1.4 Finite Element Methods Respecting Discrete Maximum Principles for Convection-Diffusion Equations

Volker John

Transferring important physical properties from a continuous model to a discrete version of this model is of utmost importance for many applications. Such properties might be conservation or balance laws that lead to constraints on solutions, like the conservation of mass for incompressible flow problems, or the guarantee of computing only physically admissible values with the discrete problem, like concentrations in the interval $[0, 1]$. There has been a long tradition in RG 3 *Numerical Mathematics and Scientific Computing* on developing, analyzing, and using such so-called *physically consistent* discretizations.

This highlight article considers discretizations of the elliptic linear second-order equation

$$-\varepsilon \Delta u + \mathbf{b} \cdot \nabla u + \sigma u = f \quad \text{in } \Omega \tag{1}$$

and its parabolic counterpart

$$\partial_t u - \varepsilon \Delta u + \mathbf{b} \cdot \nabla u + \sigma u = f \quad \text{in } (0, T] \times \Omega. \tag{2}$$

In (1) and (2), $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, is a bounded domain with Lipschitz boundary, T is a final time, $\varepsilon > 0$ is the diffusion coefficient, $\mathbf{b}$ the convection field, $\sigma \geq 0$ the reaction field, and f describes sinks and sources of the scalar quantity. Both problems have to be equipped with appropriate boundary conditions and the parabolic problem also with an initial condition. Problems (1) and (2) describe the transport of a scalar quantity, like temperature or concentration, by diffusion and convection. The reactive term arises in coupled problems that model, in addition to the transport, also chemical reactions, e.g., see [1, Chapt. 2.1.2]. The monograph [1] contains comprehensive explanations and extended references for all statements made in this highlight article.

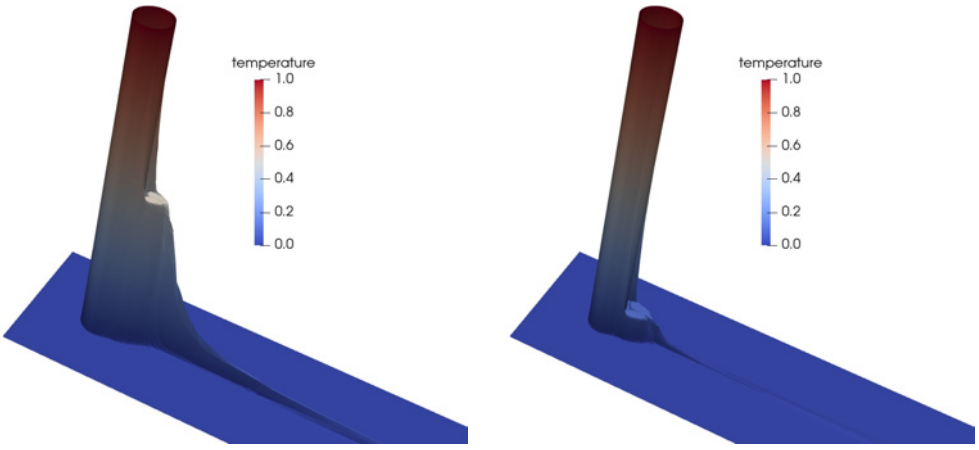


Fig. 1: *Stationary transport of temperature from a heated cylinder. The flow field is from left to right with* $\|\mathbf{b}\|_{L^\infty(\Omega)} = \mathcal{O}(1)$. *Left: Temperature distribution for* $\varepsilon = 10^{-6}$. *Right: Temperature distribution for* $\varepsilon = 10^{-8}$. *Pictures taken from [1]*

Although (1) and (2) are just linear problems, they pose challenges in their numerical solution, particularly in the convection-dominated regime. This regime is characterized by a large mesh Péclet number $\|\mathbf{b}\|_{L^\infty(\Omega)} h/\varepsilon$, where h is a characteristic scale of the mesh width. In the convection-dominated regime, solutions of (1) and (2) possess layers, compare Figure 1, which are very thin

structures with steep gradients. From asymptotic analysis it is known that layers are of size $\mathcal{O}(\varepsilon)$ or $\mathcal{O}(\sqrt{\varepsilon})$, which is, in the convection-dominated regime, usually much smaller than the affordable mesh width. Consequently, these very important structures of the solution cannot be represented on given grids. This feature is typical for multiscale problems. Hence, from the numerical point of view, problems (1) and (2) are multiscale problems in the convection-dominated regime, where the layers are the unresolved (or subgrid) scales. It is well known that the physically consistent and accurate numerical solution of multiscale problems is challenging.

Solutions of (1) and (2) satisfy maximum principles for appropriate data. These principles are of high importance from the practical point of view, since they state, e.g., that concentrations take values in $[0, 1]$ or that the temperature stays positive, and if, additionally, there are no sources of energy in Ω, the highest temperature is attained at the boundary of Ω for the steady-state problem. A discretization of (1) or (2) that is useful in practice should satisfy the discrete counterpart of the maximum principles, so-called *discrete maximum principles (DMPs)*. The satisfaction of DMPs is an important aspect of the physical consistency of a discretization. It should be noted, however, that the satisfaction of DMPs does not exclude that a numerical solution possesses small wiggles (with physically admissible values).

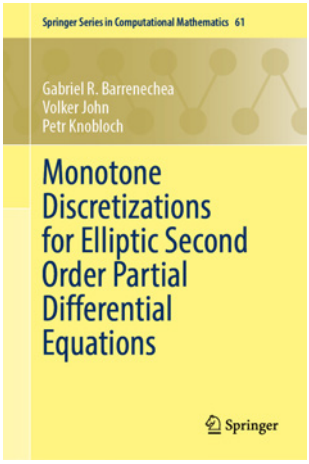


Fig. 2: *Reference [1]*

The difficulties to perform numerical simulations for convection-diffusion problems have been known for decades. Until about one decade ago, the main direction of research in Numerical Mathematics was the development of high-order methods, thus focussing on accuracy in certain norms of Sobolev and Lebesgue spaces, e.g., see the most cited monograph in this field [5]. Usually, these methods compute numerical solutions with spurious oscillations in a vicinity of layers. Early approaches for constructing DMP-preserving finite element methods, from around 1980, combine finite difference and finite volume upwind techniques with a finite element discretization of the diffusive term. Only in the first decade of this century, more sophisticated finite element methods, again adapting ideas from finite volume schemes, were proposed, e.g., in [4] for the steady-state problem (1). In the last decade, there has been an enormous development of new methods, among them the currently best performing DMP-preserving methods for (1). An overview of the state of the art of DMP-preserving finite element methods for (1) and (2) is provided in [2]. The monograph [1] contains a comprehensive presentation and numerical analysis of DMP-preserving methods for the steady-state problem (1).

This highlight article will concentrate on DMP-preserving methods for the steady-state problem (1). Since (1) is a linear boundary value problem, it seems to be natural to apply a linear discretization, i.e., a discretization that leads to a linear system of equations $A\underline{u} = \underline{g}$ with $A \in \mathbb{R}^{n\times n}, \underline{u}, \underline{g} \in \mathbb{R}^n$. If one uses Lagrange finite elements with the standard basis functions, then the vector $\underline{u}$ contains values of the finite element solution $u_h(\mathbf{x})$ at certain points in $\overline{\Omega}$, the so-called *nodes*. There is a classical theory for the satisfaction of global DMPs for the values of $\underline{u}$ from the 1970s. A widely used sufficient criterion states two conditions: The matrix A should have nonnegative row sums, and it should be a monotone matrix, i.e., A^{-1} has only nonnegative entries. A proper subset of monotone matrices is the class of M-matrices. An alternative theory is based on the concept of matrices of nonnegative type that are matrices with nonnegative row sums and with nonpositive off-diagonal entries. This concept allows also to investigate local DMPs.

The extension of a DMP property from the values in the nodes to the finite element solution is only

possible for the lowest order conforming finite elements P_1 and Q_1. For all other finite element functions, values away from the nodes cannot be bounded by the values in the nodes.

Already for the simplest case of (1), the Poisson problem, i.e., $\varepsilon = 1, \mathbf{b} = \mathbf{0}, \sigma = 0$, it turns out that the standard Galerkin discretization leads to restrictions on the mesh if the satisfaction of DMPs is desired. For P_1 finite elements in two dimensions, this restriction is almost equivalent to the requirement that the triangulation is Delaunay. In three dimensions, the set of admissible meshes is neither a subset of the class of Delaunay triangulations nor vice versa. For Q_1 and P_2 finite elements, the latter in two dimensions, global DMPs (P_2 for the values in the nodes) can be proved only for very special grids. And for P_3 finite elements in two dimensions, it can be shown that DMPs are not satisfied even on special grids. This situation for the Poisson problem explains why the development of DMP-preserving methods for (1) and (2) has been concentrated on P_1 finite elements.

For convection-dominated problems, there are several linear discretizations that satisfy DMPs or are at least monotone (positivity-preserving). One class contains the already mentioned finite element upwind methods. In the late 1980s and 1990s, there were several proposals to formulate, in two dimensions, a counterpart of the Scharfetter–Gummel finite volume scheme in the framework of finite element methods. However, these proposals have not been used in practice.

Since in the convection-dominated regime one has to deal with a multiscale problem from the numerical point of view, an appropriate strategy for computing numerical solutions that are both accurate and DMP-preserving is to apply different techniques for the resolved scales and the subgrid scales. This strategy leads to nonlinear discretizations since the distribution of the scales in the domain depends on the solution. The first nonlinear discretization was proposed already in the 1980s, the Mizukami–Hughes upwind method. It defines the local upwind direction in such a way that the discretization locally behaves like a method with a matrix of nonnegative type. This upwind direction might depend on the numerical solution. However, numerical studies in [1] show that this method is usually not competitive with modern algebraically stabilized schemes. In the 1990s, many of so-called *spurious oscillations at layers diminishing (SOLD)* or shock-capturing methods, which augment a standard linear stabilized method with a nonlinear term, have been proposed. But none of them consistently succeeds in removing the spurious oscillations of the linear method, which were solely reduced to some extent. Only the development of algebraically stabilized discretizations, with the initial contribution [4] for steady-state problems, offered an approach for computing accurate and DMP-preserving solutions for convection-dominated problems.

Most of the known algebraically stabilized methods fit in the framework of algebraic flux correction (AFC) schemes. Consider P_1 or Q_1 finite element spaces with the standard nodal basis $\{\phi_i\}_{i=1}^n$. Let, without loss of generality, the nodes be ordered such that the $(n-m)$, $m < n$, Dirichlet values come last. Denote by

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

the bilinear form of a variational formulation of (1). Then, the first step of the AFC schemes consists in assembling the matrix $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n \times n}$ of the Galerkin finite element method, i.e., $a_{ij} = a(\phi_j, \phi_i)$, $i, j = 1, \ldots, n$. Let $\underline{u} \in \mathbb{R}^n$ be the solution vector, $\underline{g} \in \mathbb{R}^m$ the vector from assembling the source term, and $\underline{u}^{\mathrm{b}} \in \mathbb{R}^{n-m}$ the vector of the Dirichlet values. Then, the general

form of an AFC scheme is given as follows:

$$\begin{aligned} \sum_{j=1}^{n} a_{ij} u_j + \sum_{j=1}^{n} b_{ij}(\underline{u})\,(u_j - u_i) &= g_i, \quad i = 1, \ldots, m, \\ u_i &= u_i^{\mathrm{b}}, \quad i = m+1, \ldots, n, \end{aligned} \tag{3}$$

with a solution-dependent matrix $B(\underline{u}) = (b_{ij}(\underline{u}))_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$ that should satisfy

$$b_{ij}(\underline{u}) = b_{ji}(\underline{u}), \quad i, j = 1, \ldots, n, \qquad b_{ii}(\underline{u}) = -\sum_{j\neq i} b_{ij}(\underline{u}) \quad i = 1, \ldots, n.$$

The symmetry of B guarantees that the AFC scheme is conservative, which is another important aspect of the physical consistency of the discretization. In AFC schemes, the term $B(\underline{u})$ is defined with the help of an artificial diffusion matrix $D \in \mathbb{R}^{n\times n}$, which can be easily computed from A by

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\} \quad \forall\, i \neq j, \quad d_{ii} = -\sum_{j\neq i} d_{ij} \quad i = 1, \ldots, n,$$

and a matrix consisting of limiters $(\alpha_{ij}(\underline{u}))_{i,j=1}^{n}$, $\alpha_{ij} \in [0, 1]$, which is assumed to be symmetric, so that

$$b_{ij}(\underline{u}) = \big(1 - \alpha_{ij}(\underline{u})\big)\, d_{ij} \quad \forall\, i \neq j, \qquad b_{ii}(\underline{u}) = -\sum_{j\neq i} b_{ij}(\underline{u}) \quad i = 1, \ldots, n.$$

Inserting this expression in (3), defining the so-called *algebraic fluxes* $f_{ij} = d_{ij}(u_j - u_i)$, $i, j = 1, \ldots, n$, and rearranging terms leads to a problem of the form

$$\begin{aligned} \sum_{j=1}^{n} (a_{ij} + d_{ij})\, u_j &= g_i + \sum_{j=1}^{n} \alpha_{ij}(\underline{u})\, f_{ij} \quad i = 1, \ldots, m, \\ u_i &= u_i^{\mathrm{b}}, \qquad\qquad\qquad\qquad\quad i = m+1, \ldots, n. \end{aligned} \tag{4}$$

By construction, the matrix on the left-hand side of (4) is an M-matrix with nonnegative row sums. With setting all limiters $\alpha_{ij}(\underline{u}) = 0$, one obtains a linear problem that corresponds to discretization with high artificial diffusion. AFC methods are designed such that limiters close to 1 are chosen in smooth subregions away from layers, to recover the high order of the Galerkin method, and limiters close to 0 are proposed in subregions including layers, so that the numerical diffusion becomes effective to suppress spurious oscillations. After the pioneering paper [4], whose proposal is nowadays called *Kuzmin limiter*, starting from 2016, a number of new and improved limiters have been designed. Most of these algebraic limiters can be computed directly from the given matrices and vectors. Thus, their implementation is independent of the dimension d of the problem.

The first comprehensive numerical analysis of AFC schemes was presented in [3]. The existence of a solution can be shown, using Brouwer's fixed-point theorem, if $\alpha_{ij}(\underline{u})(u_j - u_i)$ is a continuous function of $\underline{u}$. This property is satisfied for all known algebraic limiters. A uniqueness result is not available for any of the known limiters. As the primary goal, all limiters were designed in such a way that the discretization satisfies DMPs. For several of the more recently proposed limiters, this property holds without any restriction on an admissible mesh. The paper [3] contains also the first convergence analysis of AFC schemes. A consistency error is committed through the introduction of the algebraic stabilization. Using only that the limiters satisfy $\alpha_{ij} \in [0, 1]$ gives an error bound

for an energy norm that is augmented by a contribution from the algebraic stabilization that is of order 0.5 for the convection-dominated regime and that even does not tend to zero for the diffusion-dominated case. On the one hand, numerical studies in [3] demonstrate that both bounds are sharp within the assumptions for the analysis. But on the other hand, numerical experience shows that the order of convergence for smooth solutions is usually better. Whether or not the optimal (interpolation) order can be observed depends on the concrete limiter and on properties of the family of triangulations. The numerical analysis for providing a theoretical basis of these observations is open. Nevertheless, the numerical analysis of AFC schemes has been further developed and unified in recent years. For instance, the convergence analysis and the analysis for several algebraic limiters could be extended to general Lagrange finite elements, see [1].

From the practical point of view, it has been found that some AFC schemes cause high costs (many iterations) for solving the nonlinear discrete problems. In applications, however, convection-diffusion-reaction problems are usually a part of a coupled system, which is nonlinear anyway, so that the nonlinearity introduced by algebraically stabilized methods should not substantially increase the computational costs for solving the system.

In [1], many discretizations for (1) were studied numerically for an example that models the transport of (a concentration of) a species through $\Omega = (0,1)^2$. The solution is depicted in Figure 3. In regions with high concentration, the problem is defined to be convection-dominated, in a small neighborhood of these regions it is reaction-dominated, and away from high concentrations, it is diffusion-dominated. Simulations were performed on a family of Delaunay triangulations that were generated with a mesh generator that is freely available for academic purposes. For assessing the accuracy, reference cross-sections were computed from a numerical solution on a very fine grid. Most methods that were studied are DMP-preserving, the others at least monotone.

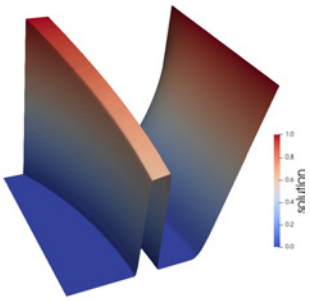
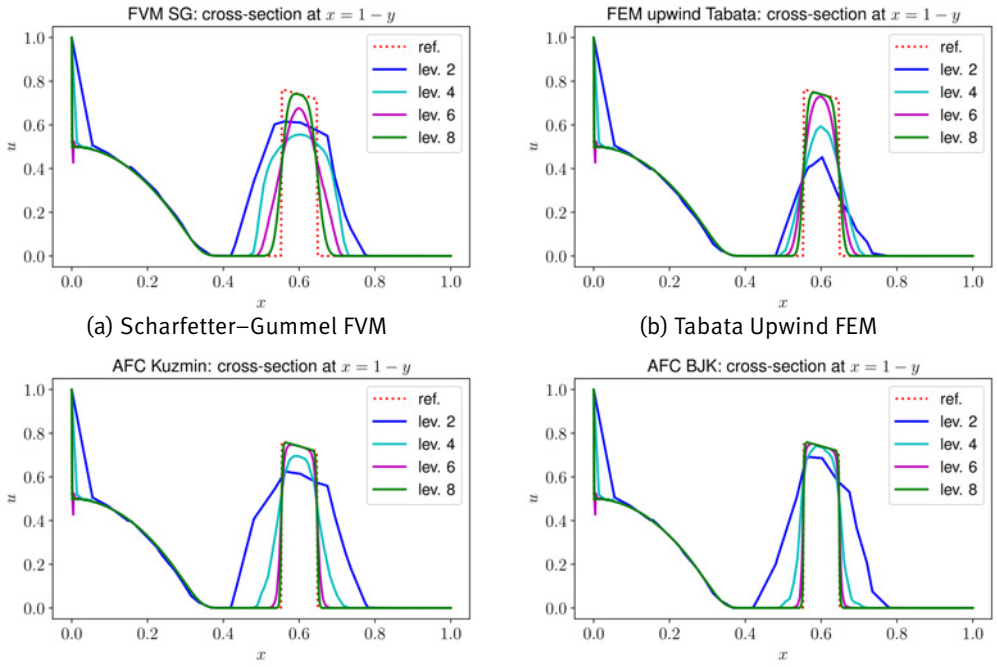


Fig. 3: *Solution of the problem that models the transport of a concentration through a domain; the inlet ($x = 0$) is on top and the outlet ($x = 1$) at the bottom*



(a) Scharfetter–Gummel FVM

(b) Tabata Upwind FEM

(c) AFC method with Kuzmin limiter [4]

(d) AFC method with BJK limiter

Fig. 4: *Problem that models the transport of a concentration through a domain; cross-sections at $x = 1 - y$ for two linear discretizations (top) and two AFC schemes (bottom) and for different levels of mesh refinement. Pictures taken from [1]*

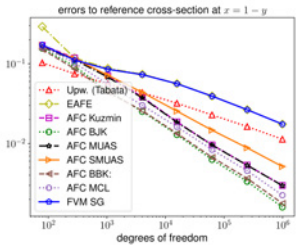


Fig. 5: *Problem that models the transport of a concentration through a domain; errors for cross-section at $x = 1 - y$. Picture taken from [1]*

None of the studied methods computed unphysical values, i.e., values not contained in $[0, 1]$. The evaluation with respect to accuracy was qualitatively the same for all cross-sections. Exemplarily, results for the cross-section $x = 1 - y$ are presented in Figures 4 and 5. As expected, the nonlinear discretizations gave more accurate solutions where the layers are usually much sharper than for the linear discretizations. It can be also observed that within both the class of linear discretizations and AFC schemes, there are noticeable differences in the accuracy for different methods. The most important conclusion from this numerical study is that the AFC schemes are well suited for the solution of convection-diffusion-reaction problems where different regimes are present in different parts of the domain.

Considering the following three aspects: order of convergence for smooth solutions, steepness of layers in numerical solutions, and efficiency for solving the nonlinear problems, there is currently no AFC method that performs best with respect to all of these aspects.

The concept of algebraic stabilizations can also be used within finite element methods for the time-dependent problem (2). These methods are called *FEM-FCT (flux-corrected transport) schemes*. There are even linear variants of FEM-FCT schemes that have been proved to be a good compromise of accuracy and efficiency in several application projects of RG 3.

Members of RG 3 have been working on different aspects of algebraically stabilized methods in recent years, often with collaborators. Two new limiters (BJK (Barrenechea, John, and Knobloch) and MUAS (Monotone Upwind-type Algebraically Stabilized) in Figures 4 and 5) were developed. Both of them satisfy DMPs for arbitrary elliptic second order problems on arbitrary simplicial triangulations. Approaches for solving the nonlinear problems were identified that are more efficient than previously used ones. An a posteriori error estimator for AFC methods was developed. It was clarified how AFC methods have to be applied on adaptively refined grids with hanging nodes. Comprehensive numerical studies provided a better insight in the advantages and shortcomings of several methods. And finally, for algebraic stabilizations of the parabolic problem (2), results concerning the existence and uniqueness of a discrete solution were obtained.

References

[1] G.B. Barrenechea, V. John, P. Knobloch, *Monotone Discretizations for Elliptic Second Order Partial Differential Equations*, vol. **61** of Springer Series in Computational Mathematics, Springer, Cham, 2025.

[2] ———, *Finite element methods respecting the discrete maximum principle for convection-diffusion equations*, SIAM Rev., **66** (2024), pp. 3–88.

[3] ———, *Analysis of algebraic flux correction schemes*, SIAM J. Numer. Anal., **54** (2016), pp. 2427–2451.

[4] D. Kuzmin, *Algebraic flux correction for finite element discretizations of coupled systems*, in: Proceedings of the Int. Conf. on Computational Methods for Coupled Problems in Science and Engineering, Barcelona, M. Papadrakakis, E. Oñate, B. Schrefler eds., CIMNE, 2007, pp. 653–656.

[5] H.-G. Roos, M. Stynes, L. Tobiska, *Robust Numerical Methods for Singularly Perturbed Differential Equations*, vol. 24 of Springer Series in Computational Mathematics, Springer, Berlin, 2008.