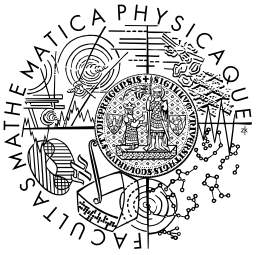


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Jakub Kurnas

Volba zastavovacích kritérií pro metody Newtonova typu

Katedra numerické matematiky

Vedoucí bakalářské práce: prof. RNDr. Vít Dolejší, Ph.D., DSc.

Studijní program: Matematika

Studijní obor: Obecná matematika

Praha 2017

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Volba zastavovacích kritérií pro metody Newtonova typu

Autor: Jakub Kurnas

Katedra: Katedra numerické matematiky

Vedoucí bakalářské práce: prof. RNDr. Vít Dolejší, Ph.D., DSc., Katedra numerické matematiky

Abstrakt: Formulujeme příklady parciálních diferenciálních rovnic, jejichž diskretizací se dostáváme k nelineárním soustavám rovnic algebraických. Nastiňujeme diskretizaci nespojitou Galerkinovou metodou, formulujeme pojmy diskretizační, algebraická chyba. Odvozujeme Newtonovu metodu pro řešení nelineárních algebraických soustav pomocí sekvence lineárních problémů, modifikujeme jí a zabýváme se její implementací. Za pomoci zavedených chyb formulujeme zastavovací kritéria pro metodu Newtonova typu a popisujeme, jak vyvážit přesnost řešení algebraického systému a původní parciální diferenciální rovnice. Odvozené ilustrujeme praktickými výpočty a provádíme několik základních pozorování týkajících se řešení různých soustav algebraických rovnic různými modifikacemi Newtonovy metody.

Klíčová slova: metody Newtonova typu, zastavovací kritéria, soustavy lineárních a nelineárních algebraických rovnic

Title: The choice of the stopping criteria for Newton-like methods

Author: Jakub Kurnas

Department: Department of Numerical Mathematics

Supervisor: prof. RNDr. Vít Dolejší, Ph.D., DSc., Department of Numerical Mathematics

Abstract: We formulate examples of partial differential equations which can be solved through their discretization and subsequent solution of derived algebraic system. A brief summary of Discontinuous Galerkin Discretization is given as well as definitions of algebraic and discretization errors. We derive the Newton method, which solves nonlinear algebraic systems by solving a sequence of linear problems, we modify the method and examine implementation options. We define stopping criteria for the Newton-like method using aforementioned errors and we explain how to keep accuracy of the solution of derived algebraic system and the original partial differential equation in balance. We present numerical experiments to illustrate theoretical background and mention several basic properties of the Newton-like method.

Keywords: Newton-like methods, stopping criteria, systems of linear and nonlinear algebraic equations

Děkuji rodině a profesoru Dolejšímu za trpělivost a ochotu.

Obsah

Úvod

Metody Newtonova typu jsou nástrojem pro řešení soustav nelineárních algebraických rovnic, které převádějí na posloupnosti lineárních algebraických problémů generující aproximace řešení. Jak vysvětlujeme v této práci, takové soustavy vznikají při řešení praktických problémů, jako například při řešení nelineárních parciálních diferenciálních rovnic. Hlavním cílem tohoto textu je popsat základní pojmy spojené s metodami Newtonova typu a v sekci o numerických experimentech ilustrovat, jak se jednoduché teoretické úvahy o konvergenci těchto metod projevují v praxi. Klíčovým pojmem jsou „zastavovací kritéria“, tedy parametry, podle kterých iterativní algoritmus rozhoduje, kdy dosáhl dostatečně přesného řešení.

V prvních dvou kapitolách uvedeme příklady parciálních diferenciálních rovnic, konkrétně obecnou rovnici *konvekce-difuze-reakce*, *Navierovy-Stokesovy* rovnice popisující proudění vazké stlačitelné tekutiny a *Richardsovu* rovnici popisující proudění tekutiny v porézním prostředí. Stručně nastíníme diskretizaci těchto rovnic *Nespojitou Galerkinovou metodou*, pomocí které přecházíme od problému řešení těchto rovnic k problému řešení soustav nelineárních algebraických rovnic, k čemuž se využívá metod Newtonova typu.

Těžištěm práce je potom část o Newtonově metodě: odvozujeme Newtonovu metodu nejprve ve skalárním případě a následně ji zobecňujeme na n-rozměrný prostor. Dále formulujeme Picardovu metodu pro hledání pevného bodu funkce a popisujeme její vztah k Newtonově metodě.

Ústředním tématem studia iterativních metod jsou jejich konvergenční vlastnosti. Nejinak tomu je v naší práci: popisujeme, jakými pojmy se popisuje přesnost a „efektivita“ iterativních metod, Newtonovu i Picardovu metodu modifikujeme; důležitou modifikací je tzv. *Andersonova akcelerace*, která má zajistit jejich větší robustnost a rychlejší konvergenci. Formulujeme zastavovací kritéria našich metod. Dále formulujeme několik tvrzení o jejich konvergenci a vysvětlujeme problémy, se kterými se můžeme potkat při aplikaci teoretické analýzy konvergence na praktické problémy.

V sekci o numerických experimentech potom zkoumáme, jaké vlastnosti vykazují definované metody při jejich použití na soustavy nelineárních rovnic vzešlých z diskretizace parciálních diferenciálních rovnic. Konkrétně, sledujeme chování *algebraické* a *diskretizační* chyby, potažmo dalších vlastností jako délky výpočtu v závislosti na rozsahu a povaze řešených problémů.

1. Příklady parciálních diferenciálních rovnic

V následujícím zformulujeme obecnou parciální diferenciální rovnici, popisující evoluční proces s konvekcí, difuzí a reakcí. Dále uvedeme dva konkrétní příklady: Navierovu-Stokesovu rovnici popisující proudění vazké nestlačitelné tekutiny a Richardsovu rovnici popisující proudění v porézním prostředí.

Rovnice, které zde uvedeme, se řeší v jisté výpočetní oblasti $Q_T = \Omega \times (0,T) \in \mathbb{R}^d \times \mathbb{R}$, tj. v prostorové oblasti Ω a na časovém intervalu $(0,T)$: aby byly plně zadány, je potřeba předepsat jisté *okrajové* a *počáteční* podmínky. Ty značně závisí na řešeném problému. Okrajové podmínky popisují chování funkce na hranici výpočetní oblasti; počáteční podmínky popisují stav systému v čase $t = 0$.

1.1 Obecná rovnice konvekce-difuze-reakce

Obecnou soustavu n rovnic konvekce-difuze-reakce můžeme formulovat na dané výpočetní oblasti Q_T následujícím způsobem (v souladu s [1]):

$$\frac{\partial \boldsymbol{\xi}(\boldsymbol{u})}{\partial t} - \sum_{i=1}^{d} \frac{\partial}{\partial x_i} \left(\sum_{j=1}^{d} \mathcal{H}_{ij}(\boldsymbol{u}) \frac{\partial \boldsymbol{u}}{\partial x_j} \right) + \sum_{i=1}^{d} \frac{\partial \boldsymbol{f}_i(\boldsymbol{u})}{\partial x_i} + \boldsymbol{s}(\boldsymbol{u}) = \boldsymbol{g}. \tag{1.1}$$

Tuto soustavu řešíme pro $\boldsymbol{u} : Q_T \to \mathbb{R}^n$, přičemž máme zadány funkce $\boldsymbol{\xi}, \boldsymbol{s} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathcal{H}_{ij} \in \mathbb{R}^{n \times n}, i,j = 1,\ldots,d$, $\boldsymbol{f} = (\boldsymbol{f}_1,\ldots,\boldsymbol{f}_d)$, $\boldsymbol{f}_i : \mathbb{R}^n \to \mathbb{R}^n$, $i = 1,\ldots,d$, $\boldsymbol{g} : Q_T \to \mathbb{R}^n$. První člen rovnice popisuje závislost na čase, druhý člen difuzi, třetí člen konvekci, čtvrtý člen reakci a pravá strana zdroj sil působících na systém.

Různou volbou parametrů dostáváme rovnice popisující různé fyzikální procesy. Můžeme popsat vazké, respektive nevazké, stlačitelné proudění ($\boldsymbol{\xi} = id$, $\boldsymbol{s} = 0$, $\boldsymbol{g} = 0$, $\mathcal{H} \equiv 0$, respektive $\mathcal{H} \neq 0$, viz sekce 1.2), kde vazkost odpovídá „vnitřnímu tření“ tekutiny. Dále můžeme modelovat proudění tekutiny v porézním prostředí, tj. např. prosakování vody do země ($\boldsymbol{f} = 0, \boldsymbol{s} \neq 0$, viz sekce 1.3). Jako poslední příklad uvedeme proudění mělké vody ($\boldsymbol{\xi} = id, \mathcal{H} = 0, \boldsymbol{g} = 0$), které popisuje proudění kapalin v místech, kde je její plocha výrazně větší než hloubka, například na pobřeží moře.

1.2 Navierovy-Stokesovy rovnice

Formulaci Navierových-Stokesových rovnic přejímame z článku [2]. Popisujeme jimi vazké stlačitelné proudění tekutiny, např. proudění vzduchu kolem profilu křídla letadla. Systém Navierových-Stokesových rovnic lze psát jako

$$\frac{\partial \mathbf{w}}{\partial t} + \sum_{i=1}^{d} \frac{\partial \mathbf{f}_i(\mathbf{w})}{\partial x_i} = \sum_{i=1}^{d} \frac{\partial \mathbf{R}_i(\mathbf{w}, \nabla \mathbf{w})}{\partial x_i} \quad \text{v } Q_T,$$

kde $\mathbf{w} = (\rho,\rho v_1,\ldots,\rho v_d,e)^T$ je neznámý stavový vektor (ρ je hustota, $\mathbf{v} = (v_1,\ldots,v_d)$ je vektor rychlosti a e je energie), $\mathbf{f}_i : \mathbb{R}^{d+2} \to \mathbb{R}^{d+2}$, $i = 1,\ldots,d$ a $\mathbf{R}_i : \mathbb{R}^{2d} \times \mathbb{R}^{2(d+2)} \to \mathbb{R}^4$, $i = 1,\ldots,d$ reprezentují vazké, respektive nevazké toky.

1.3 Richardsova rovnice

Richardsova rovnice popisuje proudění kapalin v porézním prostředí a její formulaci přejímáme z článku [3]. Poznamenejme, že jde o speciální případ rovnice konvekce-difuze-reakce. Buď tedy

$$\frac{\partial\theta(h)}{\partial t} - \nabla\cdot(Kk_r(h)\nabla(h+z)) = 0, \tag{1.2}$$

kde h je normovaný tlak tekutiny, $\theta(h) = s(h)\phi$ obsah vody se saturací $s(h)$ a porozitou materiálu ϕ, K je hydraulická vodivost (veličina popisující jak snadno se může tekutina pohybovat skrz póry či trhliny materiálu), $k_r(h)$ je relativní prostupnost vody vzhledem ke vzduchu, t je čas, z je vertikální souřadnice.

2. Diskretizace rovnice konvekce-difuze-reakce

Cílem této kapitoly bude ilustrovat způsob, jakým diskretizací parciálních diferenciálních rovnic pomocí *Nespojité Galerkinovy metody* (*Discontinuous Galerkin Method*, v dalším též DG metoda) docházíme k nelineárním algebraickým problémům, které se řeší metodami Newtonova typu. Důsledné provedení diskretizace by bylo nad rámec bakalářské práce, omezíme se proto zejména na zavedení pojmů, které budou užitečné v dalších kapitolách. Vycházíme zejména z článku [2].

Mějme výpočetní oblast Q_T, na které budeme řešit systém parciálních diferenciálních rovnic (1.1) s vhodnými okrajovými a počátečními podmínkami. V následujícím popíšeme, jak se nespojitou Galerkinovou diskretizací dá jejich řešení převést na problém řešení soustav nelineárních rovnic.

2.1 Diskretizace výpočetní oblasti

Uvažme dělení $0 = t_0 < t_1 < \ldots < t_r = T$ intervalu $(0,T)$, $r \in \mathbb{N}$. Jednotlivé podintervaly dělení označíme $I_m := (t_{m-1}, t_m]$, velikost časového kroku $\tau_m := t_m - t_{m-1}$, $m = 1, \ldots, r$ a délku nejdelšího z nich $\tau := \max_{m=1,\ldots,r} \tau_m$. Položme dále $\mathcal{I}_\tau := \{I_m\}_{m=1}^r$. Pak $\mathcal{I}_\tau$ je dělení časového intervalu.

Uvažme nyní časovou hladinu t_m, $m = 0, \ldots, r$ a na ní triangulaci $\mathcal{T}_{h_m,m}$, tj. konečný systémem disjunktních, uzavřených trojúhelníků pokrývající oblast $\bar{\Omega}$ $\forall m = 1, \ldots, r$. Index h_m označuje diametr nejvetšího z trojúhelníků triangulace příslušné dané časové hladině, tj. $h_m := \max_{K \in \mathcal{T}_{h_m,m}} \mathrm{diam}(K)$. Konečně, zavádíme označení pro množinu prostorových dělení na všech časových hladinách $\mathcal{T}_h := \{\mathcal{T}_{h_m,m}\}_{m=1}^r$, kde $h := \max_{m=1,\ldots,r} h_m$ udává diametr největšího trojúhelníku z triangulací na všech časových hladinách.

Tedy, dvojice $(\mathcal{T}_h, \mathcal{I}_\tau)$ potom tvoří časoprostorové dělení výpočetní oblasti Q_T.

2.2 Přechod k nelineární soustavě rovnic

Na diskretizované výpočetní oblasti budeme chtít pomocí nespojité Galerkinovy diskretizace reprezentovat původní rovnici jistou formou, která bude definovaná na (nekonečnědimenzionálním) *Sobolevově prostoru*. Zároveň budeme chtít řešení původní rovnice aproximovat na (konečnědimenzionálním) prostoru funkcí, jejichž restrikce na jednotlivé časoprostorové elementy, tj. prvky (K,I_m), kde $K \in \mathcal{T}_{h,m}$, $m \in \{1, \ldots, r\}$, budou polynomiální funkce daného stupně.

2.2.1 Funkční prostory

Pro časovou, resp. prostorovou aproximaci budeme obecně uvažovat rozdílné nejvyšší stupně polynomů q, resp. p. Označme nejdříve prostor polynomiálních

funkcí stupně nejvýše $r \in \mathbb{N}_0$ na množině $M \subset \mathbb{R}^n$ jako

$$P^r(M) := \left\{ \sum_{|\alpha| \leq r} a_\alpha x^\alpha \mid x \in M, \quad a_\alpha \in \mathbb{R},\ \alpha \in \mathbb{N}_0^n \right\}.$$

Potom můžeme zavést množinu vektorových funkcí, jejichž zúžení na jednotlivé prvky dělení $(\mathcal{T}_h,\mathcal{I}_\tau)$ jsou polynomy nejvýše daného stupně:

$$S_{h,p}^{\tau,q} := \{\boldsymbol{v} : Q_T \to \mathbb{R}^n; \boldsymbol{v}|_{K\times I_m} \in [P^p(K) \times P^q(I_m)]^n \\ \forall K \in \mathcal{T}_{h,m} \forall m \in \{0,1,\ldots,r\}\}.$$

Jelikož máme pro původní rovnici (1.1) zadány počáteční podmínky, je výhodné provádět některé úvahy na jednotlivých časových hladinách. Tomu by měla odpovídat i volba našich prostorů. Zavedeme proto prostor funkcí, jejichž zúžení na prvky dělení $\mathcal{T}_{h,m} \times I_m$ na dané časové hladině t_m jsou polynomiální vektorové funkce daného stupně, tj. pro $m = 0,1,\ldots,r$

$$S_{m,h,p}^{\tau,q} := \{\boldsymbol{v} : \Omega \times I_m \to \mathbb{R}^n; \boldsymbol{v}|_{K\times I_m} \in [P^p(K) \times P^q(I_m)]^n \ \forall K \in \mathcal{T}_{h,m}\}.$$

V následujícím budeme používat pojem *Sobolevův prostor* $H^2(M)$. Jedná se o prostor funkcí, jejichž zobecněné derivace druhého řádu náleží prostoru $L^2(M)$ (podrobná definice pojmů viz např. [4]). Poznamenejme, že nám tento prostoro poskytuje i příslušnou normu $\|\cdot\|_{H^2(M)}$. Dále můžeme definovat *po částech Sobolevův prostor na dané triangulaci* $\mathcal{T}_{h,m}$ oblasti Ω jako prostor funkcí, jejichž zúžení na jednotlivé trojúhelníky triangulace je sobolevovská funkce:

$$H^2(\mathcal{T}_{h,m}) = \left\{\boldsymbol{v} : \Omega \to \mathbb{R}^n; \boldsymbol{v}\,|_K \in H^2(K)\ \forall K \in \mathcal{T}_{h,m}\right\}.$$

Dále definujeme *po částech Sobolevův prostor na daném dělení* $H^1(\mathcal{I}_\tau,H^2(\mathcal{T}_{h,m}))$ pro dané dělení $(\mathcal{T}_h,\mathcal{I}_\tau)$ oblasti Q_T. Nejprve definujme následující normu pro $\boldsymbol{u} \in H^2(\mathcal{T}_{h,m})$:

$$\| \boldsymbol{u} \|^2_{H^1(I_m,H^2(\mathcal{T}_{h,m}))} = \int_{I_m} (\| \boldsymbol{u} \|^2_{H^2(\mathcal{T}_{h,m})} + \| \partial_t \boldsymbol{u} \|^2_{H^2(\mathcal{T}_{h,m})}),$$

kde

$$\| \boldsymbol{u} \|^2_{\mathcal{T}_{h,m}} = \sum_{K\in\mathcal{T}_{h,m}} \| \boldsymbol{u} \|^2_{H^2(K)}.$$

Pak můžeme zavést kýžený po částech Sobolevův prostor:

$$\boldsymbol{u} \in H^1(\mathcal{I}_\tau,H^2(\mathcal{T}_{h,m})) \Leftrightarrow \sum_{m=1}^{r} \| \boldsymbol{u} \|_{H^1(I_m,H^2(\mathcal{T}_{h,m}))} < \infty.$$

Přesné odůvodnění toho, že uvedené definice daných prostorů jsou pro nás užitečné, by vycházelo z teorie o DG metodách, a vzhledem k povaze práce jej nebudeme uvádět. Pro tento text je však podstatné, že nad takovýmito prostory lze provést diskretizaci parciální diferenciální rovnice (1.1), jak bude ukázáno v dalším, a že užívané Sobolevovy prostory jsou nekonečné dimenze, což bude podstatné při definování chyby diskretizace ve třetí kapitole.

2.2.2 Diskretizace PDR

Diskretizace rovnice (1.1) na m-té časové hladině, $m = 1,\dots,r$ vede k definici formy $A^m_{h,\tau} : H^1(\mathcal{I}_\tau,H^2(\mathcal{T}_{h,m})) \times H^1(\mathcal{I}_\tau,H^2(\mathcal{T}_{h,m})) \to \mathbb{R}$, která je nelineární k první a lineární k druhé proměnné. Navíc, tato forma je *konzistentní* s původní rovnicí, tedy splňuje definici níže. Tyto skutečnosti, vycházející z teorie o DG metodě, považujeme v tomto textu za fakta.

Definice 1. *Buď $\boldsymbol{u} : Q_T \to \mathbb{R}^n$ řešení rovnice* (1.1). *Řekneme, že forma $A^m_{h,\tau}$ odvozená Galerkinovou diskretizací z rovnice* (1.1) *je s touto rovnicí* ***konzistentní****, pokud platí*

$$A^m_{h,\tau}(\boldsymbol{u},\boldsymbol{\psi}) = 0 \quad \forall \boldsymbol{\psi} \in H^1(\mathcal{I}_\tau,H^2(\mathcal{T}_{h,m})), \tag{2.1}$$

Tedy, konzistence formy zaručuje jistý vztah mezi odvozenou formou a řešením původní rovnice. Máme-li k dispozici formu $A^m_{h,\tau}$ a prostor polynomiálních funkcí $S^{\tau,q}_{h,p}$, můžeme definovat *přibližné řešení* původní parciální diferenciální rovnice.

Definice 2. *Řekneme, že funkce $\boldsymbol{u}_{h,\tau} \in S^{\tau,q}_{h,p}$ je* ***přibližné řešení*** *problému* (1.1), *pokud platí*

$$A^m_{h,\tau}(\boldsymbol{u}_{h,\tau},\boldsymbol{\psi}) = 0 \quad \forall \boldsymbol{\psi} \in S^{\tau,q}_{h,p},\ m = 1,\dots,r. \tag{2.2}$$

Jsme-li tedy schopni vyřešit problém (2.2), dostáváme pak „v jistém smyslu“ aproximaci řešení původní rovnice, tj. platí $\boldsymbol{u}_{h,\tau} \approx \boldsymbol{u}$. Jinými slovy, formy $A^m_{h,\tau}$ „reprezentují“ rovnici (1.1) nad jistým funkčním prostorem $H^1(\mathcal{I}_\tau,H^2(\mathcal{T}_{h,m}))$ na daném časoprostorovém dělení. Přibližné řešení $\boldsymbol{u}_{h,\tau}$, které díky této formě můžeme definovat, uvažované v jistém polynomiálním prostoru $S^{\tau,q}_{h,p}$, je však zatíženo *diskretizační chybou*. Zároveň však teorie o DG metodě zaručuje, že tato chyba bude pro dostatečně jemné časoprostorové dělení libovolně malá.

Hledejme nyní přibližné řešení postupně na různých časových hladinách t_m. Nechť tedy $\boldsymbol{u}^m_{h,\tau} := \boldsymbol{u}_{h,\tau}|_{\mathcal{T}_{h,m}\times I_m}$, $m = 1,\dots,r$. Pak platí $\boldsymbol{u}^m_{h,\tau} \in S^{\tau,q}_{m,h,p}$ a systém (2.2) přirozeně přechází na posloupnost systémů

$$A^m_{h,\tau}(\boldsymbol{u}^m_{h,\tau},\boldsymbol{\psi}^m) = 0 \quad \forall \boldsymbol{\psi}^m \in S^{\tau,q}_{m,h,p},\ m = 1,\dots,r. \tag{2.3}$$

Uvažujme v daném prostoru $S^{\tau,q}_{m,h,p}$, který má konečnou dimenzi $N_m \in \mathbb{N}$, bázi $\{\boldsymbol{\psi}^m_i\}_{i=1}^{N_m}$. Je možné odvodit, že pro dimenzi prostoru (a tedy počet prvků báze) bude platit $N_m = \frac{1}{2}n(q+1)(p+1)(p+2)\#\mathcal{T}_{h,m}$, kde $\#\mathcal{T}_{h,m}$ odpovídá počtu prvků dělení $\mathcal{T}_{h,m}$ na m-té časové hladině (viz [2](3.4)). Pak můžeme řešení $\boldsymbol{u}^m_{h,\tau} \in S^{\tau,q}_{m,h,p}$ psát jako

$$\boldsymbol{u}^m_{h,\tau}(\boldsymbol{x},t) = \sum_{i=1}^{N_m} u^m_i \boldsymbol{\psi}^m_i(\boldsymbol{x},t),\ \boldsymbol{x} \in \Omega,$$

kde $u^m_i \in \mathbb{R}$ jsou reálné bázové koeficienty. Pak platí

$$A^m_{h,\tau}(\sum_{i=1}^{N_m} u^m_i \boldsymbol{\psi}^m_i,\boldsymbol{\psi}^m_j) = 0, \quad \forall j = 1,\dots,N_m,\ m = 1,\dots,r.$$

Je intuitivní, že přibližné řešení $\boldsymbol{u}^m_{h,\tau}$ a tedy i forma $A^m_{h,\tau}$ budou záviset na předcházející časové hladině, konkrétně na aproximaci $\boldsymbol{u}^{m-1}_{h,\tau}$.

Definujme algebraickou reprezentaci zavedené formy $A^m_{h,\tau}$:

Definice 3. *Mějme* $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{N_m}) \in \mathbb{R}^{N_m}$ *a uvažme* m*-tou formu* $A^m_{h,\tau}$, $m = 0,1,\ldots,r$, *danou Galerkinovou diskretizací rovnice* (1.1).

1. *Pro formu* $A^m_{h,\tau}$ *definujeme funkci* $F^m_j : \mathbb{R}^{N_m} \to \mathbb{R}$,

$$F^m_j(\boldsymbol{\xi}) := A^m_{h,\tau}(\sum_{i=1}^{N_m} \xi_i \boldsymbol{\psi}^m_i, \boldsymbol{\psi}^m_j), \quad j = 1,\ldots,N_m,\ m = 1,\ldots,r,$$

2. ***Algebraickou reprezentaci formy*** $A^m_{h,\tau}$ *definujeme jako vektorovou funkci* $\boldsymbol{F}^m : \mathbb{R}^{N_m} \to \mathbb{R}^{N_m}$, *pro níž platí*

$$\boldsymbol{F}^m(\boldsymbol{\xi}) := \{F^m_j(\boldsymbol{\xi})\}_{j=1}^{N_m}, \quad m = 1,\ldots,r.$$

Tedy, naším cílem při hledání přibližného řešení $\boldsymbol{u}^m_{h,\tau}$ je najít takové $\boldsymbol{\xi} \in \mathbb{R}^{N_m}$, pro nějž bude platit

$$\boldsymbol{F}^m(\boldsymbol{\xi}) = \boldsymbol{0}, \quad m = 1,\ldots,r. \tag{2.4}$$

Přibližné řešení na m-té časové hladině pak dostaneme jako $\boldsymbol{u}^m_{h,\tau} = \sum_{i=1}^{N_m} \xi_i \boldsymbol{\psi}^m_i$. Tím se dostáváme od hledání řešení parciální diferenciální rovnice (1.1) k řešení posloupnosti nelineárních algebraických rovnic (2.4). Toto řešení budeme hledat vhodnými numerickými metodami. Konkrétněji, problém budeme řešit iteračními metodami, které generují posloupnost aproximací řešení $\boldsymbol{\xi}_k$, $k = 0,1,\ldots$, takových, že platí $\boldsymbol{\xi}_k \to \boldsymbol{\xi}, k \to \infty$. Při takovém postupu je vhodné mít na paměti, že vektor $\boldsymbol{\xi}$ odpovídá „přibližnému“ řešení $\boldsymbol{u}_{h,\tau}$ rovnice (1.1).

2.3 Matice toku

V části o metodách Newtonova typu budeme používat Jacobián funkce $\boldsymbol{F}^m$. Zdůvodníme v ní, že pro nás bude výhodné pracovat s jistou aproximací tohoto Jacobiánu, kterou budeme označovat jako *matici toku*. V následujícím odstavci stručně naznačíme, jak se taková matice toku $\mathbb{C}$ aproximující Jacobián může zavést.

Připomeňme nejprve definici Jacobiánu.

Definice 4. *Buď* $D_G \subset \mathbb{R}^N$ *otevřená množina a* $\boldsymbol{G} : D_G \to \mathbb{R}^N$ *vektorová funkce. Pak Jacobiánem funkce* $\boldsymbol{G} = (\boldsymbol{G}_1,\ldots,\boldsymbol{G}_N)$ *v bodě* $\boldsymbol{\xi} \in D_G$ *definujeme jako*

$$\boldsymbol{G}'(\boldsymbol{\xi}) := \begin{pmatrix} \frac{\partial \boldsymbol{G}_1}{\partial \xi_1}(\boldsymbol{\xi}) & \cdots & \frac{\partial \boldsymbol{G}_1}{\partial \xi_N}(\boldsymbol{\xi}) \\ \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{G}_N}{\partial \xi_1}(\boldsymbol{\xi}) & \cdots & \frac{\partial \boldsymbol{G}_N}{\partial \xi_N}(\boldsymbol{\xi}) \end{pmatrix} = \left\{ \frac{\partial \boldsymbol{G}^m_i}{\partial \xi_j}(\boldsymbol{\xi}) \right\}_{i,j=1}^{N}.$$

Mějme formu $A^m_{h,\tau}(\boldsymbol{u},\boldsymbol{\psi})$ danou Galerkinovou diskretizací rovnice (1.1). Připomeňme, že tato forma je nelineární vzhledem k první a lineární vzhledem k druhé proměnné. V řadě aplikací, včetně těch zmíněných v první kapitole, lze tuto formu vhodnou DG diskretizací vyjádřit jako

$$A^m_{h,\tau}(\boldsymbol{u},\boldsymbol{\psi}) = \mathbb{C}(\boldsymbol{u},\boldsymbol{u},\boldsymbol{\psi}) + \boldsymbol{d}(\boldsymbol{u},\boldsymbol{\psi}), \quad \forall \boldsymbol{\psi},\boldsymbol{u} \in S^{\tau,q}_{m,h,p},$$

kde $\mathbb{C}$ je forma lineární k druhé a třetí proměnné a $\boldsymbol{d}$ forma lineární k druhé proměnné (konkrétní tvar viz [2] . Forma $\mathbb{C}$ reprezentuje linearizaci formy $A^m_{h,\tau}$; forma $\boldsymbol{d}$ je nulová pro většinu bázových funkcí prostoru $S^{\tau,q}_{m,h,p}$. Pro zafixované hodnoty $\hat{\boldsymbol{u}}$ v prvních proměnných funkcí $\mathbb{C}, \boldsymbol{d}$ a vyjádřením $\boldsymbol{u} = \sum_{k=1}^{N_m} u_k \boldsymbol{\psi}^m_k$ funkce $\boldsymbol{u}$ v příslušné bázi prostoru $S^{\tau,q}_{m,h,p}$ dostáváme vztah

$$\frac{\partial \boldsymbol{F}_i}{\partial \xi_j} = \frac{\partial}{\partial u_j} A^m_{h,\tau}(\boldsymbol{u}, \boldsymbol{\psi}_i) \approx \frac{\partial}{\partial u_j}(\mathbb{C}(\hat{\boldsymbol{u}}, \sum_{k=1}^{N_m} u_k \boldsymbol{\psi}^m_k, \boldsymbol{\psi}_i) + \boldsymbol{d}(\hat{\boldsymbol{u}}, \boldsymbol{\psi}_i)) =$$

$$= \sum_{k=1}^{N_m} \{\frac{\partial}{\partial u_j}(u_k \mathbb{C}(\hat{\boldsymbol{u}}, \boldsymbol{\psi}^m_k, \boldsymbol{\psi}_i))\} + \frac{\partial}{\partial u_j} \boldsymbol{d}(\hat{\boldsymbol{u}}, \boldsymbol{\psi}_i) = \mathbb{C}(\hat{\boldsymbol{u}}, \boldsymbol{\psi}_j, \boldsymbol{\psi}_i).$$

Tedy, jelikož jsme definovali $\boldsymbol{F}^m_i(\boldsymbol{\xi}) = A^m_{h,\tau}(\sum_{k=1}^{N_m} \xi_k \boldsymbol{\psi}^m_k, \boldsymbol{\psi}^m_i)$, dostáváme pro hledaný Jacobián funkce $\boldsymbol{F}^m$ přibližný vztah

$$\frac{\partial \boldsymbol{F}^m_i}{\partial \xi_j}(\boldsymbol{\xi}) \approx \mathbb{C}_{i,j}(\boldsymbol{\xi}),$$

kde definovaná matice $\mathbb{C}(\boldsymbol{\xi}) := \{\mathbb{C}_{i,j}(\boldsymbol{\xi})\}^{N_m}_{i,j=1} := \mathbb{C}(\hat{\boldsymbol{u}}, \boldsymbol{\psi}^m_j, \boldsymbol{\psi}^m_i)$ je kýžená matice toku. Poznamenejme, že takto definovaná matice $\mathbb{C}$ má stejnou blokovou strukturu jako původní Jacobiho matice $\{\frac{\partial \boldsymbol{F}_i}{\partial \xi_j}\}$, ve které jednotlivé bloky odpovídají prvkům prostorové diskretizace na m-té časové hladině $K \in \mathcal{T}_{h_m,m}$, přesněji, že matice sestává z diagonálních bloků doplněných o mimodiagonální bloky, odpovídající po řádcích sousednosti elementů triangulace. Matice tedy bude řídká, tj. bude mít mnoho nulových prvků.

3. Řešení soustav nelineárních rovnic

V této kapitole se budeme zabývat řešením soustav nelineárních algebraických rovnic (2.4). Mějme nelineární funkci $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$, $n \in \mathbb{N}$ a uvažme problém hledání řešení systému

$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}. \tag{3.1}$$

Chtěli bychom popsat iterativní algoritmus aproximující přesné řešení $\boldsymbol{x}^* \in \mathbb{R}^n$ této soustavy - tj. takový bod, pro který platí $\boldsymbol{F}(\boldsymbol{x}^*) = \boldsymbol{0}$ - na základě vhodné počáteční aproximace řešení $\boldsymbol{x}_0 \in \mathbb{R}^n$. Tj. chceme, aby algoritmus generoval posloupnost bodů $\boldsymbol{x}_k \in \mathbb{R}^n$, pro které bude platit

$$\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}^*,$$

přičemž bod $\boldsymbol{x}_k$ bude dán pomocí bodu $\boldsymbol{x}_{k-1}$. Základní myšlenkou konstrukce algoritmu, který bude aproximovat kořen $\boldsymbol{x}^*$, je zobecnění klasické Newtonovy metody pro skalární funkce, kterou uvedeme v následujícím.

3.1 Klasická Newtonova metoda

3.1.1 Analytické odvození

Uvaž (jako např. v [5](1.1)) funkci $f : \mathbb{R} \to \mathbb{R}$, $f \in C^2(\mathbb{R})$, a počáteční odhad $x_0 \in \mathbb{R}$ přesného řešení $x^* \in \mathbb{R}$, čili takového čísla, pro které platí $f(x^*) = 0$ (tj. uvažujeme systém (3.1) pro $n = 1$). Chceme najít posloupnost $\{x_k\}_{k=1}^{\infty} \subset \mathbb{R}$, pro kterou bude platit

$$\lim_{k\to\infty} x_k = x^*.$$

Označme $\Delta x := x^* - x_0$. Použitím Taylorova rozvoje dostáváme

$$0 = f(x_0 + \Delta x) = f(x_0) + f'(x_0)\Delta x + O(|\ \Delta x\ |^2)$$

a zanedbáním členů vyššího než lineárního řádu docházíme ke vztahu

$$0 \approx f(x_0) + f'(x_0)\Delta x.$$

Pravá strana je potom pro pevné x_k lineární funkcí, a proto můžeme za předpokladu $f'(x_k) \neq 0$ snadno nalézt Δx, pro které se rovná nule. Označme ho Δx_1, tedy

$$\Delta x_1 := -\frac{f(x_0)}{f'(x_0)}.$$

Pak můžeme definovat první aproximaci řešení jako $x_1 := x_0 + \Delta x_1$. Obecně, polož

$$\Delta x_{k+1} := -\frac{f(x_k)}{f'(x_k)},\ k = 0,1,\ldots.$$

Potom pokládáme $(k+1)$. aproximaci řešení jako

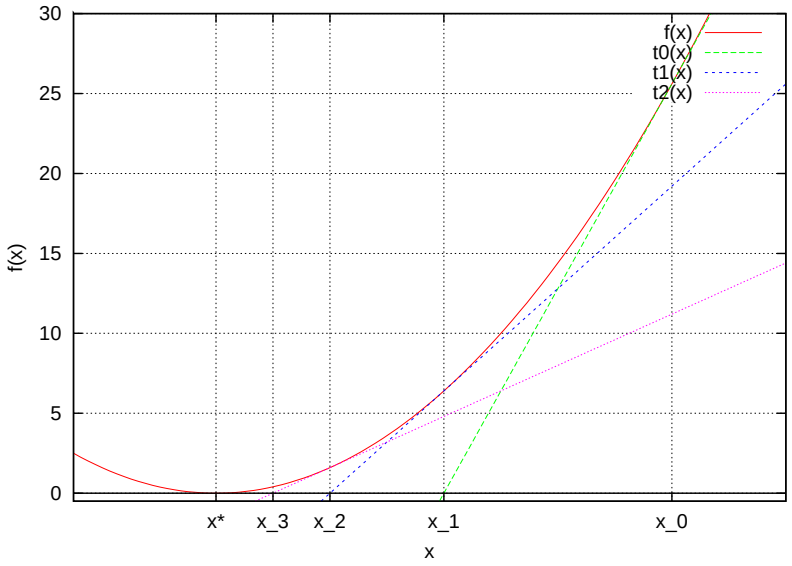
$$x_{k+1} := x_k + \Delta x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0,1,\ldots, \tag{3.2}$$

čímž dostáváme předpis pro iterativní algoritmus, který bude za jistých předpokladů konvergovat k hledanému kořenu x^*.

3.1.2 Geometrické odvození

Newtonova metoda v $\mathbb{R}$ má také názornou geometrickou interpretaci. Máme-li danou funkci $f : \mathbb{R} \to \mathbb{R}$, řešení problému $f(x) = 0$ odpovídá hledání průsečíku funkce f s x-ovou osou. Pokud existuje počáteční odhad $x_0 \in \mathbb{R}$ kořenu našeho problému, můžeme zkusit v daném bodě zkonstruovat tečnu k naší funkci a za další aproximaci řešení zvolit průsečík této tečny s osou x. Tato metoda bude skutečně pro vhodné funkce a vhodný počáteční odhad konvergovat k hledanému řešení; specielně, je známo, že metoda bude konvergovat pro globálně konvexní či konkávní funkci f i pro špatný počáteční odhad.

Uvedené ilustrujeme v obrázku 3.1. Máme-li zadán počáteční odhad x_0 kořenu x^* funkce f, můžeme sestrojit tečnu $t_0(x)$ k zadané funkci v bodu $[x_0,f(x_0)]$, jako novou aproximaci řešení x_1 označit průsečík této tečny s osou x a analogicky postupovat dále. V našem případě je vidět, že takovýto postup konverguje ke kořenu. Poznamenejme, že popsaná metoda přesně odpovídá vzorci (3.2).



Obrázek 3.1: Geometrické odvození Newtonovy metody.

3.1.3 Konvergence Newtonovy metody

Nejprve, pro obecnou iterativní metodu $\mathcal{M}$, která generuje posloupnost $\{x_k\}_{k=1}^{\infty}$ aproximující přesné řešení x^*, definujeme asymptotický řád konvergence.

Definice 5. *Řekneme, že* ***asymptotický řád konvergence metody $\mathcal{M}$ je $n \in \mathbb{N}$,*** *existuje-li číslo $C \in \mathbb{R}$ takové, že*

$$\lim_{k\to\infty} \frac{|x_k - x^*|}{|x_{k+1} - x^*|^n} = C$$

Potom můžeme formulovat větu o konvergenci a jejím řádu pro Newtonovu metodu ve skalárním případě.

Věta 1 ([6], str. 123)**.** *Mějme funkci $f : [a,b] \subset \mathbb{R} \to \mathbb{R}$, $f \in C^2([a,b])$. Buď $x^* \in [a,b]$ její kořen, tj. $f(x^*) = 0$, a nechť $f'(x^*) \neq 0$. Uvažme dále metodu (3.2) aproximující přesné řešení x^* posloupností $\{x_k\}_{k=1}^{\infty}$. Pokud „x_0 je dostatečně blízko x^*“, pak:*

1. *tato metoda konverguje k přesnému řešení, tedy $\lim_{k\to\infty} |x_k - x^*| = 0$;*
2. *asymptotický řád konvergence této metody je dva.*

Poznamenejme, že vágní formulace „x_0 je dostatečně blízko x^*“ ve znění věty by byla opodstatněna důkazem - zmiňovanou „blízkost“ chápeme ve smyslu existence jistého okolí $U(x^*)$, na kterém máme díky dostatečné hladkosti funkce f k dispozici rovnost $|\frac{1}{2}\frac{f''(\xi_k)}{f'(x_k)}| = C$ pro jistou konstantu $C \in \mathbb{R}$. Náleží-li počáteční aproximace tomuto okolí, pak metoda konverguje.

3.2 Newtonova metoda v $\mathbb{R}^n$

V následujícím odstavci uvedeme zobecnění metody ze sekce (3.1) pro systémy rovnic. Hlavní myšlenka je přejata z knihy [5]. Mějme nyní $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$ a uvažme problém hledání přesného řešení $\boldsymbol{x}^* \in \mathbb{R}^n$ soustavy $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}$. Situace je oproti skalárnímu případu složitější. Uvažme opět k-tou aproximaci řešení $\boldsymbol{x}_k \in \mathbb{R}^n$ a položme $\delta\boldsymbol{x}_k := \boldsymbol{x}^* - \boldsymbol{x}_k$. Podobně jako ve skalárním případě pak můžeme za předpokladu dostatečné hladkosti použít na funkci $\boldsymbol{F}$ Taylorův rozvoj pro funkce více proměnných:

$$\boldsymbol{0} = \boldsymbol{F}(\boldsymbol{x}_k + \delta\boldsymbol{x}_k) = \boldsymbol{F}(\boldsymbol{x}_k) + \boldsymbol{F}'(\boldsymbol{x}_k)\delta\boldsymbol{x}_k + O(\|\delta\boldsymbol{x}_k\|^2),$$

kde výraz $\boldsymbol{F}'(\boldsymbol{x}_k)$ odpovídá Jacobiánu funkce $\boldsymbol{F}$ v bodě $\boldsymbol{x}_k$. Zanedbáním členů vyššího než lineárního řádu opět docházíme ke vztahu

$$\boldsymbol{0} \approx \boldsymbol{F}(\boldsymbol{x}_k) + \boldsymbol{F}'(\boldsymbol{x}_k)\delta\boldsymbol{x}_k,\ k = 0,1,\ldots.$$

Mějme k dispozici počáteční odhad řešení $\boldsymbol{x}_0 \in \mathbb{R}^n$. Pokud bychom chtěli novou aproximaci $\boldsymbol{x}_{k+1}$ určovat pomocí předpisu analogickému skalárnímu případu (3.2), museli bychom vypočítat inverzní matici $\boldsymbol{F}'(\boldsymbol{x}_k)^{-1}$, což je z numerického pohledu drahá a nežádoucí procedura. Proto zavádíme člen $\boldsymbol{d}_k \in \mathbb{R}^n$, pro který platí

$$\boldsymbol{F}'(\boldsymbol{x}_k)\boldsymbol{d}_k = -\boldsymbol{F}(\boldsymbol{x}_k),\ k = 0,1,\ldots,$$

což představuje pro pevné $\boldsymbol{x}_k$ soustavu lineárních algebraických rovnic. Můžeme tedy definovat Newtonovu metodu v $\mathbb{R}^n$.

Definice 6. *Mějme funkci* $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$, $\boldsymbol{F} \in C^1(\mathbb{R}^n)$ *a nechť* $\boldsymbol{x}_0 \in \mathbb{R}^n$ *je daná počáteční aproximace řešení.* ***Newtonovou metodou v*** $\mathbb{R}^n$ *rozumíme iterativní proces aproximující řešení problému* $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}$, *který získá novou aproximaci* $\boldsymbol{x}_{k+1} \in \mathbb{R}^n$ *ze známé předchozí aproximace* $\boldsymbol{x}_k \in \mathbb{R}^n$ *jako*

$$\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \boldsymbol{d}_{k+1}, \tag{3.3}$$

kde člen $\boldsymbol{d}_{k+1}$ *najde jako řešení lineární soustavy rovnic*

$$\boldsymbol{F}'(\boldsymbol{x}_k)\boldsymbol{d}_{k+1} = -\boldsymbol{F}(\boldsymbol{x}_k). \tag{3.4}$$

Tím jsme přešli od problému hledání řešení nelineární soustavy (3.1) k posloupnosti lineárních problémů (3.3), (3.4), které řešíme pro $\boldsymbol{d}_{k+1}$: dopočítáním tohoto členu získáváme novou aproximaci řešení $\boldsymbol{x}_{k+1}$. Podobně jako ve skalárním případu vycházíme z počátečního odhadu $\boldsymbol{x}_0$ pro který metoda za jistých předpokladů bude konvergovat (viz sekce 3.2.3).

Způsob řešení lineární soustavy (3.4) záleží na povaze řešeného probému. Rozlišujeme *přímé* a *nepřímé Newtonovy metody*. Zatímco přímé metody výpočet členu $\boldsymbol{d}_{k+1}$ uskutečňují eliminací, nepřímé zavádějí do každého kroku Newtonovy metody vnitřní iterační proces, který kýžený člen aproximuje. Jak jsme vysvětlili v sekci 2.3, v situaci, kdy řešené rovnice získáváme diskretizací parciálních diferenciálních rovnic, je vhodné používat metody nepřímé.

Metod pro uskutečnění vnitřního iteračního procesu je mnoho. V našich numerických experimentech budeme používat metodu GMRES s blokvým ILU(0) předpodmíněním (minimalizace residua; podrobné odvození metody viz [7, Kapitola 9.3.1]). Předpodmíněním myslíme následující: řešíme-li systém $Ax = b$, předpodmíněnou soustavou je soustava $PAx = Pb$, kde P je taková matice, pro kterou má matice PA lepší konvergenční vlastnosti než-li A. Lineární řešič zastavujeme, když je předpodmíněné residuum $P(Ax - b)$ stokrát menší než-li počáteční předpodmíněné residuum.

3.2.1 Zastavovací kritéria

Závěrečným krokem zavedení zobecněné Newtonovy metody je určení *zastavovacích kritérií*. Ta se snaží vyvážit požadavky na efektivitu a přesnost výpočtu. V literatuře se často používá tzv. „residuální přístup“, založený na sledování chyby $\|\mathbf{F}(\mathbf{x}_k)\|$. Pokud bude chyba dostatečně malá, tedy bude platit $\|\mathbf{F}(\mathbf{x}_k)\| \leq \mathrm{TOL}$, kde TOL je předem určená tolerance na algebraickou chybu, algoritmus ukončíme.

Tento přístup je problematický. Řešíme-li například soustavy algebraických rovnic vzniklé diskretizací parciálních diferenciálních rovnic, snažíme se primárně aproximovat řešení daných diferenciálních rovnic. V takovém případě je pro nás zbytečné získávat přesnější řešení algebraické soustavy, pokud nám nepřináší přesnější řešení původního problému. Proto někteří autoři definují tzv. „funkce odhadující chyby“, jejichž cílem je v jistém smyslu „vyvážit“ chyby vzniklé časovou diskretizací, prostorovou diskretizací a algebraické chyby vzniklé při aproximaci řešení algebraické soustavy. Tyto funkce zavedeme v kontextu diskretizace rovnice konvekce-difuze-reakce, která byla naznačena v prvních dvou kapitolách.

Abychom mohli tyto chyby definovat, uvažme tři typy řešení:

- $\boldsymbol{u} : Q_T \to \mathbb{R}^n$ buď **přesné řešení** úlohy (1.1),
- $\boldsymbol{u}_h \in S_{h,p}^{\tau,q}$ buď **přibližné řešení** dané (2.2),
- $\tilde{\boldsymbol{u}}_h \in S_{h,p}^{\tau,q}$ buď **spočtené přibližné řešení** problému (2.4), které (2.2) nesplňuje přesně, neboť jsme jej počítali metodami Newtonova typu.

Dále, zavádíme pro $\boldsymbol{u} \in H^1(\mathcal{I}_\tau,H^2(\mathcal{T}_{h,m}))$ normu

$$\| \boldsymbol{u} \|_X^2 = \| \boldsymbol{u} \|_{L^2(Q_T)}^2 + \| \nabla \boldsymbol{u} \|_{L^2(Q_T)}^2 + \| \partial_t \boldsymbol{u} \|_{L^2(Q_T)}^2 .$$

Pak můžeme definovat následující chyby.

Definice 7. *Mějme řešení $\boldsymbol{u}$, $\boldsymbol{u}_{h,\tau}$ a $\tilde{\boldsymbol{u}}_{h,\tau}$ zavedená výše. Uvažme prostory $H^1(\mathcal{I}_\tau,H^2(\mathcal{T}_{h,m}))$ a $S_{h,p}^{\tau,q}$ a mějme formu $A_{h,\tau}^m$ odvozenou z rovnice* (1.1) *pomocí nespojité Galerkinovy metody. Pak definujeme*

1. ***algebraickou chybu***

$$\epsilon_a(\tilde{\boldsymbol{u}}_{h,\tau}) := \sup_{\boldsymbol{\psi} \in S_{h,p}^{\tau,q}, \boldsymbol{\psi} \neq 0} \frac{A_{h,\tau}^m(\tilde{\boldsymbol{u}}_{h,\tau},\boldsymbol{\psi}) - A_{h,\tau}^m(\boldsymbol{u}_{h,\tau},\boldsymbol{\psi})}{\| \boldsymbol{\psi} \|_X};$$

2. ***chybu diskretizace***

$$\epsilon_s(\tilde{\boldsymbol{u}}_{h,\tau}) := \sup_{\boldsymbol{\psi} \in H^1(\mathcal{I}_\tau,H^2(\mathcal{T}_{h,m})), \boldsymbol{\psi} \neq 0} \frac{A_{h,\tau}^m(\tilde{\boldsymbol{u}}_{h,\tau},\boldsymbol{\psi}) - A_{h,\tau}^m(\boldsymbol{u},\boldsymbol{\psi})}{\| \boldsymbol{\psi} \|_X}.$$

Všimněme si, že při výpočtu máme k dispozici pouze řešení $\tilde{\boldsymbol{u}}_{h,\tau}$. Navíc, v případě chyby diskretizace ϵ_s uvažujeme supremum přes nekonečný prostor. Tedy, tyto chyby jsou v průběhu výpočtu nevypočitatelné. Abychom mohli spočítat alespoň jejich odhad, uvažme následující. Připomeňme, že forma $A_{h,\tau}^m$, kterou jsme zavedli v druhé kapitole, je konzistentní s původní rovnicí (1.1) (tj. platí vztah (2.1)). Konzistence pak zaručuje následující vztah pro chybu diskretizace:

$$\epsilon_s(\tilde{\boldsymbol{u}}_{h,\tau}) = \sup_{\boldsymbol{\psi} \in H^1(\mathcal{I}_\tau,H^2(\mathcal{T}_{h,m})), \boldsymbol{\psi} \neq 0} \frac{A_{h,\tau}^m(\tilde{\boldsymbol{u}}_{h,\tau},\boldsymbol{\psi})}{\| \boldsymbol{\psi} \|_X}.$$

Tím jsme vyjádřili chybu diskretizace pomocí výrazu, který máme při výpočtu k dispozici. Na druhou stranu, stále v něm uvažujeme supremum přes nekonečnědimenzionální prostor $H^1(\mathcal{I}_\tau,H^2(\mathcal{T}_{h,m}))$. Pro aproximaci chyby diskretizace použijeme supremum přes jistý, dostatečně velký ale konečnědimenzionální podprostor daného po částech Sobolevova prostoru, např. $S_{h,p+1}^{\tau,q+1}$.

Podobně, díky definici přibližného řešení $\boldsymbol{u}_{h,\tau}$ (tedy platnosti vztahu (2.2)) dostáváme vztah pro chybu aproximace

$$\epsilon_a(\tilde{\boldsymbol{u}}_{h,\tau}) = \sup_{\boldsymbol{\psi} \in S_{h,p}^{\tau,q}, \boldsymbol{\psi} \neq 0} \frac{A_{h,\tau}^m(\tilde{\boldsymbol{u}}_{h,\tau},\boldsymbol{\psi})}{\| \boldsymbol{\psi} \|_X}.$$

Pak můžeme zavést funkce, které budou poskytovat jejich spodní odhad.

Definice 8. *Mějme řešení* $\boldsymbol{u}$, $\boldsymbol{u}_{h,\tau}$ *a* $\tilde{\boldsymbol{u}}_{h,\tau}$ *zavedená výše a mějme formu* $A^m_{h,\tau}$ *odvozenou z rovnice* (1.1) *pomocí nespojité Galerkinovy metody. Pak definujeme*

1. ***odhad algebraické chyby***

$$\eta_a(\tilde{\boldsymbol{u}}_{h,\tau}) := \sup_{\boldsymbol{\psi}\in S^{\tau,q}_{h,p},\boldsymbol{\psi}\neq 0} \frac{A^m_{h,\tau}(\tilde{\boldsymbol{u}}_{h,\tau},\boldsymbol{\psi})}{\| \boldsymbol{\psi} \|_X};$$

2. ***odhad chyby diskretizace***

$$\eta_s(\tilde{\boldsymbol{u}}_{h,\tau}) := \sup_{\boldsymbol{\psi}\in S^{\tau,q+1}_{h,p+1},\boldsymbol{\psi}\neq 0} \frac{A^m_{h,\tau}(\tilde{\boldsymbol{u}}_{h,\tau},\boldsymbol{\psi})}{\| \boldsymbol{\psi} \|_X}.$$

Využíváme přitom již zmíněného faktu, že námi definované polynomiální prostory $S^{\tau,q}_{h,p}$ jsou konečné dimenze. Uvedených suprem se tedy nabývá, definice dává dobrý smysl a hlavně uvedené odhady lze v průběhu výpočtu dopočítat. Navíc si můžeme všimnout, že platí $\epsilon_d = \eta_d$.

Nyní můžeme zastavací kritéria formulovat následovně: výpočet Newtonovy metody ukončíme, pokud funkce odhadující algebraickou chybu poklesne dostatečně ve vztahu k funkci odhadující chybu diskretizace, tedy pokud bude platit $\eta_a < C\eta_s$, kde $C \in (0,1]$ je uživatelem předepsaná konstanta. Tím získáváme zastavovací kritéria pro metodu Newtonova typu, které mají vztah k původně řešenému problému.

3.2.2 Modifikace Newtonovy metody v $\mathbb{R}^n$

Výše jsme odvodili zobecněnou Newtonovu metodu ve tvaru (3.3),(3.4). V následující sekci se blíže zamyslíme nad její implementací.

Matice toku Při pohledu na vzorec (3.4) si můžeme povšimnout toho, že v každém kroku naší iterativní metody musíme počítat Jacobián $\mathbf{F}'(\mathbf{x}_k)$. To však nemusí být z několika důvodů výhodné. Předně, numerický výpočet Jacobiánu bude drahý a může být problém ho numericky spočítat efektivně a přesně. Dále, z geometrické interpretace Newtonovy metody (ilustrovaná v sekci 3.1.2.) je vidět, že Jacobián nám udává „směr“ v prostoru $\mathbb{R}^n$, ve kterém hledáme další iteraci aproximace řešení: ve skalárním případě přechází Jacobián v derivaci skalární funkce, tedy určuje sklon její tečny; ve více dimenzích Jacobián určuje sklony tečných nadrovin k funkci. Můžeme se ptát, jak moc se sklon této tečné nadroviny mění při přechodu mezi jednotlivými kroky aproximace, a zdali je potřeba tento sklon aktualizovat v každém kroku výpočtu. Menší přesnost sklonu by pak mohla být v procesu vyvážena tím, že si ušetříme drahý výpočet Jacobiánu.

Tyto úvahy nás vedou k definici tzv. *matice toku* $\mathbb{C} \in \mathbb{R}^{n\times n}$, která bude „vhodnou aproximací“ Jacobiánu, tj. bude platit $\mathbb{C}(\mathbf{x}_k) \approx \mathbf{F}'(\mathbf{x}_k)$, a zároveň bude její výpočet co nejlevnější. Jak jsme ilustrovali v sekci (2.3), řešíme-li nelineární systémy vzniklé diskretizací parciálních diferenciálních rovnic pomocí nespojité Galerkinovy metody, taková matice se dá v rámci diskretizace skutečně zavést.

Tlumicí koeficient Dalším problémem Newtonovy metody je nutnost mít počáteční aproximaci $\boldsymbol{x}_0$ dostatečně blízko řešení $\boldsymbol{x}^*$. Pokud takovou aproximaci k dispozici nemáme, můžeme definovat

$$\mathbf{x}_{k+1} := \mathbf{x}_k + \lambda_k \mathbf{d}_{k+1},$$

kde $\lambda_k \in (0,1]$ je *tlumicí koeficient* („damping coefficient"), který umožňuje konvergenci algoritmu i pro špatný počáteční odhad $\mathbf{x}_0$. Zmenšením parametru λ_k docílíme toho, že budeme počítat další krok algoritmu v bodu $\mathbf{x}_{k+1}$, který bude o něco blíže bodu z předcházející iterace. Vzhledem k tomu, že pracujeme místo s Jacobiánem funkce $\mathbf{F}'(\mathbf{x}_k)$ s její aproximací $\mathbb{C}(\mathbf{x}_k)$, kterou navíc neaktualizujeme v každém kroku výpočtu, je z geometrického náhledu přirozené předpokládat, že pokud nás aktuální volba $\mathbf{x}_k$ neposouvá k výsledku, mohla by „tlumená" volba vykazovat lepší vlastnosti.

Tedy, odvodili jsme metodu Newtonova typu, kterou budeme označovat jako *modifikovanou Newtonovu metodu*. Odvozené shrneme v následující definici.

Definice 9. *Mějme danou funkci $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$ a počáteční odhad $\boldsymbol{x}_0 \in \mathbb{R}^n$ přesného řešení $\boldsymbol{x}^* \in \mathbb{R}^n$, pro které platí $\boldsymbol{F}(\boldsymbol{x}^*) = \mathbf{0}$. **Modifikovanou Newtonovou metodou** rozumíme iterativní algoritmus, který generuje posloupnost bodů $\boldsymbol{x}_k \in \mathbb{R}^n$ aproximující přesné řešení následujícím způsobem: je-li $\mathbb{C}(\boldsymbol{x}_k)$ matice aproximující Jacobián $\boldsymbol{F}'(\boldsymbol{x}_k)$, $\lambda_k \in (0,1]$ tlumicí koeficient a $\boldsymbol{d}_{k+1} \in \mathbb{R}^n$, pak novou aproximaci $\boldsymbol{x}_{k+1}$ získáme z předchozí jako*

$$\boldsymbol{x}_{k+1} := \boldsymbol{x}_k + \lambda_k \boldsymbol{d}_{k+1}, \tag{3.5}$$

kde $\boldsymbol{d}_{k+1}$ najdeme jako řešení soustavy rovnic

$$\mathbb{C}(\boldsymbol{x}_k)\boldsymbol{d}_{k+1} = -\boldsymbol{F}(\boldsymbol{x}_k). \tag{3.6}$$

Zrekapitulujme na závěr průběh výpočtu modifikované Newtonovy metody:

- řešíme soustavu $\mathbb{C}(\boldsymbol{x}_k)\boldsymbol{d}_{k+1} = -\boldsymbol{F}(\boldsymbol{x}_k)$ metodou GMRES s blokovým ILU(0) předpodmíněním. Jako počáteční podmínka pro ni slouží buď počáteční odhad $\boldsymbol{x}_0$, nebo řešení z předchozího kroku Newtonovy iterace $\boldsymbol{x}_k$.

- Zavádíme *monitorovací funkci* $\theta_k := \frac{\|\mathbf{F}(\mathbf{x}_{k+1})\|}{\|\mathbf{F}(\mathbf{x}_k)\|}$, která udává, zdali nás nová aproximace přibližuje k řešení problému. Inicializujeme $\lambda_k = 1$. Z rovnosti $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \lambda_k \boldsymbol{d}_{k+1}$ dopočítáme novou aproximaci řešení $\boldsymbol{x}_{k+1}$. Je-li $\theta_k < 1$, pokračujeme k další Newtonově iteraci. V opačném případě nastavíme $\lambda_k := \frac{\lambda_k}{2}$ a vracíme se k předchozímu kroku.

- Dosáhne-li damping parametr λ_k minimální předpsané hodnoty, aktualizujeme matici toku $\mathbb{C}(\boldsymbol{x}_k)$. Tedy, obecně neaktulizujeme tuto matici v každém kroku algoritmu - ukazuje se, že je výhodnější provést více Newtonových aproximací na jedné časové hladině než v každém kroku provádět výpočet matice toku.

- Výpočet ukončujeme vhodnými zastavovacími kritérii, které závisí na kontextu řešeného problému, viz sekce 3.2.1.

3.2.3 Konvergence zobecněné Newtonovy metody

V literatuře existují různá tvrzení popisující vlastnosti konvergence modifikované Newtonovy metody. Pro ilustraci jednu z těchto vět zformulujeme. Povšimněme si, že ve větě budeme předpokládat znalost horních odhadů na jisté veličiny - tyto horní odhady jsou ve výpočetní praxi obtížně získatelné, neboť jejich výpočet zpravidla bývá drahý (viz [5], str. 11). I proto je propojení teoretické analýzy a praktického ověření konvergence metod Newtonova typu problematické.

Věta 2 ([5], str. 56)**.** *Buď $\boldsymbol{F} : D \to \mathbb{R}^n$, $\boldsymbol{F} \in C^1(D)$ hladká funkce, kde $D \subset \mathbb{R}^n$ je otevřená, konvexní množina. Mějme k dispozici počáteční odhad $\boldsymbol{x}_0 \in D$, $\mathbb{C}$ aproximaci Jacobiánu funkce $\boldsymbol{F}$, tj. nechť platí $\mathbb{C} \approx \boldsymbol{F}'$, a konstanty $\alpha, \omega_0, \delta_0, \delta_1, \delta_2 \geq 0$ takové, že $\forall \boldsymbol{x},\boldsymbol{y} \in D$ platí následující podmínky:*

1. $\| \mathbb{C}(\boldsymbol{x}_0)^{-1}\boldsymbol{F}(\boldsymbol{x}_0) \| \leq \alpha$;
2. $\| \mathbb{C}(\boldsymbol{x}_0)^{-1}(\boldsymbol{F}'(\boldsymbol{y}) - \boldsymbol{F}'(\boldsymbol{x})) \| \leq \omega_0 \| \boldsymbol{y} - \boldsymbol{x} \|$;
3. $\| \mathbb{C}(\boldsymbol{x}_0)^{-1}(\boldsymbol{F}'(\boldsymbol{x}) - \mathbb{C}(\boldsymbol{x})) \| \leq \delta_0 + \delta_1 \| \boldsymbol{x} - \boldsymbol{x}_0 \|$;
4. $\| \mathbb{C}(\boldsymbol{x}_0)^{-1}(\mathbb{C}(\boldsymbol{x}) - \mathbb{C}(\boldsymbol{x}_0)) \| \leq \delta_2 \| \boldsymbol{x} - \boldsymbol{x}_0 \|$,

kde $\delta_0 < 1$, $\sigma := max(\omega_0,\delta_1 + \delta_2)$,$h := \frac{2\alpha\sigma}{(1-\delta_0)^2} \leq 1$. Nechť konečně

$$\bar{S}(\boldsymbol{x}_0,\rho) \subset D, \text{ kde } \rho := \frac{2\alpha}{1-\delta_0}/(1+\sqrt{1-h}).$$

Pak posloupnost $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ generovaná modifikovanou Newtonovou metodou (9) je dobře definovaná, zůstává v $\bar{S}(\boldsymbol{x}_0,\rho)$ a konverguje k přesnému řešení $\boldsymbol{x}^*$, pro nějž $\boldsymbol{F}(\boldsymbol{x}^*) = \boldsymbol{0}$. Položíme-li

$$\bar{h} := \frac{\omega_0}{\sigma}h, \ \rho_{\pm} := \frac{2\alpha}{1-\delta_0}/(1 \mp \sqrt{1-\bar{h}}),$$

je přesné řešení $\boldsymbol{x}^* \in \bar{S}(\boldsymbol{x}_0,\rho_-)$ určeno jednoznačně v $\bar{S}(\boldsymbol{x}_0,\rho) \cup (D \cap S(\boldsymbol{x}_0,\rho_+))$.

3.3 Metody pevného bodu a Andersonova akcelerace

Vraťme se k odvození Newtonovy metody ve skalárním případě (viz (3.1)). Můžeme si všimnout, že problém

$$x_{k+1} := x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0,1,\ldots,$$

řešený se znalostí počátečního odhadu $x_0 \in \mathbb{R}$, lze přeformulovat na metodu hledání pevného bodu funkce $g : \mathbb{R} \to \mathbb{R}$ dané jako

$$\begin{aligned} g(x) &:= x - \frac{f(x)}{f'(x)}, \\ x_{k+1} &:= g(x_k), \quad k = 0,1,\ldots. \end{aligned}$$

Tím získáváme možnost řešit náš problém dvěmi rozdílnými metodami, které mohou mít pro různé typy problémů různé numerické vlastnosti, a tedy jejich aplikace může být různě výhodná. Podobně můžeme uvažovat metodu hledání pevného bodu pro nelineární systémy vzešlé z diskretizace parciálních diferenciálních rovnic, jak si naznačíme v následujícím (pro ilustraci více k přechodu od např. Richardsovy rovnice - popisující proudění v porézním prostředí - k problémům hledání pevného bodu např. v článku [3]).

3.3.1 Picardova metoda

V následující sekci se krátce zastavíme u jiného způsobu řešení systémů nelineárních rovnic. Mějme funkci $\boldsymbol{G} : \mathbb{R}^n \to \mathbb{R}^n$ a uvažme problém pevného bodu

$$\boldsymbol{x} = \boldsymbol{G}(\boldsymbol{x}). \tag{3.7}$$

Máme-li k dispozici počáteční odhad $\boldsymbol{x}_0 \in \mathbb{R}^n$, můžeme definovat iterativní schéma pro řešení tohoto problému jako

$$\boldsymbol{x}_{k+1} := \boldsymbol{G}(\boldsymbol{x}_k), \quad k = 0,1,\ldots. \tag{3.8}$$

Schéma tohoto typu označujeme jako *Picardovo*.

Banachova věta o pevném bodu nám říká, že pokud je funkce $\boldsymbol{G}$ kontrakce, schéma bude konvergovat k přesnému řešení. Nejprve definujme pojem *kontrakce*.

Definice 10. *Řekneme, že funkce $\boldsymbol{G} : M \subset \mathbb{R}^n \to \mathbb{R}^n$ je kontrakce, pokud*

$$\exists\, 0 < L < 1 \,\forall\, \boldsymbol{x},\boldsymbol{y} \in M : \| \boldsymbol{G}(x) - \boldsymbol{G}(y) \| < L \| \boldsymbol{x} - \boldsymbol{y} \| .$$

Potom můžeme formulovat Banachovu větu o pevném bodu v $\mathbb{R}^n$.

Věta 3 ([6], str. 139)**.** *Buď $\boldsymbol{G} : M \subset \mathbb{R}^n \to \mathbb{R}^n$ kontrakce. Potom má tato funkce právě jeden pevný bod $\boldsymbol{x}^*$, tedy*

$$\exists!\, \boldsymbol{x}^* \in M : \boldsymbol{G}(\boldsymbol{x}^*) = \boldsymbol{x}^*.$$

Navíc, Picardova metoda (3.8) *k tomuto pevnému bodu konverguje, tj. pro jí generovanou posloupnost $\boldsymbol{x}_k$ platí*

$$\lim_{k\to\infty} \| \boldsymbol{x}_k - \boldsymbol{x}^* \| = 0.$$

V praktických problémech, jako například v těch vzniklých diskretizací parciálních diferenciálních rovnic, ovšem funkce $\boldsymbol{G}$ kontrakcí obecně není. Podobně jako v případě Newtonovy metody je teoretická analýza konvergence Picardovy metody složitá a její aplikace v praxi je problematická. Jak uvádí např. [3], numerické experimenty ukazují, že hlavními úskalími Picardovy metody jsou v porovnání s metodami Newtonova typu menší robustnost a pomalejší konvergence. Existují ovšem postupy, kterými lze tyto nedostatky redukovat. Jedním z nich je *Andersonova akcelerace*.

3.3.2 Andersonova akcelerace

V následující sekci čerpáme z článku [8]. Definujme dále *residuum* $\boldsymbol{g}_k := \boldsymbol{x}_k - \boldsymbol{G}(\boldsymbol{x}_k)$. Cílem Picardovy metody je docílit toho, aby norma $\| \boldsymbol{g}_k \|$ byla „dostatečně malá“. Můžeme si všimnout, že k definici nové aproximace řešení se používá pouze aproximace z předcházející iterace. Myšlenkou Andersonovy akcelerace je definovat novou iteraci $\boldsymbol{x}_k$ jako nejmenší afinní kombinaci m_k předcházejících aproximací $\boldsymbol{x}_{k-1}, \ldots, \boldsymbol{x}_{k-m_k-1}$. Tedy, k-tou Picardovu iteraci (3.8) nahradíme problémem najít $\alpha = (\alpha_0, \ldots, \alpha_{m_k})$ tak, že

$$\min_{\alpha, \sum_{i=0}^{m_k} \alpha_i = 1} \| \sum_{i=0}^{m_k} \alpha_i \boldsymbol{g}_{k-i} \|_{\ell^2}, \tag{3.9}$$

tedy problémem nejmenších čtverců. Pro $\mathbb{R}^2$ odpovídá minimalizace residua přes afinní kombinace hledání takového bodu v m_k-úhelníku, jehož vrcholy jsou předcházející aproximace řešení, pro který bude residuum nejmenší. Novou aproximaci $\boldsymbol{x}_{k+1}$ pak definujeme jako

$$\boldsymbol{x}_{k+1} := \sum_{i=0}^{m_k} \alpha_i \boldsymbol{G}(\boldsymbol{x}_{k-i}). \tag{3.10}$$

Položíme-li $\gamma_1, \ldots, \gamma_{m_k}$ taková, že platí vztahy

$$\begin{aligned} \alpha_0 &= 1 - \gamma_1 \\ \alpha_1 &= \gamma_1 - \gamma_2 \\ &\vdots \\ \alpha_{m_k} &= \gamma_{m_k}, \end{aligned}$$

pak platí také

$$\begin{aligned} \sum_{i=0}^{m_k} \alpha_i g_{k-1} &= (1-\gamma_1) g_k + (\gamma_1 - \gamma_2) g_{k-1} + \ldots + \gamma_{m_k} g_{k-m_k} = \\ &= g_k + \gamma_1 (g_{k-1} - g_k) + \gamma_2 (g_{k-2} - g_{k-1}) + \ldots + \gamma_{m_k} (g_{k-m_k} - g_{k-m_k+1}) = \\ &= g_k + \sum_{i=1}^{m_k} \gamma_i (g_{k-i} - g_{k-i+1}). \end{aligned} \tag{3.11}$$

Tedy, je možné místo minimalizace výrazu (3.9) řešit tentýž problém pro výraz (3.11), přičemž budeme pokládat

$$\boldsymbol{x}_{k+1} = \boldsymbol{G}(\boldsymbol{x}_k) + \sum_{i=1}^{m_k} \gamma_i (\boldsymbol{G}(\boldsymbol{u}_{k-i}) - \boldsymbol{G}(\boldsymbol{u}_{k-i+1})).$$

V našem případě řešení algebraických soustav rovnic vzniklých diskretizací rovnice konvekce-difuze-reakce můžeme pak uvedené aplikovat následujícím způsobem. Newtonova metoda řeší problém (3.6). Ekvivalentně můžeme řešit problém pevného bodu pro novou funkci $\boldsymbol{G} : \mathbb{R}^n \to \mathbb{R}^n$

$$\boldsymbol{G}(\boldsymbol{x}_k) := -\boldsymbol{F}(\boldsymbol{x}_k)\mathbb{C}^{-1}(\boldsymbol{x}_k) + \boldsymbol{x}_k = \boldsymbol{x}_{k+1}. \tag{3.12}$$

Tedy, na řešení nového problému používáme Picardovu metodu, na kterou následně můžeme aplikovat Andersonovu akceleraci: spočítáme vhodnou váhu α předcházejících aproximací (3.9) a novou aproximaci řešení položíme jako (3.10).

Problémem je, jak vyčíslit funkci $\boldsymbol{G}$ - všimněme si, že v jejím předpisu se vyskytuje inverz matice toku. Ten však v praxi počítat nechceme. Proto položíme

$$\boldsymbol{G}(\boldsymbol{x}_k) := \boldsymbol{x}_{k+1}^{\text{Newton}},$$

kde $\boldsymbol{x}_{k+1}^{\text{Newton}}$ je $(k+1)$. aproximace získána modifikovanou Newtonovou metodou (9).

Tímto získáváme další modifikaci Newtonovy metody.

Definice 11. *Mějme danou funkci $\boldsymbol{F} : \mathbb{R}^n \to \mathbb{R}^n$ a počáteční odhad $\boldsymbol{x}_0 \in \mathbb{R}^n$ přesného řešení $\boldsymbol{x}^* \in \mathbb{R}^n$, pro které platí $\boldsymbol{F}(\boldsymbol{x}^*) = \boldsymbol{0}$. Buď $\mathbb{C}(\boldsymbol{x}_k)$ matice aproximující Jacobián $\boldsymbol{F}'(\boldsymbol{x}_k)$, $\lambda_k \in (0,1]$ tlumicí koeficient a $\boldsymbol{d}_{k+1} \in \mathbb{R}^n$. Označme $\{\boldsymbol{x}_k^{Newton}\}_{k=1}^{\infty}$ posloupnost bodů generovanou modifikovanou Newtonovou metodou (9). Položme $\boldsymbol{G}(\boldsymbol{x}_k) := \boldsymbol{x}_{k+1}^{Newton}$. Konečně, buď $\alpha = (\alpha_0, \ldots, \alpha_{m_k})$, $m_k < k$. Pak **modifikovanou Newtonovou metodou s Andersonovou akcelerací** rozumíme metodu generující novou aproximaci $\boldsymbol{x}_k$ přesného řešení $\boldsymbol{x}^*$ předpisem (3.10), kde váhu α nalezneme minimalizací residua $\boldsymbol{g}_k := \boldsymbol{x}_k - \boldsymbol{G}(\boldsymbol{x}_k)$, tj. vyřešením problému (3.9).*

4. Numerické experimenty

V numerických experimentech řešíme následující problém: je dána rovnice (1.2) kterou řešíme softwarem Adgfem v prostoru na jednotkovém čtverci, tj. $\Omega = (0,1) \times (0,1)$, a do času $T = 0.5$, tj. výpočetní oblasti $Q_T = \Omega \times (0, 0.5)$. Metodami popsanými v prvních dvou kapitolách budeme tuto rovnici na dané časové hladině diskretizovat, čímž dojdeme k systému nelineárních algebraických rovnic, které budeme řešit metodami Newtonova typu. Výpočet Newtonovy metody řídíme monitorovací funkcí θ_k, podle které rozhodujeme, jestli přijímáme novou aproximaci řešení systému $\boldsymbol{x}_{k+1}$ - tu získáváme metodou GMRES. Ukončujeme ho, platí-li mezi algebraickou chybou η_a a chybou diskretizace η_s vztah $\eta_a < 0.001\,\eta_s$, popřípadě dosáhne-li metoda 30 iterací. Následně se posouváme k diskretizaci rovnice na další časové hladině a postup se opakuje; celkem výpočet provedeme pro pět časových kroků.

V našich experimentech nastavujeme několik parametrů. Předně, máme možnost spustit modifikovanou Newtonovu metodu (9), nebo modifikovanou Newtonovu metodu s Andersonovou akcelerací (11). Dále máme tři parametry ovlivňující velikost řešeného problému: nastavujeme stupeň polynomiální aproximace v čase ($q = 0,1,2$), prostoru ($p = 0,1,2,3$) a triangulaci, kterou pokrýváme oblast Ω (volíme mezi triangulací obsahující 250 a 1000 prvků, označujeme „grid 250“, respektive „grid 1000“). Posledním nastavovaným parametrem je délka časového kroku τ - dělení časového intervalu bude v našem případě rovnoměrné. Pro příliš velký časový krok budeme ztrácet přesnost řešení, pro příliš malý časový krok zase efektivitu výpočtu. Konkrétně budeme uvažovat kroky $\tau \in \{0.1, 0.02, 0.001\}$.

4.1 Pozorování 1 - Závislost na zvolené metodě

Nejprve porovnejme efektivitu výpočtu v závislosti na tom, jestli použijeme modifikovanou Newtonovu metodu nebo modifikovanou Newtonovu metodu s An-

τ	**Grid**	**Metoda**	**0,1**	**1,1**	**2,1**	**0,3**	**1,2**	**2,2**	**1,3**	**2,3**
0.1	250	Newton	4.87	4.85	7.50	17.14	20.4	34.93	30.96	41.77
		Anderson	3.96	6.77	10.14	45.36	22.59	33.14	69.77	99.41
	1000	Newton	71.4	50.6	61.57	168.22	117.32	556.94	261.93	310.36
		Anderson	55.92	71.65	67.96	182.26	199.06	694.85	268.83	416.61
0.02	250	Newton	4.44	3.86	5.59	11.67	6.19	15.6	25.33	36.41
		Anderson	3.84	4.88	7.35	18.6	8.48	15.49	19.24	46.52
	1000	Newton	25.34	26.05	65.84	86.77	184.04	764.56	116.5	348.62
		Anderson	23.48	34.13	42.21	130.42	102.91	482.3	182.38	334.29
0.005	250	Newton	2.10	3.21	4.48	9.20	8.13	22.94	18.04	22.31
		Anderson	2.75	4.41	5.64	12.64	9.34	15.30	19.01	26.82
	1000	Newton	15.56	20.94	63.0	69.33	79.56	137.32	205.5	232.73
		Anderson	19.56	27.90	55.64	99.72	77.04	81.22	160.10	235.84

Tabulka 4.1: Doba výpočtu [s] v závislosti na velikosti časového kroku τ, síti, metodě a stupni polynomů q,p po řadě v čase, prostoru.

dersonovou akcelerací. Ukazuje se, že efektivita toho kterého algoritmu závisí zejména na velikosti řešeného problému a volbě velikosti časového kroku τ.

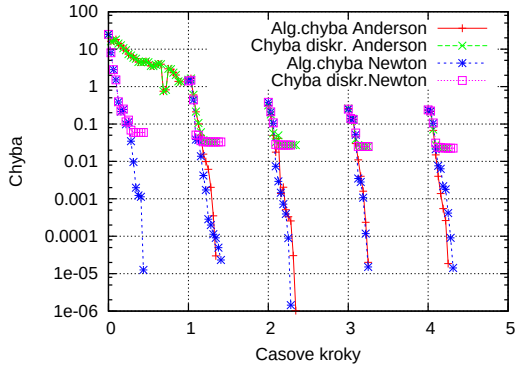
Délky výpočtů v závislosti na velikosti problému a zvolené metodě jsou uvedeny tabulce 4.1. Velikosti matic a počty nenulových prvků v nich jsou k nahlédnutí v tabulce 4.2.

Z dat vidíme, že diskretizační chyby se po pěti krocích pro obě metody srovnávají. Lišit se však může vývoj chyb v předcházejících krocích. Jak ilustruje obrázek 4.1a, problém metody s Andersonovou akcelerací je několikráte špatná konvergence v prvním kroku výpočtu - můžeme se domnívat, že pro špatný počáteční odhad zanášíme akcelerací do výpočtu informaci z několika nepřesných řešení, a tím zpomalujeme konvergenci. Můžeme si také všimnout, že výpočetní časy s nejdelším časovým krokem $\tau = 0.1$ jsou lepší u Newtonovy metody bez akcelerace. Tomu odpovídá vývoj zavedených chyb, viz obrázek 4.1b. V situaci $q = 0$, $p = 1$, $\tau = 0.02$ na gridu 1000 pozorujeme o něco pomalejší konvergenci neakcelerované metody, viz 4.1c - to se odráží i v porovnání příslušných výpočetních časů, kdy je neakcelerovaná metoda o něco pomalejší. Na největším uvažovaném problému s nejjemnější sítí, tedy při $q = 2$, $p = 3$, $\tau = 0.005$ na gridu 1000 pozorujeme (viz obrázek 4.1d), že neakcelerovaná metoda nezkonverguje ve třetím kroku, zatímco akcelerovaná v kroku prvním. Výsledné časy výpočtu se takřka neliší. Konečně, na hrubším prostorovém dělení o 250 prvcích si můžeme všimnout podobných dob výpočtu, s vyjímkou situace, kdy aplikujeme akcelerovanou metodu s krokem $\tau = 0.1$, nízkým stupněm polynomů v čase a vysokým v prostoru, tj. $q \in \{0,1\}$, $p = 3$. Obecně se zdá být rychlejší při problémech s delším časovým krokem využít neakcelerované metody.

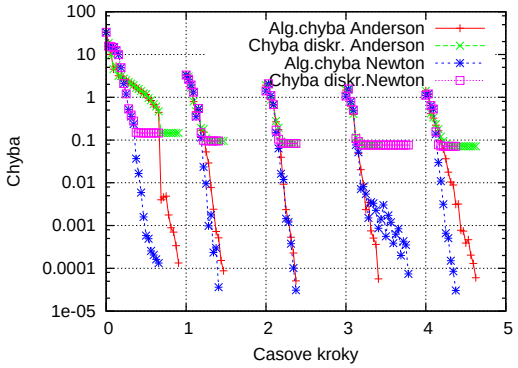
4.2 Pozorování 2 - Závislost na velikosti řešených matic

Zásadní vliv na rychlost konvergence a velikost odvozených chyb má velikost řešených matic. Je rozumné očekávat, že čím větší problém řešíme, tím delší bude výpočetní doba a tím lepší aproximaci řešení dostaneme, jak ilustrujeme v obrázku 4.2a, 4.2b, 4.2c (po řade problémy s počtem nenulových prvků 9 936, 162 432,4 060 800). V části 2.2.2 jsme uvedli vzorec pro řád řešených matic v závislosti na q, p a zvolené triangulaci. Náš řešicí software nám poskytuje údaj o tom, kolik nenulových prvků řešené matice mají. Závislost počtu nenulových prvků a velikosti matice v závislosti na stupni aproximace a triangulaci uvádíme v tabulce 4.2.

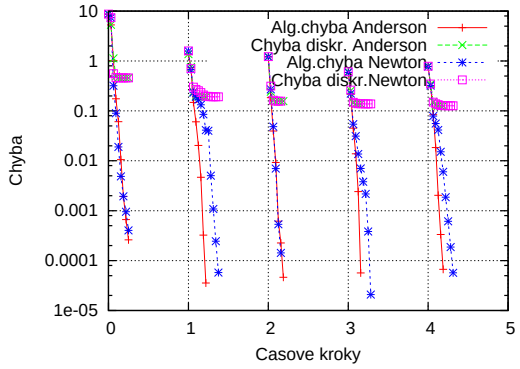
Přesto nastávají situace, kdy větší systém zkonverguje rychleji než systém menší. Jako příklad uvedeme situaci, kdy řešíme problém metodou s Andersonovou akcelerací na triangulaci o 1000 prvcích s časovým krokem $\tau = 0.005$. Položíme-li $p = 1$, $q = 2$, dostáváme systém o 649 728 nenulových prvcích. Pro $q = 0$, $p = 3$ dostáváme systém o 451 200 nenulových prvcích. Přesto, výpočetní čas v prvním případě je $77.04s$ a v druhém $99.72s$. V grafu 4.2d vidíme vývoj chyb při daném nastavení. Nabízí se myšlenka, že je výhodnější volit podobné stupně polynomiální aproximace v prostoru a čase.



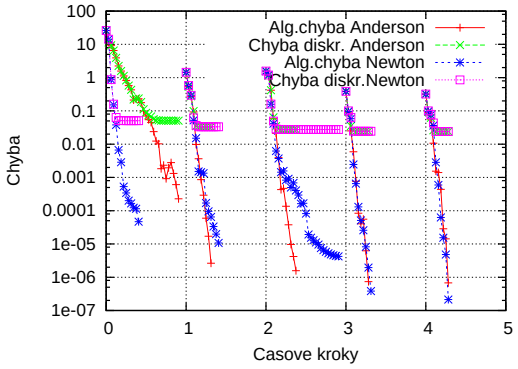
(a) $q = 0$, $p = 3$, $\tau = 0.005$, grid 1000



(b) $q = 2$, $p = 2$, $\tau = 0.1$, grid 1000



(c) $q = 0$, $p = 1$, $\tau = 0.02$, grid 1000

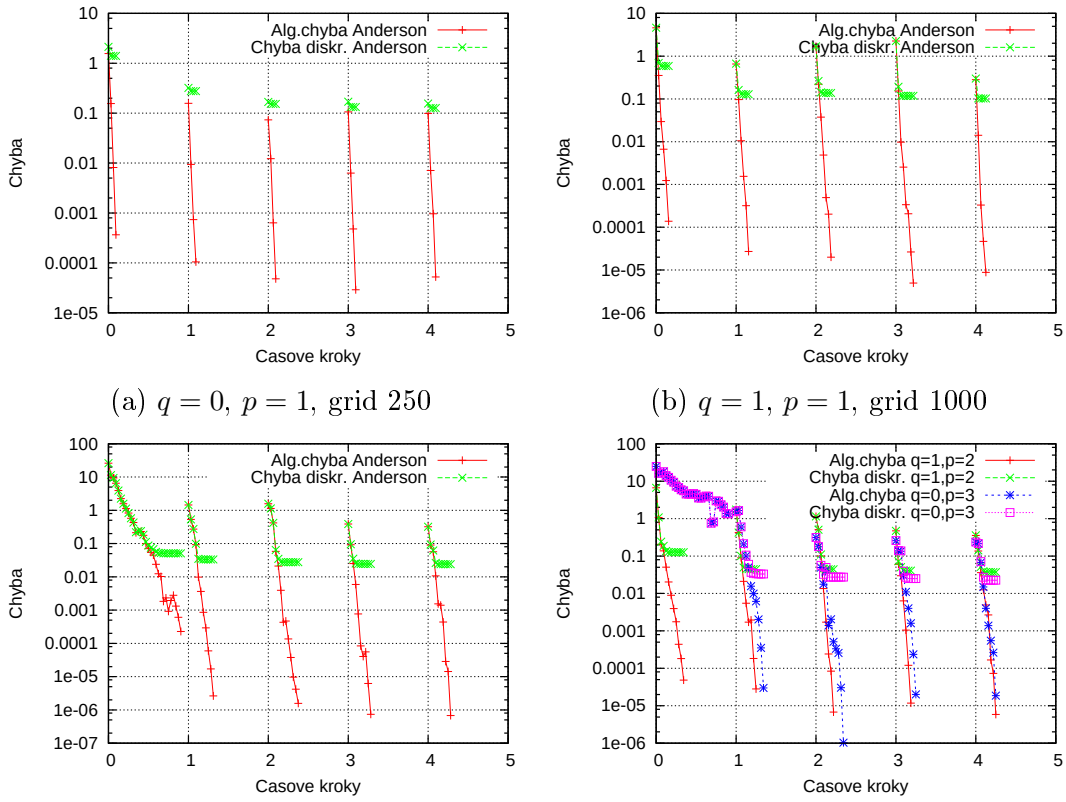


(d) $q = 2$, $p = 3$, $\tau = 0.005$, grid 1000

Obrázek 4.1: Porovnání algebraické, diskretizační chyby.

q	**p**	**grid**	**# nenulových prvků**	**# všech prvků**
0	1	250	9 936	2 250 000
0	3	250	110 400	25 000 000
1	1	250	39 744	9 000 000
1	2	250	158 967	36 000 000
1	3	250	441 600	100 000 000
2	1	250	9 936	20 250 000
2	2	250	357 696	81 000 000
2	3	250	993 600	225 000 000
0	1	1000	40 608	36 000 000
0	3	1000	451 200	400 000 000
1	1	1000	162 432	144 000 000
1	2	1000	649 728	576 000 000
1	3	1000	1 804 800	1 600 000 000
2	1	1000	365 472	324 000 000
2	2	1000	1 461 888	1 296 000 000
2	3	1000	4 060 800	3 600 000 000

Tabulka 4.2: Počet nenulových prvků matic v závislosti na triangulaci, stupni polynomů



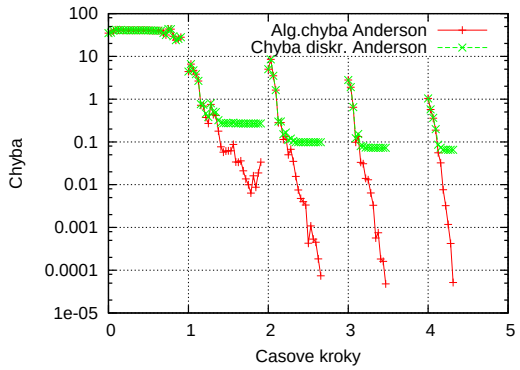
(a) $q = 0$, $p = 1$, grid 250

(b) $q = 1$, $p = 1$, grid 1000

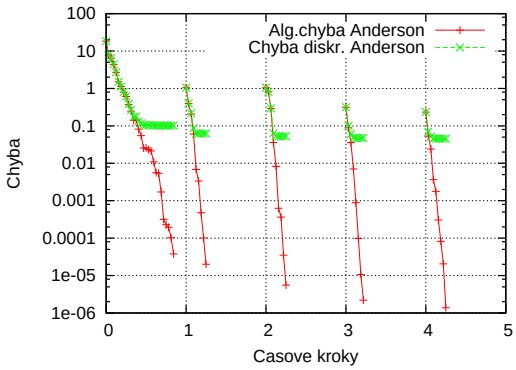
(c) $q = 2$, $p = 3$, grid 1000

(d) $q \in \{0,1\}$, $p \in \{2,3\}$, grid 1000

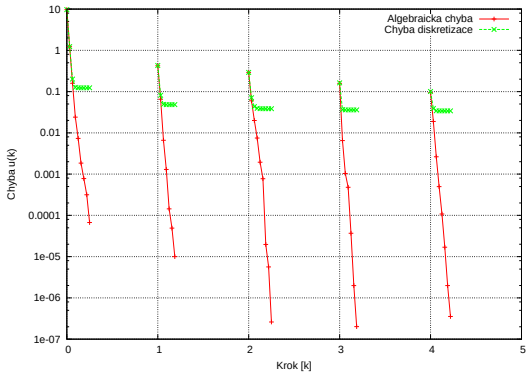
Obrázek 4.2: Porovnání algebraické a diskretizační chyby pro metodu s Andersonovou akcelerací, $\tau = 0.005$, rozdílnou triangulací a rozdílným stupněm polynomiální aproximace q v čase a p v prostoru.



(a) $\tau = 0.1$



(b) $\tau = 0.02$



(c) $\tau = 0.005$

Obrázek 4.3: Porovnání algebraické, diskretizační chyby pro Andersonovou metodu s $p = 2$, $q = 3$, grid 250, τ různá.

4.3 Pozorování 3 - Závislost na délce časového kroku τ

Neuvažejeme-li případy, kdy stupeň polynomiální aproximace v prostoru je výrazně jiný než v čase, můžeme si všimnout, že přesnost aproximace i čas výpočtu klesají s kratším časovým krokem τ. To se zdá být logické - pro kratší časový krok dostáváme z Galerkinovy diskretizace přesnější formu $A^m_{h,\tau}$ pročež odvozený algebraický problém konverguje rychleji a navíc dostáváme nižší diskretizační chybu. Pro metodu s Andersonovou akcelerací toto chování ilustrujeme na obrázku 4.3.

4.4 Pozorování 4 - Vhodnost zvolených zastavovacích ktitérií

Cílem našeho výpočtu je primárně spočítat co nejpřesnější aproximaci řešení původní rovnice (1.1) v co nejkratším čase. K vyvážení těchto dvou požadavků jsme zavedli zastavovací kritérium jako podmínku na poměr diskretizační a algebraické chyby. Naším cílem je, aby se diskretizační chyba ve všech krocích výpočtu ustálila, přičemž bychom chtěli, aby byl výpočet co nejdříve po ustálení ukončen.

Z uvedených grafů můžeme odečíst, že naše podmínka $\eta_a < 0.001\eta_s$ funguje

dobře - pakliže v některém z kroků Newtonovy metody je diskretizační chyba ustálena v mnoha krocích (viz např. třetí krok Newtonovy metody v 4.1d), mohlo by oslabení požadavku v některém z dalších kroku (viz pátý krok tamtéž) zapříčinit to, že by se chyba ustálit nestihla.

Závěr

Tato práce prezentuje v teoretické části základní myšlenky volby zastavovacích kritérií pro metody Newtonova typu aplikované na nelineární algebraické problémy vzešlé z diskretizace parciálních diferenciálních rovnic konvekce-difuze-reakce. Předně, odvodili jsme zobecněnou Newtonovu metodu (9) a její v praxi užívané modifikace, jmenovitě zavedení tlumicího koeficientu, matice toku a Andersonovy akcelerace. Zavedli jsme pojmy diskretizační a algebraická chyba a nastínili zdůvodnění jejich definic. Opodstatnili jsme zavedení zastavovacích kritérií, která by se měla snažit vyvážit chybu vzniklou diskretizací parciálních diferenciálních rovnic a algebraickou chybu vzniklou řešením nelineárních algebraických soustav iterativní zobecněnou Newtonovou metodou.

Numerickými experimenty jsme pak zkoumali chování odvozených zobecněných Newtonových metod pro různé nelineární algebraické problémy. V první řadě jsme ověřili, že naše volba zastavovacích kritérií coby tisícinásobného poklesu algebraické oproti diskretizační chybě se jeví jako vhodná. Dále jsme sledovali a ilustrovali základní charakteristiky výpočtů Newtonovy metody, tedy délku výpočtu a dosaženou chybu diskretizace, v závislosti na parametrech diskretizace původní parciální diferenciální rovnice.

Seznam použité literatury

[1] V. Dolejší. Private communication. 2017.

[2] V. Dolejší, F. Roskovec, and M. Vlasák. Residual based error estimates for the space-time discontinuous galerkin method applied to the compressible flows. *Computers Fluids*, 117:304–324, 2015.

[3] P.A. Lott, H.F. Walker, C.S. Woodward, and U.M. Yank. An accelerated picard method for nonlinear systems related to variably saturated flow). *Advance in Water Resources*, 38:92–101, 2011.

[4] J. Papež. *Estimation of the algebraic error and stopping criteria in numerical solution of partial diferential equations.* Master thesis, Charles University in Prague, Faculty of Mathematics and Physics, 2011.

[5] P. Deuflhard. *Newton Methods for Nonlinear Problems.* Druhé opravené vydání. Springer, Berlin, 2004.

[6] A. Greenbaum and T.P. Chartier. *Numerical Methods: Design, Analysis, and Computer Implementation of Algorithms.* Princeton University Press, 3.2.2012 edition, 2012.

[7] J.D. Tebbens, I. Hnětýnkova, M. Plešinger, T. Strakoš, and P. Tichý. *Analýza metod pro maticové výpočty.* První. Matfyzpress, Praha, 2012.

[8] H.F. Walker. Anderson acceleration for fixed-point iterations. *SIAM J. Numer. Anal.*, 49:1715–1735, 2011.