A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems on arbitrary meshes. ⋆

D. Kuzmin [a,*] , M. J. Shashkov [b], D. Svyatskiy [b]

[a] *Institute of Applied Mathematics, Dortmund University of Technology Vogelpothsweg 87, D-44227, Dortmund, Germany*

[b] *Mathematical Modeling and Analysis Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545*

Abstract

Nonlinear constrained finite element approximations to anisotropic diffusion problems are considered. Starting with a standard (linear or bilinear) Galerkin discretization, the entries of the stiffness matrix are adjusted so as to enforce sufficient conditions of the discrete maximum principle (DMP). An algebraic splitting is employed to separate the contributions of negative and positive off-diagonal coefficients which are associated with diffusive and antidiffusive numerical fluxes, respectively. In order to prevent the formation of spurious undershoots and overshoots, a symmetric slope limiter is designed for the antidiffusive part. The corresponding upper and lower bounds are defined using an estimate of the steepest gradient in terms of the maximum and minimum solution values at surrounding nodes. The recovery of nodal gradients is performed by means of a lumped-mass L_2 projection. The proposed slope limiting strategy preserves the consistency of the underlying discrete problem and the structure of the stiffness matrix (symmetry, zero row and column sums). A positivity-preserving defect correction scheme is devised for the nonlinear algebraic system to be solved. Numerical results and a grid convergence study are presented for a number of anisotropic diffusion problems in two space dimensions.

Key words: anisotropic diffusion, discrete maximum principle, nonnegativity constraints, finite element method, gradient recovery, slope limiting
1991 MSC: 35B50, 35J25 65N30, 65N22

⋆ LA-UR 08-05077

* Corresponding author.
Email addresses: kuzmin@math.uni-dortmund.de (D. Kuzmin), shashkov@lanl.gov (M. J. Shashkov), dasvyat@lanl.gov (D. Svyatskiy).

1 Introduction

Some partial differential equations of elliptic type are known to possess a remarkable property: if the source term is nonpositive, then the solution attains its maximum on the boundary. This statement is known as the *maximum principle* and provides a useful criterion for the analysis of numerical approximations. A finite element scheme that does not generate spurious extrema in the interior of the domain is said to satisfy the *discrete maximum principle* (DMP) which makes it possible to prove uniform convergence of an approximate solution to the exact one [4]. Sufficient conditions of the DMP can be formulated using the concept of *monotone operators* and, in particular, *M-matrices* which play an important role in numerical linear algebra [3,23]. The inverse of a monotone matrix is nonnegative, and so is the solution of the algebraic system for any nonnegative right-hand side. A numerical scheme that enjoys this property is called *positivity-preserving*. The DMP criterion is more stringent since it requires that the row sums of the stiffness matrix be zero for all interior nodes.

Monotonicity constraints are difficult to satisfy at the discrete level. In many cases, the above sufficient conditions impose severe restrictions on the choice of basis functions and on the geometric properties of the mesh. For a triangulation of *acute* or *nonobtuse* type (all angles smaller than or equal to $\pi/2$) the piecewise-linear finite element approximation of the Poisson equation produces a discrete operator that proves to be an M-matrix [4,12,13]. In the case of bilinear finite elements, it is sufficient to require that all quadrilaterals be of *nonnarrow* type (aspect ratios smaller than or equal to $\sqrt{2}$ [6]). However, the geometric approach to DMP verification fails in the case of higher-order finite elements [9], singularly perturbed convection-diffusion equations [15], and anisotropic diffusion problems [17]. As a consequence, nonphysical local extrema can and do occur when a steep gradient cannot be resolved properly on a given mesh.

An ideal discretization must be conservative, consistent, and more than first-order accurate for smooth data. Moreover, it must satisfy the DMP on arbitrary meshes, even if convective effects are strong and/or the diffusion tensor is heterogeneous/anisotropic [17]. Unfortunately, no linear scheme can meet all of these requirements. A possible remedy is to adjust the coefficients of the discrete problem in an adaptive way. In the case of hyperbolic conservation laws, this can be accomplished using flux/slope *limiters* to achieve monotonicity. Many nonlinear high-resolution schemes are based on this design principle. However, the idea of blending linear approximations of high and low order cannot be directly transferred to the case of anisotropic diffusion problems since the structure of numerical fluxes is different from the hyperbolic case.

Flux limiting for diffusion operators was addressed in a number of recent publications [2,10,24] but many open questions remain. Monotone finite volume schemes for anisotropic diffusion problems are available [16,17] but they are merely positivity-preserving and, generally, do not satisfy the DMP. Moreover, their derivation is based on a design philosophy which is not applicable to finite element approximations. The

method proposed in [18] is based on constrained optimization and requires *a priori* knowledge of the solution bounds. The solution of this constrained optimization problem can be very expensive as the number of unknowns increases.

The *algebraic flux correction* paradigm described in [14,15] provides a general framework for the design of monotone discretizations on unstructured meshes. In the present paper, we extend this methodology to anisotropic diffusion problems and enforce the DMP using a combination of algebraic and geometric criteria. The new algorithm constrains an antidiffusive flux using the coefficients of a discrete gradient to derive the upper and lower bounds. The structure of these bounds resembles that for *flux-corrected transport* (FCT) algorithms [15,25] but the limiting process is intended to control a local slope, as in the case of *symmetric limited positive* (SLIP) schemes [11].

The outline of this paper is as follows. In Section 2, we formulate the continuous maximum principle to be emulated at the discrete level. The Galerkin finite element discretization and the discrete maximum principle are dealt with in Sections 3 and 4, respectively. The algebraic splitting described in Section 5 leads to the new slope limiting technique introduced in Section 6. A positivity-preserving defect correction scheme for the iterative treatment of nonlinear antidiffusive terms is proposed in Section 7. Numerical experiments and grid convergence studies for several 2D test problems are performed in Section 8 which is followed by a summary and conclusions.

2 Continuous problem

We consider a mathematical model that describes the steady diffusive transport of a generic scalar $u(\mathbf{x})$ in a bounded domain $\Omega \subset \mathbb{R}^{\mathrm{D}}$, where $\mathrm{D} = 2$ or 3, with piecewise smooth boundary Γ. The rate of transport is given by the vector-valued flux function

$$\mathbf{v}(\mathbf{x}) = -\mathcal{D}(\mathbf{x})\nabla u(\mathbf{x}), \quad \forall \mathbf{x} \in \Omega, \tag{1}$$

where $\mathcal{D}(\mathbf{x}) = \{d_{ij}\}$ is a piecewise-constant (possibly anisotropic) diffusion tensor. In the context of flows in porous media, the above formula is known as the *Darcy law* in which the functions $\mathbf{v}$ and u represent the velocity and pressure fields, respectively.

The divergence of $\mathbf{v}$ is associated with a linear differential operator $\mathcal{L}$ defined by

$$\mathcal{L}u(\mathbf{x}) = \nabla \cdot \mathbf{v} = -\nabla \cdot (\mathcal{D}\nabla u), \quad \forall \mathbf{x} \in \Omega. \tag{2}$$

In D space dimensions, the above formula can be written in the equivalent form

$$\mathcal{L}u(\mathbf{x}) = -\sum_{i=1}^{\mathrm{D}} \frac{\partial}{\partial x_i} \left(\sum_{j=1}^{\mathrm{D}} d_{ij} \frac{\partial u}{\partial x_j} \right), \quad \forall \mathbf{x} \in \Omega. \tag{3}$$

The so-defined operator $\mathcal{L}$ is called *uniformly elliptic* in the domain Ω if the diffusion tensor $\mathcal{D}(\mathbf{x})$ is symmetric and positive definite at each point $\mathbf{x} \in \Omega$. Under this assumption, the following *maximum principle* [21] holds for any $u \in C^2(\Omega) \cap C(\bar{\Omega})$

$$\mathcal{L}u(\mathbf{x}) \leq 0, \quad \forall \mathbf{x} \in \Omega \quad \Rightarrow \quad \max_{\mathbf{x}\in\bar{\Omega}} u(\mathbf{x}) = \max_{\mathbf{x}\in\Gamma} u(\mathbf{x}). \tag{4}$$

The basic idea of a proof is as follows [7,13]. Let $u_\Gamma^{\max} = \max_{\mathbf{x}\in\Gamma} u(\mathbf{x})$ and introduce $w = \max\{u - u_\Gamma^{\max}, 0\}$. Consider a subdomain $\Omega_* \subset \Omega$ in which $w = u - u_\Gamma^{\max} \geq 0$ and $w = 0$ on the boundary. Integration by parts using Green's formula yields

$$\int_{\Omega_*} w(\mathcal{L}u)\, dx = \int_{\Omega_*} \nabla w \cdot (\mathcal{D}\nabla u)\, dx = \int_{\Omega_*} \nabla w \cdot (\mathcal{D}\nabla w)\, dx. \tag{5}$$

Since $\mathcal{L}u \leq 0$ and $w \geq 0$ in Ω, the left-hand side of this relation cannot be positive. On the other hand, the right-hand side is nonnegative due to the positive definiteness of $\mathcal{D}$. This can only be the case if $w \equiv 0$, whence $u(\mathbf{x}) \leq u_\Gamma^{\max}$, $\forall \mathbf{x} \in \bar{\Omega}$. $\square$

Applying (4) to the negative of u, one obtains the complementary *minimum principle*

$$\mathcal{L}u(\mathbf{x}) \geq 0, \quad \forall \mathbf{x} \in \Omega \quad \Rightarrow \quad \min_{\mathbf{x}\in\bar{\Omega}} u(\mathbf{x}) = \min_{\mathbf{x}\in\Gamma} u(\mathbf{x}). \tag{6}$$

Continuous maximum and minimum principles, as stated above, are valid for second-order elliptic problems with Dirichlet or mixed boundary conditions [13]. In the former case, they give an *a priori* estimate of $u \in C^2(\Omega) \cap C(\bar{\Omega})$ in terms of the prescribed boundary data $g \in C^0(\Gamma)$. The corresponding boundary value problem is given by

$$\begin{cases} -\nabla \cdot (\mathcal{D}\nabla u) = q, & \text{in } \Omega, \\ \qquad\qquad u = g & \text{on } \Gamma, \end{cases} \tag{7}$$

where q is a given function that represents a volumetric source or sink of u. By virtue of (4) and (6), the maximum and minimum principles can be formulated as follows.

Theorem 1. *The solution $u \in C^2(\Omega) \cap C(\bar{\Omega})$ of problem (7) attains its maxima (minima) on the boundary Γ if q is nonpositive (nonnegative) in Ω*

$$q \leq 0 \quad \Rightarrow \quad \max_{\mathbf{x}\in\bar{\Omega}} u(\mathbf{x}) = \max_{\mathbf{x}\in\Gamma} g(\mathbf{x}), \tag{8}$$

$$q \geq 0 \quad \Rightarrow \quad \min_{\mathbf{x}\in\bar{\Omega}} u(\mathbf{x}) = \min_{\mathbf{x}\in\Gamma} g(\mathbf{x}). \tag{9}$$

Traditionally, Theorem 1 is called the maximum principle, regardless of the sign of q.

Corollary 1. *If $q = 0$ then the solution $u \in C^2(\Omega) \cap C(\bar{\Omega})$ of (7) is bounded by*

$$q = 0 \quad \Rightarrow \quad \min_{\mathbf{x}\in\Gamma} g(\mathbf{x}) \leq u(\mathbf{x}) \leq \max_{\mathbf{x}\in\Gamma} g(\mathbf{x}), \qquad \forall \mathbf{x} \in \Omega. \tag{10}$$

Corollary 2. *The solution $u \in C^2(\Omega) \cap C(\bar{\Omega})$ of problem (7) preserves the sign of the boundary data $g \in C^0(\Gamma)$ if q and g are nonpositive (nonnegative) in Ω*

$$q \leq 0, \quad g \leq 0 \quad \Rightarrow \quad u \leq 0 \quad \text{in } \Omega, \tag{11}$$
$$q \geq 0, \quad g \geq 0 \quad \Rightarrow \quad u \geq 0 \quad \text{in } \Omega. \tag{12}$$

This criterion is commonly referred to as *nonnegativity* or *positivity preservation*. It ensures that positive sources cannot create negative concentrations or temperatures. In the case of steady heat conduction with $q = 0$, a nonuniform temperature distribution can only be maintained by supplying and removing some heat on the boundary.

A similar maximum principle can be formulated for elliptic equations equipped with mixed boundary conditions of Dirichlet-Neumann type, whereby the diffusive flux is prescribed on a boundary part Γ_N. For further information on maximum principles for elliptic boundary value problems we refer to [3,12,13,21].

3 Finite element discretization

The finite element method is based on a weak form of the continuous problem

$$a(u, w) = (q, w), \tag{13}$$

where $a(\cdot, \cdot)$ is a bilinear form, u is the weak solution, w is any admissible test function, and $(\cdot, \cdot)$ is the usual shorthand notation for the scalar product in $L_2(\Omega)$

$$(u, w) = \int_\Omega u(\mathbf{x}) w(\mathbf{x})\, dx. \tag{14}$$

The bilinear form associated with the weak form of the diffusion equation reads

$$a(u, w) = \int_\Omega \nabla w \cdot (\mathcal{D}\nabla u)\, d\mathbf{x} - \int_\Gamma w\, \mathbf{n} \cdot (\mathcal{D}\nabla u)\, ds, \tag{15}$$

where $\mathbf{n}$ denotes the unit outward normal to Γ. Let $u \in V = \{v \in H^1(\Omega) \mid u = g \text{ on } \Gamma\}$ and $w \in W = \{w \in H^1(\Omega) \mid w = 0 \text{ on } \Gamma\}$. Due to the requirement that $w = 0$ on Γ,

the contribution of the surface integral to (15) vanishes, whence

$$a(u,w) = \int_\Omega \nabla w \cdot (\mathcal{D}\nabla u)\, d\mathbf{x}. \tag{16}$$

Let $\{\varphi_i\}$ be a set of piecewise-polynomial basis functions spanning a finite-dimensional subspace of V. Then an approximate solution u_h of problem (13) is given by

$$u_h(\mathbf{x}) = \sum_j u_j \varphi_j(\mathbf{x}). \tag{17}$$

Using $u = u_h$ and $w = \varphi_i$ in (13), one obtains a system of linear algebraic equations

$$\sum_j a(\varphi_j, \varphi_i) u_j = (\varphi_i, q), \qquad \forall i. \tag{18}$$

This linear system can be written in matrix form as $Au = b$, where $A = \{a_{ij}\}$ is the global *stiffness matrix* and $b = \{b_i\}$ is the *load vector* with coefficients

$$a_{ij} = a(\varphi_j, \varphi_i), \qquad b_i = (\varphi_i, q). \tag{19}$$

Consistency requires that the sum of basis functions be equal to 1 at every point

$$\sum_j \varphi_j(\mathbf{x}) = 1, \qquad \forall \mathbf{x} \in \bar{\Omega}. \tag{20}$$

Due to the fact that $a(1, \varphi_i) \equiv 0$, the stiffness matrix A has zero row sums

$$\sum_j a_{ij} = 0, \qquad \forall i \tag{21}$$

and the vector of nodal values $u = \{u_i\}$ is defined up to an arbitrary additive constant. The uniqueness of the solution to the discrete problem is provided by the Dirichlet boundary conditions to be implemented in a strong sense as explained below.

4 Discrete maximum principle

Consider an arbitrary discretization of the diffusion equation that can be expressed in terms of m discrete nodal values $u_i,\ i = 1, \ldots, m$. For the time being, we do not distinguish between interior and boundary nodes. That is, the global $m \times m$ stiffness matrix $A = \{a_{ij}\}$ is assembled as if Neumann boundary conditions were prescribed.

If the coefficients of A are given by (16) and (19), then this sparse matrix is symmetric positive definite with zero row and column sums. Moreover, it is *irreducible*, which means that its directed graph is strongly connected ([23], p. 20). In other words, for any pair of indices i and j there exists a sequence of nonzero matrix entries

$$a_{ik_1} \neq 0,\ a_{k_1k_2} \neq 0, \ldots,\ a_{k_sj} \neq 0. \tag{22}$$

This property is dictated by the nature of elliptic problems in which a disturbance introduced at a single point may affect the solution in the whole domain.

The sparsity pattern of A depends on the underlying mesh, on the choice of basis functions, and on the numbering of nodes. Let the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be located in the interior of the domain Ω and $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_m$ lie on the boundary Γ. This convention implies that A and $u = [u_1, \ldots, u_m]^T$ can be partitioned in block form as

$$A = \begin{bmatrix} A_{\Omega\Omega} & A_{\Omega\Gamma} \\ A_{\Gamma\Omega} & A_{\Gamma\Gamma} \end{bmatrix}, \quad u = \begin{bmatrix} u_\Omega \\ u_\Gamma \end{bmatrix}. \tag{23}$$

The subscripts Ω and Γ refer to row/column numbers in the range $1, \ldots, n$ and $n+1, \ldots, m$, respectively. For instance, $u_\Omega = [u_1, \ldots, u_n]^T$ is the vector of unknowns, whereas $u_\Gamma = [u_{n+1}, \ldots, u_m]^T$ is the vector of boundary values such that

$$u_\Gamma = g. \tag{24}$$

In view of (24), the algebraic system for the vector of unknowns u_Ω can be written as

$$A_{\Omega\Omega} u_\Omega = b_\Omega - A_{\Omega\Gamma} g, \tag{25}$$

where b_Ω denotes the contribution of the source/sink q to equations for internal points.

In a practical implementation, it is convenient to build the Dirichlet boundary conditions into the global stiffness matrix A and solve the extended linear system

$$\bar{A} u = b, \tag{26}$$

where the $m \times m$ matrix $\bar{A}$ and the right-hand side b are defined as

$$\bar{A} = \begin{bmatrix} A_{\Omega\Omega} & A_{\Omega\Gamma} \\ 0 & I \end{bmatrix}, \quad b = \begin{bmatrix} b_\Omega \\ g \end{bmatrix}. \tag{27}$$

The first n rows of $\bar{A}$ are the same as those of A, whereas the last $m - n$ rows are replaced by those of an identity matrix. If the block $A_{\Omega\Omega}$ is nonsingular, then so is $\bar{A}$.

A nonsingular matrix $\bar{A}$ is said to be *monotone* if $\bar{A}^{-1} \geq 0$. Here and below, inequalities are meant to hold elementwise for all indices, unless a range of relevant index values is specified explicitly. By virtue of (27), the inverse of $\bar{A}$ is given by

$$\bar{A}^{-1} = \begin{bmatrix} A_{\Omega\Omega}^{-1} & -A_{\Omega\Omega}^{-1} A_{\Omega\Gamma} \\ 0 & I \end{bmatrix} \tag{28}$$

and has no negative entries if $A_{\Omega\Omega}^{-1} \geq 0$ and $A_{\Omega\Gamma} \leq 0$. If $\bar{A}$ is monotone and $\bar{a}_{ij} \leq 0$ for all $j \neq i$, then $\bar{A}$ is called an *M-matrix*. Any matrix obtained by setting certain off-diagonal entries of $\bar{A}$ to zero is also an M-matrix ([23], p. 85).

Clearly, it is usually not feasible to calculate the inverse of $\bar{A}$ or $A_{\Omega\Omega}$ and check the sign of its coefficients. Instead, the following well-known set of sufficient conditions can be used to prove monotonicity and the M-matrix property ([23], p. 85).

Theorem 2. *If $A_{\Omega\Omega}$ is an irreducibly diagonally dominant $n \times n$ matrix with $a_{ii} > 0$ for all $i = 1, \ldots, n$ and $a_{ij} \leq 0$ for all $i \neq j$, then $A_{\Omega\Omega}^{-1} > 0$.*

Note that the last inequality is strict, although $A_{\Omega\Omega}^{-1} \geq 0$ would suffice for $A_{\Omega\Omega}$ to be an M-matrix. By definition of an irreducibly diagonally dominant matrix ([23], p. 23), it is supposed to be irreducible and $|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$ for all $i = 1, \ldots, n$, with strict inequality for at least one i. Since $a_{ij} \leq 0$ for all $i \neq j$, the row sums of $A_{\Omega\Omega}$ should be nonnegative ($\sum_j a_{ij} \geq 0$) and strictly positive ($\sum_j a_{ij} > 0$) for at least one row i.

Corollary 3. *Under the conditions of Theorem 2, the block $A_{\Omega\Omega}$ is an M-matrix and*

$$A_{\Omega\Omega} u_\Omega \geq 0 \quad \Rightarrow \quad u_\Omega \geq 0. \tag{29}$$

The M-matrix property of $A_{\Omega\Omega}$ ensures fast convergence of iterative solvers and makes it possible to prove a discrete counterpart of the *a priori* estimates (8)–(9).

We will say that the *discrete maximum principle* (DMP) holds for problem (26) if

$$b \geq 0 \quad \Rightarrow \quad u \geq 0, \tag{30}$$

whereas for $b_\Omega \leq 0$ a positive maximum is attained on the boundary [3,5,13], i.e.,

$$\max_i u_i \leq \max_j \{0, g_j\}. \tag{31}$$

In a similar vein, a negative minimum is required to occur on Γ if $b_\Omega \geq 0$. If there are no sources or sinks ($b_\Omega = 0$), then the nodal values are bounded by

$$\min_j \{0, g_j\} \leq u_i \leq \max_j \{0, g_j\}, \qquad \forall i. \tag{32}$$

In the case of linear and bilinear finite elements, the interpolant u_h satisfies a local maximum principle in each cell, i.e., it is bounded by the nodal values associated with the vertices of the mesh. Therefore, the DMP for nodal values implies that for u_h.

Many other definitions of DMP can be found in the literature. One of the most general formulations, which includes (31) as a special case, is as follows [22]

$$\max_i u_i \le \max_{j\in N_+} u_j, \tag{33}$$

where $N_+ = \{1 \le j \le m \mid b_j > 0\}$. The right-hand side is taken to be zero if $N_+ = \emptyset$.

Theorem 3. *The discrete maximum principle holds if* $A_{\Omega\Omega}^{-1} \ge 0$, $A_{\Omega\Gamma} \le 0$, *and*

$$\sum_{j=1}^{m} a_{ij} \ge 0, \qquad i = 1, \ldots, n. \tag{34}$$

A proof of this theorem is as follows. Due to (28), the discrete operator $\bar{A}$ enjoys the M-matrix property under the above conditions. It follows that $\bar{A}^{-1} \ge 0$ and

$$b \ge 0 \quad \Rightarrow \quad u = \bar{A}^{-1} b \ge 0. \tag{35}$$

If $b_\Omega \le 0$, let $u_\Gamma^{\max} = \max_j\{0, g_j\}$, take $w = u - u_\Gamma^{\max}$, and invoke (34) to prove that

$$\sum_{j=1}^{m} a_{ij} w_j = \sum_{j=1}^{m} a_{ij} u_j - u_\Gamma^{\max} \sum_{j=1}^{m} a_{ij} = b_i - u_\Gamma^{\max} \sum_{j=1}^{m} a_{ij} \le 0, \quad \forall i = 1, \ldots, n. \tag{36}$$

Since $A_{\Omega\Omega}$ is monotone, it follows that $w_\Omega + A_{\Omega\Omega}^{-1} A_{\Omega\Gamma} w_\Gamma \le 0$. Furthermore, $w_\Gamma \le 0$ and $A_{\Omega\Gamma} \le 0$, so $w_\Omega \le 0$, which implies (31). For a proof of (33) we refer to [22]. $\square$

If the first n rows of the global matrix A have zero row sums, then we can take $w = u - \max_j g_j$. Arguing as above, we deduce that $w_\Omega = A_{\Omega\Omega}^{-1}[b_\Omega - A_{\Omega\Gamma} w_\Gamma]$ and, consequently, a discrete counterpart of *a priori* estimate (8) holds for $b_\Omega \le 0$

$$u_i \le \max_j g_j, \qquad \forall i = 1, \ldots, n. \tag{37}$$

The corresponding minimum principle for $b_\Omega \ge 0$ can be proved in the same way.

Theorem 4. *The solution of (26) satisfies (37) if* $b_\Omega \le 0$ *and the first* n *rows of the irreducible stiffness matrix* A *satisfy the following sufficient conditions*

- diagonal coefficients are strictly positive

$$a_{ii} > 0, \qquad \forall i = 1, \ldots, n, \tag{38}$$

- off-diagonal coefficients are nonpositive

$$a_{ij} \le 0, \qquad \forall j \ne i, \quad 1 \le j \le m, \tag{39}$$

- the first n row sums are equal to zero

$$\sum_{j=1}^{m} a_{ij} = 0, \qquad \forall i = 1, \ldots, n. \tag{40}$$

The proof follows from Theorems 2 and 3 reinforced by condition (40) which implies weak diagonal dominance and consistency. If $g = const$ and $b_\Omega = 0$, then $u \equiv g$ and the exact solution of the boundary value problem (7) is recovered in this case.

Conditions (38)–(40) are rather restrictive but provide a useful tool for the analysis and design of numerical schemes. In the case of linear and bilinear finite element approximations of the Poisson equation, they are satisfied automatically on a suitably designed mesh (triangles of *acute/nonobtuse* type, quadrilaterals of *nonnarrow* type [6,12,13]). However, the involved geometric constraints turn out too restrictive in the case of problem (7) with an anisotropic diffusion tensor. If the stiffness matrix A fails to satisfy (38)–(40) on a given mesh, then a violation of the discrete maximum principle is possible. As a rule, it originates in regions in which the gradients of the solution are steep and not aligned with the orientation of mesh edges, so that the discretization method is unable to capture them properly. On the other hand, no spurious maxima or minima are generated if the numerical solution is well-resolved.

5 Algebraic splitting

In the present paper, we treat the sufficient DMP conditions (39)–(40) as algebraic constraints to be imposed at the discrete level. The basic idea is to check the coefficients of A and adjust them, if necessary, subject to the following design principles

- perturbation to the residual of (26) admits a conservative flux decomposition;
- the discrete maximum principle holds for the solution of the perturbed system;

The methodology to be presented is based on the *algebraic flux correction* paradigm that was originally developed for convection-dominated transport problems [15].

As already mentioned, the Galerkin approximation based on (16) and (19) gives rise to a symmetric stiffness matrix $A = \{a_{ij}\}$ with zero row and column sums

$$a_{ij} = a_{ji}, \qquad \sum_i a_{ij} = \sum_j a_{ij} = 0. \tag{41}$$

The coefficients of this matrix satisfy (40) but may violate conditions (38) and (39) if the computational mesh and/or the diffusion tensor $\mathcal{D}$ are anisotropic.

Following Stoyan [22], we split A so as to extract the 'bad' part $A^+ = \{a_{ij}^+\}$ with

$$a_{ij}^+ := \max\{0, a_{ij}\}, \qquad \forall j \neq i. \tag{42}$$

The diagonal coefficients of A^+ are defined so that it has zero row and column sums

$$a_{ii}^+ := -\sum_{j\neq i} a_{ij}^+. \tag{43}$$

The complement $A^- := A - A^+$ represents the 'good' part of the stiffness matrix

$$A = A^+ + A^-. \tag{44}$$

By virtue of (41)–(43), the $i-$th element of the vector $A^{\pm}u$ is given by

$$(A^{\pm}u)_i = \sum_j a_{ij}^{\pm} u_j = \sum_{j\neq i} a_{ij}^{\pm}(u_j - u_i), \qquad \forall i = 1, \ldots, n \tag{45}$$

and can be expressed in terms of numerical fluxes from one node into another

$$(A^{\pm}u)_i = -\sum_{j\neq i} f_{ij}^{\pm}, \qquad f_{ij}^{\pm} = a_{ij}^{\pm}(u_i - u_j). \tag{46}$$

Adopting the convention that is commonly used in the context of 1D conservation laws, we will call a flux of the form $f_{ij} = a_{ij}(u_i - u_j)$ *diffusive* if $a_{ij} \leq 0$ and *antidiffusive* if $a_{ij} \geq 0$. The fluxes $f_{ij}^{\pm}$ and $f_{ji}^{\pm}$ have the same magnitude and opposite signs

$$f_{ij}^{\pm} + f_{ji}^{\pm} = 0. \tag{47}$$

This property provides discrete conservation. Every pair of fluxes can be associated with an edge of the graph of A, i.e., with a pair of nonzero off-diagonal coefficients.

The $i-$th element of the residual vector $r = b - Au$ admits the representation

$$r_i = b_i + \sum_{j\neq i}[f_{ij}^+ + f_{ij}^-], \qquad i = 1, \ldots, n. \tag{48}$$

In order to enforce the discrete maximum principle, the contribution of positive off-diagonal coefficients to the residual vector (48) may need to be reduced. To this end, every raw antidiffusive flux f_{ij}^+ is replaced by its limited counterpart

$$\bar{f}_{ij}^+ = \alpha_{ij} a_{ij}^+ (u_i - u_j), \qquad 0 \leq \alpha_{ij} \leq 1. \tag{49}$$

The solution-dependent correction factors $\alpha_{ij} = \alpha_{ji}$ should be chosen so that there exist a pair of nonnegative coefficients $\beta_{ij} \geq 0$ and $\beta_{ji} \geq 0$ such that

$$\bar{f}_{ij}^{+} = \beta_{ij}(u_k - u_i), \qquad \bar{f}_{ji}^{+} = \beta_{ji}(u_l - u_j) \tag{50}$$

for a pair of nodes $k \neq i$ and $l \neq j$. In essence, this representation means that the limited antidiffusive fluxes $\bar{f}_{ij}^{+}$ and $\bar{f}_{ji}^{+}$ have the same effect as diffusive fluxes from other nodes. Assuming that the modified matrix remains irreducible, criterion (50) implies the discrete maximum principle for the solution of the perturbed system. A practical approach to the construction of skew-symmetric fluxes $\bar{f}_{ij}^{+} = -\bar{f}_{ji}^{+}$ that satisfy the above monotonicity constraint is presented in the next section.

In essence, the correction step (49) corresponds to the following modification of A^{+}

$$\bar{a}_{ii}^{+} := -\sum_{j \neq i} \bar{a}_{ij}^{+}, \qquad \bar{a}_{ij}^{+} = \alpha_{ij} a_{ij}^{+}, \qquad \forall j \neq i. \tag{51}$$

The original matrix A^{+} is recovered for $\alpha_{ij} \equiv 1$, whence the correction factors should be as close to 1 as possible without violating the discrete maximum principle. The flux-corrected scheme remains consistent if $\alpha_{ij} \to 1$ as the mesh size h goes to zero.

6 Slope limiting

Slope limiting amounts to reducing the magnitude of the solution difference $u_i - u_j$ so as to enforce the DMP. Let $\bar{s}_{ij}$ be a limited counterpart of $u_i - u_j$ such that

$$f_{ij}^{+} = a_{ij}^{+} \bar{s}_{ij}, \qquad \bar{s}_{ij} = \alpha_{ij}(u_i - u_j). \tag{52}$$

To find the right value of $\bar{s}_{ij}$, we need the (approximate) solution gradients at nodes i and j. In the case of linear or bilinear finite elements, a continuous approximation to nodal gradients can be obtained, e.g., using the lumped-mass L_2-projection, cf. [19]

$$\hat{\nabla} u_i = \frac{1}{m_i} \sum_k \mathbf{c}_{ik} u_k, \tag{53}$$

where $m_i = (1, \varphi_i)$ is a diagonal entry of the lumped mass matrix and $\mathbf{c}_{ik} = (\varphi_i, \nabla \varphi_k)$.

If the degrees of freedom are associated with edges/faces of the computational mesh, then the nodal gradients can be recovered in a similar way, e.g., using the Crouzeix-Raviart / Rannacher-Turek elements to approximate ∇u. Alternatively, they can be defined as an average of the mean values for the two cells sharing the edge/face [19].

Due to (20) the gradients of basis functions have zero sum at every point $\mathbf{x} \in \bar{\Omega}$, so that $\mathbf{c}_{ii} = -\sum_{k \neq i} \mathbf{c}_{ik}$ and the right-hand side of (53) can be written as

$$\hat{\nabla} u_i = \frac{1}{m_i} \sum_{k \neq i} \mathbf{c}_{ik} (u_k - u_i). \tag{54}$$

For any pair of nodes i and j, a usable approximation to the difference $u_i - u_j$ is

$$s_{ij} = (\mathbf{x}_i - \mathbf{x}_j) \cdot \hat{\nabla} u_i. \tag{55}$$

Introducing the maximum and minimum nodal values of u over the stencil of i

$$u_i^{\max} = \max\{0, u_i + \max_{j \neq i}(u_j - u_i)\}, \tag{56}$$

$$u_i^{\min} = \min\{0, u_i + \min_{j \neq i}(u_j - u_i)\}, \tag{57}$$

the extrapolated slope s_{ij} can be estimated in terms of the steepest gradients

$$\mu_{ij}(u_i^{\min} - u_i) \leq s_{ij} \leq \mu_{ij}(u_i^{\max} - u_i), \tag{58}$$

$$\mu_{ij} = \frac{1}{m_i} \sum_{k \neq i} |\mathbf{c}_{ik} \cdot (\mathbf{x}_i - \mathbf{x}_j)|. \tag{59}$$

A similar estimate is obtained for the one-sided approximation

$$s_{ji} = (\mathbf{x}_j - \mathbf{x}_i) \cdot \hat{\nabla} u_j. \tag{60}$$

The final formula for $\bar{s}_{ij}$ incorporates some features of symmetric limited positive (SLIP) schemes [11], flux-corrected transport (FCT) algorithms [15,25], and geometric high-resolution schemes based on the Barth-Jespersen slope limiter [1]. We take

$$\bar{s}_{ij} = \begin{cases} \min\{2\mu_{ij}(u_i^{\max} - u_i), u_i - u_j, 2\mu_{ji}(u_j - u_j^{\min})\}, & \text{if } u_i > u_j, \\ \max\{2\mu_{ij}(u_i^{\min} - u_i), u_i - u_j, 2\mu_{ji}(u_j - u_i^{\max})\}, & \text{if } u_i < u_j. \end{cases} \tag{61}$$

As the mesh is refined, the difference between the local slopes shrinks and $\bar{s}_{ij}$ approaches $u_i - u_j$ which corresponds to a consistent finite element approximation. Moreover, the slope limiter is designed to be *linearity preserving*, i.e., $\bar{s}_{ij} = u_i - u_j$ if u is a linear function. The discrete maximum principle follows from the fact that $\bar{f}_{ij} = a_{ij}^+ \bar{s}_{ij}$ can be written in the form (50), where $u_k = u_i^{\max}$ or $u_k = u_i^{\min}$ and

$$0 \leq \beta_{ij} \leq 2\mu_{ij} a_{ij}^+. \tag{62}$$

Likewise, the limited antidiffusive flux $\bar{f}_{ji} = -\bar{f}_{ij}$ admits an equivalent representation of the form (50) with $0 \le \beta_{ji} \le 2\mu_{ji}a_{ji}^{+}$ and $u_l = u_j^{\min}$ or $u_l = u_j^{\max}$.

If i is an interior node and j is a node on the boundary, i.e., $1 \le i \le n < j \le m$, then the coefficients a_{ji}^{+} and a_{jj}^{+} belong to the blocks $A_{\Gamma\Omega}$ and $A_{\Gamma\Gamma}$, respectively. These blocks pose no hazard to the DMP since the last m rows are replaced by rows of the identity matrix in (27). Therefore, only the antidiffusive flux f_{ij}^{+} into node i needs to be constrained and the following one-sided slope limiting strategy is in order

$$\bar{s}_{ij} = \begin{cases} \min\{2\mu_{ij}(u_i^{\max} - u_i), u_i - u_j\}, & \text{if } u_i > u_j, \\ \max\{2\mu_{ij}(u_i^{\min} - u_i), u_i - u_j\}, & \text{if } u_i < u_j. \end{cases} \tag{63}$$

Remark. It is worth mentioning that the quality of the slope-limited Galerkin scheme depends on that of the underlying gradient recovery method. In the case of a heterogeneous diffusion tensor, the gradient is discontinuous but the standard lumped-mass L_2 projection (53) samples data from both sides of an internal interface, which may result in slow grid convergence (see below). Alternatively, problems with discontinuous coefficients can be treated using a special (ENO-like) gradient recovery technique.

Example. In one space dimension, the diffusion tensor $\mathcal{D}$ reduces to a scalar coefficient and cannot be anisotropic. In this case, the piecewise-linear Galerkin approximation on a uniform mesh of size h satisfies the DMP unconditionally, i.e., $A^{+} \equiv 0$ and $f_{ij}^{+} \equiv 0$. Although slope limiting is redundant, it is instructive to derive the 1D counterpart of (61) to illustrate the implications of the proposed limiting strategy.

The lumped-mass L_2-projection (53) with $m_i = h$ and $c_{i\pm1/2} = \pm 1/2$ reduces to the second-order accurate central difference approximation of the nodal gradient

$$u_i' = \frac{1}{2}\left[\frac{u_i - u_{i-1}}{h} + \frac{u_{i+1} - u_i}{h}\right]. \tag{64}$$

The local maxima and minima of the grid function u are given by

$$u_i^{\max} = \max\{u_{i-1}, u_i, u_{i+1}\}, \qquad u_i^{\min} = \min\{u_{i-1}, u_i, u_{i+1}\}. \tag{65}$$

Estimate (58) with $\mu_{ij} = 1$ and $j = i + 1$ yields the upper and lower bounds

$$u_i^{\min} - u_i \le hu_i' \le u_i^{\max} - u_i. \tag{66}$$

The one-dimensional version of formula (61) can be written as follows

$$\bar{s}_{ij} = \text{MINMOD}\{2(u_{i-1} - u_i), u_i - u_{i+1}, 2(u_{i+1} - u_{i+2})\}, \tag{67}$$

where three-parameter MINMOD function is defined as [11]

$$\text{MINMOD}\{a,b,c\} = \begin{cases} \min\{a,b,c\}, & \text{if} \quad a>0,\ b>0,\ c>0, \\ \max\{a,b,c\}, & \text{if} \quad a<0,\ b<0,\ c<0, \\ 0, & \text{otherwise.} \end{cases} \tag{68}$$

If u_i or u_{i+1} is a local maximum or minimum then the corresponding slope ratio is negative and $\bar{s}_{ij} = 0$. Otherwise, the slope limiter returns $u_i - u_{i+1}$ or a slope of the same sign and smaller magnitude. Hence, it is activated only if two neighboring slopes have opposite signs and/or their magnitudes differ by a factor of 2 and more.

7 Defect correction

After slope limiting, the corrected antidiffusive flux $\bar{f}_{ij}^{+} = a_{ij}^{+}\bar{s}_{ij}$ is added to the sum

$$\bar{f}_i^{+} = \sum_{j \neq i} \bar{f}_{ij}^{+}, \quad \forall i = 1, \ldots, n. \tag{69}$$

The resulting vector $\bar{f}_\Omega^{+}$ represents the admissible contribution of A^{+} to the residual

$$\bar{r}_\Omega = b_\Omega - A_{\Omega\Omega}^{-} u_\Omega - A_{\Omega\Gamma}^{-} g + \bar{f}_\Omega^{+}, \tag{70}$$

and the slope-limited solution u_Ω is implicitly defined by the requirement that $\bar{r}_\Omega = 0$.

Since the correction factors $\alpha_{ij} = \bar{s}_{ij}/(u_i - u_j) \in [0,1]$ depend on the *a priori* unknown slopes, the corresponding algebraic system is nonlinear and must be solved iteratively. Consider a sequence of approximations $u^{(l)},\ l = 0, 1, \ldots, L$. The current iterate $u^{(l)}$ can be used to update the residual $\bar{r}_\Omega^{(l)}$ and compute $u_\Omega^{(l+1)}$ as follows

$$\begin{bmatrix} u_\Omega^{(l+1)} \\ u_\Gamma^{(l+1)} \end{bmatrix} = \begin{bmatrix} u_\Omega^{(l)} \\ u_\Gamma^{(l)} \end{bmatrix} + \begin{bmatrix} \bar{A}_{\Omega\Omega}^{(l)} & \bar{A}_{\Omega\Gamma}^{(l)} \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} \bar{r}_\Omega^{(l)} \\ 0 \end{bmatrix}, \tag{71}$$

where $\bar{A}^{(l)}$ is a suitable 'preconditioner' to be defined below. Each cycle of this fixed-point defect correction scheme involves the solution of the auxiliary linear system

$$\begin{bmatrix} \bar{A}_{\Omega\Omega}^{(l)} & \bar{A}_{\Omega\Gamma}^{(l)} \\ 0 & I \end{bmatrix} \begin{bmatrix} v_\Omega^{(l+1)} \\ v_\Gamma^{(l+1)} \end{bmatrix} = \begin{bmatrix} \bar{r}_\Omega^{(l)} \\ 0 \end{bmatrix} \tag{72}$$

followed by the explicit solution update $u^{(l+1)} = u^{(l)} + v^{(l+1)}$. The iteration process continues until the norm of the nonlinear residual $r^{(l+1)}$ becomes small enough.

At the first outer iteration, we take $u_\Omega^{(0)} = 0$, $u_\Gamma^{(0)} = g$, and $\bar{A}^{(0)} = A^-$ or $\bar{A}^{(0)} = A$. The latter choice yields the unconstrained Galerkin solution that may violate the discrete maximum principle. On the other hand, the operator $\bar{A}_{\Omega\Omega}^{(0)} = A_{\Omega\Omega}^-$ enjoys the M-matrix property, so the DMP holds but the initial guess $u_\Omega^{(1)}$ may be very poor.

The simplest preconditioner for $l = 1, \ldots, L$ is, again, the linear operator $\bar{A}^{(l)} = A^-$. It does not need to be reassembled, and the auxiliary system (72) can be solved efficiently due to the $M-$matrix property of $\bar{A}_{\Omega\Omega}^{(l)}$. However, the convergence of outer iterations tends to be very slow, or even fail, if the anisotropy effects are too strong. Moreover, only the fully converged solution is guaranteed to satisfy the DMP.

The convergence behavior of the defect correction scheme (71) can be improved using underrelaxation ([20], p. 67). Let the diagonal entries of $\bar{A}^{(l)}$ be redefined as

$$\bar{a}_{ii}^{(l)} := \bar{a}_{ii}^{(l)}/\omega_i, \qquad \forall i = 1, \ldots, n, \tag{73}$$

where $0 < \omega_i < 1$ is a free parameter that makes $\bar{A}^{(l)}$ strictly diagonally dominant.

Implicit underrelaxation of the form (73) speeds up the convergence of inner iterations for the linear system (72) and makes it possible to constrain the solution changes so as to stabilize the residual (70) of the nonlinear problem. If fixed-point iteration (71) converges then the final solution $u = u^{(L)}$ does not depend on the choice of $\bar{A}^{(l)}$ and ω_i but the convergence history and the properties of intermediate approximations do. The best value of the relaxation parameter ω_i is problem-dependent and must be determined by trial and error, which restricts the practical utility of this approach.

In order to secure the convergence of outer iterations, it is worthwhile to design the preconditioner $\bar{A}^{(l)}$ so that every solution update is positivity-preserving. In accordance with one of Patankar's 'four basic rules' ([20], pp. 36–39), we treat the antidiffusive flux $\bar{f}_{ij}^+ = \beta_{ij}(u_k - u_i)$ as a source term and linearize it as follows

$$\bar{f}_{ij}^+ = \beta_{ij}^{(l)}(u_k^{(l)} - u_i^{(l+1)}). \tag{74}$$

Since $\beta_{ij}^{(l)} \geq 0$, this 'negative-slope linearization' [20] preserves positivity if we take

$$\bar{A}^{(l)} = A^- + \mathrm{diag}\{\sigma_i^{(l)}\}, \qquad \sigma_i^{(l)} = \sum_{j \neq i} \beta_{ij}^{(l)}. \tag{75}$$

Note that it is necessary to update the diagonal part of the preconditioner $\bar{A}^{(l)}$ at each outer iteration. Alternatively, we can invoke estimate (62) and consider

$$\bar{A}^{(l)} = A^- + \mathrm{diag}\{\sigma_i\}, \qquad \sigma_i = 2\sum_{j\neq i} \mu_{ij} a_{ij}^+. \tag{76}$$

Both definitions enhance the diagonal dominance of $\bar{A}^{(l)}$ and can be interpreted as 'relaxation through inertia' [20] for A^- or selective lumping for A. To prove positivity preservation, let us cast the linear system associated with (71) into the form

$$\begin{bmatrix} \bar{A}^{(l)}_{\Omega\Omega} & \bar{A}^{(l)}_{\Omega\Gamma} \\ 0 & I \end{bmatrix} \begin{bmatrix} u^{(l+1)}_{\Omega} \\ u^{(l+1)}_{\Gamma} \end{bmatrix} = \begin{bmatrix} \bar{b}^{(l)}_{\Omega} \\ g \end{bmatrix}. \tag{77}$$

Due to (70) and (71) the right-hand side $\bar{b}^{(l)}_{\Omega}$ of this linear system is given by

$$\bar{b}_{\Omega} = \bar{r}_{\Omega} + \bar{A}^{(l)}_{\Omega\Omega} u_{\Omega} + \bar{A}^{(l)}_{\Omega\Gamma} g = b_{\Omega} + \mathrm{diag}(\bar{A}^{(l)}_{\Omega\Omega} - A^-_{\Omega\Omega}) u_{\Omega} + \bar{f}^+_{\Omega}. \tag{78}$$

By construction, the $i-$th element of $\bar{b}^{(l)}_{\Omega}$ admits the representation

$$\bar{b}^{(l)}_i = b_i + u^{(l)}_i \sum_{j\neq i} (\bar{\beta}^{(l)}_{ij} - \beta^{(l)}_{ij}) + u^{\max}_i \sum_{u_i > u_j} \beta^{(l)}_{ij} + u^{\min}_i \sum_{u_i < u_j} \beta^{(l)}_{ij}, \tag{79}$$

where $\bar{\beta}^{(l)}_{ij} = \beta^{(l)}_{ij}$ or $\bar{\beta}^{(l)}_{ij} = 2\mu_{ij} a^+_{ij} \geq \beta^{(l)}_{ij}$ for (75) and (76), respectively. As a result, all coefficients in (79) are nonnegative. Since $\bar{A}^{(l)}_{\Omega\Omega}$ is an M-matrix and $\bar{A}^{(l)}_{\Omega\Gamma} \leq 0$

$$b_{\Omega} \geq 0, \quad u^{(l)} \geq 0 \quad \Rightarrow \quad \bar{b}_{\Omega} \geq 0 \quad \Rightarrow \quad u^{(l+1)} \geq 0. \tag{80}$$

Positivity preservation is a valuable asset since intermediate solutions remain free of undershoots and residuals converge to the machine zero in a monotone fashion. However, since the column sums of $\bar{A}^{(l)}$ are nonzero, only the final solution is guaranteed to be conservative. This is in contrast to preconditioning by $\bar{A}^{(l)} = A^-$ or $\bar{A}^{(l)} = A$, whereby each solution update is conservative but may fail to be positivity-preserving.

Remarkably, the sign of $u^{(l)}$ is preserved even if the preconditioner $\bar{A}^{(l)}$ is taken to be the diagonal part of (75) or (76). In this case, the $i-$th equation can be written as

$$\bar{a}^-_{ii} u^{(l+1)}_i = \bar{r}^{(l)}_i + (\bar{a}^{(l)}_{ii} - a^-_{ii}) u^{(l)}_i = \bar{b}_i - \sum_{j\neq i} a^-_{ij} u^{(l)}_j, \tag{81}$$

where $\bar{r}^{(l)}_i$ and $\bar{b}_i$ are given by equations (70) and (79), respectively. By definition, the coefficients a^-_{ij} are nonnegative, so positivity preservation follows from that for $\bar{b}_{\Omega}$.

The use of a diagonal preconditioner $\bar{A}^{(l)}$ leads to a fully explicit solution strategy but the convergence of outer iterations is very slow. The implicit version equipped with (75) or (76) converges faster but, nevertheless, thousands of flux/defect correction steps may be required to achieve the prescribed tolerance on a fine mesh. As in the case of linear systems which result from the discretization of elliptic problems, convergence rates deteriorate as the mesh is refined. In many cases, the nonlinearity of the slope-limited finite element discretization results in a high overhead cost.

In light of the above, a nonlinear full approximation storage / full multigrid (FAS-FMG) solution strategy lends itself to the numerical treatment of the problem at hand. The slowly converging fixed-point iteration (71) constitutes a usable smoother that can be preconditioned by the diagonal or upper/lower triangular part of $\bar{A}^{(l)}$. The use of an ILU decomposition (in conjunction with appropriate renumbering) is also feasible, and its existence is guaranteed by the M-matrix property of $\bar{A}^{(l)}_{\Omega\Omega}$.

8 Numerical examples

In this section, we test the ability of the proposed method to enforce the DMP for problem (7) with an anisotropic diffusion tensor. Also, we present a grid convergence study for test problems with smooth and discontinuous data. The difference between the numerical solution u_h and the exact solution u is measured in the norms

$$||u - u_h||_2 = \sqrt{\sum_i m_i |u(\mathbf{x}_i) - u_i|^2}, \qquad ||u - u_h||_\infty = \max_i |u(\mathbf{x}_i) - u_i|, \tag{82}$$

where $m_i = (1, \varphi_i)$ denotes a diagonal coefficient of the lumped matrix or, equivalently, the area of the control volume associated with the mesh point $\mathbf{x}_i$. The defect correction scheme preconditioned by (75) is employed in all numerical tests.

8.1 Nonsmooth solutions

To begin with, we consider two examples that represent a challenge to the conventional Galerkin discretization and illustrate the benefits of slope limiting. The computational domain for the first test problem (TP1) is $\Omega = (0,1)^2 \backslash [4/9, 5/9]^2$, as depicted in Fig. 1a. The outer and inner boundary of Ω are denoted by Γ_0 and Γ_1, respectively. Let the following Dirichlet boundary conditions be imposed on $\Gamma = \Gamma_0 \cup \Gamma_1$

$$u = -1 \quad \text{on } \Gamma_0, \qquad u = 1 \quad \text{on } \Gamma_1. \tag{83}$$

The diffusion tensor $\mathcal{D}$ is a symmetric positive definite matrix given by the formula

$$\mathcal{D} = \mathcal{R}(-\theta) \begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix} \mathcal{R}(\theta), \qquad \mathcal{R}(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \tag{84}$$

where $k_1 = 100$ and $k_2 = 1$ are the diffusion coefficients associated with the axes of the Cartesian coordinate system rotated by the angle $\theta = -\pi/6$. The source term is taken to be zero ($q \equiv 0$). By the continuous maximum principle, the (unknown) exact solution to problem (7) is bounded by the Dirichlet boundary values (83). However, the diffusion tensor (84) is highly anisotropic, which may result in a violation of the discrete maximum principle even on a regular mesh of acute/nonnarrow type.

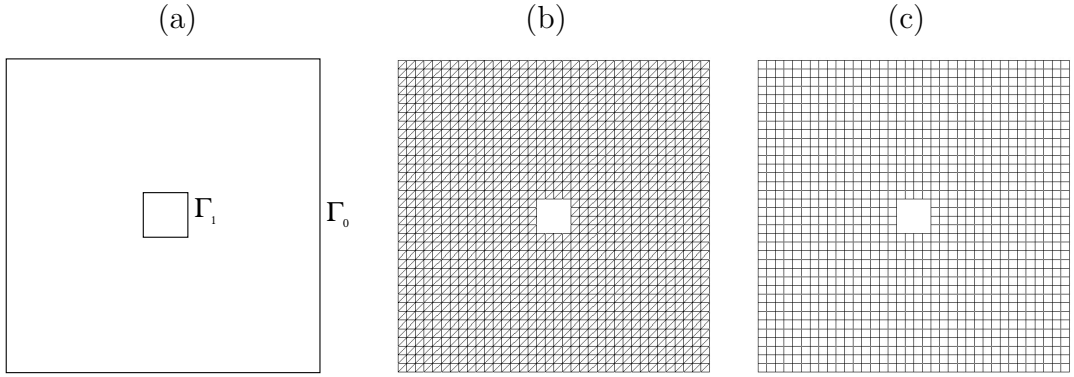

Fig. 1. TP1: (a) computational domain Ω, (b) triangular mesh, (c) quadrilateral mesh.

The verification of the DMP property is performed for linear and bilinear finite element discretizations on two uniform meshes (see Fig. 1b-c). In both cases, the total number of nodes is 1,360. The number of mesh elements equals 2,560 for the triangular mesh and 1,280 for the quadrilateral one. The numerical solutions computed on these meshes by the standard Galerkin method are displayed in Fig. 2. Both of them attain correct maximum values but exhibit spurious minima that fall below the theoretical lower bound $u_{\min} = -1$ by about 5%. Although the undershoots are relatively small, they might be totally unacceptable in some situations. For example, if the scalar variable u is responsible for phase transitions, such undershoots can trigger a nonphysical process. Since it is rather difficult to 'repair' a DMP-violating solution [18], it is worthwhile to use a scheme that does not produce undershoots/overshoots in the first place. The constrained Galerkin solutions computed on the same meshes using the slope limiter proposed in this paper are presented in Fig. 3. Both of them satisfy the DMP perfectly, and no other side effects are observed. Note that nonlinear schemes which force the solution to be positive are not applicable to this test problem.

The second test problem (TP2) stems from a benchmark suite for anisotropic diffusion problems on general grids ([8], Test 9: anisotropy and wells). The diffusion tensor is given by (84) with $k_1 = 1$, $k_2 = 10^{-3}$, and $\theta = 67.5°$. As before, the source term is zero. The computational domain $\Omega = (0,1)^2 \backslash (\bar{\Omega}_{4,6} \cup \bar{\Omega}_{8,6})$ has two square holes that

(a) $u_{\min} = -1.055,\ u_{\max} = 1.0$ (b) $u_{\min} = -1.054,\ u_{\max} = 1.0$

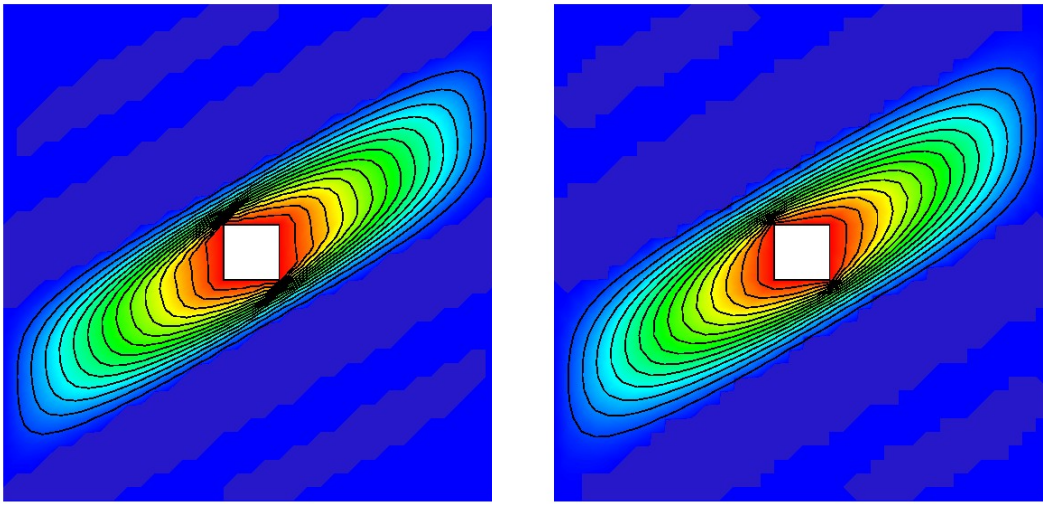

Fig. 2. TP1: unconstrained solutions, (a) triangular mesh, (b) quadrilateral mesh.

(a) $u_{\min} = -1.0,\ u_{\max} = 1.0$ (b) $u_{\min} = -1.0,\ u_{\max} = 1.0$

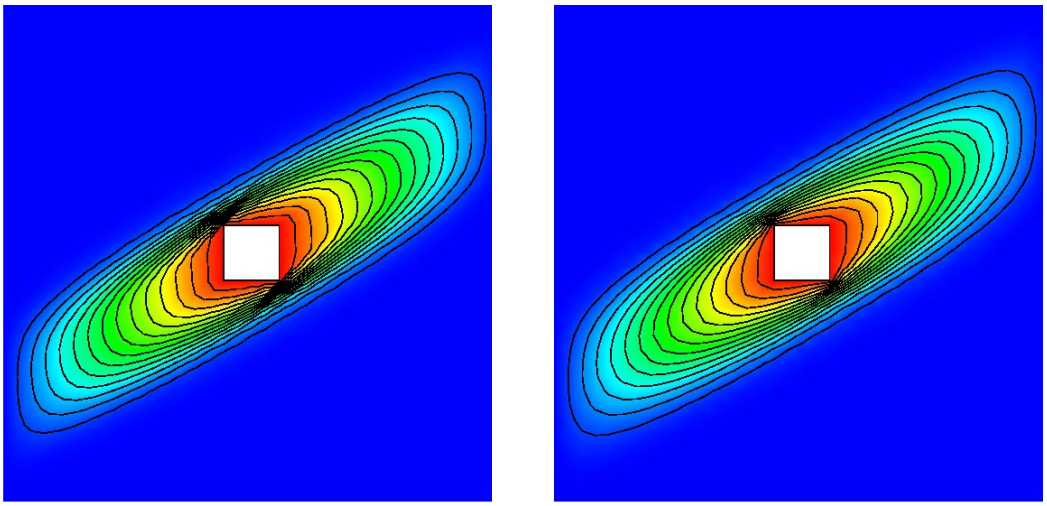

Fig. 3. TP1: constrained solutions, (a) triangular mesh, (b) quadrilateral mesh.

correspond to cells $(4, 6)$ and $(8, 6)$ of a uniform grid with 11×11 cells. The Dirichlet boundary conditions prescribed on $\Gamma_1 = \partial\bar{\Omega}_{4,6}$ and $\Gamma_2 = \partial\bar{\Omega}_{8,6}$ are as follows

$$u = 0 \quad \text{on } \Gamma_1, \qquad u = 1 \quad \text{on } \Gamma_2. \tag{85}$$

Homogeneous Neumann boundary conditions are applied at the outer boundary Γ_0 of Ω. For a detailed description of this benchmark problem we refer to [8].

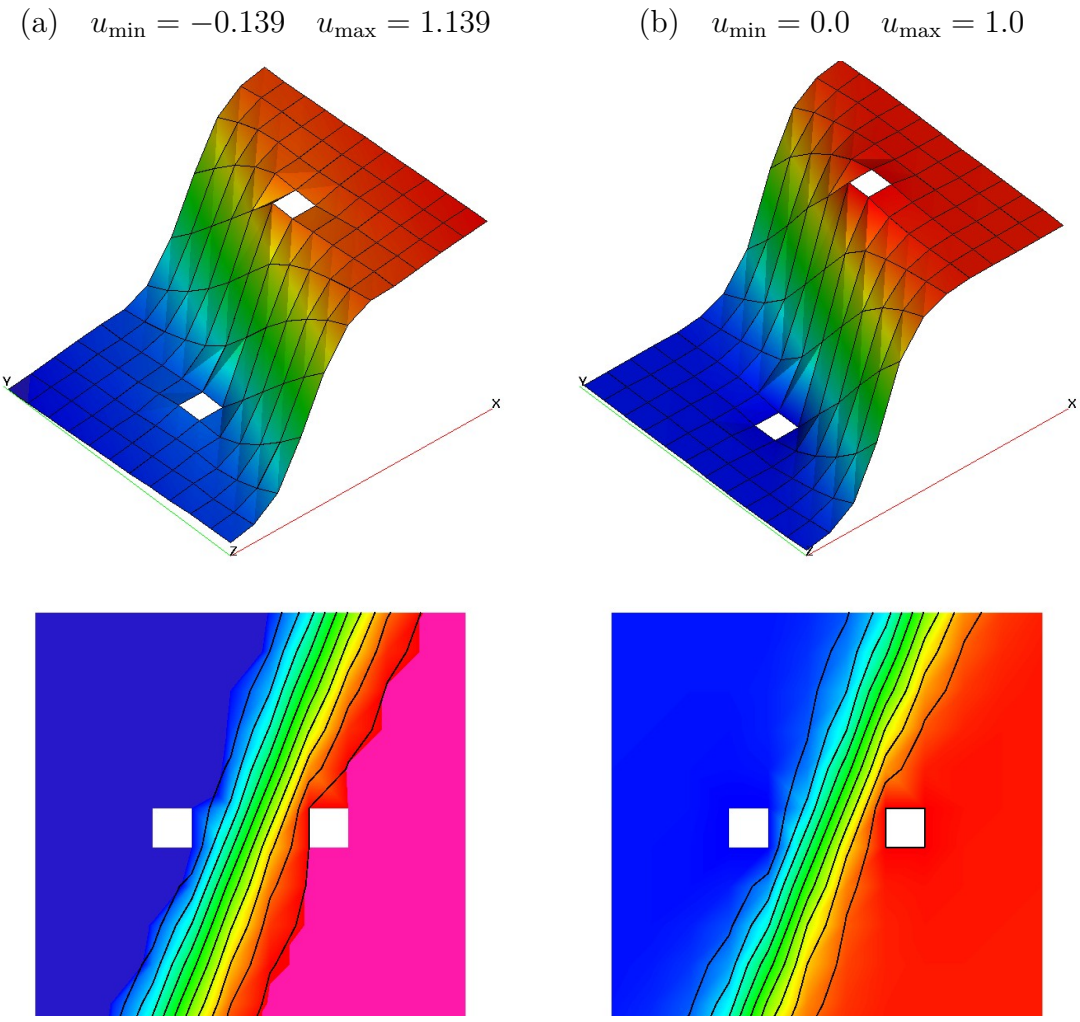

Fig. 4. TP2: bilinear elements, (a) unconstrained solution, (b) constrained solution.

The numerical solutions obtained with 11×11 bilinear finite elements are shown in Fig. 4. On such a coarse mesh, the unconstrained Galerkin method produces undershoots and overshoots of about 14%. Other discretization methods presented and compared in [8] behave in the same way, whereas our slope-limited solution is uniformly bounded by the Dirichlet boundary values, as required by the maximum principle.

8.2 Smooth solutions

In this subsection, we study the approximation properties of the proposed technique as applied to problems with smooth solutions. Usually even the conventional Galerkin scheme does not violate the discrete maximum principle for this type of problems. Thus, no slope limiting is actually required for smooth data. The goal of the numerical experiments to be performed is to compare the accuracy and convergence behavior of the constrained nonlinear scheme to those of the underlying Galerkin discretization.

The diffusion tensor and source term for the third test problem (TP3) are given by

$$\mathcal{D} = \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix}, \qquad q(x,y) = 50.5 \sin(\pi x) \sin(\pi y). \tag{86}$$

With these parameter settings, the exact solution to the Dirichlet problem (7) is

$$u(x,y) = \frac{1}{2\pi^2} \sin(\pi x) \sin(\pi y). \tag{87}$$

In accordance with this formula, homogeneous Dirichlet boundary conditions are imposed. The problem is solved on a sequence of distorted triangular and quadrilateral meshes. Given a uniform grid with spacing h, its distorted counterpart is generated by applying random perturbations to the Cartesian coordinates of internal nodes

$$x := x + \alpha \xi_x h \qquad y := y + \alpha \xi_y h, \tag{88}$$

where ξ_x and ξ_y are random numbers with values in the range from -0.5 to 0.5. The parameter $\alpha \in [0, 1]$ quantifies the degree of distortion. The default value $\alpha = 0.4$ was adopted to introduce sufficiently strong grid deformations without tangling.

In this test, the results produced by the standard Galerkin scheme and by its slope-limited counterpart are optically indistinguishable. The two diagrams in Fig. 5 show the solutions computed using linear and bilinear finite elements with $h = 1/16$. The corresponding grid convergence study is presented in Tables 1–2. On coarse meshes, the slope limiter tends to 'clip' smooth peaks, which is a well-known drawback of such methods [15]. To alleviate the undesirable decay of admissible local extrema, the sufficient conditions of DMP should be replaced by a weaker monotonicity constraint.

(a) $u_{\min} = 0.0,\ u_{\max} = 0.05$ (b) $u_{\min} = 0.0,\ u_{\max} = 0.05$

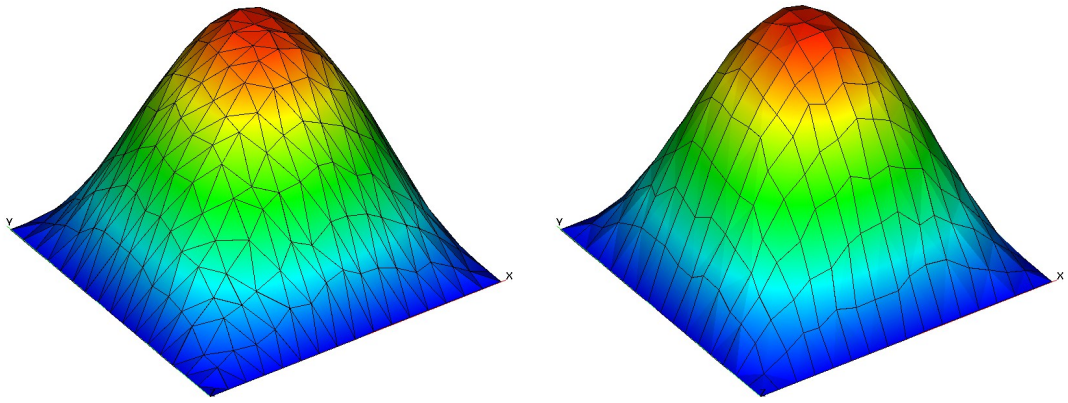

Fig. 5. TP3: numerical solutions, (a) triangular mesh, (b) quadrilateral mesh.

Table 1
TP3: grid convergence study for the unconstrained Galerkin scheme.

	triangular meshes		quadrilateral meshes	
h	$\|\|u-u_h\|\|_2$	$\|\|u-u_h\|\|_\infty$	$\|\|u-u_h\|\|_2$	$\|\|u-u_h\|\|_\infty$
1/16	0.158E-03	0.576E-03	0.113E-03	0.396E-03
1/32	0.445E-04	0.154E-03	0.270E-04	0.113E-03
1/64	0.112E-04	0.473E-04	0.693E-05	0.351E-04
1/128	0.320E-05	0.140E-04	0.176E-05	0.789E-05
1/256	0.820E-06	0.467E-05	0.441E-06	0.231E-05

Table 2
TP3: grid convergence study for the constrained Galerkin scheme.

	triangular meshes		quadrilateral meshes	
h	$\|\|u-u_h\|\|_2$	$\|\|u-u_h\|\|_\infty$	$\|\|u-u_h\|\|_2$	$\|\|u-u_h\|\|_\infty$
1/16	0.293E-03	0.136E-02	0.265E-03	0.103E-02
1/32	0.656E-04	0.407E-03	0.616E-04	0.337E-03
1/64	0.146E-04	0.121E-03	0.104E-04	0.847E-04
1/128	0.321E-05	0.140E-04	0.204E-05	0.211E-04
1/256	0.826E-06	0.467E-05	0.468E-06	0.642E-05

As the mesh is refined and resolution improves, the slope limiter is gradually deactivated, and the error norms approach those for the linear Galerkin discretization. The results presented in Tables 1–2 indicate that slope limiting does not degrade the order of convergence, and peak clipping becomes less pronounced on finer meshes.

8.3 Heterogeneous diffusion

The last example (TP4) is designed to test the ability of a discretization technique to handle problems with discontinuous coefficients. Let the diffusion tensor $\mathcal{D}$ be a piecewise-constant function defined in the square domain $\Omega = (0,1)^2$ as follows

$$\mathcal{D}(x,y) = \begin{cases} \mathcal{D}_1, & \text{if } x < 0.5, \\ \mathcal{D}_2, & \text{if } x > 0.5, \end{cases} \qquad \mathcal{D}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathcal{D}_2 = \begin{pmatrix} 10 & 3 \\ 3 & 1 \end{pmatrix}. \tag{89}$$

This heterogeneous diffusion tensor has a jump in value and direction of anisotropy across the line $x = 0.5$. The source term q is also discontinuous

$$q(x,y) = \begin{cases} 4.0, & \text{if } x < 0.5, \\ -5.6, & \text{if } x > 0.5. \end{cases} \tag{90}$$

For $\mathcal{D}$ and q defined as above, the analytical solution of problem (7) is given by

$$u(x,y) = \begin{cases} 1 - 2y^2 + 4xy + 2y + 6x, & \text{if } x \le 0.5, \\ b_2 y^2 + c_2 xy + d_2 x + e_2 y + f_2, & \text{if } x > 0.5. \end{cases} \tag{91}$$

Substitution into (7) yields the following values of the involved coefficients

$$b_2 = -2, \quad c_2 = \frac{4(\mathcal{D}_2(1,2)+1)}{\mathcal{D}_2(1,1)}, \quad d_2 = \frac{6 - 4\mathcal{D}_2(1,2)}{\mathcal{D}_2(1,1)}, \tag{92}$$

$$e_2 = \frac{4\mathcal{D}_2(1,1) - 2\mathcal{D}_2(1,2) - 2}{\mathcal{D}_2(1,1)}, \quad f_2 = \frac{4\mathcal{D}_2(1,1) + 2\mathcal{D}_2(1,2) - 3}{\mathcal{D}_2(1,1)}. \tag{93}$$

Again, the discretization is performed using linear and bilinear finite elements on distorted meshes. These meshes are constructed as explained in the previous subsection but nodes that belong to the line $x = 0.5$ are shifted in the $y-$direction only. The unconstrained Galerkin solutions for $h = 1/16$ are presented in Fig. 6. Their constrained counterparts look the same but a comparison of the error norms presented in Tables 1–2 reveals significant differences between the convergence histories of the slope-limited version on triangular and quadrilateral meshes. Although the solution consists of two smooth patches, its gradient is discontinuous across the internal interface $x = 0.5$. Moreover, the corresponding kink in the solution profile makes the outcome of the slope limiting procedure highly mesh-dependent. Note that the solution is smooth along the y-axis and piecewise-smooth along the $x-$axis. This is why the constrained and unconstrained solutions coincide on quadrilateral meshes.

On the other hand, some edges of the triangular mesh are directed skew to the kink so that the corresponding solution differences are large, whereas the distance to the nearest local maximum or minimum, as defined in (56)–(57), is small. This places a heavy burden on the slope limiter which is forced to reject a large percentage of the antidiffusive flux in accordance with (61). The approximation of discontinuous gradients by means of the standard L_2 projection (53) can also be responsible for the relatively slow convergence on distorted triangular meshes. In summary, this test problem turns out to be very easy or rather difficult, depending on the orientation of mesh edges. It was included to identify the limitations of the proposed limiting strategy, discuss their ramifications, and illustrate the need for further research.

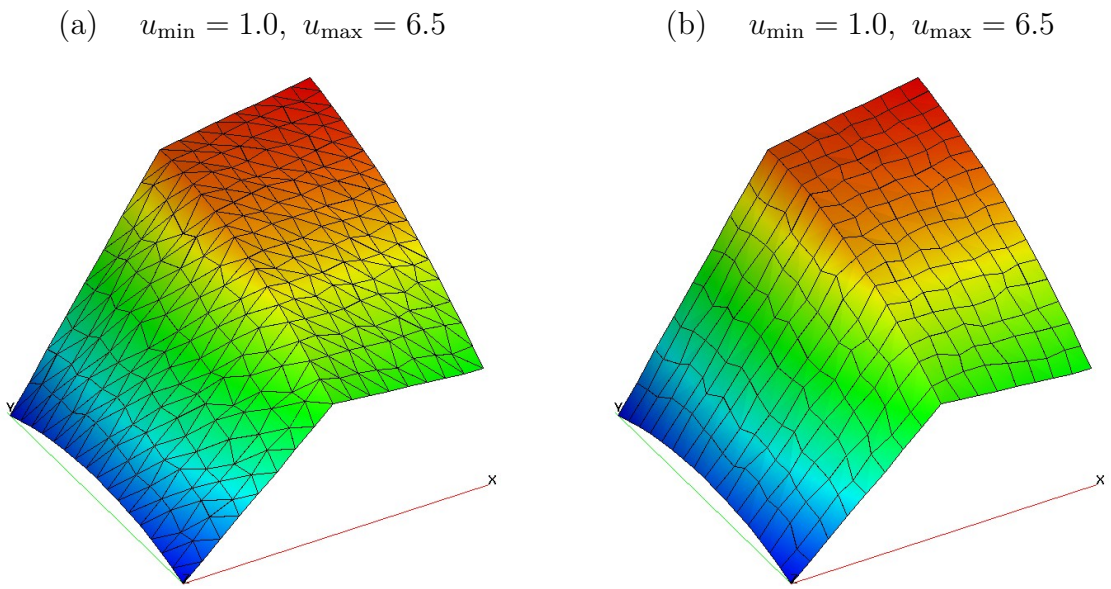

Fig. 6. TP4: numerical solutions, (a) triangular mesh, (b) quadrilateral mesh.

Table 3
TP4: grid convergence study for the unconstrained Galerkin scheme.

	triangular meshes		quadrilateral meshes	
h	$\|\|u-u_h\|\|_2$	$\|\|u-u_h\|\|_\infty$	$\|\|u-u_h\|\|_2$	$\|\|u-u_h\|\|_\infty$
1/16	0.101E-02	0.337E-02	0.473E-03	0.167E-02
1/32	0.328E-03	0.133E-02	0.154E-03	0.636E-03
1/64	0.841E-04	0.374E-03	0.426E-04	0.242E-03
1/128	0.211E-04	0.107E-03	0.109E-04	0.514E-04
1/256	0.551E-05	0.351E-04	0.286E-05	0.166E-04

Table 4
TP4: grid convergence study for the constrained Galerkin scheme.

	triangular meshes		quadrilateral meshes	
h	$\|\|u-u_h\|\|_2$	$\|\|u-u_h\|\|_\infty$	$\|\|u-u_h\|\|_2$	$\|\|u-u_h\|\|_\infty$
1/16	0.244E-02	0.155E-01	0.473E-03	0.167E-02
1/32	0.161E-02	0.187E-01	0.154E-03	0.636E-03
1/64	0.101E-02	0.109E-01	0.426E-04	0.242E-03
1/128	0.281E-03	0.438E-02	0.109E-04	0.514E-04
1/256	0.140E-03	0.214E-02	0.286E-05	0.166E-04

9 Summary and conclusions

The objective of the present paper was to review some important qualitative properties of exact solutions to elliptic problems and preserve these properties at the discrete level. After a brief overview of continuous maximum principles, linear and bilinear finite element discretizations of the stationary diffusion equation were considered. Sufficient conditions of the discrete maximum principle were formulated. The residual of the algebraic system was decomposed into internodal fluxes associated with positive and negative coefficients of the global stiffness matrix. Flux correction was performed using a new slope limiter based on gradient reconstruction. The resulting nonlinear problem was solved by an iterative defect correction scheme equipped with a positivity-preserving preconditioner. A numerical study was performed for several 2D test problems with anisotropic diffusion tensors and discontinuous coefficients.

The presented computational results demonstrate that the new approach to slope limiting rules out the formation of spurious undershoots and overshoots, while preserving the accuracy and consistency of the underlying discretization in the case of sufficiently smooth data. An extension to quadratic and/or mixed finite elements appears to be feasible and will be addressed in future work. Further research is required to accelerate the slowly converging defect correction scheme and reduce the significant overhead cost incurred by the nonlinearity of the constrained approximation. Another important open problem is the design of optimal slope limiters for problems that feature smooth local extrema and/or jumps across internal interfaces.

Acknowledgments.

The authors thank Dr. K. Lipnikov (LANL), Prof. Yu. Kuznetsov (University of Houston), and Dr. Yu. Vassilevski (INM, Moscow) for fruitful discussions on the paper topic and many useful comments. The work of the second and third author was carried out under the auspices of the National Nuclear Security Administration of the U.S. Department of Energy at Los Alamos National Laboratory under Contract No. DE-AC52-06NA25396 and the DOE Office of Science Advanced Scientific Computing Research (ASCR) Program in Applied Mathematics Research.

References

[1] T. Barth, D. C. Jespersen, The design and application of upwind schemes on unstructured meshes, AIAA paper 89-0366, 27th AIAA Aerospace Sciences Meeting, Reno, NV, USA, 1989.

[2] J.-M. Beckers, H. Burchard, E. Deleersnijder, P. P. Mathieu, Numerical discretization of rotated diffusion operators in ocean models, Monthly Weather Review 28 (8) (2000) 2711–2733.

[3] P. G. Ciarlet, Discrete maximum principle for finite-difference operators, Aequationes Math. 4 (1970) 338–352.

[4] P. G. Ciarlet, P.-A. Raviart, Maximum principle and convergence for the finite element method, Comput. Meth. Appl. Mech. Engrg. 2 (1973) 17–31.

[5] R. Codina, A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation, Comput. Methods Appl. Mech. Engrg. 110 (1993) 325–342.

[6] I. Faragó, R. Horváth, S. Korotov, Discrete maximum principle for linear parabolic problems solved on hybrid meshes, Appl. Numer. Math. 53 (2005) 249–264.

[7] D. Gilbarg, N. S. Trudinger, Elliptic Partial Differential Equations of Second Order (2nd edition), Grudlehren der Mathematischen Wissenschaften 224, Springer, 1983.

[8] R. Herbin and F. Hubert, Benchmark on discretization methods for anisotropic diffusion problems on general grids, in: R. Eymard, J.-M. Herard (Eds.), Finite Volumes for Complex Applications V, 2008, pp. 659–692.

[9] W. Höhn, H.-D. Mittelmann, Some remarks on the discrete maximum-principle for finite elements of higher order. Computing 27 (2)(1981) 145–154.

[10] W. Hundsdorfer, C. Montijn, A note on flux limiting for diffusion discretizations. IMA J. Numer. Anal. 24 (4) (2004) 635–642.

[11] A. Jameson, Analysis and design of numerical schemes for gas dynamics 1 Artificial diffusion, upwind biasing, limiters and their effect on accuracy and multigrid convergence, Int. J. Comput. Fluid Dynamics, 4 (1995) 171–218.

[12] J. Karátson, S. Korotov, M. Křížek, On discrete maximum principles for nonlinear elliptic problems, Math. Comp. Simulation 76 (2007) 99–108.

[13] J. Karátson, S. Korotov, Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions, Numer. Math. 99 (2005) 669–698.

[14] D. Kuzmin, On the design of general-purpose flux limiters for implicit FEM with a consistent mass matrix. I. Scalar convection, J. Comput. Phys. 219 (2006) 513–531.

[15] D. Kuzmin, M. Möller, Algebraic flux correction I. Scalar conservation laws, in: D. Kuzmin, R. Löhner, S. Turek (Eds.), Flux-Corrected Transport: Principles, Algorithms, and Applications, Springer, Berlin, 2005, pp. 155–206.

[16] C. Le Potier, Schema volumes finis monotone pour des operateurs de diffusion fortement anisotropes sur des maillages de triangle non structures. C. C. Acad. Sci. Paris, Ser. I, 341 (2005), 787-792.

[17] K. Lipnikov, M. Shashkov, D. Svyatskiy, Yu. Vassilevski, Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes, J. Comp. Phys. 227 (2007) 492–512.

[18] R. Liska, M. Shashkov, Enforcing the discrete maximum principle for linear finite element solutions of second-order elliptic problems, Commun. Comp. Phys. 3 (4) (2007) 852–877.

[19] M. Möller, D. Kuzmin, Adaptive mesh refinement for high-resolution finite element schemes, Int. J. Numer. Meth. Fluids, 52 (2006) 545–569.

[20] S. V. Patankar, Numerical Heat Transfer and Fluid Flow. McGraw-Hill, New York, 1980.

[21] M. H. Protter, H. F. Weinberger, Maximum Principles in Differential Equations, Prentice-Hall, 1967.

[22] G. Stoyan, On maximum principles for monotone matrices. Linear Algebra Appl. 78 (1986) 147-161.

[23] R. S. Varga, Matrix Iterative Analysis. Prentice-Hall, Englewood Cliffs, 1962.

[24] M. Xue, High-order monotonic numerical diffusion and smoothing. Monthly Weather Review 128 (8) (2000) 2853-2864.

[25] S. T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids, J. Comput. Phys. 31 (1979) 335–362.