frac{1}{6}u_{,xxx}(x,t)h^3 + ...) =$$
$$u_{,xx}(x,t)h^2 + \frac{1}{12}u_{,xxxx}(x,t)h^4.$$

Uvážíme-li, že $u_{,t} = u_{,xx}$, dostáváme dosazením právě uvedených Taylorových rozvojů

$$\varepsilon_{h,\tau}(x,t) = \underbrace{\frac{1}{2}u_{,tt}(x,t)\tau - \frac{1}{12}u_{,xxxx}(x,t)h^2}_{\text{hlavní část chyby diskretiace}} + \underbrace{\ldots}_{\text{členy vyššího řádu}}.$$

Podobně jako v úvodním příkladu je však i zde pro zkonstruování odhadu chyby diskretizace výhodnější uvažovat konečné Taylorovy rozvoje.

$$u(x,t+\tau) = u(x,t) + u_{,t}(x,t)\tau + \frac{1}{2}u_{,tt}(x,\eta)\tau^2, \quad \eta \in (t,t+\tau),$$

$$u(x+h,t) = u(x,t)+u_{,x}(x,t)h+\frac{1}{2}u_{,xx}(x,t)h^2+\frac{1}{6}u_{,xxx}(x,t)h^3+\frac{1}{24}u_{,xxxx}(\xi,t)h^2, \quad \xi \in (x-h,h+h).$$

Za předpokladu dostatečné hladkosti počáteční podmínky a její konzistence s okrajovými podmínkami existují konstanty M_1, M_2 tak, že $|u_{,tt}| \le M_1$, $|u_{,xxxx}| \le M_2$ na $[0,1]\times[0,T]$. Potom platí

$$|\varepsilon_{h,\tau}(x,t)| \le \frac{M_1}{2}\tau + \frac{M_2}{12}h^2 = \frac{\tau}{2}\left[M_1 + \frac{1}{6\mu}M_2\right]. \tag{1.65}$$

Pro pevný poměr časového a prostorového kroku μ se $\varepsilon_{h,\tau}(x,t)$ pro $\tau \to 0$ chová jako $\mathcal{O}(\tau)$. Explicitní schéma (1.58) je tedy prvního řádu přesnosti, což znamená, že s poloviční časovým krokem bude i chyba diskretizace poloviční.

Pro zajímavost uveďme, že v tomto konkrétním případě lze jednoduchým trikem dosáhnout i druhého řádu přesnosti. Z rovnosti $u_{,t} = u_{,xx}$ totiž plyne $u_{,tt} = u_{,xxt} = (u_{,t})_{,xx} = u_{,xxxx}$, takže

$$\varepsilon_{h,\tau}(x,t) = \frac{1}{2}\left(1 - \frac{1}{6\mu}\right)u_{,xxxx}(x,t)\tau + \mathcal{O}(\tau^2)$$

a pro $\mu = 1/6$ je schéma druhého řádu přesnosti. Jde však o velmi speciální situaci, se kterou se v obecnějších případech nesetkáme.

1.2.2 Konvergence explicitního schématu

Zatím jsme ukázali, že explicitní schéma (1.58) je prvního řádu přesnosti. V tomto odstavci se budeme zabývat otázkou konvergence explicitního schématu, tedy otázkou, zda pro zjemňující se síť (tzn. $\tau \to 0$ a $h \to 0$) budeme dostávat stále přesnější výsledky, až v limitním případě dostaneme přesné řešení. Odpověď na tuto otázku formulujeme v následující větě:

Věta 1.1 (O konvergenci explicitního schématu)**.** *Uvažujme posloupnost $(h_i,\tau_i) \to (0,0)$ pro $i \to \infty$ a předpokládejme, že existuje i_0 tak, že pro všechna $i > i_0$ platí $\mu_i \equiv \frac{\tau_i}{h_i^2} \le \frac{1}{2}$. Nechť $T > 0$ a existují $M_1, M_2 \in \mathbb{R}$ tak, že $|u_{,tt}| \le M_1$ a $|u_{,xxxx}| \le M_2$ na $[0,1]\times[0,T]$. Pak pro libovolný bod $(x,t) \in [0,1]\times[0,T]$ a libovolnou posloupnost $(j_i,n_i) \in \mathbb{N}$ takovou, že $j_i h_i \to x$, $n_i\tau_i \to t$, konvergují aproximace $U_{j_i}^{n_i}$ generované explicitním diferenčním schématem (1.58) k řešení $u(x,t)$, přičemž tato konvergence je stejnoměrná v $[0,1]\times[0,T]$.*

Důkaz Uvažujme libovolnou dvojici h, τ a libovolný bod $(x_j, t_n) \in (0,1)\times(0,T)$. Analogicky jako v úvodním příkladu taženého-tlačeného prutu definujme chybu aproximace jako $e_j^n = U_j^n - u(x_j, t_n)$. Dosazením do (1.58) dostaneme

$$e_j^{n+1} = e_j^n + \mu(e_{j+1}^n - 2e_j^n + e_{j-1}^n) - \tau\varepsilon_j^n = (1-2\mu)e_j^n + \mu e_{j+1}^n + \mu e_{j-1}^n - \tau\varepsilon_j^n, \qquad (1.66)$$

kde $\varepsilon_j^n = \varepsilon_{h,\tau}(x_j, t_n)$. Definujeme-li maximální chybu v uzlech v rámci n-té časové hladiny jako

$$||e^{n+1}||_\infty = \max_{l=0,\dots J} |e_l^n|, \qquad (1.67)$$

pak "znormováním" obou stran (1.66) postupně dostaneme (přitom používáme vlastnosti normy (2.166 a 2.167)

$$\begin{aligned} ||e^{n+1}||_\infty &= ||(1-2\mu)e_j^n + \mu e_{j+1}^n + \mu e_{j-1}^n - \tau\varepsilon_j^n||_\infty \\ &\le \underbrace{(|1-2\mu| + 2\mu)}_{=1 \text{ pro } \mu\le\frac{1}{2}} ||e^n||_\infty + \tau||\varepsilon_j^n||_\infty. \end{aligned} \qquad (1.68)$$

Jelikož dále pro počáteční podmínku platí $e_l^0 = U_l^0 - u^0(x_l)\ \forall l$, dostáváme za podmínky $\mu \le \frac{1}{2}$ rekurzivně z (1.68)

$$||e^n||_\infty \le \tau \sum_{k=0}^{n-1} ||\varepsilon^k||_\infty \le t_n \max_{k=0,\dots,n-1} ||\varepsilon^k||_\infty, \qquad (1.69)$$

kde v poslední nerovnosti triviálně místo součtu n čísel bereme n-krát největší z nich. Použitím dříve dokázaného odhadu pro chybu diskretizace (1.65) dostaneme požadovaný odhad pro chybu aproximace

$$|U_j^n - u(x_j, t_n)| \le T\left(\frac{M_1}{2}\tau + \frac{M_2}{12}h^2\right). \qquad (1.70)$$

Protože funkce u je na $[0,1]\times[0,T]$ spojitá (to plyne z existenci jejích derivací) a odhad (1.70) nezávisí na x, ani na t, je konvergence stejnoměrná.

Použitím dostatečně jemných sítí lze tedy docílit libovolné přesnosti. Podmínka $\mu \le \frac{1}{2}$ však požaduje, abychom se zmenšujícím se prostorovým krokem h zjemňovali i časový krok, a to velmi dramaticky - dokonce dvakrát rychleji. To je velmi omezující. Navíc z uvedeného není jasné, co se stane, když podmínka $\mu \le \frac{1}{2}$ splněna nebude. Odpověď na tuto otázku dává následující kapitola.

1.2.3 Fourierova analýza chyby pro explicitní schéma

V kapitole (1.2) jsme prezentovali Fourierovu metodu řešení problému (1.38) - (1.40) ze které vyplynulo, že jeho řešení lze psát ve tvaru Fourierovy řady

$$u(x,t) = \sum_{m=1}^{\infty} a_m \mathrm{e}^{-\lambda_m t}\sin(m\pi x), \qquad (1.71)$$

kde z počáteční podmínky

$$u(x,0) = u^0(x) = \sum_{m=1}^{\infty} a_m \sin(m\pi x), \tag{1.72}$$

plynul vztah pro koeficienty a_m jakožto Fourierovy koeficienty v rozvoji $u^0(x)$ do sinové řady.

$$a_m = 2\int_0^1 u^0(x)\sin(m\pi x). \tag{1.73}$$

Řešení (1.71) i počáteční podmínku (1.72) lze zapsat místo funkce sinus pomocí komplexních exponenciálních funkcí jako

$$u(x,t) = \sum_{m=-\infty}^{\infty} A_m \mathrm{e}^{\mathrm{i}m\pi x-(m\pi)^2 t}, \quad u(x,0) = u^0(x) = \sum_{m=-\infty}^{\infty} A_m \mathrm{e}^{\mathrm{i}m\pi x}. \tag{1.74}$$

Pokud dodefinujeme počáteční podmínku i pro $x \in [-1,0]$ jako $u^0(x) = -u^0(x)$[8], pak pro koeficienty A_m platí

$$A_m = \frac{1}{2}\int_{-1}^{1} u^0(x)\mathrm{e}^{-\mathrm{i}m\pi x}. \tag{1.75}$$

Snadno se ověří , že platí $A_m = -A_{-m} = -\mathrm{i}\frac{a_m}{2}$. Dále budeme předpokládat, že Fourierova řada pro počáteční podmínku v (1.74) je absolutně konvergentní, tedy že

$$\sum_{m=-\infty}^{\infty} |A_m \mathrm{e}^{\mathrm{i}m\pi x}| < \sum_{m=-\infty}^{\infty} |A_m| < \infty, \tag{1.76}$$

neboť $|\mathrm{e}^{\mathrm{i}m\pi x}| \leq 1$ pro všechna x.[9] Postačující podmínkou je konzistence počátečních a okrajových podmínek (tj. musí platit $u^0(0) = u^0(1) = 0$), dále absolutní spojitost [10] funkce $u^0(x)$ na $[0,1]$ a derivace $(u^0)'$ musí ležet v prostoru $L^2(0,1)$[11].

Z předchozího víme, že funkce tvaru $\mathrm{e}^{\mathrm{i}m\pi x-(m\pi)^2 t} = \mathrm{e}^{\mathrm{i}m\pi}\mathrm{e}^{-(m\pi)^2 t}$ řeší problém (1.38) - (1.40). V analogii k označení uzlů síťe uvažujme $x = jh$ a $t = n\tau$. Označíme-li $k = m\pi$, můžeme se ptát, zda analogicky $\mathrm{e}^{\mathrm{i}kjh-k^2 n\tau} = \mathrm{e}^{\mathrm{i}kjh}\mathrm{e}^{-k^2 n\tau}$ řeší explicitní diferenční schéma (1.58). Zavedeme-li ještě následující označení

$$\lambda \approx \mathrm{e}^{-k^2\tau}, \tag{1.77}$$

[8]Tedy jako lichou funkci. Připomínáme, že funkce $f(x)$ je lichá, pokud $f(x) = -f(-x)$ a sudá, pokud $f(x) = f(-x)$.

[9]To plyne například z Eulerova vzorce: $\mathrm{e}^{\mathrm{i}m\pi x} = \cos(m\pi x) + \mathrm{i}\sin(m\pi x)$. Tedy $\mathrm{e}^{\mathrm{i}m\pi x}$ leží na jednotkové kružnici.

[10]Existuje několik ekvivalentních definic absolutně spojitých funkcí. Například funkce f je absolutně spojitá na $[a,b]$, pokud existuje (Lebesgueovsky) integrovatelná funkce $g(x)$ tak, že pro všechna $x \in [a,b]$ platí $f(x) = f(a) + \int_a^x g(t)\mathrm{d}t$. Potom i platí, že $g(x) = f'(x)$ pro skoro všechna $x \in [a,b]$ (až na množinu míry 0). Požadavek absolutní spojitosti je velmi silný a lze dokázat, že absolutně spojitá funkce je i stejnoměrně spojitá, a tedy i spojitá.

[11]Pro více informací o prostorech funkcí viz například Dodatek (2.7).

zajímá nás vlastně, zda je možné přibližnou hodnotu funkce $u(x,t)$ v uzlu (x_j, t_n) vyjádřit ve tvaru

$$U_j^n = \mathrm{e}^{\mathrm{i}kjh}\lambda^n. \tag{1.78}$$

Pokud má být (1.78) řešením explicitního schématu (1.58), musíme dosazením dostat identitu. Dosaďme tedy (1.78) do (1.58), čímž obdržíme

$$\mathrm{e}^{\mathrm{i}kjh}\lambda^{n+1} = \mathrm{e}^{\mathrm{i}kjh}\lambda^n[1 + \mu(\mathrm{e}^{\mathrm{i}kh} - 2 + \mathrm{e}^{-\mathrm{i}kh})]. \tag{1.79}$$

Z rovnosti (1.79) plyne po vydělení výrazem $\mathrm{e}^{\mathrm{i}kjh}\lambda$ následující vztah pro λ

$$\lambda \equiv \lambda(k) = 1 + \mu(\mathrm{e}^{\mathrm{i}kh} - 2 + \mathrm{e}^{-\mathrm{i}kh}), \tag{1.80}$$

což lze dále upravit použitím Eulerova vzorce ve tvaru $\mathrm{e}^{\mathrm{i}kh} = \cos(kh) + \mathrm{i}\sin(kh)$ takto

$$\begin{aligned}\lambda(k) = 1 + \mu(\mathrm{e}^{\mathrm{i}kh} - 2 + \mathrm{e}^{-\mathrm{i}kh}) &= 1 + \mu(\cos(kh) + \mathrm{i}\sin(kh) - 2 + \cos(kh) - \mathrm{i}\sin(kh)) \\ &= 1 - 2\mu(1 - \cos(kh)),\end{aligned}$$

což dále použitím vzorce $\cos(kh) = \cos\left[2\left(\frac{kh}{2}\right)\right] = \cos^2\left(\frac{kh}{2}\right) - \sin^2\left(\frac{kh}{2}\right)$ a vzorce $\cos^2\left(\frac{kh}{2}\right) + \sin^2\left(\frac{kh}{2}\right) = 1$ upravíme na

$$\begin{aligned}\lambda(k) = 1 - 2\mu(1 - \cos(kh)) &= 1 - 2\mu\left(1 - \cos^2\left(\frac{kh}{2}\right) + \sin^2\left(\frac{kh}{2}\right)\right) \\ &= 1 - 2\mu\left(\sin^2\left(\frac{kh}{2}\right) + \sin^2\left(\frac{kh}{2}\right)\right) = 1 - 4\mu\sin^2\left(\frac{kh}{2}\right).\end{aligned}$$

Koeficient $\lambda(k)$ se nazývá zesilující (amplifikační) faktor příslušného členu Fourierovy řady. Přibližné řešení U_j^n můžeme tedy zapsat analogickým způsobem, jako v případě přesného řešení $u(x,t)$ jako

$$U_j^n = \sum_{m=-\infty}^{\infty} A_m \mathrm{e}^{\mathrm{i}m\pi jh}[\lambda(m\pi)]^n. \tag{1.81}$$

Možnost přibližného zápisu řešení plyne z omezenosti amplifikačního faktoru $\lambda(m\pi)$ a jednak z konvergence řady (1.76), což znamená, že (1.81) konverguje absolutně. Jelikož každý člen řady (1.81) řeší diferenční rovnici (1.58), řeší ji i její součet. Pro $n = 0$ se řada (1.81) redukuje na řadu pro počáteční podmínku v (1.74) s $x = jh$, a tedy součet řady (1.81) splňuje počáteční podmínku (1.40). Jelikož dále $A_m[\lambda(m\pi)]^n = -A_{-m}[\lambda(-m\pi)]^n$, a tudíž pro $j = 0$ a $j = J$ je součet řady (1.81) roven nule, platí i okrajové podmínky (1.39) a zápis (1.81) je tedy oprávněný.

Vidíme, že $\lambda(m\pi)$ závisí kromě velikosti prostorového kroku h také na m, což je frekvence (nebo také mód) příslušného členu Fourierovy řady. Nízké frekvence (členy s nízkým m) v řadě (1.81) dobře aproximují odpovídající členy řady (1.74) pro přesné řešení. To snadno zjistíme, pokud přesné i aproximované výrazy rozvineme pomocí Taylorova rozvoje

$$\begin{aligned}\mathrm{e}^{-k^2\tau} &= 1 - k^2\tau + \frac{1}{2}k^4\tau^2 - ... \\ \lambda(k) = 1 - 2\mu(1 - \cos(kh)) &= 1 - 2\mu\left[\frac{1}{2}(kh)^2 - \frac{1}{24}(kh)^4 + ...\right] \\ &= 1 - k^2\tau + \frac{1}{12}k^4\tau h^2 - ...,\end{aligned} \tag{1.82}$$

kde vidíme, že $\mathrm{e}^{-k^2\tau}$ a příslušná aproximace $\lambda(k)$ se liší až od třetího členu Taylorova rozvoje. Tyto rozvoje mimo jiné mohou sloužit jako alternativní prostředek vyšetřování chyby diskretizace $\varepsilon_{h,\tau}(x_j,t_n)$, kterou nyní dosazením (1.81) do (1.74) lze psát ve tvaru

$$\begin{aligned}
\varepsilon_{h,\tau}(x_j,t_n) =& \frac{u(x_j,t_n+\tau)-u(x_j,t_n)}{\tau} - \frac{u(x_j+h,t_n)-2u(x_j,t_n)+u(x_j-h,t_n)}{h^2} \\
&= \sum_{m=-\infty}^{\infty} A_m \mathrm{e}^{\mathrm{i}m\pi x_j-(m\pi)^2 t_n}\left[\frac{\mathrm{e}^{-(m\pi)^2\tau}-1}{\tau} - \frac{\overbrace{\mathrm{e}^{\mathrm{i}kh}-2+\mathrm{e}^{-\mathrm{i}kh}}^{2\cos(m\pi h)-2}}{h^2}\right] \\
&= \sum_{m=-\infty}^{\infty} A_m \mathrm{e}^{\mathrm{i}m\pi x_j-(m\pi)^2 t_n}\frac{1}{\tau}\left[\mathrm{e}^{-(m\pi)^2\tau}\underbrace{-1+2\mu(1-\cos(m\pi h))}_{-\lambda(m\pi)}\right] \qquad (\mu=\frac{\tau}{h^2}) \\
&= \sum_{m=-\infty}^{\infty} \frac{\mathrm{e}^{-(m\pi)^2\tau}-\lambda(m\pi)}{\tau} A_m \mathrm{e}^{\mathrm{i}m\pi x_j-(m\pi)^2 t_n} \\
&= \sum_{m=-\infty}^{\infty} \frac{(1-k^2\tau+\frac{1}{2}k^4\tau^2)-(1-k^2\tau+\frac{1}{12}k^4\tau h^2)+\mathcal{O}(\tau^3)}{\tau} A_m \mathrm{e}^{\mathrm{i}m\pi x_j-(m\pi)^2 t_n} \quad \text{(Taylor)} \\
&= \sum_{m=-\infty}^{\infty} \left[\frac{1}{2}k^4\left(\tau-\frac{1}{6}h^2\right)+\mathcal{O}(\tau^2)\right] A_m \mathrm{e}^{\mathrm{i}m\pi x_j-(m\pi)^2 t_n},
\end{aligned}$$

z čehož vidíme, že je schéma prvního řádu přesnosti. Pokud by ovšem $6\tau = h^2$, pak by schéma bylo druhého řadu přesnosti, což je stejný výsledek, který nám vyšel výše, viz (1.66). Snadno lze zjistit následující odhad, který se nám bude hodit později.

$$\begin{aligned}
|\lambda(k)-\mathrm{e}^{-k^2\tau}| &= \left|\frac{1}{12}k^4\tau h^2\cos\xi - \frac{1}{2}k^4\tau^2\mathrm{e}^{-\xi}\right| \qquad ,\xi\in(0,\tau) \\
&\le \left(\frac{1}{12}\frac{h^2}{\tau}+\frac{1}{2}\right)k^4\tau^2 = C(\mu)k^4\tau^2 \quad ,|\cos\xi|\le 1, |\mathrm{e}^{-\xi}|\le 1, \frac{h^2}{\tau}=\frac{1}{\mu}. \quad (1.83)
\end{aligned}$$

Vyjádření (**??**) nám poskytuje ještě více informací. Nejdříve se však podívejme na následující odhad chyby aproximace. Problém zformulujme jako větu.

Věta 1.2. *Chyba $e_j^n = U_j^n - u(x_j,t_n)$ zůstává při pevných τ, h omezená pro $n\to\infty$, a to pro libovolnou počáteční podmínku u^0 splňující (1.76), právě když* $|\lambda(m\pi)|\le 1,\ \ \forall m\in\mathbb{N}$.

Důkaz

Jelikož přesné řešení $u(x_j,t_n)$ je pro $t\to\infty$ omezené (to plyne z konvergence řad (1.50) a (1.52)), stačí se zabývat omezeností přibližného řešení U_j^n pro $n\to\infty$. Pro důkaz ekvivalence je třeba dokázat, že z jednoho tvrzení plyne druhé a naopak. Nechť tedy nejprve $|\lambda(m\pi)|\le$

1, $\forall m \in \mathbb{N}$. Potom ale z (1.81) plyne

$$|U_j^n| \leq \sum_{m=-\infty}^{\infty} |A_m||\lambda(m\pi)|^n \qquad (|\mathrm{e}^{\mathrm{i}m\pi jh}| \leq 1)$$
$$\leq \sum_{m=-\infty}^{\infty} |A_m| < \infty. \qquad (|\lambda(m\pi)| \leq 1 \text{ dle předp., (1.76)})$$

Nechť naopak platí $|U_j^n| < \infty$ a zároveň $\exists m_0 \in \mathbb{N}; |\lambda(m_0\pi)| > 1$. Zvolme počáteční podmínku $u^0(x) = \sin(m_0\pi x)$. Pak je její aproximace $U_j^n = \sin(m_0\pi x_j)[\lambda(m_0\pi)]^n$ a tedy $|U_j^n| \to \infty$ pro $n \to \infty$, což je spor a důkaz je hotov.

Nyní je vidět význam podmínky $\mu \leq \frac{1}{2}$. Je-li tato podmínka splněna, pak platí $|\lambda(m\pi)| \leq 1$ a podle předchozí věty zůstává přibližné řešení omezené. To plyne z

$$\lambda(m\pi) = 1 - 4\mu\sin^2\left(\frac{m\pi h}{2}\right) \leq 1 - 4\mu,$$

odkud je vidět, že $1 > \lambda(m\pi) > -1$. Je-li však $\mu > \frac{1}{2}$, pak pro některé frekvence m bude $\lambda(m\pi) < -1$ a velikost členů řady (1.81) odpovídajících těmto frekvencím s postupujícím časem poroste nade všechny meze. Při bližším pohledu zjistíme, že $\lambda(m\pi) < -1$ pro $m = (2l+1)J,\ l \in \mathbb{Z}$, neboť potom $\lambda(m\pi) = 1 - 4\mu\sin^2\left(\frac{\pi}{2} + l\pi\right) = 1 - 4\mu$. V principu by sice bylo teoreticky možné zvolit počáteční podmínku tak, aby $A_m = 0$ právě pro tato m, je to však velmi speciální situace, která v obecnějších problémech nemusí nastat, a navíc při numerickém řešení by i u těchto frekvencí (které odpovídají problematickým m) vlivem zaokrouhlování vznikly nenulové hodnoty, které by postupem času díky $[\lambda(m\pi)]^n$ neomezeně rostly, a došlo by k destrukci přibližného řešení.

V dosavadní analýze jsme uvažovali v reprezentaci přibližného řešení nekonečné řady (1.81), protože to bylo výhodné pro porovnávání s přesným řešením při vyšetřování chyby. Při praktickém počítání to však není možné. Není to však ani nutné, neboť na diskrétní (a tedy konečné) síti existuje pouze konečné množství frekvencí. To lze snadno vidět následující úvahou, neboť z 2π-periodicity $\mathrm{e}^{\mathrm{i}x} = \cos(x) + \mathrm{i}\sin(x)$ plyne

$$\mathrm{e}^{\mathrm{i}(m+2Jl)\pi jh} = \mathrm{e}^{\mathrm{i}[m\pi jh + 2\pi lj]} = \mathrm{e}^{\mathrm{i}m\pi jh} \tag{1.84}$$

a frekvence m a $m + 2Jl$ od sebe nelze rozeznat. To však znamená, že na dané síti existuje pouze $2J$ nezávislých frekvencí, které zvolíme pro jednoduchost jako

$$m = -(J-1), -(J-2), ..., -1, 0, 1, ...J. \tag{1.85}$$

Funkci U_j^n tedy můžeme popsat jako lineární kombinaci funkcí tvaru $\mathrm{e}^{\mathrm{i}m\pi jh}[\lambda(m\pi)]^n$, kde m probíhá pouze posloupnost (1.85). Přitom nejvyšší frekvence, která se na síti vyskytuje, je $m = J$. Funkci $\mathrm{e}^{\mathrm{i}m\pi jh}[\lambda(m\pi)]^n$ můžeme rozložit na "prostorovou" část $\mathrm{e}^{\mathrm{i}m\pi jh}$ a "časovou" část $[\lambda(m\pi)]^n$. Pro nejvyšší frekvenci $m = J$ se potom prostorová část rovná $\mathrm{e}^{\mathrm{i}\pi j} = \cos(j\pi) + \mathrm{i}\sin(j\pi)$, z čehož je patrné, že v po sobě jdoucích uzlech j střídavě nabývá hodnot ± 1. Jelikož pro $m = J$ se amplifikační faktor rovná $\lambda(J\pi) = 1 - 4\mu\sin^2\left(\frac{\pi}{2}\right) = 1 - 4\mu$ (připomínáme,

že $h = 1/J$), poroste pro $\mu > \frac{1}{2}$ časová část $[\lambda(J\pi)]^n$ nejrychleji a zcela přebije řešení. To se přesně stalo v příkladu na Obr. 1.6.

Fourierova metoda je nesmírně silný nástroj pro vyšetřování schémat MKD. Jako další příklad jejího použití si ukážeme, jak bychom s její pomocí dokázali konvergenci explicitního schématu. Velkou výhodou oproti důkazu (1.1) je skutečnost, že nemusíme předpokládat dostatečnou hladkost řešení a také stejnoměrnou omezenost $u_{,xxxx}$ a $u_{,tt}$. Jediným nutným předpokladem je absolutní konvergence řady (1.52) pro počáteční podmínku. To zejména znamená, že počáteční podmínka nemusí být hladká. Předpokládejme, že $\mu \leq \frac{1}{2}$ je pevné. Nechť dále $\varepsilon > 0$ je libovolné a $m_0 \in \mathbb{N}$ takové, že

$$\sum_{|m|\leq m_0} |A_m| \leq \frac{\varepsilon}{4}. \tag{1.86}$$

Potom pro chybu aproximace postupně dostáváme

$$\begin{aligned}
|e_j^n| = \left|U_j^n - u(x_j,t_n)\right| &= \left|\sum_{m=-\infty}^{\infty} A_m \mathrm{e}^{\mathrm{i}m\pi x_j}\left[(\lambda(m\pi))^n - \mathrm{e}^{-(m\pi)^2 t_n}\right]\right| \quad \text{(dosazení z (1.74) a (1.81))} \\
&\leq \frac{\varepsilon}{2} + \sum_{|m|\leq m_0} |A_m| n \left|(\lambda(m\pi)) - \mathrm{e}^{-(m\pi)^2\tau}\right| \\
&\leq \frac{\varepsilon}{2} + n\tau^2 C(\mu)\pi^4 \sum_{|m|\leq m_0} |A_m| m^4. \qquad \text{(plyne z odhadu (1.83))}
\end{aligned}$$

Pro dostatečně malá τ je $|e_j^n| < \varepsilon$ a to $\forall\ (x_j,t_n) \in [0,1]\times[0,T]$, neboť $n\tau \leq T$.[12]

Viděli jsme, že aby bylo explicitní diferenční schéma (1.58) stabilní, musí platit $\tau \leq \frac{h^2}{2}$, svazující časový a prostorový krok. To je velmi silné omezení, neboť pokud budeme chtít přesnější řešení (a tedy budeme zmenšovat prostorový krok), budeme muset velmi výrazně zmenšit i časový krok a výpočet tak bude trvat velmi dlouho. Tento nedostatek lze odstranit použitím zpětné diference pro diskretizaci časové derivace, což vede na implicitní schéma, kterému je věnována následující kapitola.

1.2.4 Implicitní schéma

V této kapitole si ukážeme implicitní diferenční schéma pro řešení nestacionárního vedení tepla v jedné prostorové dimenzi, tedy pro řešení problému (1.38) - (1.40). Implicitní schéma se od explicitního liší zdánlivě nepatrně - pouze místo dopředné časové diference pro časovou derivaci používáme zpětnou. Má to však zcela zásadní vliv jak na vlastnosti schématu (zejména stabilitu), tak na konkrétní výpočet. Implicitní schéma má tvar

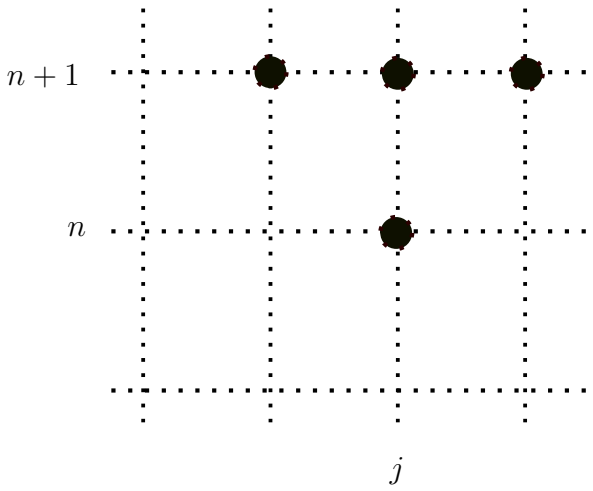
$$\frac{U_j^{n+1} - U_j^n}{\tau} = \frac{U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}}{h^2}, \qquad (\text{tj. } \Delta_{-t}U_j^{n+1} = \mu\delta_x^2 U_j^{n+1}) \tag{1.87}$$

[12] V první nerovnosti tohoto důkazu jsme použili následující skutečnost. Pokud platí $|\lambda_1| \leq 1$ a $|\lambda_2| \leq 1$, pak platí $|\lambda_1^n - \lambda_2^n| \leq n|\lambda_1 - \lambda_2|$. To plyne z binomické věty. Je totiž $(\lambda_1 - \lambda_2)\sum_{j=0}^{n-1}\lambda_1^{n-1-j}\lambda_2^j = \sum_{j=0}^{n-1}\lambda_1^{n-j}\lambda_2^j - \sum_{j=0}^{n-1}\lambda_1^{n-(j+1)}\lambda_2^{j+1} = \lambda_1^n - \lambda_2^n$.

Na skutečnost, že jsme zvolili zpětnou diferenci pro čas se lze dívat také tak, že "podmínku rovnováhy" (v našem případě je přesnější hovořit o bilanci tepla) sestavujeme na konci časového kroku, tedy tam, kam se chceme na časové ose teprve dostat. Přepsáno do podoby vhodné pro výpočet vypadá schéma následovně

$$-\mu U_{j-1}^{n+1} + (1+2\mu)U_j^{n+1} - \mu U_{j+1}^{n+1} = U_j^n, \quad j = 1, 2, ..., J-1, \; n = 0, 1, ... \tag{1.88}$$

Ke schématu pochopitelně opět patří okrajové podmínky, které však mají stejný tvar, jako u schématu explicitního. Podíváme-li se na právě uvedené schéma (1.88), je z praktického hlediska zásadní rozdíl v tom, že nyní nelze určit hodnotu U_j^{n+1} na časové hladině $n+1$ pro žádný uzel j, neboť rovnice (1.88) obsahuje ještě další dvě neznámé hodnoty. Pro ilustraci je to znázorněno na Obr. (1.7). Není tedy jiné cesty, než získat všechny hodnoty v časové hladině $n+1$ současně, a to řešením soustavy $J-1$ rovnic o $J-1$ neznámých (předpis (1.88) vlastně není nic jiného, než "složkový", nebo "indexový" zápis soustavy lineárních algebraických rovnic). Je však třeba říci, že tato soustava je tridiagonální a její řešení je tedy velmi levné (lze použít např. tzv. Thomasův algoritmus).



Obr. 1.7: *Ilustrace implicitního schématu. Hodnoty v čase t_{n+1} lze vypočitat pouze z najednou řešením soustavy algebraických rovnic.*

1.2.5 Fourierova analýza chyby pro implicitní schéma

Jak jsme již uvedli v úvodu k implicitnímu schématu a jak bylo patrné ze závěru vyšetřování schématu explicitního, je nejdůležitější výhodou (implicitního schématu) fakt, že časové kroky mohou být mnohem delší, neboť zde neexistuje žádná omezující podmínka na μ, potažmo na volbu časového kroku τ. Celkové výpočetní nároky, a to i přesto, že musíme v každém časovém kroku řešit soustavu lineárních rovnic, značně poklesnou, neboť časový krok může být i řádově větší, a to zejména při velmi malém prostorovém kroku h. Že zde skutečně žádné takové omezení není, ukážeme opět pomocí Fourierovy analýzy.

V analogii k postupu pro explicitní schéma dosadíme přibližné řešení ve tvaru $U_j^n = \mathrm{e}^{\mathrm{i}kjh}\lambda^n$ do (1.88), čímž dostaneme

$$-\mu\mathrm{e}^{\mathrm{i}k(j-1)h}\lambda^{n+1} + (1+2\mu)\mathrm{e}^{\mathrm{i}kjh}\lambda^{n+1} - \mu\mathrm{e}^{\mathrm{i}k(j+1)h}\lambda^{n+1} = \mathrm{e}^{\mathrm{i}kjh}\lambda^n, \tag{1.89}$$

což po vydělení $\mathrm{e}^{\mathrm{i}kjh}\lambda^n$ dává

$$\lambda(-\mu\mathrm{e}^{-\mathrm{i}kh} + (1+2\mu) - \mu\mathrm{e}^{\mathrm{i}kh}) = 1. \tag{1.90}$$

Jelikož podobně jako v odvozování u explicitního schématu platí pomocí Eulerova vzorce $\mathrm{e}^{-\mathrm{i}kh} + \mathrm{e}^{\mathrm{i}kh} = 2\cos(kh) = 2 - 4\sin^2\frac{kh}{2}$, dostáváme, že

$$\lambda = \frac{1}{1 + 4\mu\sin^2\frac{kh}{2}}. \tag{1.91}$$

Protože výraz $1+4\mu\sin^2\frac{kh}{2}$ je vždy větší nebo roven nule, leží pro libovolné $\mu > 0$ amplifikační faktor $\lambda \in (0,1)$. Tedy schéma (1.88) je stabilní nezávisle na volbě τ a h. V takovém případě říkáme, že je schéma nepodmíněně stabilní.

Chyba diskretizace se stanoví v případě implicitního schématu úplně stejně, jako v případě schématu explicitního. Diferenční náhrada prostorové derivace je stejná, v čase se použije zpětná diference místo dopředné. Ta je však stejného řádu chyby, jak je snadno vidět z následujícího.

$$\Delta_{-t}u(x,t) = u(x,t) - u(x,t-\tau) = u_{,t}(x,t)\tau - \frac{1}{2}u_{,tt}(x,t)\tau^2 + \frac{1}{6}u_{,ttt}(x,t)\tau^3 + ...$$

Pro centrální diferenci druhého řádu jsme již dříve odvodili vztah

$$\delta_x^2 u(x,t) = u_{,xx}(x,t)h^2 + \frac{1}{12}u_{,xxxx}(x,t)h^4. \tag{1.92}$$

Spojením pak vychází chyba diskretizace

$$\varepsilon_{h,\tau}(x,t) = -\frac{1}{2}u_{,tt}(x,t)\tau - \frac{1}{12}u_{,xxxx}(x,t)h^2 + ...,$$

což vede stejně jako v případě explicitního schématu na chybu řádu $\mathcal{O}(\tau)$.

Poznámka Mohlo by nás napadnout, že když centrální diference jsou druhého řádu přesnosti (zatím to víme jen o diferenci druhého řádu, ale můžeme prozradit, že totéž platí i pro centrální diferenci prvního řádu - to čtenář snadno ověří pomocí Taylorova rozvoje), mohli bychom použitím centrální diference na časovou derivaci docílit schématu, které bude druhého řádu přesnosti jak v čase, tak v prostoru. Podívejme se tedy na schéma

$$\frac{U_j^{n+1} - U_j^{n-1}}{2\tau} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2}. \qquad (\text{tj. } \frac{\Delta_{0t}U_j^n}{2\tau} = \frac{\delta_x^2 U_j^n}{h^2}). \tag{1.93}$$

Toto schéma je dvoukrokové, neboť se v něm vyskytují tři časové hladiny. Jelikož na počátku známe jen počáteční data, je možné první krok vypočíst například explicitním schématem uvedeným výše. Tím získáme z počátečních dat první časovou hladinu a můžeme již počítat pomocí schématu (1.93). Ukázat, že je centrální diferenceΔ_{0t} druhého řádu přesnosti je snadné,

a přenecháváme to čtenáři za cvičení. Celkem vychází, že je schéma (1.93) řádu $\mathcal{O}(\tau^2 + h^2)$. Podívejme se nyní na Fourierovu analýzu schématu (1.93). Za tím účelem do schématu dosaďme $U_j^n = \mathrm{e}^{\mathrm{i}kjh}\lambda^n$. Po chvíli snažení (které je však zcela analogické předchozí analýze - vykrácení $\mathrm{e}^{\mathrm{i}kjh}\lambda^{n-1}$, použití dříve odvozeného vztahu $\mathrm{e}^{\mathrm{i}kh} - 2 + \mathrm{e}^{-\mathrm{i}kh} = 1 - 4\mu\sin^2\left(\frac{kh}{2}\right)$ a drobných úpravách) obdržíme následující kvadratickou rovnici pro λ

$$\lambda^2 + 8\mu\sin\left(\frac{1}{2}kh\right)\lambda - 1 = 0. \tag{1.94}$$

Podrobnější analýzou (zapojením diferenčních rovnic), která přesahuje rámec tohoto textu lze ukázat, že je vždy jeden kořen $\lambda > 1$, a tedy že schéma (1.93) je vždy nestabilní (a tedy z praktického hlediska k ničemu). To však neznamená, že každé dvoukrokové schéma je nestabilní. Jako příklad uveďme schéma

$$\frac{U_j^{n+1} - U_j^{n-1}}{2\tau} = \frac{\theta\delta_x^2 U_j^{n+1} + (1-\theta)\delta_x^2 U_j^n}{h^2}, \tag{1.95}$$

které je druhého řádu přesnosti a pro $\theta \leq \frac{1}{2}$ je stabilní, pokud $\mu \leq \frac{1}{4(1-\theta)}$, což však znamená podmíněnou stabilitu, takže jsme si rovněž příliš nepomohli. Optimální řešení se skrývá pod formulací, kterou se budeme zabývat v následujícím odstavci. Jde o tzv. θ-schéma.

1.2.6 θ-schéma

V této kapitole se budeme zabývat patrně nejpoužívanějším schématem pro řešení (nejen) úlohy (1.38) - (1.40). Jde o tzv. θ-schéma, které je v jistém smyslu něčím mezi explicitním a implicitním schématem. Přesněji řečeno jde o schéma, kde časovou derivaci nahrazujeme dopřednou diferencí, a pro druhou derivaci v prostoru volíme (konvexní) lineární kombinaci explicitního a implicitního schématu. Schéma má následující podobu

$$\frac{U_j^{n+1} - U_j^n}{\tau} = \frac{\theta\delta_x^2 U_j^{n+1} + (1-\theta)\delta_x^2 U_j^n}{h^2}, \tag{1.96}$$

kde parametr θ leží v intervalu $[0, 1]$. Pro $\theta = 0$ přitom dostáváme explicitní schéma, $\theta = 1$ pak plně implicitní schéma. Opět je možné schéma přepsat do podoby vhodnější pro výpočet, které má pro libovolné $\theta \in (0, 1]$ tvar

$$-\theta\mu U_{j-1}^{n+1} + (1+2\mu)U_j^{n+1} - \theta\mu U_{j+1}^{n+1} = [1 + (1-\theta)\mu\delta_x^2]U_j^n, \quad j = 1, 2, ..., J-1,\ n = 0, 1, ... \tag{1.97}$$

Podívejme se nejdříve na chybu diskretizace θ-schématu, kterou definujeme stejně jako v předchozích kapitolách jako rozdíl levé a pravé strany (1.96). V tomto případě je z praktických důvodů vhodnější vyhodnocovat v čase $t_{n+\frac{1}{2}}$, tzn. v polovině mezi časy n a $n+1$. Definujme tedy

$$\varepsilon_j^{n+\frac{1}{2}} = \varepsilon_{h,\tau}(x_j, t_{n+\frac{1}{2}}) = \frac{u(x_j, t_n + \tau) - u(x_j, t_n)}{\tau} - \frac{\theta\delta_x^2 u(x_j, t_{n+1}) + (1-\theta)\delta_x^2 u(x_j, t_n)}{h^2}. \tag{1.98}$$

Označme pro jednoduchost centrální diferenci s polovičním krokem v čase $t_{n+\frac{1}{2}}$ jako

$$\frac{\delta_t u\left(x_j, t_{n+\frac{1}{2}}\right)}{\tau} = \frac{u(x_j, t_n + \tau) - u(x_j, t_n)}{\tau}.$$

Potom můžeme psát (1.98) v libovolném bodě (x, t) jako

$$\varepsilon_{h,\tau}(x,t) = \frac{\delta_t u\,(x,t)}{\tau} - \left(\theta - \frac{1}{2}\right)\frac{\delta_t \delta_x^2 u(x,t)}{h^2} - \frac{1}{2}\frac{\delta_x^2 u(x, t+\frac{\tau}{2}) + \delta_x^2 u(x, t-\frac{\tau}{2})}{h^2}, \qquad (1.99)$$

kde jsme pouze vhodně přeskupili členy (1.98). Jelikož vyšetřování chyby diskretizace rozvojem do Taylorovy řady jsme prováděli již výše a postup je stále týž, budeme postupovat již rychleji. Taylorovy rozvoje jednotlivých členů vyskytujících se ve schématu jsou

$$\begin{aligned}\delta_t u(x,t) = u(x, t+\frac{\tau}{2}) - u(x, t-\frac{\tau}{2}) &= \sum_{k\geq 0}\frac{1}{k!}\frac{\partial^k u}{\partial t^k}(x,t)\left(\frac{\tau}{2}\right)^k - \sum_{k\geq 0}\frac{1}{k!}\frac{\partial^k u}{\partial t^k}(x,t)\left(-\frac{\tau}{2}\right)^k \\ &= 2\sum_{\substack{k\geq 0\\ \text{k liché}}}\frac{1}{k!}\frac{\partial^k u}{\partial t^k}(x,t)\left(\frac{\tau}{2}\right)^k = u_{,t}(x,t) + \frac{1}{24}u_{,ttt}(x,t)\tau^3 + ...,\end{aligned}$$

dále

$$\frac{1}{2}\left[u(x, t+\frac{\tau}{2}) + u(x, t-\frac{\tau}{2})\right] = \sum_{\substack{k\geq 0\\ \text{k sudé}}}\frac{1}{k!}\frac{\partial^k u}{\partial t^k}(x,t)\left(\frac{\tau}{2}\right)^k = u(x,t) + \frac{1}{8}u_{,tt}(x,t)\tau^2 + ...,$$

a jak jsme ukázali výše při vyšetřování chyby diskretizace explicitního schématu,

$$\delta_x^2 u(x,t) = u_{,xx}(x,t)h^2 + \frac{1}{12}u_{,xxxx}(x,t)h^4 + \frac{2}{6!}u_{,xxxxxx}(x,t)h^6 + ...$$

Tedy celkem dostáváme (pro jednoduchost vynecháme bod (x, t), ve kterém výrazy vyhodnocujeme)

$$\begin{aligned}\varepsilon_{h,\tau}(x,t) = &\left[u_{,t} + \frac{1}{24}u_{,ttt}\tau^2 + ...\right] - \left(\theta - \frac{1}{2}\right)\left[u_{,xxt} + \frac{1}{12}u_{xxxxt}h^2\tau + ...\right] + \\ &- \left[u_{,xx} + \frac{1}{12}u_{,xxxx}h^2 + \frac{2}{6!}u_{,xxxxxx}(x,t)h^4 + \frac{1}{8}u_{xxtt}\tau^2 + ...\right], \qquad (1.100)\end{aligned}$$

z čehož vychází, že chyba $\varepsilon_{h,\tau}(x,t) = \mathcal{O}(\tau + h^2)$, což je obecně stejný řád přesnosti, jako u explicitní metody.[13]. Při vhodné volbě parametru θ však můžeme dosáhnout i lepší přesnosti. Například pro $\theta = \frac{1}{2}$ dostáváme $\varepsilon_{h,\tau}(x,t) = \mathcal{O}(\tau^2 + h^2)$. Schéma (1.97) je pro tuto speciální volbu θ druhého řádu přesnosti v prostoru i v čase. Toto schéma se pak nazývá Crankovo-Nicholsonové.

Podívejme se nyní na vyšetření stability θ-schématu. Tu provedeme opět pomocí Fourierovy analýzy.

[13]Přesná analýza při použití konečných Taylorových rozvojů by vedla na následující závěr. Pokud existují konstanty $M_1 - M_3$ tak, že $|u_{,tt}| \leq M_1$, $u_{,xxxx}| \leq M_2$, $u_{,xxt}| \leq M_3$, potom platí, že $\varepsilon_{h,\tau}(x,t) \leq \left(\frac{M_1}{4} + M_3\right)\tau + \frac{M_2}{6}h^2 + \left|\theta - \frac{1}{2}\right|\left(M_3\tau + \frac{M_2}{6}h^2\right)$.

1.2.7 Fourierova analýza chyby pro θ-schéma

Stabilitu θ-schématu vyšetříme dosazením $U_j^n = \mathrm{e}^{\mathrm{i}kjh}\lambda^n$ do (1.97). Tím obdržíme (po zkrácení výrazem $\mathrm{e}^{\mathrm{i}kjh}\lambda$)

$$\begin{aligned}\lambda - 1 &= \mu[\theta\lambda(\mathrm{e}^{\mathrm{i}kh} - 2 + \mathrm{e}^{-\mathrm{i}kh}) + (1-\theta)(\mathrm{e}^{\mathrm{i}kh} - 2 + \mathrm{e}^{-\mathrm{i}kh})] \\ &= \mu[\theta\lambda + (1-\theta)]\left(-4\sin^2\frac{kh}{2}\right),\end{aligned}$$

z čehož dále plyne

$$\lambda = \frac{1 - 4(1-\theta)\mu\sin^2\frac{kh}{2}}{1 + 4\theta\mu\sin^2\frac{kh}{2}}. \tag{1.101}$$

Jelikož $\theta \in [0,1]$, $\mu > 0$ a $\sin^2 x > 0$, je zřejmé, že $\lambda < 1$ a nestabilita tedy může nastat pouze pokud by $\lambda < -1$ (dokázali jsme totiž, že schéma je stabilní právě tehdy, pokud $|\lambda| < 1$). To ovšem díky (1.101) může nastat pouze tehdy, pokud

$$1 - 4(1-\theta)\mu\sin^2\frac{kh}{2} < -1 - 4\theta\mu\sin^2\frac{kh}{2}, \tag{1.102}$$

což platí právě pokud

$$(1-2\theta)\mu\sin^2\frac{kh}{2} > \frac{1}{2}. \tag{1.103}$$

Bude-li $(1-2\theta)\mu > \frac{1}{2}$, bude nejrychleji oscilující člen (tedy člen s nejvyšší frekvencí), pro který je $kh = \pi$, nestabilní. V tom je zahrnutý i dřívější výsledek pro explicitní schéma, neboť θ-schéma je pro $\theta = 0$ totožné, jako explicitní schéma. Zároveň je z (1.103) patrné, že plně implicitní schéma při $\theta = 1$ není nestabilní pro žádnou hodnotu μ, protože je na levé straně záporné číslo a nemůže tedy nastat $\lambda < -1$. Celkem jsme tedy získali následující podmínky pro stabilitu θ-schématu (1.97)

$$\text{Je-li } \theta \in \left[0, \frac{1}{2}\right), \text{ pak je } \theta\text{-schéma stabilní} \Leftrightarrow \quad \mu \le \frac{1}{2(1-2\theta)} \tag{1.104}$$

$$\text{Je-li } \theta \in \left[\frac{1}{2}, 1\right], \text{ pak je } \theta\text{-schéma stabilní} \quad \forall\, \mu. \tag{1.105}$$

V prvním případě jde o podmíněnou, ve druhém o nepodmíněnou stabilitu.

Výše uvedené schéma Crankovo-Nicholsonové je tedy nepodmíněně stabilní (tzn. není zde žádná závislost mezi časovým a prostorovým krokem) a můžeme tak psát, že je řádu $\mathcal{O}(\tau^2)$. Jsme pomocí něj schopni počítat s velmi dobrou přesností a zároveň malými výpočetními nároky. Nejlepší volba parametru θ však závisí na řešeném problému a obecně není jasné, která hodnota je opravdu nejlepší.

Nyní se podíváme na alternativní prostředek pro získaní podmínek stability. Ten je založen na tzv. diskrétním principu maxima.

1.2.8 Diskrétní princip maxima a θ-schéma

Diskrétní princip maxima dává alternativní postup, jak získat podmínky stability daného schématu. Je založen na myšlence, že pokud neexistují v řešené oblasti vnitřní zdroje, pak přibližné řešení v žádném časovém kroku nemůže být nikde uvnitř řešené oblasti menší, resp. větší, než nejmenší, resp. největší hodnota z okrajových podmínek na hranici a počáteční podmínky. Z toho plyne, že přibližné řešení zůstává po celou dobu omezené a tedy stabilní. Celý problém zformulujeme jako následující větu.

Věta 1.3. *Mějme θ-schéma (1.97). Nechť platí $\theta \in [0,1]$ a $\mu(1-\theta) \leq \frac{1}{2}$. Potom přibližné hodnoty $\{U_j^n\}$, počítané tímto schématem, splňují*

$$U_{min} \leq U_j^n \leq U_{max}, \tag{1.106}$$

kde

$$U_{min} = \min\{U_0^m,\ 0 \leq m \leq n;\ \ U_j^0,\ 0 \leq j \leq J;\ \ U_J^m,\ 0 \leq m \leq n\} \tag{1.107}$$

$$U_{max} = \max\{U_0^m,\ 0 \leq m \leq n;\ \ U_j^0,\ 0 \leq j \leq J;\ \ U_J^m,\ 0 \leq m \leq n\}. \tag{1.108}$$

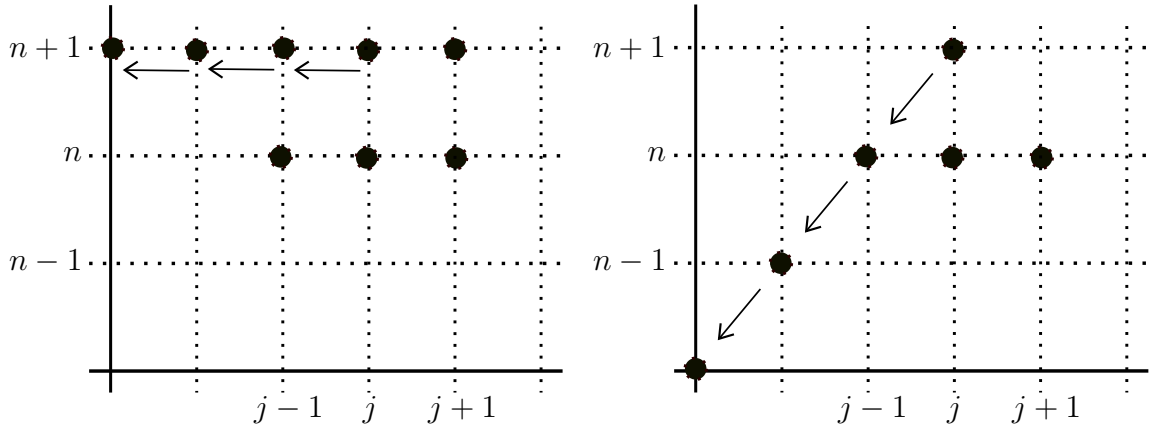
Důkaz
Abychom toto tvrzení dokázali, přepišme nejprve θ-schéma (1.97) do následujícího tvaru

$$(1+2\theta\mu)U_j^{n+1} = \theta\mu(U_{j-1}^{n+1} + U_{j+1}^{n+1}) + (1-\theta)\mu(U_{j-1}^n + U_{j+1}^n) + [1-2(1-\theta)\mu]U_j^n. \tag{1.109}$$

Dále si všimněme, že koeficienty u jednotlivých U na pravé straně (1.109) jsou všechny nezáporné (to plyne z toho, že $\theta \in [0,1]$ a $\mu \geq 0$) a jejich součet je roven $(1+2\theta\mu)$, tedy je roven koeficientu na levé straně (1.109). Nyní v rozporu s tvrzením věty předpokládejme, že U nabývá svého maxima ve vnitřním bodě a nechť je tohoto maxima nabýváno v bodě U_j^{n+1}. Jelikož tato hodnota se nikde na pravé straně (1.109) nevyskytuje, musí pro všechny hodnoty platit, že jsou menší nebo rovny, než U_j^{n+1} (jelikož nemohou být větší, než maximum). Protože ale součet koeficientů u těchto hodnot na pravé straně (1.109) je roven koeficientu na levé straně u U_j^{n+1}, musí nutně platit, že pokud je příslušný koeficient nenulový, pak $U = U_j^{n+1}$ v každém z pěti bodů na pravé straně (1.109), neboť jinak by se obě strany této rovnosti nemohly rovnat. Pro $\theta \neq 0$ tedy sestrojíme v dané časové hladině $n+1$ posloupnost hodnot U až se dostaneme na hranici, tj. $U_j^{n+1} = U_0^{n+1} = U_J^{n+1}$, jak je ilustrováno na Obr. (1.8) vlevo. Pokud je $\theta = 0$ pak se jedná o explicitní schéma a lze analogickou posloupnost sestrojit diagonálně (neboť se tyto hodnoty na diagonále musejí ze stejného důvodu, jako v případě $\theta \neq 0$, rovnat), jak je ilustrováno na Obr. (1.8) vpravo. Tedy v obou případech nakonec dostáváme $U_j^{n+1} = U_{max}$. Obdobnou úvahu lze provést i pro minimum a důkaz je tedy hotov.

V závěru této kapitoly ještě uvedeme větu o konvergenci θ-schématu. Opět tento problém zformulujeme jako větu.

Věta 1.4. *Uvažujme posloupnost $(h_i, \tau_i) \to (0,0)$ pro $i \to \infty$ a předpokládejme, že existuje i_0 tak, že pro všechna $i > i_0$ platí $\mu_i(1-\theta) \equiv \frac{\tau_i}{h_i^2} \leq \frac{1}{2}$. Nechť $T > 0$ a existují $M_1, M_2 \in \mathbb{R}$*



Obr. 1.8: Ilustrace sestrojení posloupnosti hodnot U k hranici výpočetní oblasti pro $\theta \neq 0$ a $\theta = 0$

tak, že $|u_{,tt}| \leq M_1$ *a* $|u_{,xxxx}| \leq M_2$ *na* $[0,1] \times [0,T]$ *(tj. chyba diskretizace* θ*-schématu (1.97) konverguje stejnoměrně k nule na* $[0,1] \times [0,T]$*). Nechť chyba v okrajových a počátečních podmínkách rovněž konverguje stejnoměrně k nule pro* $i \to \infty$*. Pak pro libovolný bod* $(x,t) \in [0,1] \times [0,T]$ *a libovolnou posloupnost* $(j_i, n_i) \in \mathbb{N}$ *takovou, že* $j_i h_i \to x,\ n_i \tau_i \to t$*, konvergují aproximace* $U_{j_i}^{n_i}$ *generované* θ*-schématem (1.97) k řešení* $u(x,t)$ *s konzistentními okrajovými a počátečními podmínkami, přičemž tato konvergence je stejnoměrná v* $[0,1] \times [0,T]$.

Důkaz
Chybu aproximace jsme již dříve definovali jako $e_j^n = U_j^n - u(x_j, t_n)$. Uvažujme libovolnou dvojici h, τ a libovolný bod $(x_j, t_n) \in (0,1) \times (0,T)$. Dosazením e_j^n do (1.97) dostaneme analogicky jako v případě Věty o konvergenci explicitního schématu

$$(1+2\theta\mu)e_j^{n+1} = \theta\mu(e_{j-1}^{n+1}+e_{j+1}^{n+1})+(1-\theta)\mu(e_{j-1}^{n}+e_{j+1}^{n})+[1-2(1-\theta)\mu]e_j^n-\tau\varepsilon_j^{n+\frac{1}{2}}. \quad (1.110)$$

Předpokládejme nyní, že chyby v počátečních a okrajových podmínkách jsou nulové, tj. $e_j^0 = 0$ pro $j = 0, ..., J$, $e_0^n = e_J^n = 0$ pro $n = 0, 1, 2, ...$ a označme maximální chybu v uzlech v rámci n-té časové hladiny jako

$$||e^n||_\infty = \max_{l=0,...J} |e_l^n|, \quad (1.111)$$

a chybu diskretizace

$$||\varepsilon^{n+\frac{1}{2}}||_\infty = \max_{l=0,...J} |\varepsilon_j^{n+\frac{1}{2}}|, \quad (1.112)$$

pak analogicky jako v případě Věty o konvergenci explicitního shcématu dostaneme

$$(1+2\theta\mu)||e^{n+1}||_\infty \leq 2\theta\mu||e^{n+1}||_\infty + ||e^n||_\infty + \tau||\varepsilon^{n+\frac{1}{2}}||_\infty, \quad (1.113)$$

z čehož dále plyne

$$||e^{n+1}||_\infty \leq ||e^n||_\infty + \tau||\varepsilon^{n+\frac{1}{2}}||_\infty, \quad (1.114)$$

a tedy rekurentně

$$||e^{n+1}||_\infty \leq \tau \sum_{m=0}^{n-1} ||\varepsilon^{m+\frac{1}{2}}||_\infty \leq n\tau \max_{m=0,...,n-1} ||\varepsilon^{m+\frac{1}{2}}||_\infty \to 0, \tag{1.115}$$

pro $i \to \infty$ a tedy chyba aproximace konverguje k nule.

Předpokládejme nyní, že chyby v počátečních a okrajových podmínkách jsou nenulové, tj.

$$e_j^0 = \eta_j^0, \;\; j = 0, ..., J, \quad e_0^n = \eta_0^n, \;\; e_J^n = \eta_J^n, \;\; n = 0, 1, 2, ... \tag{1.116}$$

Potom lze chybu aproximace rozložit na dvě části $e_j^n = \bar{e}_j^n + \tilde{e}_j^n$, kde $\bar{e}_j^n$ splňuje (1.110) s homogenními okrajovými podmínkami a $\tilde{e}_j^n$ splňuje (1.109) a (1.116). Potom ovšem $||\bar{e}_j^n||_\infty$ splňuje právě dokázaný odhad (1.115) a $\tilde{e}_j^n$ splňuje předpoklady věty o diskrétním principu maxima a tedy platí

$$||\tilde{e}_j^n||_\infty \leq \max\{|\eta_0^m|, \; m = 0, ..., n-1, \;\; |\eta_j^0|, \; j = 0, ..., J, \;\; |\eta_J^m|, \; m = 0, ..., n-1, \} \tag{1.117}$$

a $||\tilde{e}_j^n||_\infty$ zůstává omezená a jelikož dle předpokladu chyby v počátečních a okrajových podmínkách konvergují k nule, konverguje i $||\tilde{e}_j^n||_\infty$ k nule a důkaz je hotov.

Máme tedy k dispozici dvě metody určování stability diferenčních schémat - Fourierovy analýzu a diskrétní princip maxima. Jejich porovnáním zjistíme, že podmínky pro platnost diskrétního principu maxima jsou mnohem více omezující, než podmínky nutné pro Fourierovu analýzu. Například pro $\theta = \frac{1}{2}$ dostáváme z $\mu(1-\theta) \leq \frac{1}{2}$ nutně $\mu \leq 1$. Přitom z Fourierovy analýzy plyne, že při $\theta = \frac{1}{2}$ už žádnou podmínku na μ nepotřebujeme. Princip maxima má však tu výhodu, že ho zle snadno aplikovat i na úlohy s nekonstantními koeficienty, jak uvidíme v další kapitole. Snadné je však odvodit pouze postačující podmínky stability (nutné podmínky jsou mnohem složitější a obvykle není jasné, jak je získat).

1.2.9 Obecnější lineární rovnice

V této kapitole se podíváme na obecnější tvar lineárních parabolických rovnic. Začneme přitom jednoduchým případem vedení tepla, kde bude prostorově i časově závislá difuzivita. Takovou úlohu můžeme matematicky popsat následovně.

$$u_{,t} = bu_{,xx}, \quad \forall t > 0, \;\; x \in (0, 1), \tag{1.118}$$

kde $b = b(x, t) > 0$ má fyzikální význam difuzivity. Použitím explicitního schématu na rovnici (1.118) dostaneme následující schéma

$$U_j^{n+1} = U_j^n + \frac{\tau}{h^2} b_j^n (U_{j+1}^n - 2U_j^n + U_{j-1}^n), \tag{1.119}$$

kde $b_j^n = b(x_j, t_n)$ je difuzivita b vyhodnocená v bodě (x_j, t_n). Rozvojem jednotlivých členů do Taylorových rozvojů bychom stejně jako dříve mohli vyšetřit chybu diskretizace, která by v tomto případě měla tvar

$$\varepsilon_{h,\tau}(x, t) = \frac{1}{2} u_{,tt}\tau - \frac{1}{12} b(x, t) u_{,xxxx} h^2 + ... \tag{1.120}$$

a schéma je tedy prvního řádu přesnosti v čase a druhého v prostoru. Konvergence schématu (1.119) se dokáže opět analogicky, jako pro případ $b = 1$, pouze podmínka stability má v tomto případě tvar

$$\frac{\tau}{h^2} b(x,t) \leq \frac{1}{2}. \tag{1.121}$$

Odhad chyby aproximace potom je

$$|e_j^n| = |U_j^n - u(x_j, t_n)| \leq T \left(\frac{M_1}{2}\tau + \frac{BM_2}{12} h^2 \right), \tag{1.122}$$

kde konstanta $B > b(x,t), \quad \forall (x,t) \in [0,1] \times [0,T]$ a konstanty M_1, M_2 mají stejný význam, jako při vyšetřování explicitního schématu (1.58).

Pro úlohu (1.118) lze definovat i θ-schéma, avšak existuje více možností, jak to provést. Jedna z možností je

$$U_j^{n+1} - U_j^n = \frac{\tau}{h^2} b^* [\theta \delta_x^2 U_j^{n+1} + (1-\theta)\delta_x^2 U_j^n], \tag{1.123}$$

kde b^* je vhodná hodnota b. Je možné položit například $b^* = b_j^{n+\frac{1}{2}}$. V tom případě je potom možné dokázat chybu diskretizace stejným způsobem, jako u θ-schématu pro $b = 1$, avšak výsledek bude přenásoben faktorem b. Důkaz konvergence by při použití diskrétního principu maxima probíhal také stejným způsobem, jen podmínka stability by nabývala tvaru

$$\frac{\tau}{h^2}(1-\theta) b(x,t) \leq \frac{1}{2}. \tag{1.124}$$

Místo $b^* = b_j^{n+\frac{1}{2}}$ je však možné volit například $b^* = \frac{1}{2}(b_j^{n+1} + b_j^n)$, což vede ke stejnému řádu konvergence.

Nejobecnější tvar lineární parabolické rovnice 2. řádu je

$$u_{,t} = b u_{,xx} - a u_{,x} + cu + d, \quad \forall t > 0, \;\; x \in (0,1), \tag{1.125}$$

kde $a = a(x,t)$, $b = b(x,t)$, $c = c(x,t)$, $d = d(x,t)$ jsou dané funkce, přičemž $b > 0$. Rovnicí (1.125) můžeme popisovat rozložení (například koncentraci) u zkoumané látky v nějakém médiu. Jednotlivé členy (1.125) pak odpovídají různým fyzikálním procesům šíření této látky v rámci média. Interpretace těchto členů může být následující. Člen $bu_{,xx}$ představuje difúzi, která popisuje, jak se zkoumaná látka rozptyluje v médiu. Funkce a je tzv. konvektivní rychlost, neboli rychlost, s jakou médium proudí a tím unáší zkoumanou látku. Člen $au_{,x}$ odpovídá konvekci, kdy se rozložení zkoumané látky u mění s proudem média, ve kterém se vyskytuje. Člen cu odpovídá tzv. reakci. Tento člen se může vyskytovat v například v případech, kdy se zkoumaná látka vytváří v důsledku chemické reakce. Konečně člen $d(x,t)$ představuje vnitřní zdroj (například tepla). Rovnice (1.125) v tomto nejobecnějším tvaru se nazývá nestacionární rovnice konvekce-difúze-reakce. Její úplná analýza je náročnější a přesahuje rámec tohoto textu. Podíváme se pouze na některé základní výsledky.

Nejjednodušší způsob, jak diskretizovat rovnici (1.125), je použít explicitní schéma. To je přirozené uvažovat ve tvaru

$$\frac{U_j^{n+1}-U_j^n}{\tau} = b_j^n \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2} - a_j^n \frac{U_{j+1}^n - U_{j-1}^n}{2h} + c_j^n U_j^n + d_j^n, \tag{1.126}$$

kde jsme použili na všechny prostorové derivace centrální diference a příslušná chyba diskretizace je tedy $\varepsilon_j^n = \mathcal{O}(\tau + h^2)$. Označíme-li nyní $\mu_j^n = \frac{\tau}{h^2} b_j^n$ a $\nu_j^n = \frac{\tau}{h} a_j^n$, pak můžeme schéma (1.126) přepsat do tvaru

$$U_j^{n+1} = U_j^n + \mu_j^n (U_{j+1}^n - 2U_j^n + U_{j-1}^n) - \frac{\nu_j^n}{2}(U_{j+1}^n - U_{j-1}^n) + \tau c_j^n U_j^n + \tau d_j^n, \tag{1.127}$$

který by se hodil pro výpočet. Pro vyšetření chyby aproximace bychom mohli použít diskrétního principu maxima. Dosazením $e_j^n = U_j^n - u(x_j, t_n)$ do (1.127) nejprve získáme

$$\begin{aligned} e_j^{n+1} &= e_j^n + \mu_j^n (e_{j+1}^n - 2e_j^n + e_{j-1}^n) - \frac{\nu_j^n}{2}(e_{j+1}^n - e_{j-1}^n) + \tau c_j^n e_j^n - \tau \varepsilon_j^n, \\ &= (1 - 2\mu_j^n + \tau c_j^n) e_j^n + \left(\mu_j^n - \frac{1}{2}\nu_j^n\right) e_{j+1}^n + \left(\mu_j^n + \frac{1}{2}\nu_j^n\right) e_{j-1}^n - \tau \varepsilon_j^n. \end{aligned} \tag{1.128}$$

Abychom nyní mohli použít diskrétního principu maxima (a získat tak omezenost chyby aproximace a tím konvergenci), museli bychom zajistit, aby koeficienty na pravé straně rovnice (1.128) jsou nezáporné a že jejich součet není větší, než 1 (jelikož koeficient na levé straně je roven 1). K tomu je třeba, aby platilo

$$\frac{1}{2}|n u_j^n| \le \mu_j^n, \quad 2\mu_j^n - \tau c_j^n \le 1, \quad c_j^n \le 0. \tag{1.129}$$

Speciálně tedy musí být (viz první podmínka (1.129))

$$h\left(\frac{|a_j^n|}{2b_j^n}\right) \le 1, \tag{1.130}$$

což prostřednictvím druhé podmínky dává i omezení na τ, neboť

$$\tau \le \frac{h^2}{2b_j^n - h^2 c_j^n}. \tag{1.131}$$

Jelikož v mnoha případech praktických problémů bývá koeficient u konvektivního členu výrazně větší, než u členu difúzivního, tj. $\dfrac{|a_j^n|}{b_j^n} << 1$, vyžaduje (1.130) velmi malé prostorové a (1.131) pak i velmi malé časové kroky. To je však velmi nevýhodné. Lze to však snadno napravit vhodnější volbou diferenční náhrady prostorové derivace u konvektivního členu $a u_{,x}$. Místo centrální diference použité v (1.126), použijeme zpětnou/dopřednou diferenci v závislosti na znaménku konvektivní rychlosti a.

$$u_{,x}(x_j, t_n) \approx \begin{cases} \frac{U_j^n - U_{j-1}^n}{h}, & \text{je-li } a(x_j, t_n) \ge 0, \\ \frac{U_{j+1}^n - U_j^n}{h}, & \text{je-li } a(x_j, t_n) < 0. \end{cases} \tag{1.132}$$

Vzhledem k výše uvedenému fyzikálnímu významu funkce $a(x,t)$ jakožto rychlosti proudění média, které unáší zkoumanou látku u (přesněji řečeno její koncentraci), má tato nově zavedená diskretizace jednoduchý význam. K diskretizaci $u_{,x}(x_j,t_n)$ využíváme (v závislosti na znaménku a) informaci o hodnotách u z té strany, odkud se do bodu x_j v čase t_n látka pohybuje. V této souvislosti hovoříme o diskretizaci typu "upwind" (proti větru). Předpokládejme nyní pro jednoduchost, že $a(x,t) > 0$ a $c(x,t) = 0$. Potom má explicitní schéma s "upwind" diskretizací tvar

$$\frac{U_j^{n+1} - U_j^n}{\tau} = b_j^n \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2} - a_j^n \frac{U_j^n - U_{j-1}^n}{h} + d_j^n, \tag{1.133}$$

což dává chybu aproximace ve tvaru

$$\begin{aligned} e_j^{n+1} &= e_j^n + \mu_j^n (e_{j+1}^n - 2e_j^n + e_{j-1}^n) - \nu_j^n (e_j^n - e_{j-1}^n) - \tau \varepsilon_j^n, \\ &= (1 - 2\mu_j^n - \nu_j^n) e_j^n + \mu_j^n e_{j+1}^n + \left(\mu_j^n + \nu_j^n\right) e_{j-1}^n - \tau \varepsilon_j^n. \end{aligned} \tag{1.134}$$

Aby všechny koeficienty na pravé straně byly nezáporné, potřebujeme pouze

$$2\mu_j^n + \nu_j^n \leq 1. \tag{1.135}$$

Stabilita tedy nevyžaduje žádné omezení na prostorový krok h. Cenou za to však je, že je schéma pouze $\varepsilon_j^n = \mathcal{O}(\tau + h)$, neboť jsme použili zpětnou diferenci.

1.3 Eliptické problémy - stacionární vedení tepla ve 2D

V této kapitole se zaměříme na jiný typ PDR, a sice na lineární eliptické PDR 2. řádu. Pro více informací o klasifikaci PDR odkazujeme čtenáře například na Dodatek (2.8). Budeme uvažovat následující úlohu.

$$-\Delta u = f, \quad x \in \Omega \equiv [0,1]^2, \tag{1.136}$$

$$u = 0, \quad x \in \partial\Omega. \tag{1.137}$$

Připomínáme, že symbol Δ představuje Laplaceův operátor, takže první rovnici v můžeme přepsat jako

$$-\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f.$$

Úloha zkoumaná v této kapitole tedy není závislá na čase. Rovnice (1.136) popisuje celou řadu problémů. Za všechny jmenujme například ustálené vedení tepla ve dvou dimenzích, nebo rovinnou úlohu pružnosti. S oběma těmito problémy se čtenáři již setkali například v předmětu NAK 1. [14]

Pokud chceme úlohu (1.136) řešit pomocí MKD, je stejně jako v případě jedné prostorové proměnné nutné nejprve diskretizovat oblast, na které tuto rovnici řešíme. Jinými slovy je třeba definovat příslušnou síť, na které potom zadefinujeme diferenční operátory, kterými aproximujeme operátory diferenciální. Pro jednoduchost úlohu řešíme na čtvercové oblasti $[0,1]^2$, v rámci které vytvoříme čtvercovou síť s J intervaly v každém směru. Tím vzniknou uzly $(x_r, y_s) \equiv (rh, sh),\ r,s \in \{0,1,...,J\},\ h = \frac{1}{J}$.[15] Označme nyní

$$J_\Omega = \{(x_r, y_s) \equiv (rh, sh),\ r,s \in \{1,...,J-1\} \tag{1.138}$$

množinu vnitřních uzlů a

$$J_{\partial\Omega} = \{(x_r, 0),\ (x_r, 1),\ (0, y_s),\ (1, y_s),\ r,s \in \{0,...,J\} \tag{1.139}$$

množinu hraničních uzlů. Druhé derivace v každé prostorové proměnné nyní můžeme nahradit centrálními diferencemi druhého řádu. Označíme-li podobně jako dříve hodnotu přibližného řešení $U_{r,s} \approx u(x_r, y_s)$, dostáváme tak

$$-\frac{\delta_x^2 U_{r,s}}{h^2} - \frac{\delta_y^2 U_{r,s}}{h^2} = f_{r,s}, \quad r,s = 1,...,J-1, \tag{1.140}$$

$$U_{r,0} = U_{r,J} = U_{0,s} = U_{J,s} = 0, \quad r,s = 0,...,J, \tag{1.141}$$

[14] V obou případech bychom však museli rovnici modifikovat tak, aby měla fyzikální význam. V obou případech by bylo možné rovnici přepsat do tvaru $-\mathrm{div}\,(\mathbb{A}\nabla u) = f$, kde $\mathbb{A}$ by v případě vedení tepla byla matice difuzivity materiálu, v případě rovinné úlohy pružnosti pak matice tuhosti materiálu.

[15] Obecně by bylo možné definovat prostorový krok jiné délky ve směrech x a y. Tím pádem bychom měli různý počet uzlů J_x a J_y a příslušné prostorové kroky $h_x = \frac{1}{J_x}$ a $h_y = \frac{1}{J_y}$. Pro snížení počtu indexů však uvažujeme stejnou délku kroku v obou směrech.

což lze přepsat jako

$$\frac{4U_{r,s} - U_{r+1,s} - U_{r-1,s} - U_{r,s+1} - U_{r,s-1}}{h^2} = f_{r,s}, \quad r,s = 1,...,J-1, \tag{1.142}$$

$$U_{r,0} = U_{r,J} = U_{0,s} = U_{J,s} = 0, \quad r,s = 0,...,J. \tag{1.143}$$

Máme tak k dispozici $(J-1)^2$ rovnic. Označme dále diskrétní Laplaceův operátor

$$(L_h U)_{r,s} \equiv L_h U_{r,s} = \frac{4U_{r,s} - U_{r+1,s} - U_{r-1,s} - U_{r,s+1} - U_{r,s-1}}{h^2}. \tag{1.144}$$

Později ukážeme, že tento oprerátor splňuje princip maxima

$$(L_h U_P \leq 0, \ \ \forall P \in J_\Omega) \Rightarrow \max_{P \in J_\Omega} U_P \leq \max\{0, \max_{Q \in J_{\partial\Omega}} U_Q\}, \tag{1.145}$$

tj. že se maximálních (a nezáporných) hodnot U_P nabývá na hranici oblasti Ω. Podobně jako u parabolických úloh (a také analogicky postupu, který jsme použili u jednoduchého příkladu v úvodní kapitole) můžeme principu maxima využít k odhadu chyby aproximace. Definujme proto nejprve chybu diskretizace jako

$$\varepsilon_{r,s} = L_h u(x_r, y_s) - f_{r,s}. \tag{1.146}$$

Použitím Taylorova rozvoje dostaneme analogicky, jako u parabolických úloh

$$|\varepsilon_{r,s}| \leq \frac{1}{12}(M_1 + M_2)h^2, \tag{1.147}$$

kde $M_1 = \max_{\bar{\Omega}} \left|\frac{\partial^4 u}{\partial x^4}\right|$, $M_2 = \max_{\bar{\Omega}} \left|\frac{\partial^4 u}{\partial y^4}\right|$. Chyba aproximace $e_{r,s} = U_{r,s} - u(x_r, y_s)$ tedy splňuje

$$L_h e_{r,s} = -\varepsilon_{r,s}, \quad r,s = 1,...,J-1 \tag{1.148}$$

$$e_{r,0} = e_{r,J} = e_{0,s} = e_{J,s} = 0, \quad r,s = 0,...,J. \tag{1.149}$$

To můžeme snadno zapsat ve tvaru

$$L_h e_P = -\varepsilon_P, \ \ \forall P \in J_\Omega, \quad e_P = 0, \ \ \forall P \in J_{\partial\Omega}. \tag{1.150}$$

Podobně jako v příkladu v úvodu definujme k odvození odhadu chyby aproximace vhodnou srovnávací funkci

$$\Phi(x,y) = \left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2. \tag{1.151}$$

Tato funkce není určena jednoznačně a obecně není její nalezení snadné. Musí být volena tak, aby byla všude nezáporná a zároveň chceme, aby na hranici nabývala co nejmenších hodnot. To následně ovlivní velikost konstanty u odhadu chyby aproximace. V našem případě má srovnávací funkce Φ nulové čtvrté derivace, takže chyba diskretizace (1.147) vychází rovna nule. Tím pádem se v případě funkce Φ nedopouštíme žádné chyby, když Laplaceův operátor nahradíme diskrétním Laplaceovým operátorem (1.144). Vychází proto

$$L_h \Phi_P = (-\Delta\Phi)_P = -4, \quad \forall P \in J_\Omega. \tag{1.152}$$

Označme nyní pomocnou funkci Ψ v bodě P jako

$$\Psi_P = e_P + \frac{1}{4}\frac{h^2}{12}(M_1 + M_2)\Phi_P. \tag{1.153}$$

Tato síťová funkce je zkonstruována velmi důmyslně jako kombinace Φ a chyby diskretizace ε tak, aby ve všech bodech byly splněny předpoklady diskrétního principu maxima. Po dosazení do diskrétního Laplaceova operátoru (1.144) totiž vychází

$$L_h\Psi_P = L_h e_P - \frac{h^2}{12}(M_1 + M_2) = -\varepsilon_P - \frac{h^2}{12}(M_1 + M_2) \leq 0, \quad \forall P \in J_\Omega. \tag{1.154}$$

Nyní je vidět důvod, proč byla funkce Ψ zvolena právě takto. Díky odhadu (1.147) máme v (1.154) zaručenu poslední nerovnost. Je tedy splněna podmínka pro platnost diskrétního principu maxima a podle (1.145) platí

$$\Psi_P \leq \frac{1}{4}\frac{h^2}{12}(M_1 + M_2)\max_{Q\in J_{\partial\Omega}} \Phi_Q = \frac{1}{8}\frac{h^2}{12}(M_1 + M_2), \quad \forall P \in J_\Omega. \tag{1.155}$$

Jelikož dle definice Ψ platí $e_P \leq \Psi_P$, dostáváme

$$U_P - u_P \leq \frac{1}{96}(M_1 + M_2)h^2. \tag{1.156}$$

Označíme-li nyní $\Psi_P = e_P - \frac{1}{4}\frac{h^2}{12}(M_1 + M_2)\Phi_P$, získáme obdobný odhad pro $-(U_P - u_P)$. Celkem tedy platí (po zpětném přeznačení)

$$|U_{r,s} - u(x_r, y_s)| \leq \frac{1}{96}(M_1 + M_2)h^2, \tag{1.157}$$

což je odhad, který jsme hledali. Vidíme, že přibližné řešení kvadraticky konverguje k přesnému řešení.

1.3.1 Obecnější lineární eliptická rovnice

Podívejme se nyní na následující úlohu, která je zobecněním úlohy z předchozího odstavce.

$$-\mathrm{div}(a\nabla u) = f, \quad (x, y) \in \Omega \tag{1.158}$$

$$\alpha_0 u + \alpha_1 \frac{\partial u}{\partial n} = g, \quad (x, y) \in \partial\Omega, \tag{1.159}$$

kde Ω je omezená oblast v $\mathbb{R}^2$, n je jednotkový vektor vnější normály k hranici Ω, $a = a(x, y) \geq a_0 > 0$ je hladká funkce a α_0, α_1 jsou konstanty splňující $\alpha_0 \geq 0,\ \alpha_1 \geq 0,\ \alpha_0+\alpha_1 > 0$. Nejprve vysvětlíme význam použitých symbolů. Operátor divergence vektoru je definován jako součet derivací tohoto vektoru podle všech proměnných. Jelikož gradient ∇u funkce u je vektor o složkách $\frac{\partial u}{\partial x}$ a $\frac{\partial u}{\partial y}$, vychází po aplikaci operátoru divergence

$$\mathrm{div}(a\nabla u) = \frac{\partial}{\partial x}\left(a(x, y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(a(x, y)\frac{\partial u}{\partial y}\right). \tag{1.160}$$

Okrajová podmínka (1.159) má reálný fyzikální význam. Kombinuje Dirichletovu a Neumannovu okrajovou podmínku, kterou čtenáři znají již z předmětu NAK 1. Například pro $\alpha_1 = 0$ dostáváme $\alpha_0 u = g$, což je klasická Dirichletova podmínka, ve které předepisujeme na hranici hodnotu hledané funkce. Naopak při $\alpha_0 = 0$ dostáváme $\alpha_1 \frac{\partial u}{\partial n} = \alpha_1 \nabla u \cdot n = g$, což je Neumannova okrajová podmínka, kde předepisujeme tok jisté veličiny. Studenti znají z předmětu NAK1 úlohu dvourozměrného ustáleného vedení tepla. V ní by funkce a by odpovídala difuzivitě, u by byla funkce teploty a Neumannova okrajová podmínka by předepisovala tepelný tok na hranici ve tvaru $q \cdot n = \bar{q}$, kde $\bar{q}$ je teplo, které opouští oblast v daném bodě hranice. Z Fourierova zákona víme, že $q = -a\nabla u$ a tedy $\bar{q} = -a(\nabla u) \cdot n = -a\frac{\partial u}{\partial n}$, což je po přeznačení podmínka totožná s naší okrajovou podmínkou (1.159).

Označíme-li $b = -a_{,x}$ a $c = -a_{,y}$, můžeme rovnici (1.158) zapsat ve tvaru

$$-a\Delta u + bu_{,x} + cu_{,y} = f. \tag{1.161}$$

Oblast Ω diskretizujme pravidelnou obdélníkovou sítí s krokem h_x ve směru x a h_y ve směru y. V uzlech nesousedících s hranicí lze rovnici (1.161) diskretizovat použitím centrálních diferencí, čímž získáme

$$-a_{r,s}\left[\frac{\delta_x^2 U_{r,s}}{h_x^2} + \frac{\delta_y^2 U_{r,s}}{h_y^2}\right] + b_{r,s}\frac{\Delta_{0x} U_{r,s}}{h_x} + c_{r,s}\frac{\Delta_{0y} U_{r,s}}{h_y} = f_{r,s}. \tag{1.162}$$

Snadno se ukáže, že chyba diskretizace je druhého řádu v h_x a h_y (používáme všude centrální diference). Aby platil diskrétní princip maxima, musely by být koeficienty v bodech sousedících s bodem o souřadnicích (r, s) nekladné. Rozepsáním (1.162) však zjistíme, že koeficient u $U_{r-1,s}$ je $-\frac{1}{2h_x^2}[2a_{r,s} + b_{r,s}h_x]$ a koeficient u $U_{r+1,s}$ je $-\frac{1}{2h_x^2}[2a_{r,s} - b_{r,s}h_x]$. Muselo by tedy platit $|b_{r,s}|h_x \leq 2a_{r,s}$ a podobně pro koeficienty u $U_{r,s+1}$ a $U_{r,s-1}$ by muselo být $|c_{r,s}|h_y \leq 2a_{r,s}$. V místech, kde je a malé, ale rychle se mění (tzn. její první derivace jsou velké) bychom proto museli volit velmi jemnou síť, což není výhodné.

Je proto vhodnější diskretizovat přímo rovnici (1.158), což lze provést následovně

$$-\left[\frac{\delta_x(a\delta_x U)}{h_x^2} + \frac{\delta_y(a\delta_y U)}{h_y^2}\right]_{r,s} = f_{r,s}, \tag{1.163}$$

neboli rozepsáním

$$-\frac{1}{h_x^2}\left[a_{r+\frac{1}{2},s}(U_{r+1,s} - U_{r,s}) - a_{r-\frac{1}{2},s}(U_{r,s} - U_{r-1,s})\right] + \tag{1.164}$$

$$-\frac{1}{h_y^2}\left[a_{r,s+\frac{1}{2}}(U_{r,s+1} - U_{r,s}) - a_{r,s-\frac{1}{2}}(U_{r,s} - U_{r,s-1})\right] = f_{r,s}. \tag{1.165}$$

Nyní vidíme, že koeficienty mají všechny záporné znaménko a tedy princip maxima platí bez jakéhokoliv omezení na síť. Situace je problematičtější, pokud by oblast Ω byla zakřivená. Těmto otázkám se však v našem textu nebudeme věnovat. Zbývá ještě dokázat diskrétní princip maxima, jehož platnost jsme dosud předpokládali. Tomu bude věnována poslední kapitola v rámci eliptických rovnic.

1.3.2 Chyba aproximace a diskrétní princip maxima

V této kapitole dokážeme platnost diskrétního principu maxima pro operátor získaný diskretizací lineární eliptické úlohy. Předpokládejme tedy, že diskretizací úlohy (1.158) - (1.159) získali v každém bodě $P \in J_\Omega$ síťovou rovnici

$$L_h U_P = f_P + g_P, \tag{1.166}$$

kde g_P jsou hodnoty vzniklé Neumanovými okrajovými podmínkami, které jsme zavedli v předchozí kapitole. Nadále budeme předpokládat, že v množině uzlů J_Ω mohou být i hraniční uzly, ve kterých je předepsána Neumannova okrajová podmínka (dosud jsme v ní uvažovali pouze vnitřní uzly). V množině $J_{\partial\Omega}$ tedy budou pouze ty hraniční uzly, ve kterých je předepsána Dirichletova okrajová podmínka. Jinými slovy, v množině J_Ω jsou ty uzly, pro které máme nějakou rovnici. Zaveďme navíc následující předpoklady.

- Pro každé $P \in J_\Omega$ definujme množinu $\mathcal{N}_P$ uzlů, které jsou hraniční k uzlu P (mohou ležet i v $J_{\partial\Omega}$). Přitom požadujme

$$L_h U_P = c_P U_P - \sum_{Q \in \mathcal{N}_P} c_{PQ} U_Q, \quad \text{kde } c_{PQ} > 0, \;\; c_P \geq \sum_{Q \in \mathcal{N}_P} c_{PQ}. \tag{1.167}$$

- Množina J_Ω je souvislá v tom smyslu, že

$$\forall P, Q \in J_\Omega \; \exists P_1, ..., P_m \in J_\Omega \text{ tak, že } \;\; P_1 \in \mathcal{N}_P, \; P_2 \in \mathcal{N}_{P_1}, ..., \; P_m \in \mathcal{N}_{P_{m-1}}, \; Q \in \mathcal{N}_{P_m}. \tag{1.168}$$

- Na části hranice musí být předepsány okrajové Dirichletovy podmínky, neboli

$$\exists P \in J_\Omega, \; Q \in J_{\partial\Omega} \text{ tak, že } \; Q \in \mathcal{N}_P. \tag{1.169}$$

První z předpokladů je podmínka na koeficienty, kterou jsme při aplikaci diskrétního principu maxima vždy požadovali, tj. součet koeficientů vpravo musel být menší, nebo roven koeficientu vlevo. Druhý předpoklad požaduje, aby pro každé dva body uvnitř oblasti Ω existovala posloupnost navzájem sousedních bodů, kterými se od jednoho bodu dostaneme ke druhému. Třetí pak požaduje existenci bodu uvnitř oblasti Ω, který má souseda na hranici s Dirichletovou okrajovou podmínkou. Při platnosti právě uvedených předpokladů platí následující věta.

Věta 1.5 (Diskrétní princip maxima)**.**
Mějme operátor definovaný v (1.166). Pokud pro všechna $P \in J_\Omega$ je $L_h U_P \leq 0$, pak platí

$$\max_{P \in J_\Omega} U_P \leq \max\{0, \max_{q \in J_{\partial\Omega}} U_Q\}. \tag{1.170}$$

Důkaz
Označme maximální hodnotu síťové funkce U uvnitř oblasti Ω, a na její hranici, jako $M_\Omega = \max_{P \in J_\Omega} U_P, \;\; M_{\partial\Omega} = \max_{Q \in J_{\partial\Omega}} U_Q$. Je-li $M_\Omega \leq 0$, pak tvrzení platí. Nechť tedy $M_\Omega \geq 0$ a

předpokládejme v rozporu s tvrzením věty, že $M_\Omega > M_{\partial\Omega}$ [16]. Označme jako $P^* \in J_\Omega$ ten bod, pro který $M_\Omega = U_{P^*}$. Potom dle předpokladu věty platí

$$M_\Omega = U_{P^*} \leq \frac{1}{c_{P^*}} \sum_{Q \in \mathcal{N}_{P^*}} c_{P^*Q} U_Q \leq M_\Omega. \tag{1.171}$$

To je vzhledem k předpokladu (1.167) možné jedině tehdy, pokud pro všechna $Q \in \mathcal{N}_{P^*}$ platí $U_Q = M_\Omega$ a všude v (1.171) jsou rovnosti. Ze souvislosti J_Ω (předpoklad (1.168)) plyne, že $U_P = M_\Omega$ ve všech bodech J_Ω. Protože však díky předpokladu (1.169) alespoň jeden bod z J_Ω sousedí s nějakým $Q \in J_{\partial\Omega}$, musí platit $M_\Omega = M_{\partial\Omega}$, což je spor s $M_\Omega > M_{\partial\Omega}$ a důkaz je hotov.

Jako jednoduchý důsledek dostáváme, že

$$L_h U_P \geq 0 \ \ \forall P \in J_\Omega \ \ \Rightarrow \ \ \min_{P \in J_\Omega} U_P \geq \min\{0, \min_{Q \in J_{\partial\Omega}} U_Q\}. \tag{1.172}$$

a

$$L_h U_P = 0 \ \ \forall P \in J_\Omega \ \ \Rightarrow \ \ \max_{P \in J_\Omega} |U_P| \leq \max_{Q \in J_{\partial\Omega}} |U_Q|. \tag{1.173}$$

To dokážeme snadno. Z $L_h U_P \geq 0 \ \ \forall P \in J_\Omega$ totiž plyne $L_h(-U_P) \leq 0 \ \ \forall P \in J_\Omega$ a proto dle dokázaného principu maxima platí

$$\min_{P \in J_\Omega} U_P = -\max_{P \in J_\Omega}(-U_P) \geq -\max\{0, \max_{Q \in J_{\partial\Omega}}(-U_Q)\} = \min\{0, \min_{Q \in J_{\partial\Omega}} U_Q\}, \tag{1.174}$$

a první důsledek je dokázán. Pokud platí $L_h U_P = 0 \ \ \forall P \in J_\Omega$, pak platí zároveň oba právě dokázané nerovnosti

$$\max_{P \in J_\Omega} U_P \leq \max_{Q \in J_{\partial\Omega}} |U_Q| \quad \text{a zároveň} \quad \min_{P \in J_\Omega} U_P \geq -\max_{Q \in J_{\partial\Omega}} |U_Q|, \tag{1.175}$$

z čehož dostáváme pro všechna $P \in J_\Omega$ nerovnost $|U_P| \leq \max_{Q \in J_{\partial\Omega}} |U_Q|$. Pokud platí pro všechny body z J_Ω, platí i pro maximum a druhý důsledek je dokázán.

Dosavadní dílčí výsledky nám nyní umožňují dokázat následující tvrzení.

Věta 1.6.
Úloha (1.166) má právě jedno řešení.

Důkaz
Úloha (1.166) představuje soustavu lineárních algebraických rovnic se čtvercovou maticí. Z lineární algebry je známo, že existence jednoznačného řešení takové soustavy rovnic je ekvivalentní požadavku, aby příslušná homogenní soustava rovnic měla pouze triviální řešení. Příslušnou homogenní soustavu rovnic $L_h U_P = 0$ dostaneme, pokud pro všechna $P \in J_\Omega$ bude $f_P = g_P = 0$. Potom však $\max_{Q \in J_{\partial\Omega}} |U_Q| = 0$ pro všechna $Q \in J_{\partial\Omega}$ a ze druhého důsledku diskrétního principu maxima dokázaného výše plyne $U_P = 0$ pro všechna $P \in J_\Omega$, což jsme chtěli dokázat.

[16] Přesně řečeno předpokládáme, že platí první část implikace a zároveň negace druhé části. Z výrokové logiky je známo, že negací implikace $A \Rightarrow B$, je tvrzení $A \wedge \neg B$.

V následujícím odstavci předvedeme pro úlohu (1.166) zcela obecně důkaz, který jsme použili v příkladu v úvodu k MKD a také v této kapitole pro dokázání konvergence. Definujme chybu diskretizace a chybu aproximace analogickým způsobem, jako doposud

$$\varepsilon_P = L_h u_P - f_P - g_P, \quad e_P = U_P - u_P, \tag{1.176}$$

kde $u_P = u(P)$. Kombinací obou vztahů vychází

$$L_h e_P = -\varepsilon_P, \quad \forall P \in J_\Omega. \tag{1.177}$$

Předpokládejme, že Dirichletovy okrajové podmínky jsou předepsány přesně, tj. že platí $e_P = 0, \ \forall P \in J_{\partial\Omega}$. Potom platí následující věta.

Věta 1.7. *Nechť Φ je nezáporná síťová funkce definovaná na $J_\Omega \cup J_{\partial\Omega}$, splňující*

$$L_h \Phi_P \leq -1, \quad \forall P \in J_\Omega. \tag{1.178}$$

Potom platí

$$\max_{P \in J_\Omega} |e_P| \leq \left(\max_{Q \in J_{\partial\Omega}} \Phi_Q \right) \left(\max_{P \in J_\Omega} |\varepsilon_P| \right). \tag{1.179}$$

Důkaz
Položme $D = \max_{P \in J_\Omega} |\varepsilon_P|$. Pak platí $L_h(D\Phi_P + e_P) = DL_h\Phi_P - \varepsilon_P \leq -D - \varepsilon_P \leq 0$, a tudíž podle diskrétního principu maxima platí

$$e_P \leq D\Phi_P + e_P \leq \max\{0, \max_{Q \in J_{\partial\Omega}} (D\Phi_Q + e_Q)\} = \max_{Q \in J_{\partial\Omega}} (\Phi_Q) D. \tag{1.180}$$

Aplikujeme-li diskrétní princip maxima na $D\Phi_P - e_P$, dostaneme analogický odhad pro $-e_P$ a důkaz je hotov.

Zatímco odhad chyby diskretizace je poměrně snadný, nalezení srovnávací funkce Φ nemusí být vždy jednoduché. Tato funkce není určena jednoznačně a cílem je nalézt takové Φ, aby $\max_{Q \in J_{\partial\Omega}}(\Phi_Q)$ bylo co nejmenší. Zároveň tento obecný postup vysvětluje možná trochu tajemnou volbu srovnávací funkce Φ v konkrétních příkladech uvedených dříve. Z uvedené věty tedy pro úlohu (1.166) plyne odhad

$$|U_P - u_P| \leq Ch^2. \tag{1.181}$$

Poznamenejme, že pokud bychom měli zakřivenou hranici, platil by tento odhad pouze ve vnitřních (tzv. regulárních) uzlech sítě, zatímco v uzlech ležících u hranice bychom dostali pouze $|U_P - u_P| \leq Ch$. To plyne z faktu, že chyba diskretizace je v těchto uzlech obecně jen prvního řádu přesnosti. Podrobnější analýzou problému u zakřivené hranice se zde však nebudeme zabývat.

1.4 Hyperbolické problémy

Obsahem této kapitoly je vyšetřování numerických schémat pro hyperbolické rovnice. S tímto typem rovnic se studenti již setkali v předmětu Dynamika stavebních konstrukcí. Rovnice podélně kmitajícího prutu (2.246) je typickým příkladem hyperbolické rovnice. Více informací o členění PDR najde čtenář v Dodatku 2.8. V tomto textu se budeme zabývat dvěma typy hyperbolických rovnic. Prvním z nich je transportní rovnice, která má význam například při popisu fyzikálních zákonů zachování. Druhou rovnicí, kterou se zde budeme zabývat, je vlnová rovnice, která jako speciální případ zahrnuje problémy probírané v Dynamice stavebních konstrukcí.

1.4.1 Transportní rovnice

Transportní rovnice je nejjednodušší parciální diferenciální rovnice, ve které vstupují pouze první derivace. Tato rovnice má tvar

$$u_{,t} + a u_{,x} = 0, \quad x \in \mathbb{R}, \;\; t > 0, \tag{1.182}$$

kde funkce a může obecně záviset i na hledané funkci, tj. $a = a(u)$. V tom případě je rovnice (1.182) nelineární. K rovnici této je třeba definovat počáteční podmínku. Tu označíme následovně.

$$u(x, 0) = u^0(x). \tag{1.183}$$

Má-li úloha (1.182) hladké řešení, pak lze definovat tzv. charakteristiky $x = x(t)$ vztahem

$$x'(t) = a(u(x(t), t)), \tag{1.184}$$

kde čárkou značíme obyčejnou časovou derivaci. Použitím pravidla o derivování složené funkce dostaneme

$$\frac{d}{dt} u(x(t), t) = \left(u_{,t} + x'(t) u_{,x}\right)(x(t), t) = \left(u_{,t} + a(u) u_{,x}\right)(x(t), t) = 0, \tag{1.185}$$

neboť u je řešením (1.182). To znamená, že hledaná funkce u je podél těchto charakteristik konstantní a vzhledem k definici charakteristik (1.184) jsou tyto charakteristiky přímky. Charakteristika procházející v čase $t = 0$ bodem x_0 je tedy přímka se sklonem

$$\frac{dx}{dt} = a(u^0(x_0)). \tag{1.186}$$

Řešení u proto můžeme vyjádřit pomocí implicitního vztahu

$$u(x, t) = u^0(x - a(u(x, t), t)), \tag{1.187}$$

kde jsme využili jednoduché úvahy $x' = at \;\Rightarrow\; x = at + x_0$ a tedy $x_0 = x - at$. V obecném případě se však charakteristiky mohou protnout a vztah (1.187) tak platí pouze do okamžiku jejich protnutí.

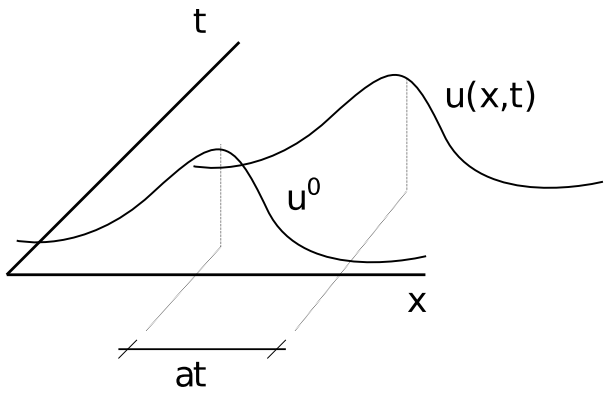
Uvažujme dále pouze lineární úlohu

$$u_{,t} + au_{,x} = 0, \quad x \in \mathbb{R}, \;\; t > 0, \quad u(x,0) = u^0(x), \;\; x \in \mathbb{R}, \tag{1.188}$$

kde funkce $a = a(x,t)$ už nezávisí na u. Přesto, že tato rovnice představuje nejjednodušší PDR, její numerické řešení není zdaleka triviální, jak uvidíme na v následujících odstavcích. Charakteristiky jsou v tomto případě křivky, splňující $x'(t) = a(x(t),t)$. Je-li funkce a lipschitzovsky spojitá[17] v x a spojitá v t, pak se charakteristiky neprotnou. Řešení u je podél charakteristik opět konstantní. Sestrojením těchto charakteristik tedy získáme řešení, neboť $u(x(t),t) = u^0(x(0))$. Je-li dokonce a konstantní, pak jsou charakteristiky rovnoběžné přímky $x - at = const$ a řešení je dáno vztahem

$$u(x,t) = u^0(x - at). \tag{1.189}$$

Pro $a = const$ se rovnice (1.188) nazývá jednosměrná vlnová rovnice. Řešení u v čase t je tedy rovno počáteční podmínce u^0 posunuté o vzdálenost $|a|t$ (doprava pro $a > 0$, doleva pro $a < 0$). Parametr a se nazývá rychlost šíření vlny podél charakteristik. Řešení jednosměrné vlnové rovnice lze tedy považovat za vlnu šířící se rychlostí a beze změny tvaru, viz Obr 1.9. Nadále budeme předpokládat, že $a = const$.



Obr. 1.9: *Šíření vlny podél charakteristik.*

Podívejme se nyní blíže na řešení rovnice (1.188) metodou konečných diferencí. Opět budeme uvažovat konstantní prostorový krok h a časový krok τ. Řešení u úlohy (1.188) budeme aproximovat hodnotou $U_j^n \approx u(x_j, t_n)$, kde $x_j = jh$ a $t_n = n\tau$. Máme různé možnosti, jak diskretizovat časovou a prostorovou derivaci v (1.188).

Uvažujme nejprve nejjednodušší explicitní schéma, kde pro časovou i prostorovou derivaci použijeme dopřednou diferenci.

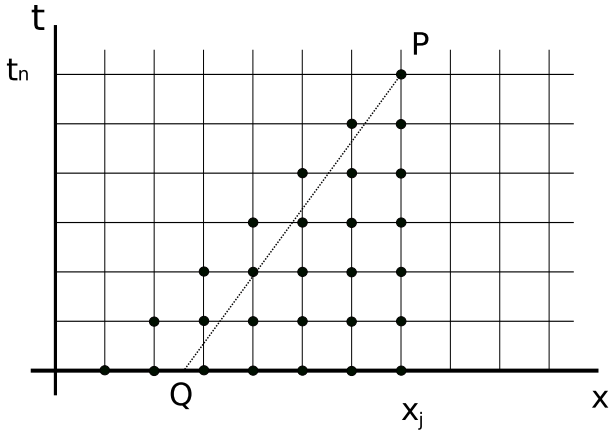
$$\frac{U_j^{n+1} - U_j^n}{\tau} + a\frac{U_j^n - U_{j-1}^n}{h}, \tag{1.190}$$

[17]Funkce f je lipschitzovsky spojitá, pokud existuje kladná konstanta M tak, že pro všechna x, y z definičního oboru platí $|f(x) - f(y)| \leq M|x - y|$. Snadno lze ukázat, že z lipschitzovské spojitosti plyne (dokonce absolutní) spojitost.

které můžeme zavedením koeficientu $\nu = \frac{a\tau}{h}$ přepsat do podoby vhodnější pro výpočet

$$U_j^{n+1} = \nu U_{j-1}^n + (1+\nu)U_j^n. \tag{1.191}$$

Na Obr. 1.10 vidíme, že hodnota U_j^{n+1} závisí na dvou hodnotách z předchozí časové hladiny, každá z těchto hodnot opět na na dvou hodnotách z časové hladiny t_{n-1} atd. Celkem tedy hodnota U_j^{n+1} závisí na hodnotách obsažených ve zobrazeném trojúhelníku vymezeném bodem (x_j, t_{n+1}) a hodnotami $x_{j-n-1}, x_{j-n}, ..., x_{j-1}, x_j$ v čase $t = 0$. Tento trojúhelník se nazývá oblast závislosti U_j^{n+1}, nebo bodu obecněji diferenčního schématu v (x_j, t_{n+1}). Víme, že pro řešení u platí podle (1.189) $u(P) = u^0(Q)$, kde Q je průsečík přímky $t = 0$ a charakteristiky $x - at = const$ procházející bodem P. Úsečka PQ je oblastí závislosti rovnice (1.188). Tyto oblasti závislosti původní rovnice a numerického schématu spolu musí souviset. Jaký vztah je mezi nimi třeba říká následující velice významná věta.



Obr. 1.10: *Ilustrace CFL podmínky. Oblast závislosti diferenčního schématu a dané PDR.*

Věta 1.8 (CFL podmínka (Courant, Friedrichs, Lewy (1928)))**.** *Nutná podmínka konvergence diferenčního schématu je, aby oblast závislosti PDR ležela uvnitř oblasti závislosti diferenčního schématu.*

Pro schéma (1.191) je CFL podmínka splněna, leží-li bod Q na přímce $t = 0$ mezi body x_{j-n-1} a x_j. To je právě tehdy, když $a \geq 0$ a zároveň $a\tau \leq h$. Uvažujme posloupnost sítí s $\frac{\tau}{h} = const$ a $h \to 0$ a nechť bod P je uzlem všech těchto sítí. Pak oblast závislosti pro všechny tyto sítě je stejná. Pokud $a > 0$ nebo $a\tau > h$, pak U_P nezávisí na hodnotách u^0 v okolí bodu Q a přibližná řešení obecně nemohou konvergovat k hodnotě u_P. Z uvedeného je zřejmé, že platí následující věta.

Věta 1.9. *Nutnou podmínkou konvergence explicitního schématu tvaru*

$$U_j^{n+1} = \alpha U_{j-1}^n + \beta U_j^n + \gamma U_{j-1}^n \tag{1.192}$$

pro rovnici (1.188) při $\frac{\tau}{h} = const$ je, aby platilo

$$\left|\frac{a\tau}{h}\right| \leq 1. \tag{1.193}$$

Někdy se CFL podmínka říká nerovnosti (1.193). Všimněme si, že pro schéma (1.192) značí poměr $\frac{h}{\tau}$ numerickou rychlost šíření vlny, neboť informace se během jednoho časového kroku rozšíří o jeden prostorový krok. Nerovnost (1.193) tedy říká, že numerická rychlost musí být větší, nebo rovná rychlosti šíření odpovídající PDR. Pokud diferenční schéma nemůže šířit informaci alespoň tak rychle, jako se šíří řešení příslušné PDR, nemůže přibližné řešení obdržené tímto schématem konvergovat k přesnému řešení. Na CFL podmínku se lze dívat jako nutnou podmínku stability diferenčního schématu. Obecně to sice není podmínka postačující, ale umožňuje s minimální námahou vyřadit ta schémata, která ji nesplňují. Teprve schémata, která CFL podmínku splňují má smysl vyšetřovat podrobněji použitím kritérií, která jsou pro stabilitu postačující.

Z dosavadních úvah plyne, že schéma (1.190) lze použít pouze v případě, že $a > 0$. Pokud je $a < 0$, neleží oblast závislosti rovnice (1.188) v oblasti závislosti schématu (1.190) a tedy není splněna CFL podmínka. Nerovnost (1.193) však napovídá, že v případě $a < 0$ bude fungovat analogické schéma, kde zvolíme dopřednou diferenci pro prostorovou derivaci. Celkově tedy dostáváme následující diskretizaci

$$\frac{U_j^{n+1} - U_j^n}{\tau} + \begin{cases} a\frac{U_j^n - U_{j-1}^n}{h}, & \text{pokud } a > 0, \\ a\frac{U_{j+1}^n - U_j^n}{h}, & \text{pokud } a < 0,. \end{cases} \tag{1.194}$$

Pokud by a nebylo konstantní, pak ve schématu (1.194) značí $a = a(x_j, t_n)$. Jedná se o diskretizaci typu "upwind", se kterou jsme se již setkali u schématu (1.133). Pokud platí $|a|\tau \leq h$, splňuje schéma (1.194) CFL podmínku.

Obraťme nyní pozornost k Fourierově analýze schématu (1.194). V případě hyperbolických rovnic je matematické pozadí Fourierovy analýzy poněkud složitější, neboť rovnici řešíme na celé reálně ose a jde tedy o Cauchyovu úlohu. [18] Na rozdíl od parabolických úloh, kde jsme k Fourierově analýze používali Fourierových řad (v analogii k řešení rovnice (1.35) - (1.37) Fourierovou metodou), zde je nutné se uchýlit k Fourierově transormaci.Formálně lze však postupovat stejně, jako v případě parabolických úloh, kde jsme přibližné řešení příslušného diferenčního schématu hledali ve tvaru $U_j^n = \mathrm{e}^{\mathrm{i}\xi jh}\lambda^n$. V případě hyperbolických úloh budeme řešení hledat formálně ve stejném tvaru

$$U_j^n = e^{i\xi jh}\lambda^n, \tag{1.195}$$

kde ovšem nyní $\xi \in \mathbb{R}$, nikoli jen $\xi \in \mathbb{Z}$, jako v parabolických rovnicích. Další postup je však analogický. Dosaďme tedy (1.195) do schématu (1.194). Pro $a > 0$ a při použití dříve zavedeného označení $\nu = \frac{a\tau}{h}$ dostáváme

$$\mathrm{e}^{\mathrm{i}\xi jh}\lambda^{n+1} = \mathrm{e}^{\mathrm{i}\xi jh}\lambda^n(1-\nu) + \nu\mathrm{e}^{\mathrm{i}\xi(j-1)h}\lambda^n, \tag{1.196}$$

z čehož po zkrácení $\mathrm{e}^{\mathrm{i}\xi jh}\lambda^n$ plyne

$$\lambda(\xi) = 1 - \nu + \nu\mathrm{e}^{-\mathrm{i}\xi h}. \tag{1.197}$$

[18] V případě eliptických úloh, kterým je věnována předchozí kapitola, řešíme příslušnou rovnici na dané omezené oblasti, na jejíž hranici je předepsána okrajová podmínka. Takové úloze říkáme Dirichletova úloha. V případě, že rovnici řešíme na neomezené oblasti, kde nejsou předepsány žádné okrajové podmínky, úloha se nazývá Cauchyova.

Pro zápornou rychlost šíření $a < 0$ dostaneme zcela analogicky

$$\lambda(\xi) = 1 + \nu - \nu \mathrm{e}^{\mathrm{i}\xi h}. \tag{1.198}$$

To lze zapsat jednotně jako

$$\lambda(\xi) = 1 - |\nu| + |\nu|\mathrm{e}^{\pm\mathrm{i}\xi h} = 1 - |\nu| + |\nu|\cos(\xi h) \pm \mathrm{i}|\nu|\sin(\xi h). \tag{1.199}$$

Vidíme, že $\lambda(\xi)$ je komplexní číslo. Velikost $|z|$ komplexního čísla $z = a + \mathrm{i}b$ je dána vztahem $|z|^2 = a^2 + b^2$. Velikost amplifikačního faktoru $\lambda(\xi)$ tedy můžeme spočíst následujícími upravami

$$|\lambda(\xi)|^2 = (1 - |\nu|)^2 + 2(1 - |\nu|)|\nu|\cos(\xi h) + \underbrace{|\nu|^2\cos^2(\xi h) + |\nu|^2\sin^2(\xi h)}_{|\nu|^2} \tag{1.200}$$

$$= 1 - 2|\nu| + 2|\nu|^2 + 2(1 - |\nu|)|\nu|\cos(\xi h) \tag{1.201}$$

$$= 1 - 2|\nu|(1 - |\nu|)\underbrace{[1 - \cos(\xi h)]}_{2\sin^2\left(\frac{\xi h}{2}\right)} \tag{1.202}$$

$$= 1 - 4|\nu|(1 - |\nu|))\sin^2\left(\frac{\xi h}{2}\right), \tag{1.203}$$

z čehož plyne, že že pro splnění $|\lambda(\xi)| \leq 1$ je třeba, aby $|\nu| \in [0, 1] \Leftrightarrow \nu \in [-1, 1]$. To je přesně stejný výsledek, který jsme obdrželi z CFL podmínky.

Chyba diskretizace $\varepsilon_j^n = U_j^n - u(x_j, t_n)$ se určí stejným způsobem, jako u parabolických úloh rozvojem příslušných diferenčních náhrad do Taylorovy řady. Chyba diskretizace schématu (1.194) je tak řádu $\mathcal{O}(\tau + h)$, jelikož používá pouze dopředné, či zpětné diferenční náhrady. Abychom docílili lepší chyby diskretizace, je třeba užít centrální diference. Podívejme se tedy na schéma

$$\frac{U_j^{n+1} - U_j^n}{\tau} + a\frac{U_{j+1}^n - U_{j-1}^n}{2h}. \tag{1.204}$$

Přenecháváme čtenáři k procvičení, že chyba diskretizace je v tomto případě $\varepsilon_{h,\tau} = \mathcal{O}(\tau + h^2)$. Zajímavější je však Fourierova analýza. Dosaďme tedy $U_j^n = e^{\mathrm{i}\xi jh}\lambda^n$ do (1.204). Tím získáme

$$\mathrm{e}^{\mathrm{i}\xi jh}\lambda^{n+1} = \mathrm{e}^{\mathrm{i}\xi jh}\lambda^n - \frac{\nu}{2}\left(\mathrm{e}^{\mathrm{i}\xi(j+1)h}\lambda^n - \mathrm{e}^{\mathrm{i}\xi(j-1)h}\lambda^n\right), \tag{1.205}$$

což po vydělení $e^{\mathrm{i}\xi jh}\lambda^n$ dává

$$\lambda(\xi) = 1 - \frac{\nu}{2}\left(\mathrm{e}^{\mathrm{i}\xi h} - \mathrm{e}^{-\mathrm{i}\xi h}\right) \tag{1.206}$$

$$= 1 - \frac{\nu}{2}\left(\cos(\xi h) + \mathrm{i}\sin(\xi h) - \cos(\xi h) + \mathrm{i}\sin(\xi h)\right) \tag{1.207}$$

$$= 1 - \mathrm{i}\nu\sin(\xi h). \tag{1.208}$$

Absolutní hodnota $\lambda(\xi)$ je však vždy větší než 1, neboť $|\lambda(\xi)| = \sqrt{1 + \nu^2\sin^2(\xi h)} > 1$ pro každé $\xi \in \mathbb{R}$, pro které $\sin(\xi h) \neq 0$. Pro každý proces zjemňování je tedy schéma nestabilní. Z toho je vidět, že CFL podmínka je opravdu pouze nutnou podmínkou, nikoliv však postačující.

Dalším používaným schématem je schéma Laxovo-Friedrichsovo:

$$\frac{U_j^{n+1} - \frac{1}{2}\left(U_{j+1}^n + U_{j-1}^n\right)}{\tau} + a\frac{U_{j+1}^n - U_{j-1}^n}{2h}. \tag{1.209}$$

V Laxově-Fridrichsově schématu se podíváme blíže i na chybu diskretizace.

$$\begin{aligned}
\varepsilon_{h,\tau} &= \frac{u(x,t+\tau) - \frac{1}{2}(u(x+h,t)+u(x-h,t))}{\tau} + a\frac{u(x+h,t)-u(x-h,t)}{2h} \\
&= \frac{u(x,t) + u_{,t}(x,t)\tau + \frac{1}{2}u_{,tt}(x,t)\tau^2 + \mathcal{O}(\tau^3) - \frac{1}{2}(\overbrace{2u(x,t) + u_{,xx}(x,t)h^2 + \mathcal{O}(h^4)}^{\text{liché se odečetly, sude jsou 2x}})}{\tau} \\
&= +a\frac{\overbrace{2u_{,x}(x,t)h + \frac{2}{6}u_{,xxx}(x,t)h^3 + \mathcal{O}(h^5)}^{\text{sudé se odečetly, liché jsou 2x}}}{2h} \\
&= \frac{1}{2}u_{,tt}(x,t)\tau - \frac{1}{2}u_{,xx}(x,t)\frac{h^2}{\tau} + \mathcal{O}(\tau^2) + \mathcal{O}\left(\frac{h^4}{\tau}\right) + \frac{a}{6}u_{,xxx}(x,t)h^2 + \mathcal{O}(h^4) \\
&= \mathcal{O}(\tau + h^2), \qquad \text{pokud } \frac{h^2}{\tau} \to 0.
\end{aligned} \tag{1.210}$$

Podmínku pro chybu diskretizace ve tvaru $\frac{h^2}{\tau} \to 0$ nazýváme podmínkou konzistence. Pokud tato podmínka je splněna, pak je Laxovo-Friedrichsovo schéma konzistentní s PDR (1.188). V opačném případě přebývající člen $\frac{1}{2}u_{,xx}(x,t)\frac{h^2}{\tau}$ způsobí, že schéma nebude konvergovat (protože $\varepsilon_{h,\tau}$ nepůjde k nule).

Fourierova analýza chyby dává v případě Laxova-Friedrichsova schématu stejnou podmínku stability, jako CFL podmínka. Dosadíme-li totiž $U_j^n = \mathrm{e}^{\mathrm{i}\xi jh}\lambda^n$ do (1.209), pak po zkrácení výrazem $\mathrm{e}^{\mathrm{i}\xi jh}\lambda^n$ dostaneme

$$\lambda(\xi) = \frac{1}{2}(\mathrm{e}^{\mathrm{i}\xi h} + \mathrm{e}^{-\mathrm{i}\xi h}) - \frac{\nu}{2}(\mathrm{e}^{\mathrm{i}\xi h} - \mathrm{e}^{-\mathrm{i}\xi h}) \tag{1.211}$$

$$= \cos(\xi h) - \mathrm{i}\nu\sin(\xi h). \tag{1.212}$$

Velikost amplifikačního parametru $\lambda(\xi)$ pak plyne z výpočtu

$$|\lambda(\xi)|^2 = \cos^2(\xi h) + \nu^2\sin^2(\xi h) = 1 + (\nu^2 - 1)\sin^2(\xi h). \tag{1.213}$$

Pokud má platit $\lambda(\xi) \leq 1$, musí být $(\nu^2 - 1)\sin^2(\xi h) \leq 0$, což je právě tehdy, je-li $|\nu| \leq 1$. Podrobnější analýza by byla založena na Fourierově transformaci a zabývala by se tzv. disperzí, což je jev, kdy schéma šíří různou rychlostí vlny o různé frekvenci, v důsledku čehož mohou rovněž nastat oscilace a nestability v řešení. Tato analýza však již překračuje rámec tohoto textu.

Posledním schématem, na které se v rámci této kapitoly podíváme, je schéma Laxovo-Wendroffovo. Toto schéma pro zajímavost uvedeme ve tvaru vhodném pro řešení rovnice (1.188) s nenulovou pravou stranou, tj.

$$u_{,t} + au_{,x} = f, \quad x \in \mathbb{R}, \ t > 0, \tag{1.214}$$

kde $f = f(x,t)$ je daná funkce. Schéma přitom vychází z následujících úvah. Z (1.214) plyne $u_{,t} = f - au_{,x}$. Pokud rovnici (1.214) zderivujeme jednou podle času a zvlášť podle prostorové proměnné, dostaneme tyto pomocné vztahy.

$$u_{,xt} + au_{,xx} = f_{,x} \quad \Rightarrow \quad u_{,xt} = f_{,x} - au_{,xx} \tag{1.215}$$

$$u_{,tt} + au_{,xt} = f_{,t} \quad \Rightarrow \quad u_{,tt} = f_{,t} - au_{,xt}. \tag{1.216}$$

Kombinací obou těchto vztahů pak obdržíme $u_{,tt} = f_{,t} - a(f_{,x} - au_{,xx})$. Těchto vztahů nyní využijeme k odvození Laxova-Wendroffova schématu. Předně si ještě připomeňme, že

$$u(x,t+\tau) = u(x,t) + u_{,t}(x,t)\tau + \frac{1}{2}u_{,tt}(x,t)\tau^2 + \mathcal{O}(\tau^3). \tag{1.217}$$

Dosazením právě odvozených pomocných vztahů do tohoto rozvoje po troše námahy dostaneme

$$\begin{aligned} &u(x,t+\tau) - u(x,t) = \Delta_{+t}u \\ &= (f - au_{,x})\tau + \frac{1}{2}(f_{,t} - a(f_{,x} - au_{,xx}))\tau^2 + \mathcal{O}(\tau^3) \\ &= \frac{\tau}{2}(f + f + f_{,t}\tau) - a\tau u_{,x} + \frac{a^2}{2}u_{,xx}\tau^2 - \frac{a}{2}f_{,x}\tau^2 + \mathcal{O}(\tau^3) \\ &= -\frac{a\tau}{2h}\Delta_{0x}u + \mathcal{O}(\tau h^2) + \frac{a^2\tau^2}{2h^2}\delta_x^2 u - \frac{a\tau^2}{4h}\Delta_{0x}f + \frac{\tau}{2}[f(x,t+\tau) + f(x,t)] + \mathcal{O}(\tau^3), \end{aligned} \tag{1.218}$$

neboli po dosazení $U_j^n \approx u(x_j, t_n)$ dostáváme

$$\frac{U_j^{n+1} - U_j^n}{\tau} + a\frac{U_{j+1}^n - U_{j-1}^n}{2h} - \underbrace{\frac{a^2\tau}{2}\frac{U_{j+1}^n - 2U_j^n + U_{j+1}^n}{h^2}}_{\text{přidaná difuzivita}} = \frac{1}{2}(f_j^{n+1} + f_j^n) - \frac{a\tau}{4h}(f_{j+1}^n - f_{j-1}^n). \tag{1.219}$$

Přenecháváme čtenáři za cvičení, že chyba diskretizace Laxova-Wendroffova schématu je řádu $\mathcal{O}(\tau^2 + h^2)$. Podívejme se ještě krátce na Fourierovu analýzu. Dosazením $U_j^n = \mathrm{e}^{\mathrm{i}\xi jh}\lambda^n$ do (1.219) vychází po několika úpravách

$$\lambda(\xi) = 1 - \mathrm{i}\nu\sin(xih) - 2\nu^2\sin^2\left(\frac{\xi h}{2}\right), \tag{1.220}$$

z čehož dostáváme postupně

$$\begin{aligned} |\lambda(\xi)|^2 &= 1 - 4\nu^2\sin^2\left(\frac{\xi h}{2}\right) + 4\nu^4\sin^4\left(\frac{\xi h}{2}\right) + \nu^2\sin^2(\xi h) \\ &= 1 - 4\nu^2\sin^2\left(\frac{\xi h}{2}\right) + 4\nu^4\sin^4\left(\frac{\xi h}{2}\right) + 4\nu^2 sin^2\left(\frac{\xi h}{2}\right)\left(1 - sin^2\left(\frac{\xi h}{2}\right)\right) \\ &= 1 + 4\nu^2(\nu^2 - 1)\sin^4\left(\frac{\xi h}{2}\right). \end{aligned} \tag{1.221}$$

Vidíme tedy, že schéma je stabilní za podmínky $|\nu| \leq 1$, což je opět totožný výsledek, který dává CFL podmínka. Schéma (1.219) je jako u předchozích případů možné přepsat do podoby vhodnější pro výpočet. S použitím značení $\nu = \frac{a\tau}{h}$ lze dostat následující podobu Laxova-Wendroffova schématu

$$U_j^{n+1} = U_j^n - \frac{\nu}{2}(U_{j+1}^n - U_{j-1}^n) + \frac{\nu^2}{2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n) + \frac{\tau}{2}(f_j^{n+1} + f_j^n) - \frac{a\tau}{4h}(f_{j+1}^n - f_{j-1}^n). \tag{1.222}$$

Pro porovnání jednotlivých schémat na konkrétních numerických příkladech odkazujeme čtenáře na interaktivní pomůcky vytvořené k tomuto textu v rámci rozšířených přednášek z předmětu NAK 1, které jsou dostupné na webových stránkách http://mech.fsv.cvut.cz/nak. Na příkladech uvedených na této adrese si čtenář může prohlédnout a změnami parametrů zkoumat chování jednotlivých schémat. Na tomto místě učiníme několik komentářů ke zmíněným schématům.

Explicitní schéma typu "upwind" (1.194) bylo stabilní pro $|\nu| \in [0, 1]$ a řádu $\mathcal{O}(\tau + h)$. Spuštěním animací na uvedených webových stránkách čtenář zjistí, že dochází postupem času ke značnému útlumu šířící se vlny oproti přesnému řešení. Nedochází však k žádným oscilacím. Laxovo-Friedrichsovo schéma je při splnění podmínky konzistence lepšího řádu přesnosti, útlum v amplitudě je však ještě větší. Schéma s centrální diferencí v prostorové proměnné je stejného řádu přesnosti, jako Laxovo-Friedrichsovo, je však silně nestabilní a vykazuje velké oscilace pro jakoukoliv hodnotu ξ. Podíváme-li se na schéma (1.219) pozorněji (předpokládejme teď nulovou pravou stranu), všimneme si, že vypadá stejně, jako předchozí schéma s centrální diferencí v prostoru. Kromě toho však obsahuje ještě diferenční náhradu za druhou prostorovou derivaci, která se v rovnici(1.214) nevyskytuje. Vzpomeneme-li si na parabolické úlohy, vysvitne, že tento přídavný člen modeluje difuzi. Vlivem toho se utlumí oscilace, které jsme viděli v předchozím schématu. Tyto oscilace však stále přetrvávají, máme však přesnost $\mathcal{O}(\tau^2 + h^2)$. Podrobnější analýza, založena na Fourierově transformaci, tyto oscilace dokáže vysvětlit pomocí popisu jevů, jako je tzv. disperze, což je jev, kdy schéma šíří různou rychlostí vlny o různé frekvenci, v důsledku čehož mohou rovněž nastat oscilace a nestability v řešení. Tato analýza však již překračuje rámec tohoto textu.

1.4.2 Vlnová rovnice

Hyperbolická rovnice, kterou se budeme nyní zabývat má tvar

$$\frac{\partial^2 u}{\partial^2 t} = a^2 \frac{\partial^2 u}{\partial^2 x} \tag{1.223}$$

kde a značí rychlost vlny. Tato rovnice se někdy také nazývá rovnicí struny, neboť popisuje šíření vlnění v napnuté struně. Stejný popis mají jiné jevy vlnění v mechanice, optice, či elektrodynamice. K této rovnici je samozřejme třeba přidat okrajové a počáteční podmínky. V tomto případě je však třeba předepsat nejen počáteční hodnotu funkce $u(x, 0) = c(x)$, ale i hodnotu derivace podle času $u_{,t}(x, 0) = d(x)$. Řešení rovnice struny metodou sítí opět spočívá v nahrazení derivací podílem diferencí. Aproximaci hledané funkce v bodě x_j a čase t_n označíme jako

$$u(x_j, t_n) \approx U_j^n \tag{1.224}$$

Použijeme centrální diference na časovou i prostorovou derivaci

$$\frac{\partial^2 U_j}{\partial^2 x} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2} \tag{1.225}$$

$$\frac{\partial^2 U_j}{\partial^2 t} = \frac{U_j^{n+1} - 2U_j^n + U_j^{n-1}}{\tau^2} \tag{1.226}$$

Po dosazení (1.225) a (1.226) do (1.223) dostáváme diferenční rovnici

$$\frac{U_j^{n+1} - 2U_j^n + U_j^{n-1}}{\tau^2} = a^2 \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2} \tag{1.227}$$

ze které můžeme vyjádřit U_j^{n+1}

$$U_j^{n+1} = 2U_j^n - U_j^{n-1} + (\mu)^2 \left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right) \tag{1.228}$$

kde jsme podíl $\frac{a\tau}{h}$ označili μ. Rovnicí (1.228) jsme vyjádřili řešení v čase t_{n+1} pomocí řešení v předchozích krocích t_n a t_{n-1}, jedná se tedy o metodu explicitní. Navíc vidíme, že nám nestačí znát řešení pouze v předchozím čase, ale je třeba znát dva předchozí kroky. To činí problémy v prvním časovém kroku t_1, kdy se hodnota v předchozím kroku t_0 získá z počáteční podmínky, ale hodnotu v čase t_{-1} neznáme. Obvykle se stanoví pomocí předpokladu konstantní rychlosti:

$$U_j^1 = c(x_j) + \tau d(x_j) \tag{1.229}$$

Pro odvození chyby diskretizace dosadíme přesné řešení u do (1.228) a použijeme Taylorův rozvoj:

$$u_{,tt} + \frac{1}{12}\tau^2 u_{,tttt} + ... = a^2 \left(u_{,xx} + \frac{1}{12}h^2 u_{xxxx} + ...\right) \tag{1.230}$$

Pokud původní rovnici (1.223) zderivujeme dvakrát podle času máme vztah

$$u_{,tttt} = a^2 u_{,xxtt} \tag{1.231}$$

Pokud původní rovnici (1.223) zderivujeme dvakrát podle prostorové proměnné dostáváme

$$u_{,ttxx} = a^2 u_{,xxxx} \tag{1.232}$$

Spojením rovnic (1.231), (1.232) a s využitím zámeny derivací vychází

$$u_{,tttt} = a^2 u_{,xxtt} = a^2 u_{,ttxx} = a^4 u_{,xxxx} \tag{1.233}$$

Tento výsledek dosadíme do Taylorova rozvoje (1.230)

$$u_{,tt} + \frac{1}{12}\tau^2 a^4 u_{,xxxx} + ... = a^2 \left(u_{,xx} + \frac{1}{12}h^2 u_{xxxx} + ...\right) \tag{1.234}$$

Tím dostáváme chybu diskretizace

$$\varepsilon_h \leq \frac{1}{12}(\tau^2 a^4 - h^2 a^2) u_{,xxxx} = \mathcal{O}(\tau^2 + h^2) \tag{1.235}$$

Jak již víme z předchozích kapitol, explicitní metody bývají netabilní. Provedeme proto Fourierově analýze stability. Hledejme řešení U_j^n ve tvaru

$$U_j^n = \mathrm{e}^{\mathrm{i}kjh}\lambda^n \tag{1.236}$$

Po dosazení do (1.228) a vydělení výrazem $\mathrm{e}^{\mathrm{i}kjh}$ a λ^n dostáváme

$$\lambda^2 = \left(2 + \mathrm{e}^{\mathrm{i}kh} - 2 + \mathrm{e}^{-\mathrm{i}kh}\right)\lambda + 1 \tag{1.237}$$

využijeme nám již známý vztah

$$\mathrm{e}^{\mathrm{i}kh} - 2 + \mathrm{e}^{-\mathrm{i}kh} = 2(\cos kh - 1) = -4\sin^2\frac{kh}{2} \tag{1.238}$$

vychází rovnice, ze které spočítáme amplifikační faktor

$$\lambda^2 = \left(2 - 4\mu^2\sin^2\frac{kh}{2}\right)\lambda + 1 \tag{1.239}$$

Kořeny této kvadratické rovnice mají tvar

$$\lambda_{1,2} = \alpha \pm \sqrt{\alpha^2 - 1} \quad \textit{kde} \quad \alpha = 1 - 2\mu^2\sin^2\frac{kh}{2} \tag{1.240}$$

kde $\alpha = 1 - 2\mu^2\sin^2\left(\frac{kh}{2}\right)$. Všimněme si, že pro $\alpha > 1$ bude absolutní hodnota alespoň jednoho kořene kvadratické rovnice větší než jedna. Pokud vezmeme $|\alpha| \leq 1$ amplifikační faktory budou komplexní čísla

$$\lambda_{1,2} = \alpha \pm \mathrm{i}\sqrt{1 - \alpha^2} \tag{1.241}$$

o velikosti

$$|\lambda_{1,2}|^2 = |\alpha|^2 \pm |1 - \alpha^2| = 1 \tag{1.242}$$

z čehož plyne, že schéma je podmíněně stabilní s podmínkou stability

$$-1 \leq \alpha \leq 1 \tag{1.243}$$

neboli

$$-1 \leq 1 - 2\mu^2\sin^2\left(\frac{kh}{2}\right) \leq 1 \tag{1.244}$$

upravíme první podmínku na tvar

$$1 \geq \mu^2\sin^2\left(\frac{kh}{2}\right) \tag{1.245}$$

aby tedy nerovnost platila pro libovolné k musí platit

$$\mu^2 \leq 1 \tag{1.246}$$

Druhá nerovnost má tvar

$$1 - 2\mu^2\sin^2\left(\frac{kh}{2}\right) \leq 1 \tag{1.247}$$

Po úpravách

$$\mu^2 \sin^2\left(\frac{kh}{2}\right) \geq 0 \tag{1.248}$$

Tato nerovnost musí opět platit pro libovolné k:

$$\mu^2 \geq 0 \tag{1.249}$$

jak vidíme, takto podmínka je vždy splněna. Pomocí definice $\mu = \dfrac{a\tau}{h}$ můžeme podmínku stability (1.246) přepsat jako

$$\frac{|a|\tau}{h} \leq 1 \tag{1.250}$$

Tato nerovnost nám omezuje poměr parametrů časové a prostorové diskretizace. Vidíme, že se zvyšující se rychlostí vlny musíme zmenšovat časový krok vzhledem k prostorové diskretizaci. Jinou možností je volit hrubší prostorovou diskretizaci, to však není vhodné vzhledem k chybě diskretizace. Pokud definujeme rychlost numerického schématu jako $a_n = \frac{h}{\tau}$, podmínka (1.250) nám říká, že rychlost numerického schématu musí být větší nebo stejná než skutečná rychlost.

Z důvodu podmíněné stability explicitního schématu se často pro časovou integraci používá Newmarkovo schéma. Newmarkova metoda je založená na následujících vztazích mezi posuny, rychlostmi a zrychleními

$$U_j^{n+1} = U_j^n + \tau \dot{U}^n + \frac{1}{2}\tau^2 \left((1-2\beta)\ddot{U}_j^n + 2\beta \ddot{U}_j^{n+1}\right) \tag{1.251}$$

$$\dot{U}_j^{n+1} = \dot{U}^n + \tau \left((1-\gamma)\ddot{U}_j^n + \gamma \ddot{U}_j^{n+1}\right) \tag{1.252}$$

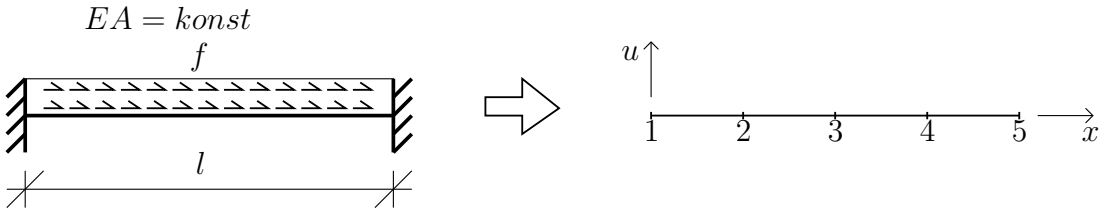
Newmark navrhl parametry tak, aby schéma bylo nepodíněně stabilní, konkrétní hodnoty jsou $\beta = \dfrac{1}{4}$ a $\gamma = \dfrac{1}{2}$.

2 Metoda konečných prvků

2.1 Úvod

Obsahem této kapitoly je matematický úvod do metody konečných prvků (MKP). S ní se studenti již setkali v předmětu Numerická Analýza Konstrukcí 1 (NAK1), kde však byl kladen důraz především na výpočetní postupy a techniky vedoucí k nalezení příslušného přibližného řešení. Již méně pozornosti bylo věnováno vyšetřování vlastností použitých numerických schémat, jejich konvergence, či dokonce vyšetřování existence a jednoznačnosti řešení výchozí úlohy. Bez těchto znalostí si však nemůžeme být jisti, zda námi získané výsledky jsou správné (v žádném rozumném smyslu), nebo jaké chyby oproti přesnému řešení jsme se dopustili. Zmíněné problémy a jejich řešení jsou náplní této kapitoly. V ní se podíváme na vybrané problémy řešené v předmětu NAK1 z obecnějšího hlediska jakožto na variační formulaci eliptické úlohy, definujeme je v matematicky exaktním smyslu a dokážeme, že takto definovaný problém má jednoznačně určené řešení. Dále dokážeme, za jakých podmínek a s jakým řádem příslušná přibližná řešení konvergují k řešení přesnému.

Stejně jako u metody konečných diferencí začneme s úlohou taženého-tlačeného prutu, viz Obr. 2.1. Pokud uvažujeme konstantní tuhost a průřezovou plochu po celé délce prutu, můžeme úlohu popsat eliptickou diferenciální rovnicí druhého řádu. Pokud uvažujeme kon-



Obr. 2.1: *Příklad 1 - Zadání*

stantní normálovou tuhost i průřezovou plochu po celé délce prutu, můžeme úlohu popsat eliptickou diferenciální rovnicí druhého řádu

$$EAu_{,xx} = f, \tag{2.1}$$

kde $u_{,xx}$ značí druhou derivaci podle x. Tato rovnice má být splněna v každém bodě intervalu $(0, l)$. Pro určení jednoznačného řešení je ještě třeba přidat okrajové podmínky, které (dle Obr. 2.1) uvažujeme ve formě

$$u(0) = 0 \qquad N(l) = EAu_{,x}(l) = 0. \tag{2.2}$$

V tomto jednoduchém případě je možné rovnici (2.1) vyřešit přesně. Dvojnásobnou integrací postupně dostaneme, že obecné řešení funkce posunů u má tvar

$$u = \frac{1}{EA}\left[fx^2 + C_1 x + C_2\right]. \tag{2.3}$$

Po uplatnění okrajových podmínek pak dostáváme

$$u = \frac{1}{EA}\left[fx^2 + flx\right]. \tag{2.4}$$

Aby rovnice (2.1) vůbec měla smysl (tzn. zejména, aby v ní uvedené výrazy byly korektně definovány) je třeba, aby hledaná funkce měla v celém intervalu $(0, l)$ spojité druhé derivace. Navíc, abychom mohli splnit okrajové podmínky, potřebujeme spojitost ve funkčních hodnotách až do krajních bodů. Matematicky vyjádřeno klademe na hledanou funkci u požadavky $u \in C^2(0,l) \cap C^0[0,l]$, díky čemuž však musí být též $f \in C^0(0,l)$. To jsou velice silné požadavky, které v praktických aplikacích (zejména ve složitějších úlohách, než je náš ilustrativní příklad) nemusejí být splněny. Rovnici (2.1) by pak nebylo možno vůbec řešit. Bylo by proto užitečné pojem řešení této rovnice vhodným způsobem modifikovat (rozšířit) tak, aby požadavky na hledanou funkci byly slabší (a tedy aby bylo snazší ji najít), ale zároveň ne příliš slabé - musí to být řešení v nějakém rozumném smyslu. Navíc bychom chtěli, aby množina řešení původního problému byla podmnožinou množiny řešení modifikovaného problému - tedy aby nová formulace byla s tou starou konzistentní. Jedna z cest vede přes formulaci tzv. slabého řešení.

V předmětu NAK1 se studenti seznámili s postupem, jak vypočíst přibližné řešení pomocí MKP. Výchozím bodem přitom byl pojem slabé formulace problému (2.1). Postup k získání slabé formulace byl následující: Vezměme libovolnou funkci $v \in C^\infty(0,l)$ takovou, že $v(0) = 0$.[19] Vynásobme s ní rovnici (2.1) a výslednou rovnici zintegrujme po délce prutu. Po provedení integrace per-partes na člen obsahující druhé derivace a uplatnění okrajových podmínek dostaneme

$$\int_0^l EAu_{,x}v_{,x}\mathrm{d}x = \int_0^l fv\mathrm{d}x, \quad \forall v \in C^\infty(0,l),\ v(0) = 0, \tag{2.5}$$

což je zmíněná slabá formulace rovnice (2.1). Slabé formulaci (2.5) se někdy říká variační formulace, protože funkce v může být libovolná. V mechanice se (2.5) obvykle nazývá princip virtuálních prací. Funkce v jsou pak takzvaná virtuální přemístění. K této souvislosti se ještě vrátíme později. Řešení rovnice (2.5) se nazývá slabé řešení. Zaveďme nyní v duchu pozdějšího výkladu následující značení.

$$a(u,v) := \int_0^l EAu_{,x}v_{,x}\mathrm{d}x, \qquad (f,v) := \int_0^l fv\mathrm{d}x, \tag{2.6}$$

Potom můžeme slabou formulaci (2.5) zapsat jako

$$a(u,v) = (f,v), \quad \forall v \in C^\infty(0,l),\ v(0) = 0. \tag{2.7}$$

Později ukážeme, že $a(u,v)$ je bilineární forma na jistém prostoru funkcí a (f,v) představuje lineární funkcionál na tomto prostoru. Pouhým pohledem na (2.5) je intuitivně jasné, že na

[19] $C^\infty(0,l)$ značíme prostor nekonečně hladkých (majících spojité všechny derivace) funkcí. Požadavek $v(0) = 0$ má za následek fakt, že funkce z tohoto prostoru splňují Dirichletovské okrajové podmínky rovnice (2.1). Pro podrobnější informaci odkazujeme čtenáře na přednášky z Matematiky 4.

hledanou funkci u jsou kladeny menší požadavky, než v případě řešení problému (2.1) - je třeba, aby funkce u měla pouze první derivace. Situace je však o něco složitější, neboť požadované derivace jsou derivace pouze ve slabém smyslu. S tím souvisí i definice prostorů funkcí, ve kterých je nutné hledat řešení u, pokud úlohu formulujeme pouze ve slabém smyslu, tedy jako (2.5). Hledané prostory jsou tzv. Sobolevovy prostory, v našem případě prostor $W_2^1(0,l)$. Funkce patřící do $W_2^1(0,l)$ jsou, zhruba řečeno, integrovatelné ve druhé mocnině a stejnou vlastnost mají i jejich derivace. Přesněji řečeno pro ně platí $\int_0^l u^2 \mathrm{d}x < \infty$ a $\int_0^l (u')^2 \mathrm{d}x < \infty$, kde derivace je opět uvažována ve slabém smyslu. Uvedený prostor $W_2^1(0,l)$ přitom budeme značit obvykle $H^1(0,l)$. Pro ilustraci prostoru $H^1(0,l)$ lze říci, že funkce v něm obsažené již musí být spojité (ovšem až na množiny míry nula).[20] Označme nyní pro potřeby tohoto příkladu

$$V = \{v \in H^1(0,l),\ v(0) = 0\}. \tag{2.8}$$

Potom je možné slabou formulaci úlohy přepsat takto: Najděme $u \in V$ tak, že

$$a(u,v) = (f,v), \quad \forall v \in V. \tag{2.9}$$

První otázka, kterou je třeba zodpovědět, je, zda a za jakých podmínek je řešení slabé formulace (2.9) řešením původní (silné) formulace (2.1) (opačný problém je snadný a ověří se prostým dosazením). Mějme tedy funkci u, která řeší (2.9) a předpokládejme o ní, že $u \in C^2(0,l) \cap C^0([0,l])$. O funkci f představující zatížení předpokládejme, že je spojitá, tedy že $f \in C^0([0,l])$. Potom platí, že u řeší (2.1).

Vezměme libovolnou funkci $v \in V$. Potom integrací per-partes postupně dostáváme

$$\begin{aligned}(f,v) = a(u,v) = \int_0^l (-u_{,xx})v\mathrm{d}x + u_{,x}(l)v(l) \ \Rightarrow \\ \int_0^l (u_{,xx} + f)v\mathrm{d}x - u_{,x}(l)v(l) = 0\end{aligned} \tag{2.10}$$

Je tedy třeba dokázat, že

$$(i)\ \ u_{,xx} + f = 0, \tag{2.11}$$

$$(ii)\ \ u_{,x}(1) = 0. \tag{2.12}$$

Jelikož (2.10) musí platit pro libovolnou funkci $v \in V$, lze (2.11), (2.12) dokázat její speciální volbou. Pro důkaz (2.11) zvolme například $v(x) = (x - x_1)^2(x - x_2)^2$ uvnitř $[x_1, x_2] \subset [0,l]$ a $v(x) = 0$ mimo $[x_1, x_2]$. Jelikož $w(x) = u_{,xx} + f$ je dle předpokladu spojitá funkce, můžeme bez újmy na obecnosti předpokládat, že na intervalu $[x_1, x_2]$ nemění znaménko. Potom ale

$$(w,v) = \int_0^l [u_{,xx} + f](x - x_1)^2(x - x_2)^2\mathrm{d}x = 0 \ \Rightarrow\ u_{,xx} + f = 0, \tag{2.13}$$

protože $v(x) = (x - x_1)^2(x - x_2)^2 > 0$ na $[x_1, x_2]$ a integrál z kladné spojité funkce na uzavřeném intervalu je roven nule právě tehdy, když je identicky roven nule integrand. Pro

[20] Pro přesné definice a stručnou informaci odkazujeme čtenáře na Dodatek 2.7, případně na rozšířené přednášky z Matematiky 4 vedené doc. Nekvindou.

důkaz (2.12) volme například $v(x) = x/l$. Jelikož už máme $u_{,xx} + f = 0$, plyne z (2.10), že $u_{,x} = 0$, a důkaz je hotov.

Další z důležitých otázek je, zda slabá formulace problému (2.9) má řešení a zda je toto řešení jednoznačně určeno vstupními daty. To je mimořádně důležitá otázka, neboť ze slabé formulace vycházející metoda konečných prvků konstruuje přibližná řešení tohoto problému, od nichž očekáváme, že se zjemňující se sítí budou konvergovat k řešení přesnému. Pokud by toto řešení neexistovalo nebo nebylo určeno jednoznačně, přibližná řešení by neměla valnou hodnotu. Na otázku existence a jednoznačnosti řešení problému (2.9) se podíváme ve větší obecnosti v následující sekci.

2.2 Variační formulace eliptických problémů

Náplní této sekce je podání důkazu existence a jednoznačnosti řešení slabé (variační) formulace lineárních eliptických problémů a zasazení problematiky řešené v předmětu NAK1 do obecnějšího kontextu. Celý problém rozdělíme na případ, kdy příslušná bilineární forma $a(u, v)$ je symetrická a nesymetrická. V obou případech však formální zápis bude vypadat stejně. Zavedeme proto následující

Abstraktní variační problém (AVP).
Hledáme $u \in V$ takové, že

$$a(u, v) = F(v), \quad \forall v \in V, \tag{2.14}$$

kde V je reálný Hilbertův prostor s normou $||\cdot||_V$,
$a : V \times V \to \mathbb{R}$ je bilineární forma, která je spojitá a koercivní na V, tj.

$$\exists M, \alpha > 0 : \qquad |a(u, v)| \leq M||u||_V||v||_V, \quad \forall u, v \in V, \tag{2.15}$$

$$a(u, u) \geq \alpha||u||_V^2, \quad \forall u \in V \tag{2.16}$$

a $F(v) = (f, v)$ je spojitý lineární funcionál na V.

2.2.1 Symetrický variační problém

V této části se budeme zabývat AVP pro případ, že příslušná bilineární forma $a(u, v)$ je symetrická. V tom případě se existence a jednoznačnost řešení AVP opírá o Riezsovu větu o reprezentaci. Dále ukážeme souvislost s problémem minimalizace kvadratického funkcionálu, který v aplikacích v mechanice představuje potenciální energii systému.

Nechť tedy máme Hilbertův prostor V a výše definovaný AVP, kde navíc bilineární forma $a(u, v)$ je symetrická, tj. platí $a(u, v) = a(v, u)$. Předně ukážeme, že $a(u, v)$ je skalární součin na V, jinými slovy $(V, a(\cdot, \cdot))$ tvoří Hilbertův prostor. Ze symetrie a linearity $a(u, v)$ plynou vlastnosti $(i) - (iii)$ z definice skalárního součinu z Dodatku 2.5. Z koercivity $a(u, v)$ plyne, že pokud $a(v, v) = 0$, pak nutně $v = 0$ a $a(u, v)$ je tedy skalární součin. Jím indukovaná norma se nazývá energetická norma, kterou značíme $||v||_E = \sqrt{a(v, v)}$. Lze dokázat, že prostor $(V, ||\cdot||_E)$ je úplný, a tedy Banachův. Z toho je zřejmé, že $(V, a(\cdot, \cdot))$ je Hilbertův prostor.[21]

[21] Hilbertův prostor není nic jiného, než Banachův prostor se skalárním součinem, kde je příslušná norma tímto skalárním součinem indukována.

Z uvedených úvah plyne, že existence a jednoznačnost řešení AVP pro případ, kdy $a(u,v)$ je symetrická, je přímým důsledkem Riezsovy věty o reprezentaci[22]. Ta říká, že libovolný spojitý lienární funkcionál $F(v)$ na Hilbertově prostoru V může být vyjádřen pomocí pevně zvoleného prvku $u \in V$ jako skalární součin (u,v). Jelikož jsme výše ukázali, že symetrická bilineární forma $a(u,v)$ tvoří skalární součin na V, pak pro danou funkci f (čímž je dán spojitý lineární funkcionál $(f,v) = F(v) \in V^*$)[23] existuje právě jeden prvek $u \in V$ tak, že $a(u,v) = (f,v)$ pro všechna $v \in V$, a u je tedy řešením symetrického AVP.

Souvislost s minimalizací funkcionálu energie

V případě symetrického variačního problému lze postupovat také jinak. Ukážeme, že řešení symetrického variačního problému je ekvivalentní hledání minima jistého kvadratického funkcionálu. V aplikacích v mechanice tento funkcionál představuje celkovou potenciální energii systému. Tento fakt ukážeme na jednoduchém příkladu taženého-tlačeného prutu. Nejprve však zmíněný kvadratický funkcionál definujeme a dokážeme, že problém nalezení funkce, ve které tento funkcionál nabývá svého minima, je ekvivalentní řešení původního symetrického AVP. Zformulujme celý problém jako větu.

Věta 2.1 (Ekvivalence řešení symetrického AVP a minimalizace funcionálu energie)**.** *Nechť bilineární forma z AVP je symetrická. Pak řešení AVP je ekvivalentní minimalizaci kvadratického funkcionálu*

$$J(v) = \frac{1}{2}a(v,v) - (f,v). \tag{2.17}$$

Důkaz
Pro důkaz ekvivalence dvou tvrzení je třeba dokázat, že z jednoho plyne druhé a naopak. Nechť tedy nejdříve u je řešením AVP. Tedy platí $a(u,v) = (f,v)$ pro všechna $v \in V$. Pak ovšem pro libovolné $v \in V$ platí

$$\begin{aligned} J(u+v) &= \frac{1}{2}a(u+v,u+v) - (f,u+v) = \\ &= \frac{1}{2}a(u,u) + a(u,v) + \frac{1}{2}a(v,v) - (f,u) - (f,v) = \\ &= J(u) + \frac{1}{2}a(v,v), \end{aligned} \tag{2.18}$$

kde jsme použili linearitu a symetrii $a(u,v)$ a dále předpokladu, že $a(u,v) = (f,v)$. Z (2.18) a koercivity $a(u,v)$ plyne, že

$$J(u+v) - J(u) \geq \alpha||u||_V^2 \geq 0, \tag{2.19}$$

a tedy hodnota funkcionálu J pro libovolnou jinou funkci je vždy větší, než pro u. Platí tedy

$$J(u) = \inf_{v \in V} J(v), \tag{2.20}$$

[22] Přesné znění Riezsovy věty o reprezentaci a její důkaz najde čtenář v Dodatku (2.6)
[23] V^* značí duální prostor k V. Pro více informací, viz Dodatek (2.6)

neboli u minimalizuje J a první implikace je dokázána.

Nechť nyní naopak platí, že u minimalizuje funkcionál J. Potom ale pro libovolnou funkci v a libovolné číslo $\varepsilon > 0$ platí

$$J(u + \varepsilon v) \geq J(u). \tag{2.21}$$

Rozepsáním stejným způsobem, jako v (2.18) dostáváme

$$J(u) + \varepsilon[a(u,v) - (f,v)] + \frac{\varepsilon^2}{2}a(v,v) \geq J(u) \tag{2.22}$$

$$\Rightarrow \quad a(u,v) - (f,v) \geq -\frac{\varepsilon}{2}a(v,v). \tag{2.23}$$

Pokud provedeme stejný postup pro $J(u - \varepsilon v) \geq J(u)$, obdržíme analogicky

$$a(u,v) - (f,v) \leq \frac{\varepsilon}{2}a(v,v). \tag{2.24}$$

Celkem jsme tedy získali následující oboustranný odhad

$$-\frac{\varepsilon}{2}a(v,v) \leq a(u,v) - (f,v) \leq \frac{\varepsilon}{2}a(v,v). \tag{2.25}$$

Jelikož v byla libovolná funkce a ε libovolné kladné číslo, dostaneme přechodem k limitě pro $\varepsilon \to 0$

$$a(u,v) = (f,v), \quad \forall v \in V, \tag{2.26}$$

tedy u řeší AVP a důkaz je hotov.

Pro ilustraci se podíváme, jak vypadá kvadratický funkcionál J pro náš úvodní příklad taženého-tlačeného prutu. Jak bylo řečeno výše, má funkcionál v aplikacích na mechaniku význam celkové potenciální energie. Ta se přitom skládá z potenciální energie vnitřní a vnější. Hustota vnitřní potenciální energie sil se rovná součinu napětí a deformace v daném bodě, hustota vnější potenciální energie pak součinu posunu a daného objemového zatížení v témže bodě (v našem příkladu je objemové zatížení reprezentováno rovnoměrným spojitým zatížením f). Posčítáním příspěvků této energie přes celý objem prutu získáme celkovou potenciální energii. Matematicky vyjádřeno tedy je

$$E_i = \int_0^l A\varepsilon\sigma \mathrm{d}x \tag{2.27}$$

$$E_e = -\int_0^l Auf \mathrm{d}x, \tag{2.28}$$

kde A je plocha prutu, E_i je celková vnitřní energie prutu a E_e jeho vnější energie. Do vnější energie by se obecně počítal i vliv předepsané síly například na konci konzoly. V našem příkladu je však síla nulová. Přitom normálovou deformaci střednice zde značíme ε, normálové napětí σ. Celková potenciální energie nosníku je potom součtem E_i a E_e

$$E = \int_0^l \varepsilon\sigma \mathrm{d}x - \int_0^l uf \mathrm{d}x. \tag{2.29}$$

Pomocí materiálových a geometrických vztahů lze (2.29) přepsat jako

$$E = \int_0^l A \underbrace{u_{,x}}_{\varepsilon} \underbrace{E u_{,x}}_{\sigma} \mathrm{d}x - \int_0^l Auf \mathrm{d}x. \tag{2.30}$$

Za funkci u lze do (2.30) dosadit libovolnou funkci, která vyhovuje kinematickým okrajovým podmínkám (tj. vyhovuje způsobu uložení konců prutu). Energie určená (2.30) však bude nabývat svého minima právě pro takovou funkci u, která vedle kinematických okrajových podmínek splňuje podmínky rovnováhy. Že tomu tak musí být, lze ověřit stejným postupem, který jsme použili v důkazu ekvivalence řešení AVP a minimalizace energetického funkcionálu. Stacionární bod funkcionálu J pak odpovídá principu virtuálních přemístění, známému z mechaniky.

2.2.2 Nesymetrický variační problém

V případě, že bilineární forma $a(u, v)$ není symetrická, nelze použít postup uvedený pro symetrický problém, neboť se v něm podstatně využívá faktu, že $a(u, v)$ je skalární součin na V. Odpověď na otázku existence i jednoznačnosti řešení AVP v obecném (nesymetrickém) případě dává Laxovo-Milgramovo lemma. Jeden způsob, jak toto lemma dokázat, se opírá o tzv. "Princip kontraktivního zobrazení", který vzhledem k jeho významu i v jiných oblastech (například numerické metody řešení rovnic) uvádíme i s důkazem v Dodatku (2.5). Tento princip, který se někdy také nazývá "Věta o pevném bodě", zhruba řečeno říká, že máme-li zobrazení na normovaném lineárním prostoru $T : V \to V$, které "zkracuje délky", pak existuje prvek $u \in V$, který se zobrazí sám na sebe a kterému se říká "pevný bod" zobrazení T. Přesněji řečeno, pokud pro všechny dvojice $v_1, v_2 \in V$ a konstantu M, $0 \leq M < 1$ platí

$$||Tv_1 - Tv_2|| \leq M||v_1 - v_2||, \tag{2.31}$$

pak existuje právě jedno $u \in V$ tak, že $Tu = u$. Nyní již můžeme uvést slíbené Laxovo-Milgramovo lemma.

Laxovo-Milgramovo Lemma. *Nechť je dán Hilbertův prostor V, na něm spojitá, koercivní bilineární forma $a(\cdot, \cdot)$ a lineární funkcionál $F(\cdot)$. Potom existuje právě jedno $u \in V$ tak, že*

$$a(u, v) = F(v), \quad \forall v \in V. \tag{2.32}$$

Důkaz
Předně si uvědomme, že pomocí bilineární formy $a(\cdot, \cdot)$ lze definovat lineární funkcionál na V. Pro pevně zvolený prvek $u \in V$ je totiž pro všechna $v \in V$ zobrazení $Au : V \mapsto R,\ Au(v) = a(u, v)$ funkcionál na V . Díky linearitě $a(\cdot, \cdot)$ snadno dostaneme, že je $Au(\cdot)$ lineární. Ze spojitosti $a(\cdot, \cdot)$ dostáváme

$$|Au(v)| = |a(u, v)| \leq M||u||_V||v||_V, \quad \forall v \in V \tag{2.33}$$

a tedy $Au(\cdot)$ je omezený, a tím pádem spojitý. Je tedy $Au \in V^*$, neboli Au patří do duálu k V. Takové zobrazení však můžeme definovat pro libovolný prvek $u \in V$. To nás přivádí k zobrazení $A : V \mapsto V^*,\ A(u) = Au = a(u, \cdot) \in V^*$, které každému prvku $u \in V$ přiřadí

příslušný funkcionál $Au \in V^*$ a je evidentně lineární. Je tedy $A \in \mathcal{L}(V, V^*)$. Podle vztahu (2.190) v Dodatku (2.6) určíme normu Au jako

$$||Au||_{V^*} = \sup_{v \in V \setminus \{0\}} \frac{|Au(v)|}{||v||_V}. \tag{2.34}$$

Podle (2.33) potom ale platí

$$||Au||_{V^*} = \sup_{v \in V \setminus \{0\}} \frac{|Au(v)|}{||v||_V} \leq M \sup_{v \in V \setminus \{0\}} \frac{||u||_V ||v||_V}{||v||_V} = M||u||_V. \tag{2.35}$$

Tím pádem ale podle vztahu (2.192) v Dodatku (2.6) je zobrazení A rovněž spojité, neboť podle (2.35) platí

$$||A||_{\mathcal{L}(V,V^*)} = \sup_{u \in V \setminus \{0\}} \frac{||A(u)||_{V^*}}{||u||_V} \leq M. \tag{2.36}$$

Nyní přijde zásadní argument celého důkazu. Podle Riezsovy věty o reprezentaci (uvedené a dokázané v Dodatku (2.6)) existuje mezi prostory V a V^* vzájemně jednoznačné zobrazení, které jsme v tomtéž dodatku označili jako $\tau : V^* \mapsto V$. K funkcionálu Au tak existuje jednoznačně určený prvek $\tau(Au) \in V$ takový, že pro všechna $v \in V$ platí

$$Au(v) = (\tau(Au), v). \tag{2.37}$$

Rovněž však i k funkcionálu F z předpokladů Laxova-Milgramova lemmatu musí existovat jednoznačně určený prvek $\tau(F) \in V$ tak, že

$$F(v) = (\tau(F), v). \tag{2.38}$$

Jelikož z definice Au pro všechna $v \in V$ platí $Au(v) = a(u, v)$, pak původní AVP problém hledání prvku $u \in V$ tak, aby $a(u, v) = F(v)$, můžeme převést na problém rovnosti

$$(\tau(Au), v) = (\tau(F), v), \quad \forall v \in V. \tag{2.39}$$

To je však ekvivalentní řešení rovnice

$$\tau(Au) = \tau(F) \quad \text{ve } V, \tag{2.40}$$

nebo ekvivalentně, jelikož τ je vzájemně jednoznačné zobrazení, rovnice

$$Au = F \quad \text{ve } V^*. \tag{2.41}$$

My budeme řešit první z těchto rovnic, a sice pomocí "Principu kontraktivního zobrazení", uvedeného a dokázaného v Dodatku (2.5). Sestrojíme proto následující zobrazení $T : V \mapsto V$ předpisem

$$Tv := v - \rho(\tau(Au) - \tau(F)), \quad \forall v \in V, \ \rho > 0. \tag{2.42}$$

Pokud by T bylo kontraktivní zobrazení, pak by mělo podle "Principu kontraktivního zobrazení" právě jeden pevný bod, tj. existoval by právě jeden prvek $u \in V$ tak, že

$$Tu = u - \rho(\tau(Au) - \tau(F)) = u, \tag{2.43}$$

neboli $\rho(\tau(Au) - \tau(F)) = 0$. Tento (jednoznačně určený) prvek $u \in V$ by byl řešením našeho AVP a byli bychom hotovi. Stačí tedy dokázat, že existuje takové $\rho > 0$, že výše definované zobrazení T je kontraktivní.

Vezměme tedy libovolné $v_1, v_2 \in V$ a označme pro zjednodušení $v = v_1 - v_2 \in V$. Potom platí

$$\begin{aligned}
||Tv_1 - Tv_2||_V^2 &= ||v_1 - v_2 - \rho(\tau Av_1 - \tau Av_2)||_V^2 \\
&= ||v - \rho(\tau Av)||_V^2 && (\tau \text{ i } A \text{ jsou lineární}) \\
&= ||v||_V^2 - 2\rho(\tau Av, v) + \rho^2||\tau Av||_V^2 && (\text{norma indukována skalárním součinem}) \\
&= ||v||_V^2 - 2\rho Av(v) + \rho^2 Av(\tau Av) && (\text{definice } \tau) \\
&= ||v||_V^2 - 2\rho a(v, v) + \rho^2 a(v, \tau Av) && (\text{definice } A) \\
&\leq ||v||_V^2 - 2\rho\alpha||v||_V^2 + \rho^2 M||v||_V||\tau Av||_V && (\text{spojitost a koercivita } a(\cdot,\cdot)) \\
&\leq (1 - 2\rho\alpha + \rho^2 M^2)||v||_V^2 && (\text{izometrie } \tau \text{ a omezenost } A) \\
&\leq C||v_1 - v_2||_V^2,
\end{aligned}$$

a tedy pro $C < 1$ je zobrazení T kontraktivní[24]. Stačí tedy najít ρ tak, aby

$$\begin{aligned}
(1 - 2\rho\alpha + \rho^2 M^2) &< 1 \\
\rho(\rho M^2 - 2\alpha) &< 0,
\end{aligned}$$

tedy $\rho \in (0, 2\alpha/M^2)$ a důkaz je hotov.

Dokázali jsme tedy, že AVP má jednoznačné řešení i v případě, že příslušná bilineární forma je nesymetrická. Můžeme se tedy věnovat otázkám, jak sestrojit přibližná numerická řešení u_h a také se ptát, zda a jak konvergují k přesnému řešení u. Tyto otázky jsou náplní zbývajících kapitol o MKP.

2.2.3 Galerkinova metoda

V této sekci se podíváme na Galerkinovu metodu, která představuje velmi vhodnou (a také obvyklou) metodu, jak získat přibližné řešení AVP. Základní idea je stejná, jako u všech numerických metod. Místo řešení spojitého problému (AVP), tedy hledání funkce $u \in V$, budeme řešit diskrétní problém, kde budeme hledat funkci $u_h \in V_h$. Přitom V_h je vhodně zkonstruovaný konečně dimenzionální (uzavřený) podprostor původního prostoru V. Jak uvidíme, díky konečné dimenzi V_h se podaří celý problém nalezení přibližného řešení u_h převést na řešení soustavy algebraických rovnic, kterou (alespoň v principu) není žádný problém vyřešit. Precizujme tedy postup diskretizace.

Každému konečně dimenzionálnímu podprostoru $V_h \subset\subset V$ přiřadíme diskrétní problém nalézt $u_h \in V_h$ tak, že platí

$$a(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_h. \tag{2.44}$$

[24]Izometrické zobrazení zachovává délky. Jako argument v poslední nerovnosti tedy používáme následující odhad: $||\tau Av||_V = ||Av||_V \leq M||v||_V$, kde poslední nerovnost plyne z omezenosti A.

Z Laxova-Milgramova lemmatu uvedeného v předchozí části plyne, že takto definovaný diskrétní problém má právě jedno řešení u_h, které nazveme diskrétním řešením. Poznamenejme, že je-li příslušná bilineární forma $a(\cdot,\cdot)$ symetrická, lze ekvivalentně řešit diskrétní problém

$$J(u_h) = \inf_{v_h \in V_h} J(v_h). \tag{2.45}$$

Tato alternativní definice diskrétního problému se nazývá Ritzova metoda.

Označme nyní $\{\varphi_1, ..., \varphi_N\}$ bázi prostoru V_h[25]. Potom lze libovolný prvek $u_h \in V_h$ vyjádřit jako lineární kombinaci prvků báze ve tvaru $u_h = \sum_{j=1}^{N} u_j \varphi_j$. K tomu, abychom určili diskrétní řešení u_h ,tedy stačí spočítat koeficienty lineární kombinace $u_j, \quad j = 1..N$. Seřaďme tyto koeficienty do vektoru $U = (u_1, ..., u_N)^T$. Dosaďme nyní $u_h = \sum_{j=1}^{N} u_j \varphi_j$ do (2.44), čímž dostaneme

$$a(\sum_{j=1}^{N} u_j \varphi_j, v_h) = (f, v_h), \quad \forall v_h \in V_h. \tag{2.46}$$

Rovnice (2.46) má platit pro všechna $v_h \in V_h$. Jelikož však každý prvek v_h lze vyjádřit pomocí prvků báze $\{\varphi_1, ..., \varphi_N\}$, lze rovnost ekvivalentně požadovat pouze pro všechny prvky báze V_h. Navíc můžeme sumu vytknout ven z bilineární formy (právě díky linearitě) a dostat tak

$$\sum_{j=1}^{N} a(\varphi_j, \varphi_i) u_j = (f, \varphi_i), \quad \forall \varphi_i, i = 1..N. \tag{2.47}$$

Pokud nyní označíme $K_{ij} = a(\varphi_j, \varphi_i)$ a $F_i = (f, \varphi_i)$, můžeme (2.47) přepsat jako

$$\sum_{j=1}^{N} K_{ij} u_j = F_i, \quad i = 1..N, \tag{2.48}$$

což není nic jiného, než soustava lineárních rovnic $KU = F$ pro hledaný vektor koeficientů lineární kombinace U. Řešení diskrétního problému (2.44) je tedy ekvivalentní řešení soustavy rovnic (2.48). Přitom pokud příslušná bilineární forma v AVP byla symetrická, je i matice K symetrická.[26] Pro libovolný vektor $V = (v_1, ..., v_N)^T$ je díky koercivitě bilineární formy $a(\cdot,\cdot)$ $V^T K V = \sum_{j=1}^{N} v_i a_{ij} v_j = a(\sum_{j=1}^{N} \varphi_j v_j, \sum_{i=1}^{N} \varphi_i v_i) \geq \alpha || \sum_{i=1}^{N} \varphi_i v_i ||_V^2 > 0$, tedy matice K je pozitivně definitní[27], a tedy regulární. Z toho plyne, že vždy existuje jednoznačně určená inverzní matice K^{-1} a soustava (2.48) je jednoznačně řešitelná. Matice K se obvykle nazývá matice tuhosti, jak je známo z předmětu NAK1.

2.2.4 Obecné úvahy o konvergenci MKP

Nyní obraťme pozornost k otázce, zda a jak konvergují diskrétní řešení u_h z konečně dimenzionálních prostorů V_h k přesnému řešení $u \in V$. Jak uvidíme v dalších kapitolách, index h

[25] Konečně dimenzionální prostor má pochopitelně konečně prvkovou bázi.

[26] To je podstatná výhoda oproti metodě konečných diferencí, kdy výsledná soustava rovnic až na výjimky symetrická není. Symetrie matice K má pozitivní důsledky při řešení soustavy (2.48).

[27] Připoměňme, že matice $A_{n\times n}$ je pozitivně definitní, pokud pro libovolný nenulový vektor x platí $x^T A x > 0$.

v označení podprostoru V_h se vztahuje k velikosti prvku příslušné konečně prvkové sítě. Čím jemnější síť je, tím menší je tato hodnota h a zároveň tím větší je dimenze podprostoru V_h. Z praktických výpočtů v předmětu NAK1 víme, že pokud danou úlohu vypočteme na jemnější síti, pak dostaneme (obvykle) přesnější výsledek. Vyvstává otázka, zda pokud půjdeme s $h \to 0+$, budeme se s diskrétním řešením u_h blížit k řešení přesnému, tedy zda $u_h \to u$. Pokud se na věc podíváme z jiného pohledu, pak jde o to, zda se zmenšujícím se h, a tedy s rostoucí dimenzí podprostorů V_h, budou tyto podprostory aproximovat prostor V tak dobře, aby

$$\lim_{h\to 0+} ||u - u_h||_V = \lim_{N\to\infty} ||u - \sum_{i=1}^{N} \varphi_i u_i||_V = 0. \tag{2.49}$$

Odhad chyby (neboli vzdálenosti přesného a diskrétního řešení ve smyslu normy prostoru V) $||u - u_h||_V$ dává následující velice důležité

Céaovo lemma. *Nechť je definován AVP a jemu přiřazený diskrétní problém (2.44). Potom existuje konstanta C nezávislá na V_h taková, že*

$$||u - u_h||_V \leq C \inf_{v_h \in V_h} ||u - v_h||_V. \tag{2.50}$$

Důkaz
Začněme následujícím velice důležitým pozorováním. V původním AVP jde o nalezení $u \in V$, aby pro všechna $v \in V$ platilo

$$a(u, v) = (f, v),$$

a tedy speciálně (jelikož má-li původní AVP platit pro všechna $v \in V$, musí platit i pro $v_h \in V_h$)

$$a(u, v_h) = (f, v_h).$$

V příslušném diskrétním problému pak hledáme $u_h \in V_h$ tak, že pro všechna $v_h \in V_h$ platí

$$a(u_h, v_h) = (f, v_h).$$

Odečtením posledních dvou rovností dostáváme následující vztah, kterému se říká tzv. Galerkinovská ortogonalita

$$a(u - u_h, v_h) = 0. \tag{2.51}$$

Galerkinovská ortogonalita (2.51) nám říká, že má li být u_h diskrétním řešením, pak musí být vektor $e_h = u - u_h$ označující chybu mezi přesným a diskrétním řešením "kolmý" na všechny prvky podprostoru V_h. Slovo kolmý jsme schválně dali do uvozovek, neboť jde o kolmost ve smyslu skalárního součinu pouze v případě, že je bilineární forma a symetrická a určuje tak skalární součin na V. I v případě, že je však nesymetrická, dává (2.51) jakýsi analogický vztah.

Z koercivity $a(\cdot, \cdot)$ nyní dostáváme

$$\alpha||u - u_h||_V^2 \leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) \leq M||u - u_h||_V||u - v_h||_V,$$

kde rovnost uprostřed jsme získali přičtením a odečtením v_h ve druhé složce a a s využitím Galerinovské ortogonality (2.51). Poslední nerovnost plyne ze spojitosti a. Po vydělení α a $||u - u_h||_V$ dostáváme

$$||u - u_h||_V \leq C||u - v_h||_V, \tag{2.52}$$

kde $C = \frac{M}{\alpha}$, což je hledaná nerovnost, která platí pro všechna $v_h \in V_h$. Tedy musí platit i pro infimum

$$||u - u_h||_V \leq C \inf_{v_h \in V_h} ||u - v_h||_V \tag{2.53}$$

a důkaz je hotov.

Nerovnost (2.50) v Céaově lemmatu je klíčová pro všechny odhady chyb, které z ní vycházejí. V konkrétních odhadech pak jde "jen" o to, jakým způsobem zkonstruujeme podprostor V_h. Céaovo lemma nám pak umožňuje převést odhad chyby $||u - u_h||_V$ na problém schopnosti prostorů V_h aproximovat původní prostor V. Jinými slovy převádí problém na vyhodnocení vzdálenosti $\mathrm{dist}(u, V_h) = \inf_{v_h \in V_h} ||u - v_h||_V$ mezi funkcí $u \in V$ a podprostorem V_h. V této souvislosti doporučujeme čtenáři, aby si prostudoval Dodatek (2.6.1), který se týká Hilbertových prostorů, a zejména pojmu kolmosti, existence nejbližšího prvku k dané množině a ortogonálních rozkladů.

Obsahem dalších odstavců bude konstrukce odhadů chyb pro konkrétní volbu podprostoru V_h. Předesíláme, že metoda konečných prvků volí tyto prostory velmi speciálně jako prostory po částech polynomiálních funkcí.

2.3 Apriorní odhad chyby

Budeme se zabývat otázkou, jak dobře dokážeme aproximovat funkci $u \in V$, která řeší AVP (2.2), pomocí libovolného v_h z konečněprvkového prostoru V_h. K zodpovězení této otázky budeme potřebovat několik vět spíše technického charakteru. Pro jednoduchost se nejdříve zaměříme na jednorozměrnou úlohu pružnosti, která již byla popsána v úvodu této kapitoly.

Ze Ceova lemmatu víme, že abychom odhadli chybu $||u - v_h||_V$, stačí pracovat s $\inf_{v_h \in V_h} ||u - v_h||_V$, neboli stačí odhadnout normu $||u - v_h||_V$ pro libovolné $v_h \in V_h$. Přirozenou volbou funkce v_h je orthogonalní projekce přesného řešení u na prostor V_h, tato volba ale není vhodná, protože zkonstruování orthogonální projekce u na prostor V_h je velice obtížné. My zvolíme za v_h interpolaci přesného řešení $\Pi(u)$, a to navíc takovou, aby její restrikce na každý element byla Lagrangeova či Hermitova[28] interpolace u, neboli

$$||u - u_h||_V \leq C||u - v_h||_V = C||u - \Pi(u)||_V \tag{2.54}$$

$$\Pi(u)|_K = \Pi_K(u|_K) \tag{2.55}$$

kde $\Pi_K(u|_K)$ značí Lagrangeovu či Hermitovu interpolaci funkce u na elementu K. Výhodou takto zvolené interpolace je, že chybu řešení můžeme zapsat pomocí odhadů chyby na jednotlivých elementech

$$||u - u_h||_{1,\Omega} \leq C||u - \Pi(u)||_{1,\Omega} = C\left(\sum_K ||u - \Pi(u)||^2_{1,K}\right)^{\frac{1}{2}} = \tag{2.56}$$

$$= C\left(\sum_K ||u - \Pi_K(u|_K)||^2_{1,K}\right)^{\frac{1}{2}} \tag{2.57}$$

[28] Zjednodušeně řečeno Lagrangeova interpolace využívá k interpolaci pouze uzlových hodnot, zatímco Hermitova i jejich derivací.

Stačí nám tedy odhadnout chybu řešení lokálně na každém prvku, což je velice výhodné a je možné odhadnout chybu pouze na referenčním konečném prvku a tento odhad ztransformovat na prvek skutečný. V další části se proto budeme zabývat chybou interpolace na konečném prvku. Začneme s matematickou definicí konečného prvku.

2.3.1 Konečný prvek

Definice Konečný prvek v $\mathbb{R}^n$ je trojice (K, Σ, P), kde

- $K \subset \mathbb{R}^n$ s lipschitzovskou[29] hranicí a neprázdným vnitřkem
- Σ je množina N spojitých lineárních forem $\mathbb{L}_i : C^\infty(K) \mapsto \mathbb{R}^n, 1 \leq i \leq N$
- P je prostor polynomů s bází $\varphi_i, i = 1, 2, ..., N$.

Pojem konečného prvku objasníme pomocí lineárního Lagrangeovského prvku v jedné dimenzi

Definice *Lineární Lagrangeovký 1D prvek*

- K: Úsečka s krajními body A a B
- Σ je množina tvořena dvěma lineárními formami
  - $\mathbb{L}_1(u) = u(A)$
  - $\mathbb{L}_2(u) = u(B)$

  které nám funkci u zobrazí na hodnoty funkce u v krajních bodech A a B. Lineárním formám $\mathbb{L}_i$ říkáme stupně volnosti.
- P: Prostor polynomů prvního stupně $P_1([A, B])$, s bázovými funkcemi
  - $\varphi_1 = \dfrac{x_B - x}{x_B - x_A}$
  - $\varphi_2 = \dfrac{x - x_A}{x_B - x_A}$

  Bázové funkce φ_i jsou definovány tak, aby

$$\mathbb{L}_i(\varphi_j) = \delta_{ij} \tag{2.58}$$

neboli

$$\mathbb{L}_1(\varphi_1) = \varphi_1(x_A) = \frac{x_B - x_A}{x_B - x_A} = 1 \tag{2.59}$$

$$\mathbb{L}_1(\varphi_2) = \varphi_2(x_A) = \frac{x_A - x_A}{x_B - x_A} = 0 \tag{2.60}$$

[29] pro vysvětlení výrazu lipschitzovská hranice odkazujeme na přednášky z Matematiky 4

Stejným způsobem se můžeme přesvědčit o platnost vztahu (2.58) pro druhou bázovou funkci, tuto úlohu přenecháváme čtenáři jako cvičení. Interpolaci Π_K funkce v můžeme pomocí bázových funkcí φ_1, φ_2 a stupňů volnosti $\mathbb{L}_1$, $\mathbb{L}_2$ zapsat jako

$$\Pi_K(v) = L_1(v)\varphi_1 + L_2(v)\varphi_2 = v(x_A)\varphi_1 + v(x_B)\varphi_2 \tag{2.61}$$

tedy součet hodnoty funkce v v bodě A přenásobené první bázovou funkcí a hodnoty funkce v v bodě B přenásobené druhou bázovou funkcí. Z definice Π_K vyplívá

$$\Pi_K(p) = p \quad \forall p \in P_1(K) \tag{2.62}$$

neboli interpolace zachovává polynomy stupně jedna.

Příkladem Hermitovského prvku je nosníkový prvek založený na Navierově-Bernoulliho hypotéze

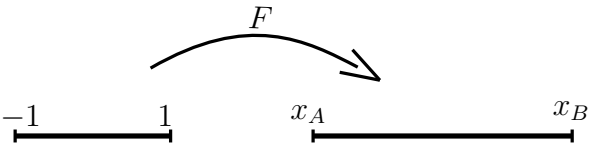
Definice *Nosníkový prvek*

- K: Úsečka s krajními body A a B
- Σ je množina tvořena čtyřmi lineárními formami
  - $\mathbb{L}_1(w) = w(A)$
  - $\mathbb{L}_2(w) = w(B)$
  - $\mathbb{L}_3(w) = w_{,x}(A)$
  - $\mathbb{L}_4(w) = w_{,x}(B)$

  funkcemi průhybu w v bodech A a B a jejich derivacemi w, x, které mají význam potočení.
- P: Prostor polynomů třetího stupně $P_3([A, B])$, s bázovými funkcemi
  - $\varphi_1 = 1 - \frac{(3x^2)}{(x_B - x_A)^2} + \frac{(2x^3)}{(x_B - x_A)^3}$
  - $\varphi_2 = \frac{(3x^2)}{(x_B - x_A)^2} - \frac{(2x^3)}{(x_B - x_A)^3}$
  - $\varphi_3 = x - \frac{2x^2}{(x_B - x_A)} + \frac{x^3}{(x_B - x_A)^2}$
  - $\varphi_4 = -\frac{x^2}{(x_B - x_A)} + \frac{x^3}{(x_B - x_A)^2}$

2.3.2 Referenční konečný prvek v 1D

Koncept referenčího prvku je velice užitečný jak při počítačové implementaci, tak při odhadování chyby řešení, neboť nám umožňuje pracovat pouze s referenčním prvkem a následně odhady chyb převádět na prvek skutečný. V případě tlačeného-taženého prvku zavedeme přirozenou souřadnici ξ, která je rovna -1 v uzlovém bodě A a 1 v uzlovém bodě B. Hodnotu skutečné souřadnice x vypočítáme pomocí zobrazení $F : \hat{K} \to K$, kde $\hat{K}$ značí referenční prvek a K značí skutečný prvek, Zobrazení F je definováno pomocí přirozené souřadnice ξ a souřadnic uzlových bodů A a B se souřadnicemi x_A a x_B:



Obr. 2.2: 1D referenční prvek

$$x = F(\xi) = \frac{1}{2}(1-\xi)x_A + \frac{1}{2}(1+\xi)x_B \tag{2.63}$$

Tento vztah je také možné přepsat jako

$$F(\xi) = N_1(\xi)x_A + N_2(\xi)x_B \tag{2.64}$$

kde

$$N_1(\xi) = \frac{1}{2}(1-\xi) \tag{2.65}$$

a

$$N_2(\xi) = \frac{1}{2}(1+\xi) \tag{2.66}$$

jsou dobře známé bázové funkce lineárního tlačeného-taženého prvku v přirozených souřadnicích. Dosazením můžeme ověřit hodnotu souřadnice x v uzlových bodech

$$x(-1) = \frac{1}{2}(1+1)x_A + \frac{1}{2}(1-1)x_B = x_A \tag{2.67}$$

$$x(1) = \frac{1}{2}(1-1)x_A + \frac{1}{2}(1+1)x_B = x_B \tag{2.68}$$

a uprostřed prvku

$$x(0) = \frac{1}{2}(1)x_A + \frac{1}{2}(1)x_B = \frac{x_A + x_B}{2} \tag{2.69}$$

2.3.3 Odhad chyby aproximace pro 1D úlohu

Přikročíme k formulaci věty o odhadu seminormy funkce na skutečném prvku pomocí seminormy na prvku referenčním a naopak

Věta 2.2. *Pro libovolnou funkci* $v \in H^m(K)$ *platí*

$$|\hat{v}|_{m,\hat{K}} \leq C\left(\frac{h_k}{2}\right)^{m-\frac{1}{2}} |v|_{m,K} \tag{2.70}$$

$$|v|_{m,K} \leq C\left(\frac{2}{h_K}\right)^{m-\frac{1}{2}} |\hat{v}|_{m,\hat{K}} \tag{2.71}$$

kde K značí skutečný prvek, $\hat{K}$ značí prvek referenční a $\hat{v} = v \circ F$

Důkaz: K důkazu této věty použijeme větu o substituci s pravidlem o derivaci složené funkce. Nejdříve vyjádříme derivaci funkce $\hat{v}$ vzhledem k souřadnicím ξ na referenčním prvku pomocí derivace v podle x

$$\frac{\mathrm{d}\hat{v}}{\mathrm{d}\xi} = \frac{\mathrm{d}v}{\mathrm{d}x} J \tag{2.72}$$

kde J je Jakobián zobrazení (2.63) definovaný jako

$$J = \frac{\mathrm{d}x}{\mathrm{d}\xi} = \frac{x_1}{2} + \frac{x_2}{2} = \frac{h_K}{2} \tag{2.73}$$

kde h_K značí délku prvku K. Obecně můžeme vyjádřit m-tou derivaci funkce $\hat{v}$ pomocí derivace funkce v

$$\frac{\mathrm{d}^m\hat{v}}{\mathrm{d}\xi^m} = \frac{\mathrm{d}^m v}{\mathrm{d}x^m} J^m \tag{2.74}$$

Po dosazení rovnice (2.73) do (2.74) dostáváme vyjádření derivace $\hat{v}$ na referenčním prvku pomocí derivace v na skutečném prvku

$$\frac{\mathrm{d}^m\hat{v}}{\mathrm{d}\xi^m} = \frac{\mathrm{d}^m v}{\mathrm{d}x^m} \left(\frac{h_K}{2}\right)^m \tag{2.75}$$

Můžeme přejít k odhadu L^2-normy m-té derivace funkce $\hat{v}$ pomocí L^2-normy m-té derivace funkce v. Vyjdeme z definice L^2-normy, do které dosadíme rovnici (2.75) a s využitím věty o substituci převedeme integrál z prvku $\hat{K}$ na prvek K. K tomu budeme opět potřebovat Jakobián zobrazení z referenčního na skuteční prvek, který nám udává vztah mezi diferenciály souřadnic

$$\mathrm{d}x = J\mathrm{d}\xi = \frac{h_K}{2}\mathrm{d}\xi \tag{2.76}$$

Nakonec dostáváme

$$\sqrt{\int_{-1}^{1} \left(\frac{\mathrm{d}^m\hat{v}}{\mathrm{d}\xi^m}\right)^2 \mathrm{d}\,\xi} = \sqrt{\int_A^B \left(\frac{\mathrm{d}^m v}{\mathrm{d}x^m}\right)^2 \left(\frac{h_K}{2}\right)^{2m} \frac{2}{h_K}\mathrm{d}x} \tag{2.77}$$

Z pravé strany rovnice vytkneme členy obsahující h_K

$$\sqrt{\int_{-1}^{1} \left(\frac{\mathrm{d}^m\hat{v}}{\mathrm{d}\xi^m}\right)^2 \mathrm{d}\,\xi} = \sqrt{\left(\frac{h_K}{2}\right)^{2m} \frac{2}{h_K}} \sqrt{\int_A^B \left(\frac{\mathrm{d}^m v}{\mathrm{d}x^m}\right)^2 \mathrm{d}x} \tag{2.78}$$

integrál na levé i na pravé straně můžeme zapsat pomocí seminormy

$$|\hat{v}|_{m,\hat{K}} = \left(\frac{h_K}{2}\right)^{m-\frac{1}{2}} |v|_{m,K} \tag{2.79}$$

což je první tvrzení s $C = 1$.
Při důkazu druhého tvrzení postupujeme zcela obdobně, pouze vyměníme $\hat{K}$ za K a naopak.

$$\sqrt{\int_A^B \left(\frac{\mathrm{d}^m v}{\mathrm{d}x^m}\right)^2 \mathrm{d}\,x} = \sqrt{\int_{-1}^{1} \left(\frac{\mathrm{d}^m\hat{v}}{\mathrm{d}\xi^m}\right)^2 \left(\frac{2}{h_K}\right)^{2m} \frac{h_K}{2}\mathrm{d}\xi} \tag{2.80}$$

Z pravé strany rovnice vytkneme členy obsahující h_K

$$\sqrt{\int_A^B \left(\frac{\mathrm{d}^m v}{\mathrm{d}x^m}\right)^2 \mathrm{d}\,x} = \sqrt{\left(\frac{2}{h_K}\right)^{2m} \frac{h_K}{2}} \sqrt{\int_{-1}^1 \left(\frac{\mathrm{d}^m \hat{v}}{\mathrm{d}\xi^m}\right)^2 \mathrm{d}\xi} \tag{2.81}$$

což můžeme pomocí seminorem přepsat jako

$$|v|_{m,K} = \left(\frac{h_K}{2}\right)^{\frac{1}{2}-m} |\hat{v}|_{m,\hat{K}} \tag{2.82}$$

dostáváme druhé tvrzení s $C = 1$.

Ještě než přejdeme k vyslovení věty pro odhad chyby aproximace, budeme potřebovat následující větu:

Věta 2.3. *Existuje konstanta $C > 0$ taková, že pro libovolnou funkci $\hat{v} \in H^{k+1}(\hat{K})$ platí*

$$\inf_{p \in P_k(\hat{K})} ||\hat{v} + p||_{k+1,\hat{K}} \leq C|\hat{v}|_{k+1,\hat{K}} \tag{2.83}$$

pokud tedy k funkci $\hat{v} \in H^{k+1}$ přičteme libovolný polynom stupně k, můžeme H^{k+1}-normu tohoto součtu odhadnout pomocí H^{k+1} seminormy funkce $\hat{v}$.

Důkaz tohoto tvrzení je poměrně složitý, proto ho zde neuvádíme a čtenáře odkazujeme na odbornou literaturu. Nyní jsme již připraveni k vyslovení klíčové věty pro odhad chyby aproximace konečného prvku.

Věta 2.4. *Nechť interpolace $\Pi_{\hat{K}}$ zachovává polynomy k-tého stupně, neboli*

$$\Pi_{\hat{K}}(p) = p \quad \forall p \in P_k(\hat{K}) \tag{2.84}$$

pak platí

$$|v - \Pi_K(v)|_{1,K} \leq C h_K^k |v|_{k+1,K} \quad \forall v \in H^{k+1}(K) \tag{2.85}$$

Důkaz: Prvním krokem důkazu věty bude přetransformování seminormy $|v - \Pi(v)|_{1,K}$ na referenční prvek použitím věty (2.2)

$$|v - \Pi_K(v)|_{1,K} \leq C \left(\frac{2}{h_K}\right)^{m-\frac{1}{2}} |v - \Pi_{\hat{K}}(\hat{v})|_{1,\hat{K}} \tag{2.86}$$

na pravou stranu (2.86) přičteme a odečteme polynom p stupně k a dále využijeme toho, že interpolační operátor zachovává polynomy k-tého stupně

$$v - \Pi_{\hat{K}}(\hat{v}) = v + p - p - \Pi_{\hat{K}}(\hat{v}) = \hat{v} + p - \Pi_{\hat{K}}(\hat{v} + p) = (I - \Pi_{\hat{K}})(\hat{v} + p) \tag{2.87}$$

což platí pro všechna $\hat{v} \in H^{k+1}(\hat{K})$ a $p \in P_k(\hat{K})$. I značí identitu, jakožto spojité zobrazení z $H^{k+1}(\hat{K})$ do $H^1(\hat{K})$. Následovně dosadíme (2.87) do (2.86)

$$|v - \Pi_K(v)|_{1,K} \leq C \left(\frac{2}{h_K}\right)^{m-\frac{1}{2}} |(I - \Pi_{\hat{K}})(\hat{v} + p)|_{1,\hat{K}} \tag{2.88}$$

seminormu z pravé strany (2.88) odhadneme pomocí H^1-normy

$$|(I-\Pi_{\hat{K}})(v+p)|_{1,\hat{K}} \leq ||(I-\Pi_{\hat{K}})(v+p)||_{1,\hat{K}} \tag{2.89}$$

Při odhadu pravé strany vyjdeme z definice normy spojitého zobrazení

$$||(I-\Pi_{\hat{K}})||_{\mathcal{L}(H^{k+1}(\hat{K})),(H^1(\hat{K}))} = \sup_{||\eta||\neq 0} \frac{||(I-\Pi_{\hat{K}})(\eta)||_{1,\hat{K}}}{||\eta||_{k+1,\hat{K}}} \quad \forall \eta \in H^{k+1}(\hat{K}) \tag{2.90}$$

Jistě platí

$$||(I-\Pi_{\hat{K}})||_{\mathcal{L}(H^{k+1}(\hat{K})),(H^1(\hat{K}))} = \sup_{||\eta||\neq 0} \frac{||(I-\Pi_{\hat{K}})(\eta)||_{1,\hat{K}}}{||\eta||_{k+1,\hat{K}}} \geq \frac{||(I-\Pi_{\hat{K}})(\hat{v}+p)||_{1,\hat{K}}}{||\hat{v}+p||_{k+1,\hat{K}}} \quad \forall \eta \in H^{k+1}(\hat{K}), \tag{2.91}$$

po vynásobení $||\hat{v}+p||$ dostáváme

$$||(I-\Pi_{\hat{K}})||_{\mathcal{L}(H^{k+1}(\hat{K})),(H^1(\hat{K}))}||\hat{v}+p||_{1,\hat{K}} \geq ||(I-\Pi_{\hat{K}})(\hat{v}+p)||_{1,\hat{K}} \tag{2.92}$$

tento výsledek dosadíme do (2.89) a odhadneme seminormu rozdílu funkce v a její interpolace

$$|\hat{v}-\Pi_{\hat{K}}(\hat{v})|_{1,\hat{K}} \leq ||(I-\Pi_{\hat{K}})(\hat{v}+p)||_{1,\hat{K}} \leq ||I-\Pi_{\hat{K}}||_{\mathcal{L}(H^{k+1}(\hat{K})),(H^1(\hat{K}))}||v+p||_{k+1,(\hat{K})} \quad \forall p \in P_k(\hat{K}) \tag{2.93}$$

jelikož tato nerovnost platí pro všechny polynomy stupně k, jistě platí jistě i pro jejich infimum a nerovnost můžeme přepsat jako

$$|v-\Pi_{\hat{K}}(v)|_{1,\hat{K}} \leq C||I-\Pi||_{\mathcal{L}_{(H^{k+1}(\hat{K})),(H^1(\hat{K}))}} \inf_{\forall p \in P_k(\hat{K})} ||v+p||_{k+1,\hat{K}} \tag{2.94}$$

Normu $||(I-\Pi)||_{\mathcal{L}_{(H^2(\hat{K})),(H^1(\hat{K}))}}$ zahrneme do konstanty C

$$|\hat{v}-\Pi_{\hat{K}}(\hat{v})|_{1,\hat{K}} \leq \hat{C} \inf_{\forall p \in P_k(\hat{K})} ||\hat{v}+p||_{k+1,\hat{K}} \tag{2.95}$$

Pravou stranu odhadneme pomocí věty (2.3)

$$|\hat{v}-\Pi_K(\hat{v})|_{1,\hat{K}} \leq C|\hat{v}|_{k+1,\hat{K}} \tag{2.96}$$

a tento výsledek dosadíme do (2.86)

$$|v-\Pi_K(v)|_{1,K} \leq C\left(\frac{2}{h_K}\right)^{m-\frac{1}{2}} |\hat{v}|_{k+1,\hat{K}} \tag{2.97}$$

posledním krokem je převedení pravé strany zpět na element K, což můžeme provést použitím odhadu (2.70)

$$|v-\Pi_K(v)|_{1,K} \leq C\left(\frac{2}{h_K}\right)^{m-\frac{1}{2}} |v|_{k+1,\hat{K}} < C\left(\frac{h_K}{2}\right)^{k} |v|_{k+1,K} \tag{2.98}$$

Dokázali jsme tedy, že chyba interpolace prvku je řádu $\mathcal{O}(h_K^k)$, abychom dostali chybu aproximace řešení, stačí za funkci v dosadit u, tedy řešení AVP(2.2) a posčítat interpolační chyby přes všechny prvky, viz (2.56). Výsledná chyba má tvar

$$|(u - \Pi(u))|_{1,\Omega} \leq Ch^k |u|_{k+1,\Omega} \quad \forall u \in H^{k+1}(\Omega) \tag{2.99}$$

kde $h = max_K h_k$ a Ω je oblast na které řešíme AVP, v tomto případě interval $[0, L]$. Výsledná chyba aproximace metody konečných prvků je $\mathcal{O}(h^k)$, kde h je délka nejdelšího prvku a parametr k souvisí s regularitou přesného řešení. Pokud je přesné řešení $u \in H^2(\Omega)$ dostáváme

$$|(u - \Pi(u))|_{1,\Omega} \leq Ch|u|_{2,\Omega} \tag{2.100}$$

Tento odhad vypadá na první pohled hůře než odhad, který jsme dostali v metodě konečných diferencí. Musíme si ale uvědomit, že jsme nedostali odhad chyby aproximace funkce u, ale její derivace podle prostorové proměnné, která v má v našem případě fyzikální význam deformace. Dostali jsme tedy chybu aproximace deformace řádu $\mathcal{O}(h)$. Otázkou je, jak odhadnout chybu aproximace pro posuny, tedy odhad v L^2 normě. K tomu slouží takzvaný Aubinův-Nietscheho trik.

Věta 2.5. *Hledáme $\phi_g \in H^1$ tak, že pro $g \in L^2$ platí*

$$a(v, \phi_g) = (g, v) \quad \forall v \in H^1 \tag{2.101}$$

kde a je bilineární forma definovaná v AVP(2.2). Pak

$$||u - u_h||_0 \leq M||u - u_h||_1 \sup_{g \in L^2} \left(\frac{1}{||g||_0} \inf_{v_h \in V_h} ||\phi_g - v_h|| \right) \tag{2.102}$$

Tento problém je tzv. duální k AVP(2.2).

Důkaz: Zopakujme problém duálním k AVP

$$a(v, \phi) = (g, v) \tag{2.103}$$

který platí pro $\forall v \in H^1$, musí tedy platit i pro řešení AVP u, my však zkusíme za v dosadit rozdíl mezi přesným a přibližným řešením $u - u_h$. Ještě připomeneme tzv. Galerkinovskou ortogonalitu: $a(u - u_h, v_h) = 0 \quad v_h \in V_h$

$$(g, u - u_h)_H = a(u - u_h, \phi) = a(u - u_h, \phi) + a(u - u_h, v_h) = \tag{2.104}$$

$$= a(u - u_h, \phi - v_h) \leq M||u - u_h||_1 ||\phi_g - v_h||_1 \tag{2.105}$$

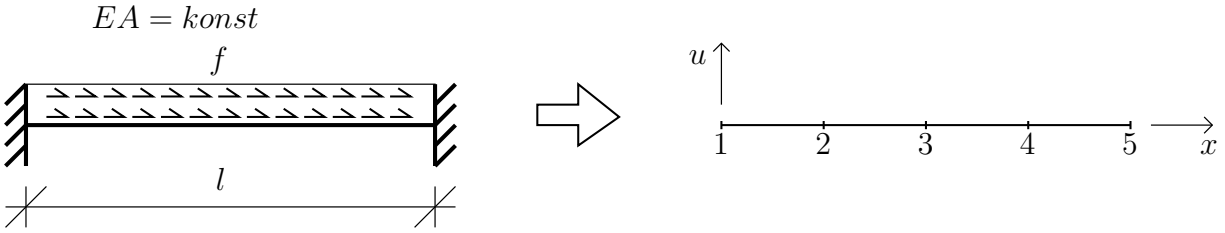
AN

kde M je konstanta spojitosti bilineární formy $a(\cdot, \cdot)$. Nyní vyjádříme normu $||u - u_h||_0$ jako

$$||u - u_h||_0 = \sup_{g \in L^2} \frac{(g, u - u_h)}{||g||_1} \tag{2.106}$$

a za $(g, u - u_h)$ dosadíme odhad (**??**)

$$||u - u_h||_0 = M||u - u_h||_1 \sup_{g \in H} \left(\frac{1}{||g||_0} \inf_{v_h \in V_h} ||\phi_g - v_h||_V \right) \tag{2.107}$$



Obr. 2.3: Zadání

Věta 2.6. Odhad chyby v L^2 normě *Necht je adjungovaný variační problém regulární, neboli*

- $\phi_g \in H^2([0,L]) \quad \forall g \in L^2([0,L])$
- $\exists C > 0 : ||\phi_g||_{2,([0,L])} \leq C||g||_{0,([0,L])} \quad \forall g \in L^2([0,L])$

Potom

$$||u - u_h||_{0,\Omega} \leq Ch^{k+1}|u|_{k+1,\Omega} \tag{2.108}$$

Důkaz: K důkazu této věty nejdříve odhadneme poslední člen rovnice (2.107) pomocí rovnice (2.155)

$$\inf_{v_h \in V_h} ||\phi_g - v_h||_{1,\Omega} \leq ||\phi_g - \Pi(\phi_g)||_{1,\Omega} \leq Ch|\phi_g|_{2,([0,L])} \leq Ch||g||_{0,([0,L])} \tag{2.109}$$

kde poslední odhad plyne z regularity adjungovaného variačního problému. S tímto výsledkem a s využitím Aubinova-Nietzcheho triku již není problém dokázat odhad chyby aproximace v L^2 normě. Zopakujme výsledek (2.107)

$$||u - u_h||_{0,\Omega} \leq M||u - u_h||_1 \sup_{g \in L^2} \left(\frac{1}{||g||_0} \inf_{v_h \in V_h} ||\phi_g - v_h||_V \right) \tag{2.110}$$

ve kterém odhadneme poslední člen pomocí (2.109) a následně použijeme odhad (2.155)

$$||u - u_h||_{0,\Omega} \leq Ch||u - u_h||_{1,[(0,L)]} \leq Ch^{k+1}||u - u_h||_{k+1,[(0,L)]} \tag{2.111}$$

Jak vidíme, tak nakonec dostáváme stejný řád aproximace jako v MKD. Musíme však poznamenat, že tento odhad platí pro funkce $u \in H^{k+1}$. Otázka hladkosti řešení je poměrně komplikovaný problém, který závisí na celé řadě faktorů, jako je například hladkost pravé strany (zatížení), hladkost okrajovým podmínek atd. Tyto informace nám umožňují volit teoreticky optimální konstrukci prostotu V_h. Jak jsme ukázali, pokud v je $u \in H^2$, je chyba aproximace posunů řádu $\mathcal{O}(h^2)$ a deformací $\mathcal{O}(h)$ a nemá smysl používat jiné než lineární konečné prvky, pokud je však $u \in H^3$ chyba aproximace je řádu $\mathcal{O}(h^3)$ a použití kvadratických prvků může být výhodné.

2.3.4 Příklad

Použití metody konečných prvků ilustrujeme na přikladu taženého prutu, který byl řešen v úvodu kapitoly popisující MKD, zadání viz Obr.1.3. Rovnice popisující tento problém má tvar

$$EAu_{,xx} + f_x = 0 \tag{2.112}$$

s okrajovými podmínkami

$$u(0) = 0 \quad u(L) = 0 \tag{2.113}$$

Příslušný Abstraktní variační problém zní

$$\int_O^L EAu_{,x}v_{,x}\mathrm{d}x + \int_O^L f_x v\mathrm{d}x = 0 \quad \forall v \in H_0^1([0,L]) \tag{2.114}$$

Vyřešíme tento problém metodou konečných prvků. Budeme tedy hledat $u_h \in V_h$, kde V_h je nějaký vhodně zkonstruovaný podprostor prostoru H_0^1. K sestrojení V_h využijeme lineární Lagrangeovský konečný prvek popsaný v předchozí kapitole. Rozdělme interval [0,L] ekvidistantně na $N+1$ prvků s dělícími body $x_0, x_1, ..., x_{N+1}$, které nazýváme uzly a $h = \frac{L}{N+1}$ značí délku prvku. Zaveďme funkce $\varphi_i(x)$

$$\varphi(x) = \begin{cases} 0 \qquad \textit{pokud } x \in (0, x_{i-1}) \\ \frac{x-x_{i-1}}{h} \quad \textit{pokud } x \in (x_{i-1}, x_i) \\ \frac{x_{i+1}-x_i}{h} \quad \textit{pokud } x \in (x_i, x_{i+1}) \\ 0 \quad \textit{pokud } x \in (x_{i+1}, L) \end{cases} \tag{2.115}$$

a konečněprvkový prostor V_h definujeme jako lineární obal funkcí φ_i. Definic konečně prvkového prostoru můžeme zapsat jako $V_h = \{v_h \in C([0,L]); v_h|_{x_{i-1},x_i} \in P_1([x_{i-1}, x_i]), v_h(0) = v_h(L) = 0\} \quad \forall i = 1, 2, ..., N$, je to tedy prostor spojitých funkcí na intervalu $[0, L]$ s nulovými hodnotami na kraji intervalu, jejichž restrikce na jednotlivé elementy jsou lineární polynomy. Jelikož funkce φ tvoří bázi prostoru V_h, můžeme psát, viz kapitola o Galerkinově metodě(**??**)

$$\sum_{j=1}^{N} a(\varphi_i, \varphi_j)u_j = (f, \varphi_j) \quad \forall \varphi_i = 1, 2, ..., N \tag{2.116}$$

K řešení úlohy použijeme stejnou síť jako MKD. Spočtěme členy $a(\varphi_i, \varphi_j)$. Díky volbě bázových funkcí bude

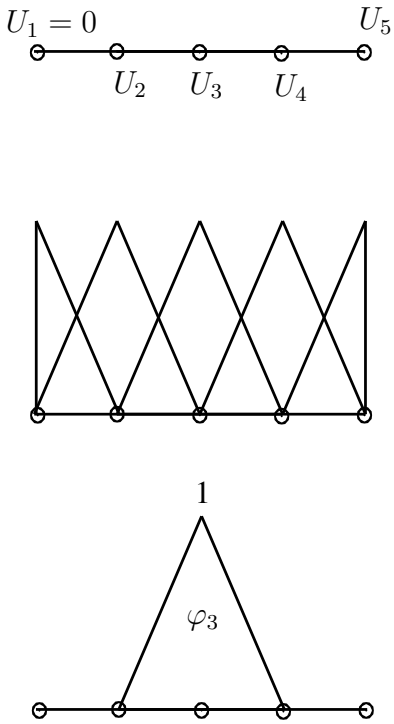
$$a(\varphi_{i+k}, \varphi_i) = 0 \quad \forall k > 1 \tag{2.117}$$

a

$$a(\varphi_{i-k}, \varphi_i) = 0 \quad \forall k > 1 \tag{2.118}$$

což je patrné z obrázku (**??**). Pro výpočet matice tuhosti stačí spočítat tři různé členy a pravou stranu

- $$a(\varphi_{i+1}, \varphi_i) = EA \int_{x_i}^{x_{i+1}} -\frac{1}{h}\frac{1}{h} = -\frac{EA}{h} \tag{2.119}$$



Obr. 2.4: Bázové funkce

- $$a(\varphi_i, \varphi_i) = EA \int_{x_i}^{x_{i+1}} \frac{1}{h}\frac{1}{h} + EA \int_{x_{i-1}}^{x_i} \frac{1}{h}\frac{1}{h} = 2\frac{EA}{h} \tag{2.120}$$

- $$a(\varphi_{i-1}, \varphi_i) = EA \int_{x_{i-1}}^{x_i} \frac{1}{h}\frac{1}{h} = -\frac{EA}{h} \tag{2.121}$$

- $$(f, \varphi_i) = \int_{x_{i-1}}^{x_i} = f\varphi_i = fh \tag{2.122}$$

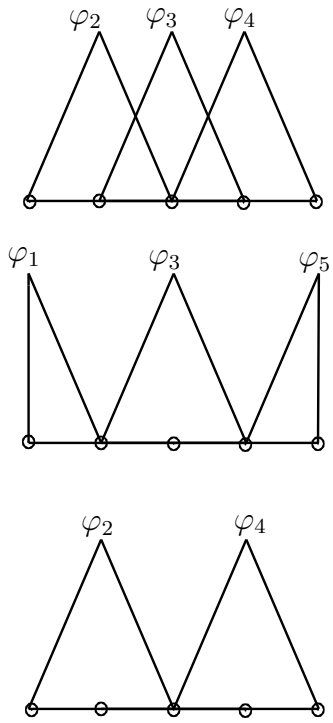
a nakonec dostáváme

$$\frac{EA}{h}\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}\begin{pmatrix} U_2 \\ U_3 \\ U_4 \end{pmatrix} = \begin{pmatrix} fh \\ fh \\ fh \end{pmatrix}. \tag{2.123}$$

Vidíme, že jsme obdrželi stejnou soustavu jako v metodě konečných diferencí a navíc i při použití metody konečných prvků dostáváme v uzlech sítě přesné řešení.

To však platí jen v jednodimenzionálním případě při použití lineárních konečných prvků. Obecně vede metoda konečných prvků na jinou matici než metoda sítí, navíc v obecných případech není při výpočtu metodou konečných prvků rovnice v uzlech splněna přesně. Na obrázku xz vidíme chybu řešení v různých normách.

Na ose x je vynesen počet prvků sítě a na ose y je chyba řešení v semilogaritmickém měřítku. Je vidět, že L^2-norma přibližného řešení konverguje k přesnému řešení kvadraticky,



Obr. 2.5

neboli při použití dvojnásobného počtu prvků se chyba řešení zmenší čtyřikrát. V druhém případě je vykreslena chyba deformací. Zde již nedostáváme tak rychlou konvergenci, což jsme však očekávali a je to zcela v souladu s odvozenými výsledky.

Otázkou stále zůstává, zda odhady, které jsme dokázali pro jednodimenzionální úlohu jsou platné i pro obecnější úlohy. Zaměříme se na dvoudimenzionální úlohu. Nejprve zavedeme lineární Lagrangeovský trojúhelníkový prvek

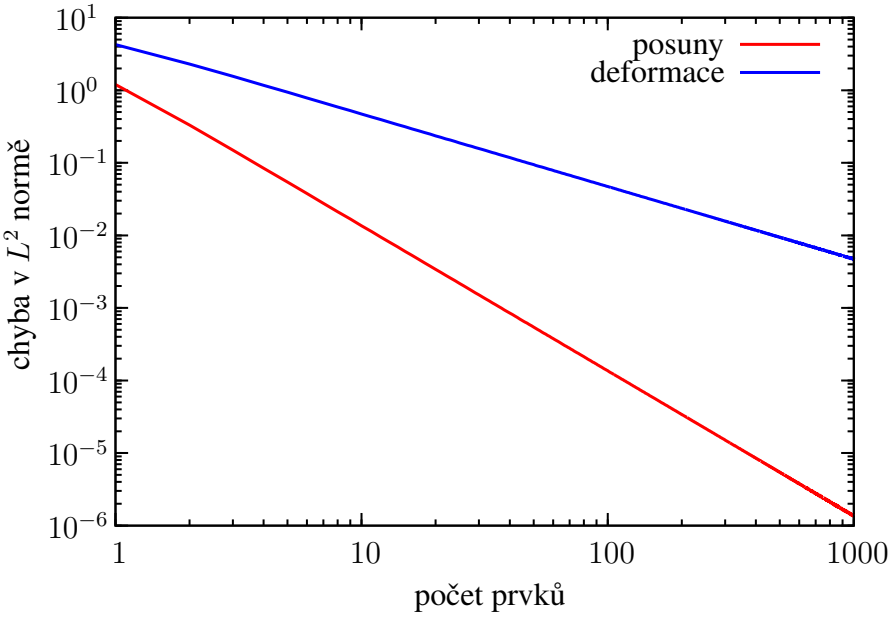
Definice *Lineární trojúhelníkový konečný prvek*

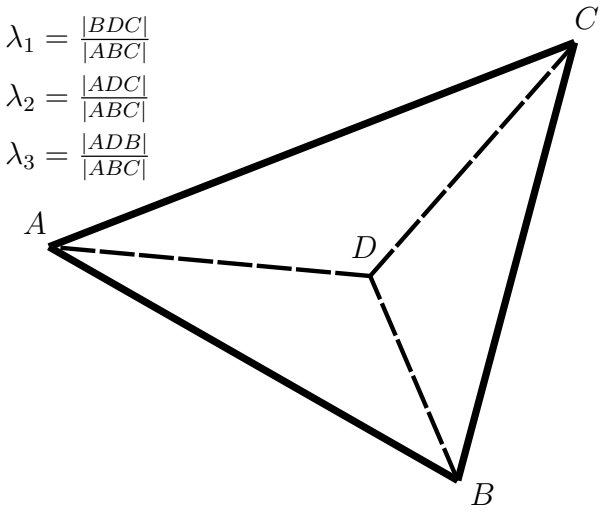
- K je trojúhelník s vrcholy A, B a C
- Σ je množina tvořena třemi lineárními formami $\mathbb{L}_1(u) = u(A) \quad \mathbb{L}_2(u) = u(B) \quad \mathbb{L}_3(u) = u(C)$
- P: Prostor polynomů prvního stupně $P_1(K)$. Bázové funkce lineárního trojúhelníkového prvku zavedeme v části věnované referenčnímu trojúhelníkovému prvku.

2.3.5 Referenční trojúhelníkový konečný prvek

Začneme s takzvanými barycentrickými souřadnicemi.[30], které jsou bázovými funkcemi lineárního Lagrangeovského referenčního trojúhelníkového prvku.

[30]barycentrické souřadnice se také někdy nazývají plošné souřadnice pro jejich geometrický význam viz (**??**)





Obr. 2.6: Barycentrické souřadnice: Geometrický význam

Barycentrické souřadnice $\lambda_1(x)$,$\lambda_2(x)$ a $\lambda_3(x)$ libovolného bodu $x \in \mathbb{R}^2$ vzhledem k vrcholu A,B a C jsou definovány vztahy

$$\sum_{j=1}^{3} \lambda_j(x) x_j = x \quad \sum_{j=1}^{3} \lambda_j(x) = 1 \tag{2.124}$$

Platí, že $\lambda_1, \lambda_2, \lambda_3 \in P_1$ a $\lambda_i(a_j) = \delta_{ij} \quad i,j = 1,2,3$ Zobrazení F (**??**) z referenčního konečného prvku na prvek skutečný je definované následovně

$$x = F_1(\lambda_1, \lambda_2) = x_A + \lambda_1(x_B - x_A) + \lambda_2(x_C - x_A) \tag{2.125}$$

$$y = F_2(\lambda_1, \lambda_2) = y_A + \lambda_1(y_B - y_A) + \lambda_2(y_C - y_A) \tag{2.126}$$

F můžeme zapsat jako

$$F = B_K \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \tag{2.127}$$

kde B_K je matice definovaná jako

$$B_K = \begin{pmatrix} x_B - x_A & x_C - x_A \\ y_B - y_A & y_C - y_A \end{pmatrix} \tag{2.128}$$

Věta 2.7. *Normu matice B_K lze odhadnout pomocí průměru elementu h_K a $\hat{\rho}$, tedy poloměru koule vepsané do referenčního trojúhelníka $\hat{K}$.*

$$||B_K|| \leq \frac{h_k}{\hat{\rho}} \tag{2.129}$$

Normu inverzni matice B_K^{-1} lze naopak odhadnout pomocí průmeru referenčního prvku $\hat{K}$ a ρ, což je poloměr koule vepsané do prvku K.

$$||B_K^{-1}|| \leq \frac{\rho}{\hat{h}} \tag{2.130}$$

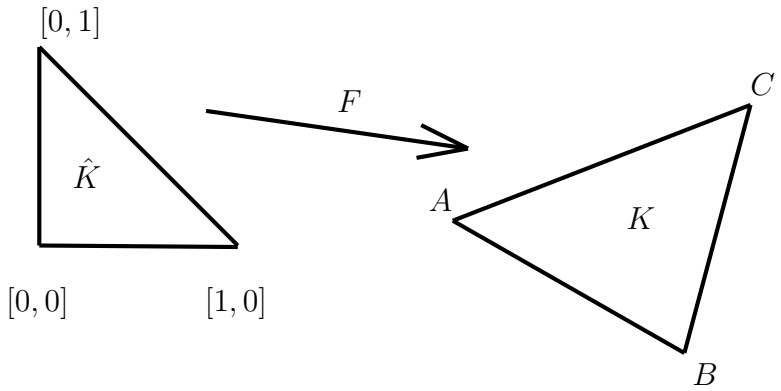
Důkaz: Začneme s definicí normy matice

$$||B_K|| = \sup_{||\eta||=1} ||B_K\eta|| = \frac{1}{\hat{\rho}} = \sup_{||\eta||=\hat{\rho}} ||B_K\eta|| \tag{2.131}$$

Vybereme dva body $\hat{p}$ a $\hat{q}$ na referenčním prvku, aby $\eta = p - q$, neboli $||\hat{p} - \hat{q}|| = \hat{\rho}$. Takové dva body jistě existují, jelikož $\hat{\rho}$ je poloměr kružnice vepsané do referenčního trojúhelníku. Promítneme $\hat{p}$ a $\hat{q}$ na skutečný prvek: $p = B\hat{p}$, $q = B\hat{q}$. Vzdálenost těchto bodů můžeme odhadnout parametrem h_K, který značí délku nejdelší strany trojúhelníku K, neboli dostáváme

$$||B_K|| = \frac{1}{\hat{\rho}}||B_K\eta|| = \frac{1}{\hat{\rho}}||B_K(\hat{p} - \hat{q})|| = \frac{1}{\hat{\rho}}||p - q|| \leq \frac{h_K}{\hat{\rho}} \tag{2.132}$$

Odhad $||B_K^{-1}||$ je zcela analogický a čtenář ho jistě zvládne sám.



Obr. 2.7: Referenční trojúhelníkový prvek

2.3.6 Odhad chyby aproximace pro 2D úlohy

Dále postupujeme stejně jako pro jednodimenzionální případ. Potřebujeme vědět jak odhadnout L^m-seminormu funkce na skutečném prvku pomocí L^m-seminormy na prvku referenčním a naopak

Věta 2.8. *Pro libovolnou funkci* $v \in H^m(K)$ *platí*

$$|\hat{v}|_{m,\hat{K}} \leq C||B_K||^m \, |\det(B_K)|^{-\frac{1}{2}} |v|_{m,K} \tag{2.133}$$

$$|v|_{m,K} \leq C||B_K^{-1}||^m \, |\det(B_K)|^{\frac{1}{2}} |\hat{v}|_{m,\hat{K}} \tag{2.134}$$

Důkaz: Na rozdíl od jednodimenzionální úlohy se omezíme na $m = 2$. Jak uvidíme tato volba je pro naše účely zcela dostačující. Postupovat budeme analogicky s $1D$ případem. Nejprve pomocí věty o substituci a zobrazení F přetransformujeme seminormu z referenčního elementu na element skutečný. Potřebujeme přetransformovat druhé derivace funkce v

$$\frac{\partial^2 \hat{v}}{\partial\lambda_1 \partial\lambda_1} = \frac{\partial}{\partial\lambda_1}\left(\frac{\partial v}{\partial x}\frac{\partial F_1}{\partial\lambda_1} + \frac{\partial v}{\partial y}\frac{\partial F_2}{\partial\lambda_1}\right) = \frac{\partial^2 v}{\partial x^2}\frac{\partial F_1}{\partial\lambda_1}\frac{\partial F_1}{\partial\lambda_1} + \frac{\partial^2 v}{\partial x \partial y}\frac{\partial F_2}{\partial\lambda_1}\frac{\partial F_1}{\partial\lambda_1} + \frac{\partial^2 v}{\partial y \partial x}\frac{\partial F_1}{\partial\lambda_1}\frac{\partial F_2}{\partial\lambda_1} \tag{2.135}$$

$$+\frac{\partial^2 v}{\partial y^2}\frac{\partial F_2}{\lambda_1}\frac{\partial F_2}{\partial\lambda_1} = \frac{\partial^2 v}{\partial x^2}B_{11}B_{11} + \frac{\partial^2 v}{\partial x \partial y}B_{21}B_{11} + \frac{\partial^2 v}{\partial y \partial x}B_{11}B_{21} + \frac{\partial^2 v}{\partial y^2}B_{21}B_{21} \tag{2.136}$$

Jistě platí, že $|B_{ij}| \leq ||B||$, což je na první pohled vidět pokud uvažujeme například řádkovou normu matice[31], která je definovaná jako maximum ze součtů absolutních hodnot prvků matice v jednotlivých řádcích. Potom můžeme odhadnout členy B_{ij} vystupující u druhých derivací funkce v pomocí normy matice B_K, dostáváme odhad

$$\frac{\partial^2 \hat{v}}{\partial\lambda_1 \partial\lambda_1} \leq 4||B_K||^2 \sum_{i+j=2} \frac{\partial^2 v}{\partial x^i \partial y^j} \tag{2.137}$$

[31]Norma matice může být definována několika způsoby. Všechny tyto normy jsou si v jistém smyslu ekvivalentní

a pro součet všech druhých derivací funkce $\hat{v}$

$$\sum_{i+j=2} \frac{\partial^2 \hat{v}}{\partial \lambda_1^i \partial \lambda_2^j} \leq 16||B_K||^2 \sum_{i+j=2} \frac{\partial^2 v}{\partial x^i \partial y^j} \tag{2.138}$$

kde $\sum_{i+j=2} \frac{\partial^2 v}{\partial x^i \partial y^j}$ značí součet všech druhých derivací funkce v. Výsledky pro další derivace dostaneme cyklickou záměnou indexů. Nyní můžeme přistoupit k odhadu seminormy funkce $\hat{v}$

$$|\hat{v}|_{2,\hat{K}} = \sqrt{\int_{\hat{K}} \left(\sum_{i+j=2} \frac{\partial^2 \hat{v}}{\partial \lambda_1^i \partial \lambda_2^j} \right)^2 \mathrm{d}\lambda} \leq \sqrt{\int_K \left(\sum_{i+j=2} \frac{\partial^2 v}{\partial x^i \partial y^j} 16||B_K||^2 \left(\frac{1}{J} \right)^2 \right)^2 \mathrm{d}x} \tag{2.139}$$

kde $J = \det(B)$ je Jakobián zobrazení mezi prvky $\hat{K}$ a K. Vytkneme člen $16||B||^2|\det(B_K)|^{-\frac{1}{2}}$ před odmocninu a integrál zapíšeme pomocí seminormy

$$|\hat{v}|_{2,\hat{K}} = \sqrt{\int_{\hat{K}} \left(\sum_{i+j=2} \frac{\partial^2 \hat{v}}{\partial \lambda_1^i \partial \lambda_2^j} \right)^2 \mathrm{d}\lambda} \leq 16||B||^2 J^{-\frac{1}{2}} \sqrt{\int_K \left(\sum_{i+j=2} \frac{\partial^2 v}{\partial x^i \partial y^j} \right)^2 \mathrm{d}x} = \tag{2.140}$$

$$= 16||B||^2|\det(B_K)|^{-\frac{1}{2}}|v|_{2,K} \tag{2.141}$$

Tím jsme dokázali první tvrzení. Druhé tvrzení se dokáže zcela analogicky a přenecháváme ho čtenáři jako cvičení.

Další věty jsou již s $1D$ případem zcela analogické, pro úplnost je zde stručně znovu uvádíme.

Věta 2.9. *Existuje konstanta $C > 0$ taková, že pro libovolnou funkci $v \in H^2(\hat{K})$ platí*

$$\inf_{p \in P_1(\hat{K})} ||v + p||_{2,\hat{K}} \leq C|v|_{2,\hat{K}} \tag{2.142}$$

Pro důkaz tohoto tvrzení opět čtenáře odkazujeme na odbornou literaturu.

Věta 2.10. *Nechť interpolace $\Pi_{\hat{K}}$ zachovává polynomy 1. stupně, neboli*

$$\Pi_{\hat{K}}(p) = p \quad \forall p \in P_1(\hat{K}) \tag{2.143}$$

pak platí

$$|(v - \Pi_K(v))|_{1,K} \leq C\frac{h_K^2}{\rho_K}|v|_{2,K} \quad \forall v \in H^2(K) \tag{2.144}$$

Důkaz: V prvním kroku odhadneme seminormu rozdílu mezi funkcí v a její interpolací na konečném prvku K pomocí seminormy rozdílu mezi funkcí v a její interpolací na referenčním konečném prvku $\hat{K}$ viz věta (2.134)s $m = 1$

$$|(v - \Pi_K(v))|_{1,K} \leq C||B^{-1}||\,|\mathrm{det}B|^{\frac{1}{2}}|(v - \Pi_{\hat{K}}(\hat{v}))|_{1,\hat{K}} \tag{2.145}$$

K důkazu této věty, stejně jako v $1D$ nejdříve využijeme to, že interpolační operátor zachovává polynomy 1. stupně. Zapíšeme rozdíl mezi funkcí a její interpolací. Následně k funkci $\hat{v}$ i k její interpolaci přičteme polynom stupně 1

$$\hat{v} - \Pi_{\hat{K}}(\hat{v}) = v + p - p - \Pi_{\hat{K}}(\hat{v}) = \hat{v} + p - \Pi_{\hat{K}}(\hat{v} + p) = (I - \Pi_{\hat{K}})(\hat{v} + p) \tag{2.146}$$

což platí pro všechna $\hat{v} \in H^2(\hat{K})$ a $p \in P_k(\hat{K})$, následně odhadneme seminormu rozdílu funkce v a její interpolace

$$|(v - \Pi_{\hat{K}}(\hat{v}))|_{1,\hat{K}} \leq C||\hat{v} + p||_{2,\hat{K}} \quad \forall p \in P_1(K) \tag{2.147}$$

kde jsme normu $||(I - \Pi)||_{\mathcal{L}(H^2([-1,1])),(H^1([-1,1]))}$ zahrnuli do konstanty C. Pro odhad pravé strany použijeme větu (2.9)

$$|(v - \Pi_{\hat{K}}(\hat{v}))|_{1,\hat{K}} \leq \tilde{C}|\hat{v}|_{2,\hat{K}} \tag{2.148}$$

Pravou stranu ztransformujeme zpět z $\hat{K}$ na K

$$|(v - \Pi_K(v))|_{1,K} \leq C||B^{-1}||\,||B||^2\,|\mathrm{det}B|^{\frac{1}{2}}\mathrm{det}B|^{-\frac{1}{2}}|(v - \Pi_K(v))|_{2,K} \quad \forall v \in H^2(K) \tag{2.149}$$

a normu matice B a její inverze odhadneme pomocí (2.129) a (2.130)

$$|(v - \Pi_K(v))|_{1,K} \leq C\frac{\hat{h}}{\rho}\left(\frac{h_K}{\hat{\rho}}\right)^2 |v - \Pi_K(v))|_{2,K} \quad \forall v \in H^2(K) \tag{2.150}$$

kde $\frac{\hat{h}}{\hat{\rho}^2}$ značí poměr mezi délkou nejdelší strany referenčního prvku a poloměrem kružnice vepsané do $\hat{K}$, který je jistě konstantní a můžeme ho zahrnout do C. Nakonec dostáváme výsledný odhad

$$|(v - \Pi_K(v))|_{1,K} \leq C\left(\frac{h_K^2}{\rho}\right)\,|(v - \Pi_K(v))|_{2,K} \quad \forall v \in H^2(K) \tag{2.151}$$

Dále budeme uvažovat pouze tzv. regulární sítě, neboli sítě pro které existuje konstanta $\beta > 0$ nezávislá na h_K, ρ_K a taková, že pro každý element platí

$$\frac{h_K}{\rho_K} \leq \beta \quad \textit{pro} \quad h \to 0 \tag{2.152}$$

kde $h = \max_K h_K$ To znamená, že se zjemňováním sítě nedochází k degeneraci trojúhelníků v úsečky. Potom můžeme odhad (2.153) přepsat jako

$$|(v - \Pi_K(v))|_{1,K} \leq Ch_K\,|(v - \Pi_K(v))|_{2,K} \quad \forall v \in H^2(K) \tag{2.153}$$

Dokázali jsme, že chyba interpolace prvku je řádu $\mathcal{O}(h_K)$, čímž jsme dostali stejný výsledek jako v případně jednorozměrné úlohy pružnosti a chybu aproximace dostaneme dosazení řešení AVP u za v a posčítáním chyby interpolace pře jednotlivé elementy, viz (2.56). Výsledná chyba aproximace má tvar

$$|(u - \Pi(u))|_{1,\Omega} \leq Ch|u|_{2,\Omega} \quad \forall v \in H^2(\Omega) \tag{2.154}$$

kde $h = max_K h_k$ a Ω je oblast na které řešíme AVP. Výsledná chyba aproximace metody konečných prvků je $\mathcal{O}(h^k)$, kde h je délka nejdelšího prvku a parametr k souvisí s regularitou přesného řešení. Jak jsme již zmínili regularita přesného řešení je velice komplikovaný problém, omezili jsme se proto pouze na $u \in H^2(\Omega)$ a lineární konečné prvky, neboť je to nejčastěji řešený problém v inženýrské praxi. Odhad v L^2-normě obdržíme opět pomocí Aubinova-Nietscheova triku,

$$|(u - \Pi(u))|_{0,\Omega} \leq Ch^2 |u|_{2,\Omega} \quad \forall u \in H^2(\Omega) \tag{2.155}$$

Jak vidíme, výsledné odhady jsou stejné jako pro jednodimenzionální úlohu a platí dokonce pro jakýkoliv problém popsaný AVP, pod kterým se může skrývat řada fyzikálních interpretací, jako úloha stacionárního vedení tepla, úloha pružnosti a další.

Dodatky

2.4 Vektorové prostory

Označme $\mathbb{R}$ těleso reálných čísel (v případě, že bychom potřebovali těleso komplexních čísel $\mathbb{C}$, výslovně to uvedeme). Prvkům tělesa říkáme skaláry. Vektorovým prostorem (nebo také lineárním prostorem) rozumíme neprázdnou množinu V (její prvky se nazývají vektory), na které jsou definovány operace sčítání vektorů a násobení vektoru skalárem. Tyto operace mají následující vlastnosti:

Každé dvojici vektorů $u, v \in V$ je jednoznačně přiřazen prvek $u + v \in V$ tak, že $u + v = v + u$. Je-li ještě $w \in V$, pak platí $(u + v) + w = u + (v + w)$. Těmto vztahům říkáme komutativní a asociativní zákon. Ve V existuje jednoznačně určený vektor 0 (nulový vektor, nebo také počátek) takový, že $0+u = u$ pro každé $u \in V$. Dále je každému $u \in V$ jednoznačně přiřazen vektor $(-u)$ tak, že $u + (-u) = 0$. Vektoru $(-u)$ se říká opačný vektor k vektoru u.

Každému vektoru $u \in V$ a skaláru $\alpha \in \mathbb{R}$ je jednoznačně přiřazen prvek $\alpha u \in V$ tak, že $1u = u$, $\alpha(\beta u) = (\alpha\beta)u$, a jsou navíc splněny distributivní zákony

$$\alpha(u + v) = \alpha u + \alpha v \tag{2.156}$$

$$(\alpha + \beta)u = \alpha u + \beta u. \tag{2.157}$$

Příklady vektorových prostorů:

- Těleso reálných čísel s běžnými operacemi sčítání a násobení tvoří vektorový prostor.
- Mějme uzavřený interval $[a, b]$. Množina všech spojitých funkcí definovaných na $[a, b]$ spolu s operací sčítání funkcí (funkce sčítáme bodově) tvoří vektorový prostor. To lze snadno domyslet, neboť součet dvou spojitých funkcí je jistě spojitá funkce a násobek spojité funkce je rovněž spojitá funkce. Nulovým prvkem je konstantní funkce všude rovná nule, existence opačného prvku je zřejmá. Tento prostor obvykle značíme $C([a, b])$, někdy také $C^0([a, b])$. Při tomto druhém značení je obecně $C^k([a, b])$ prostor funkcí majících spojité derivace až do řádu k.
- Množina všech čtvercových matic typu $n \times n$ tvoří vektorový prostor.

Podprostorem vektorového prostoru V je neprázdná množina W, která je uzavřená vzhledem k operacím sčítání vektorů a násobení vektoru skalárem, tj. jsou-li $u, v \in W$, pak $u + v \in W$ a je-li $u \in W$ a $\alpha \in \mathbb{R}$, pak $\alpha u \in W$. Pokud W je podprostorem ve V, značíme tuto skutečnost $W \subset\subset V$.

Zavedený pojem (lineárního) vektorového prostoru je velmi důležitý a představuje abstraktní strukturu mnoha konkrétních objektů. Asi nejdůležitější vlastností, plynoucí z lineární struktury, je platnost principu superpozice, se kterým se studenti setkali na mnoha místech v různých předmětech mechaniky, kde je často základním stavebním kamenem. Za všechny příklady jmenujme silovou, či deformační metodu pro výpočet prutových konstrukcí ze Stavební mechaniky 3, nebo odvozování funkcí poddajnosti a relaxačních funkcí v předmětu Přetváření a porušování materiálů.

Pro popis a zkoumání vlastností numerických metod (metody konečných prvků a metody konečných diferencí) však budeme potřebovat zavést ještě další struktury, stejně důležité, které s lineární strukturou obecně nemusí souviset. V našem případě však většinou budou. Jsou to pojmy související s měřením vzdálenosti mezi prvky, velikostí prvků a také s měřením úhlů mezi nimi (speciálně pak zavedení pojmu kolmosti). Studenti se s nimi patrně již setkali v kurzech matematiky, pro osvěžení je však stručně připomeneme.

2.5 Metrika, norma a skalární součin

Nejprve si připomeneme pojem metrického prostoru. Ačkoliv obecné metrické prostory nebudeme přímo potřebovat, je vhodné začít právě jimi, protože jsou ještě relativně intuitivní a mnoho vlastností metrických prostorů lze poté přenést i na obecnější a hůře představitelné prostory.

Metrika. *Mějme libovolnou neprázdnou množinu V. Metrikou na množině V rozumíme funkci (nikoliv nutně (bi)lineární!) $\rho : V \times V \mapsto \mathbb{R}$ takovou, ze pro všechna $u, v, w \in V$ platí*

$$
\begin{aligned}
&(i)\ \rho(u,v) \geq 0, \quad \rho(u,v) = 0 \Leftrightarrow u = v \\
&(ii)\ \rho(u,v) = \rho(v,u) \\
&(iii)\ \rho(u,w) \leq \rho(u,v) + \rho(v,w).
\end{aligned}
$$

Uspořádanou dvojici (V, ρ) nazýváme metrickým prostorem.

Číslo $\rho(u, v)$ se nazývá vzdálenost u a v. Vlastnost (i) říká, že vzdálenost nemůže být záporná a rovná nule může být jen tehdy, jsou-li prvky totožné. Vlastnost (ii) vyjadřuje symetrii, tedy že vzdálenost u od v je stejná, jako v od u, a konečně vlastnost (iii) je známá trojúhelníková nerovnost. Pokud bude jasné, o jaký prostor se jedná, budeme místo například (V, ρ), psát jen V.

Příklady různých metrik (a metrických prostorů):

- Mějme množinu $\mathbb{R}^n$. Přirozeným zobecněním vzdálenosti z dvou a třírozměrných prostorů je následující definice (euklidovské) vzdálenosti bodů $x = (x_1, ..., x_n) \in \mathbb{R}^n, y = (y_1, ..., y_n) \in \mathbb{R}^n$

$$\rho_2(x,y) := \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} = \sqrt{(x_1 - y_1)^2 + ... + (x_n - y_n)^2}. \tag{2.158}$$

  Množinu $\mathbb{R}^n$, na které definujeme euklidovskou vzdálenost, nazýváme n-rozměrným euklidovským prostorem.

- Na množině $\mathbb{R}^n$ je však možné definovat metriku i jiným způsobem. Významné jsou především tzv. "maximová" metrika ρ_∞ a "součtová" metrika ρ_1:

$$\rho_\infty(x,y) := \max\{|x_1 - y_1|, ..., |x_n - y_n|\}, \tag{2.159}$$

$$\rho_1(x,y) := |x_1 - y_1| + ... + |x_n - y_n|. \tag{2.160}$$

Podstatné je, že analogickým způsobem lze metriku zavést i na obecnějších, a hlavně nekonečně rozměrných množinách/prostorech (prostorem obvykle nazýváme množinu, na které je definována nějaká struktura, ať už linearní v případě vektorových prostorů, metrická v připadě těch metrických, nebo jiná, jak uvidíme dále). Prvky těchto množin (prostorů) jsou zpravidla funkce a není obvykle možné si je dostatečně jasně představit, jako například $\mathbb{R}^n$ (pro $n = 2, 3$). Je však možné na tyto prostory aplikovat poznatky získané abstrakcí obecně platných tvrzení, která jsou v prostorech $\mathbb{R}^2, \mathbb{R}^3$ zřejmá z geometrického názoru. Díky této "geometrizaci" je možné přirozeně zavést původně geometrické pojmy jako vzdálenost (metrika), velikost (norma) a kolmost, případně odchylka úhlů (skalární součin).

Uvažujme množinu X všech reálných funkcí, spojitých na $[0, 1]$. Otázkou je, jak definovat na takovémto prostoru (který je mimo jiné zjevně lieární, neboť součet dvou spojitých funkcí na daném intervalu i násobek takové funkce je opět spojitá funkce na témže intervalu) vzdálenost dvou funkcí f, g. Nejpřirozenější volbou by asi byla definice odpovídající "maximové" metrice:

$$\rho_\infty(f, g) := \max |f(x) - g(x)| : x \in [0, 1]. \tag{2.161}$$

Vzdálenost funkcí by tak byla dána maximálním rozdílem funkčních hodnot těchto funkcí na celém intervalu $[0, 1]$. Vidíme, že tato metrika je analogií metriky ρ_∞ zavedené na $\mathbb{R}^n$. Na X můžeme stejně tak definovat metriky, které jsou analogiemi metrik $\rho_1(x, y)$, $\rho_2(x, y)$:

$$\rho_1(f, g) := \int_0^1 |f(x) - g(x)| \, \mathrm{d}x, \tag{2.162}$$

$$\rho_2(f, g) := \sqrt{\int_0^1 (f(x) - g(x))^2 \, \mathrm{d}x}. \tag{2.163}$$

Máme-li metrický prostor V pak pro každé dvě množiny $M, N \subset V$ definujeme jejich vzdálenost jako

$$\mathrm{dist}(M, N) := \inf\{\rho(u, v); u \in M, v \in N\}. \tag{2.164}$$

Speciálně tedy vzdálenost bodu $x \in P$ od množiny $M \subset V$ je rovna

$$\mathrm{dist}(x, M) := \inf\{\rho(x, u); u \in M\}. \tag{2.165}$$

Jedním z nejdůležitějších pojmů pro nás však nebude obecný metrický prostor, ale normovaný lineární prostor. Ten je vždycky i prostorem metrickým, neboť jak uvidíme dále, vždy je na něm implicitně zavedena i metrika. Normované lineární prostory tedy lze chápat jako podmnožinu metrických prostorů.

Norma. *Normovaným lineárním prostorem rozumíme vektorový prostor V, na kterém je definována funkce $||\cdot|| : V \mapsto \mathbb{R}$ taková, že pro všechna $u, v \in V$ a všechna $\alpha \in \mathbb{R}$ platí:*

$$(i)\ ||u|| \geq 0, \quad ||u|| = 0 \Leftrightarrow u = 0$$

$$(ii)\ ||\alpha u|| = |\alpha| ||u|| \tag{2.166}$$

$$(iii)\ ||u + v|| \leq ||u|| + ||v||. \tag{2.167}$$

Funkce $||\cdot|| : V \mapsto \mathbb{R}$ se nazývá norma na V a uspořádaná dvojice $(V, ||\cdot||)$ normovaný lineární prostor.

Otázkou nyní může být, jaký je vztah mezi normou a metrikou, definovanou na lineárním (vektorovém) prostoru. Lze snadno ověřit (stačí ověřit vlastnosti $(i), (ii), (iii)$ z definice metriky), že v každém normovaném lineárním prostoru je přirozeně definována i metrika. Každá norma na lineárním prostoru V totiž přirozeně definuje metriku rovností

$$\rho(u, v) := ||u - v||. \tag{2.168}$$

V tomto případě říkáme, že je metrika "generovaná" nebo "indukovaná" normou. Obráceně to však neplatí. Metrika definovaná na lineárním prostoru nemusí obecně s jeho lineární strukturou vůbec souviset. Metrika generovaná normou má však následující vlastnosti

$$\rho(u + w, v + w) := ||u + w - v - w|| = ||u - v|| = \rho(u, v), \tag{2.169}$$
$$\rho(\alpha u, \alpha v) := ||\alpha u - \alpha v|| = |\alpha| \, ||u - v|| = \alpha\rho(u, v). \tag{2.170}$$

První z nich se nazývá invariance vůči translaci (posunutí), druhá pak homogenita (obecněji homogenita prvního řádu). Z invariance vůči translaci plyne jednoduchou úvahou $\rho(u, v) = \rho(0, v - u)$, což znamená, že k určení vzdálenosti dvou vektorů stačí znát vzdálenost jejich rozdílu od počátku, tedy jinými slovy k určení metriky (vzdálenosti) stačí znát velikost $||w|| = \rho(w, 0)$ libovolného vektoru w.

Pokud je metrický prostor (V, ρ) úplný[32], pak říkáme, že příslušný normovaný lineární prostor $(V, ||\cdot||)$ je Banachův prostor. Pokud budeme chtít zdůraznit, že daná norma je definovaná na prostoru V, budeme tuto skutečnost značit $||\cdot||_V$.

Na tomto místě uvedeme významnou větu, kterou využíváme v důkazu Laxova-Milgramova lemmatu v odstavci (2.2.2). Jedná se o tzv. "Větu o pevném bodě", nebo také o "Princip kontraktivního zobrazení". Toto tvrzení platí v libovolném úplném metrickém prostoru. My ho však uvedem v kontextu normovaných lineárních prostorů.

Princip kontraktivního zobrazení.
Mějme Banachův prostor V a na něm zobrazení $T : V \to V$, které splňuje

$$||Tv_1 - Tv_2|| \leq M||v_1 - v_2|| \tag{2.171}$$

pro všechny dvojice $v_1, v_2 \in V$ a konstantu M, $0 \leq M < 1$. Pak existuje právě jedno $u \in V$ tak, že $Tu = u$, které se nazývá pevným bodem zobrazení T.

Důkaz Je třeba dokázat, že takový prvek existuje a že existuje jednoznačně. Nejprve se podívejme na důkaz existence. Vezměme libovolný prvek $v_0 \in V$ a definujme posloupnost

$$v_1 = Tv_0, \; v_2 = Tv_1, \; ..., v_{k+1} = Tv_k, ... \tag{2.172}$$

[32]Metrický prostor nazýváme úplný, pokud je v něm každá cauchyovská posloupnost prvků konvergentní (tedy má v tomto prostoru limitu). Posloupnost prvků $\{x_n\}$ se nazývá cauchyovská, pokud pro každé $\varepsilon \geq 0$ existují n, m tak, že $\rho(x_n, x_m) < \varepsilon$. V našem případě tedy $||x_n - x_m|| < \varepsilon$. Pro bližší informace odkazujeme čtenáře na přednášky z výběrové matematiky, vedené doc. Nekvindou.

Potom dle předpokladu platí

$$||v_{k+1} - v_k||_V = ||Tv_k - Tv_{k-1}||_V \leq M||v_k - v_{k-1}||_V, \tag{2.173}$$

a tedy indukcí

$$||v_k - v_{k-1}||_V \leq M^{k-1}||v_1 - v_0||_V. \tag{2.174}$$

Pro libovolné $N > n$ potom postupně platí

$$\begin{aligned}
||v_N - v_n||_V &= \left\| \sum_{k=n+1}^{N} (v_k - v_{k-1}) \right\|_V \\
&\leq ||v_1 - v_0||_V \sum_{k=n+1}^{N} M^{k-1} \qquad \text{(Trojúhelníková nerovnost, odhad (2.174))} \\
&\leq \frac{M^n}{1-M}||v_1 - v_0||_V \qquad \text{(Součet geometrické řady)} \\
&= \frac{M^n}{1-M}||Tv_0 - v_0||_V,
\end{aligned}$$

z čehož plyne, že $\{v_n\}$ je cauchyovská posloupnost. Jelikož prostor V je Banachův (a tedy úplný), je $\{v_n\}$ konvergentní ve V. Označme tedy její limitu

$$v := \lim_{n\to\infty} v_n. \tag{2.175}$$

Potom ale platí (posun indexu o jedničku v konvergenci nehraje žádnou roli)

$$\begin{aligned}
v &= \lim_{n\to\infty} v_{n+1} \\
&= \lim_{n\to\infty} Tv_n \\
&= T(\lim_{n\to\infty} v_n) \qquad \text{(Zobrazení T je spojité - Heineho definice)} \\
&= Tv,
\end{aligned}$$

z čehož plyne, že v je hledaný pevný bod zobrazení T.

Že takový prvek může být jen jeden, plyne sporem z následujícího. Nechť tedy existují dva různé prvky $v_1, v_2 \in V$ takové, že $Tv_1 = v_1, Tv_2 = v_2$. Potom ale z definiční vlastnosti T je

$$||Tv_1 - Tv_2||_V \leq M||v_1 - v_2||_V \tag{2.176}$$

pro nějaké $0 \leq M < 1$. Jelikož ale dle předpokladu $||Tv_1 - Tv_2||_V = ||v_1 - v_2||_V$, musí být zároveň

$$||v_1 - v_2||_V \leq M||v_1 - v_2||_V. \tag{2.177}$$

To je ale možné jen tehdy, je-li $||v_1 - v_2||_V = 0$, neboť $M < 1$. Tedy z vlastnosti normy plyne $v_1 = v_2$, pevný prvek je určen jednoznačně a důkaz je hotov.

Dalším mimořádně důležitým pojmem je skalární součin a s ním související unitární prostor. Jde o přirozené zobecnění nástrojů pro měření odchylek vektorů, speciálně pak určování kolmosti vektorů.

Skalární součin. *Unitárním prostorem rozumíme lineární (vektorový) prostor V, na kterém je definovaná symetrická, pozitivně definitní bilineární forma. Jinými slovy existuje funkce $s : V \times V \mapsto \mathbb{R}$ přiřazující každému u a $v \in V$ číslo $s(u, v)$ tak, že pro všechna $u, v \in V$ a všechna $\alpha \in \mathbb{R}$ platí*

$$\begin{aligned}
&(i)\ s(u,v) = s(v,u)\\
&(ii)\ s(\alpha u, v) = \alpha s(v,u)\\
&(iii)\ s(u+v,w) = s(u,w) + s(v,w)\\
&(iv)\ x \neq 0 \Rightarrow s(u,u) \geq 0.
\end{aligned}$$

Bilineární formu $s(u, v)$ nazýváme skalárním součinem. Uspořádanou dvojici (V, s) pak nazýváme unitárním prostorem, nebo také prostorem se skalárním součinem.

Pokud nebude hrozit nedorozumění, budeme skalární součin prvků u a v obvykle značit (u, v). Řekneme, že dva prvky $u, v \in V$ jsou vzájemně kolmé, pokud jejich skalární součin $(u, v) = 0$. Tuto skutečnost značíme $u \perp v$. Dvě podmnožiny P, Q jsou vzájemně kolmé, pokud platí $(u, v) = 0$ pro všechna $u \in P$ a $v \in Q$.

Každý skalární součin na lineárním prostoru V přirozeně indukuje (určuje) na V normu rovností

$$||u||_V := \sqrt{(u,u)}. \tag{2.178}$$

Vlastnost normy (i) plyne z vlastnosti (iv) skalárního součinu a faktu, že $(0, 0) = 0$. Vlastnost (ii) plyne z $||\alpha u|| = \sqrt{(\alpha u, \alpha u)} = \sqrt{\alpha^2}||u|| = |\alpha|\ ||u||$. Trojúhelníková nerovnost (iii) je důsledkem tzv. Schwartzovy nerovnosti, kterou vzhledem k její důležitosti uvádíme samostatně:

Schwartzova nerovnost.

$$|(u,v)| \leq ||u||_V ||v||_V. \tag{2.179}$$

Potom totiž

$$\begin{aligned}
||u+v||_V^2 &= (u+v, u+v) = ||u||_V^2 + ||v||_V^2 + 2(u,v)\\
&\leq ||u||_V^2 + ||v||_V^2 + 2||u||_V||v||_V = (||u||_V + ||v||_V)^2
\end{aligned}$$

a po odmocnění dostáváme požadovanou vlastnost (iii) pro normu. Zbývá uvést

Důkaz Schwartzovy nerovnosti
Z definice skalárního součinu platí pro libovolné $u, v \in X, \alpha \in \mathbb{R}$

$$0 \leq (\alpha u - v, \alpha u - v) = \alpha^2 ||u||_V^2 - 2\alpha(u,v) + ||v||_V^2, \tag{2.180}$$

kde vpravo je kvadratický polynom v proměnné α, který má nekladný diskriminant, tj platí $4(u, v)^2 - 4||u||_V^2||v||_V^2 \leq 0$. Z toho pak snadno plyne $(u, v)^2 \leq ||u||_V^2||v||_V^2$, což po odmocnění dává Schwartzovu nerovnost, a důkaz je hotov.

Na každém unitárním prostoru je tudíž implicitně zadána norma, a tedy i metrika. Pokud příslušný metrický prostor je úplný (tedy příslušný normovaný lineární prostor Banachův), pak unitární prostor V nazýváme Hilbertovým prostorem.

Uveďme pro doplnění ještě dvě další důležité nerovnosti, které platí v Hilbertových prostorech. První z nich je známá Pythagorova věta. Jsou-li $u, v \in V$ a platí-li $u \perp v$, pak je

$$||u||_V^2 + ||v||_V^2 \leq ||u+v||_V^2. \tag{2.181}$$

To lze ověřit jednoduchým výpočtem z definice indukované normy a předpokladu kolmosti u a v. Je totiž $||u+v||_V^2 = (u+v, u+v) = ||u||_V^2 + 2(u,v) + ||v||_V^2 = ||u||_V^2 + ||v||_V^2$.

Druhá důležitá nerovnost se nazývá rovnoběžníkové pravidlo. Pro libovolné prvky $u, v \in V$ platí

$$||u+v||_V^2 + ||u-v||_V^2 = 2(||u||_V^2 + ||v||_V^2). \tag{2.182}$$

Ověření lze udělat přímým výpočtem a přenecháváme ho čtenáři. Poznamenejme, že pokud v obecném Banachově prostoru platí rovnoběžníkové pravidlo, pak v něm lze zavést skalární součin tak, aby norma na něm zavedená splývala s tou, kterou tento skalární součin indukuje.

Spojitost normy a skalárního součinu.

Norma i skalární součin definovaný v této kapitole jsou spojitá zobrazení. Přesně řečeno, máme-li Hilbertův prostor X, pak zobrazení

$$u \mapsto ||u||_V, \quad u \in V, \tag{2.183}$$

$$\{u, v\} \mapsto (u, v), \quad u, v \in V \tag{2.184}$$

jsou spojitá. Tvrzení plyne z toho, že konvergentní posloupnost je omezená[33] a ze Schwartzovy nerovnosti. Nechť tedy $u_n \to u$ a $v_n \to v$. Potom je

$$|(u_n, v_n) - (u, v)| \leq |(u_n, v_n - v) + (u_n - u, v)| \leq ||u_n||_V \underbrace{||v_n - v||_V}_{\to 0} + \underbrace{||u_n - u||_V}_{\to 0} ||v||_V,$$

kde v první nerovnosti jsme užili linearitu skalárního součinu a přičtení/odečtení výrazu (u_n, v), druhá nerovnost je důsledkem Schwartzovy nerovnosti. Jelikož u_n je omezená, dostáváme spojitost skalárního součinu. Pro spojitost normy stačí použít právě dokázanou spojitost skalárního součinu a definici indukované normy. Jelikož již víme, že $(u_n, u_n) \to (u, u)$ a $(u, u) = ||u||_V^2$, dostáváme, že $||u_n||_V^2 \to ||u||_V^2$, a tedy $||u_n||_V \to ||u||_V$ a norma je spojitá.

2.6 Lineární a bilineární zobrazení

Lineárním zobrazením (někdy též homomorfismem) vektorového prostoru V do vektorového prostoru V' rozumíme takové zobrazení F prostoru V do V' (někdy píšeme $F(\,\cdot\,) : V \mapsto V'$), že pro všechna $u, v \in V$ a všechny sklaláry $\alpha, \beta \in \mathbb{R}$ platí

$$F(\alpha u + \beta v) = \alpha F(u) + \beta F(v). \tag{2.185}$$

[33] To je vidět snadno. Nechť například $\{x_n\}$ konverguje k x. Potom existuje $\varepsilon > 0$, že od jistého n platí $||x_n - x|| < \varepsilon$. Z trojúhelníkové nerovnosti $||x_n|| < \varepsilon + ||x||$ a $\{x_n\}$ je tedy omezená. Poznamenáváme, že zde používáme Heineho definici spojitosti - F je spojité, pokud z $x_n \to x$ plyne $F(x_n) \to F(x)$.

Pokud je prostor V' těleso (například reálných čísel), pak zobrazení F říkáme lineární forma, někdy také lineární funkcionál. Příklady lineárních zobrazení:

- Mějme výše uvedený vektorový prostor spojitých funkcí $C^0([a,b])$. Zobrazení, které každé funkci přiřadí její integrál od a do b, je lineární funkcionál (forma) definovaný na $C^0([a,b])$. Formálně vyjádřeno jde o zobrazení $F : C^0([a,b]) \mapsto \mathbb{R}$ s předpisem

$$F(f) = \int_a^b f \mathrm{d}x, \quad f \in C^0([a,b]). \tag{2.186}$$

- Na tomtéž prostoru definujme zobrazení předpisem

$$F(f) = \max\{|f(x)| : x \in (a,b)\}, \quad f \in C^0([a,b]). \tag{2.187}$$

  Toto zobrazení je lineární funkcionál (forma) na prostoru $C^0([a,b])$. Dále ukážeme, že obě právě definovaná zobrazení lze chápat jako tzv. normy na prostoru $C^0([a,b])$.

Pokud máme na normovaném lineárním prostoru V definován lineární funkcionál $F(\cdot)$, pak řekneme, že je omezený, tj. pokud $\exists\, C < \infty$ tak, že

$$|F(u)| \leq C||u||_V, \quad \forall u \in V. \tag{2.188}$$

Z linearity $F(\cdot)$ pak plyne snadno i spojitost, neboť pro všechna u, $v \in V$ máme $|F(u+v)| = |F(u) + F(v)| \leq C||u+v||_V$. Prostor všech lineárních omezených funkcionálů na prostoru V označujeme V^* a nazýváme prostorem duálním k V

$$V^* = \{F(\cdot) : V \mapsto \mathbb{R}; \text{F je lineární a omezený}\}. \tag{2.189}$$

Norma na prostoru V^* se definuje jako supremum "velikostí" prvků zobrazených pomocí F přes jednotkovou kouli ve V

$$||F||_{V^*} = \sup_{u \in V \setminus \{0\}} \frac{|F(u)|}{||u||_V}. \tag{2.190}$$

Analogicky však definujeme normu libovolného lineárního zobrazení. Množinu všech lineárních zobrazení z V do V' značíme $\mathcal{L}(V, V')$. Není těžké ukázat, že $\mathcal{L}(V, V')$ je lineární vektorový prostor. Normu na tomto prostoru definujeme opět jako supremum přes jednotkovou kouli ve V.

$$||F||_{\mathcal{L}(V,V')} = \sup\{||F(u)||; ||u||_V \leq 1\}. \tag{2.191}$$

Vzhledem k (2.188) a (2.190) lze tedy říci, že lineární zobrazení $F \in \mathcal{L}(V, V')$ je spojité, pokud je omezené, tj. pokud

$$||F||_{\mathcal{L}(V,V')} \leq C||u||_V, \quad \forall u \in V.. \tag{2.192}$$

Množina prvků, které se zobrazením F zobrazí na nulový prvek (v případě funkcionálu na nulu), se nazývá jádro lineárního zobrazení (homomorfismu) F a značí se $\mathrm{Ker}\ F$. Množinu všech prvků, na které se zobrazí nějaký prvek pomocí F, nazýváme obrazem lineárního zobrazení (homomorfismu) F a značí se $\mathrm{Im}\ F$. Přesně vyjádřeno

$$\mathrm{Ker}\ F = \{u \in V; F(u) = 0\}, \tag{2.193}$$

$$\mathrm{Im}\ F = \{F(u); u \in V\} = F(V). \tag{2.194}$$

Z linearity F se snadno ukáže, že $\mathrm{Ker}\ F$ i $\mathrm{Im}\ F$ tvoří podprostory. Pokud je F spojitý, pak jsou $\mathrm{Ker}\ F$ i $\mathrm{Im}\ F$ uzavřené množiny (a tedy tvoří úplné podprostory ve V).

Kartézským součinem množin X a Y rozumíme množinu všech uspořádaných dvojic tvaru (x, y), kde $x \in X$ a $y \in Y$, a značíme ji $X \times Y$. Formálně zapsáno

$$X \times Y = \{(x, y);\ x \in X,\ y \in Y\}. \tag{2.195}$$

Bilineární formou rozumíme zobrazení $a(\cdot, \cdot) : V \times V \mapsto \mathbb{R}$ takové, že pro všechny $u, v, w \in V$ a všechny sklaláry $\alpha, \beta \in \mathbb{R}$ platí

$$a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w) \tag{2.196}$$
$$a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(v, w). \tag{2.197}$$

Právě uvedené vztahy vyjadřují linearitu v první, resp. druhé složce zobrazení a.

Pokud máme na normovaném lineárním prostoru V definovánu bilineární formu $a(\cdot, \cdot)$, pak řekneme, že je omezená (nebo ekvivalentně spojitá) pokud $\exists\ C < \infty$ tak, že

$$|a(u, v)| \leq C||u||_V||v||_V, \quad \forall u, v \in V, \tag{2.198}$$

symetrická, pokud

$$a(u, v) = a(v, u), \quad \forall u, v \in V, \tag{2.199}$$

pozitivně definitní, pokud

$$a(v, v) > 0, \quad \forall v \in V, \tag{2.200}$$

a koercivní na $Y \subset\subset V$, pokud $\exists\ \alpha > 0$ tak, že

$$|a(u, u)| \geq \alpha||u||_V^2, \quad \forall u \in Y. \tag{2.201}$$

Nyní uvedeme důležitou Riezsovu větu o reprezentaci, o kterou se opírá důkaz existence a jednoznačnosti řešení symetrického variačního problému, definovaného v kapitole (2.2.1).

Riezsova věta o reprezentaci. *Nechť V je Hilbertův prostor. Ke každému lineárnímu funkcionálu $F(\cdot)$ definovanému na V existuje jednoznačně určený prvek $u \in V$ takový, že*

$$F(v) = (u, v), \quad \forall v \in V. \tag{2.202}$$

Navíc norma funkcionálu $F(\cdot)$ ve V^* je rovná normě tohoto prvku u ve V

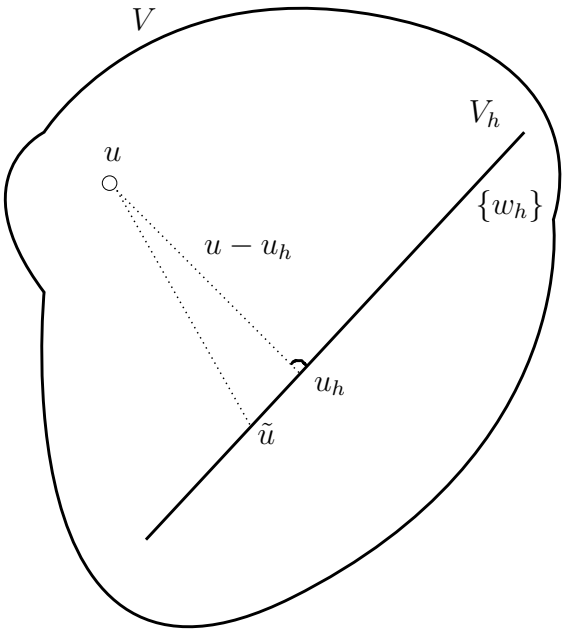
$$||F||_{V^*} = ||u||_V. \tag{2.203}$$

Než přistoupíme k důkazu Rieszovy věty, uvedeme některé vlastnosti Hilbertových prostorů, které souvisejí s existencí skalárního součinu a z něj plynoucího pojmu kolmosti.

2.6.1 Ortogonální projekce a metoda nejmenších čtverců

Obsahem této části je bližší pohled na ortogonalitu v Hilbertových prostorech. Přitom nám půjde o dvě skutečnosti. První z nich je založena na jednoduché geometrické představě. Mějme libovolný Hilbertův prostor V a v něm daný prvek (funkci) u. Chceme v úplném podprostoru $V_h \subset\subset V$ najít prvek, který bude danému prvku u nejblíže. Na základě geometrického názoru (viz Obr. 2.8) představuje nejbližší prvek pata kolmice spuštěná z bodu u (jde pochopitelně pouze o zjednodušení, ovšem podobná "geometrizace" nám velmi pomáhá pochopit jinak abstraktní problémy). Hledaný prvek u_h tedy musí mít tu vlastnost, aby prvek, který vnikne jako rozdíl $u - u_h$ (který pochopitelně leží ve V), byl kolmý na podprostor V_h (tedy na každý prvek tohoto podprostoru). Jednak si ukážeme, že takový prvek existuje, a dále ukážeme, že existuje jednoznačně. V dalším si pak předvedeme, jak takový prvek prakticky sestrojit - k tomu uvedeme metodu "Ortogonální projekce" a metodu "Nejmenších čtverců". Tento problém souvisí také s aproximací nějaké funkce pomocí daného souboru jiných, jednodušších (případně vhodnějších) funkcí, jejichž lineární kombinace ho v nějakém smyslu nejlépe nahrazuje.

Druhá věc, kterou ukážeme, je, že pomocí pojmu kolmosti lze Hilbertův prostor jednoznačně rozložit na tzv. direktní součet dvou podprostorů. Tyto podprostory jsou přitom vzájemně kolmé, tj. každý prvek z jednoho podprostoru je kolmý na každý prvek z druhého podprostoru.



Obr. 2.8: *Metoda nejmenších čtverců - Schéma*

Začněmě následující charakterizací nejbližšího prvku. Nechť V je Hilbertův prostor a $V_h \subset\subset V$ jeho úplný podprostor, $u_h \in V_h$ a $u \in V$. Potom platí, že

$$||u - u_h||_V = \mathrm{dist}(u, V_h) \quad \Leftrightarrow \quad u - u_h \perp V_h. \tag{2.204}$$

Abychom tuto ekvivalenci dokázali, musíme ukázat, že z jednoho tvrzení plyne druhé a naopak. Předpokládejme tedy nejdříve, že $||u-u_h||_V = \mathrm{dist}(u,V_h)$. Zvolme nyní libovolné $v_h \in V_h$ a $\varepsilon \in \mathbb{R}$. Jelikož $u_h + \varepsilon v_h \in V_h$, platí

$$||u-u_h||_V^2 \leq ||u-(u_h+\varepsilon v_h)||_V^2 = ||u-u_h||_V^2 - 2\varepsilon(u-u_h,v_h) + \varepsilon^2||v_h||_V^2, \quad (2.205)$$

kde jsme použili předpoklad, že u_h je nejbližší prvek. Zbytek plyne z roznásobení (lze přepsat přes skalární součin, protože $||u||_V^2 = (u,u)$). Z toho však plyne po vydělení ε (pro $\varepsilon = 0$ tvrzení platí dle předpokladu) nerovnost

$$-\varepsilon||v_h||_V^2 \leq 2(u-u_h,v_h) \leq \varepsilon||v_h||_V^2. \quad (2.206)$$

Jelikož však ε bylo libovolné, dostáváme přechodem k $\varepsilon \to 0$, že $(u-u_h,v_h) = 0$ a tedy $u-u_h \perp V_h$.

Nechť nyní naopak $u-u_h \perp V_h$. Potom pro libovolné $v_h \in V_h$ je podle Pythagorovy věty

$$||u-v_h||_V^2 = ||(u-u_h)+(u_h-v_h)||_V^2 = ||u-u_h||_V^2 + ||u_h-v_h||_V^2 \geq ||u-u_h||_V^2,$$

a tedy u_h je nejbližší prvek k u ze všech prvků V_h a důkaz je hotov. Zatím tedy víme, že má-li být nějaký prvek z V_h nejbližším prvkem k zadanému prvku $u \in V$, pak rozdíl $u-u_h$ musí být kolmý ke všem prvkům z V_h. Nyní vyvstává otázka, zda a kolik takových prvků ve V_h existuje. Odpověď je následující.

Nechť V je Hilbertův prostor a $V_h \subset\subset V$ jeho úplný podprostor. Potom existuje právě jeden prvek $u_h \in V_h$ tak, že

$$||u-u_h||_V = \mathrm{dist}(u,V_h). \quad (2.207)$$

K důkazu tohoto tvrzení nejdříve vyloučíme triviální případ, kdy $u \in V_h$. Potom zřejmě $u = u_h$ je jediným prvkem, který požadovanou rovnost splňuje. Nechť tedy $u \notin V_h$. Přitom můžeme bez újmy na obecnosti pomocí posunutí (transformací) $h \mapsto u-h$ předpokládat, že $u = \mathbf{0}$[34]. Označíme-li nyní $d := \mathrm{dist}(\mathbf{0},V_h)$, hledáme vlastně ve V_h prvek s nejmenší normou (což je totéž, jako prvek nejblíže počátku), tedy s vlastností $||u_h||_V = d$. Přitom je nutně $d > 0$, neboť jinak by prvek $\mathbf{0}$ ležel ve V_h, což jsme na začátku vyloučili.

Abychom dokázali, že takový prvek opravdu existuje, zkonstruujeme posloupnost prvků $\{v_i\}$ tak, aby $||v_i||_V \to d$, neboť $d = \inf\{||v||_V,\ \ v \in V_h\}$. Stačí ukázat, že $\{v_i\}$ je cauchyovská, protože V_h je úplný podprostor, a tedy každá cauchyovská posloupnost v něm má limitu. Jelikož (jak jsme dokázali výše) je norma spojitá funkce, musí se tato limita rovnat našemu hledanému prvku, tedy $||u_h||_V = \lim_{i\to\infty}||v_i||_V = d$. Abychom ukázali, že $\{v_i\}$ je cauchyovská, využijeme faktu, že $(v_i+v_j)/2 \in V_h$ (díky tomu, že V_h je podprostor) a rovnoběžníkového pravidla dokázaného výše.

$$||v_i-v_j||_V^2 = 2(||v_i||_V^2 + ||v_j||_V^2) - 4||\frac{1}{2}(v_i+v_j)||_V^2 \leq 2(||v_i||_V^2 + ||v_j||_V^2) - 4d^2 \to 0,$$

[34]Nulový vektor značíme tučně, aby bylo zřejmé, že jde o vektor.

pro $i, j \to \infty$ a tedy $\{v_i\}$ je cauchyovská a hledaný prvek s nejmenší normou ve V_h existuje. Že může být nejvýše jeden, plyne z následující úvahy. Nechť v_1, v_2 splňují oba hledanou rovnost. Potom $(v_i + v_j)/2 \in V_h$ a opět z rovnoběžníkového pravidla máme

$$||v_1 - v_2||_V^2 = 2(||v_1||_V^2 + ||v_2||_V^2) - 4||\frac{1}{2}(v_1 + v_2)||_V^2 \leq 2(d^2 + d^2) - 4d^2 \leq 0.$$

Jelikož ale norma je nezáporná funkce, musí nutně být $v_1 = v_2$ a důkaz je hotov.

Zatím jsme tedy ukázali, že máme-li podprostor $V_h \subset\subset V$, pak v něm existuje jednoznačně určený prvek, který je nejblíže nějakému zadanému pvku $u \in V$. Nyní se podíváme na to, jak takový prvek prakticky nalézt v případě, že je příslušný podprostor $V_h \subset\subset V$ konečně dimenzionální[35]. V této souvislosti uvedeme jednak metodu "Ortogonální projekce" a pak také metodu "Nejmenších čtverců".

Metoda ortogonální projekce

Mějme tedy libovolný Hilbertův prostor V a v něm daný prvek (funkci) u. Chceme v podporstoru $V_h \subset\subset V$ najít prvek, kterým bude danému prvku u nejblíže. Ukázali jsme, že hledaný prvek u_h tedy musí mít tu vlastnost, aby rozdíl $u - u_h$ byl kolmý na celý podprostor V_h (tedy na každý prvek tohoto podprostoru).

Každý prvek $w \in V_h$ se však dá jednoznačně vyjádřit jako lineární kombinace prvků báze V_h. Proto stačí, aby rozdíl $u - u_h$ byl kolmý ke všem prvkům báze V_h. Matematicky vyjádřeno jde o následující vztah (připomínáme, že kulatými závorkami značíme skalární součin, a ten je roven nule v případě, že jsou dva prvky vzájemně kolmé)

$$(u - u_h, w) = 0, \quad \forall w \in V_h. \tag{2.208}$$

Oba prvky u_h a w leí v prostoru V_h. Pokud označíme $\{w_i\}_{i=1}^n$ bázi tohoto prostoru, můžeme tyto prvky rozepsat jako lineární kombinaci prvků této báze. Tím dostaneme

$$(u - \sum_{i=1}^{n} \alpha_i w_i, w_j) = 0, \quad \forall w_j, j = 1..n, \tag{2.209}$$

kde platnost rovnosti požadujeme nikoliv pro všechna $w \in V_h$, ale pouze pro všechny prvky báze V_h (doporučujeme čtenáři, aby si promyslel, že je tento požadavek ekvivalentní s platností pro všechny prvky). Další úpravy plynou z vlastností skalárního součinu, které jsou uvedeny v předchozí kapitole. Z (2.209) postupně dostáváme

$$(\sum_{i=1}^{n} \alpha_i w_i, w_j) = \sum_{i=1}^{n} \alpha_i \underbrace{(w_i, w_j)}_{K_{ij}} = \underbrace{(u, w_j)}_{F_j}, \quad \forall w_j, j = 1..n, \tag{2.210}$$

kde jsme označili K_{ij} a F_j skalární součin dvou prvků báze $\{w_i\}_{i=1}^n$, resp. součin prvku báze a původního aproximovaného prvku u. Rovnost (2.210) lze tedy kompatněji zapsat jako

$$\sum_{i=1}^{n} K_{ij}\alpha_i = F_j, \quad \forall j = 1..n, \tag{2.211}$$

[35] Poznamenejme, že každý konečně dimenzionální prostor je úplný.

což není nic jiného, než alternativní zápis pro soustavu rovnic $Kr = F$, která má v plném rozepsání tvar

$$\begin{pmatrix} (w_1,w_1) & (w_1,w_2) & \dots & (w_1,w_n) \\ (w_2,w_1) & (w_2,w_2) & \dots & (w_2,w_n) \\ \dots & \dots & \dots & \dots \\ (w_n,w_1) & (w_n,w_2) & \dots & (w_n,w_n) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} (u,w_1) \\ (u,w_2) \\ \vdots \\ (u,w_n) \end{pmatrix}. \tag{2.212}$$

Poznamenejme, že vzhledem k lineární nezávislosti prvků báze $\{w_i\}_{i=1}^n$ je determinant soustavy (2.212) nenulový a tato soustava je tedy jednoznačně řešitelná. Determinant této soustavy se někdy nazývá Grammův determinant.

Metoda nejmenších čtverců

Mějme opět daný Hilbertův prostor V a předpokládejme, že je v něm norma generována skalárním součinem. Nechť opět $V_h \subset\subset V$ je konečně dimenzionální podprostor a $u \in V$ libovolný prvek. V metodě nejmenších čtverců jde o nalezení takového prvku $u_h \in V_h$, že jeho vzdálenost od zadaného prvku u je nejmenší ze všech možných prvků $\tilde{u} \in V_h$, viz Obr. 2.8. Přitom budeme požadovat minimální druhou mocninu (square = čtverec, odtud název metody) této vzdálenosti, což ovšem vede ke stejnému výsledku. Jinými slovy, hledáme takový prvek u_h, že

$$||u - u_h||^2 = \min_{\tilde{u} \in V_h} ||u - \tilde{u}||^2. \tag{2.213}$$

Jelikož opět prvek $\tilde{u}$ leží ve V_h, můžeme ho vyjádřit pomocí prvků báze, tj. $\tilde{u} = \sum_{i=1}^n \alpha_i w_i$. Pokud zavedeme označení $F = ||u - \tilde{u}||^2$, pak řešíme úlohu minimalizace funkce

$$F(\alpha_1, ..., \alpha_n) = ||u - \sum_{i=1}^n \alpha_i w_i||^2.$$

Pokud má tato funkce nabývat svého minima v nějakém bodě $(\alpha_1, ..., \alpha_n)$, pak v tomto bodě musí platit

$$\frac{\partial F(\alpha_1, ..., \alpha_n)}{\partial \alpha_i} = 0, \quad \forall i = 1..n.$$

Jelikož je dle předpokladu norma na V generována skalárním součinem, můžeme psát

$$F(\alpha_1, ..., \alpha_n) = (u - \sum_{i=1}^n \alpha_i w_i, u - \sum_{j=1}^n \alpha_j w_j) = (u,u) - 2(u, \sum_{i=1}^n \alpha_i w_i) + (\sum_{i=1}^n \alpha_i w_i, \sum_{j=1}^n \alpha_j w_j).$$

$$= (u,u) - 2(u, \sum_{i=1}^n \alpha_i w_i) + \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j (w_i, w_j). \tag{2.214}$$

Derivace (2.214) podle k-té proměnné má potom tvar

$$\frac{\partial F(\alpha_1, ..., \alpha_n)}{\partial \alpha_k} = -2(u, w_k) + 2\alpha_k (w_k, w_k) + 2 \sum_{\substack{i=1 \\ i \neq k}}^n \alpha_i (w_k, w_i)$$

$$= -2(u, w_k) + 2 \sum_{i=1}^n \alpha_i (w_k, w_i) = 0, \quad \forall k = 1..n. \tag{2.215}$$

Vztah (2.215) je ovšem totožná soustava rovnic, jakou jsme dostali v metodě ortogonální projekce, viz (2.210). Obě metody tedy vedou na stejnou soustavu rovnic za předpokladu, že norma v prostoru V je generována skalárním součinem.

Ortogonální rozklad

Nechť opět $V_h \subset\subset V$ je úplný podprostor. Definujme nyní zobrazení $P_{V_h} : V \mapsto V_h$, které prvku u přiřadí takový prvek $P_{V_h}u$, pro který platí $||u - P_{V_h}u||_V = \mathrm{dist}(u, V_h)$. Výše jsme dokázali, že takový prvek je jednoznačně určen. Toto zobrazení P_{V_h} nazýváme ortogonální projekcí prostoru V na podprostor V_h.[36] Lze dokázat, že P_{V_h} je lineární zobrazení, $P^2 = P$, $||Pu||_V \leq ||u||_V$ pro všechna $u \in V$ a platí následující vztahy

$$\mathrm{Ker}\ P_{V_h} = V_h^\perp, \quad \mathrm{Im}\ P_{V_h} = V_h. \tag{2.216}$$

Význam symbolů je následující. Definujme množinu těch $v \in V$, které jsou kolmé k danému prvku $u \in V$, jako $u^\perp$. Tedy $u^\perp = \{v \in V; (v, u) = 0\}$. Průnik $V_h^\perp = \bigcap_{u_h \in V_h} u_h^\perp$ je potom množina těch prvků $v \in V$, které jsou kolmé na každý prvek $u_h \in V_h$, neboli na celý prostor V_h. Tedy $V_h^\perp = \{v \in V; (v, u_h) = 0, \ \forall u_h \in V_h\}$.

Že je $V_h^\perp$ uzavřený podprostor plyne z toho, že je podprostor ve V (linearita - podprostor) a ze spojitosti skalárního součinu dokázaného výše (uzavřenost). Podprostor $V_h^\perp$ se nazývá ortogonální doplněk V_h ve V. Linearita P_{V_h} vyplývá z toho, jak jsme v dané množině charakterizovali nejbližší prvek k danému prvku a toho, že takový prvek jednoznačně existuje. Vezmeme-li totiž $u, v \in V$, pak pro libovolné $u_h \in V_h$ máme

$$(u + v - (P_{V_h}u + P_{V_h}v, u_h) = (u - P_{V_h}u, u_h) + (v - P_{V_h}v, u_h) = \mathbf{0}, \tag{2.217}$$

protože $u - P_{V_h}u \perp V_h$ (tak jsme charakterizovali nejbližší prvek). Potom ale nutně

$$||u + v - (P_{V_h}u + P_{V_h}v, u_h)||_V = \mathrm{dist}(u + v, V_h), \tag{2.218}$$

a protože nejbližší prvek k $u + v$ ve V_h je určen jednoznačně a je jím prvek $P_{V_h}(u + v)$, musí platit $P_{V_h}(u + v) = P_{V_h}u + P_{V_h}v$. Analogicky bychom dokázali, že $P_{V_h}(\alpha u) = \alpha P_{V_h}u$, a tedy P_{V_h} je lineární. Pro každý prvek $u \in V$ triviálně platí $u = (u - P_{V_h}u) + P_{V_h}u$. Jelikož ale zároveň $u - P_{V_h}u \perp P_{V_h}u$, pak z Pythagorovy věty dostáváme

$$||u||_V^2 = ||u - P_{V_h}u||_V^2 + ||P_{V_h}u||_V^2 \geq ||P_{V_h}u||_V^2, \tag{2.219}$$

a tedy $||P_{V_h}u||_V^2 \leq ||u||_V^2$ a P_{V_h} je omezený (a tedy spojitý) operátor. Jelikož pro všechna $u \in V$ platí $P_{V_h}u \in V_h$ a pro všechna $u_h \in V_h$ je $P_{V_h}u_h = u_h$, musí být $P_{V_h}(P_{V_h}u) = P_{V_h}^2u = P_{V_h}u$ a P_{V_h} je projekce.

Pro důkaz rovností (2.216) je třeba dokázat, že je-li prvek v jedné množině, je i ve druhé a opačně (podobně jako u důkazu ekvivalence). Je-li tedy $u \in \mathrm{Ker}\ P_{V_h}$, pak je podle definice $P_{V_h}u = \mathbf{0}$. V tom případě $u = u - P_{V_h}u \in V_h^\perp$. Je-li naopak $u \in V_h^\perp$, je $u \perp V_h$, a tedy i $u - \mathbf{0} \perp V_h$. Z definice nejbližšího prvku tím pádem musí být $||u - \mathbf{0}||_V = \mathrm{dist}(u, V_h)$ a podle

[36] Projekcí nazýváme každý spojitý lineární operátor $P : V \mapsto V$, pro který platí $P^2 = P$. Této vlastnosti se říká idempotence.

definice operátoru P_{V_h} musí být $P_{V_h}u = \mathbf{0}$. Potom ale $u \in \mathrm{Ker}\ P_{V_h}$ a platí tedy $\mathrm{Ker}\ P_{V_h} = V_h^\perp$. Druhá z rovností v (2.216) je zřejmá z definice P_{V_h} a důkaz je hotov.

Z dosavadního výkladu vyplývá v podstatě následující. Máme-li Hilbertův prostor V a $V_h \subset\subset V$ je jeho úplný podprostor, pak lze libovolný prvek $u \in V$ psát jednoznačným způsobem jako

$$u = u_{V_h} + u_{V_h^\perp}, \tag{2.220}$$

kde $u_{V_h} \in V_h$ a $u_{V_h^\perp} \in V_h^\perp$. S pomocí výše uvedených skutečností lze tento rozklad dokázat již snadno. Dle předpokladu víme, že V_h je uzavřený podprostor a uzavřenost $V_h^\perp$ jsme dokázali výše. Dále pokud by pro nějaké $u \in V$ platilo $u \in V_h \cap V_h^\perp$, muselo by být $(u, u) = 0$, a tedy $u = \mathbf{0}$. Tudíž platí, že $V_h \cap V_h^\perp = \{\mathbf{0}\}$. Jelikož dále je $u = P_{V_h}u + (u - P_{V_h}u)$, přičemž $P_{V_h}u \in V_h$ a $u - P_{V_h}u \in V_h^\perp$, je rozklad jednoznačný. [37]

Důsledkem pak je fakt, že je-li V Hilbertův prostor a $V_h \subset\subset V$ je jeho vlastní (tj. $V_h \neq V$) uzavřený podprostor, pak v ortogonálním doplňku $V_h^\perp$ existuje nenulový prvek.

Nyní můžeme přistoupit k důkazu Riezsovy věty o reprezentaci, formulované v Dodatku (2.6).

Důkaz Riezsovy věty o reprezentaci

Začněme důkazem existence. Definujme proto v duchu předchozího výkladu $V_h := \{v \in V; F(v) = 0\} = \mathrm{Ker}\ F$. Podle výše dokázaného je V_h uzavřený podprostor ve V, a navíc lze každý prvek $u \in V$ psát ve tvaru $u = u_{V_h} + u_{V_h^\perp}$, kde $u_{V_h} \in V_h$ a $u_{V_h^\perp} \in V_h^\perp$. Bez újmy na obecnosti lze předpokládat, že $V_h^\perp \neq \{\mathbf{0}\}$ (jinak totiž $F \equiv 0$ a lze brát $u = \mathbf{0}$).

Vezměme libovolné nenulové $z \in V_h^\perp$. Potom platí $F(z) \neq 0$, neboť v opačném případě by muselo být $z \in V_h$, a tedy $z \in V_h \cap V_h^\perp = \{\mathbf{0}\}$, a to je spor s nenulovostí z. Zvolme libovolný prvek $v \in V$ a definujme číslo $\beta = F(v)/F(z)$. Potom z linearity$F$ je

$$F(v - \beta z) = F(v) - \beta F(z) = 0, \tag{2.221}$$

kde jsme pouze použili definici čísla β. To ovšem znamená, že $v - \beta z \in V_h$, neboť tak jsme podprostor V_h definovali. Tím pádem je podle předchozího označení $v - \beta z = v_{V_h}$ a $\beta z = v_{V_h^\perp}$, a tedy musí být $\beta z \in V_h^\perp$. Zvolme nyní prvek

$$u := \frac{F(z)}{||z||_V^2} z. \tag{2.222}$$

Předně poznamenejme, že $u \in V_h^\perp$ (neboť $z \in V_h^\perp$). Potom následující výpočet dává

$$\begin{aligned}
(u, v) &= (u, (v - \beta z) + \beta z) \\
&= (u, v - \beta z) + (u, \beta z) \\
&= (u, \beta z) && (u \in V_h^\perp, v - \beta z \in V_h) \\
&= \beta \frac{F(z)}{||z||_V^2}(z, z) && \text{(definice prvku u 2.222)} \\
&= \beta F(z) && \text{(definice normy)} \\
&= F(v), && \text{(definice } \beta)
\end{aligned}$$

[37] Tento rozklad jinými slovy říká, že podprostory V_h a $V_h^\perp$ tvoří direktní součet, tj. že platí $V = V_h \oplus V_h^\perp$.

a tedy u je náš hledaný prvek z V.

Zbývá dokázat rovnost $||F||_{V^*} = ||u||_V$. Z definice prvku u plyne, že

$$||u||_V := \frac{|F(z)|}{||z||_V}. \tag{2.223}$$

Nyní z definice normy lineárního funkcionálu (2.190) dostáváme postupně

$$\begin{aligned}
||F||_{V^*} &= \sup_{u \in V \setminus \{0\}} \frac{|F(u)|}{||u||_V} \\
&= \sup_{u \in V \setminus \{0\}} \frac{|(u,v)|}{||u||_V} \\
&\leq ||u||_V && \text{(Schwartzova nerovnost)} \\
&= \frac{|F(z)|}{||z||_V} \\
&\leq ||F||_{V^*}. && \text{(konkrétní prvek dá menší hodnotu, než supremum)}
\end{aligned}$$

Tedy máme $||F||_{V^*} \leq ||u||_V \leq ||F||_{V^*}$ a důkaz je hotov.

Uvědomme si, že Riezsova věta o reprezentaci umožňuje v jistém smyslu ztotožnit prostory V a V^*, mezi kterými díky této větě existuje vzájemně jednoznačné zobrazení (izometrie) $\tau : V^* \mapsto V$. Toho využijeme v důkazu Laxova-Milgramova lemmatu v sekci (2.2.2).

2.7 O prostorech funkcí

V této sekci uvedeme definice některých důležitých prostorů funkcí a přehled jejich nejdůležitějších vlastností. Některých těchto prostorů jsme se již dotkli v předchozí kapitole při definování pojmů metriky, normy a skalárního součinu, které v podstatě vytvářejí strukturu těchto prostorů a určují i některé jejich vlastnosti. Následující podkapitoly nemají nahrazovat příslušné přednášky z matematiky, slouží jen jako shrnutí pojmů a zdroj odkazů pro výklad zejména v kapitole o metodě konečných prvků. Pro zájemce o hlubší studium zde zmíněných pojmů odkazujeme na přednášky z matematiky a příslušnou literaturu.

Lebesgueovy prostory

V kontextu prostorů funkcí, které vystupují v teorii parciálních diferenciálních rovnic jsou základními prostory tzv. Lebesgueovy prostory. V dalším textu tohoto odstavce budeme předpokládat, že Ω je omezená, lebesgueovsky měřitelná podmnožina $\mathbb{R}^n$ a f je lebesgueovsky měřitelná funkce.[38]

Lebesgueův integrál funkce f přes množinu Ω značíme

$$\int_\Omega f(x)\mathrm{d}x. \tag{2.224}$$

[38]Na tomto místě předpokládáme, že čtenář zná z matematiky základy teorie míry a Lebesgueova integrálu. Dostačující jsou například rozšířené přednášky z předmětu Matematika 4, vedené doc. Nekvindou.

Připomínáme, že pro hodnotu integrálu (2.224) nehrají roli množiny (lebesgueovsky) míry nula. Nechť nyní $1 \leq p < \infty$ je reálné číslo. Potom můžeme definovat normu funkce f jako

$$||f||_{L^p(\Omega)} := \left(\int_\Omega |f(x)|^p \mathrm{d}x \right)^{\frac{1}{p}}. \tag{2.225}$$

Pokud $p = \infty$, definujeme

$$||f||_{L^\infty(\Omega)} := \operatorname{ess\,sup}\{|f(x)| \; ; \;\; x \in \Omega\}, \tag{2.226}$$

kde $\operatorname{ess\,sup}$ značí esenciální supremum. To se liší od běžného suprema pouze tím, že nebere v potaz množiny míry nula. S použitím právě zavedených norem lze definovat Lebesgueovy prostory jako následující množiny funkcí

$$L^p(\Omega) := \{f \; ; \;\; ||f||_{L^p(\Omega)<\infty}\}. \tag{2.227}$$

Poznamenejme, že prostor $L^p(\Omega)$ je ve skutečnosti prostorem tříd ekvivalence funkcí, které se vzájemně liší nejvýše na množině míry nula. Je však obvyklé tento prostor brát tak, jako bychom měli pouze běžný prostor funkcí, ve kterém z každé třídy ekvivalence bereme nějakého reprezentanta.

Nyní uvedeme některé důležité nerovnosti, které se často využívají v kapitole o metodě konečných prvků. První z nich je tzv. Minkowského nerovnost.

Minkowského nerovnost. *Mějme* $1 \leq p \leq \infty$ *a funkce* $f, g \in L^p(\Omega)$. *Potom platí*

$$||f + g||_{L^p\Omega} \leq ||f||_{L^p(\Omega)} + ||g||_{L^p(\Omega)}. \tag{2.228}$$

Minkowského nerovnost vlastně není ničím jiným, než trojúhelníkovou nerovností v L^p prostorech. Další nerovností, z hlediska četnosti užití asi nejdůležitější, je Hölderova nerovnost.

Hölderova nerovnost. *Mějme* $1 \leq p, q \leq \infty$ *takové, že splňují* $1 = \frac{1}{p} + \frac{1}{q}$. *Nechť dále* $f \in L^p(\Omega), \; g \in L^q(\Omega)$. *Potom platí*

$$||fg||_{L^1\Omega} \leq ||f||_{L^p(\Omega)} ||g||_{L^q(\Omega)}. \tag{2.229}$$

Na závěr uveďme, že prostory L^p jsou úplné normované lineární prostory, a tedy Banachovy. Pro $p = 2$ je prostor $L^2(\Omega)$ dokonce Hilbertův, neboť v tom případě jeho norma indukuje skalární součin

$$(f, g) = \int_\Omega f(x) g(x) \mathrm{d}x. \tag{2.230}$$

Pro obecné L^p prostory to však neplatí.

Sobolevovy prostory

V teorii PDR jsou velmi důležité Sobolevovy prostory. Než si je zadefinujeme, podíváme se stručně na pojem tzv. slabé derivace funkce. Tento pojem se vyskytuje ve slabé formulaci diferenciální rovnice, například v (2.7). Důvod je ten, že příslušné funkce hledáme v prostorech, kde hodnoty funkcí v konečně (nebo i spočetně) mnoha bodech nehrají roli. Jelikož pojem klasické derivace se vztahuje právě ke konkrétnímu bodu, je z hlediska této teorie nevhodný a je třeba najít obecnější definici derivace. Mějme stejně jako v předchozím odstavci omezenou podmnožinu $\Omega \subset \mathbb{R}^n$. Označme $\mathcal{D}(\Omega)$ vektorový prostor nekonečně diferencovatelých funkcí, majících kompaktní nosič v Ω [39] a $L^1_{loc}(\Omega)$ prostor všech lokálně integrovatelných funkcí na Ω. [40] Řekneme, že funkce $f \in L^1_{loc}(\Omega)$ má slabou derivaci podle i-té proměnné, pokud existuje funkce$g^i \in L^1_{loc}(\Omega)$ taková, že

$$\int_\Omega g^i(x)\phi(x)\mathrm{d}x = -\int_\Omega f(x)\frac{\partial\phi(x)}{\partial x_i}\mathrm{d}x, \qquad \forall\phi(x) \in \mathcal{D}(\Omega). \tag{2.231}$$

Pokud má funkce f slabou derivaci podle i-té proměnné, pak funkce $g^i \in L^1_{loc}(\Omega)$ v uvedené rovnosti je určena jednoznačně. [41] V tom případě značíme tuto funkci g^i jako $D^i f$. Má-li funkce f slabé derivace podle všech proměnných, pak řekneme, že je slabě diferencovatelná na Ω a její slabou derivaci značíme $\nabla f := (D^1 f, ..., D^n f)$.

Má-li funkce f spojité parciální derivace $\frac{\partial f}{\partial x_i}$ na Ω, pak platí $g^i = \frac{\partial f}{\partial x_i}$, tedy klasická a slabá derivace splývají. Analogickým způsobem lze definovat i derivace vyšších řádů.

Nyní můžeme přistoupit k definici Sobolevových prostorů. Mějme $1 \leq p \leq \infty$ a otevřenou množinu $\Omega \subset \mathbb{R}^n$. Řekneme, že funkce $f \in L^p(\Omega)$ leží v Sobolevově prostoru $W^{1,p}(\Omega)$, pokud pro všechna $i = 1, ..., n$ existuje slabá derivace $D^i f$ a platí $D^i f \in L^p(\Omega)$. Jinými slovy funkce f leží v Sobolevově prostoru, pokud leží v Lebesgueově prostoru $L^p(\Omega)$ ona sama, i její slabá derivace. Poznamenejme, že prostor $W^{1,p}(\Omega)$ sestává opět z tříd ekvivalentních funkcí, které však ztotožňujeme s vybraným reprezentantem.

V prostoru $W^{1,p}(\Omega)$ definujeme normu následujícím způsobem. Pokud $1 \leq p < \infty$ pak

$$||f||_{W^{1,p}(\Omega)} := \left(\int_\Omega |f(x)|^p + |\nabla f|^p \mathrm{d}x\right)^{\frac{1}{p}}. \tag{2.232}$$

V případě $p = \infty$ definujeme

$$||f||_{W^{1,\infty}(\Omega)} := \max\{||f(x)||_{L^\infty(\Omega)},\ ||\nabla f(x)||_{L^\infty(\Omega)}\}, \tag{2.233}$$

Sobolevovy prostory jsou úplné a tedy Banachovy. Navíc v případě, že $p = 2$, je $W^{1,2}(\Omega)$ rovněž Hilbertův. Analogicky lze definovat i prostory $W^{k,p}(\Omega)$. K tomu nejprve definujme tzv.

[39] Množinu nekonečně difeencovatelných funkcí už jsme v kapitole o MKP označili jako $C^\infty(\Omega)$. Funkce z $\mathcal{D}(\Omega)$ jsou ty funkce z $C^\infty(\Omega)$, pro které $\mathrm{supp} f(x) \subset \Omega$. Přitom $\mathrm{supp} f(x) = \{x \in \Omega;\ f(x) \neq 0\}$, tedy množina, kde je $f(x)$ je nenulová.

[40] Funkce z $L^1_{loc}(\Omega)$ patří do $L^1(K)$, kde K je libovolná uzavřená omezená podmnožina Ω. Požadavek $f \in L^1_{loc}(\Omega)$ je tedy slabší, než $f \in L^1(\Omega)$, neboť platí $L^1(\Omega) \subset L^1_{loc}(\Omega)$. Opačná inkluze však neplatí.

[41] To plyne snadno z následující úvahy. Nechť jsou g, h dvě funkce takové, že pro ně platí (2.231). Potom odečtením obou rovností dostáváme $\int_\Omega (g-h)\phi \mathrm{d}x = 0$ pro každou funkci z $\mathcal{D}(\Omega)$, což je však možné jen tehdy, je-li $g = h$ skoro všude.

multiindex α jakožto n-tici nezáporných přirozených čísel $(\alpha_1, ..., \alpha_n)$. Délka multiindexu α je definovaná jako

$$|\alpha| = \sum_{i=1}^{n} \alpha_i = k. \tag{2.234}$$

Potom můžeme derivaci k-tého řádu označit jako

$$D^{|\alpha|}\phi = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n}. \tag{2.235}$$

Analogicky k definici slabé derivace můžeme nyní definovat slabou derivaci k-tého řádu funkce f jako takovou funkci g^α, pro kterou

$$\int_\Omega g^\alpha(x)\phi(x)\mathrm{d}x = (-1)^{|\alpha|}\int_\Omega f(x)D^{|\alpha|}\phi(x)\mathrm{d}x, \qquad \forall\phi(x) \in \mathcal{D}(\Omega). \tag{2.236}$$

Funkci g^α pak obvykle značíme $D^{|\alpha|}f$. Sobolevův prostor $W^{k,p}(\Omega)$ potom obsahuje funkce, pro které

$$||f||_{W^{k,p}(\Omega)} := \left(\sum_{|\alpha|\leq k} ||D^\alpha f||^p_{L^p(\Omega)}\right)^{\frac{1}{p}} \tag{2.237}$$

v případě, že $1 \leq p \leq \infty$ a

$$||f||_{W^{k,\infty}(\Omega)} := \max_{|\alpha|\leq k} ||D^\alpha f(x)||_{L^\infty(\Omega)}, \tag{2.238}$$

v případě, že $p = \infty$. Poznamenjme, že v případě prostoru $W^{k,2}(\Omega)$ tyto prostory obvykle značíme $H^k(\Omega)$ (jelikož to jsou Hilbertovy prostory). Normy na nich pak značíme $||\cdot||_{k,\Omega}$. V tomto duchu pak normu na $L^2(\Omega)$ značíme $||\cdot||_{0,\Omega}$.

Poslední pojem, který je třeba definovat, je tzv. seminorma na prostoru $H^k(\Omega)$. Ta je definovaná stejně, jako norma na $H^k(\Omega)$ s tím rozdílem, že bereme pouze nejvyšší derivace. Tuto seminormu značíme jednoduchými svislými čarami $|\cdot|_{k,\Omega}$. Seminorma $f \in H^k(\Omega)$ je tedy rovna

$$|f|_{k,\Omega} := \left(\sum_{|\alpha|=k} ||D^\alpha f||^2_{L^2(\Omega)}\right)^{\frac{1}{2}}. \tag{2.239}$$

2.8 Klasifikace PDR

V této sekci se budeme věnovat klasifikaci PDR. Přitom se pro jednoduchost omezíme na lineární PDR 2. řádu ve dvou proměnných. Každou takovou rovnici lze obecně zapsat ve tvaru $Lu = G$,

$$A\frac{\partial^2 u}{\partial x_1^2} + B\frac{\partial^2 u}{\partial x_1 \partial x_2} + C\frac{\partial^2 u}{\partial x_2^2} + D\frac{\partial u}{\partial x_1} + E\frac{\partial u}{\partial x_2} + Fu = G, \tag{2.240}$$

kde A, B, C, D, E, F, G jsou funkce x_1, x_2. Zmíněné dělení je dáno znaménkem diskriminantu $\Delta = B^2 - 4AC$. Rovnice (2.240) se nazývá

$$\text{eliptická} \quad \Leftrightarrow \quad \Delta < 0 \tag{2.241}$$

$$\text{parabolická} \quad \Leftrightarrow \quad \Delta = 0 \tag{2.242}$$

$$\text{hyperbolická} \quad \Leftrightarrow \quad \Delta > 0. \tag{2.243}$$

Stojí za povšimnutí, že uvedené dělení závisí pouza na koeficientech stojících u druhých derivací. Přitom libovolná z proměnných x_1, x_2 může představovat čas! Zdůvodnění zavedeného označení lze hledat v analytické geometrii. Množina bodů v rovině tvoří kvadratickou křivku, pokud souřadnice těchto bodů vyhovují rovnici

$$Ax_1^2 + Bx_1x_2 + Cx_2^2 + Dx_1 + Ex_2 + F = G. \tag{2.244}$$

Odpovídající kvadratické křivka je přitom elipsa, parabola, nebo hypeprbola právě tehdy, když Δ je záporné, rovné nule nebo kladné. Na základě této analogie bylo zavedeno i dělení pro PDR.

Příklady jednotlivých typů rovnic:

- Problém vedení tepla ve 2D je popsán následující rovnicí

$$\lambda \left(\frac{\partial^2 T}{\partial x_1^2} + \frac{\partial^2 T}{\partial x_2^2} \right) + Q = 0, \tag{2.245}$$

  kde vodivost λ předpokládáme kostantní. Porovnáním s obecným zápisem PDR 2. řádu (2.240) zjistíme, že $A = 1, B = 0, C = 1$ a tedy $\Delta = B^2 - 4AC = -4 < 0$ a rovnice je tedy eliptická.

- Problém podélného kmitání prutu je popsán rovnicí

$$\rho \frac{\partial^2 u}{\partial t^2} - \frac{\partial}{\partial x} \left(EA \frac{\partial u}{\partial x} \right) - f_x = 0, \tag{2.246}$$

  kde ρ a EA jsou konstanty popisující objemovou hustotu a tuhost prutu v tahu/tlaku. Opět porovnáním s obecným zápisem PDR zjistíme (pozor, zde je jedna proměnná časová, druhá prostorová - viz výše), že $A = 1, B = 0, C = -1$, diskriminant $\Delta = B^2 - 4AC = 4 > 0$ a rovnice je tedy hyperbolická.

- Příkladem parabolické rovnice může být časově závislý problém vedení tepla, jmenovitě

$$\rho c_v \frac{\partial T}{\partial t} - \frac{\partial}{\partial x} \left(\lambda \frac{\partial T}{\partial x} \right) + Q = 0, \tag{2.247}$$

  kde ρ je hustota a c_v je tepeln8 kapacita materiálu. Analogicky jako výše vidíme, že $A = 1, B = 0, C = 0, D = 0, E = -1, F = 0$ a tedy $\Delta = B^2 - 4AC = 0$ a dle definice jde tedy o úlohu parabolickou.

Uvedené dělení PDR není samoúčelné. Výběr příkladů rovnic není náhodný, tyto rovnice představují totiž tři základní fyzikální problémy - šíření vln, difúzi a ustálené problémy. Matematické metody řešení těchto problému jsou značně odlišné. Každou lineární PDR 2. řádu lze lineární transformací souřadnic převést na jeden z těchto tří typů. Konkrétně lze rovnici (2.240) pomocí transformace

$$\xi = \alpha x_2 + \beta x_1 \tag{2.248}$$

$$\eta = \gamma x_2 + \delta x_1, \tag{2.249}$$

kde $\alpha, \beta, \gamma\, \delta$ jsou reálná čísla, převést na tzv. kanonický tvar:

$$\frac{\partial^2 u}{\partial \xi^2} + \frac{\partial^2 u}{\partial \eta^2} = \Phi(\xi, \eta, u, u_\xi, u_\eta) \qquad \text{kanonický tvar eliptické rovnice} \tag{2.250}$$

$$\frac{\partial^2 u}{\partial \eta^2} = \Psi(\xi, \eta, u, u_\xi, u_\eta) \qquad \text{kanonický tvar parabolické rovnice} \tag{2.251}$$

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = \Theta(\xi, \eta, u, u_\xi, u_\eta) \qquad \text{kanonický tvar hyperbolické rovnice.} \tag{2.252}$$

Ve vztazích (2.250) - (2.252) představují Φ, Ψ, Θ funkce nových nezávisle proměnných ξ, η, závisle proměnné u a jejích prvních derivací, zde označovaných jako u_ξ, u_η. Z těchto vztahů je okamžitě vidět, že diskriminant je po řadě menší než nula, roven nule a větší než nula, jak odpovídá původní definici eliptické, parabolické a hyperbolické rovnici.

2.9 Několik obecných poznámek o numerických metodách

Jak již bylo zmíněno v úvodu, nalezení analytického řešení dané PDR není ve většině případů možné. Buď není možné nalézt řešení vůbec, nebo případně jen obecné řešení, ale už ne příslušné partikulární, kde může být složitá funkční závislost na okrajových podmínkách. Teorie diferenciálních rovnic se tak zabývá většinou jen otázkami existence, jednoznačnosti a regularity (závislosti řešení na vstupních datech, například z hlediska spojitosti) řešení. Nedává však návody, jak takovéto řešení prakticky nalézt. Tyto otázky řeší numerická matematika.

V numerických metodách jde o nalezení aproximace u_n přesného řešení u a o stanovení chyby, které se dopustíme při nahrazení přesného řešení jeho aproximací. Idea při hledání přibližného řešení je přitom ve všech numerických metodách pro řešení diferenciálních rovnic stejná. Původní spojitý (a tedy nekonečně dimenzionální) problém hledání funkce, která ve všech bodech dané oblasti vyhovuje zadaným rovnicím se transformuje (diskretizuje) na problém pouze konečně dimenzionální, kde hledáme funkci (aproximaci řešení), která vyhovuje zadaným rovnicím (které mohou být transformací rovněž dotčeny - viz předchozí kapitoly o MKP a MKD) pouze v konečně mnoha bodech. Index n u přibližného řešení u_n odkazuje obvykle na dimenzi transformovaného (přibližného) problému. Schematicky je možno situaci znázornit následovně:

$$\begin{aligned} \mathcal{P}(u, g) = 0 \quad & \text{Puvodní formulace PDR} \\ \downarrow \quad & \text{Numerická metoda} \\ \mathcal{P}_n(u_n, g_n) = 0. \quad & \text{Numerické schéma pro řešení PDR} \end{aligned}$$

Zde jsme jako g_n označili aproximaci vstupních dat (pravé strany) g a $\mathcal{P}_n$ funkční vztah charakterizující aproximaci původního problému, označeného $\mathcal{P}$. Konkrétní příklady numerických schémat a jejich analýzy jsou uvedeny v předchozích kapitolách o MKP a MKD. Zde vyzdvihneme pouze hlavní rysy a vlastnosti numerických metod.

O numerické metodě (přesněji spíše o konkrétním numerickém schématu, ale zde obecně budeme luvit o numerické metodě) řekneme, že je konvergentní, pokud v dané normě platí

$$||u - u_n|| \to 0 \quad pro \quad n \to \infty. \tag{2.253}$$

To zhruba řečeno znamená, že zvyšováním dimenze prostoru, ze kterého bereme aproximační funkce, se můžeme teoreticky dostat libovolně blízko k přesnému řešení (ve smyslu dané normy). Převedeno do exaktního "epsilon-delta" jazyka, numerická metoda je konvergentní, právě když

$$(\forall \varepsilon > 0)(\exists n_0 \in \mathbb{R})(\exists \delta > 0)(\forall n > n_0)(\forall g_n, ||g - g_n|| < \delta)(||u - u_n|| < \varepsilon). \tag{2.254}$$

Přímé ověření této podmínky (tzn. přímo z definice) však pro konkrétní případy numerických metod nemusí být jednoduché. Proto je vhodné mít k dispozici jiné podmínky, které konvergenci rovněž zaručují, avšak jejich ověření je snazší. Ukazuje se, že stačí ověřit platnost následujících dvou pojmů - konzistence a stabilita. Numerické metoda se nazývá konzistentní, pokud

$$\mathcal{P}_n(u, g) = 0 \quad \forall n \geq 1. \tag{2.255}$$

Podmínka konzistence požaduje, aby funkce u, která je řešením původního problému, byla řešením i aproximovného problému (konkrétního numerického schématu). Někdy se jako konzistence uvádí slabší podmínka

$$\mathcal{P}_n(u, g) \to 0 \quad pro \quad n \to \infty, \tag{2.256}$$

která konzistenci požaduje pro limitní aproximovaný problém (pokud s dimenzí jdeme do nekonečna). V tom případě se pak numerické podmínce splňující (2.255) říká plně konzistentní.

Numerická metoda se nazývá stabilní, pokud "malým" změnám vstupních dat (pravé straně) odpovídají "malé" změny řešení. Přesně řečeno jde o splnění následující podmínky:

$$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall \delta g_n, ||\delta g_n|| < \delta)(||\delta u_n|| < \varepsilon), \quad \forall n \geq 1. \tag{2.257}$$

Malé změně funkce (ať už řešení, nebo vstupních dat) se říká perturbace. V předchozí definici stability $u_n + \delta u_n$ je řešení perturbovaného problému

$$\mathcal{P}_n(u_n + \delta u_n, g_n + \delta g_n) = 0. \tag{2.258}$$

Hledanou ekvivalentní podmínkou pro zajištění konvergence je Laxova-Richtermayerova věta, která říká, že:

Věta (Lax-Richtermayer). *Pokud je numerická metoda konzistentní, pak je konvergentní právě tehdy, když je stabilní.*

Volba konkrétní metody pak závisí na dalších aspektech, jako je rychlost konvergence, vypočetní náročnost a další.