Weierstrass Institute for
Applied Analysis and Stochastics

Leibniz
Association

Algebraic Finite Element Stabilizations for Convection-Diffusion Equations

Volker John (WIAS and Freie Universität Berlin)
joint work with Gabriel R. Barrenechea (Glasgow), Petr Knobloch (Prague), Abhinav Jha (FUB)

Mohrenstrasse 39 · 10117 Berlin · Germany · Tel. +49 30 20372 0 · www.wias-berlin.de
Workshop on Computational Modeling and Numerical Analysis 2019, LNCC - Petrópolis, 25-28 February 2019

Outline of the talk

1 Convection-Diffusion-Reaction Equations

- Ω – bounded domain in $\mathbb{R}^d$, $d \in \{2,3\}$
- steady-state convection-diffusion-reaction equations

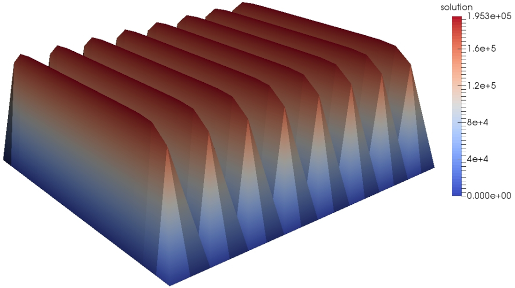
$$-\varepsilon\Delta u + \mathbf{b}\cdot\nabla u + cu = f \quad \text{in } \Omega$$

  - boundary conditions
- time-dependent convection-diffusion-reaction equations

$$\partial_t u - \varepsilon\Delta u + \mathbf{b}\cdot\nabla u + cu = f \quad \text{in } (0,T]\times\Omega$$

  - initial condition
  - boundary conditions
- model for transport of species (concentration, temperature, . . .)
  - diffusive transport
  - convective transport
- convection-dominated case $\varepsilon \ll \|\mathbf{b}\|_{L^\infty(\Omega)}$ of interest in applications
  - typical feature: layers

- Galerkin finite element discretization: numerical solution globally polluted with large spurious oscillations



- $\Longrightarrow$ stabilization necessary

- classical stabilizations: add terms to Galerkin finite element discretization
- most popular method: Streamline-Upwind Petrov–Galerkin (SUPG) method, [1,2]
  - stabilization in streamline direction with additional term

$$\sum_{K\in\mathcal{T}_h}(-\varepsilon\,\Delta u_h + \mathbf{b}\cdot\nabla u_h + c\,u_h - f, y_h\,\mathbf{b}\cdot\nabla v_h)_K$$

  - a standard parameter choice

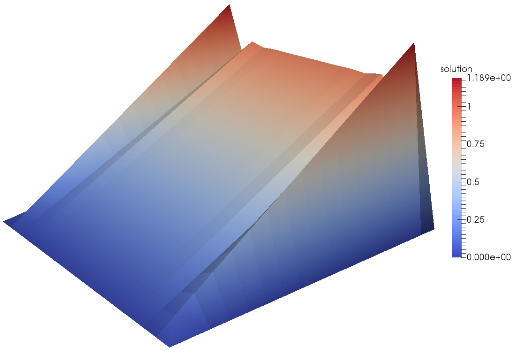
$$y_h|_K = \frac{h_K}{2\,p\,|\mathbf{b}|}\,\xi(Pe_K) \ \text{ with } \ \xi(\alpha) = \coth\alpha - \frac{1}{\alpha}\,,\ Pe_K = \frac{|\mathbf{b}|\,h_K}{2\,p\,\varepsilon}$$

- advantages
  - numerical analysis available
  - higher order of convergence in appropriate norms for higher order finite elements

[1] Hughes, Brooks; Finite Element Methods for Convection Dominated Flows, 19 – 35, 1979

[2] Brooks, Hughes; Comput. Methods Appl. Mech. Engrg. 32, 199 – 259, 1982

- typical result in numerical simulations



- (strong) spurious oscillations in vicinity of layers
  - not tolerable in many applications

- comprehensive numerical assessments of stabilized finite element methods
  - steady-state problems [1]
  - time-dependent problems [2]
- results
  - algebraic stabilizations from [3,4,5] showed very good results
  - comparative study from [2]: FEM–FCT schemes. These were clearly the best schemes.
  - comparative study from [1]: From the more modern approaches which were included in this study, FEMTVD (AFC) stands out somewhat by suppressing under- and overshoots . . .
  - moderate smearing of layers

[1] Augustin, Caiazzo, Fiebach, Fuhrmann, J., Linke, Umla; Comput. Methods Appl. Mech. Engrg. 200, 3395 – 3409, 2011

[2] J., Schmeyer; Comput. Methods Appl. Mech. Engrg. 198, 475 – 494, 2008

[3] Kuzmin; Proc. Int. Conf. Comp. Meth. for Coup. Prob. in Sci. and Engrg., CIMNE, 2007

[4] Kuzmin, Möller; in Flux-Corrected Transport: Principles, Algorithms and Applications, 155 – 206, 2005

[5] Kuzmin; J. Comput. Phys. 228, 2517 – 2534, 2009

- starting point: algebraic linear system of equations of Galerkin discretization

$$\mathbb{A}U = G \quad \mathbb{A} \in \mathbb{R}^{n\times n}$$

- define symmetric matrix $\mathbb{D}$ with

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\}, \; i \neq j, \quad d_{ii} = -\sum_{i\neq j} d_{ij}$$

- equivalent system

$$(\mathbb{A} + \mathbb{D})\, U = G + \mathbb{D}U$$

  - $\mathbb{A} + \mathbb{D}$ is an M-matrix

- decomposition into fluxes

$$(\mathbb{D}U)_i = \sum_{j\neq i} f_{ij} = \sum_{j\neq i} d_{ij}\,(u_j - u_i)$$

- ansatz for algebraic stabilization scheme

$$((\mathbb{A}+\mathbb{D})\,U)_i = G_i + \sum_j \alpha_{ij} f_{ij}, \quad i = 1,\ldots,M$$

  - limiter $\alpha_{ij} \in [0,1]$
  - $\alpha_{ij} = 1$ for all i, j: original Galerkin discretization
  - $\alpha_{ij} = 0$ for all i, j: corresponds to low order discretization (very diffusive)
  - $\{\alpha_{ij}\}$ depend usually on solution $\Longrightarrow$ nonlinear discretization
- difficulties
  - appropriate choice of α_{ij}
  - numerical analysis: completely different construction as all other stabilized finite element schemes
- advantage
  - implementation independent of the dimension (if limiter do not depend on the grid)

2 Numerical Analysis of Algebraic Stabilizations

- first numerical analysis in [1]
  - 1d problem without assuming $\alpha_{ij} \neq \alpha_{ji}$
  - no conservation
  - construction of examples without solution possible
  - subproblems in fixed point iteration have unique solution
  - redefinition of limiters
    - nonlinear problem has solution
    - discrete maximum principle (DMP) only approximately satisfied (order of a small regularization parameter)
- main conclusion: symmetry of limiter also desirable from mathematical point of view

[1] Barrenechea, J., Knobloch; IMA J. Numer. Anal. 35, 1729 – 1756, 2015

- numerical analysis for multi-dimensional problems in [1]
- starting point: linear system of equations

$$\begin{aligned} \sum_{j=1}^{N} a_{ij} u_j &= g_j,\ i = 1, \ldots, M, \\ u_i &= u_i^{\mathrm{b}},\ i = M+1, \ldots, N \end{aligned}$$

  - assumption: $\mathbb{A}$ is positive definite
- rewrite the system with limiters

$$\begin{aligned} \sum_{j=1}^{N} a_{ij} u_j + \sum_{j=1}^{N} (1-\alpha_{ij})\, d_{ij}\, (u_j - u_i) &= g_j,\ i = 1, \ldots, M, \\ u_i &= u_i^{\mathrm{b}},\ i = M+1, \ldots, N \end{aligned}$$

  - symmetric limiter: $\alpha_{ij} = \alpha_{ji}$

[1] Barrenechea, J., Knobloch; SIAM J. Numer. Anal. 54, 2427 – 2451, 2016

- solvability of nonlinear problem:
  - let $\alpha_{ij} : \mathbb{R}^N \to [0,1]$ be such that

$$\Phi_{ij} = \alpha_{ij}(u_1, \ldots, u_N)(u_j - u_i)$$

    is a continuous function of $u_1, \ldots, u_N$
  - $\Longrightarrow$ there is a solution of nonlinear problem
  - proof: based on Brouwer's fixed point theorem

- solvability of nonlinear problem:
  - let $\alpha_{ij} : \mathbb{R}^N \to [0,1]$ be such that

$$\Phi_{ij} = \alpha_{ij}(u_1, \ldots, u_N)(u_j - u_i)$$

    is a continuous function of $u_1, \ldots, u_N$
  - $\Longrightarrow$ there is a solution of nonlinear problem
  - proof: based on Brouwer's fixed point theorem
- corollary: there is a unique solution of the linear system with $\alpha_{ij} \in [0,1]$, $i, j = 1, \ldots, N$

- criterion for continuity condition:
  - let $\alpha_{ij} : \mathbb{R}^N \to [0, 1]$ satisfy

$$\alpha_{ij}(U) = \frac{A_{ij}(U)}{|u_j - u_i| + B_{ij}(U)} \quad \forall\, U \equiv (u_1, \ldots, u_N) \in \mathbb{R}^N,\ u_i \neq u_j$$

    - $A_{ij}, B_{ij} : \mathbb{R}^N \to [0, \infty)$ are nonnegative functions
    - continuous at any point $U \in \mathbb{R}^N$ with $u_i \neq u_j$
  - $\Longrightarrow \Phi_{ij}(U) := \alpha_{ij}(U)(u_j - u_i)$ is continuous function of $u_1, \ldots, u_N$ on $\mathbb{R}^N$

- Kuzmin limiter [1] (standard)
  - using ideas from [2]
  - compute for all pairs $i, j \in \{1, \dots, N\}$

$$\begin{aligned} P_i^+ &:= P_i^+ + \max\{0, f_{ij}\}\,, \; P_i^- := P_i^- - \max\{0, f_{ji}\} && \text{if } a_{ji} \le a_{ij}\,, \\ Q_i^+ &:= Q_i^+ + \max\{0, f_{ji}\}\,, \; Q_i^- := Q_i^- - \max\{0, f_{ij}\} && \text{if } i < j\,, \\ Q_j^+ &:= Q_j^+ + \max\{0, f_{ij}\}\,, \; Q_j^- := Q_j^- - \max\{0, f_{ji}\} && \text{if } i < j \end{aligned}$$

  - compute

$$R_i^+ := \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \quad R_i^- := \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}, \quad i = 1, \dots, N$$

  - set at Dirichlet nodes

$$R_i^+ := 1\,, \quad R_i^- := 1\,, \quad i = M+1, \dots, N$$

[1] Kuzmin; Proc. Int. Conf. Comp. Meth. Coupl. Prob. Sci. Engrg., CIMNE 1 – 5, 2007

[2] Zalesak; J. Comp. Phys. 31, 335 – 362, 1979

- Kuzmin limiter [1] (cont.)
  - for any $i, j \in \{1, \ldots, N\}$ such that $a_{ji} \le a_{ij}$ set

$$\alpha_{ij} := \begin{cases} R_i^+ & \text{if } f_{ij} > 0\,, \\ 1 & \text{if } f_{ij} = 0\,, \\ R_i^- & \text{if } f_{ij} < 0\,, \end{cases} \qquad \alpha_{ji} := \alpha_{ij}$$

[1] Kuzmin; Proc. Int. Conf. Comp. Meth. Coupl. Prob. Sci. Engrg., CIMNE 1 – 5, 2007

- Kuzmin limiter [1] (cont.)
  - for any $i, j \in \{1, \dots, N\}$ such that $a_{ji} \le a_{ij}$ set

$$\alpha_{ij} := \begin{cases} R_i^+ & \text{if } f_{ij} > 0\,, \\ 1 & \text{if } f_{ij} = 0\,, \\ R_i^- & \text{if } f_{ij} < 0\,, \end{cases} \qquad \alpha_{ji} := \alpha_{ij}$$

- α_{ij} are such that $\alpha_{ij}(u_1, \dots, u_N)(u_j - u_i)$ are Lipschitz-continuous functions of $u_1, \dots, u_N$ on $\mathbb{R}^N$
  - proof based on rewriting limiters and deriving representation that fits into the criterion of continuity with

$$A_{ij} = \frac{1}{|d_{ij}|} \begin{cases} \min\{-P_i^-, -Q_i^-\} & \text{if } u_i < u_j\,, \\ \min\{P_i^+, Q_i^+\} & \text{if } u_i > u_j\,, \end{cases}$$

$$B_{ij} = \frac{1}{|d_{ij}|} \begin{cases} -P_i^- & \text{if } u_i < u_j\,, \\ P_i^+ & \text{if } u_i > u_j\,. \end{cases}$$

[1] Kuzmin; Proc. Int. Conf. Comp. Meth. Coupl. Prob. Sci. Engrg., CIMNE 1 – 5, 2007

- discrete maximum principle can be proved
  - if $\sum_{j=1}^{N} a_{ij} \geq 0$, then for any $i \in \{1,\ldots,M\}$

$$g_i \leq 0 \quad \Rightarrow \quad u_i \leq \max_{j\neq i,\, a_{ij}\neq 0} u_j \quad \text{for } u_i \geq 0 \quad \Rightarrow \quad u_i \leq \max_{j\neq i,\, a_{ij}\neq 0} u_j^+$$
$$g_i \geq 0 \quad \Rightarrow \quad u_i \geq \min_{j\neq i,\, a_{ij}\neq 0} u_j \quad \text{for } u_i \leq 0 \quad \Rightarrow \quad u_i \geq \min_{j\neq i,\, a_{ij}\neq 0} u_j^-$$

  - if $\sum_{j=1}^{N} a_{ij} = 0$, then for any $i \in \{1,\ldots,M\}$

$$g_i \leq 0 \quad \Rightarrow \quad u_i \leq \max_{j\neq i,\, a_{ij}\neq 0} u_j$$
$$g_i \geq 0 \quad \Rightarrow \quad u_i \geq \min_{j\neq i,\, a_{ij}\neq 0} u_j$$

- convergence
- flux correction scheme is equivalent to variational problem

  find $u_h \in W_h$ such that $u_h(x_i) = u_b(x_i)$, $i = M+1, \ldots, N$, and

$$a_h(u_h, v_h) + d_h(u_h; u_h, v_h) = \langle g, v_h \rangle \quad \forall\, v_h \in V_h$$

  - V_h – finite element space with homogeneous Dirichlet boundary conditions
  - W_h – finite element space with prescribed Dirichlet boundary conditions
  - $a_h(\cdot,\cdot)$ – approximation of bilinear form of continuous problem with

$$a_h(v_h, v_h) \geq C_a \|v_h\|_a^2 \quad \forall\, v_h \in V_h$$

  - stabilization

$$d_h(w_h; z_h, v_h) = \sum_{i,j=1}^{N} (1 - \alpha_{ij}(w_h))\, d_{ij}\, (z_j - z_i)\, v_i \quad \forall\, w_h, z_h, v_h \in W_h$$

- convergence (cont.)
- Cauchy–Schwarz inequality holds

$$|d_h(w;z,v)|^2 \leq d_h(w;z,z)\, d_h(w;v,v) \quad \forall\, w,z,v \in C(\overline{\Omega})$$

- natural norm on V_h

$$\|v_h\|_h := \Big(C_a\, \|v_h\|_a^2 + d_h(u_h;v_h,v_h)\Big)^{1/2}, \quad v_h \in V_h$$

- convergence (cont.)
- Cauchy–Schwarz inequality holds

$$|d_h(w;z,v)|^2 \le d_h(w;z,z)\, d_h(w;v,v) \quad \forall\, w,z,v \in C(\overline{\Omega})$$

- natural norm on V_h

$$\|v_h\|_h := \Big(C_a\, \|v_h\|_a^2 + d_h(u_h;v_h,v_h)\Big)^{1/2}, \quad v_h \in V_h$$

- abstract error estimate (Strang-type) can be derived

$$\|u-u_h\|_h \le C_a^{1/2}\, \|u - i_h u\|_a + \sup_{v_h\in V_h} \frac{a(u,v_h) - a_h(i_h u, v_h)}{\|v_h\|_h} + (d_h(u_h; i_h u, i_h u))^{1/2}$$

  - interpolation error
  - consistency error

- convergence (cont.)
- application of abstract approach to steady-state convection-diffusion reaction equations

$$a(u,v) = \varepsilon\,(\nabla u, \nabla v) + (\mathbf{b}\cdot\nabla u, v) + (c\,u, v)$$

with

$$\nabla\cdot\mathbf{b} = 0\,, \quad c \geq \sigma_0 \geq 0 \quad \text{in } \Omega$$

- P_1 finite elements
- discrete bilinear form by using mass lumping

$$(c\,u_h, v_h) = \sum_{i=1}^{M} (c\,u_h, \varphi_i)\,v_i \approx \sum_{i=1}^{M} (c, \varphi_i)\,u_i\,v_i \quad \forall\ u_h \in W_h,\ v_h \in V_h$$

  - matrix $\mathbb{D}$ becomes independent of reaction
  - consistency error from mass lumping

$$\left|(c\,u_h, v_h) - \sum_{i=1}^{M} (c, \varphi_i)\,u_i\,v_i\right| \leq C\,h\,\|c\|_{0,\infty,\Omega}\,|u_h|_{1,\Omega}\,\|v_h\|_{0,\Omega}$$

- convergence (cont.)
- norm from coercivity of bilinear form

$$\|v\|_a^2 = \varepsilon\,|v|_{1,\Omega}^2 + \sigma_0\,\|v\|_{0,\Omega}^2$$

- convergence (cont.)
- norm from coercivity of bilinear form

$$\|v\|_a^2 = \varepsilon\,|v|_{1,\Omega}^2 + \sigma_0\,\|v\|_{0,\Omega}^2$$

- interpolation error

$$\|u - i_h u\|_a \le C\,(\varepsilon + \sigma_0\,h^2)^{1/2}\,h\,|u|_{2,\Omega}$$

- convergence (cont.)
- norm from coercivity of bilinear form

$$\|v\|_a^2 = \varepsilon \, |v|_{1,\Omega}^2 + \sigma_0 \, \|v\|_{0,\Omega}^2$$

- interpolation error

$$\|u - i_h u\|_a \le C \, (\varepsilon + \sigma_0 \, h^2)^{1/2} \, h \, |u|_{2,\Omega}$$

- first consistency error ($\sigma_0 > 0$)

$$\sup_{v_h \in V_h} \frac{a(u, v_h) - a_h(i_h u, v_h)}{\|v_h\|_h} \le C \, (\varepsilon + \sigma_0^{-1} \, \{\|\mathbf{b}\|_{0,\infty,\Omega}^2 + \|c\|_{0,\infty,\Omega}^2\})^{1/2} \, h \, \|u\|_{2,\Omega}$$

  - additional dependency on ε^{-1} if $\sigma_0 = 0$

- convergence (cont.)
- second consistency error: only with the assumptions $\alpha_{ij} \in [0,1]$, $\alpha_{ij} = \alpha_{ji}$

$$d_h(w_h; i_h u, i_h u)^{1/2} \le C\left(\varepsilon + \|\mathbf{b}\|_{0,\infty,\Omega}\, h\right)^{1/2} |i_h u|_{1,\Omega} \quad \forall\, w_h \in W_h,\ u \in C(\overline{\Omega})$$

  - convergence order lost already in first step of the proof

$$\begin{aligned} d_h(w_h; i_h u, i_h u) &= \sum_{\substack{i,j=1\\ i<j}}^{N} (1-\alpha_{ij}(w_h))\,|d_{ij}|\,[u(x_i)-u(x_j)]^2 \\ &\le \sum_{T\in\mathcal{T}_h} \sum_{x_i,x_j\in T} |d_{ij}|\,[u(x_i)-u(x_j)]^2 \\ &\le \ldots \end{aligned}$$

- convergence (cont.)
- final estimate

$$\begin{aligned}\|u-u_h\|_h &\le C\left(\varepsilon+\sigma_0^{-1}\left\{\|\mathbf{b}\|^2_{0,\infty,\Omega}+\|c\|^2_{0,\infty,\Omega}\right\}+\sigma_0 h^2\right)^{1/2} h\,\|u\|_{2,\Omega}\\ &\quad +C\left(\varepsilon+\|\mathbf{b}\|_{0,\infty,\Omega}\,h\right)^{1/2}|i_h u|_{1,\Omega}\,.\end{aligned}$$

  - in general only order 0.5 in convection-dominated regime
  - in general no convergence in diffusion-dominated regime

[1] Barrenechea, J., Knobloch; SIAM J. Numer. Anal. 54, 2427 – 2451, 2016

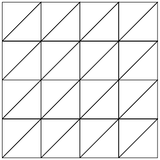
- convergence (cont.)
- final estimate

$$\begin{aligned}\|u-u_h\|_h \quad &\le \quad C\left(\varepsilon+\sigma_0^{-1}\left\{\|\mathbf{b}\|_{0,\infty,\Omega}^2+\|c\|_{0,\infty,\Omega}^2\right\}+\sigma_0 h^2\right)^{1/2} h\,\|u\|_{2,\Omega} \\ &\quad +C\left(\varepsilon+\|\mathbf{b}\|_{0,\infty,\Omega}\,h\right)^{1/2}|i_h u|_{1,\Omega}\,.\end{aligned}$$

  - in general only order 0.5 in convection-dominated regime
  - in general no convergence in diffusion-dominated regime
  - numerical studies in [1] with $\alpha_{ij}=0.5$: estimate is sharp within the assumptions of the analysis
  - refined analysis of diffusion-dominated regime on special types of grids proves better results
    - all angles of the triangles smaller than $\pi/2$: first order convergence
    - all angles of the triangles smaller or equal than $\pi/2$: order convergence 0.5

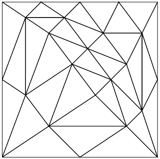
[1] Barrenechea, J., Knobloch; SIAM J. Numer. Anal. 54, 2427 – 2451, 2016

- convergence (cont.)
- results with Kuzmin limiter, convection-dominated case
  - arithmetic mean value of $\{1 - \alpha_{ij}(u_h)\}$ tends almost linearly to 0
  - optimal order of convergence only on Friedrichs–Keller type grids



l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$\|e_h\|_h$	ord.
3	5.457e−3	1.85	2.287e−1	1.10	1.114e−1	0.97
4	1.408e−3	1.95	1.074e−1	1.09	5.319e−2	1.07
5	3.493e−4	2.01	5.113e−2	1.07	2.472e−2	1.11
6	8.652e−5	2.01	2.546e−2	1.01	1.158e−2	1.09
7	2.152e−5	2.01	1.321e−2	0.95	5.533e−3	1.07
8	5.357e−6	2.01	6.822e−3	0.95	2.685e−3	1.04

- convergence (cont.)
- results with Kuzmin limiter, convection-dominated case
  - reduced order of convergence on irregular grids



l	$\|\|e_h\|\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$\|\|e_h\|\|_h$	ord.
3	6.125e−3	1.61	3.202e−1	0.71	9.209e−2	1.06
4	2.216e−3	1.47	2.244e−1	0.51	4.493e−2	1.04
5	9.946e−4	1.16	1.821e−1	0.30	2.226e−2	1.01
6	4.993e−4	0.99	1.559e−1	0.22	1.125e−2	0.98
7	2.519e−4	0.99	1.375e−1	0.18	5.682e−3	0.98
8	1.277e−4	0.98	1.231e−1	0.16	2.874e−3	0.98

- summary
  - first numerical analysis (error estimates, convergence) of algebraic stabilizations in [1]
  - obtained much more insight into these methods
    - in particular into their shortcomings
  - convergence in standard norms generally not optimal
  - supported by numerical examples
  - order of convergence depends on the used type of grid

[1] Barrenechea, J., Knobloch; SIAM J. Numer. Anal. 54, 2427 – 2451, 2016

3 Algebraic Stabilization with Linearity Preservation

- linearity preservation: stabilization vanishes if the solution is a first order polynomial
- Kuzmin limiter not linearity preserving on general meshes
- definition of a new limiter in [1]
  - is linearity preserving

[1] Barrenechea, J., Knobloch; M3AS 27, 525 – 548, 2017

- definition of the limiter
- set for any $i \in \{1, \ldots, M\}$

$$u_i^{\max} := \max_{j \in S_i \cup \{i\}} u_j \,, \quad u_i^{\min} := \min_{j \in S_i \cup \{i\}} u_j \,, \quad q_i := \gamma_i \sum_{j \in S_i} d_{ij} \,,$$

with $\gamma_i > 0$

- define

$$P_i^+ := \sum_{j \in S_i} f_{ij}^+ \,, \; P_i^- := \sum_{j \in S_i} f_{ij}^- \,, \; Q_i^+ := q_i \left(u_i - u_i^{\max}\right) , \; Q_i^- := q_i \left(u_i - u_i^{\min}\right)$$

- define

$$R_i^+ := \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \quad R_i^- := \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}$$

- P_i^+ or P_i^- vanishes, set $R_i^+ := 1$ or $R_i^- := 1$

- definition of the limiter (cont.)
- define

$$\widetilde{\alpha}_{ij} := \begin{cases} R_i^+ & \text{if } f_{ij} > 0\,, \\ 1 & \text{if } f_{ij} = 0\,, \\ R_i^- & \text{if } f_{ij} < 0\,, \end{cases} \qquad i = 1,\ldots,M,\ j = 1,\ldots,N$$

- set

$$\alpha_{ij} := \min\{\widetilde{\alpha}_{ij}, \widetilde{\alpha}_{ji}\}\,, \quad i,j = 1,\ldots,M\,,$$
$$\alpha_{ij} := \widetilde{\alpha}_{ij}\,, \qquad i = 1,\ldots,M,\ j = M+1,\ldots,N$$

- DMP
  - assume

$$\sum_{j=1}^{N} a_{ij} \geq 0\,, \qquad i = 1, \ldots, M$$

  - assume

$$\text{there exists } j \in \{1, \ldots, N\},\, j \neq i \;:\quad a_{ij} < 0 \quad \text{or} \quad a_{ij} < a_{ji}$$

    - typically satisfied for finite element discretizations of convection-diffusion equations
  - $\Longrightarrow$ DMP satisfied

- DMP
  - assume
$$\sum_{j=1}^{N} a_{ij} \geq 0\,, \qquad i = 1,\ldots,M$$
  - assume
$$\text{there exists } j \in \{1,\ldots,N\},\, j \neq i \;: \quad a_{ij} < 0 \quad \text{or} \quad a_{ij} < a_{ji}$$
    - typically satisfied for finite element discretizations of convection-diffusion equations
  - $\Longrightarrow$ DMP satisfied
- limiter is of the form
$$\Phi_{ij}(U) := \alpha_{ij}(u_1,\ldots,u_N)(u_j - u_i)$$
  and it is a continuous functions of $u_1,\ldots,u_N$ on $\mathbb{R}^N$
  - $\Longrightarrow$ existence of solution of nonlinear discrete problem
  - $\Longrightarrow$ unique solution of linearized problem

- convergence
  - same analysis and results as for Kuzmin limiter

- convergence
  - same analysis and results as for Kuzmin limiter
- linearity preservation: with appropriate choice of parameter γ_i
  - patch around vertex x_i

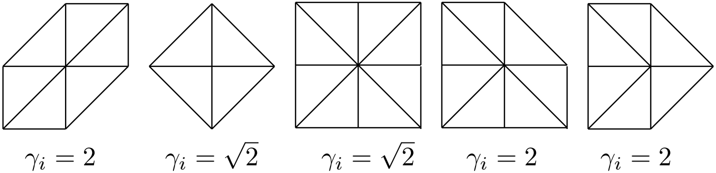
$$\Delta_i = \operatorname{supp} \varphi_i$$

  - Δ_i^{conv} convex hull of Δ_i
  - if

$$\gamma_i = \frac{\max\limits_{x_j \in \partial\Delta_i} |x_i - x_j|}{\operatorname{dist}(x_i, \partial\Delta_i^{\mathrm{conv}})}, \quad i = 1, \ldots, M$$

    then algebraic stabilization scheme is linearity preserving
  - same property for larger values of γ_i

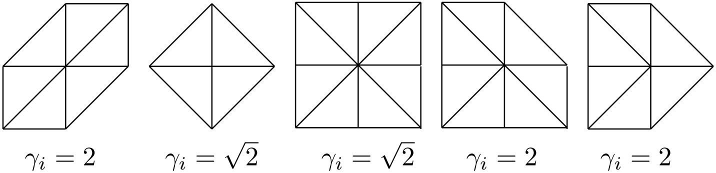
- linearity preservation (cont.)
  - examples



  - value for general patch in 2d easily to compute
  - easy to compute upper bound for value in 3d

[1] Allende, Barrenechea, Rankin; SIAM J. Sci. Comput. 39, A1903 – A2927, 2017

- linearity preservation (cont.)
  - examples



  - value for general patch in 2d easily to compute
  - easy to compute upper bound for value in 3d
- all results hold for arbitrary simplicial grids
- in particular: DMP + linearity preservation + optimal convergence (numerical experience) in diffusion-dominated regime, e.g., Poisson equation
- open problem: how to use linearity preservation in numerical analysis?
- fully computable a posteriori error estimator in [1]

[1] Allende, Barrenechea, Rankin; SIAM J. Sci. Comput. 39, A1903 – A2927, 2017

4 Connection to Edge-Based Stabilizations

- edge-based stabilizations already proposed in [1]: continuous interior penalty (CIP) method
  - linear discretization
- link between AFC schemes and nonlinear edge-based stabilizations established in [2]
  - different tools in the analysis of AFC schemes can be applied
  - unified analysis of both schemes possible [3]
    - existence of a solution
    - minimal conditions for validity of DMP
    - finite element error estimates

[1] Burman, Hansbo; CMAME 193, 1437 – 1453, 2004

[2] Barrenechea, Burman, Karakatsani; Numer. Math. 135, 521 – 545, 2017

[3] Barrenechea, J., Knobloch, Rankin; SeMA Journal 75, 655 – 685, 2018

- link to edge-based stabilizations: P_1 finite elements

$$
\begin{aligned}
& d_h(u_h; v_h, w_h) \\
= \quad & \sum_{i>j}(1-\alpha_{ij}(u_h))d_{ij}(v_j - v_i)w_i + \sum_{i<j}(1-\alpha_{ij}(u_h))d_{ij}(v_j - v_i)w_i \\
\overset{\text{chg. i,j}}{=} \quad & \sum_{i>j}(1-\alpha_{ij}(u_h))d_{ij}(v_j - v_i)w_i + \sum_{i>j}(1-\alpha_{ji}(u_h))d_{ji}(v_i - v_j)w_j \\
\overset{\text{symm.}}{=} \quad & \sum_{i>j}(1-\alpha_{ij}(u_h))d_{ij}(v_j - v_i)(w_i - w_j) \\
= \quad & \sum_{E\in\mathcal{E}_h}(1-\alpha_E(u_h))|d_E|(v_h(x_{E,1}) - v_h(x_{E,2}))(w_h(x_{E,1}) - w_h(x_{E,2})) \\
= \quad & \sum_{E\in\mathcal{E}_h}(1-\alpha_E(u_h))|d_E|h_E\left(\nabla v_h\cdot \boldsymbol{t}_E, \nabla w_h \cdot \boldsymbol{t}_E\right)_E
\end{aligned}
$$

  - index E denotes quantities on edge E that connects $x_{E,1}$ and $x_{E,2}$

5 Numerical Studies on Accuracy for Different Limiters

- limiters
  - Kuzmin limiter [1]
  - BJK limiter [2], linearity preserving
  - BBK limiter [3], edge-based
    - numerical studies in [4]: less accurate than the other two limiters

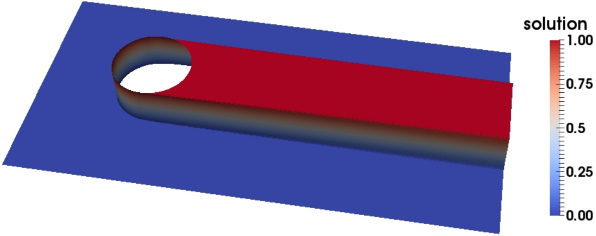
[1] Kuzmin; Proc. Int. Conf. Comp. Meth. Coupl. Prob. Sci. Engrg., CIMNE 1 – 5, 2007

[2] Barrenechea, J., Knobloch; M3AS 27, 525 – 548, 2017

[3] Barrenechea, Burman, Karakatsani; Numer. Math. 135, 521 – 545, 2017

[4] Barrenechea, J., Knobloch, Rankin; SeMA Journal 75, 655 – 685, 2018

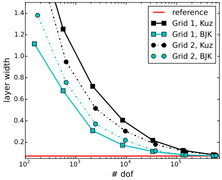
- 2d Hemker problem [1]
  - $\varepsilon = 10^{-4}, \mathbf{b} = (1,0)^T, c = f = 0$
  - reference solution



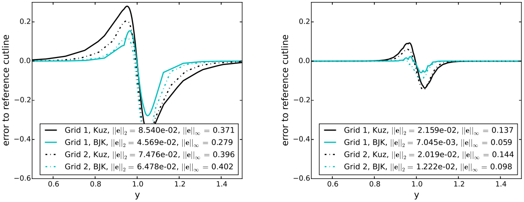
  - Grid 1: structured, Grid 2: unstructured
  - P_1 finite elements

[1] Barrenechea, J., Knobloch, Rankin; SeMA Journal 75, 655 – 685, 2018

- 2d Hemker problem, representative results from [1]
  - smearing of the interior layer



  - error at cutlines, different refinement levels



[1] Barrenechea, J., Knobloch, Rankin; SeMA Journal 75, 655 – 685, 2018

- 2d Hemker problem
  - [1]: results with BJK limiter considerably more accurate
  - but: [2]: nonlinear problems for BJK limiter and $\varepsilon = 10^{-6}$ not solvable on fine grids
    - within prescribed maximal number of iterations
    - details: see next part of the talk
- experience so far (also with other examples): if nonlinear problems for BJK limiter can be solved, one gets the most accurate solutions among all studied limiters

[1] Barrenechea, J., Knobloch, Rankin; SeMA Journal 75, 655 – 685, 2018

[2] Jha, J.; submitted 2018

6 Numerical Studies on Solvers for Different Limiters

- limiters
  - Kuzmin limiter [1]
  - BJK limiter [2], linearity preserving
- limiters depend on discrete solution $\Longrightarrow$ nonlinear problems
- first studies in [3]
  - simple academic examples in 2d
  - P_1 finite elements
  - initial iterate (Zero, Galerkin solution, SUPG solution, Upwind FE solution) does not possess much impact on number of iterations
    - here: SUPG solution initial iterate

[1] Kuzmin; Proc. Int. Conf. Comp. Meth. Coupl. Prob. Sci. Engrg., CIMNE 1 – 5, 2007

[2] Barrenechea, J., Knobloch; M3AS 27, 525 – 548, 2017

[3] Jha, J.; Proc. BAIL 2018, to appear

- given iterate $u^{(m)}$
- fixed point iteration with changing matrix

$$\begin{aligned}\sum_{j=1}^{N} a_{ij}\,\tilde{u}_j^{(m+1)} + \sum_{j=1}^{N}\left(1-\alpha_{ij}^{(m)}\right) d_{ij}\,\left(\tilde{u}_j^{(m+1)} - \tilde{u}_i^{(m+1)}\right) &= g_i,\\ \tilde{u}_i^{(m+1)} &= u_i^b\end{aligned}$$

- fixed point iteration with fixed matrix: using

$$\sum_{j=1}^{N}(1-\alpha_{ij})d_{ij}(u_j-u_i) = \sum_{j=1}^{N} d_{ij}u_j - u_i\underbrace{\sum_{j=1}^{N} d_{ij}}_{=0} - \sum_{j=1}^{N}\alpha_{ij}d_{ij}(u_j-u_i),$$

gives

$$\begin{aligned}\sum_{j=1}^{N}(a_{ij}+d_{ij})\tilde{u}_j^{(m+1)} &= g_i + \sum_{j=1}^{N}\alpha_{ij}^{(m)} f_{ij}^{(m)}, \quad & i=1,\ldots,M,\\ \tilde{u}_i^{(m+1)} &= u_i^b, & i=M+1,\ldots,N\end{aligned}$$

- fixed point iterations
  - fixed point iteration with fixed matrix
    - matrix is M-matrix
    - with sparse direct solver: factorization only once needed
  - fixed point iteration with changing matrix
    - more implicit approach, hope for better convergence properties
  - general fixed point iteration by linear combination

$$
\begin{aligned}
& \sum_{j=1}^{N} (a_{ij} + d_{ij})\, \tilde{u}_j^{(m+1)} - \omega_{\mathrm{fp}} \sum_{j=1}^{N} \alpha_{ij}^{(m)} d_{ij} \left(\tilde{u}_j^{(m+1)} - \tilde{u}_i^{(m+1)} \right) \\
& \quad = \quad g_i + (1 - \omega_{\mathrm{fp}}) \sum_{j=1}^{N} \alpha_{ij}^{(m)} f_{ij}^{(m)}, \quad i = 1, \ldots, M, \\
\tilde{u}_i^{(m+1)} & \quad = \quad u_i^b, \qquad\qquad i = M+1, \ldots, N
\end{aligned}
$$

- formal Newton method
  - formal derivation of Jacobian

$$DF\left(\underline{u}^{(m)}\right)_{ij}$$

$$= \begin{cases} a_{ij} + d_{ij} - \alpha_{ij}^{(m)} d_{ij} - \sum_{k=1}^{N} \frac{\partial \alpha_{ik}^{(m)}}{\partial u_j} d_{ik} \left(u_k^{(m)} - u_i^{(m)}\right) & \text{if } i \neq j, \\ a_{ii} + d_{ii} + \sum_{\substack{j=1 \\ j\neq i}}^{N} \alpha_{ij}^{(m)} d_{ij} - \sum_{k=1}^{N} \frac{\partial \alpha_{ik}^{(m)}}{\partial u_i} d_{ik} \left(u_k^{(m)} - u_i^{(m)}\right) & \text{if } i = j \end{cases}$$

- formal Newton method: how to deal with non-smooth cases?
- discussion only for Kuzmin limiter
  - involves maxima and minima of two arguments, one of them is constant

1. non-regularized approach
   - take one-sided derivative w.r.t. constant, i.e., set value to zero
2. regularized approach
   - replace maximum for some $\sigma > 0$ by [1]

$$\max_\sigma(x, y) = \frac{1}{2}\left(x + y + \sqrt{(x - y)^2 + \sigma}\right)$$

   - we did not regularized the limiter in the equation, only in the iteration matrix, since
     - in our opinion: solution should not depend on solver
     - analytical results from literature not longer applicable

[1] Badia, Bonilla: CMAME 313, 133–158, 2017

- general form of the matrix

$$\underbrace{\underbrace{\underbrace{a_{ij} + d_{ij}}_{\text{fp, const. matrix}} \quad - \omega_{\text{fp}} \alpha_{ij} d_{ij}}_{\text{fp, changing matrix}} + \omega_{\text{jac}} (\text{term with der. of } \alpha_{ij})}_{\text{formal Newton}}, \quad i \neq j$$

  - similar for diagonal entries
  - neglect entries of formal Jacobian that did not fit in sparsity pattern
  - some more modifications for regularized Newton approach
- iteration

$$\underline{u}^{(m+1)} = \underline{u}^{(m)} + \omega^{(m)} \left(\underline{\tilde{u}}^{(m+1)} - \underline{u}^{(m)} \right)$$

  - adaptive choice of damping parameter as proposed in [1]

[1] J., Knobloch: CMAME 197, 1997–2014, 2008

- further algorithmic components
  - Anderson acceleration of fixed point methods [1]
    - gives second order information

[1] Walker, Ni; SIAM J. Numer. Anal. 49, 1715 – 1735, 2011

[2] Badia, Bonilla; CMAME 313, 133 – 158, 2017

[3] Jha, J.; Comput. Math. Appl., in revision, 2019

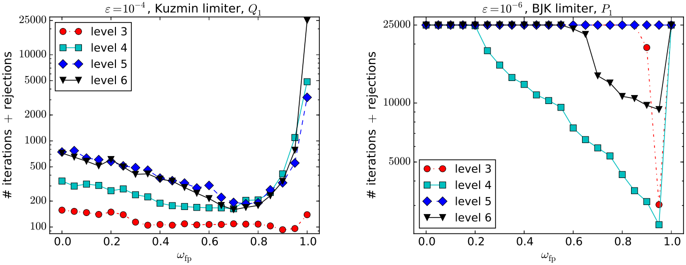
- further algorithmic components
  - Anderson acceleration of fixed point methods [1]
    - gives second order information
  - projection to admissible values after each iteration as proposed in [2]
    - DMP holds only for solution of nonlinear problem
    - projection should ensure this property for intermediate iterates such that early termination of iteration is possible
    - projection can be performed only if admissible values are known a priori
    - projection is simply a truncation
    - experience [3]:
      - often no big impact on number of iterations
      - one example: no convergence with projection; convergence without

[1] Walker, Ni; SIAM J. Numer. Anal. 49, 1715 – 1735, 2011

[2] Badia, Bonilla; CMAME 313, 133 – 158, 2017

[3] Jha, J.; Comput. Math. Appl., in revision, 2019

- 2d Hemker problem [1]
  - $\varepsilon \in \{10^{-4}, 10^{-6}\}$, $\mathbf{b} = (1,0)^T$, $c = f = 0$
  - Kuzmin limiter with P_1 and Q_1 finite elements
  - BJK limiter with P_1 finite elements
  - typical result for general fixed point iteration

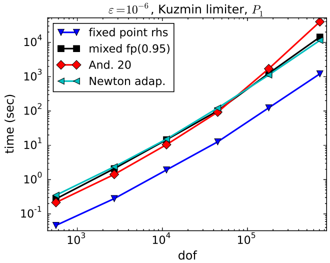


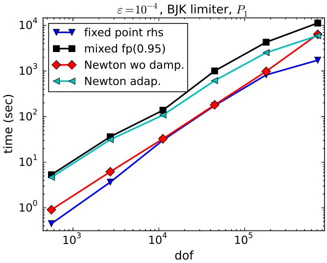
[1] Jha, J.; Comput. Math. Appl., in revision, 2019

- 2d Hemker problem, further observations (also in the other examples) [1]
  - problems with Kuzmin limiter generally easier to solve
  - Anderson acceleration
    - Kuzmin limiter: number of iterations sometimes considerably reduced, but sometimes even more iterations
    - BJK limiter: failed in all examples
  - formal Newton method without damping
    - Kuzmin limiter: failed generally
    - BJK limiter: sometimes comparably very few iterations
  - formal Newton method with damping
    - both limiters: number of iterations sometimes considerably reduced, but sometimes even more iterations

[1] Jha, J.; Comput. Math. Appl., in revision, 2019

- 2d Hemker problem, computing times for approaches with fewest number of iterations [1]

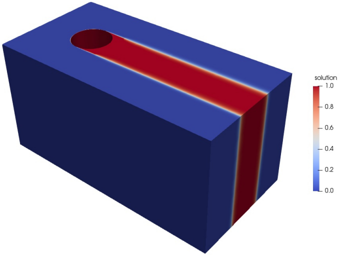


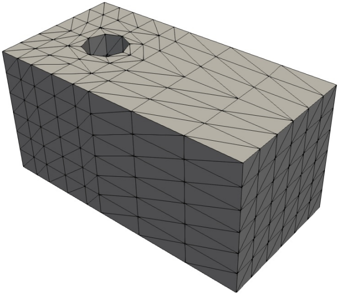


  - fixed point iteration with fixed matrix one order of magnitude faster than other methods
    - sparse direct solver UMFPACK requires only one factorization
    - getting the discrete system is very fast

[1] Jha, J.; Comput. Math. Appl., in revision, 2019

- 3d Hemker problem [1]
  - $\varepsilon \in \{10^{-4}, 10^{-6}\}, \mathbf{b} = (1,0,0)^T, c = f = 0$
  - solution for $\varepsilon = 10^{-6}$

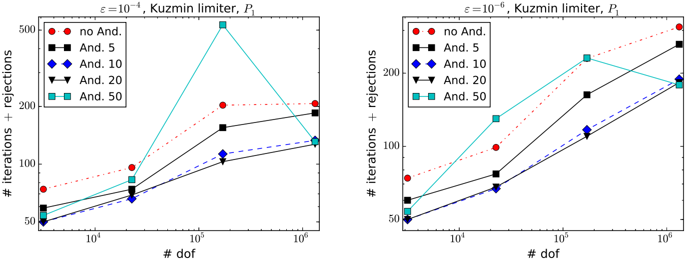




  - structured grid
  - Kuzmin limiter with P_1 and Q_1 finite elements
  - BJK limiter with P_1 finite elements

[1] Jha, J.; Comput. Math. Appl., in revision, 2019

- 3d Hemker problem [1]
  - typical impact of Anderson acceleration, Kuzmin limiter

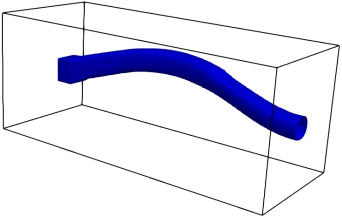


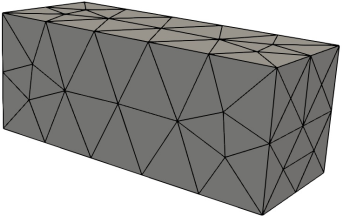
— user-chosen parameter: number of Anderson vectors

— in each iteration, eigenvalue problem of the size of the number of Anderson vectors has to be solved

[1] Jha, J.; Comput. Math. Appl., in revision, 2019

- 3d problem with non-constant convection from
  - $\varepsilon \in \{10^{-4}, 10^{-6}\}$, $\mathbf{b}$ non-constant, $c = f = 0$
  - solution for $\varepsilon = 10^{-6}$

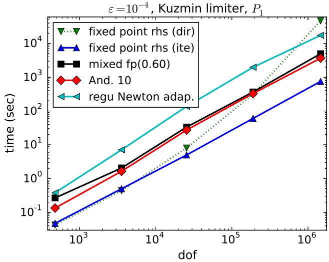


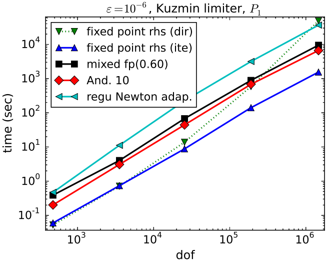


  - unstructured grid
  - Kuzmin limiter with P_1 and Q_1 finite elements
  - BJK limiter with P_1 finite elements

[1] Barrenechea, J., Knobloch, Rankin; SeMA Journal 75, 655 – 685, 2018

- 3d problem with non-constant convection, efficiency (computing times) [1]
  - linear systems solved iteratively: GMRES with right preconditioner SSOR
  - only for fixed point iteration with fixed matrix also UMFPACK





  - fixed point iteration with fixed matrix half an order of magnitude faster than other methods
    - iterative solver for linear systems very efficient (M-matrix)

[1] Jha, J.; Comput. Math. Appl., in revision, 2019

- summary [1]
  - simplest method by far most efficient in terms of computing times
    - fixed point iteration with fixed matrix
    - 2d: sparse direct solvers very efficient, only one factorization needed
    - 3d: iterative solver for linear system with M-matrix very efficient
  - number of iterations of fixed point iteration with fixed matrix usually quite large
  - more complicated methods might reduce these only sometimes considerably
  - none of the methods needed really few iterations
  - solution of the nonlinear problems is still a bottleneck for steady-state problems

[1] Jha, J.; Comput. Math. Appl., in revision, 2019

7 Outlook

- good discretization for convection-diffusion-reaction equations should [1]
  - compute sharp layers
  - not compute spurious oscillations
  - be efficient in its use

  after 40 years of research no method available that ticks all boxes !!!

[1] J., Knobloch, Novo; Comp. Visual. Sci. 19, 47 – 63, 2018

7 Outlook

- good discretization for convection-diffusion-reaction equations should [1]
  - compute sharp layers
  - not compute spurious oscillations
  - be efficient in its use

  after 40 years of research no method available that ticks all boxes !!!
- our opinion
  - algebraic stabilizations are a promising class, at least for first two issues
  - they should be augmented with geometric information
- important open problems
  - steady-state problems: analysis for special grids, analysis for anisotropic grids, efficient solvers for nonlinear problem
  - analysis for time-dependent problems

[1] J., Knobloch, Novo; Comp. Visual. Sci. 19, 47 – 63, 2018