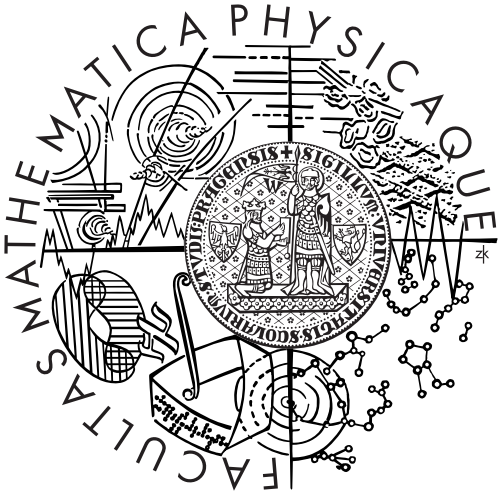
Univerzita Karlova v Praze
Matematicko–fyzikální fakulta

DIPLOMOVÁ PRÁCE



Pavel Kůs

Řešení konvektivně–difusních rovnic pomocí adaptivních metod vyšších řádů v prostoru a v čase

Katedra numerické matematiky
Vedoucí diplomové práce: Doc. RNDr. Vít Dolejší, Ph.D

Studijní obor: Výpočtová matematika

Na tomto místě bych rád poděkoval vedoucímu diplomové práce Doc. RNDr. Vítu Dolejšímu, Ph.D za jeho trpělivost, odborné rady a konzultace při tvorbě práce

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne Pavel Kůs

Obsah

Název práce: Řešení konvektivně–difusních rovnic pomocí adaptivních metod vyšších řádů v prostoru a v čase
Autor: Pavel Kůs
Katedra: Katedra numerické matematiky
Vedoucí diplomové práce: Doc. RNDr. Vít Dolejší, Ph.D
e-mail vedoucího: dolejsi@karlin.mff.cuni.cz
Abstrakt: Předmětem této práce je řešení skalární nelineární konvektivně–difusní rovnice pomocí nespojité Galerkinovy metody. Jejím cílem je implementace adaptivní volby časového kroku. Za tímto účelem jsou odvozeny 2 dostatečně stabilní metody pro řešení soustav obyčejných diferenciálních rovnic, které vzniknou prosorovou semidiskretizací po užití nespojité Galerkinovy metody. Na základě dvou přibližných řešení, získaných těmito metodami, je odvozen odhad lokální chyby diskretizace. Pomocí něj je pak volen následující časový krok tak, aby se lokální chyba co nejvíce blížila požadované předem zvolené toleranci. Je provedeno několik numerických simulací, které ověřují vlastnosti této metody.
Klíčová slova: konvektivně–difusní problémy, nespojitá Galerkinova metoda, časová adaptivita

Title: Solution of convection–diffusion equations with adaptive methods of higher order in space and time
Author: Pavel Kůs
Department: Department of Numerical Mathematics
Supervisor: Doc. RNDr. Vít Dolejší, Ph.D
Supervisor's e-mail address: dolejsi@karlin.mff.cuni.cz
Abstract: This thesis deals with solution of scalar nonlinear convection–diffusion equation with aid of discontinuous Galerkin method. It's aim is to implement an adaptive choice of time step. To do this, we derived 2 sufficiently stable methods for solution of systems of ordinary differential equations obtained by space semidicretization, which is carried out by the discontinuous Galerkin method. Using those two approximate solutions, we estimate local error of discretization. Using it, we are able to choose following time step in such way, that local error is approximately equal to given tolerance. Several numerical simulations were carried out to check properties of this method.
Key words: convection–diffusion problems, discontinuous Galerkin method, time adaptivity

Kapitola 1

Úvod

Řešení konvektivně–difusních problémů hraje významnou roli v mnoha různých vědních oborech. Patří mezi ně nejen přírodní vědy a technika (modelování proudění tekutin, hydrologie, ochrana životního prostředí, konstrukce letadel, atd.), ale i finanční matematika či zpracování obrazu. V minulosti byla vyvinuta řada metod určených k řešení těchto problémů (metoda konečných diferencí, konečných objemů, konečných prvků). Hlavní obtíž při řešení konvektivně–difusních problémů spočívá v nespojitostech či velkých gradientech v přesném řešení rovnic. Ukazuje se, že vhodnou metodou pro zachycení těchto jevů je nespojitá Galerkinova varianta metody konečných prvků (discontinuous Galerkin finite element method, DGFEM). Ta, podobně jako klasická metoda konečných prvků, rozdělí výpočetní oblast na konečné elementy a používá po částech polynomiální aproximaci na jednotlivých elementech sítě. Není však vyžadována spojitost řešení mezi sousedícími elementy. Místo toho jsou do definice přibližného řešení přidány jisté „stabilizační“ členy. Nespojitá Galerkinova metoda je popsána v článcích [1], [2] a [3], její základy jsou uvedeny v kapitole 2.

Důležitým rysem moderních numerických metod je jejich adaptivita, tedy schopnost v průběhu výpočtu odhadovat lokální chybu, nějakým způsobem na ni reagovat a tím ji udržovat v požadovaných mezích. Pro prostorovou adaptivitu se používá zjemňování sítě v místech, kde je odhad lokální chyby veliký a naopak použití hrubší sítě v místech, kde odhad lokální chyby vychází malý. Tím se šetří výpočetní čas, jemná síť je totiž použita pouze tam, kde je skutečně potřeba a tím se snižuje počet elementů sítě. Podobným způsobem můžeme využít časovou adaptivitu u nestacionárních úloh, kdy po každém kroku výpočtu odhadneme lokální chybu a na jejím základě určíme délku následujícího časového kroku. Tím se zefektivní výpočet, protože se

vždy použije jen tak velký čaový krok, jaký je potřeba a tím se šetří výpočetní čas. Dříve se časový krok volil čistě experimentálně.

Právě implementace časové adaptivity do již existujícího programového balíku pro řešení skalární nelineární konvektivně–difusní rovnice je *vlastní* náplní této práce. Pro prostorovou semidiskretizaci je použita nespojitá Galerkinova metoda. Pro časovou diskretizaci byla původně použita metoda BDF (backward difference formulae), viz článek [4]. Abychom však byli schopni v každém časovém kroku odhadnout velikost lokální chyby, potřebujeme mít k dispozici dvojici metod. Z rozdílu řešení získaných pomocí nich pak odhadneme velikost chyby. Tyto metody (označeny jako BDF2) jsou odvozeny v kapitole 3. Na rozdíl od článku [10], kde je použita dvojice implicitní a explicitní metody, jsou obě naše metody implicitní. Řešený problém je totiž typu „stiff“ a jeho řešení vyžaduje metody s dostatečně širokou oblastí stability, což explicitní metody jistě nejsou. Zkoumání stability našich metod je věnována kapitola 3.4.

Algoritmus adaptivní volby časového kroku a programová realizace jsou popsány v kapitole 4. Některé numerické výsledky získané pomocí nové metody jsou uvedeny v kapitole 5. Jsou zde vyšetřovány jednak vlastnosti metody BDF2 a také srovnávána efektivita adaptivní a neadaptivní metody. Obojí je prováděno na příkladech obyčejných diferenciálních rovnic i skalární nelineární konvektivně–difusní rovnice.

Kapitola 2

Nespojitá Galerkinova metoda

2.1 Spojitý problém

Předpokládejme, že $\Omega \subset I\!R^d,\ d = 2, 3$, je omezená oblast tvaru mnohoúhelníku (pokud $d = 2$) nebo mnohostěnu (pokud $d = 3$) s Lipschitzovsky spojitou hranicí $\partial\Omega$ a $T > 0$. Položme $Q_T = \Omega \times (0, T)$. $\overline{\Omega}$ značí uzávěr a $\partial\Omega$ hranici oblasti Ω. Uvažujme následující nestacionární, nelineární, konvektivně–difusní problém: Najít $u : Q_T \to I\!R$ takové, že

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} = \varepsilon\, \Delta u + g \quad \text{in } Q_T, \tag{2.1}$$

$$u\,\Big|_{\partial\Omega\times(0,T)} = u_D, \tag{2.2}$$

$$u(x, 0) = u^0(x), \quad x \in \Omega. \tag{2.3}$$

Předpokládejme, že data úlohy splňují následující podmínky:

$$\begin{aligned}
&\text{a)} \quad f_s \in C^1(I\!R),\ f_s(0) = 0,\ s = 1, \ldots, d, \\
&\text{b)} \quad \varepsilon > 0, \\
&\text{c)} \quad g \in C([0, T]; L^2(\Omega)), \\
&\text{d)} \quad u_D \text{ je stopa funkce } u^* \in C([0, T]; H^1(\Omega)) \cap L^\infty(Q_T) \\
&\qquad\ \ \text{na } \partial\Omega \times (0, T), \\
&\text{e)} \quad u^0 \in L^2(\Omega).
\end{aligned} \tag{2.4}$$

Používáme obvyklé značení pro prostory funkcí (viz např. [11]): $L^p(\Omega)$, $L^p(Q_T)$ značí Lebesgueův prostor , $W^{k,p}(\Omega)$, $H^k(\Omega) = W^{k,2}(\Omega)$ jsou Sobolevovy prostory, $L^p(0, T; X)$ je Bochnerův prostor funkcí p-integrovatelných na intervalu $(0, T)$ s hodnotami v Banachově prostoru X, $C([0, T]; X)$ $(C^1([0, T]; X))$

je prostor spojitých (spojitě diferencovatelných) zobrazení intervalu $[0,T]$ do X. Symbolem $H_0^1(\Omega)$ značíme podprostor všech funkcí z $H^1(\Omega)$ s nulovými stopami na $\partial\Omega$.

Předpoklad že $f_s(0) = 0$, $s = 1, \ldots, d$ neznamená žádnou ztrátu obecnosti, jak je vidět z rovnice (2.1). Funkce f_s, zvané toky, reprezentují konvektivní členy, $\varepsilon > 0$ je difusní koeficient. Difusní člen může být obecně komplikovanější, v některých případech dokonce nelineární. Je také možné uvažovat smíšené okrajové podmínky (na části hranice je předepsaná Dirichletova a na části Neumannova okrajová podmínka). Zde pro jednoduchost uvažujeme Dirichletovu podmínku na celé hranici.

2.1.1 Slabé řešení

Dostatečně hladká funkce bodově splňující (2.1) – (2.3) se nazývá *klasické řešení*. Pro použití v metodě konečných prvků je potřeba zavést slabé řešení. K tomu použijeme následující značení:

$$(u, v) = \int_\Omega uv \,\mathrm{d}x, \quad u,\, v \in L^2(\Omega) \tag{2.5}$$

(skalární součin v $L^2(\Omega)$),

$$\|u\|_{L^2(\Omega)} = (u, u)^{1/2} \tag{2.6}$$

(norma v $L^2(\Omega)$),

$$a(u, v) = \varepsilon \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \quad u,\, v \in H^1(\Omega), \tag{2.7}$$

$$b(u, v) = \int_\Omega \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} v \,\mathrm{d}x, \quad u \in H^1(\Omega) \cap L^\infty(\Omega),\ v \in L^2(\Omega), \tag{2.8}$$

$$\|u\|_{H^1(\Omega)} = \left(\int_\Omega \left(|u|^2 + |\nabla u|^2 \right) \mathrm{d}x \right)^{1/2}, \quad u \in H^1(\Omega), \tag{2.9}$$

(norma v $H^1(\Omega)$),

$$|u|_{H^1(\Omega)} = \left(\int_\Omega |\nabla u|^2 \,\mathrm{d}x \right)^{1/2}, \quad u \in H^1(\Omega), \tag{2.10}$$

(seminorma v $H^1(\Omega)$). Je známo, že $|\cdot|_{H^1(\Omega)}$ je norma na $H_0^1(\Omega)$ ekvivalentní s $\|\cdot\|_{H^1(\Omega)}$.

Definice Řekneme, že funkce u je *slabým řešením* problému (2.1)–(2.3), pokud jsou splněny následující podmínky

$$\text{a)} \qquad u - u^* \in L^2(0,T;H_0^1(\Omega)), \quad u^* \in L^\infty(Q_T), \tag{2.11}$$

$$\text{b)} \qquad \frac{\mathrm{d}}{\mathrm{d}t}(u(t),v) + b(u(t),v) + a(u(t),v) = (g(t),v)$$

pro každé $v \in H_0^1(\Omega)$ ve smyslu distribucí na $(0,T)$,

$$\text{c)} \qquad u(0) = u_0 \quad \text{na } \Omega.$$

Symbolem $u(t)$ značíme funkci na Ω takovou, že $u(t)(x) = u(x,t)$, $x \in \Omega$.

S pomocí postupů obsažených v [8] a [9] se dá dokázat existence a jednoznačnost slabého řešení. To navíc splňuje podmínku $\partial u/\partial t \in L^2(Q_T)$. (2.11), b) může být tedy přepsáno jako

$$\left(\frac{\partial u(t)}{\partial t}, v\right) + b(u(t),v) + a(u(t),v) = (g(t),v) \tag{2.12}$$

pro každé $v \in H_0^1(\Omega)$ a skoro všechna $t \in (0,T)$.

2.2 Prostor funkcí po částech Sobolevovských

Nechť $\mathcal{T}_h$ ($h > 0$) značí triangulaci uzávěru $\overline{\Omega}$ oblasti Ω na konečný počet uzavřených trojúhelníků (pokud $d = 2$) nebo čtyřstěnů (pokud $d = 3$) K se vzájemně disjunktními vnitřky. Položme $h = \max_{K\in\mathcal{T}_h} \operatorname{diam}(K)$. Všechny elementy z $\mathcal{T}_h$ očíslujeme tak, že $\mathcal{T}_h = \{K_i\}_{i\in I}$, kde I je vhodná indexová množina. Dva elementy K_i, $K_j \in \mathcal{T}_h$ nazýváme *sousedy*, pokud je jejich průnik neprázdná otevřená část jejich hran. V takovém případě položíme $\Gamma_{ij} = \Gamma_{ji} = \partial K_i \cap \partial K_j$. Pro $i \in I$ klademe $s(i) = \{j \in I; K_j \text{ je soused } K_i\}$. Hranice $\partial\Omega$ je tvořena konečným počtem hran elementů K_i hraničících s $\partial\Omega$. Všechny tyto hrany označíme jako S_j, kde $j \in I_b$ je vhodná indexová množina a položíme $\gamma(i) = \{j \in I_b; S_j \text{ je hrana } K_i\}$, $\Gamma_{ij} = S_j$ pro $K_i \in \mathcal{T}_h$ takové, že $S_j \subset \partial K_i$, $j \in I_b$. Pro K_i neobsahující žádnou hranu S_j položíme $\gamma(i) = \emptyset$. Dále položíme $S(i) = s(i) \cup \gamma(i)$ a $\boldsymbol{n}_{ij} = ((n_{ij})_1, \ldots, (n_{ij})_d)$ značí jednotkovou vnější normálu k ∂K_i na hraně Γ_{ij}.

Na triangulaci $\mathcal{T}_h$ definujeme tak zvaný *prostor funkcí po částech Sobolevovských (broken Sobolev space)*

$$H^k(\Omega,\mathcal{T}_h) = \{v; v|_K \in H^k(K)\ \forall\, K \in \mathcal{T}_h\}, \tag{2.13}$$

s normou

$$\|v\|_{H^k(\Omega,\mathcal{T}_h)} = \left(\sum_{K\in\mathcal{T}_h} \|v\|^2_{H^k(K)}\right)^{1/2} \tag{2.14}$$

a seminormou

$$|v|_{H^k(\Omega,\mathcal{T}_h)} = \left(\sum_{K\in\mathcal{T}_h} |v|^2_{H^k(K)} \right)^{1/2} \tag{2.15}$$

kde $H^k(K) = W^{k,2}(K)$ značí (klasický) Sobolevův prostor na elementu K. Pro $v \in H^1(\Omega, \mathcal{T}_h)$ položíme

$$\begin{aligned} v|_{\Gamma_{ij}} &= \text{stopa funkce } v|_{K_i} \text{ na } \Gamma_{ij}, \\ v|_{\Gamma_{ji}} &= \text{stopa funkce } v|_{K_j} \text{ na } \Gamma_{ji}, \\ \langle v\rangle_{\Gamma_{ij}} &= \frac{1}{2}\left(v|_{\Gamma_{ij}} + v|_{\Gamma_{ji}}\right), \\ [v]_{\Gamma_{ij}} &= v|_{\Gamma_{ij}} - v|_{\Gamma_{ji}}. \end{aligned} \tag{2.16}$$

Je zřejmé, že $\langle v\rangle_{\Gamma_{ij}} = \langle v\rangle_{\Gamma_{ji}}$, ale $[v]_{\Gamma_{ij}} = -[v]_{\Gamma_{ji}}$. Naopak $[v]_{\Gamma_{ij}}\boldsymbol{n}_{ij} = [v]_{\Gamma_{ji}}\boldsymbol{n}_{ji}$.

2.3 Prostorová semidiskretizace

Nejdříve diskretizujeme rovnici (2.1) vzhledem k prostorovým složkám. Použijeme takzvanou Galerkinovu metodu s nesymetrickou stabilizací difusních členů a vnitřní penalizací (NIPG), která sice nedává optimální apriorní řád konvergence v L^2-normě, ale její výhodda spočívá v tom, že je koercivní pro každý kladný penalizační koeficient σ, viz [5], [6]. To je důležité pro případ Navier-Stokesových rovnic, kde je numerická analýza nemožná a volba σ je spíše heuristická. Podrobnější popis NIPG může být nalezen například v [2], [3]. Zde uvedeme pouze definici přibližného řešení. Pro $u,\, v \in H^2(\Omega, \mathcal{T}_h)$, $u \in L^\infty(\Omega)$ definujeme formy

$$\begin{aligned} a_h(u,\varphi) &= \varepsilon \sum_{i\in I} \left\{ \int_{K_i} \nabla u\cdot\nabla\varphi \,\mathrm{d}x - \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \left(\langle\nabla u\rangle\cdot\boldsymbol{n}_{ij}[\varphi] - \langle\nabla\varphi\rangle\cdot\boldsymbol{n}_{ij}[u]\right) \mathrm{d}S \right. \\ &\qquad \left. - \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} \left(\nabla u\cdot\boldsymbol{n}_{ij}\,\varphi \,\mathrm{d}S - \nabla\varphi\cdot\boldsymbol{n}_{ij}\,u\right) \mathrm{d}S \right\}, \\ b_h(u,\varphi) &= \sum_{i\in I} \left\{ \sum_{j\in S(i)} \int_{\Gamma_{ij}} H\left(u|_{\Gamma_{ij}}, u|_{\Gamma_{ji}}, \boldsymbol{n}_{ij}\right)\varphi|_{\Gamma_{ij}} \,\mathrm{d}S - \int_{K_i} \sum_{s=1}^{d} f_s(u)\,\frac{\partial\varphi}{\partial x_s}\,\mathrm{d}x \right\}, \\ J_h^\sigma(u,\varphi) &= \sum_{i\in I} \left\{ \sum_{\substack{j\in s(i)\\ j<i}} \int_{\Gamma_{ij}} \sigma[u]\,[\varphi]\,\mathrm{d}S + \sum_{j\in\gamma(i)} \int_{\Gamma_{ij}} \sigma\,u\,\varphi\,\mathrm{d}S \right\}, \\ \ell_h(\varphi)(t) &= \int_\Omega g(t)\,\varphi\,\mathrm{d}x + \varepsilon \sum_{i\in I}\sum_{j\in\gamma} \int_{\Gamma_{ij}} (\nabla\varphi\cdot\boldsymbol{n}_{ij}\,u_D + \sigma\,u_D\,\varphi)\,\mathrm{d}S, \end{aligned}$$

kde σ je definována jako $\sigma|_{\Gamma_{ij}} = 1/\mathrm{diam}(\Gamma_{ij})$, $j \in S(i),\ i \in I$.

2.3.1 Numerický tok

Konvektivní hraniční členy jsou aproximovány pomocí *numerického toku* $H = H(u, v, \boldsymbol{n})$ který je znám z metody konečných objemů, viz např. [7]. Předpokládejme, že numerický tok má následující vlastnosti:

1. $H(u, v, \boldsymbol{n})$ je definováno v $\mathbb{R}^2 \times \boldsymbol{S}_1$, kde $\boldsymbol{S}_1 = \{\boldsymbol{n} \in \mathbb{R}^d; |\boldsymbol{n}| = 1\}$ a je Lipschitzovsky spojité vzhledem k u a v, tedy existuje konstanta $C_1 > 0$ taková, že

$$\begin{aligned} |H(u, v, \boldsymbol{n}) - H(u^*, v^*, \boldsymbol{n})| &\leq C_1(|u - u^*| + |v - v^*|), \\ u, v, u^*, v^* &\in \mathbb{R} \quad , \quad \boldsymbol{n} \in \boldsymbol{S}_1 \end{aligned} \tag{2.17}$$

2. $H(u, v, \boldsymbol{n})$ je konzistentní:

$$H(u, u, \boldsymbol{n}) = \sum_{s=1}^{d} f_s(u) n_s, \quad u \in \mathbb{R}, \boldsymbol{n} = (n_1 \dots n_d) \in \boldsymbol{S}_1 \tag{2.18}$$

3. $H(u, v, \boldsymbol{n})$ je konzervativní:

$$H(u, v, \boldsymbol{n}) = -H(v, u, \boldsymbol{n}), \quad u, v \in \mathbb{R}, \boldsymbol{n} \in \boldsymbol{S}_1 \tag{2.19}$$

V numerických experimentech byl použit numerický tok ve tvaru

$$H(u, v, \boldsymbol{n}) = \begin{cases} \sum_{s=1}^{2} f_s(u_1) n_s, & \text{pro } A > 0 \\ \sum_{s=1}^{2} f_s(u_2) n_s, & \text{pro } A \leq 0 \end{cases}, \tag{2.20}$$

kde $f_s(x) = x^2/2$, $A = \sum_{s=1}^{2} f_s'(\bar{u}) n_s$, $\bar{u} = 1/2(u + v)$.

2.3.2 Semidiskrétní řešení

Pro jednoduchost označme součet všech lineárních forem z (2.17) jako

$$A_h(u(t), v) \equiv a_h(u(t), v) + \varepsilon J_h^{\sigma}(u(t), v) - \ell_h(v)(t), \quad u(t), v \in H^2(\Omega, \mathcal{T}). \tag{2.21}$$

Přibližné řešení problému (2.1)-(2.3) hledáme v prostoru nespojitých, po částech polynomiálních funkcí S_h definovaných jako

$$S_h = S^{p,-1}(\Omega, \mathcal{T}_h) = \{v; v|_K \in P^p(K) \; \forall \, K \in \mathcal{T}_h\}, \tag{2.22}$$

kde p je kladné celé číslo a $P^p(K)$ značí prostor všech polynomů na K stupně nejvýše p. Zřejmě platí $S_h \subset H^2(\Omega, \mathcal{T})$.

Nyní můžeme zavést *semidiskrétní problém*.

Definice Funkce u_h je *semidiskrétní řešení* problému (2.1)-(2.3), pokud

$$\text{a)} \quad u_h \in C^1([0,T]; S_h), \tag{2.23}$$

$$\text{b)} \quad \left(\frac{\partial u_h(t)}{\partial t}, \varphi_h\right) + b_h(u_h(t), \varphi_h) + A_h(u_h(t), \varphi_h) = 0 \quad \forall\, \varphi_h \in S_h,\ \forall\, t \in (0,T),$$

$$\text{c)} \quad u_h(0) = u_h^0,$$

kde $u_h^0 \in S_h$ značí S_h-aproximaci počáteční podmínky u^0.

Výše uvedený diskrétní problém byl odvozen za pomoci *metody přímek*. Získaná soustava obyčejných diferenciálních rovnic (2.23), a) – c) musí být řešena pomocí vhodné numerické metody, čímž se zabýváme v kapitole 3.

Kapitola 3

BDF2 metoda

Naším cílem je odvodit dostatečně přesnou a efektivní numerickou metodu pro soustavu obyčejných diferenciálních rovnic (2.23), a) – c). Tato soustava je typu „stiff“, pro odvození časo-prostorové diskretizace je proto třeba použít vhodnou metodu s dostatečně velkou oblastí stability. V [4] jsou k diskretizaci použity metody typu BDF (backward difference formulae). V této práci tento postup rozšiřujeme o adaptivní volbu časového kroku, k čemuž potřebujeme být schopni odhadnout lokální chybu metody v každém časovém kroku. Z tohoto důvodu doplníme původní BDF metodu další implicitní metodou stejného řádu přesnosti a z rozdílu přibližných řešení získaných pomocí obou metod budeme odhadovat lokální diskretizační chybu a volit časový krok.

3.1 Odvození koeficientů

Uvažujme následující soustavu obyčejných diferenciálních rovnic s neznámou funkcí $y : (0,T) \to I\!R^m$

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = F(t,y), \qquad y(0) = y^0 \tag{3.1}$$

kde $y^0 \in I\!R^m$ a $F : (0,T) \times I\!R^m \to I\!R^m$. Odvodíme n krokovou metodu s proměnným časovým krokem. Nechť $0 = t_0 < t_1 < t_2 < \ldots t_r = T$ je dělení intervalu $(0,T)$, $\tau_k \equiv t_k - t_{k-1},\ k = 1,\ldots,r$, $\theta_k = \tau_k/\tau_{k-1}$, $k = 1 \ldots r$. Symbol y_k značí přibližnou hodnotu řešení $y(t_k)$. Zavedeme ještě řád metody.

Definice Řekneme, že metoda pro řešení soustavy obyčejných diferenciálních rovnic (3.1) je řádu m právě tehdy, když lokální chyba $y_k - y(t_k) = O(\tau_k^{m+1})$.

3.1.1 BDF2a metoda

Vyjdeme z Taylorova rozvoje funkce y se středem v bodě t_k. Předpokládejme, že $y \in C^{n+2}(0,T)$.

$$\begin{aligned} y(t_{k-1}) &= \sum_{i=0}^{n+1}(-1)^i \frac{\tau_k^i}{(i)!} y^{(i)}(t_k) + O(\tau_k^{n+2}) \qquad (3.2)\\ y(t_{k-2}) &= \sum_{i=0}^{n+1}(-1)^i \frac{(\tau_k+\tau_{k-1})^i}{(i)!} y^{(i)}(t_k) + O\Big((\tau_k+\tau_{k-1})^{n+2}\Big) \\ &\vdots \\ y(t_{k-n}) &= \sum_{i=0}^{n+1}(-1)^i \frac{(\tau_k+\cdots+\tau_{k-n+1})^i}{(i)!} y^{(i)}(t_k) + O\Big((\tau_k+\cdots+\tau_{k-n+1})^{n+2}\Big) \end{aligned}$$

Vhodnou lineární kombinací těchto rovnic získáme vztah, ve kterém nevystupují derivace $y''(t_k)$,..., $y^{(n)}(t_k)$ (máme soustavu n rovnic, můžeme z ní tedy vyeliminovat $n-1$ proměnných). Zřejmě můžeme uvažovat $O\Big((\tau_k)^{n+2}\Big) = O\Big((\tau_k+\tau_{k-1})^{n+2}\Big) = \cdots = O\Big((\tau_k+\cdots+\tau_{k-n+1})^{n+2}\Big)$. Pak z (3.2) dostáváme

$$\sum_{i=0}^{n} \alpha_i y(t_{k-i}) = \tau_k y'(t_k) + d y^{(n+1)} + O(\tau_k^{n+2}), \qquad (3.3)$$

kde

$$d = \sum_{i=1}^{n} \alpha_i (\tau_k + \cdots + \tau_{k-i+1})^{n+1}. \qquad (3.4)$$

Pro jednoduchost budeme uvažovat

$$d \approx C\tau_k^{n+1} = O(\tau_k^{n+1}) \qquad (3.5)$$

Konkrétní vztahy pro koeficienty α_i jsou uvedeny v následujících kapitolách.

Označme ještě $F_k = F\Big(t_k, y_k\Big)$. Z rovnice (3.3) získáme metodu

$$\sum_{i=0}^{n} \alpha_i y_{n-i} = \tau_k F_k \qquad (3.6)$$

Nyní odvodíme odhad lokální diskretizační chyby. Předpokládejme, že $y_{k-i} = y(t_{k-i})$, $i = 1 \dots n$. Nechť dále $F\Big(t_k, y(t_k)\Big) = F_k$. Odečteme (3.6) od (3.3)

$$\alpha_0\Big(y(t_k) - u_k\Big) = dy^{(n+1)} + O(\tau_k^{n+2}), \tag{3.7}$$

zanedbáme $O(\tau_k^{n+2})$ a máme odhad lokální diskretizační chyby

$$e_k \equiv y(t_k) - y_k \approx \frac{d}{\alpha_0} y^{(n+1)}(t_k) = O(\tau_k^{n+1}). \tag{3.8}$$

Je totiž $d = O(\tau_k^{n+1})$ a $\alpha_0 = O(1)$, jak je vidět z tabulek 3.1, 3.2 a 3.3. Hodnoty α_0 totiž závisí pouze na hodnotách θ_k, $k = 1 \dots r$ a ty jsou zřejmě řádu $O(1)$. metoda je tedy řádu n.

V kapitole 3.1.2 odvodíme další podobnou metodu. Aby byly obě metody formálně zapsány ve stejném tvaru, položíme $\gamma_0 = 1$, $\gamma_1 = 0$ a metodu BDF2a můžeme nyní zapsat jako

$$\sum_{i=0}^{n} \alpha_i y_{n-i} = \gamma_0 \tau_k F_k + \gamma_1 \tau_k F_{k-1}. \tag{3.9}$$

3.1.2 BDF2b metoda

Abychom mohli odhadovat lokální chybu diskretizace, odvodíme ještě jednu implicitní metodu stejného řádu přesnosti. Z rozdílu řešení získaných pomocí těchto dvou metod později odvodíme odhad lokální chyby diskretizace.

Tentokrát vyjdeme z Taylorova rozvoje se středem v bodě t_{k-1} :

$$\begin{aligned}
y(t_k) &= \sum_{i=0}^{n+1} \frac{\tau_k^i}{(i)!} y^{(i)}(t_{k-1}) + O(\tau_k^{n+2}) \qquad (3.10)\\
y(t_{k-2}) &= \sum_{i=0}^{n+1} (-1)^i \frac{\tau_{k-1}^i}{(i)!} y^{(i)}(t_{k-1}) + O(\tau_{k-1}^{n+2})\\
&\vdots\\
y(t_{k-n}) &= \sum_{i=0}^{n+1} (-1)^i \frac{(\tau_{k-1} + \dots + \tau_{k-n+1})^i}{(i)!} y^{(i)}(t_{k-1}) +\\
&\qquad + O\Big((\tau_{k-1} + \dots + \tau_{k-n+1})^{n+2}\Big).
\end{aligned}$$

Podobně jako v předchozí části úpravami těchto rovnic získáme

$$\sum_{i=0}^{n} \bar{\alpha}_i y(t_{k-i}) = \tau_k y'(t_{k-1}) + \bar{d} y^{(n+1)} + O(\tau_k^{n+2}), \tag{3.11}$$

kde

$$\bar{d} = \sum_{i=1}^{n} \bar{\alpha}_i (\tau_k + \cdots + \tau_{k-i+1})^{n+1}. \tag{3.12}$$

Pro jednoduchost budeme opět uvažovat

$$\bar{d} \approx C\tau_k^{n+1} = O(\tau_k^{n+1}). \tag{3.13}$$

Můžeme tedy definovat metodu

$$\sum_{i=0}^{n} \bar{\alpha}_i \bar{y}_{n-i} = \tau_k F_{k-1}. \tag{3.14}$$

Tato metoda je explicitní, zřejmě tedy bude pouze podmíněně stabilní. Pro řešení „stiff“ úloh je však širší stabilita metody zcela nezbytná. Metodu BDF2b tedy definujeme jako kombinaci metody BDF2a a předchozí explicitní metody takto :

$$\sum_{i=0}^{n} \hat{\alpha}_i \hat{y}_{n-i} = \hat{\gamma}_0 \tau_k F_k + \hat{\gamma}_1 \tau_k F_{k-1}, \tag{3.15}$$

kde $\hat{\gamma}_0 > 0$, $\hat{\gamma}_1 > 0$, $\hat{\gamma}_0 + \hat{\gamma}_1 = 1$ a $\hat{\alpha}_i = \hat{\gamma}_0 \alpha_i + \hat{\gamma}_1 \bar{\alpha}_i$ pro $i = 0 \dots n$. Odhad pro lokální diskretizační chybu metody BDF2b získáme jako $\hat{\gamma}_0(3.3) + \hat{\gamma}_1(3.11) -$ (3.15) :

$$\hat{\alpha}_0 (y(t_k) - \hat{y}_k) = \hat{d} y^{(n+1)} + O(\tau_k^{n+2}), \tag{3.16}$$

kde $\hat{d} = \hat{\gamma}_0 d + \hat{\gamma}_1 \bar{d} = O(\tau_k^{n+1})$. Po zanedbání $O(\tau_k^{n+2})$ získáme odhad pro chybu

$$\hat{e}_k \equiv y(t_k) - \hat{y}_k \approx \frac{\hat{d}}{\hat{\alpha}_0} y^{(n+1)}(t_{k-1}) \tag{3.17}$$

a metoda je opět řádu n.

3.1.3 Lokální chyba diskretizace

Ještě potřebujeme odvodit odhad lokální chyby diskretizace pomocí hodnot, které budeme mít v průběhu výpočtu k dispozici, tedy pomocí přibližných řešení získaných pomocí obou metod. Předpokládejme, že $y^{(n+1)}(t_k) \approx y^{(n+1)}(t_{k-1})$ a označme tuto hodnotu jako $y^{(n+1)}$. Odečtením (3.8) od (3.17) získáme

$$y_k - \hat{y}_k \approx \left(\frac{\hat{d}}{\hat{\alpha}_0} - \frac{d}{\alpha_0} \right) y^{(n+1)} \tag{3.18}$$

a tedy

$$y^{(n+1)} \approx \frac{1}{\left(\frac{\hat{d}}{\hat{\alpha}_0} - \frac{d}{\alpha_0} \right)} (y_k - \hat{y}_k). \tag{3.19}$$

		proměnný krok		konstantní krok	
		BDF2a	BDF2b	BDF2a	BDF2b
α_0	$\hat{\alpha}_0$	1	1	1	1
α_1	$\hat{\alpha}_1$	-1	-1	-1	-1
δ	$\hat{\delta}$	$-\frac{3}{2}$	$-\frac{1}{2}$	$-\frac{3}{2}$	$-\frac{1}{2}$

Tabulka 3.1: Hodnoty koeficientů pro $n = 1$ pro volbu $\hat{\gamma}_0 = \frac{2}{3}$, $\hat{\gamma}_0 = \frac{1}{3}$ pro proměnný a konstantní časový krok

Máme tedy odhad

$$e_k \approx \frac{d}{\alpha_0} y^{(n+1)} = \delta(y_k - \hat{y}_k), \tag{3.20}$$

kde $\delta = \frac{d}{\alpha_0} \frac{1}{\left(\frac{\hat{d}}{\hat{\alpha}_0} - \frac{d}{\alpha_0}\right)}$. Obdobně získáme

$$\hat{e}_k \approx \frac{\hat{d}}{\hat{\alpha}_0} y^{(n+1)} = \hat{\delta}(y_k - \hat{y}_k), \tag{3.21}$$

kde $\hat{\delta} = \frac{\hat{d}}{\hat{\alpha}_0} \frac{1}{\left(\frac{\hat{d}}{\hat{\alpha}_0} - \frac{d}{\alpha_0}\right)}$.

3.1.4 Finální řešení

Kombinací řešení získaných pomocí metod BDF2a a BDF2b můžeme získat řešení s vyšší asymptotickou přesností. Sečtěmě $\frac{\hat{d}}{\alpha_0 \hat{\alpha}_0}(3.7) - \frac{d}{\alpha_0 \hat{\alpha}_0}(3.16)$:

$$\frac{\hat{d}}{\hat{\alpha}_0} y_k - \frac{d}{\alpha_0} \hat{y}_k - \left(\frac{\hat{d}}{\hat{\alpha}_0} - \frac{d}{\alpha_0}\right) y(t_k) = O(\tau_k^{n+2}) \tag{3.22}$$

a tedy

$$\hat{\delta} y_k - \delta \hat{y}_k - y(t_k) = O(\tau_k^{n+2}). \tag{3.23}$$

Definujeme-li

$$\breve{y}_k = \hat{\delta} y_k - \delta \hat{y}_k, \tag{3.24}$$

je zřejmě

$$y(t_k) - \breve{y}_k = O(\tau_k^{n+2}) \tag{3.25}$$

a toto řešení je řádu $n + 1$.

		proměnný krok		konstantní krok	
		BDF2a	BDF2b	BDF2a	BDF2b
α_0	$\hat{\alpha}_0$	$\frac{2\,\theta_0+1}{\theta_0+1}$	1	$\frac{3}{2}$	1
α_1	$\hat{\alpha}_1$	$-\theta_0-1$	-1	-2	-1
α_2	$\hat{\alpha}_2$	$\frac{{\theta_0}^2}{\theta_0+1}$	0	$\frac{1}{2}$	0
δ	$\hat{\delta}$	$-2\,\frac{{\theta_0}^2+2\,\theta_0+1}{3\,\theta_0+2}$	$-\frac{\theta_0(2\,\theta_0+1)}{3\,\theta_0+2}$	$-\frac{8}{5}$	$-\frac{3}{5}$

Tabulka 3.2: Hodnoty koeficientů pro $n = 2$ pro volbu $\hat{\gamma}_0 = \frac{1}{2}$, $\hat{\gamma}_0 = \frac{1}{2}$ pro proměnný a konstantní časový krok

		proměnný krok		konstantní krok	
		BDF2a	BDF2b	BDF2a	BDF2b
α_0	$\hat{\alpha}_0$	$\frac{4\,\theta_0\theta_1+3\,{\theta_0}^2\theta_1+\theta_1+1+2\,\theta_0}{\theta_0+2\,\theta_0\theta_1+1+\theta_1+{\theta_0}^2\theta_1}$	$\frac{1}{2}\,\frac{4\,\theta_0\theta_1+3\,{\theta_0}^2\theta_1+2\,\theta_1+2+2\,\theta_0}{\theta_0+2\,\theta_0\theta_1+1+\theta_1+{\theta_0}^2\theta_1}$	$\frac{11}{6}$	$\frac{13}{12}$
α_1	$\hat{\alpha}_1$	$-\frac{\theta_0+2\,\theta_0\theta_1+1+\theta_1+{\theta_0}^2\theta_1}{1+\theta_1}$	$-\frac{1}{2}\,\frac{2+2\,\theta_1+{\theta_0}^2\theta_1}{1+\theta_1}$	-3	$-\frac{5}{4}$
α_2	$\hat{\alpha}_2$	$\frac{(\theta_1+\theta_0\theta_1+1){\theta_0}^2}{1+\theta_0}$	$\frac{1}{2}\,\frac{{\theta_0}^3\theta_1}{1+\theta_0}$	$\frac{3}{2}$	$\frac{1}{4}$
α_3	$\hat{\alpha}_3$	$-\frac{(1+\theta_0){\theta_0}^2{\theta_1}^3}{\theta_0{\theta_1}^2+\theta_0\theta_1+2\,\theta_1+1+{\theta_1}^2}$	$-\frac{1}{2}\,\frac{{\theta_0}^3{\theta_1}^3}{\theta_0{\theta_1}^2+\theta_0\theta_1+2\,\theta_1+1+{\theta_1}^2}$	$-\frac{1}{3}$	$-\frac{1}{12}$
δ		$-\frac{3\,\theta_0{}^4{\theta_1}^2+10\,\theta_0{}^3{\theta_1}^2+13\,\theta_0{}^2{\theta_1}^2+8\,\theta_0{\theta_1}^2+2\,{\theta_1}^2+5\,\theta_0{}^3\theta_1+13\,\theta_0{}^2\theta_1}{4\,\theta_0{}^2\theta_1+9\,\theta_0\theta_1+3\,\theta_0+4\,\theta_0{}^2{\theta_1}^2+6\,\theta_0{\theta_1}^2+2+4\,\theta_1+2\,{\theta_1}^2}-$ $-\frac{12\,\theta_0\theta_1+4\,\theta_1+2\,\theta_0{}^2+4\,\theta_0+2}{4\,\theta_0{}^2\theta_1+9\,\theta_0\theta_1+3\,\theta_0+4\,\theta_0{}^2{\theta_1}^2+6\,\theta_0{\theta_1}^2+2+4\,\theta_1+2\,{\theta_1}^2}$		$-\frac{39}{17}$	
$\hat{\delta}$		$-\frac{\theta_0\left(3\,\theta_0{}^3{\theta_1}^2+10\,\theta_0{}^2{\theta_1}^2+9\,\theta_0{\theta_1}^2+2\,{\theta_1}^2+5\,\theta_0{}^2\theta_1+9\,\theta_0\theta_1+3\,\theta_1+2\,\theta_0+1\right)}{4\,\theta_0{}^2\theta_1+9\,\theta_0\theta_1+3\,\theta_0+4\,\theta_0{}^2{\theta_1}^2+6\,\theta_0{\theta_1}^2+2+4\,\theta_1+2\,{\theta_1}^2}$		$-\frac{22}{17}$	

Tabulka 3.3: Hodnoty koeficientů pro $n = 3$ pro volbu $\hat{\gamma}_0 = \frac{1}{2}$, $\hat{\gamma}_0 = \frac{1}{2}$ pro proměnný a konstantní časový krok

3.2 Odvození koeficientů pro $n = 1$

Odvození koeficientů pro $n = 1$ je triviální. Nejdříve odvodíme metodu BDF2a. Vyjdeme z rovnice:

$$y(t_{k-1}) = y(t_k) - \tau_k y'(t_k) + \frac{1}{2}\tau_k^2 y''(t_k) + O(\tau_k^3) \tag{3.26}$$

Můžeme hned definovat numerické schéma

$$y_k - y_{k-1} = \tau_k F_k. \tag{3.27}$$

Za předpokladu $y_{k-1} = y(t_{k-1})$ a $F_k = F\Big(y(t_k)\Big)$ získáme odečtením předchozích rovnic odhad lokální chyby diskretizace

$$e_k \equiv y(t_k) - y_k = -\frac{1}{2}\tau_k^2 y''(t_k) + O(\tau_k^3). \tag{3.28}$$

Dále odvodíme metodu BDF2b. Tentokrát vyjdeme z rovnice

$$y(t_k) = y(t_{k-1}) + \tau_k y'(t_{k-1}) + \frac{1}{2}\tau_k^2 y''(t_{k-1}) + O(\tau_{k-1}^3). \tag{3.29}$$

Metoda odvozená z této rovnice by byla explicitní, metodu BDF2b proto odvodíme z rovnice získané kombinací rovnic (3.26) a (3.29). Kdybychom použili kombinaci $\frac{1}{2}(3.26) + \frac{1}{2}(3.29)$, získali bychom sice metodu druhého řádu přesnosti, neboť by vypadl člen s druhou derivací (předpokládáme totiž $y''(t_k) = y''(t_{k-1})$), ale právě ten potřebujeme k pozdějšímu odhadu lokální chyby diskretizace. Zvolíme proto kombinaci $\frac{2}{3}(3.26) + \frac{1}{3}(3.29)$:

$$y(t_k) = y(t_{k-1}) + \frac{2}{3}\tau_k y'(t_k) + \frac{1}{3}\tau_k y'(t_{k-1}) - \frac{1}{6}\tau_k^2 y''(t_k) + O(\tau_k^3). \tag{3.30}$$

Získáme numerické schéma

$$\hat{y}_k - \hat{y}_{k-1} = \frac{2}{3}\tau_k F_k + \frac{1}{3}\tau_k F_{k-1}. \tag{3.31}$$

Opět za předpokladu $y_{k-1} = y(t_{k-1})$, $F_k = F\Big(y(t_k)\Big)$ a $F_{k-1} = F\Big(y(t_{k-1})\Big)$ získáme odečtením předchozích rovnic odhad lokální chyby diskretizace

$$\hat{e}_k \equiv y(t_k) - \hat{y}_k = -\frac{1}{6}\tau_k^2 y''(t_k) + O(\tau_k^3). \tag{3.32}$$

Odečtením (3.32)-(3.28) získáme

$$e_k - \hat{e}_k = -\frac{1}{3}\tau_k^2 y''(t_k) + O(\tau_k^3), \tag{3.33}$$

z čehož plyne

$$y''(t_k) \approx -\frac{3}{\tau_k^2}(e_k - \hat{e}_k) = \frac{3}{\tau_k^2}(y_k - \hat{y}_k) \tag{3.34}$$

a tedy

$$\begin{aligned} e_k &\approx -\frac{3}{2}(y_k - \hat{y}_k) \\ \hat{e}_k &\approx -\frac{1}{2}(y_k - \hat{y}_k). \end{aligned} \tag{3.35}$$

Nakonec odvodíme ještě finální řešení. Odečtěme (3.28) − 3(3.32) :

$$y(t_k) - y_k - 3\Big(y(t_k) - \hat{y}_k\Big) = O(\tau_k^3), \tag{3.36}$$

z čehož plyne vzorec pro finální řešení 2. řádu :

$$\breve{y}_k = -\frac{1}{2}y_k + \frac{3}{2}\hat{y}_k. \tag{3.37}$$

3.3 Odvození koeficientů pro $n = 2$

V této části odvodíme podle obecného postupu popsaného výše koeficienty pro případ $n = 2$. Pro odvození metody BDF2a vyjdeme z rovnic

$$y(t_{k-1}) = y(t_k) - \tau_k y'(t_k) + \frac{1}{2}\tau_k^2 y''(t_k) - \frac{1}{6}\tau_k^3 y'''(t_k) + o(\tau_k^3) \tag{3.38}$$

$$\begin{aligned} y(t_{k-2}) = {} & y(t_k) - (\tau_k + \tau_{k-1})y'(t_k) + \frac{1}{2}(\tau_k + \tau_{k-1})^2 y''(t_k) - \\ & -\frac{1}{6}(\tau_k + \tau_{k-1})^3 y'''(t_k) + o\Big((\tau_k + \tau_{k-1})^3\Big) \end{aligned} \tag{3.39}$$

odečtením $(3.38)(\tau_k + \tau_{k-1})^2 - (3.39)\tau_k^2$ získáme

$$\begin{aligned} & y(t_k)\Big(\tau_k^2 - (\tau_k + \tau_{k-1})^2\Big) + (\tau_k + \tau_{k-1})^2 y(t_{k-1}) - \tau_k^2 y(t_{k-2}) = \\ & -y'(t_k)\Big(\tau_k(\tau_k + \tau_{k-1})^2 - (\tau_k + \tau_{k-1})\tau_k^2\Big) - \frac{1}{6}y'''(t_k)\Big(\tau_k^3(\tau_k + \tau_{k-1})^2 - \\ & -\tau_k^2(\tau_k + \tau_{k-1})^3\Big) + o(\tau_k^5), \end{aligned} \tag{3.40}$$

což lze upravit na

$$\begin{aligned} & \frac{2\theta_k + 1}{\theta_k + 1}y(t_k) - (1 + \theta_k)y(t_{k-1}) + \frac{\theta_k^2}{1 + \theta_k}y(t_{k-2}) = \\ & = \tau_k F(y(t_k)) - \frac{1}{6}\tau_k^3 y'''(t_k)\frac{1 + \theta_k}{\theta_k} + o(\tau_k^3). \end{aligned} \tag{3.41}$$

Můžeme tedy definovat numerické schéma BDF2a

$$\frac{2\theta_k+1}{\theta_k+1}y_k-(1+\theta_k)y_{k-1}+\frac{\theta_k^2}{1+\theta_k}y_{k-2}=\tau_k F(y_k), \tag{3.42}$$

kde $\theta_k=\tau_k/\tau_{k-1}$.

Nyní odvodíme odhad pro lokální chybu diskretizace. Předpokládejme, že platí $F(y_k)=F(y(t_k))$. Nechť dále $y_{k-1}=y(t_{k-1})$, $y_{k-2}=y(t_{k-2})$. Odečtením (3.42) od (3.41) získáme

$$\frac{2\theta_k+1}{\theta_k+1}\Big(y(t_k)-y_k\Big)=-\frac{1}{6}\tau_k^3y'''(t_k)\frac{1+\theta_k}{\theta_k} \tag{3.43}$$

a tedy

$$e_k=-\frac{1}{6}\tau_k^3y'''(t_k)\frac{(1+\theta_k)^2}{\theta_k(1+2\theta_k)} \tag{3.44}$$

Dále odvodíme metodu BDF2b. Nejdříve potřebujeme získat explicitní metodu, vyjdeme tedy z Taylorova rozvoje funkce y v bodě t_{k-1} :

$$\begin{aligned} y(t_k)&=y(t_{k-1})+\tau_k y'(t_{k-1})+\frac{1}{2}\tau_k^2y''(t_{k-1})+ \\ &+\frac{1}{6}\tau_k^3y'''(t_{k-1})+o(\tau_{k-1}^3) \end{aligned} \tag{3.45}$$

$$\begin{aligned} y(t_{k-2})&=y(t_{k-1})-\tau_{k-1}y'(t_{k-1})+\frac{1}{2}\tau_{k-1}^2y''(t_{k-1})- \\ &-\frac{1}{6}\tau_{k-1}^3y'''(t_{k-1})+o(\tau_{k-1}^3). \end{aligned} \tag{3.46}$$

Odečtením $(3.45)\tau_{k-1}^2-(3.46)\tau_k^2$ získáme

$$\begin{aligned} &\tau_{k-1}^2y(t_k)+(\tau_k^2-\tau_{k-1}^2)y(t_{k-1})-\tau_k^2y(t_{k-2})= \\ &=y'(t_{k-1})(\tau_k\tau_{k-1}^2+\tau_k^2\tau_{k-1})+\frac{1}{6}y'''(t_{k-1})(\tau_k^3\tau_{k-1}^2+\tau_{k-1}\tau_k^2)+o(\tau_k^5). \end{aligned} \tag{3.47}$$

Po úpravě získáme

$$\begin{aligned} \frac{1}{\theta_k+1}y(t_k)+(\theta_k-1)y(t_{k-1})-\frac{\theta_k^2}{\theta_k+1}y(t_{k-2})= \\ =\tau_k y'(t_{k-1})+\frac{1}{6}y'''(t_{k-1})\tau_k^3\frac{1}{\theta_k}+o(\tau_k^3) \end{aligned} \tag{3.48}$$

Metoda odvozená z předchozí rovnosti by byla explicitní a patrně by tedy nebyla nepodmíněně stabilní. Proto metodu odvoníme z rovnosti $\frac{1}{2}(3.41)+\frac{1}{2}(3.48)$. Předpokládáme, že $y'''(t_k)=y'''(t_{k-1})$.

$$y(t_k)-y(t_{k-1})=\frac{\tau_k}{2}\Big(y'(t_k)+y'(t_{k-1})\Big)-\frac{1}{12}y'''(t_{k-1})\tau_k^3+o(\tau_k^3) \tag{3.49}$$

Máme tedy metodu BDF2b

$$\hat{y}_k - \hat{y}_{k-1} = \frac{1}{2}\tau_k\Big(F(\hat{y}_k) + F(\hat{y}_{k-1})\Big) \tag{3.50}$$

a odhad lokální chyby diskretizace

$$\hat{e}_k = \frac{1}{12}y'''(t_{k-1})\tau_k^3. \tag{3.51}$$

Nyní odvodíme odhad lokální chyby diskretizace závisející pouze na přibližných řešeních získaných pomocí metod BDF2a a BDF2b. Opět předpokládáme $y'''(t_k) = y'''(t_{k-1})$. Odečtěme (3.44) od (3.51) :

$$\begin{aligned}\hat{e}_k - e_k = y(t_k) - \hat{y}_k - \Big(y(t_k) - y_k\Big) = y_k - \hat{y}_k = \\ = \frac{1}{12}\tau_k^3 y'''(t_k)\left(\frac{2(1+\theta_k)^2}{\theta_k(1+2\theta_k)} - 1\right) = \frac{1}{12}\tau_k^3 y'''(t_k)\frac{3\theta_k+2}{\theta(1+2\theta_k)}\end{aligned} \tag{3.52}$$

a tedy

$$y'''(t_k) = y'''(t_{k-1}) = \frac{12(y_k - \hat{y}_k)}{\tau_k^3}\frac{\theta_k(1+2\theta_k)}{3\theta_k+2}. \tag{3.53}$$

Dosazením do (3.44) získáme

$$e_k = -\frac{2(1+\theta_k)^2}{3\theta_k+2}(y_k - \hat{y}_k) \tag{3.54}$$

a podobně dosazením do (3.51) získáme

$$\hat{e}_k = -\frac{\theta_k(1+2\theta_k)}{3\theta_k+2}(y_k - \hat{y}_k). \tag{3.55}$$

Zbývá ještě najít finální řešení. Odečtěme $(3.51)\frac{2(1+\theta_k)^2}{\theta_k(1+2\theta_k)} - (3.44)$

$$y(t_k)\left(\frac{2(1+\theta_k)^2}{\theta_k(1+2\theta_k)} - 1\right) = -y_k + \frac{2(1+\theta_k)^2}{\theta_k(1+2\theta_k)}\hat{y}_k + 0y'''(t_k). \tag{3.56}$$

Finální řešení 3. řádu tedy definujeme jako

$$\breve{y}_k = -\frac{\theta_k(1+2\theta_k)}{3\theta_k+2}y_k + \frac{2(1+\theta_k)^2}{3\theta_k+2}\hat{y}_k. \tag{3.57}$$

Je zřejmé, že jsme získali koeficienty z tabulky 3.2.

Odvození koeficientů pro $n = 3$ se provede analogicky, je však poněkud pracnější. Hodnoty těchto koeficientů byly odvozeny pomocí symblických výpočtů v programu MAPLE a jsou uvedeny v tabulce 3.3.

3.4 Stabilita

Nyní ověříme, že naše metody jsou nepodmíněně stabilní. Uvažujme obecnou metodu typu BDF2a nebo BDF2b

$$\sum_{i=0}^{n} \alpha_i y_{n-i} = \gamma_0 \tau_k F_k + \gamma_1 \tau_k F_{k-1}. \tag{3.58}$$

K levé straně rovnice přičteme a odečteme člen $\sum_{i=0}^{n} \alpha_i y(t_{n-i})$. Dále od rovnice odečteme identitu $\gamma_0\tau_k y'(t_k)+\gamma_1\tau_k y'(t_{k-1}) = \gamma_0\tau_k F\Big(t_k, y(t_k)\Big)+\gamma_1\tau_k F\Big(t_k, y(t_{k-1})\Big)$. Rovnost zřejmě zůstane zachována. Získáme

$$\sum_{i=0}^{n} \alpha_i \Big(y_{n-i} - y(t_{n-i})\Big) + \sum_{i=0}^{n} \alpha_i y(t_{n-i}) - \gamma_0\tau_k y'(t_k) - \gamma_1\tau_k y'(t_{k-1}) =$$
$$\gamma_0\tau_k \bigg(F_k - F\Big(t_k, y(t_k)\Big)\bigg) + \gamma_1\tau_k \bigg(F_{k-1} - F\Big(t_k, y(t_{k-1})\Big)\bigg). \tag{3.59}$$

Pro jednoduchost budeme vyšetřovat pouze lineární diferenciální rovnici, kdy pravá strana $F\Big(t, y(t)\Big) = -\lambda y(t)$, $\lambda > 0$. Je tedy $F_k - F\Big(t_k, y(t_k)\Big) = F\Big(t_k, y_k\Big) - F\Big(t_k, y(t_k)\Big) = -\lambda\Big(y_k - y(t_k)\Big)$. V nelineárním případě bychom se stejným výsledkem využili Lagrangeovu větu. Bylo by pak $F\Big(t_k, y_k\Big) - F\Big(t_k, y(t_k)\Big) = F'(\xi)\Big(y_k - y(t_k)\Big)$. Označíme-li $g_k = y_k - y(t_k)$ globální chybu, můžeme psát

$$\sum_{i=0}^{n} \alpha_i g_{n-i} + \sum_{i=0}^{n} \alpha_i y(t_{n-i}) - \gamma_0\tau_k y'(t_k) - \gamma_1\tau_k y'(t_{k-1}) = \tag{3.60}$$
$$= -\gamma_0\tau_k\lambda g_k - \gamma_1\tau_k\lambda g_{k-1}.$$

Výraz $\sum_{i=0}^{n} \alpha_i y(t_{n-i}) - \gamma_0\tau_k y'(t_k) - \gamma_1\tau_k y'(t_{k-1})$ můžeme odhadnout lokální chybou diskretizace $e_k = y(t_k) - y_k$. Lokální chyba diskretizace je totiž chyba, které se dopustíme v jednom časovém kroku metody za předpokladu, že předchozí spočtené hodnoty y_i, $i = 0 \ldots n-1$ jsou rovny přesným hodnotám $y(t_i)$, $i = 0 \ldots n-1$. Stejně jako při odvození odhadu lokální chyby diskretizace předpokládáme $y'(t_k) = F\Big(t_k, y(t_k)\Big) \approx F\Big(t_k, y_k\Big) \equiv F_k$ a podobně také $y'(t_{k-1}) = F\Big(t_{k-1}, y(t_{k-1})\Big) \approx F\Big(t_{k-1}, y_{k-1}\Big) \equiv F_{k-1}$. Za těchto předpokladů můžeme psát

$$\sum_{i=0}^{n} \alpha_i y(t_{n-i}) - \gamma_0\tau_k y'(t_k) - \gamma_1\tau_k y'(t_{k-1}) = \sum_{i=1}^{n} \alpha_i y(t_{n-i}) + \alpha_0 y(t_n) -$$
$$-\gamma_0\tau_k y'(t_k) - \gamma_1\tau_k y'(t_{k-1}) = \sum_{i=1}^{n} \alpha_i y_{n-i} + \alpha_0(y_n + e_n) - \gamma_0\tau_k F_k -$$
$$-\gamma_1\tau_k F_{k-1} = \sum_{i=0}^{n} \alpha_i y_{n-i} - \gamma_0\tau_k F_k - \gamma_1\tau_k F_{k-1} + \alpha_0 e_n = \alpha_0 e_n. \tag{3.61}$$

Rovnici (3.60) můžeme s využitím předchozích úprav zapsat jako

$$\sum_{i=0}^{n} \alpha_i g_{n-i} + \alpha_0 e_n = -\gamma_0 \tau_k \lambda g_k - \gamma_1 \tau_k \lambda g_{k-1}. \tag{3.62}$$

To je diferenční rovnice pro posloupnost globálních chyb g_k. Metodu považujeme za stabilní, pokud globální chyba v čase neroste. Nejdříve budeme vyšetřovat příslušnou homogenní diferenční rovnici

$$(\alpha_0 + \gamma_0 \tau_k \lambda) g_k + (\alpha_0 + \gamma_1 \tau_k \lambda) g_{k-1} + \sum_{i=2}^{n} \alpha_i g_{n-i} = 0. \tag{3.63}$$

Charakteristickým polynomem této rovnice nazveme polynom

$$p(\xi) = (\alpha_0 + \gamma_0 \tau_k \lambda)\xi^n + (\alpha_0 + \gamma_1 \tau_k \lambda)\xi^{n-1} + \sum_{i=2}^{n} \alpha_i \xi^{n-i}. \tag{3.64}$$

Nechť má tento polynom kořeny $\xi_1 \dots \xi_r$, kde ξ_i je kořen násobnosti $\nu_i > 0$. Homogenní diferenční rovnice má potom fundamentální systém

$$\left\{ \xi_1^n, n\xi_1^n, \dots, n^{\nu_1}\xi_1^n, \quad \dots \quad , \xi_r^n, n\xi_r^n, \dots, n^{\nu_r}\xi_r^n \right\}. \tag{3.65}$$

Řešení homogenní rovnice jsou všechny lineární kombinace funkcí z tohoto fundamentálního systému. Řešení nehomogenní rovnice (3.62) získáme jako součet partikulárního řešení této rovnice a libovolného řešení homogenní rovnice. Nehomogenní rovnice se však od té homogenní liší pouze o člen $\alpha_0 e_n$, který je malý. Podle (3.8) je totiž $e_n = O(\tau_k^{n+1})$. Tento člen nemůže chování posloupnosti globálních chyb významně ovlivnit. Metoda tedy bude stabilní, pokud žádné řešení homogenní rovnice nebude růst v čase. Z tvaru fundametálního systému je patrné, že to nastane právě tehdy, když platí $|\xi_i| \leq 1$, $i = 1 \dots r$ a pro násobné kořeny ($\nu_i > 1$) dokonce $|\xi_i| < 1$.

V následujících částech tento postup aplikujeme na metody, které budeme používat v numerických experimentech (tedy metody jednokrokové, dvoukrokové a tříkrokové) a ukážeme, že jsou nepodmíněně stabilní, tedy že jsou stabilní pro jakoukoli volbu časového kroku τ_k. Postup můžeme použít pouze pro variantu metod s konstantním časovým krokem. Jedině tak totiž získáme diferenční rovnici s konstantními koeficienty, kterou jsme schopni řešit. Dá se však předpokládat, že nepodmíněná stabilita metody s konstantním časovým krokem znamená i stabilitu metody s proměnným krokem. Krok metody se totiž nebude měnit nijak výrazně a stabilita by měla zůstat zachována.

3.4.1 Stabilita pro $n = 1$

Pro případ $n = 1$ je vyšetřování stability triviální. Začneme s metodou BDF2a, tedy s rovnicí

$$y_k - y_{k-1} = \tau_k F_k. \tag{3.66}$$

Jak bylo vysvětleno v předchozí části, zkoumáme stabilitu pro lineární rovnici, tedy rovnici s pravou stranou $F\Big(t, y(t)\Big) = -\lambda y(t)$, $\lambda > 0$. Na rovnici (3.66) provedeme úpravy popsané v předchozí části

$$y_k - y_{k-1} \pm \Big(y(t_k) - y(t_{k-1})\Big) - \tau_k y'(t_k) = -\tau_k \lambda y_k + \tau_k \lambda y(t_k) \tag{3.67}$$

a tedy

$$g_k - g_{k-1} + y(t_k) - y(t_{k-1}) - \tau_k y'(t_k) = -\tau_k \lambda g_k, \tag{3.68}$$

z čehož získáme diferenční rovnici pro globální chybu $g_k \equiv y_k - y(t_k)$

$$g_k - g_{k-1} + e_k = -\tau_k \lambda g_k. \tag{3.69}$$

Její charakteristický polynom má tvar

$$p(\xi) = (1 + \tau_k \lambda)\xi - 1, \tag{3.70}$$

má tedy jediný jednoduchý kořen $\xi_1 = \frac{1}{1+\tau_k \lambda}$. Jelikož je $\lambda > 0$, je zřejmě $|\xi_1| < 1$ pro každé $\tau_k > 0$ a metoda BDF2a je tedy nepodmíněně stabilní.

Stejný postup zopakujeme i pro metodu BDF2b. V tomto případě však vyjdeme od rovnice

$$y_k - y_{k-1} = \frac{2}{3}\tau_k F_k + \frac{1}{3}\tau_k F_{k-1}, \tag{3.71}$$

na kterou opět použijeme úpravy popsané v obecné části

$$\begin{aligned} y_k - y_{k-1} \pm \Big(y(t_k) - y(t_{k-1})\Big) - \frac{2}{3}\tau_k y'(t_k) - \frac{1}{3}\tau_k y'(t_{k-1}) = \\ = -\frac{2}{3}\tau_k \lambda y_k - \frac{1}{3}\tau_k \lambda y_{k-1} + \frac{2}{3}\tau_k \lambda y(t_k) + \frac{1}{3}\tau_k \lambda y(t_{k-1}), \end{aligned} \tag{3.72}$$

z čehož získáme

$$\begin{aligned} g_k - g_{k-1} + y(t_k) - y(t_{k-1}) - \frac{2}{3}\tau_k y'(t_k) - \frac{1}{3}\tau_k y'(t_{k-1}) = \\ -\frac{2}{3}\tau_k \lambda g_k - \frac{1}{3}\tau_k \lambda g_{k-1} \end{aligned} \tag{3.73}$$

a máme opět diferenční rovnici pro globální chybu

$$g_k - g_{k-1} + e_k = -\frac{2}{3}\tau_k \lambda g_k - \frac{1}{3}\tau_k \lambda g_{k-1}, \tag{3.74}$$

jejíž charakteristický polynom má tvar

$$p(\xi) = (\frac{2}{3}\tau_k\lambda + 1)\xi + \frac{1}{3}\tau_k\lambda - 1 \tag{3.75}$$

a má jediný jednonásobný kořen

$$\xi_1 = \frac{1 - \frac{1}{3}\tau_k\lambda}{1 + \frac{2}{3}\tau_k\lambda}. \tag{3.76}$$

Protože $\lambda > 0$, platí pro libovolné $\tau_k > 0$ nerovnost $|1 - \frac{1}{3}\tau_k\lambda| < |1 + \frac{2}{3}\tau_k\lambda|$ a je tedy $|\xi_1| < 1$. Metoda BDF2b je tedy opět nepodmíněně stabilní.

3.4.2 Stabilita pro $n = 2$

Podobným způsobem vyšetříme stabilitu pro případ $n = 2$. Jak již bylo uvedeno, budeme zkoumat pouze metodu s konstantním časovým krokem. Metoda BDF2a je tedy tvaru (uvažujeme opět pouze lineární rovnici)

$$\frac{3}{2}y_k - 2y_{k-1} + \frac{1}{2}y_{k-2} = -\tau_k\lambda y_k. \tag{3.77}$$

Stejně jako v minulé části přidáme a ubereme stejné členy, aby se rovnost nezměnila

$$\begin{aligned}\frac{3}{2}y_k - 2y_{k-1} + \frac{1}{2}y_{k-2} \pm \left(\frac{3}{2}y(t_k) - 2y(t_{k-1}) + \frac{1}{2}y(t_{k-2})\right) - \\ -\tau_k y'(t_k) = -\tau_k\lambda y_k + \tau_k\lambda y(t_k),\end{aligned} \tag{3.78}$$

z čehož získáme

$$\frac{3}{2}g_k - 2g_{k-1} + \frac{1}{2}g_{k-2} + \frac{3}{2}y(t_k) - 2y(t_{k-1}) + \frac{1}{2}y(t_{k-2}) - \tau_k y'(t_k) = -\tau_k\lambda g_k \tag{3.79}$$

a opět máme diferenční rovnici pro globální chybu $g_k \equiv y_k - y(t_k)$

$$\frac{3}{2}g_k - 2g_{k-1} + \frac{1}{2}g_{k-2} + e_k = -\tau_k\lambda g_k. \tag{3.80}$$

Charakteristický polynom je tvaru

$$p(\xi) = \left(\frac{3}{2} + \tau_k\lambda\right)\xi^2 - 2\xi + \frac{1}{2} \tag{3.81}$$

s kořeny

$$\xi_1 = \frac{2 + \sqrt{1 - 2\tau_k\lambda}}{3 + 2\tau_k\lambda}, \qquad \xi_2 = \frac{2 - \sqrt{1 - 2\tau_k\lambda}}{3 + 2\tau_k\lambda}. \tag{3.82}$$

Chceme ukázat, že oba kořeny mají absolutní hodnotu menší než 1. Musíme rozlišit několik případů. Nejdříve předpokládejme, že $\tau_k\lambda > \frac{1}{2}$, tedy že $1 - 2\tau_k\lambda < 0$. Potom má charakteristický polynom $p(\xi)$ dva komplexně sdružené kořeny

$$\xi_1 = \frac{2 + i\sqrt{2\tau_k\lambda - 1}}{3 + 2\tau_k\lambda}, \qquad \xi_2 = \frac{2 - i\sqrt{2\tau_k\lambda - 1}}{3 + 2\tau_k\lambda} \tag{3.83}$$

s absolutní hodnotou

$$|\xi_1| = |\xi_2| = \sqrt{\frac{2^2 + 2\tau_k\lambda - 1}{(3 + 2\tau_k\lambda)^2}}. \tag{3.84}$$

Protože $\tau_k\lambda > 0$, platí zřejmě $3+\tau_k\lambda < 9+12\tau_k\lambda+4(\tau_k\lambda)^2$ a tedy $|\xi_1| = |\xi_2| < 1$.

Nyní naopak předpokládejme, že $\tau_k\lambda \leq \frac{1}{2}$, tedy že $1 - 2\tau_k\lambda \geq 0$. Potom má charakteristický polynom reálné kořeny (3.82). Nejdříve ukážme, že $|\xi_1| < 1$. Na tuto nerovnici budeme provádět ekvivalentní úpravy

$$\begin{aligned}
\left|\frac{2 + \sqrt{1 - 2\tau_k\lambda}}{3 + 2\tau_k\lambda}\right| &< 1 \\
\left|2 + \sqrt{1 - 2\tau_k\lambda}\right| &< |3 + 2\tau_k\lambda| \\
\sqrt{1 - 2\tau_k\lambda} &< 1 + 2\tau_k\lambda \\
1 - 2\tau_k\lambda &< 1 + 4\tau_k\lambda + 4(\tau_k\lambda)^2 \\
0 &< 6\tau_k\lambda + 4(\tau_k\lambda)^2.
\end{aligned}$$

To však zřejmě platí, protože $\tau_k\lambda > 0$

Nyní ukažme, že $|\xi_2| < 1$, tedy že

$$\left|2 - \sqrt{1 - 2\tau_k\lambda}\right| < |3 + 2\tau_k\lambda|. \tag{3.85}$$

Rozlišíme dva případy. Nejdříve nechť je $\tau_k\lambda \geq -\frac{3}{2}$, tedy $2 - \sqrt{1 - 2\tau_k\lambda} \geq 0$. Nerovnici (3.85) můžeme potom ekvivalentně zapsat jako

$$\begin{aligned}
2 - \sqrt{1 - 2\tau_k\lambda} &< 3 + 2\tau_k\lambda \\
-\sqrt{1 - 2\tau_k\lambda} &< 1 + 2\tau_k\lambda,
\end{aligned}$$

což zřejmě platí, neboť na levé straně je číslo záporné a na pravé kladné. Nyní nechť je naopak $\tau_k\lambda < -\frac{3}{2}$, tedy $2 - \sqrt{1 - 2\tau_k\lambda} < 0$. Nerovnici (3.85) můžeme tentokrát ekvivalentně zapsat jako

$$\begin{aligned}
-2 + \sqrt{1 - 2\tau_k\lambda} &< 3 + 2\tau_k\lambda \\
\sqrt{1 - 2\tau_k\lambda} &< 5 + 2\tau_k\lambda \\
1 - 2\tau_k\lambda &< 25 + 20\tau_k\lambda + 4(\tau_k\lambda)^2 \\
0 &< 24 + 22\tau_k\lambda + 4(\tau_k\lambda)^2,
\end{aligned}$$

což zjevně platí. Ukázali jsme tedy, že pro metodu BDF2a je vždy $|\xi_1| < 1$ i $|\xi_2| < 1$ pro libovolné $\tau_k > 0$. Metoda je tedy nepodmíněně stabilní.

Nyní vyšetřujme metodu BDF2b, tedy rovnici

$$y_k - y_{k-1} = \frac{1}{2}\tau_k\lambda y_k + \frac{1}{2}\tau_k\lambda y_{k-1}. \tag{3.86}$$

Tu opět upravíme na tvar

$$\begin{aligned} y_k - y_{k-1} \pm \big(y(t_k) - y(t_{k-1})\big) - \frac{1}{2}\tau_k y'(t_k) - \frac{1}{2}\tau_k y'(t_{k-1}) = \\ = -\frac{1}{2}\tau_k\lambda y_k - \frac{1}{2}\tau_k\lambda y_{k-1} + \frac{1}{2}\tau_k\lambda y(t_k) + \frac{1}{2}\tau_k\lambda y(t_{k-1}), \end{aligned} \tag{3.87}$$

z čehož získáme

$$\begin{aligned} g_k - g_{k-1} + y(t_k) - y(t_{k-1}) - \frac{1}{2}\tau_k y'(t_k) - \frac{1}{2}\tau_k y'(t_{k-1}) = \\ = -\frac{1}{2}\tau_k\lambda g_k - \frac{1}{2}\tau_k\lambda g_{k-1}. \end{aligned} \tag{3.88}$$

Máme tedy opět diferenční rovnici pro globální chybu

$$g_k - g_{k-1} + e_k = -\frac{1}{2}\tau_k\lambda g_k - \frac{1}{2}\tau_k\lambda g_{k-1} \tag{3.89}$$

s charakteristickým polynomem

$$p(\xi) = \Big(\frac{1}{2}\tau_k\lambda + 1\Big)\xi^2 + \Big(\frac{1}{2}\tau_k\lambda - 1\Big)\xi. \tag{3.90}$$

Ten má reálné kořeny

$$\xi_1 = 0, \qquad \xi_2 = \frac{1 - \frac{1}{2}\tau_k\lambda}{1 + \frac{1}{2}\tau_k\lambda}. \tag{3.91}$$

Zřejmě je $|\xi_1| < 1$. Zbývá ukázat, že $|\xi_2| < 1$. To můžeme ekvivalentně upravit jako

$$\begin{aligned} \left|\frac{2 - \tau_k\lambda}{2 + \tau_k\lambda}\right| &< 1 \\ |2 - \tau_k\lambda| &< |2 + \tau_k\lambda| \\ 4 - 4\tau_k\lambda + (\tau_k\lambda)^2 &< 4 + 4\tau_k\lambda + (\tau_k\lambda)^2 \\ 0 &< 8\tau_k\lambda, \end{aligned}$$

což zřejmě platí. I metoda BDF2b je tedy nepodmíněně stabilní.

3.4.3 Stabilita pro $n = 3$

Nakonec se zaměřme na případ $n = 3$. Metoda BDF2a pro konstantní časový krok a lineární rovnici má tvar

$$\frac{11}{6}y_k - 3y_{k-1} + \frac{3}{2}y_{k-2} - \frac{1}{3}y_{k-3} = -\tau_k \lambda y_k. \quad (3.92)$$

Opět přidáme a ubereme stejné členy, aby se rovnost nezměnila

$$\frac{11}{6}y_k - 3y_{k-1} + \frac{3}{2}y_{k-2} - \frac{1}{3}y_{k-3} \pm \left(\frac{11}{6}y(t_k) - 3y(t_{k-1}) + \right.$$
$$\left. + \frac{3}{2}y(t_{k-2}) - \frac{1}{3}y({}_t k - 3)\right) - \tau_k y'(t_k) = -\tau_k \lambda y_k + \tau_k$$

z čehož získáme

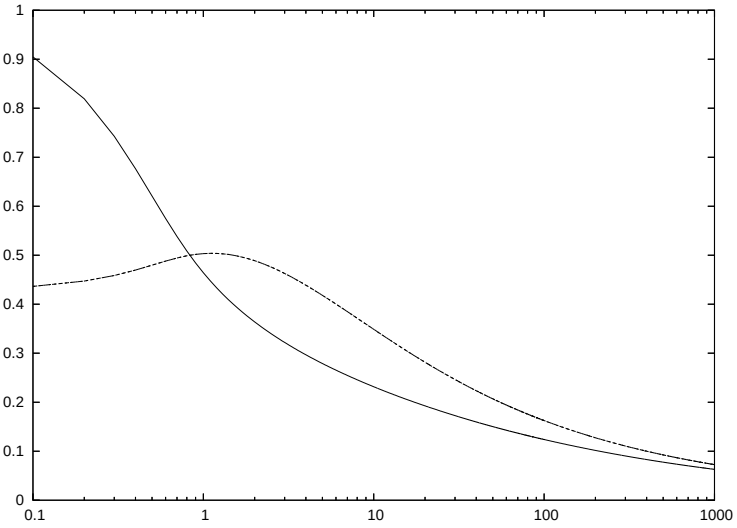
$$\frac{11}{6}g_k - 3g_{k-1} + \frac{3}{2}g_{k-2} - \frac{1}{3}g_{k-3} + \frac{11}{6}y(t_k) - 3y(t_{k-1}) \quad (3.94)$$
$$+ \frac{3}{2}y(t_{k-2}) - \frac{1}{3}y(t_{k-3}) - \tau_k y'(t_k) = -\tau_k \lambda g_k$$

a opět dostaneme diferenční rovnici pro globální chybu $g_k \equiv y_k - y(t_k)$

$$\frac{11}{6}g_k - 3g_{k-1} + \frac{3}{2}g_{k-2} - \frac{1}{3}g_{k-3} + e_k = -\tau_k \lambda g_k. \quad (3.95)$$

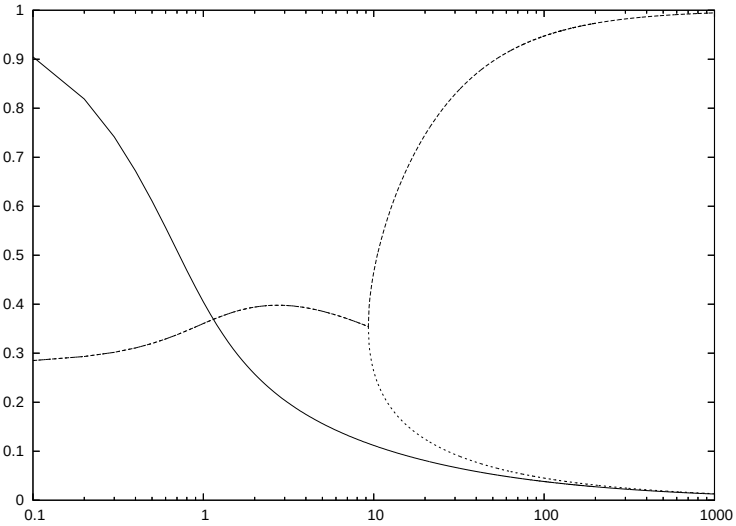
Charakteristický polynom je tvaru

$$p(\xi) = \left(\frac{11}{6} + \tau_k \lambda\right)\xi^3 - 3\xi^2 + \frac{3}{2}\xi - \frac{1}{3}. \quad (3.96)$$



Obrázek 3.1: Absolutní hodnoty kořenů charakteristického polynomu (3.96) na intervalu $[0, 1000]$. Hodnoty jsou spočteny s krokem 0.1. Dva z kořenů jsou komplexně sdružené, mají proto stejnou absolutní hodnotu.

To je polynom třetího stupně. Vyjádřit jeho kořeny v explicitním tvaru v závislosti na parametru $\lambda\tau_k$ by bylo značně komplikované. Proto se spokojíme s tím, že spočítáme hodnoty kořenů numericky pro různé hodnoty $\lambda\tau_k$ a ověříme, že jejich absolutní hodnota je menší než 1. Hodnoty absolutních hodnot kořenů charakteristického polynomu jsou zobrazeny na obrázku 3.1. Je vidět, že ve zkoumaném intervalu jsou skutečně menší než 1.



Obrázek 3.2: Absolutní hodnoty kořenů charakteristického polynomu (3.101) na intervalu $[0, 1000]$. Hodnoty jsou spočteny s krokem 0.1.

Nyní vyšetřujme metodu BDF2b, tedy rovnici

$$\frac{13}{12}y_k - \frac{5}{4}y_{k-1} + \frac{1}{4}y_{k-2} - \frac{1}{12}y_{k-3} = \frac{1}{2}\tau_k\lambda y_k + \frac{1}{2}\tau_k\lambda y_{k-1}. \tag{3.97}$$

Tu opět upravíme na tvar

$$\begin{aligned}\frac{13}{12}y_k - \frac{5}{4}y_{k-1} + \frac{1}{4}y_{k-2} - \frac{1}{12}y_{k-3} &\pm \left(\frac{13}{12}y(t_k) - \frac{5}{4}y(t_{k-1}\right)\\ &+\frac{1}{4}y(t_{k-2}) - \frac{1}{12}y(t_{k-3})\Big) - \frac{1}{2}\tau_k y'(t_k) - \frac{1}{2}\tau_k y'(t_{k-1}\\ &= -\frac{1}{2}\tau_k\lambda y_k - \frac{1}{2}\tau_k\lambda y_{k-1} + \frac{1}{2}\tau_k\lambda y(t_k) + \frac{1}{2}\tau_k\lambda y(t_{k-}\end{aligned}$$

z čehož získáme

$$\frac{13}{12}g_k - \frac{5}{4}g_{k-1} + \frac{1}{4}g_{k-2} - \frac{1}{12}g_{k-3} + \frac{13}{12}y(t_k) - \frac{5}{4}y(t_{k-1}) + \frac{1}{4}y(t_{k-2}) - \frac{1}{12}y(t_{k-3}) - \frac{1}{2}\tau_k y'(t_k) - \frac{1}{2}\tau_k y'(t_{k-1}) = -\frac{1}{2}\tau_k\lambda g_k - \frac{1}{2}\tau_k\lambda g_{k-1}. \tag{3.99}$$

Máme tedy opět diferenční rovnici pro globální chybu

$$\frac{13}{12}g_k - \frac{5}{4}g_{k-1} + \frac{1}{4}g_{k-2} - \frac{1}{12}g_{k-3} + e_k = -\frac{1}{2}\tau_k\lambda g_k - \frac{1}{2}\tau_k\lambda g_{k-1} \tag{3.100}$$

s charakteristickým polynomem

$$p(\xi) = (\frac{13}{12} + \frac{1}{2}\tau_k\lambda)\xi^3 + (-\frac{5}{4} + \frac{1}{2}\tau_k\lambda)\xi^2 + \frac{1}{4}\xi - \frac{1}{12}. \tag{3.101}$$

Opět jsme získali polynom třeího stupně, jehož kořeny bychom přímo hledali jen obtížně. Absolutní hodnoty kořenů na intervalu $[0, 1000]$ jsou zobrazeny na obrázku 3.2. Na uvedeném intervalu jsou absolutní hodnoty všech kořenů menší než 1, i když se zdá, že absolutní hodnota jednoho z kořenů k 1 konverguje.

Kapitola 4

Diskretizace v prostoru i v čase

Nyní provedeme plnou diskretizaci problému v prostoru i v čase. Vyjdeme ze semidiskrétního problému (2.23), a) – c), na který aplikujeme metodu BDF2.

4.1 Diskrétní problém

Vzhledem k tomu, že problém (2.23), a) – c) je nelineární, tak přímé použití implicitní metody vede k nutnosti řešit soustavu nelineárních rovnic, což by bylo obtížné. Při diskretizaci nelineárního členu b_h proto použijeme explicitní extrapolaci. Ta je odvozena v článku [4].

$$y_k \approx \sum_{l=1}^{n} \beta_l y_{k-l} \tag{4.1}$$

Hodnoty koeficientů $\beta_l,\ l = 1, \ldots, n$ pro $n = 1 \ldots 3$ jsou uvedeny v tabulce 4.1.

Definice Definujeme *přibližné řešení* problému (2.1)-(2.3) jako posloupnost funkcí u_h^k, $t_k \in [0, T]$, splňující podmínky

$$\begin{aligned}
&\text{a)} \quad u_h^k \in S_h, \qquad\qquad (4.2)\\
&\text{b)} \quad \frac{1}{\tau_k}\left(\sum_{l=0}^{n} \alpha_l u_h^{k-l}, v_h\right) + \gamma_0 A_h(u_h^k, v_h) + \gamma_0 b_h\left(\sum_{l=1}^{n} \beta_l u_h^{k-l}, v_h\right) + \\
&\qquad\qquad +\gamma_1 A_h(u_h^{k-1}, v_h) + \gamma_1 b_h(u_h^{k-1}, v_h) = 0 \qquad \forall\, v_h \in S_h,\ \forall\, t_k \in (0, T],\\
&\text{c)} \quad u_h^0 \text{ je } S_h \text{ aproximace } u^0,\\
&\text{d)} \quad u_h^l \in S_h,\ l = 1, \ldots, n-1 \text{ jsou získány metodami BDF2 nižších řádů,}
\end{aligned}$$

	proměnný			konstantní		
n	1	2	3	1	2	3
β_1	1	$\frac{\tau_k+\tau_{k-1}}{\tau_{k-1}}$	$\frac{(\tau_k+\tau_{k-1})(\tau_k+\tau_{k-1}+\tau_{k-2})}{\tau_{k-1}(\tau_{k-1}+\tau_{k-2})}$	1	2	3
β_2		$-\frac{\tau_k}{\tau_{k-1}}$	$-\frac{\tau_k(\tau_k+\tau_{k-1}+\tau_{k-2})}{\tau_{k-1}\tau_{k-2}}$		-1	-3
β_3			$\frac{\tau_k(\tau_k+\tau_{k-1})}{\tau_{k-2}(\tau_{k-1}+\tau_{k-2})}$			1

Tabulka 4.1: Hodnoty koeficientů $\beta_l,\ l = 1, \ldots, n$ pro proměnný a konstantní časový krok por $n = 1, 2, 3$

Funkci u_h^k nazýváme přibližným řešením na časové vrstvě t_k. Toto řešení je získané pomocí metody BDF2a. Pokud v předchozím místo koeficientů $\alpha_l,\ l = 0, \ldots, n$ použijeme koeficienty $\hat{\alpha}_l,\ l = 0, \ldots, n$ a místo γ_0, γ_1 použijeme $\hat{\gamma}_0$, $\hat{\gamma}_1$, získáme přibližné řešení $\hat{u}_h^k$ (řešení získané pomocí metody BDF2b). Finální řešení získáme jako

$$\breve{u}_h^k = \hat{\delta} u_h^k - \delta \hat{u}_h^k. \tag{4.3}$$

4.2 Časová adaptivita

Nyní můžeme zavést adaptivní volbu časového kroku. Základem je schopnost odhadu lokální chyby. K tomu použijeme rozdíl řešení získaných pomocí dvou různých metod BDF2a a BDF2b. Položme $e_h^k = u(t_h) - u_h^k$. Podle předchozího odhadujeme

$$e_h^k = \delta(u_h^k - \hat{u}_h^k).. \tag{4.4}$$

V každém kroku výpočtu jsme tedy schopni odhadnout lokální chybu. Naším cílem bude volit časové kroky tak, aby tato chyba byla stále přibližně stejná a co nejvíce se blížila předem zadané konstantě TOL. Ta reprezentuje námi předepsanou toleranci na lokální chybu. Pokud bude chyba v daném kroku výrazně větší, než je tolerance (větší než konstanta TOL2, kde TOL2>TOL), odmítneme získané řešení, zmenšíme časový krok a zopakujeme výpočet. Pokud bude chyba jen o málo větší než je tolerance (tedy mezi konstantami TOL a TOL2), řešení přijmeme a pouze zmenšíme následující časový krok. Pokud bude naopak

chyba menší než tolerance, příští časový krok prodloužíme. Označme

$$\mathrm{EST} = \|e_h^k\|_{L^2(\Omega,\mathcal{T}_h)}. \tag{4.5}$$

Naším cílem je volit časový krok τ_k tak, aby bylo EST=TOL, nebo aby alespoň tyto hodnoty byly co nejblíže. Podle odhadu (3.8) platí

$$\mathrm{EST} = O(\tau_k^{n+1}) \tag{4.6}$$

a tedy

$$\mathrm{EST} = C\tau_k^{n+1}. \tag{4.7}$$

My bychom však chtěli najít takový časový krok $\bar{\tau}_k$, aby při jeho použití ve výpočtu místo časového kroku τ_k byl odhad lokální chyby co nejblíže zadané toleranci TOL, tedy aby platilo

$$\mathrm{TOL} = C\bar{\tau}_k^{n+1}. \tag{4.8}$$

Z předchozích vztahů můžeme vyjádřit konstantu C

$$\begin{aligned} C &= \frac{\tau_k^{n+1}}{\mathrm{EST}} \\ C &= \frac{\bar{\tau}_k^{n+1}}{\mathrm{TOL}}. \end{aligned}$$

Hledané $\bar{\tau}_k$ můžeme tedy vyjádřit jako

$$\bar{\tau}_k = \tau_k \sqrt[n+1]{\frac{\mathrm{TOL}}{\mathrm{EST}}} \tag{4.9}$$

a při dalším výpočtu použijeme časový krok $\tau_{k+1} = \bar{\tau}_k$.

Úvahy provedené v této části jsou pouze přibližné, to však příliš nevadí. Výsledkem těchto úvah je pouze volba časového kroku pro další výpočet a případná chyba nijak dramaticky neovlivní výsledky. Vadit by mohla pouze volba neúměrně velkého časového kroku, pak by ovšem v dalším výpočtu byla pravděpodobně odhadnuta příliš velká chyba a tento krok by byl zamítnut a opakován s kratším časovým krokem.

4.3 Programová realizace

Metodu BDF2 a algoritmus pro adaptivní volbu časového kroku byl implementován do již existujícího programového balíku. Některé výsledky získané

pomocí nové metody a jejich srovnání s výsledky získanými pomocí původní metody je uvedeno v kapitole 5.

Při implementaci jsme provedli několik drobných změn proti zde popsaným metodám. Ve vzorci (4.2) nepočítáme výraz $b_h(u_h^{k-1}, v_h)$, místo toho je použita hodnota $b_h\left(\sum_{l=1}^{n} \beta_l u_h^{k-l-1}, v_h\right)$ z předchozí iterace. Numerické experimenty ukazují, že toto přiblížení je dostatečné a tudíž ho používáme z důvodu úspory výpočetního času.

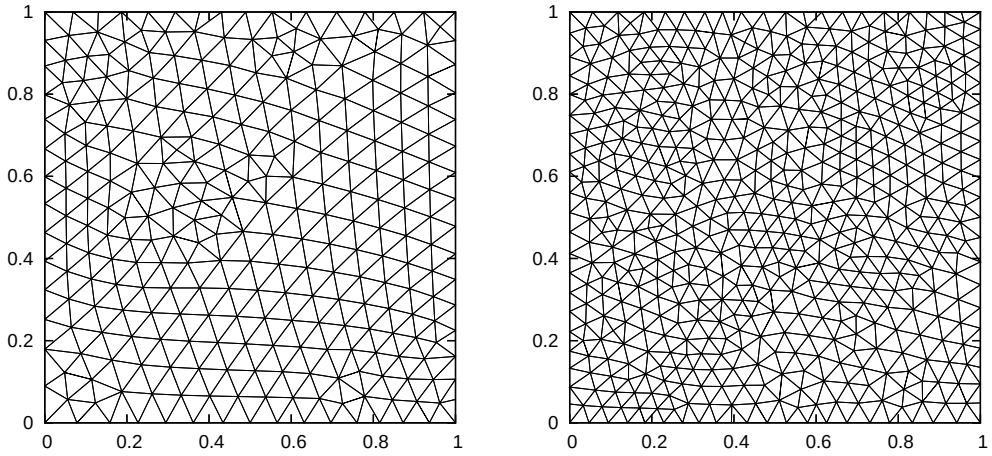
Při výpočtu s časově adaptivní volbou časového kroku se někdy stává, že algoritmus volí střídavě velmi krátký a velmi dlouhý časový krok (po výpočtu s dlouhým krokem vyjde velký odhad lokální chyby, proto se krok výrazně zkrátí, potom ovšem získáme malý odhad lokální chyby a krok se opět výrazně prodlouži, atd.). Proto se zdá vhodné omezit velikost změny časového kroku. Nový časový krok $\tau_{k+1} = \bar{\tau}_k$ se i nadále určuje podle vztahu (4.9), ale tak, aby $\tau_{k+1} < Z_1\tau_k$ a současně $\tau_{k+1} > Z_2\tau_k$. Pochopitelně není snadné určit konstanty Z_1 a Z_2, nám se osvědčila volba $Z_1 = 3$ a $Z_2 = \frac{1}{3}$.

Podobné je to i s volbou konstanty TOL2, při výpočtech jsme používali hodnoty mezi $\mathrm{TOL2} = 2 \cdot \mathrm{TOL}$ a $\mathrm{TOL2} = 5 \cdot \mathrm{TOL}$.

V kapitole 5 je při porovnávání efektivity adaptivní a neadaptivní metody uveden počet iterací potřebných k dosažení výsledku, nikoli celkový čas výpočtu. To proto, že délka iterace u obou metod je téměř stejná. Při použití adaptivní metody je sice potřeba počítat dvě řešení, musí se tedy řešit dvě soustavy lineárních rovnic, to však tvoří pouze zlomek celkové délky výpočtu. Většina doby je zabrána výpočtem prvků matic a vektorů, které se použijí pro obě metody. Navíc řešení soustavy lineárních rovnic získané první metodou se použije jako počáteční přiblížení při výpočtu druhého řešení a tím se čas opět zkrátí. Výpočet druhého řešení tedy nepředstavuje téměř žádný nárůst výpočetního času.

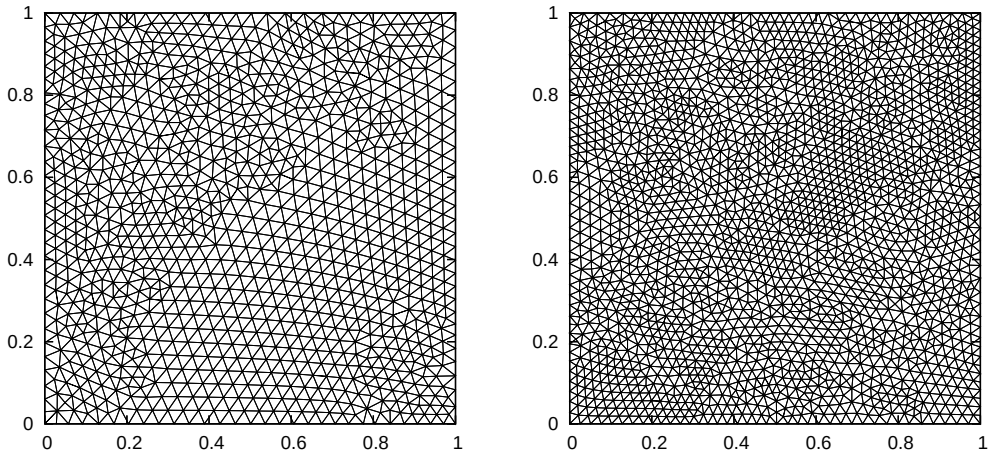
Kapitola 5

Numerické výsledky



Obrázek 5.1: Použité sítě s 504 a 1009 elementy

V této kapitole numericky ověříme vlastnosti odvozených metod. Budeme vyšetřovat jejich chování na několika různých příkladech. Vždy nejdříve pro obyčejné diferenciální rovnice a pak pro skalární rovnici (2.1), což je hlavní náplní této práce. Pro výpočty byly použity nestrukturované sítě s různým počtem elementů zobrazené na obrázcích 5.1 a 5.2. Volba sítě nemá na zde prezentované výsledky významný vliv. Všechna uvedená data byla získána s použitím sítě s 1009 elementy.



Obrázek 5.2: Použité sítě s 2043 a 4014 elementy

5.1 Experimentální řády konvergence

V této části se přesvědčíme, zda odvozené řády konvergence odpovídají skutečnosti. Podle kapitoly 3 by měla být metoda BDF2 druhého řádu pro $n = 1$, třetího řádu pro $n = 2$ a čtvrtého řádu pro $n = 3$. V této části budeme zkoumat pouze metodu s konstantním časovým krokem. Budeme volit různě dlouhé časové kroky a zjišťovat, jak velké chyby se metoda s tímto krokem dopustí.

5.1.1 Obyčejné diferenciální rovnice

Příklad 1 Nejdříve vyšetřujme experimentální řády konvergence pro obyčejnou diferenciální rovnici

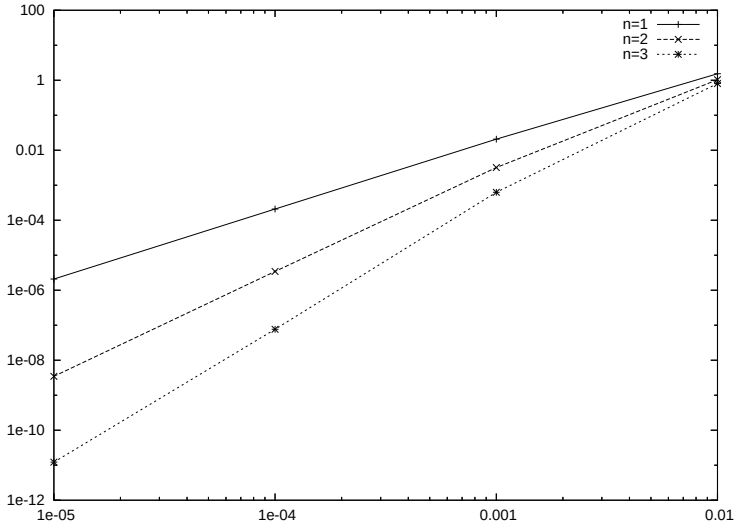
$$y' = \frac{\alpha e^{\alpha t}}{e^{\alpha} - 1} \tag{5.1}$$

s přesným řešením

$$y = \frac{e^{\alpha t} - 1}{e^{\alpha} - 1} \tag{5.2}$$

na intervalu $t \in [0, 1]$ pro volbu $\alpha = 500$.

V tabulce 5.1 jsou uvedeny experimentální řády konvergence pro obyčejnou diferenciální rovnici z příkladu 1. Podle (3.25) se metoda v každém kroku dopustí chyby $O(\tau^{n+2})$, kde τ je konstantní časový krok. Celková chyba, které se dopustí na intervalu délky T je tedy $\frac{T}{\tau} O(\tau^{n+2}) = O(\tau^{n+1})$. Z tabulky 5.1 je vidět, že experimentální řády konvergence odvozené pomocí metody nejmenších



Obrázek 5.3: Experimentální řády konvergence pro n=1, 2, 3 pro obyčejnou diferenciální rovnici (5.1)

	10^{-2}	10^{-3}	10^{-4}	10^{-5}	řád
$n=1$	1.53×10^{0}	2.07×10^{-2}	2.08×10^{-4}	2.08×10^{-6}	1.96
$n=2$	1.04×10^{0}	3.20×10^{-3}	3.45×10^{-6}	3.47×10^{-9}	2.84
$n=3$	8.06×10^{-1}	6.31×10^{-4}	7.65×10^{-8}	1.23×10^{-11}	3.64

Tabulka 5.1: Chyba metody BDF2 pro obyčejnou diferenciální rovnici (5.1) pro různé volby délky časového kroku

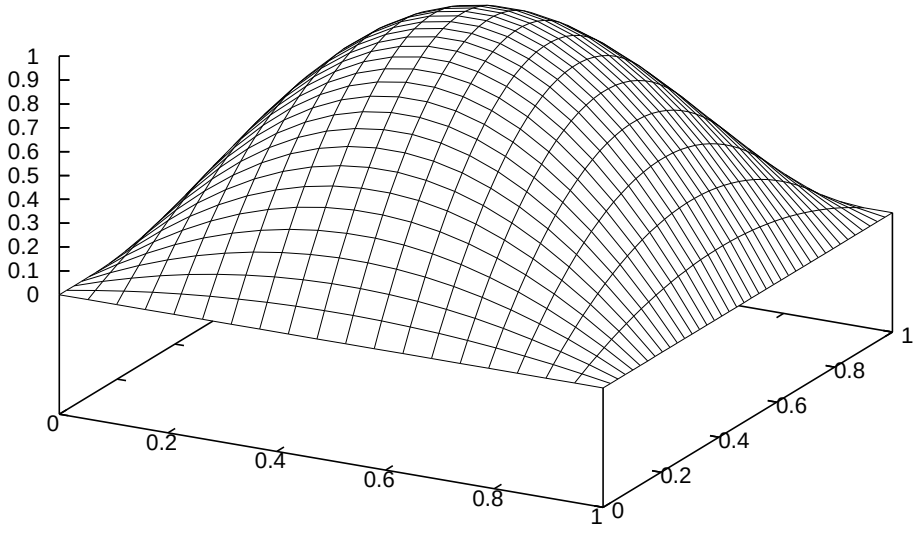
čtverců ze spočtených chyb se velmi přesně shodují s očekávanými teoretickými řády konvergence.

5.1.2 Skalární rovnice

Příklad 2 Nyní vyšetřujme experimentální řády konvergence pro parciální diferenciální rovnici (2.1) s přesným řešením

$$\bar{u} = x(1-x)y(1-y)\frac{e^{\alpha t}-1}{e^{\alpha}-1} \tag{5.3}$$

na oblasti $[0,1] \times [0,1]$ a časovém intervalu $t \in [0,1]$.



Obrázek 5.4: Řešení u z příkladu 2 v čase $t = 1$

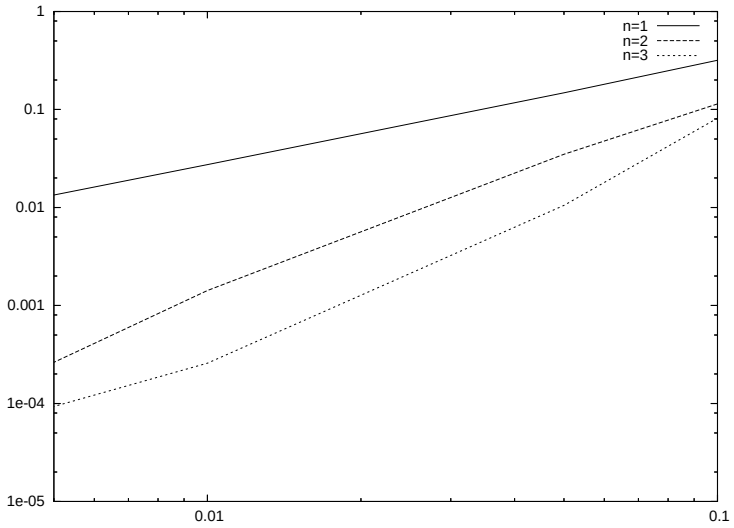
Pravou stranu g rovnice (2.1) spočítáme z přesného řešení

$$\begin{aligned}
\frac{\partial \bar{u}}{\partial t} &= x(1-x)y(1-y)\frac{\alpha e^{\alpha t}}{e^{\alpha}-1} \\
\frac{\partial \bar{u}}{\partial x} &= (1-2x)y(1-y)\frac{e^{\alpha t}-1}{e^{\alpha}-1} \\
\frac{\partial \bar{u}}{\partial y} &= x(1-x)(1-2y)\frac{e^{\alpha t}-1}{e^{\alpha}-1} \\
\frac{\partial^2 \bar{u}}{\partial x^2} &= -2y(1-y)\frac{e^{\alpha t}-1}{e^{\alpha}-1} \\
\frac{\partial^2 \bar{u}}{\partial y^2} &= -2x(1-x)\frac{e^{\alpha t}-1}{e^{\alpha}-1}.
\end{aligned}$$

Používáme $f_s(u) = \frac{u^2}{2}$, $s = 1, 2$. Položíme tedy

$$g = \frac{\partial \bar{u}}{\partial t} + \bar{u}\left(\frac{\partial \bar{u}}{\partial x} + \frac{\partial \bar{u}}{\partial y}\right) - \epsilon\left(\frac{\partial^2 \bar{u}}{\partial x^2} + \frac{\partial^2 \bar{u}}{\partial y^2}\right). \tag{5.4}$$

Řešení v čase $t = 1$ je zobrazeno na obrázku 5.4. Při numerických výpočtech byla použita hodnota $\alpha = 10$.



Obrázek 5.5: Experimentální řády konvergence pro n=1, 2, 3 pro skalární rovnici z příkladu 2. Řešení bylo získáno metodou BDF2a.

V tabulce 5.2 jsou uvedeny hodnoty chyby metody pro různé volby délky konstantního časového kroku. Při výpočtech se ukázalo, že není vhodné kombinovat řešení získaná pomocí metod BDF2a a BDF2b. Nejenom že se tím řád metody nezvýší, ale dokonce vychází hůře než jednotlivé metody. Je to zřejmě způsobeno tím, že zde nekombinujeme reálná čísla, jako v případě obyčejných diferenciálních rovnic, ale vektory řešení. V

	10^{-1}	5×10^{-2}	10^{-2}	5×10^{-3}	řád
$n = 1$	3.18×10^{-1}	1.48×10^{-1}	2.74×10^{-2}	1.34×10^{-2}	1.05
$n = 2$	1.14×10^{-1}	3.49×10^{-2}	1.42×10^{-3}	2.63×10^{-4}	2.02
$n = 3$	8.15×10^{-2}	1.05×10^{-2}	2.58×10^{-4}	9.31×10^{-5}	2.28

Tabulka 5.2: Chyba metody pro rovnici (2.1) s pravou stranou g z příkladu 2 pro různé volby délky časového kroku. Použita je metoda BDF2a, očekávané řády konvergence jsou tedy 1, 2 a 3.

tabulce jsou proto uvedena data získana použitím pouze metody BDF2a. Je vidět, že pak řády konvergence vycházejí relativně ve shodě s teorií, pouze pro $n = 3$ nedosahuje experimentální řád očekávané hodnoty 3. To je patrně dáno tím, že výsledná chyba závisí na chybě časové i prostorové diskretizace, je totiž řádu $O(h^p+\tau^n)$. Z toho je vidět, že když je τ^n malé, převládne člen h^p a zvyšování řádu přesnosti v čase nemá již žádný efekt.

5.2 Srování adaptivní metody s neadaptivní

V této části budeme porovnávat efektivitu adaptivní a neadaptivní metody pro různá zadání. Jako měřítko efektivity budemem brát počet iterací, které metoda potřebuje k získání výsledku s požadovanou přesností. U neadaptivní metody se zkouší počítat s různými délkamy časového kroku, metodou půlení intervalu se hledá taková délka, aby se výsledná chyba nelišila od požadované o více než o 1% u obyčejných diferenciálních rovnic nebo o více než 10% u skalární rovnice. U adaptivních metod hraje roli parametru místo délky časového kroku tolerance TOL.

5.2.1 Obyčejné diferenciální rovnice

		10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
konstantní	$n=1$	1375	4474	14365	45790	143641
	$n=2$	642	1425	3197	6972	15110
	$n=3$	410	855	1586	2879	5222
adaptivní	$n=1$	34	81	241	965	2520
	$n=2$	26	36	65	145	266
	$n=3$	24	29	43	70	108

Tabulka 5.3: Počty iterací potřebné k dosažení přesnosti 10^{-2} až 10^{-6} pří řešení obyčejné diferenciální rovnice (5.1)

Nejdříve opět uveďme výsledky pro obyčejné diferenciální rovnice, začněme s příkladem 1 z předchozí kapitoly. Srovnání efektivnosti adaptivní a neadaptivní metody pro rovnici (5.1) z příkladu 1 je uvedeno v tabulce 5.3. Je vidět, že použití adaptivní metody se v tomto případě projeví velmi výrazně. To je dáno volbou diferenciální rovnice. Její přesné řešení se totiž na většině intervalu téměř nemění a vystačíme zde s poměrně dlouhým časovým krokem. Teprve na konci intervalu (v blízkosti 1) začíná řešení prudce stoupat a v této části intervalu je naopak krok potřeba velmi výrazně zkrátit.

Na obrázku 5.6 je vidět vývoj délky časového kroku v průběhu výpočtu řešení příkladu 1 s použitím adaptivní metody. Je vidět, že se krok nejdříve prodlužuje (řešení rovnice se zde téměř nemění) a až v blízkosti 1 se začíná rapidně zkracovat v souvislosti s tím, jak řešení rovnice začíná prudce narůstat. Velikost časového kroku se tedy řídí velikostí derivace řešení dané rovnice.

Příklad 3 Zkoumejme ještě obyčejnou diferenciální rovnici

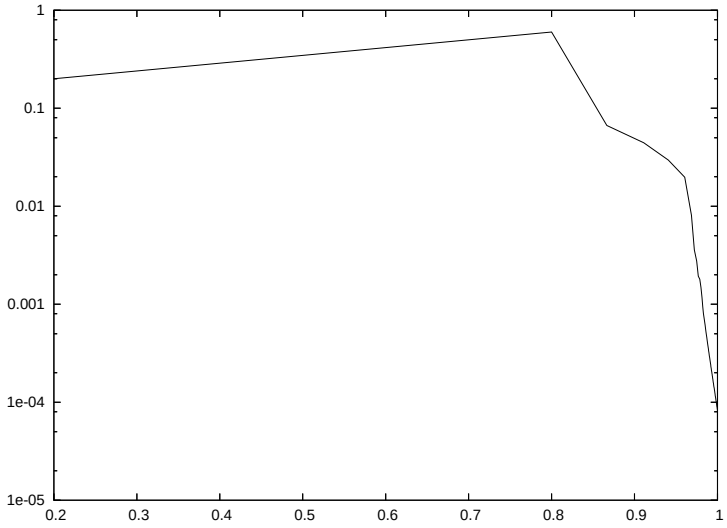
$$y'(t) = \cos\Big(t + \frac{1}{2}\sin(2t)\Big)\Big(1 + \cos(2t)\Big) \tag{5.5}$$

s přesným řešením

$$y(t) = \sin\Big(t + \frac{1}{2}sin(2t)\Big) \tag{5.6}$$

na intervalu $t \in [0, 12]$.

Srovnání efektivity neadaptivní a adaptivní metody při řešení příkladu 3 je uvedeno v tabulce 5.4. Tentokrát sice rozdíly nejsou tak velké, adaptivní

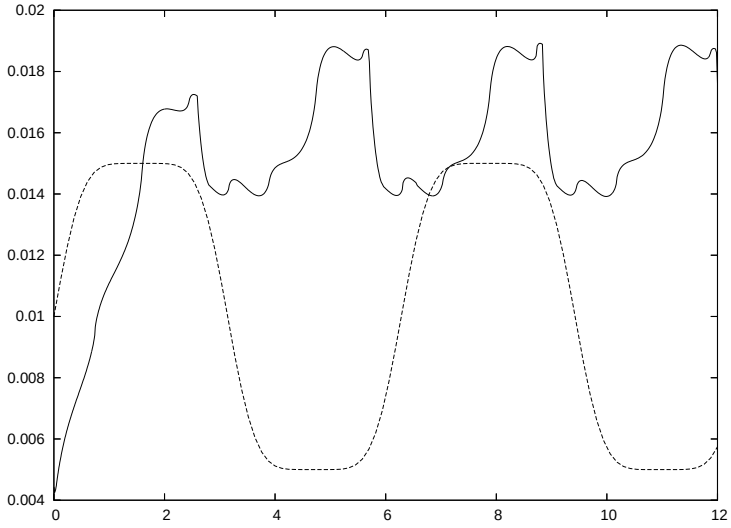


Obrázek 5.6: Velikost časového kroku (pro přehlednost v logaritmické stupnici) v průběhu výpočtu řešení rovnice (5.1) pomocí adaptivní metody s $n = 3$

		10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
konstantní	$n = 1$	402	896	4523	9894	47960
	$n = 2$	345	543	2197	3972	5110
	$n = 3$	267	553	1366	2416	5205
adaptivní	$n = 1$	450	934	3642	6832	32495
	$n = 2$	288	620	1563	2849	4386
	$n = 3$	199	385	962	1363	2374

Tabulka 5.4: Počty iterací potřebné k dosažení přesnosti 10^{-2} až 10^{-6} pří řešení obyčejné diferenciální rovnice (5.5) z příkladu 3

metoda je však i zde lepší, zejména u metod vyšších řádů. Na obrázku 5.7 je vývoj délky časového kroku při použití adaptivní metody s $n = 3$. Je vidět, že krok je delší v místech, kde se hodnota přesného řešení příliš nemění a kratší tehdy, když se mění rychleji.



Obrázek 5.7: Vývoj délky časového kroku při výpočtu řešení přikladu 3 s použitím metody s adaptivní volbou časového kroku s $n = 3$. Čárkovaně je (ve změněném měřítku, pouze pro srovnání) zobrazeno přesné řešení

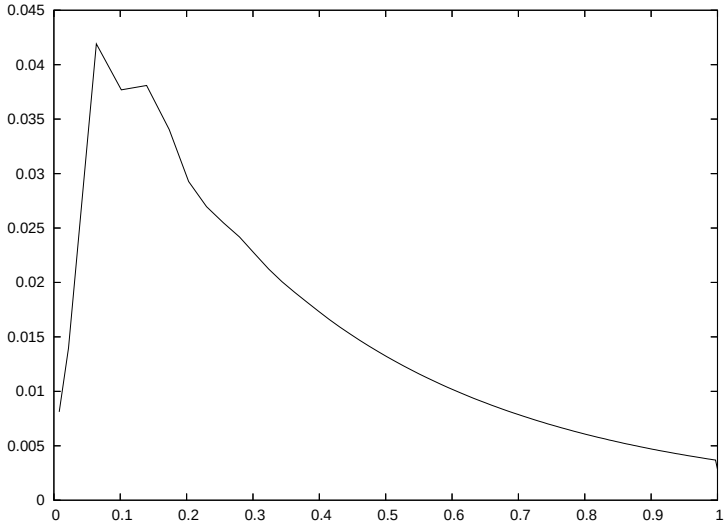
5.2.2 Skalární rovnice

V této části uvedeme srovnání efektivity neadaptivní a adaptivní metody pro skalární rovnici. Začněme s příkladem 2 z části 5.1.2.

		10^{-1}	10^{-2}	10^{-3}
konstantní	$n = 1$	7	98	> 10000
	$n = 2$	5	27	> 10000
	$n = 3$	4	18	8973
adaptivní	$n = 1$	9	29	1335
	$n = 2$	6	11	650
	$n = 3$	5	9	321

Tabulka 5.5: Počty iterací potřebné k dosažení přesnosti 10^{-1} až 10^{-3} pří řešení rovnice z příkladu 2

V tabulce 5.5 jsou uvedeny počty iterací potřebné k získání různě přesných řešení příkladu 2 pomocí neadaptivních a adaptivních metod. Je vidět,



Obrázek 5.8: Velikost časového kroku v průběhu výpočtu řešení rovnice z Příkladu 2 pomocí adaptivní metody s $n = 3$

že metoda s adaptivní volbou délky časového kroku je opět výrazně efektivnější. Důvod je vidět na obrázku 5.8. Příklad je volen jako součin časové a prostorové složky. Je to jakýsi „kopec“, který v čase roste tak, jak se zvětšuje hodnota časové slolžky. Ta je tvořena exponencielou, která se na většině intervalu téměř nemění a až v blízkosti 1 začíná prudce stoupat. Zatímco na začátku intervalu je možné brát delší časový krok, ke konci je naopak potřeba krok zjemňovat, což, jak je vidět z obrázku 5.8, adaptivní algoritmus provádi.

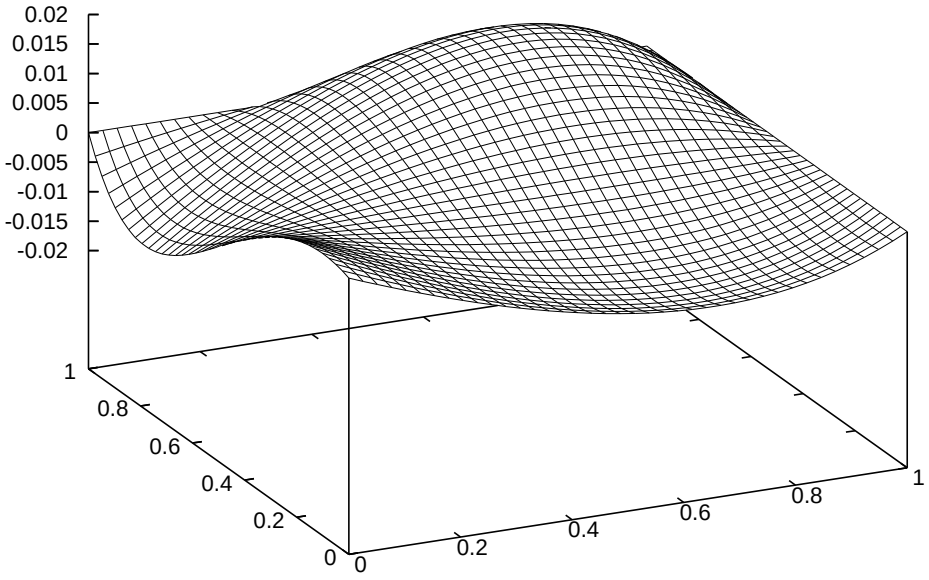
Příklad 4 Nyní se zabývejme rovnicí (2.1) s přesným řešením

$$\bar{u} = \Big(xy^2 - y^2 e^{2(x-1)} - xe^{3(y-1)} + e^{2x+3y-5}\Big)\Big(1 - e^{-\alpha t}\Big) \tag{5.7}$$

na oblasti $[0,1] \times [0,1]$ a časovém intervalu $t \in [0,1]$. Přesné řešení v čase $t = 1$ je vidět na obrázku 5.9

Podobně jako u příkladu 2 nyní spočteme pravou stranu g rovnice (2.1)

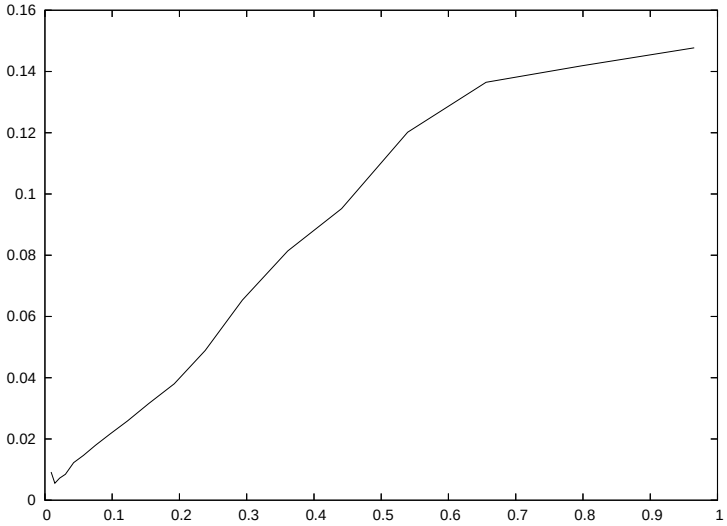
$$\begin{aligned}
\frac{\partial \bar{u}}{\partial t} &= \Big(xy^2 - y^2 e^{2(x-1)} - xe^{3(y-1)} + e^{2x+3y-5}\Big)\alpha e^{-\alpha t} \\
\frac{\partial \bar{u}}{\partial x} &= \Big(y^2 - 2y^2 e^{2(x-1)} - e^{3(y-1)} + 2e^{2x+3y-5}\Big)\Big(1 - e^{-\alpha t}\Big) \\
\frac{\partial \bar{u}}{\partial y} &= \Big(2xy - 2ye^{2(x-1)} - 3xe^{3(y-1)} + 3e^{2x+3y-5}\Big)\Big(1 - e^{-\alpha t}\Big)
\end{aligned}$$



Obrázek 5.9: Přesné řešení příkladu 4 v čase $t = 1$.

		10^{-1}	10^{-2}	10^{-3}
konstantní	$n = 1$	5	39	119
	$n = 2$	5	35	95
	$n = 3$	4	29	88
adaptivní	$n = 1$	4	13	17
	$n = 2$	4	11	14
	$n = 3$	4	10	16

Tabulka 5.6: Počty iterací potřebné k dosažení přesnosti 10^{-1} až 10^{-3} pří řešení rovnice z příkladu 4



Obrázek 5.10: Velikost časového kroku v průběhu výpočtu řešení rovnice z Příkladu 4 pomocí adaptivní metody s $n = 3$

$$\begin{aligned}\frac{\partial^2 \bar{u}}{\partial x^2} &= \left(-4y^2 e^{2(x-1)} + 4e^{2x+3y-5}\right)\left(1 - e^{-\alpha t}\right) \\ \frac{\partial^2 \bar{u}}{\partial y^2} &= \left(2x - 2e^{2(x-1)} - 9xe^{3(y-1)} + 9e^{2x+3y-5}\right)\left(1 - e^{-\alpha t}\right).\end{aligned}$$

Opět položíme

$$g = \frac{\partial \bar{u}}{\partial t} + \bar{u}\left(\frac{\partial \bar{u}}{\partial x} + \frac{\partial \bar{u}}{\partial y}\right) - \epsilon\left(\frac{\partial^2 \bar{u}}{\partial x^2} + \frac{\partial^2 \bar{u}}{\partial y^2}\right) \tag{5.8}$$

Při numerických výpočtech byla použita hodnota $\alpha = 200$.

V tabulce 5.6 je uvedeno srovnání metod s konstantní a adaptivní volbou časového kroku. Je vidět, že adaptivní metoda je opět efektivnější. Přesné řešení příkladu 4 je i tentokrát tvořeno součinem prostorové a časové složky, časová složka je však volena jinak než v příkladu 2. Zde na začátku časového intervalu ostře stoupá a na zbytku se již téměř nemění. Tomu odpovídá průběh velikosti časového kroku při výpočtu pomocí adaptivní metody, kteý je vidět z obrázku 5.10.

Kapitola 6

Závěr

V této práci jsme se zabývali řešením skalární nelineární konvektivně–difusní rovnice pomocí nespojité Galerkinovy metody. Hlavním cílem práce bylo vytvoření algoritmu pro adaptivní volbu časového kroku a jeho implementace do již existujícího programového balíku. Za tímto účelem byla odvozena dvojice implicitních metod pro řešení soustav obyčejných diferenciálních rovnic. Pomocí rozdílu hodnot řešení, získaných pomocí nich, byl navržen postup pro odhad lokální chyby v každém kroku výpočtu. Na jeho základě se určí velikost následujícího časového kroku.

Byla provedena řada numerických simulací, jak pro obyčejné diferenciální rovnice, tak pro skalární rovnici. Výsledky pro obyčejné diferenciální rovnice potvrzují očekávané řády konvergence metod a prokazují vysokou efektivitu adaptivní volby časového kroku.

Při výpočtech se skalární rovnicí se ukázalo, že není vhodné kombinovat řešení získaná pomocí obou metod v řešení s (teoreticky) vyšším řádem konvergence. Výsledné řešení totiž vychází dokonce hůře než řešení získaná pomocí jednotlivých metod. Pokud se řešení nekombinují, vycházejí experimentální řády konvergence podle očekávání. Řešení získané pomocí druhé metody se tak využije pouze při odhadu chyby a adaptivní volbě časového kroku. I zde je ovšem adaptivní metoda opět výrazně efektivnější než metoda s konstantním časovým krokem.

V další práci by bylo vhodné rozšířit zde popsanou metodu na případ Naverových–Stokesových rovnic a simulaci proudění.

Literatura

[1] V. Dolejší, M. Feistauer and J. Hosman. *Analysis of semi-implicit DGFEM for nonlinear convection-diffusion problems.* Comput. Methods Appl. Mech. Eng., (submitted).

[2] V. Dolejší and M. Feistauer. *Error estimates of the discontinuous Galerkin method for nonlinear nonstationary convection-diffusion problems.* Numer. Funct. Anal. Optim., 26(25-26):2709–2733, 2005..

[3] V. Dolejší, M. Feistauer, and V. Sobotíková. *A discontinuous Galerkin method for nonlinear convection-diffusion problems.* Comput. Methods Appl. Mech. Eng. 194:2709-2733, 2005

[4] V. Dolejší. *Higher order semi-implicit discontinuous Galerkin finite element schemes for compressible flow simulation* Proceeding of Software and Algoritms of Numerical Mathematics, 2005 (submitted)

[5] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. *Unified analysis of discontinuous Galerkin methods for elliptic problems.* SIAM J. Numer. Anal., 39(5):1749–1779, 2002.

[6] B. Cockburn. *Discontinuous Galerkin methods for convection dominated problems.* In T. J. Barth and H. Deconinck, editors, High–Order Methods for Computational Physics, Lecture Notes in Computational Science and Engineering 9, pages 69–224. Springer, Berlin, 1999.

[7] M. Feistauer, J.Felcman, and I. Straškraba. *Mathematical and Computational Methods for Compressible Flow.* Oxford University Press, Oxford, 2003.

[8] P. L. Lions. *Mathematical Topics in Fluid Mechanics.* Oxford Science Publications, 1996.

[9] K. Rektorys. *The Method of Discretization in Time and Partial Differential Equations.* Reidel, Dodrecht, 1982.

[10] P. Lotstedt, S. Soderberg, A. Ramage, L. Hemmingsson-Franden. *Implicit Solution of Hyperbolic Equations with Space–Time Adaptivity* Bit,Vol. 42, No.1, pp. 134-158, 2002

[11] A. Kufner, O. John, and S. Fučík. *Function Spaces.* Academia, Prague, 1977.