Commun. Comput. Phys.
doi: 10.4208/cicp.180615.280815a

Vol. **19**, No. 1, pp. 53-93
January 2016

A Numerical Methodology for Enforcing Maximum Principles and the Non-Negative Constraint for Transient Diffusion Equations

K. B. Nakshatrala[1,*], H. Nagarajan[2] and M. Shabouei[1]

[1] *Department of Civil & Environmental Engineering, University of Houston, Houston, Texas 77204-4003, USA.*

[2] *Department of Mechanial Engineering, Texas A&M University, College Station, TX 77843, USA.*

Received 18 June 2015; Accepted (in revised version) 28 August 2015

Abstract. Transient diffusion equations arise in many branches of engineering and applied sciences (e.g., heat transfer and mass transfer), and are parabolic partial differential equations. It is well-known that these equations satisfy important mathematical properties like maximum principles and the non-negative constraint, which have implications in mathematical modeling. However, existing numerical formulations for these types of equations do not, in general, satisfy maximum principles and the non-negative constraint. In this paper, we present a methodology for enforcing maximum principles and the non-negative constraint for transient anisotropic diffusion equation. The proposed methodology is based on the method of horizontal lines in which the time is discretized first. This results in solving steady anisotropic diffusion equation with decay equation at every discrete time-level. We also present other plausible temporal discretizations, and illustrate their shortcomings in meeting maximum principles and the non-negative constraint. The proposed methodology can handle general computational grids with no additional restrictions on the time-step. We illustrate the performance and accuracy of the proposed methodology using representative numerical examples. We also perform a numerical convergence analysis of the proposed methodology. For comparison, we also present the results from the standard single-field semi-discrete formulation and the results from a popular software package, which all will violate maximum principles and the non-negative constraint.

AMS subject classifications: 65

Key words: Numerical heat and mass transfer, maximum principles, non-negative solutions, anisotropic diffusion, method of horizontal lines, convex quadratic programming, parabolic PDEs.

*Corresponding author. *Email address:* knakshatrala@uh.edu; *Phone:* 713-743-4418 (K. B. Nakshatrala)

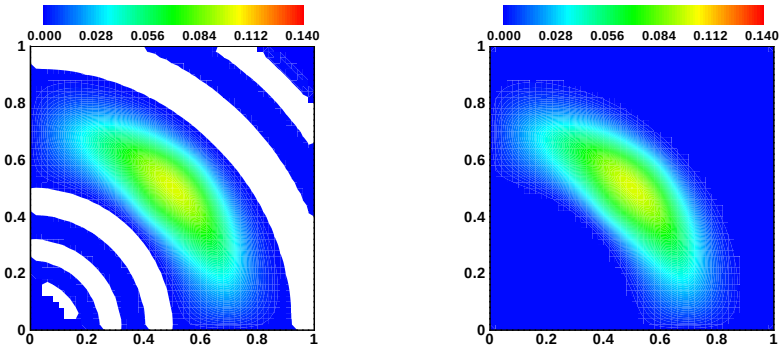


Figure 1: Diffusion in heterogeneous anisotropic medium: This figure shows the contours of the concentration under the Galerkin single-field formulation (left) and the proposed methodology (right) at time $=0.5$. The time-step is taken as $\Delta t = 0.5$, and $\text{XSeed} = \text{YSeed} = 51$. The regions that violated the non-negative constraint are indicated in white color.

1 Introduction and motivation

Certain quantities (e.g., concentration of a chemical species and absolute temperature) naturally attain non-negative values. A violation of the non-negative constraint for these quantities will imply violation of some basic tenets of physics. It is, therefore, imperative that such physical constraints are met by mathematical models and by their associated numerical formulations. In this paper, we shall focus on two popular transient mathematical models, in which physical restrictions like the non-negative constraint play a central role. The first model is based on Fick's assumption (commonly referred to as Fick's law) and the balance of mass. Fick's assumption is a simple constitutive model to describe the diffusion of a chemical species in which the flux is proportional to the negative gradient of the concentration. The second model is based on Fourier's assumption and the balance of energy, which describes heat conduction in a rigid conductor. Both these constitutive models combined with their corresponding balance laws give rise to transient diffusion equations, which are parabolic partial differential equations.

There has been tremendous progress in applied mathematics for these type of equations with respect to existence and uniqueness results, qualitative behavior of solutions, estimates, and other mathematical properties [21, 56]. In particular, it has been shown that transient diffusion equations satisfy the so-called maximum principles [56]. It will be shown in a subsequent section that the non-negative constraint can be shown as a consequence of maximum principles under certain assumptions. Analytical solutions to several problems have been documented in various monographs (e.g., see references [11, 54]). However, it should be noted that most of these solutions are for isotropic and homogeneous media, and for simple geometries. For problems involving anisotropic and heterogeneous media, and complex geometries; finding analytical solutions is not possible, and one has to resort to numerical solutions. Obtaining physically meaning-

ful numerical solutions for transient diffusion equation that satisfy maximum principles and the non-negative constraint is the main aim of this paper. It is well-known (and will be discussed in subsequent sections) that many popular numerical schemes (including the ones that are based on the finite element method) do not satisfy maximum principles and the non-negative constraint. Even for isotropic diffusion, stringent restrictions on the time-step and the computational mesh are necessary to meet these important mathematical properties.

The usual approach of solving linear second-order parabolic partial differential equations under the finite element method is to employ Galerkin formalism for spatial discretization. Several theoretical results (which include convergence proofs, a-priori estimates) for this approach can be found in the literature (e.g., see [19]). But it has been adequately documented in the literature that this approach will not satisfy maximum principles and the non-negative constraint (for example, see [28], and also the discussion in Appendix). Thus, there is a need to develop new methodologies that will satisfy important mathematical properties like maximum principles and the non-negative constraint, and thereby improve the overall predictive capabilities of current numerical schemes.

1.1 Maximum principles for diffusion equations in numerical setting

Maximum principles in the discrete setting are sometimes referred to as discrete maximum principles (DMPs). The first study on maximum principles in the context of the finite element method can be traced back to the seminal paper by Ciarlet and Raviart [17], which considered steady-state isotropic diffusion, low-order approximation, and simplicial elements. Ciarlet and Raviart points out that the single-field formulation (which is based on the Galerkin formalism) does not converge uniformly for isotropic diffusion equation unless some restrictions are placed on the mesh. In particular, they show that a sufficient condition for a three-node triangular element to converge uniformly and to meet maximum principles is that the triangle has to be acute. But this sufficient condition is valid only for steady-state isotropic diffusion equations. A more detailed account of various works can be found in [49,50,57]. Although these papers have considered steady-state diffusion equation, the discussion in these papers is applicable to transient diffusion equations. A brief summary of these three papers is as follows. In [50], a non-negative methodology for mixed finite element formulation has been proposed for steady-state diffusion equation using techniques from convex quadratic programming. The paper also studied the effect of the non-negative methodology on the element local mass balance. In [49], a methodology has been proposed for steady-state diffusion equation with decay that satisfies maximum principles and the non-negative constraint on general computational grids. (Note that the maximum principle for diffusion with decay is slightly different from the maximum principle without decay.) This methodology will be utilized later in the present paper. In [57], a systematic study on the effect of high-order approximation on the violation of maximum principles and the non-negative constraint.

In particular, it has been shown using numerical simulations that the violation of the non-negative constraint does not decrease with p-refinement. Some representative works in other areas of discretization to obtain monotone solutions include finite volume methods [40–42], and mimetic finite difference methods [43].

1.1.1 Maximum principles for transient systems

Transient diffusion equations fall in the realm of parabolic partial differential equations (PDEs), whereas steady-state diffusion equations are elliptic PDEs. For a comprehensive treatment of mathematical properties of parabolic PDEs, see [56]. Several papers have addressed maximum principles for parabolic problems in the numerical setting. In [29], flow-oriented derivatives with backward Euler have been employed to obtain non-negative solutions under finite difference and finite volume methods. A method that is commonly employed in the area of subsurface hydrology was proposed in [14]. This method is based on the standard single-field formulation but employs lumped capacity matrix. (By the standard single-field formulation we refer to the formulation obtained by employing the semi-discrete approach using method of vertical lines at integral time-steps, and Galerkin formalism for spatial discretization. See Appendix for more details.) It should be emphasized that lumping capacity matrix approach is commonly considered as a variational crime [69]. More importantly, lumped capacity matrix is not sufficient to meet maximum principles and the non-negative constraint for anisotropic diffusion even if one employs the backward Euler time-stepping scheme (e.g., see subsection 4.4 and Appendix). Reference [8] also alters the capacity matrix to preserve positivity for parabolic problems but restricts to isotropic diffusion. Other notable works are [20,22,60,63], which all focused on getting restrictions on the mesh (and in some cases on the time-step) to meet maximum principles. More importantly, they did not consider anisotropy, and such restrictions are not possible for anisotropic and heterogeneous medium.

There are several papers that considered consistent capacity matrices, but derived restrictions on the time-step to satisfy maximum principles [28, 30, 35, 45, 67]. A striking difference between the time-step restrictions with respect to numerical stability and maximum principles is that numerical stability places an upper bound on the selection of the time-step whereas maximum principles place a lower bound on the selection of the time-step. The time-step is selected based on the following inequality:

$$0 < \Delta t^{\text{MP}}_{\text{crit}} \leq \Delta t \leq \Delta t^{\text{stability}}_{\text{crit}}, \tag{1.1}$$

where $\Delta t^{\text{stability}}_{\text{crit}}$ is the critical time-step to obtain stable results, and $\Delta t^{\text{MP}}_{\text{crit}}$ is the critical time-step to satisfy maximum principles. It should be however mentioned that these works on deriving time-step restrictions have considered one-dimensional problems or isotropic media, and these conditions are not applicable otherwise. To the best of our knowledge, none of the prior works presented a methodology for transient anisotropic diffusion equations to satisfy maximum principles and the non-negative constraint on general computational grids with no further restrictions on the time-step.

Recently, Huang and co-workers have developed time-step restrictions to satisfy maximum principles for generalized α time-stepping schemes [39] and explicit Runge-Kutta method [32]. These approaches are based on a combination of method of vertical lines and DMP-based anisotropic meshes. Although these approaches can satisfy maximum principles and the non-negative constraint they still suffers from several drawbacks. First, there are restrictions on the time-step. Second, a DMP-based mesh needs to be generated. The solution procedure to generate a DMP-based mesh is nonlinear, and there is no guarantee that such a mesh exists for complex geometries [31]. Third, considerably h-refined DMP-based meshes are required to satisfy various discrete properties. This is because a coarse DMP-based mesh may not be adequate to obtain highly accurate numerical solution. The proposed methodology in this paper requires no restrictions on the time-step, and satisfies maximum principles and the non-negative constraint even on a general coarse computational grid.

1.2 Our approach

In this paper, we employ the method of horizontal lines (i.e, the Rothe method) [65] to solve transient anisotropic diffusion equation. The novelty is to convert the transient diffusion problem into solving an *appropriate form* of diffusion with decay at every time-step with non-negative forcing function and symmetric positive definite coefficient matrix. It should be emphasized that an arbitrary temporal discretization to convert the transient problem into solving elliptic partial differential equations may not inherit maximum principles and the non-negative constraint. To this end, we present several other plausible temporal discretizations in Appendix that also result in diffusion with decay. But none of these methods satisfy maximum principles and the non-negative constraint. The proposed methodology has been carefully designed so that the non-negative constraint and maximum principles are inherited at every time-step.

There are several papers in the literature that have employed the Rothe method to solve parabolic equations [9, 13, 28, 37]. These papers, except for [28], did not apply the Rothe method in the context of maximum principles. Although [28] addressed maximum principles by using the Rothe method, but the formulation is restricted to isotropic diffusion. In addition, [28] employed techniques from stabilized methods, which is different from the approach taken in this paper. In the proposed formulation, the temporal discretization using the Rothe method will give rise to inhomogeneous elliptic partial differential equation, which is solved using the approach presented in our earlier paper [49]. An attractive aspect of the proposed methodology is that there are no additional restrictions on the time-step to meet maximum principles.

In [52], the backward Euler time-stepping scheme is employed to convert a transient fast bimolecular reaction-diffusion equation to pure diffusion with decay. They then used a non-negative methodology to satisfy maximum principles and non-negative constraint on the resulting time-discretized system. Their approach is a subclass of the methods discussed in this paper. Moreover, the main focus in [52] has been reactive systems.

Herein, we also discuss various plausible approaches to meet maximum principles for transient diffusive systems, and discuss their shortcomings.

In this paper, repeated indices do not imply summation. We employ the standard notation for open, closed and half-open intervals. The continuum vectors are denoted by lower case boldface unitalicized letters, and second-order tensors will be denoted by upper case boldface normal letters (for example, vector $\mathbf{x}$ and second-order tensor $\mathbf{D}$). In the finite element context, we shall denote the vectors using lower case boldface italic letters, and the matrices are denoted using upper case boldface italic letters. For example, vector $\boldsymbol{v}$ and matrix $\boldsymbol{K}$. Other notational conventions adopted in this paper are introduced as needed.

2 Governing equations: Transient anisotropic diffusion

Let $\Omega \subset \mathbb{R}^{nd}$ be a bounded open set, where "nd" denotes the number of spatial dimensions. The boundary is denoted by $\partial\Omega$, which is assumed to be piecewise smooth. A spatial point is denoted by $\mathbf{x} \in \overline{\Omega}$, where a superposed bar denotes the set closure. The gradient and divergence with respect to $\mathbf{x}$ are denoted by $\operatorname{grad}[\cdot]$ and $\operatorname{div}[\cdot]$, respectively. Let $t \in [0,\mathcal{I}]$ denote the time, where $\mathcal{I} > 0$ denotes the length of the time interval. The concentration of a chemical species is denoted by $c(\mathbf{x},t)$. The (spatial) boundary is divided into two parts: Γ^{D} and Γ^{N} such that $\Gamma^{\mathrm{D}} \cup \Gamma^{\mathrm{N}} = \partial\Omega$ and $\Gamma^{\mathrm{D}} \cap \Gamma^{\mathrm{N}} = \emptyset$. Γ^{D} is that part of the boundary on which Dirichlet boundary condition (i.e., the concentration) is prescribed, and Γ^{N} is the part of the boundary on which Neumann boundary condition (i.e., the flux) is prescribed. The unit outward normal to the boundary is denoted by $\widehat{\mathbf{n}}(\mathbf{x})$. The governing equations for transient anisotropic diffusion can be written as:

$$\frac{\partial c(\mathbf{x},t)}{\partial t} - \operatorname{div}[\mathbf{D}(\mathbf{x})\operatorname{grad}[c(\mathbf{x},t)]] = f(\mathbf{x},t) \quad \text{in } \Omega \times (0,\mathcal{I}), \tag{2.1a}$$

$$c(\mathbf{x},t) = c_p(\mathbf{x},t) \quad \text{on } \Gamma^{\mathrm{D}} \times (0,\mathcal{I}), \tag{2.1b}$$

$$\widehat{\mathbf{n}}(\mathbf{x}) \cdot \mathbf{D}(\mathbf{x})\operatorname{grad}[c(\mathbf{x},t)] = q_p(\mathbf{x},t) \quad \text{on } \Gamma^{\mathrm{N}} \times (0,\mathcal{I}), \tag{2.1c}$$

$$c(\mathbf{x},t=0) = c_0(\mathbf{x}) \quad \text{in } \Omega, \tag{2.1d}$$

where $\mathbf{D}(\mathbf{x})$ is the diffusivity tensor, $f(\mathbf{x},t)$ is the volumetric source/sink, $c_p(\mathbf{x},t)$ is the prescribed concentration on the boundary, $q_p(\mathbf{x},t)$ is the prescribed flux on the boundary, and $c_0(\mathbf{x})$ is the prescribed initial condition. The diffusivity tensor is symmetric, and is assumed to be bounded above and uniformly elliptic. That is, there exists two constants $0 < \xi_1 \leq \xi_2 < +\infty$ such that

$$\xi_1 \mathbf{y}^{\mathrm{T}}\mathbf{y} \leq \mathbf{y}^{\mathrm{T}}\mathbf{D}(\mathbf{x})\mathbf{y} \leq \xi_2 \mathbf{y}^{\mathrm{T}}\mathbf{y} \quad \forall \mathbf{x} \in \Omega \ \text{ and } \ \forall \mathbf{y} \in \mathbb{R}^{nd}. \tag{2.2}$$

The above initial boundary value problem (2.1a)-(2.1d) is a linear parabolic partial differential equation. From the theory of partial differential equations, such equations are known to satisfy maximum principles under appropriate regularity assumptions on the input data and on the domain [62].

Remark 2.1. It should be noted that a consequence of Fickian/Fourier mathematical model is that a thermal/chemical disturbance at a point will be felt at other points instantaneously. This is because of the parabolic nature of the resulting partial differential equations. To put it differently, these mathematical models predict that the information travels at infinite speed, which is against the current accepted laws of Physics. Several modifications have been suggested in the area of heat conduction to have finite speeds for thermal disturbances, and most of these models are hyperbolic partial differential equations. Some notable works on this topic are [44], [12], and [25]. A more detailed discussion with respect to finite speed thermoelasticity can be found in [34]. It is noteworthy that hyperbolic partial differential equations do not possess maximum principles "similar" to the ones possessed by elliptic and parabolic partial differential equations. This area of research is far from settled, and is beyond the scope of this paper.

2.1 Maximum principles for parabolic equations

Maximum principles for parabolic partial differential equations can be traced back to Levi [38] and Picone [59]. A brief history and other references on maximum principles for parabolic partial differential equations can be found in [62]. Our presentation of maximum principles is similar to that of Nirenberg [53].

Let $C^m(\Omega)$ denote the set of functions defined on Ω that are continuously differentiable up to m-th order. The parabolic cylinder $\Omega_{\mathcal{I}}$ and parabolic boundary $\Gamma_{\mathcal{I}}$ are pictorially described in Fig. 2. Mathematically,

$$\Omega_{\mathcal{I}} := \Omega \times (0,\mathcal{I}) \quad \text{and} \quad \Gamma_{\mathcal{I}} := \left\{ (\mathbf{x},t) \in \overline{\Omega}_{\mathcal{I}} \,\middle|\, \mathbf{x} \in \partial\Omega \text{ or } t=0 \right\}. \tag{2.3}$$

We shall introduce the following function space with differing smoothness in the **x**- and t-variables:

$$C_1^2(\Omega_{\mathcal{I}}) := \left\{ c : \Omega_{\mathcal{I}} \to \mathbb{R} \,|\, c, \frac{\partial c}{\partial x_i}, \frac{\partial^2 c}{\partial x_i \partial x_j}, \frac{\partial c}{\partial t} \in C(\Omega_{\mathcal{I}});\ i,j=1,\cdots,nd \right\}. \tag{2.4}$$

Theorem 2.1 (Maximum principle for transient diffusion equations)**.** *Let $c(\mathbf{x},t) \in C_1^2(\Omega_{\mathcal{I}}) \cap C(\overline{\Omega}_{\mathcal{I}})$ that satisfies $\partial c/\partial t - \operatorname{div}[\mathbf{D}(\mathbf{x})\operatorname{grad}[c]] = f(\mathbf{x},t) \leq 0$ in $\Omega_{\mathcal{I}}$. Then $c(\mathbf{x},t)$ achieves its maximum on the parabolic boundary $\Gamma_{\mathcal{I}}$. That is,*

$$\max_{(\mathbf{x},t)\in\overline{\Omega}_{\mathcal{I}}} c(\mathbf{x},t) = \max_{(\mathbf{x},t)\in\Gamma_{\mathcal{I}}} c(\mathbf{x},t). \tag{2.5}$$

Similarly, if $f(\mathbf{x},t) \geq 0$ in $\Omega_{\mathcal{I}}$ then $c(\mathbf{x},t)$ achieves its minimum on $\Gamma_{\mathcal{I}}$.

Proof. A proof can be found in [56]. □

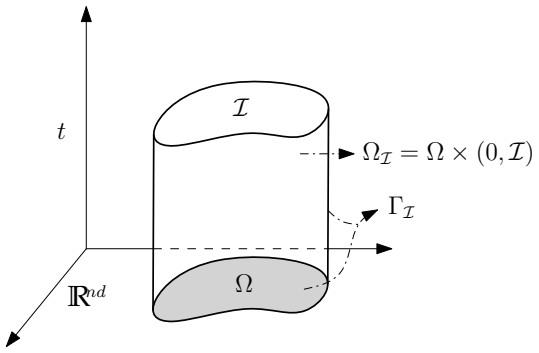


Figure 2: A pictorial description of parabolic cylinder $\Omega_{\mathcal{I}}$ and parabolic boundary $\Gamma_{\mathcal{I}}$.

Maximum principles play a central role in the study of partial differential equations. Many uniqueness theorems and powerful estimates for parabolic partial differential equations utilize maximum principles [24,56]. Maximum principles also have important physical implications in mathematical modeling, as they place restrictions on physical quantities. One such implication is the non-negative constraint. We now show that the non-negative constraint can be obtained as a consequence of the maximum principle given by Theorem 2.1. For the present discussion, assume that $\Gamma^{\mathrm{D}} = \partial\Omega$ (that is, we prescribe Dirichlet boundary conditions on the whole boundary). If $f(\mathbf{x},t) \geq 0$ (i.e., we have volumetric source), $c_p(\mathbf{x},t) \geq 0$ (i.e., we have non-negative prescribed Dirichlet boundary conditions on the whole boundary), and $c_0(\mathbf{x}) \geq 0$ (i.e., we have non-negative prescribed initial concentration); then the maximum principle given by Theorem 2.1 implies that $c(\mathbf{x},t)$ is non-negative in the whole domain and at all times. That is,

$$c(\mathbf{x},t) \geq 0 \quad \forall \mathbf{x} \in \overline{\Omega} \quad \text{and} \quad \forall t \in [0,\mathcal{I}]. \tag{2.6}$$

It should be noted that the above discussion on maximum principles and the non-negative constraint is in the continuum setting. For most practical problems (which will involve complex geometries and spatially varying coefficients), it is not possible to find analytical solutions. Therefore, one has to resort to numerical solutions. This leads to the following fundamental questions, which are addressed in this paper. *Whether numerical formulations satisfy maximum principles and the non-negative constraint for transient diffusion equation in the discrete setting. If so, under what conditions? If not, is it possible to fix a given numerical formulation to meet these important principles?*

2.2 Discrete maximum principles

The discrete analogy of maximum principles is commonly referred to as *discrete maximum principles* (DMP). Some main factors that affect numerical solutions with respect to discrete maximum principles are: (i) topology of the domain (e.g., shape of the domain, features like holes in the domain), (ii) type of mesh (e.g., Delaunay, well-centered, structured vs. unstructured), (iii) element type (simplicial vs. non-simplicial elements), (iv)

mesh size (i.e., aspect ratio), (v) medium properties (e.g., anisotropy, heterogeneity), (vi) order of approximation (i.e., low-order vs. high-order), and (vii) temporal discretization (e.g., time stepping scheme, selection of the time-step). The first six factors are equally applicable to steady anisotropic diffusion equation. Systematic studies on the effect of first five factors on maximum principles and the non-negative constraint can be found in [46, 49, 50]. The sixth factor in the context of steady diffusion equation has been discussed in [57]. The last factor (in combination with other six factors) is the subject matter of this paper.

This leads to the problem statement of this paper: *Develop a finite element methodology for linear transient tensorial diffusion equation that satisfies maximum principles and the non-negative constraint on general computational grids for low-order finite elements with no additional restrictions on the time-step.* To the best of our knowledge, such a methodology does not exist in the literature. In the next section, we shall extend the optimization-based methodologies that are presented in [49, 50] for steady diffusion equations to transient diffusion equation. We shall explicitly enforce constraints on the nodal concentrations to satisfy maximum principles and the non-negative. We shall restrict to low-order finite elements. It needs to emphasized that the proposed methodology is *not* applicable to high-order elements. The reason being that the interpolation functions could change their sign within an (finite) element, and hence enforcing non-negative constraints at nodes does not imply non-negative concentrations throughout the domain for high-order elements [57].

3 Proposed numerical methodology: Derivation and implementation details

Herein, we shall employ the method of horizontal lines (also known as the Rothe method) [65] as opposed to the commonly employed method of vertical lines [33]. The method of horizontal lines is a discretization sequence in which the time is discretized first followed by spatial discretization. To this end, we shall define two sets of time levels: *integral* and *weighted* time levels. The time interval of interest $[0,\mathcal{I}]$ is divided into N non-overlapping subintervals such that

$$[0,\mathcal{I}] = \bigcup_{n=1}^{N} [t_{n-1}, t_n], \tag{3.1}$$

where t_n $(n = 0,\cdots,N)$ are referred to as integral time levels. For convenience, we shall assume that the time-step Δt to be uniform, which implies that

$$\Delta t = \frac{\mathcal{I}}{N} \quad \text{and} \quad t_n = n\Delta t. \tag{3.2}$$

However, it should be noted that the proposed methodology can be easily extended to non-uniform time-steps. We shall apply the method of horizontal lines at weighted time

levels, which are defined as follows:

$$t_{n+\eta} := (1-\eta)t_n + \eta t_{n+1}, \tag{3.3}$$

where the parameter $\eta \in [0,1]$. The concentration and its rate at integral time levels are respectively denoted as follows:

$$c^{(n)}(\mathbf{x}) = c(\mathbf{x}, t = t_n), \tag{3.4a}$$

$$v^{(n)}(\mathbf{x}) = \frac{\partial c}{\partial t}(\mathbf{x}, t = t_n). \tag{3.4b}$$

The following notation is used to denote quantities at weighted time levels:

$$c^{(n+\eta)}(\mathbf{x}) := (1-\eta)c^{(n)}(\mathbf{x}) + \eta c^{(n+1)}(\mathbf{x}) \approx c(\mathbf{x}, t_{n+\eta}), \tag{3.5a}$$

$$v^{(n+\eta)}(\mathbf{x}) := (1-\eta)v^{(n)}(\mathbf{x}) + \eta v^{(n+1)}(\mathbf{x}) \approx \frac{\partial c}{\partial t}(\mathbf{x}, t = t_{n+\eta}), \tag{3.5b}$$

$$c_p^{(n+\eta)}(\mathbf{x}) := c_p(\mathbf{x}, t_{n+\eta}), \tag{3.5c}$$

$$f^{(n+\eta)}(\mathbf{x}) := f(\mathbf{x}, t_{n+\eta}), \tag{3.5d}$$

$$q_p^{(n+\eta)}(\mathbf{x}) := q_p(\mathbf{x}, t_{n+\eta}). \tag{3.5e}$$

3.1 Derivation

In designing the proposed methodology, attention will be exercised on two different aspects. The first aspect is to make sure that the non-negative constraint and maximum principles are preserved after both temporal and spatial discretizations. The second aspect is to achieve numerical stability in solving the resulting differential-algebraic equations. As we shall see in Subsection 3.2, we will be adding additional equations in the form of lower and upper bounds (i.e., inequality constraints). This implies that we will be dealing with differential-algebraic equations. It is important to note that numerical time integration schemes that are designed for ordinary differential equations may not be stable and accurate for solving differential-algebraic equations. This point has been discussed adequately in the literature (e.g., see Refs. [6, 26, 27]). An important work on numerical time integration of differential-algebraic equations is by Petzold [58], and the title of this paper ("Differential/algebraic equations are not ODEs") succinctly summarizes the above discussion.

We shall employ the generalized-α method for temporal discretization. The generalized-α method was first proposed for second-order transient systems in Ref. [16], and later modified for first-order transient systems in Ref. [36]. It will be shown that the entire family of time stepping schemes under the generalized-α method may not be made to satisfy maximum principles and the non-negative constraint. The following derivation will reveal that only a subset of time stepping schemes satisfying $\alpha_f = 1$ and $\alpha_m = \gamma \in (0,1]$ will be suitable.

Remark 3.1. The elimination of some time stepping schemes under the generalized-α family (e.g., explicit schemes like the forward Euler) to meet maximum principles should not be considered as a limitation of the proposed methodology. It should be interpreted that some time stepping schemes are not suitable to enforce maximum principles and the non-negative constraint. This in similar to fact that all the time stepping schemes under the Newmark family are not energy preserving [33]. If one want to preserve energy and employ a time stepping scheme from the Newmark family, the only choice will be the trapezoidal rule. Similarly, one can employ other time stepping schemes (i.e., time stepping schemes not satisfying $\alpha_f = 1$ and $\alpha_m = \gamma$) to solve transient diffusion equations, but the numerical results need not satisfy maximum principles and the non-negative constraint.

After applying the generalized-α method to the governing equations (2.1a)-(2.1c), we obtain the following equations:

$$v^{(n+\alpha_m)}(\mathbf{x}) - \operatorname{div}[\mathbf{D}(\mathbf{x})\operatorname{grad}[c^{(n+\alpha_f)}]] = f^{(n+\alpha_f)}(\mathbf{x}) \quad \text{in } \Omega, \tag{3.6a}$$

$$c^{(n+\alpha_f)}(\mathbf{x}) = c_p^{(n+\alpha_f)}(\mathbf{x}) \quad \text{on } \Gamma^{\mathrm{D}}, \tag{3.6b}$$

$$\widehat{\mathbf{n}}(\mathbf{x}) \cdot \mathbf{D}(\mathbf{x})\operatorname{grad}[c^{(n+\alpha_f)}] = q_p^{(n+\alpha_f)}(\mathbf{x}) \quad \text{on } \Gamma^{\mathrm{N}}, \tag{3.6c}$$

where the parameters $\alpha_m, \alpha_f \in [0,1]$. In addition, we have the following relationship:

$$c^{(n+1)}(\mathbf{x}) = c^{(n)}(\mathbf{x}) + \Delta t \left((1-\gamma) v^{(n)}(\mathbf{x}) + \gamma v^{(n+1)}(\mathbf{x}) \right), \tag{3.7}$$

where the parameter $\gamma \in [0,1]$. The initial condition takes the following form:

$$c^{(0)}(\mathbf{x}) = c_0(\mathbf{x}) \quad \text{in } \Omega. \tag{3.8}$$

Remark 3.2. Many popular time stepping schemes are special case of generalized-α method. For example, forward Euler ($\alpha_m = 1, \alpha_f = 1, \gamma = 0$), trapezoidal rule ($\alpha_m = 1, \alpha_f = 1, \gamma = 1/2$), and backward Euler ($\alpha_m = 1, \alpha_f = 1, \gamma = 1$).

Herein, we shall take $\alpha_m = \gamma$. This selection is intended to inherit the non-negative property for the resulting time discrete equations. The time discrete equations in terms of concentration take the following form: Find $c^{(n+\alpha_f)}(\mathbf{x})$ such that we have

$$\frac{1}{\alpha_f \Delta t} c^{(n+\alpha_f)}(\mathbf{x}) - \operatorname{div}[\mathbf{D}(\mathbf{x})\operatorname{grad}[c^{(n+\alpha_f)}]] = f^{(n+\alpha_f)}(\mathbf{x}) + \frac{1}{\alpha_f \Delta t} c^{(n)}(\mathbf{x}) \quad \text{in } \Omega, \tag{3.9a}$$

$$c^{(n+\alpha_f)}(\mathbf{x}) = c_p^{(n+\alpha_f)}(\mathbf{x}) \quad \text{on } \Gamma^{\mathrm{D}}, \tag{3.9b}$$

$$\widehat{\mathbf{n}}(\mathbf{x}) \cdot \mathbf{D}(\mathbf{x})\operatorname{grad}[c^{(n+\alpha_f)}] = q_p^{(n+\alpha_f)}(\mathbf{x}) \quad \text{on } \Gamma^{\mathrm{N}}. \tag{3.9c}$$

The above boundary value problem (3.9a)-(3.9c) is a second-order inhomogeneous elliptic partial differential equation with Dirichlet and Neumann boundary conditions.

Specifically, Eq. (3.9a) is the well-known steady-state anisotropic diffusion equation with decay, as $\alpha_f \Delta t$ will be always positive. The decay coefficient can be identified as $1/(\alpha_f \Delta t)$, and the volumetric source term is $f^{(n+\alpha_f)}(\mathbf{x}) + \frac{1}{\alpha_f \Delta t} c^{(n)}(\mathbf{x})$. This boundary value problem is also known to satisfy maximum principles and the non-negative constraint. The selection $\alpha_m = \gamma$ made it possible to preserve maximum principles and the non-negative constraint by ensuring the decay coefficient to be positive, and the volumetric source at discrete time levels to be non-negative.

It should be emphasized that an arbitrary temporal discretization will not preserve maximum principles and the non-negative constraint. An important aspect is to ensure that the resulting equation after a temporal discretization of transient diffusion equation (2.1a) is a diffusion equation with decay instead of a Helmholtz equation. Diffusion equation with decay takes the following form:

$$\alpha(\mathbf{x}) c(\mathbf{x}) - \operatorname{div}[\mathbf{D}(\mathbf{x}) \operatorname{grad}[c]] = f(\mathbf{x}), \tag{3.10}$$

with $\alpha(\mathbf{x}) \geq 0$. If $\alpha(\mathbf{x}) < 0$, the equation is referred to as Helmholtz equation. It should be noted that Helmholtz equation does not have a maximum principle similar to the one possessed by diffusion equation with decay [24]. Hence, in order to preserve maximum principles and the non-negative constraint, the temporal discretization based on the method of horizontal lines should be carried out in such a way that the resulting decay coefficient is non-negative. In Appendix, we shall outline several other ways of carrying out temporal discretization, and discuss the drawbacks of such approaches in meeting maximum principles and the non-negative constraint.

Recently, Nagarajan and Nakshatrala [49] have proposed a procedure for enforcing maximum principles and the non-negative constraint for steady diffusion with decay equation, which we shall modify to solve Eqs. (3.9a)-(3.9c). We start by applying Galerkin formalism to Eqs. (3.9a)-(3.9c). The corresponding weak form takes the following form: Find $c^{(n+\alpha_f)}(\mathbf{x}) \in \mathcal{P}_{n+\alpha_f}$ such that we have

$$\begin{aligned} &\int_\Omega w(\mathbf{x}) \frac{1}{\alpha_f \Delta t} c^{(n+\alpha_f)}(\mathbf{x}) \, \mathrm{d}\Omega + \int_\Omega \operatorname{grad}[w] \cdot \mathbf{D}(\mathbf{x}) \operatorname{grad}\left[c^{(n+\alpha_f)}\right] \mathrm{d}\Omega \\ &= \int_\Omega w(\mathbf{x}) \left(f^{(n+\alpha_f)}(\mathbf{x}) + \frac{1}{\alpha_f \Delta t} c^{(n)}(\mathbf{x}) \right) \mathrm{d}\Omega + \int_{\Gamma^{\mathrm{N}}} w(\mathbf{x}) q_p^{(n+\alpha_f)}(\mathbf{x}) \, \mathrm{d}\Gamma \quad \forall w(\mathbf{x}) \in \mathcal{Q}, \end{aligned} \tag{3.11}$$

where the function spaces $\mathcal{P}_{n+\alpha_f}$ and $\mathcal{Q}$ are defined as follows:

$$\mathcal{P}_{n+\alpha_f} := \left\{ c(\mathbf{x}) \in H^1(\Omega) \,\middle|\, c(\mathbf{x}) = c_p^{(n+\alpha_f)}(\mathbf{x}) \text{ on } \Gamma^{\mathrm{D}} \right\}, \tag{3.12a}$$

$$\mathcal{Q} := \left\{ w(\mathbf{x}) \in H^1(\Omega) \,\middle|\, w(\mathbf{x}) = 0 \text{ on } \Gamma^{\mathrm{D}} \right\}. \tag{3.12b}$$

After executing the usual steps of the finite element method, the above weak form (3.11) can be converted to a system of linear equations of the following form:

$$\boldsymbol{K} \boldsymbol{c}^{(n+\alpha_f)} = \boldsymbol{f}^{(n+\alpha_f)}, \tag{3.13}$$

where $\boldsymbol{c}^{(n+\alpha_f)} \in \mathbb{R}^{ndofs}$ denotes the unknown vector containing nodal concentrations at the weighted time level $t_{n+\alpha_f}$, $\boldsymbol{f}^{(n+\alpha_f)} \in \mathbb{R}^{ndofs}$ is a known vector, $\boldsymbol{K}$ is a symmetric and positive definite matrix, and "$ndofs$" denotes the number of (free) degrees-of-freedom. It will be shown in a subsequent section that the finite element solution obtained by solving the system of linear equations (3.13) may not satisfy maximum principles and the non-negative constraint. Using optimization-based techniques, we now modify the above solution procedure to meet these important physical constraints.

3.2 Enforcing maximum principles and the non-negative constraint

We shall denote the standard inner product on finite dimensional Euclidean spaces by $\langle \cdot ; \cdot \rangle$. We shall use the symbols $\preceq$ and $\succeq$ to denote component-wise inequalities for vectors. That is, for any two given (finite dimensional) vectors $\boldsymbol{a}$ and $\boldsymbol{b}$

$$\boldsymbol{a} \preceq \boldsymbol{b} \quad \text{means that} \quad a_i \leq b_i \ \forall i. \tag{3.14}$$

Similarly, one can define the symbol $\succeq$. The optimization problem can then be written as follows:

$$\underset{\boldsymbol{c}^{(n+\alpha_f)} \in \mathbb{R}^{ndofs}}{\text{minimize}} \quad \frac{1}{2} \left\langle \boldsymbol{c}^{(n+\alpha_f)} ; \boldsymbol{K}\boldsymbol{c}^{(n+\alpha_f)} \right\rangle - \left\langle \boldsymbol{c}^{(n+\alpha_f)} ; \boldsymbol{f}^{(n+\alpha_f)} \right\rangle \tag{3.15a}$$

$$\text{subject to} \quad c_{\min}^{(n+\alpha_f)} \mathbf{1} \preceq \boldsymbol{c}^{(n+\alpha_f)} \preceq c_{\max}^{(n+\alpha_f)} \mathbf{1}, \tag{3.15b}$$

where $\mathbf{1}$ is a vector containing ones of size $ndofs \times 1$, and $c_{\min}^{(n+\alpha_f)}$ and $c_{\max}^{(n+\alpha_f)}$ are respectively the lower and upper bounds. For enforcing maximum principles, $c_{\min}^{(n+\alpha_f)}$ and $c_{\max}^{(n+\alpha_f)}$ can be taken as follows:

$$c_{\min}^{(n+\alpha_f)} := \min \left\{ \min_{\mathbf{x} \in \Omega} c_0(\mathbf{x}), \min_{\mathbf{x} \in \partial\Omega} c_p^{(n+\alpha_f)}(\mathbf{x}) \right\}, \tag{3.16a}$$

$$c_{\max}^{(n+\alpha_f)} := \max \left\{ \max_{\mathbf{x} \in \Omega} c_0(\mathbf{x}), \max_{\mathbf{x} \in \partial\Omega} c_p^{(n+\alpha_f)}(\mathbf{x}) \right\}. \tag{3.16b}$$

For problems involving only the non-negative constraint, one can employ the following:

$$c_{\min}^{(n+\alpha_f)} = 0 \, \text{and} \, c_{\max}^{(n+\alpha_f)} = +\infty. \tag{3.17}$$

Alternatively, for enforcing the non-negative constraint, one can replace the constraint (3.15b) with the following:

$$\mathbf{0} \preceq \boldsymbol{c}^{(n+\alpha_f)}, \tag{3.18}$$

where $\mathbf{0}$ denotes the vector of size $ndofs \times 1$ containing zeros. It should be noted that the above optimization problem (3.15) belongs to *quadratic programming*. Since, for the

problem at hand, the matrix $\boldsymbol{K}$ is positive definite (which makes the objective function (3.15a) convex) the optimization problem belongs to *convex quadratic programming*. A sound mathematical theory is already in place for studying convex quadratic programming [10], and several efficient algorithms are available in the literature [10, 55, 68]. In this paper, we shall employ the built-in optimization solver available in MATLAB [4]. Some other popular packages that can handle convex quadratic programming optimization problems are GAMS [3], TAO [48], and DAKOTA [5].

Once the nodal concentrations are obtained at weighted time level, one can obtain the nodal concentrations at integral time levels as follows:

$$\boldsymbol{c}^{(n+1)} = \frac{\boldsymbol{c}^{(n+\alpha_f)} - (1-\alpha_f)\boldsymbol{c}^{(n)}}{\alpha_f}. \tag{3.19}$$

Although $\boldsymbol{c}^{(n+\alpha_f)} \succeq \boldsymbol{0}$, the nodal concentrations at integral time levels based on Eq. (3.19) need not be non-negative if $\alpha_f \neq 1$. To put it differently, one is assured of satisfying maximum principles and the non-negative constraint under the proposed methodology if $\alpha_m = \gamma \in (0,1]$ and $\alpha_f = 1$.

3.2.1 Calculation of the rate of nodal concentrations

There are two ways one could calculate the nodal rates of concentration. The first method is to directly calculate the rate of nodal concentrations at integral time levels using the following expression:

$$\boldsymbol{v}^{(n+1)} = \frac{\boldsymbol{c}^{(n+1)} - \boldsymbol{c}^{(n)} - (1-\gamma)\Delta t \boldsymbol{v}^{(n)}}{\gamma \Delta t}. \tag{3.20}$$

The second method is to first calculate the rates of concentration at weighted time levels using the following expression:

$$\boldsymbol{v}^{(n+\gamma)} = \frac{\boldsymbol{c}^{(n+1)} - \boldsymbol{c}^{(n)}}{\Delta t}. \tag{3.21}$$

The rate of nodal concentrations at the integral time levels can then be calculated using the following consistent approximation:

$$\boldsymbol{v}^{(n+1)} = \gamma \boldsymbol{v}^{(n+\gamma)} + (1-\gamma)\boldsymbol{v}^{(n+1+\gamma)}. \tag{3.22}$$

Both these consistent ways of obtaining the rates of concentration are pictorially described in Fig. 3. It should be emphasized that $\gamma = 0$ cannot be employed under the proposed methodology to meet maximum principles and the non-negative constraint. The various steps involved in the numerical implementation of the proposed methodology to satisfy maximum principles and the non-negative constraint are summarized in Algorithm 1, which could serve as a quick reference during computer code design and implementation.

Algorithm 1 Implementation of the proposed methodology based on $\alpha_f = 1$.

1: Input: Initial condition $c(\mathbf{x})$, Dirichlet boundary conditions $c_p(\mathbf{x},t)$, Neumann boundary conditions $q_p(\mathbf{x},t)$, time-step Δt, total time of interest $\mathcal{I}$, $\alpha_m = \gamma \in (0,1]$.
2: Construct initial nodal concentrations $\boldsymbol{c}^{(0)}$
3: Set $\boldsymbol{c}^{(n)} \longleftarrow \boldsymbol{c}^{(0)}$, $t \longleftarrow 0$, $n \longleftarrow 0$
4: **while** $t < \mathcal{I}$ **do**
5: Calculate $c_{\min}^{(n+1)}$ and $c_{\max}^{(n+1)}$ (see Eqs. (3.16)-(3.17))
6: Call non-negative solver to obtain $\boldsymbol{c}^{(n+1)}$

$$\begin{aligned} &\underset{\boldsymbol{c}^{(n+1)} \in \mathbb{R}^{ndofs}}{\text{minimize}} \quad \frac{1}{2}\langle \boldsymbol{c}^{(n+1)}; \boldsymbol{K}\boldsymbol{c}^{(n+1)} \rangle - \langle \boldsymbol{c}^{(n+1)}; \boldsymbol{f}^{(n+1)} \rangle \\ &\text{subject to} \quad c_{\min}^{(n+1)} \mathbf{1} \preceq \boldsymbol{c}^{(n+1)} \preceq c_{\max}^{(n+1)} \mathbf{1} \end{aligned}$$

7: If needed, obtain the rate of nodal concentrations at both integral and weighed time levels (see Eqs. (3.20)-(3.22) and Fig. 3)

$$\boldsymbol{v}^{(n+1)} = \frac{\boldsymbol{c}^{(n+1)} - \boldsymbol{c}^{(n)} - (1-\gamma)\Delta t \boldsymbol{v}^{(n)}}{\gamma \Delta t}$$

or

$$\boldsymbol{v}^{(n+\gamma)} = \frac{\boldsymbol{c}^{(n+1)} - \boldsymbol{c}^{(n)}}{\Delta t} \quad \text{and} \quad \boldsymbol{v}^{(n+1)} = \gamma \boldsymbol{c}^{(n+\gamma)} + (1-\gamma)\boldsymbol{c}^{(n+1+\gamma)}$$

8: Set $\boldsymbol{c}^{(n)} \longleftarrow \boldsymbol{c}^{(n+1)}$, $t \longleftarrow t + \Delta t$, $n \longleftarrow n+1$
9: **end while**

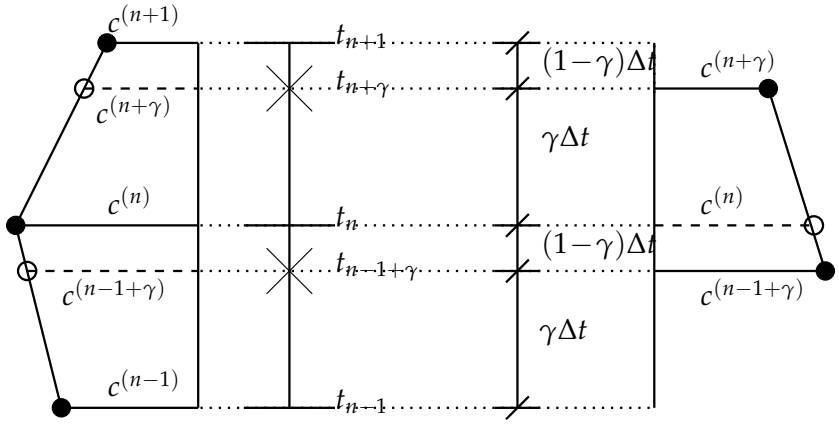


Figure 3: The left part of the figure shows the usual way of interpolating quantities at integral time levels to obtain the corresponding quantities at weighted time levels. That is, $c^{(n+\gamma)} = (1-\gamma)c^{(n)} + \gamma c^{(n+1)}$. The right part of the figure shows the interpolation of quantities at weighted time levels to obtain the corresponding quantities at integral time levels, which is adopted in this paper. That is, $c^{(n)} = \gamma c^{(n-1+\gamma)} + (1-\gamma)c^{(n+\gamma)}$. The interpolated quantities are indicated using hollow circles.

4 Representative numerical results

In all our numerical simulations we have employed low-order finite elements, and have taken $\alpha_f = 1$. It is assumed that $\alpha_m = \gamma = 1$, unless stated otherwise. The specific selection of γ does not appear in the calculation of nodal concentrations. But it will be needed to calculate the rate of nodal concentrations, which is discussed in the previous section.

4.1 One-dimensional problem with uniform initial condition

Consider the following initial boundary value problem:

$$\frac{\partial c(\mathrm{x},t)}{\partial t} - \frac{\partial^2 c(\mathrm{x},t)}{\partial \mathrm{x}^2} = 0 \quad \text{in } \Omega_{\mathcal{I}} := (0,1)\times(0,\mathcal{I}), \tag{4.1a}$$

$$\frac{\partial c(\mathrm{x}=0,t)}{\partial \mathrm{x}} = 0, \quad c(\mathrm{x}=1,t) = 0 \quad \forall t \in (0,\mathcal{I}], \tag{4.1b}$$

$$c(\mathrm{x},0) = 1 \quad \forall \mathrm{x} \in [0,1]. \tag{4.1c}$$

The exact solution is bounded between zero and unity. In the numerical simulation, we have divided the computational domain into five equal linear finite elements, and have taken the time-step to be $\Delta t = 0.001$ (which is chosen arbitrarily). Fig. 4 compares the analytical solution with the numerical solutions obtained using the single-field formulation and the proposed methodology. The single-field formulation violates the maximum principle, as the obtained numerical solution is greater than unity. The proposed methodology satisfies the maximum principle for all times. The rate of nodal concentrations under the proposed methodology are shown in Fig. 5 for various values of $\gamma = 0.1$, 0.5 and 1.0. Note that under the proposed methodology $\alpha_m = \gamma$. We have employed the second method for calculating the rates (i.e., Eqs. (3.21)-(3.22)).

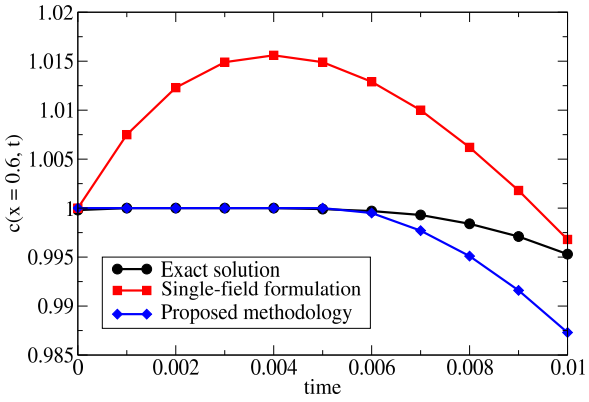


Figure 4: One dimensional problem with uniform initial condition: The numerical solution from the single-field formulation exceeds unity while the proposed methodology satisfies the maximum principle. From the maximum principle, it is known that the analytical solution is bounded above by unity.

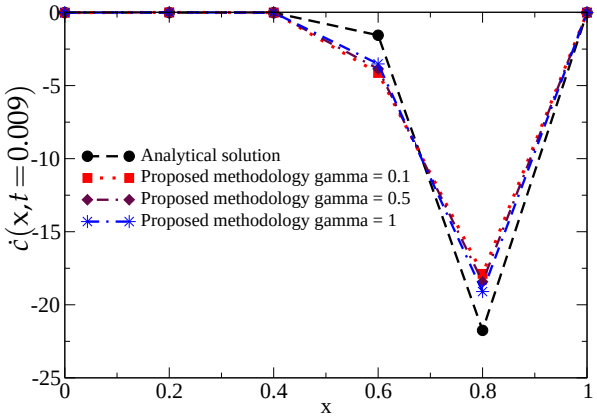


Figure 5: One dimensional problem with uniform initial condition: This figure shows the rate of nodal concentrations at $t=0.009$ as a function of x under the proposed methodology for various values of γ.

4.1.1 On critical time-steps

For one-dimensional problems, it is possible to perform further analysis on the choice of time-steps to meet numerical stability and maximum principles. The critical time-step due to Courant-Friedrichs-Levy (CFL) stability condition can be written as follows:

$$\Delta t \leq \Delta t_{\text{CFL}} = \frac{h^2}{2D}. \tag{4.2}$$

For this problem, $\Delta t_{\text{CFL}} = 0.02$, as $D=1$ and $h=0.2$. The critical time-step to satisfy maximum principles will be $\Delta t \geq \Delta t_{\text{MP}}$, where

$$\Delta t_{\text{MP}} = \frac{h^2}{6\gamma D} = \begin{cases} 6.67\times10^{-2} & \text{for } \gamma=0.1, \\ 1.33\times10^{-2} & \text{for } \gamma=0.5, \\ 6.67\times10^{-3} & \text{for } \gamma=1.0. \end{cases} \tag{4.3}$$

As one can see from this simple analysis, a time-step which meets the CFL stability condition could still violate maximum principles and the non-negative condition. To put it differently, depending on the choice of the time-stepping scheme, the critical time-step due to CFL stability condition could be either smaller or bigger than the critical time-step due to maximum principles. However, as mentioned in Introduction, a striking difference between these two critical time-steps is that the CFL condition places an upper bound on the time-step, whereas maximum principles place a lower bound on the time-step. It should be emphasized that it is not always possible to find restrictions on the time-step for two- and three-dimensional problems. More importantly, some of the more complicated numerical examples involving anisotropy (which are presented later in this section) clearly show that there may not exist a suitable time-step using which the numerical results satisfy maximum principles and the non-negative constraint.

4.2 One-dimensional problem with non-uniform initial condition

Consider the following simple one-dimensional problem with homogeneous forcing function. This problem is a modification to one of the examples given in Ref. [18]. The initial boundary value problem can be written as follows:

$$\frac{\partial c(\mathrm{x},t)}{\partial t} - \frac{\partial^2 c(\mathrm{x},t)}{\partial \mathrm{x}^2} = 0 \quad \text{in } \Omega_{\mathcal{I}} := (0,1)\times(0,\mathcal{I}), \tag{4.4a}$$

$$c(\mathrm{x}=0,t) = c(\mathrm{x}=1,t) = 0 \quad \forall t \in (0,\mathcal{I}], \tag{4.4b}$$

$$c(\mathrm{x},0) = \begin{cases} 1 & \text{if } \mathrm{x}\in[a,b], \\ 0 & \text{otherwise.} \end{cases} \tag{4.4c}$$

Herein, we have taken $a=0.4$ and $b=0.6$.

Fig. 6 shows that the numerical solution from the proposed methodology compares well point-wise with the analytical solution, and satisfies the maximum principle and the non-negative constraint. Fig. 7 shows the numerical convergence of the proposed methodology and the standard single-field formulation in L_2-norm and H^1-seminorm. Note that the convergence in L_2-norm and H^1-seminorm is in the integral sense, and it need not imply point-wise convergence. The convergence study is carried out by employing simultaneous spatial and temporal refinements satisfying the condition $\Delta t \propto (\Delta \mathrm{x})^2$. The coarsest mesh has 11 nodes, and the corresponding time-step used for this mesh is $\Delta t = 10^{-3}$.

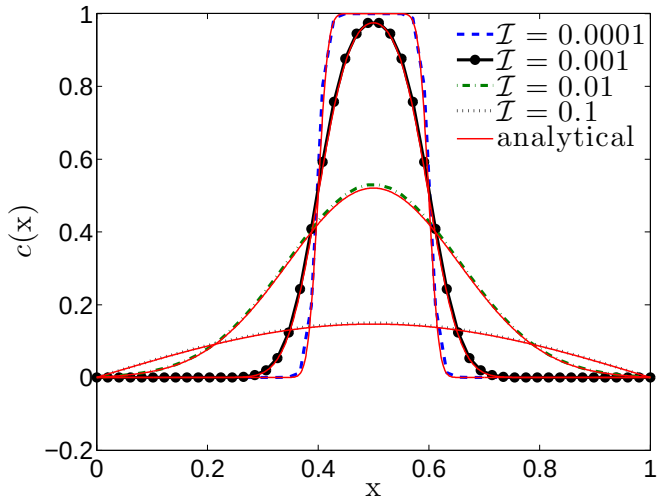
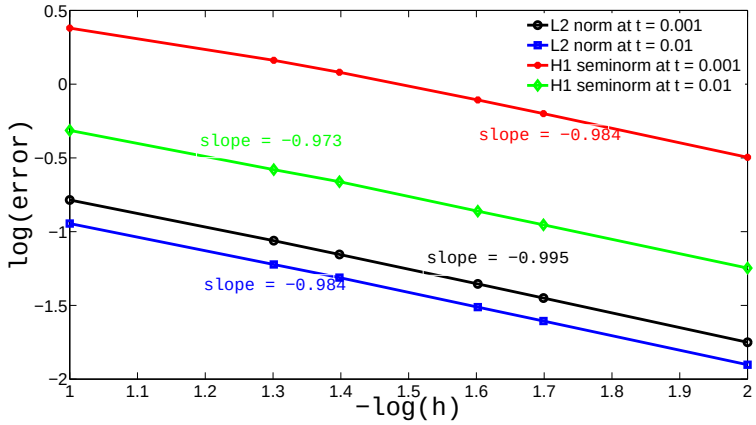
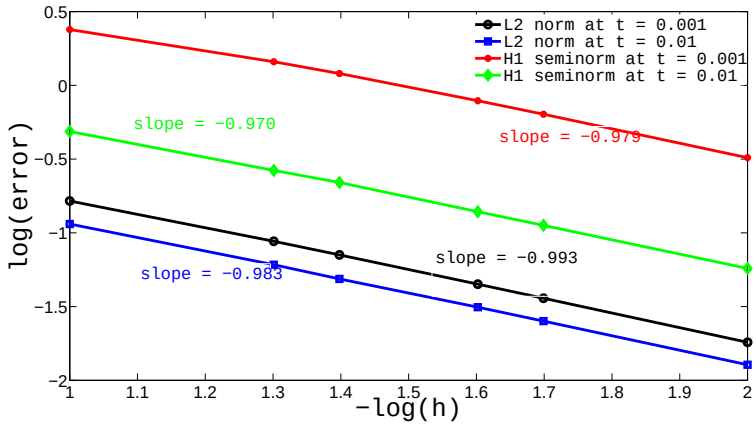


Figure 6: One-dimensional problem with non-uniform initial condition: This figure compares the concentration obtained using the proposed methodology with the analytical solution at various instants of time. For this test problem, the solution should be between zero and unity. The time step used in the numerical simulation is $\Delta t = 10^{-4}$. As one can see from the figure, the proposed methodology performed well, and it did not violate the maximum principle and the non-negative constraint.



(a) single-field formulation



(b) proposed methodology

Figure 7: One-dimensional problem with non-uniform initial condition: This figure compares the numerical convergence of the single-field formulation (top figure) and the proposed non-negative methodology (bottom figure) with simultaneous spatial and temporal refinements such that $\Delta t \propto (\Delta x)^2$. In this numerical simulation, we have taken $\gamma = 1$. The convergence is carried out at two different time levels: $t = 0.001$ and $t = 0.01$. The coarsest mesh has 11 nodes, and the corresponding time step used for this mesh is $\Delta t = 10^{-3}$. The terminal rates of convergence in L_2-norm and H^1-seminorm are also indicated in the figure.

Fig. 8 shows the variation of the minimum and maximum concentrations in the domain with respect to time for a fixed mesh but for different time-steps under the standard single-field formulation. Note that for this problem the minimum concentration should be zero, and the maximum concentration should be unity. Clearly, the results from the standard single-field formulation violated both the upper and lower bounds. Fig. 9 shows the effect of mesh refinement for a fixed time-step on the violation of the maximum principle under the single-field formulation. For a given mesh, the extent of

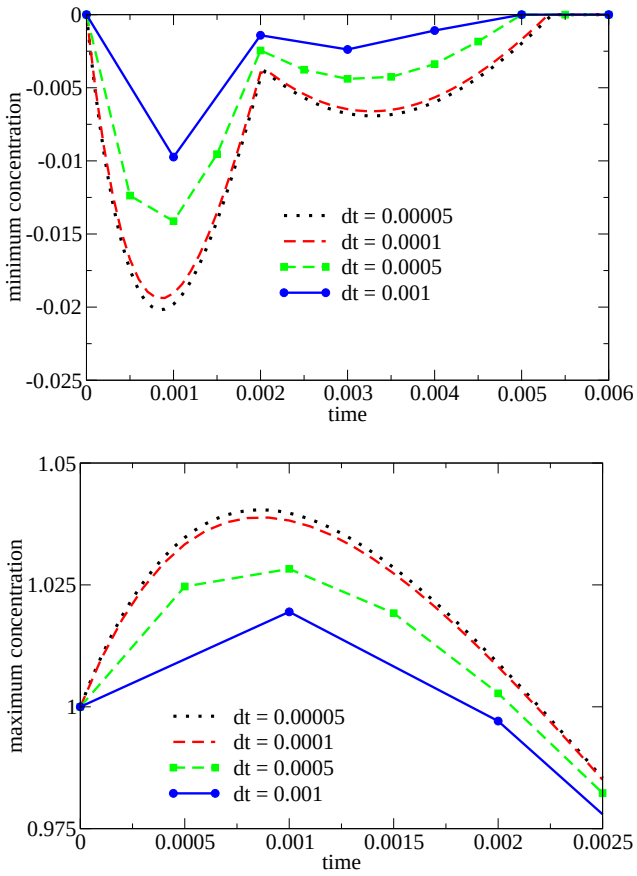


Figure 8: One-dimensional problem with non-uniform initial condition: This figure shows the variation of the *minimum* (top) and *maximum* (bottom) concentrations with respect to time for a given mesh and for different time steps under the *single-field formulation*. According to the maximum principle, the concentration should lie between 0 and 1. The domain is divided into 10 equal-sized two-node finite elements. The time steps used in the numerical simulation are indicated in the figure. The single-field formulation violated both the minimum (which is the non-negative constraint) and the maximum values that are given by the maximum principle. *It should also be noted that the violations increase in magnitude with a decrease in the time step.*

the violation will be greater for smaller time-steps. On the other hand, for a given time-step, the extent of the violation decreases with mesh refinement, which will *not* be the trend in the case of anisotropy. (For example, see the test problems given in Subsections 4.4 and 4.5.) In all the cases considered, the proposed methodology produced concentrations that satisfy the maximum principle and the non-negative constraint.

4.3 Two-dimensional problem with non-uniform initial condition

This test problem is a two-dimensional extension of the problem described earlier in Subsection 4.2. The governing equations take the following form:

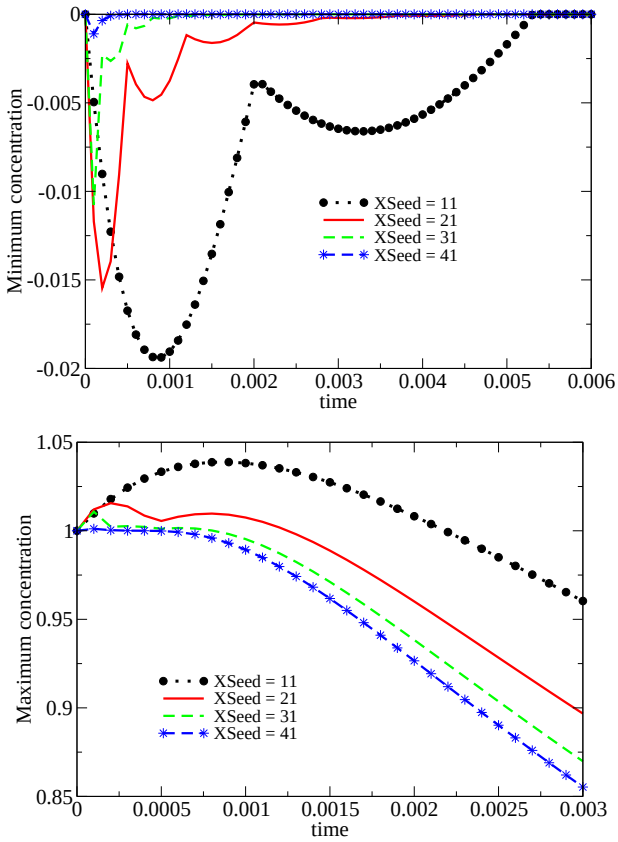


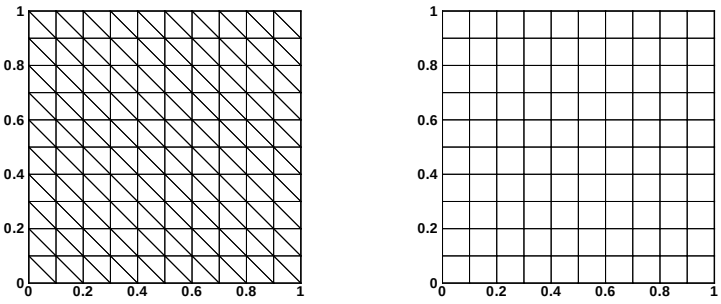
Figure 9: One-dimensional problem with non-uniform initial condition: This figure shows the variation of the *minimum* (top) and *maximum* (bottom) concentrations with respect to time for a given time step and for different mesh refinements under the *single-field formulation*. The time step is taken as $\Delta t = 10^{-4}$. According to the maximum principle, the concentration should lie between 0 and 1. Various meshes are used (XSeed=11, 21, 31 and 41). Note that XSeed denotes the number of nodes. The single-field formulation violated both the minimum (which is the non-negative constraint) and the maximum values that are given by the maximum principle. *It should be noted that the violations of the maximum principle decrease with mesh refinement if the diffusion is anisotropic [49].*

$$\frac{\partial c(\mathrm{x},\mathrm{y},t)}{\partial t} - \left(\frac{\partial^2 c(\mathrm{x},\mathrm{y},t)}{\partial \mathrm{x}^2} + \frac{\partial^2 c(\mathrm{x},\mathrm{y},t)}{\partial \mathrm{y}^2} \right) = 0 \quad \text{in } \Omega_{\mathcal{I}} := (0,1)\times(0,1)\times(0,\mathcal{I}), \tag{4.5a}$$

$$c(\mathrm{x}=0,\mathrm{y},t) = c(\mathrm{x}=1,\mathrm{y},t) = 0, \quad c(\mathrm{x},\mathrm{y}=0,t) = c(\mathrm{x},\mathrm{y}=1,t) = 0, \tag{4.5b}$$

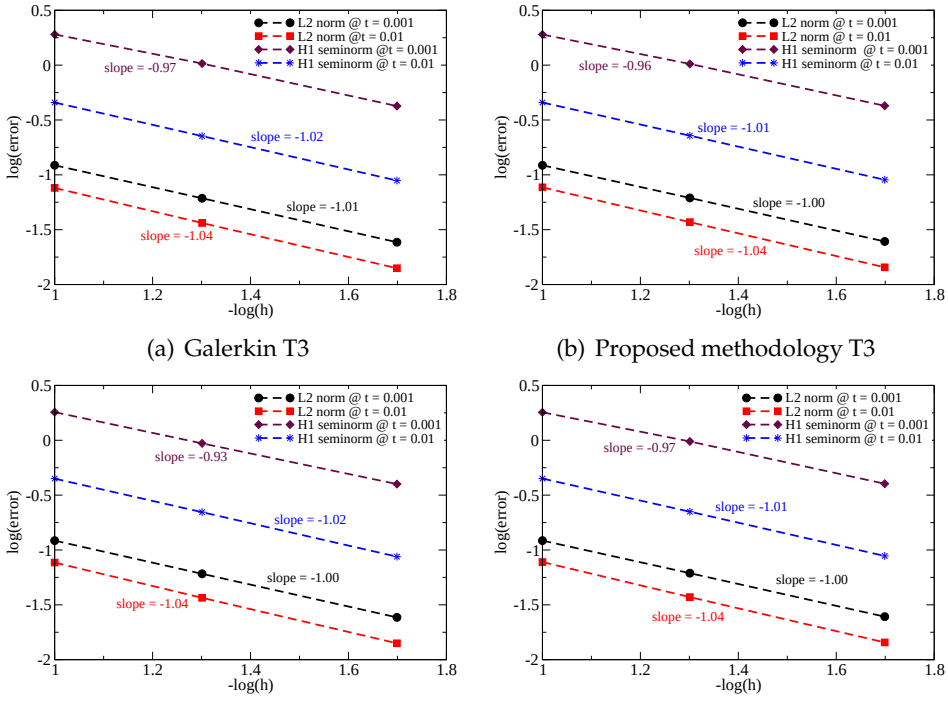
$$c(\mathrm{x},\mathrm{y},0) = \begin{cases} 1 & \text{if } \mathrm{x} \in [0.4,0.6]\times[0.4,0.6], \\ 0 & \text{otherwise.} \end{cases} \tag{4.5c}$$

Hierarchical meshes are employed in this numerical h-convergence study, which are illustrated in Fig. 10. The numerical h-convergence of the proposed methodology in L_2-norm and H^1-seminorm is shown in Fig. 11. The performance of the proposed methodology



(a) T3 finite element mesh (b) Q4 finite element mesh

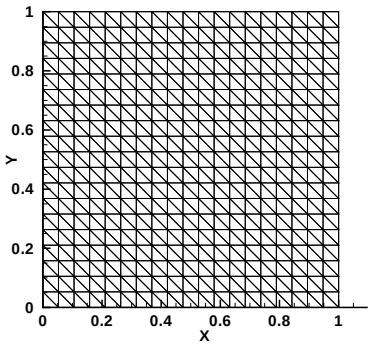
Figure 10: Two-dimensional problem with non-uniform initial condition: This figure shows the typical meshes used in the h-numerical convergence studies. The meshes in this figure have XSeed=YSeed=11, which denote the number of nodes along the x-direction and y-direction. The convergence study employs hierarchical meshes.



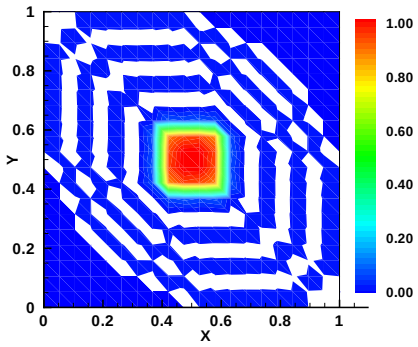
(a) Galerkin T3 (b) Proposed methodology T3

(c) Galerkin Q4 (d) Proposed methodology Q4

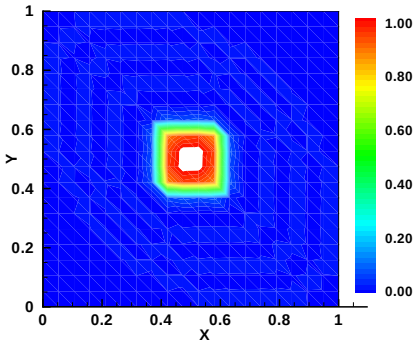
Figure 11: Two-dimensional problem with non-uniform initial condition: This figure illustrates the numerical convergence of the Galerkin single-field formulation and the proposed methodology for three-node triangular element (T3) and four-node quadrilateral element (Q4). We have taken $\gamma = 1$. The numerical convergence is performed at two time levels $t = 0.001$ and $t = 0.01$. A hierarchy of meshes are employed in the numerical study. The initial mesh has XSeed=YSeed=11, which denote the number of nodes along the x-direction and y-direction. The corresponding time step for this mesh is taken as $\Delta t = 0.001$. The mesh and the time step are simultaneously refined as $\Delta t \propto (\Delta x)^2$. The terminal rates of convergence in L_2-norm and H^1-seminorm are indicated in the figure.



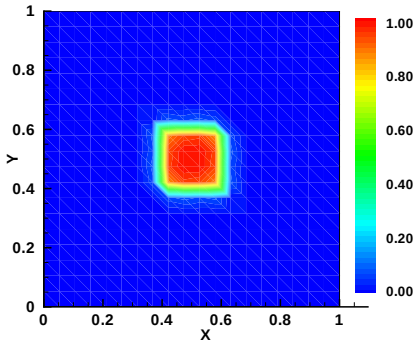
(a) 20×20 equally spaced three-node triangular mesh



(b) Violation of the non-negative constraint



(c) Violation of the maximum principle



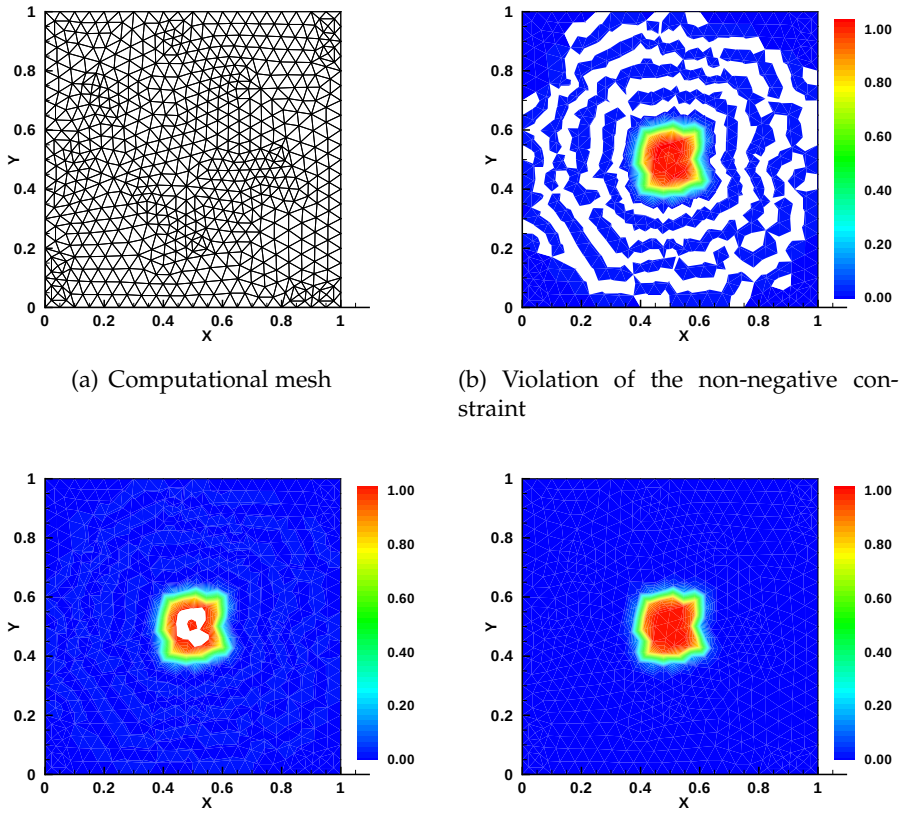
(d) Proposed methodology

Figure 12: Two-dimensional problem with non-uniform initial condition: This figure compares the concentrations obtained from the single-field formulation and the proposed methodology with the analytical solution at time level $t=\Delta t=10^{-4}$. Subfigure (b) shows that the single-field formulation violates the non-negative constraint, as 36% of nodes have negative concentrations. The obtained minimum concentration is -0.01221. Subfigure (c) shows that the single-field formulation violates the maximum principle, as 1% of nodes having concentrations greater than unity. The obtained maximum concentration is 1.02039. Subfigure (d) shows that the concentration obtained from the proposed methodology satisfies the maximum principle, and the non-negative constraint. In subfigure (b), the regions with negative concentrations are indicated in white color. In subfigure (c), the regions with concentrations greater than unity are indicated in white color.

is compared with that of the single-field formulation and MATLAB's PDE Toolbox [4] in Figs. 12 and 13.

4.4 Transient anisotropic diffusion in square plate with a hole

The computational domain is given by $\Omega := (0,1)\times(0,1) - [0.45,0.55]\times[0.45,0.55]$. The initial concentration in the domain is taken to be zero (i.e., $c_0(\mathbf{x})=0$). The volumetric source is zero (i.e., $f(\mathbf{x},t)=0$). The inner boundary is prescribed with a constant concentration of unity, and the outer boundary is prescribed with a constant concentration of zero. The



(a) Computational mesh

(b) Violation of the non-negative constraint

(c) Violation of the maximum principle

(d) Proposed methodology

Figure 13: Two-dimensional problem with non-uniform initial condition: This figure compares the numerical solutions from MATLAB's PDE Toolbox and the proposed methodology at time level $t=\Delta t=10^{-4}$. Subfigure (a) shows the computational mesh used in the numerical simulation. Subfigure (b) shows that numerical solution from the MATLAB's PDE Toolbox violates the non-negative constraint, as 40% of the nodes have negative concentrations. The regions with negative concentrations are indicated in white color. The obtained minimum concentration is -0.0339. Subfigure (c) shows that the numerical solution from MATLAB's PDE Toolbox violates the maximum principle, as 1.2% of nodes have concentrations greater than unity. The regions with concentrations greater than unity are indicated in white color. The obtained maximum concentration is 1.0397. Subfigure (d) shows that the proposed methodology satisfies the maximum principle and the non-negative constraint on the computational mesh generated by MATLAB.

diffusivity tensor is taken as follows:

$$\mathbf{D}(\mathbf{x})=\mathbf{R}\mathbf{D}_0\mathbf{R}^{\mathrm{T}}, \tag{4.6}$$

where $\mathbf{D}_0$ and the rotation tensor are, respectively, defined as follows:

$$\mathbf{D}_0=\begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix}, \tag{4.7a}$$

$$\mathbf{R}=\begin{pmatrix} +\cos(\theta) & -\sin(\theta) \\ +\sin(\theta) & +\cos(\theta) \end{pmatrix}, \tag{4.7b}$$

with the values $k_1=10$, $k_2=10^{-3}$ and $\theta=-\pi/6$. Using the maximum principle given by Theorem 2.1, it can be concluded that the concentration in the domain should be between zero and unity. This test problem is used to illustrate the following aspects:

(i) The numerical results from COMSOL [47] (which is a popular commercial finite element software package) do not satisfy the maximum principle and the non-negative constraint for transient anisotropic diffusion.

(ii) The proposed methodology satisfies the maximum principle and the non-negative constraint even on unstructured meshes with no additional restrictions on the time-step.

(iii) The approach of using the backward Euler time-stepping scheme with lumped capacity matrix does not guarantee non-negative solutions in the case of anisotropic diffusion.

Using numerical simulations it has been found that the transient solution is very close to the steady-state solution for time greater than 0.05. Therefore, the time-steps for this test problem are chosen to be smaller than or equal to 0.05 so that they are appropriate for transient analyses.

We first show the results obtained using COMSOL [47]. Two different meshes are employed in the numerical simulations, which are shown in Fig. 14. The variation of the minimum concentration with time is shown in Fig. 15, and the numerical results from COMSOL did not satisfy the non-negative constraint. Figs. 16 and 17 show the spread of the violation of the non-negative constraint and the concentration profiles using COMSOL for four-node structured mesh and three-node unstructured mesh, respectively. From these figures, the following two observations can be made:

(a) The magnitude of the violation of the non-negative constraint increases as the time-step decreases.

(b) The violation reaches a steady-state value after sufficient time, which is around $t=0.05$ for this problem. It should be emphasized that this steady-state value for minimum concentration is a significant non-negative number, and the violation of the non-negative constraint is nearly 5%.

The aforementioned problem is also solved using the proposed methodology. Fig. 18 shows the unstructured computational meshes used in the numerical simulation. The concentration profiles obtained under the proposed methodology using these computational meshes are shown in Figs. 19 and 20. Clearly, the proposed methodology satisfies the maximum principle and the non-negative constraint at all time levels. Fig. 21 clearly shows that the approach of employing the backward Euler time-stepping scheme

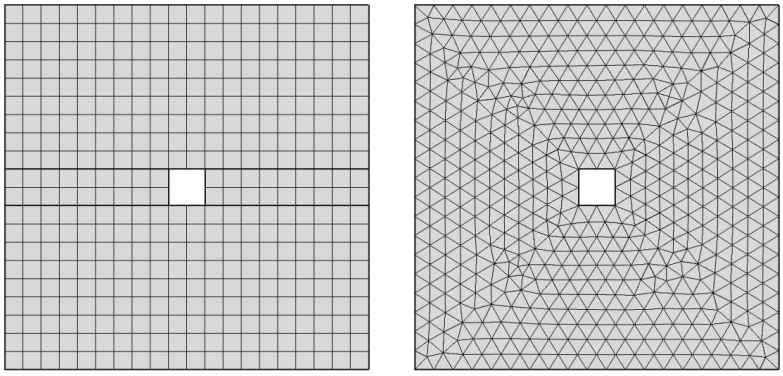


Figure 14: Anisotropic diffusion in a square plate with a hole: This figure shows the meshes employed in the numerical simulations using COMSOL [47]. The left figure shows a structured mesh based on four-node quadrilateral elements, and the right figure shows an unstructured mesh based on three-node triangular elements.

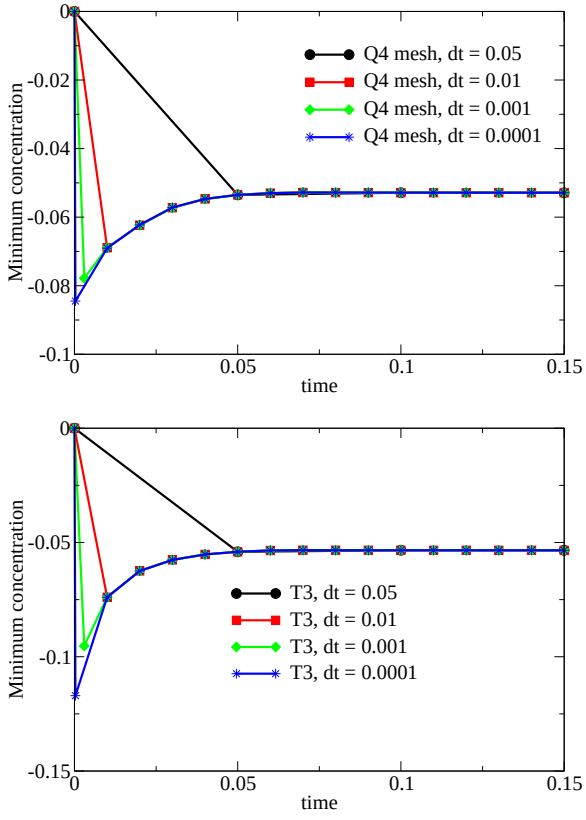
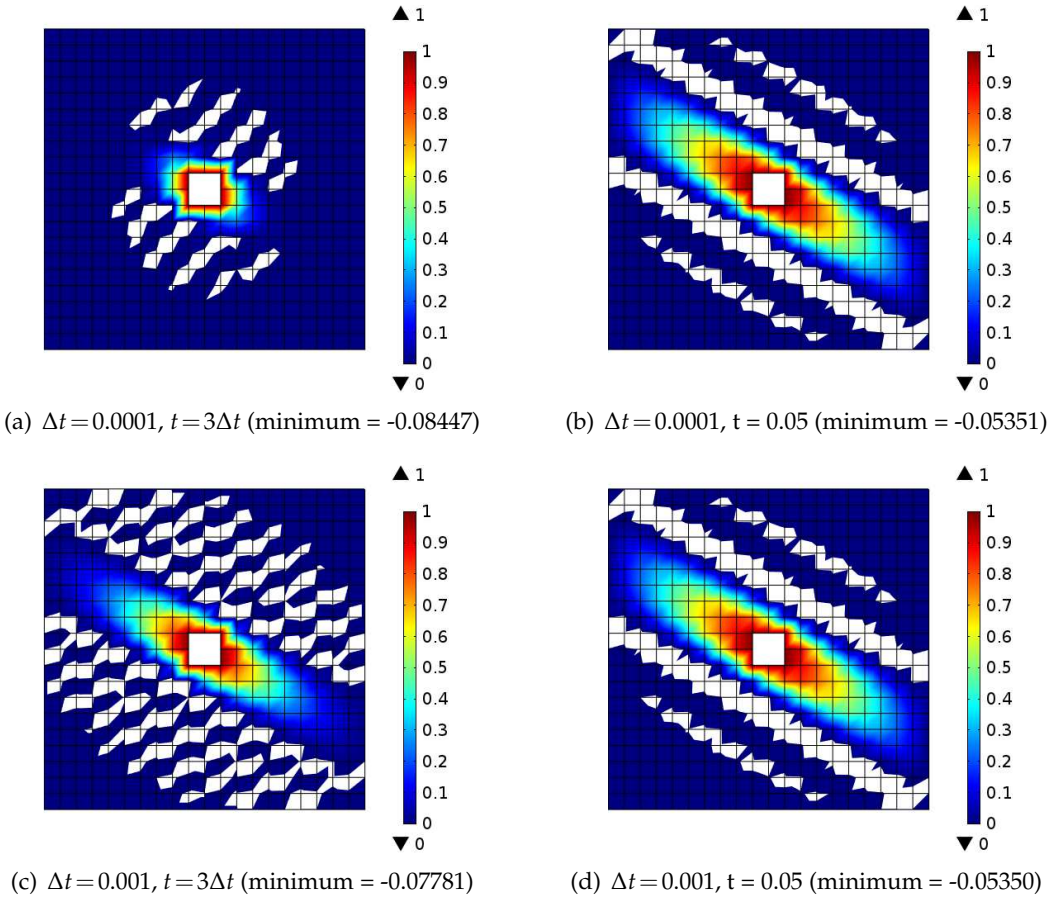


Figure 15: Anisotropic diffusion in square plate with a hole: This figure shows the variation of minimum concentration with time under the meshes shown in Fig. 14. COMSOL [47] is employed in the numerical simulation. The solution is very close to the steady-state response for time greater than 0.05.



(a) $\Delta t = 0.0001$, $t = 3\Delta t$ (minimum = -0.08447) (b) $\Delta t = 0.0001$, t = 0.05 (minimum = -0.05351)

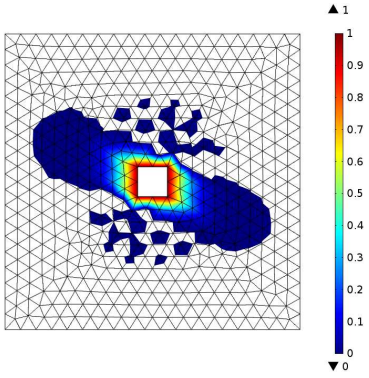
(c) $\Delta t = 0.001$, $t = 3\Delta t$ (minimum = -0.07781) (d) $\Delta t = 0.001$, t = 0.05 (minimum = -0.05350)

Figure 16: Anisotropic diffusion in square plate with a hole: Concentration profiles using COMSOL [47] by employing *structured four-node quadrilateral mesh*. The finite element mesh is also shown. The numerical results clearly violated the non-negative constraint for the concentration. The regions that violated the non-negative constraint are indicated in white color.

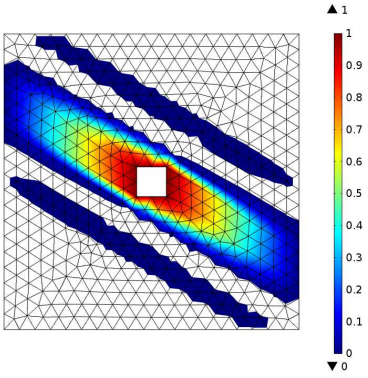
with lumped capacity matrix is not sufficient to meet the maximum principle and the non-negative constraint in the case of transient anisotropic diffusion. This approach will work in the case of transient isotropic diffusion provided some restrictions on the mesh are met, which is discussed briefly in Appendix of this paper.

4.5 Diffusion in heterogeneous anisotropic medium

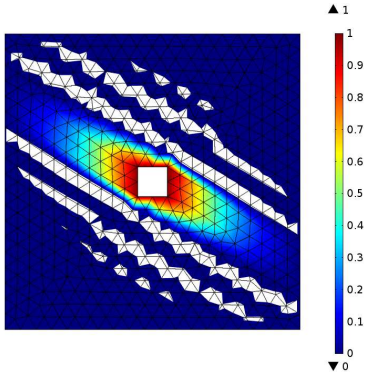
This problem considers transient diffusion in a bi-unit square domain with heterogeneous anisotropic diffusivity. Homogeneous Dirichlet boundary condition is applied on the entire boundary. The initial concentration is taken to be zero (i.e., $c_0(\mathbf{x}) = 0$). The



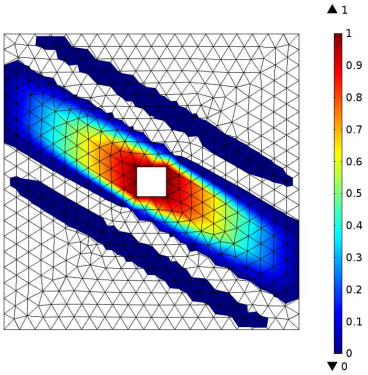
(a) $\Delta t = 0.0001$, $t = 3\Delta t$ (minimum = -0.1170)



(b) $\Delta t = 0.0001$, t = 0.05 (minimum = -0.05406)



(c) $\Delta t = 0.001$, $t = 3\Delta t$ (minimum = -0.09527)



(d) $\Delta t = 0.001$, t = 0.05 (minimum = -0.05406)

Figure 17: Anisotropic diffusion in square plate with a hole: Concentration profiles using COMSOL [47] by employing *unstructured three-node triangular mesh*. The finite element mesh is also shown. The numerical results clearly violated the non-negative constraint for the concentration. The regions that violated the non-negative constraint are indicated in white color.

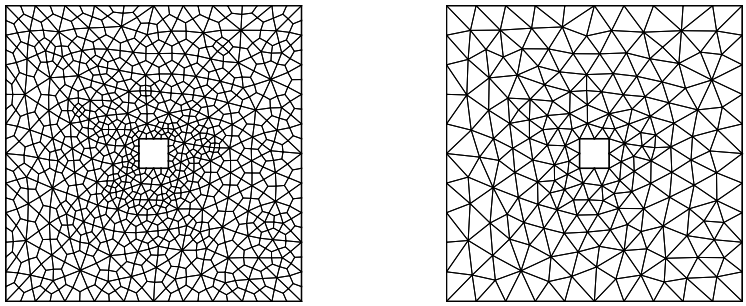
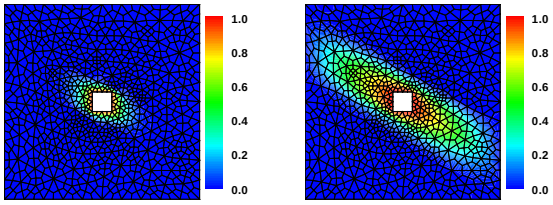
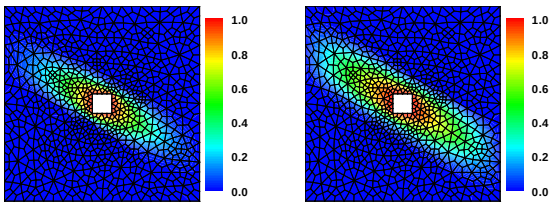


Figure 18: Anisotropic diffusion in a square plate with a hole: This figure shows the meshes employed in the numerical simulations using the proposed numerical methodology. The left figure shows an unstructured mesh based on four-node quadrilateral elements, and the right figure shows an unstructured mesh based on three-node triangular elements. The meshes are generated using GMSH [1].

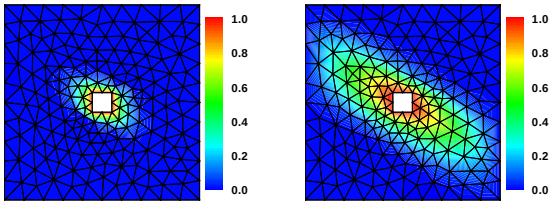


(a) $\Delta t = 0.0001$, $t = 3\Delta t$ (b) $\Delta t = 0.0001$, t = 0.05

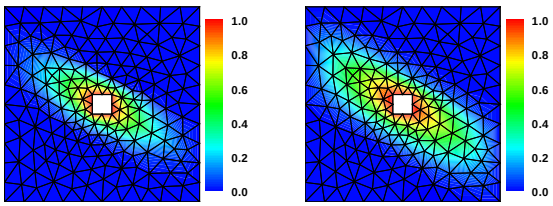


(c) $\Delta t = 0.001$, $t = 3\Delta t$ (d) $\Delta t = 0.001$, t = 0.05

Figure 19: Anisotropic diffusion in square plate with a hole: Concentration profiles using the proposed methodology by employing *unstructured four-node triangular mesh*, which is shown in Fig. 18(a). The numerical results satisfy the maximum principle and the non-negative constraint. The numerical results are visualized using Tecplot [2].

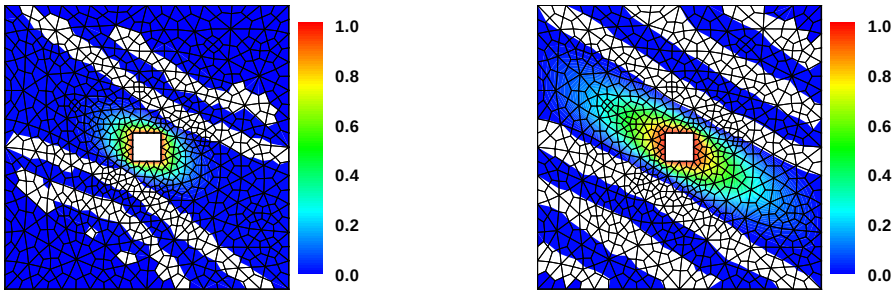


(a) $\Delta t = 0.0001$, $t = 3\Delta t$ (b) $\Delta t = 0.0001$, t = 0.05



(c) $\Delta t = 0.001$, $t = 3\Delta t$ (d) $\Delta t = 0.001$, t = 0.05

Figure 20: Anisotropic diffusion in square plate with a hole: Concentration profiles using the proposed methodology by employing *unstructured three-node triangular mesh*, which is shown in Fig. 18(b). The numerical results satisfy the maximum principle and the non-negative constraint. The numerical results are visualized using Tecplot [2].



(a) $\Delta t = 0.0001$, $t = 3\Delta t$ (minimum = -0.01024) (b) $\Delta t = 0.001$, $t = 3\Delta t$ (minimum = -0.03603)

Figure 21: Anisotropic diffusion in square plate with a hole: This figure shows the concentration profiles obtained using the backward Euler time stepping scheme ($\alpha_f = \alpha_m = \gamma = 1$) and lumped capacity matrix approach. The unstructured four-node quadrilateral mesh shown in Fig. 18(a) is used in the numerical simulation. Clearly, the numerical results do not satisfy the maximum principle and the non-negative constraint. In the case of isotropic diffusion, employing the backward Euler time-stepping scheme with lumped capacity matrix approach can be employed to satisfy maximum principles and the non-negative constraint (with some restrictions on the mesh). As it is evident from this figure, meeting these conditions is not sufficient in the case of transient anisotropic diffusion. The regions of the violation of the non-negative constraint are shown in white color. The numerical results are visualized using Tecplot [2].

volumetric source is taken as follows:

$$f(\mathbf{x},t) = \begin{cases} 1 & \text{if } (\mathrm{x},\mathrm{y}) \in [3/8,5/8]^2, \\ 0 & \text{otherwise.} \end{cases} \tag{4.8}$$

The diffusivity tensor is taken as follows:

$$\mathbf{D}(\mathbf{x}) = \begin{pmatrix} \mathrm{y}^2+\epsilon\mathrm{x}^2 & -(1-\epsilon)\mathrm{xy} \\ -(1-\epsilon)\mathrm{xy} & \epsilon\mathrm{y}^2+\mathrm{x}^2 \end{pmatrix}, \tag{4.9}$$

with $\epsilon = 0.001$. Note that the diffusivity tensor is positive definite in the open set $\Omega := (0,1)\times(0,1)$. This diffusivity tensor is widely used to test the robustness of numerical formulations in the context of maximum principles (e.g., Le Potier [61]).

Four-node quadrilateral finite elements are employed in the numerical simulation. The numerical results are generated for two different meshes (XSeed=YSeed= 51 and 101). Note that XSeed and YSeed denote the number of nodes along the x-direction and y-direction, respectively. Fig. 22 shows a typical computational mesh used in the numerical simulations with XSeed = YSeed = 51. Various time-steps ($\Delta t = 0.05, 0.1, 0.5$ and 1) are employed in the numerical simulations. The rationale behind the choice of the time-steps is that the transient solution is very close to the steady-state response for times greater than 2. Hence, any time-step bigger than the ones used in the numerical simulations does not capture the transient features of the problem, and will not be appropriate for a transient analysis. Any smaller time-step will result in bigger violation of the non-negative constraint, which will be evident from the numerical results presented in this paper.

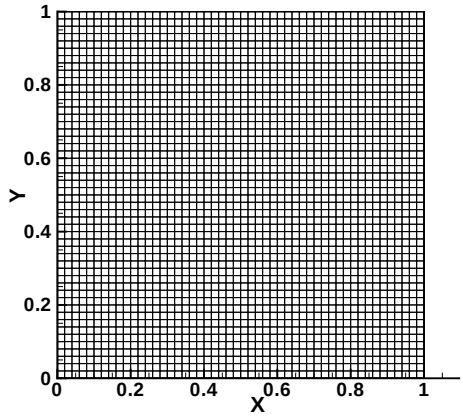


Figure 22: Diffusion in heterogeneous anisotropic medium: This figure shows a typical computational mesh used in the numerical simulations. The mesh in the figure is made of four node quadrilateral finite elements with XSeed=YSeed=51. Note that XSeed and YSeed, respectively, denote the number of nodes along the x-direction and y-direction. A similar mesh with XSeed=YSeed=101 is also used in the numerical simulations.

For the aforementioned parameters, the variation of the minimum concentration with time under the single-field formulation is shown in Fig. 23. The proposed methodology produced non-negative values for the concentration under all the considered cases, and the minimum concentration is zero. Fig. 24 compares the contours of the concentration obtained using the single-field formulation and the proposed methodology for XSeed=51 and $\Delta t = 0.5$. Even for a problem involving transient diffusion in a heterogeneous anisotropic medium, the proposed methodology did not violate the non-negative constraint. Fig. 25 compares the elapsed computational time of the proposed method-

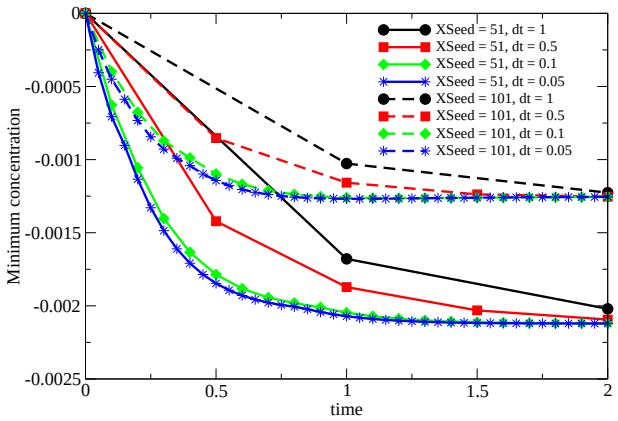


Figure 23: Diffusion in heterogeneous anisotropic medium: This figure shows the variation of the minimum concentration under the single-field formulation. Note that XSeed and YSeed denote the number of nodes along x-direction and y-direction, respectively. The proposed methodology produced non-negative solutions under all the cases considered.

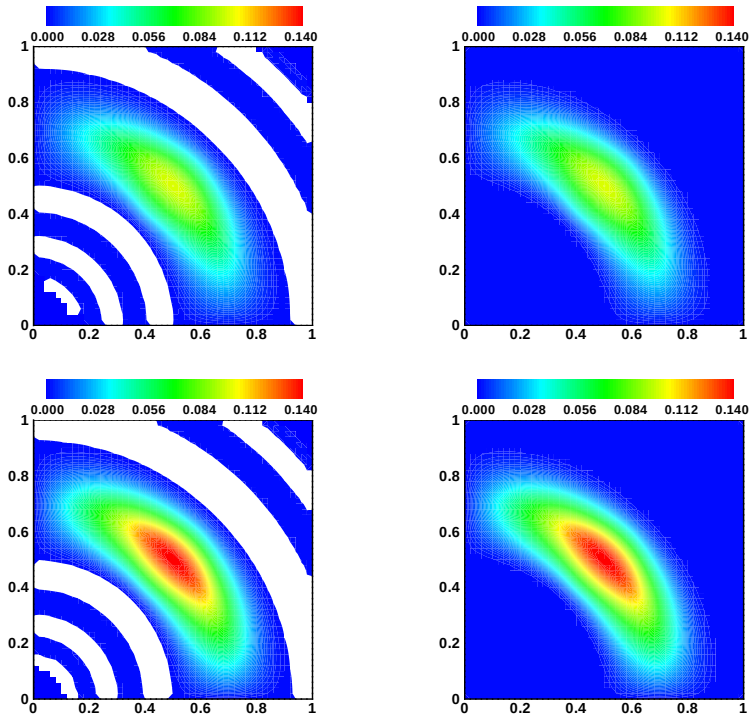


Figure 24: Diffusion in heterogeneous anisotropic medium: This figure shows the contours of the concentration under the Galerkin single-field formulation (left) and the proposed methodology (right) at time$=0.5$ (top), and time$=2$ (bottom). The time step is taken as $\Delta t = 0.5$, and $\text{XSeed} = \text{YSeed} = 51$. The regions that violated the non-negative constraint are indicated in white color.

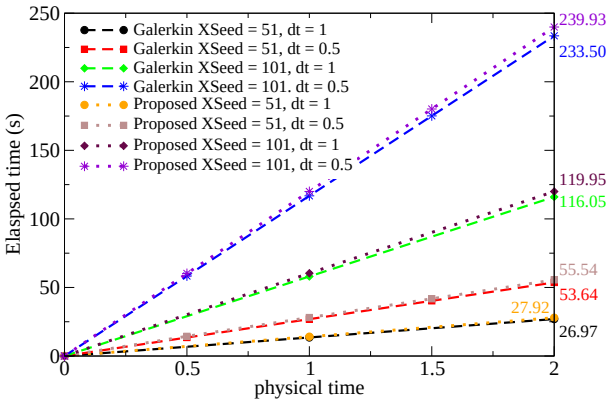


Figure 25: Diffusion in heterogeneous anisotropic medium: This figure shows the elapsed time (i.e., wall clock time) taken by the Galerkin single-field formulation and the proposed methodology for various time steps and for various meshes. The x-axis indicates the physical time, and the y-axis shows the wall clock time required to reach various time levels. The numerical simulations are carried using MATLAB R2012a [4] on Ubuntu Linux 12.04 LTS Operating System. The elapsed times are obtained using tic-toc, which is a MATLAB's built-in feature.

ology with that of the Galerkin single-field formulation. The elapsed time measured in seconds is obtained using tic-toc feature available in MATLAB [4]. From this figure, one can conclude that the additional cost incurred by the proposed methodology in meeting maximum principles and the non-negative is negligible.

5 Concluding remarks

We have presented a novel methodology for transient anisotropic diffusion equations that satisfies maximum principles and the non-negative constraint on computational grids with no additional restrictions on the time-step. The methodology has been developed using the method of horizontal lines, and techniques from convex programming. We have shown that the semi-discrete procedure based on the standard single-field formulation gives unphysical negative concentrations and violates maximum principles. Using several representative numerical examples we have shown that the proposed methodology satisfies maximum principles and the non-negative constraint on general computational grids with anisotropic and heterogeneous diffusion. The proposed methodology performs gives physically meaningful non-negative concentrations even on coarse computational grids and for small time-steps. We shall conclude the paper by discussing two possible future research endeavors in the area of discrete maximum principles. We also briefly outline potential challenges one may have to overcome in addressing these research problems.

(i) A possible future work is to incorporate advection in addition to diffusion, and devise a non-negative methodology for both steady-state and transient advection-diffusion equation. However, one cannot directly implement the procedure presented in this paper and in references [49, 50] for advection-diffusion equation, as the advection term makes the spatial differential operator non-self-adjoint.

(ii) Another interesting research problem is to devise a non-negative methodology for both steady and transient *nonlinear* diffusion-type equations.

Appendix: Plausible approaches

We now discuss other possible ways of implementing the methods of horizontal and vertical lines for transient diffusion-type equations. We will also provide reasons why these approaches may not satisfy maximum principles and the non-negative constraint. This discussion will shed light on the rationale behind the proposed methodology, and can guide future efforts in developing robust solvers for other important parabolic partial differential equations (e.g., transient diffusive-reactive systems). All the approaches presented in this appendix employ trapezoidal family of time integrators, which can be

written as follows:

$$\boldsymbol{c}^{(n+1)} = \boldsymbol{c}^{(n)} + \Delta t \left((1-\gamma)\boldsymbol{v}^{(n)} + \gamma \boldsymbol{v}^{(n+1)} \right), \tag{A.1}$$

where $\gamma \in [0,1]$. (Recall that the parameter γ used in Section 3 is different from the parameter in trapezoidal family of time integrators.) The discussion and conclusions in this appendix will hinge on the following result from Matrix Algebra. Given any vector $\boldsymbol{b} \succeq \boldsymbol{0}$, the solution of a system of linear equations of the form

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{A.2}$$

will be non-negative (i.e., $\boldsymbol{x} \succeq \boldsymbol{0}$) *if and only* if the matrix $\boldsymbol{A}$ is a monotone. (Recall that $\succeq$ denotes component-wise inequality.) A matrix is called a monotone if the matrix is invertible and all the entries of its inverse are non-negative. For further details on monotone matrices refer to the classic texts [7, 23, 66].

A.1 Method of vertical lines at integral time-steps

In this paper, this method is referred to as the *standard single-field formulation*. This is the most commonly used method for solving transient diffusion equation, and can be found in many introductory texts on finite element methods (e.g., [64,69]). The method is based on standard semi-discrete methodology and Galerkin formalism. The corresponding weak form reads: Find $c(\mathbf{x},t) \in \mathcal{P}_t$ such that we have

$$\begin{aligned} &\int_\Omega w(\mathbf{x}) \frac{\partial c(\mathbf{x},t)}{\partial t} \,\mathrm{d}\Omega + \int_\Omega \mathrm{grad}[w(\mathbf{x})] \cdot \mathbf{D}(\mathbf{x}) \mathrm{grad}[c(\mathbf{x},t)] \,\mathrm{d}\Omega \\ &= \int_\Omega w(\mathbf{x})\, f(\mathbf{x},t) \,\mathrm{d}\Omega + \int_{\Gamma^{\mathrm{N}}} w(\mathbf{x})\, q_p(\mathbf{x},t) \,\mathrm{d}\Gamma \quad \forall w(\mathbf{x}) \in \mathcal{Q}, \end{aligned} \tag{A.3}$$

where

$$\mathcal{P}_t := \{ c(\mathbf{x},t) \in H^1(\Omega) \,|\, c(\mathbf{x},t) = c_p(\mathbf{x},t) \text{ on } \Gamma^{\mathrm{D}} \}, \tag{A.4}$$

and the function space $\mathcal{Q}$ is defined previously in Eq. (3.12b). After spatial discretization using the finite element method, one obtains a system of ordinary differential equations of following form:

$$\boldsymbol{C} \frac{d\boldsymbol{c}(t)}{dt} + \boldsymbol{K}\boldsymbol{c}(t) = \boldsymbol{f}(t). \tag{A.5}$$

The capacity matrix $\boldsymbol{C}$ is symmetric and positive definite, and all the entries of the matrix are non-negative. The matrix $\boldsymbol{K}$ is symmetric and positive semi-definite. More importantly, the matrix $\boldsymbol{K}$ will not be a monotone if the medium (i.e., the diffusion process) is not isotropic. (If the medium is isotropic, it is easy to check that the matrix $\boldsymbol{K}$ is diagonally dominant, and hence it will be a monotone matrix.) If a time-stepping scheme from the

trapezoidal family is employed to solve the above ordinary differential equations, one can obtain a system of linear equations of the following form:

$$\left(\frac{1}{\gamma\Delta t}\boldsymbol{C}+\boldsymbol{K}\right)\boldsymbol{c}^{(n+1)}=\boldsymbol{f}^{(n+1)}+\frac{1}{\gamma\Delta t}\boldsymbol{C}\left(\boldsymbol{c}^{(n)}+\Delta t(1-\gamma)\boldsymbol{v}^{(n)}\right). \tag{A.6}$$

There are two potential scenarios that can contribute to the violation of the non-negative constraint and maximum principle under the method of vertical lines at integral time-steps. Firstly, the vector on the right side of Eq. (A.6) need not be non-negative, as there is no physical constraint requiring that $\boldsymbol{v}^{(n)}$ should be non-negative. Even if the volumetric source is non-negative (i.e., $\boldsymbol{f}^{(n+1)}\succeq\boldsymbol{0}$), $\boldsymbol{c}^{(n)}\succeq\boldsymbol{0}$, $\gamma\geq 0$, $\Delta t>0$, and all the entries of the capacity matrix are non-negative; the resulting vector on the right side of the above equation need not be non-negative. One possible exception is when $\gamma=1$ (that is, when the backward Euler is employed). Secondly, the matrix on the left side of Eq. (A.6) may not be a monotone. Even for an isotropic medium, the matrix will be monotone *only* if the time-step is greater than a critical time-step or by employing lumped capacity matrix. Based on the above discussion, the sufficient conditions for the method of vertical lines at integral time levels to satisfy maximum principles and the non-negative constraint are as follows:

- isotropic diffusion,
- low-order finite elements,
- backward Euler scheme (i.e., $\gamma=1$),
- lumped capacity matrix,
- select a time-step *greater* than the critical time-step, and
- place constraints on the mesh and element shapes (e.g., well-centered triangular elements, rectangular elements with aspect ratio between $1/\sqrt{2}$ and $\sqrt{2}$).

It is important to note that the above conditions are too restrictive to be able to obtain physically meaningful results for practical problems. But this method is commonly employed in many numerical simulations, and in many commercial finite element packages. Few other remarks about this method are in order.

Remark A.1. For a discussion on necessary constraints on a finite element mesh to satisfy maximum principles and the non-negative constraint, see Refs. [15, 17, 30, 49, 50]. However, all these constraints are for isotropic diffusion. It is noteworthy that, in the case of anisotropy, a computational mesh may not even exist that will ensure the satisfaction of maximum principles and the non-negative constraint.

Remark A.2. Several studies derived critical time-steps with respect to maximum principles. For example, see Refs. [35, 67]. But these derivations for critical time-steps are restricted to one-dimensional problems, isotropic diffusion, and backward Euler.

Remark A.3. It is noteworthy that there is no obvious way of modifying the non-negative formulations that has been shown recently to be successful for steady-state diffusion equations (e.g., see Refs. [49,50]) to obtain a non-negative formulation for transient diffusion equation under the method of vertical lines at integral time-steps. This is the reason why this method has not been considered as the basis in Section 3.

A.2 Method of horizontal lines at integral time-steps

By applying the method of horizontal lines at integral time levels and eliminating $v^{(n+1)}(\mathbf{x})$ using the time discretization of trapezoidal family given by Eq. (A.1), the time discretized equations take the following form:

$$\begin{aligned}&\frac{1}{\gamma\Delta t}c^{(n+1)}(\mathbf{x})-\operatorname{div}[\mathbf{D}(\mathbf{x})\operatorname{grad}[c^{(n+1)}]]\\&\qquad=f^{(n+1)}(\mathbf{x})+\frac{1}{\gamma\Delta t}\left(c^{(n)}(\mathbf{x})+(1-\gamma)\Delta t v^{(n)}(\mathbf{x})\right)\quad\text{in }\Omega,\end{aligned}\tag{A.7a}$$

$$c^{(n+1)}(\mathbf{x})=c_p^{(n+1)}(\mathbf{x})\quad\text{on }\Gamma^{\mathrm{D}},\tag{A.7b}$$

$$\widehat{\mathbf{n}}(\mathbf{x})\cdot\mathbf{D}(\mathbf{x})\operatorname{grad}[c^{(n+1)}(\mathbf{x})]=q_p^{(n+1)}(\mathbf{x})\quad\text{on }\Gamma^{\mathrm{N}}.\tag{A.7c}$$

In going from Eqs. (2.1a)-(2.1d) to Eqs. (A.7a)-(A.7c), the temporal discretization may not preserve the non-negative constraint, which should be interpreted in the following sense. One may not get a non-negative solution under Eqs. (A.7a)-(A.7c) even when the solution to the original time continuous problem given by Eqs. (2.1a)-(2.1d) is non-negative. This is again due to the fact that the right side of Eq. (A.7a) can be negative, as there is no physical constraint requiring that the rate of concentration $v^{(n)}(\mathbf{x})$ should be non-negative. However, it does not mean that the time discrete equation does not satisfy maximum principles and the non-negative equation. The above equation is diffusion with decay, and as mentioned earlier, this equation also satisfies maximum principles and the non-negative constraint. But, the requirement for the non-negative constraint is that $f^{(n+1)}(\mathbf{x})+\frac{1}{\gamma\Delta t}\left(c^{(n)}(\mathbf{x})+(1-\gamma)\Delta t v^{(n)}(\mathbf{x})\right)\geq 0$.

A.3 Method of horizontal lines at weighted time levels

We shall perform temporal discretization at the weighted time level $t_{n+\gamma}$, which gives rise to the following equations:

$$\frac{1}{\gamma\Delta t}c^{(n+\gamma)}(\mathbf{x})-\operatorname{div}[\mathbf{D}(\mathbf{x})\operatorname{grad}[c^{(n+\gamma)}]]=f^{(n+\gamma)}(\mathbf{x})+\frac{1}{\gamma\Delta t}c^{(n)}(\mathbf{x})\quad\text{in }\Omega,\tag{A.8a}$$

$$c^{(n+\gamma)}(\mathbf{x})=c_p(\mathbf{x},t_{n+\gamma})\quad\text{on }\Gamma^{\mathrm{D}},\tag{A.8b}$$

$$\widehat{\mathbf{n}}(\mathbf{x})\cdot\mathbf{D}(\mathbf{x})\operatorname{grad}[c^{(n+\gamma)}]=q_p^{(n+\gamma)}(\mathbf{x})\quad\text{on }\Gamma^{\mathrm{N}}.\tag{A.8c}$$

One can obtain nodal concentrations at weighted time levels (i.e., $\boldsymbol{c}^{(n+\gamma)}$) by employing the optimization-based solver presented in Section 3. Noting the results presented in Ref. [51] on stability issues associated with numerical time integration of differential-algebraic equations, the concentration at integral time levels is approximated in terms of corresponding quantities at weighted time levels. The interpolation scheme is pictorially described in Fig. 3, and can be mathematically written as follows:

$$\boldsymbol{c}^{(n+1)} = \gamma \boldsymbol{c}^{(n+\gamma)} + (1-\gamma)\boldsymbol{c}^{(n+1+\gamma)}. \tag{A.9}$$

The rate of concentration at weighted time levels can be calculated as follows:

$$\boldsymbol{v}^{(n+\gamma)} = \frac{\boldsymbol{c}^{(n+1)} - \boldsymbol{c}^{(n)}}{\Delta t}. \tag{A.10}$$

The corresponding quantity at integral time levels are calculated as follows:

$$\boldsymbol{v}^{(n+1)} = \gamma \boldsymbol{v}^{(n+\gamma)} + (1-\gamma)\boldsymbol{v}^{(n+1+\gamma)}. \tag{A.11}$$

The interpolation given by Eq. (A.9) is different from the usual way of interpolating the quantities at weighted time levels in terms of integral time levels. That is,

$$\boldsymbol{c}^{(n+\gamma)} = (1-\gamma)\boldsymbol{c}^{(n)} + \gamma \boldsymbol{c}^{(n+1)}. \tag{A.12}$$

Fig. 3 compares both these interpolation schemes. The only drawback of the method presented in this subsection is that it is not self-starting, as we do not have $\boldsymbol{c}^{(n-1+\gamma)}$ when $n=1$ unless $\gamma=1$. But this drawback can be easily overcome by employing the backward Euler scheme (i.e., $\gamma=1$) for the first time level, and then employ the method for subsequent time levels. Therefore, the method presented in this subsection can be considered as an alternate to the method presented in Section 3 to satisfy maximum principles and the non-negative constraint for transient diffusion-type equations.

Acknowledgments

K.B.N. and M.S. acknowledge the support from the National Science Foundation under Grant No. CMMI 1068181. K.B.N. also acknowledges the supports from the DOE Office of Nuclear Energy's Nuclear Energy University Programs (NEUP). The opinions expressed in this paper are those of the authors and do not necessarily reflect that of the sponsors.

References

[1] *Gmsh: A three-dimensional finite element mesh generator with pre- and post-processing facilities*. URL: http://www.geuz.org/gmsh/.

[2] *Tecplot 360: User's Manual*. URL: http://www.tecplot.com, Bellevue, Washington, USA, 2008.

[3] *General Algebraic Modeling System (GAMS)*. Version 23.8, GAMS Development Corporation, Washington DC, USA, 2012.
[4] *MATLAB 2012a*. The MathWorks, Inc., Natick, Massachusetts, USA, 2012.
[5] B. M. Adams, W. J. Bohnhoff, K. R. Dalbey, J. P. Eddy, M. S. Eldred, D. M. Gay, K. Haskell, P. D. Hough, and L. P. Swiler. *DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 5.2 User's Manual*. Sandia Technical Report SAND2010-2183, 2011.
[6] U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM, Philadelphia, 1998.
[7] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1987.
[8] M. Berzins. Modified mass matrices and positivity preservation for hyperbolic and parabolic PDEs. *Communications in Numerical Methods in Engineering*, 17:659–666, 2001.
[9] F. A. Bornemann. An adaptive multilevel approach to parabolic equations I. General theory and 1D implementation. *Impact of Computing in Science and Engineering*, 2:279–317, 1990.
[10] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.
[11] H. S. Carslaw and J. C. Jaeger. *Conduction of Heat in Solids*. Oxford University Press, New York, second edition, 1986.
[12] C. Cattaneo. Sur une forme de l'équation de la chaleur éliminant le paradoxe d'une propagation instantanée. *Comptes Rendus*, 247:431–433, 1958.
[13] R. Chapko and R. Kress. Rothe's method for the heat equation and boundary integral equations. *Journal of Integral Equations and Applications*, 09:47–69, 1997.
[14] C. M. Chen and V. Thomee. The lumped mass finite element method for a parabolic problem. *Journal of the Australian Mathematical Society*, 26:329–354, 1985.
[15] I. Christie and C. Hall. The maximum principle for bilinear elements. *International Journal for Numerical Methods in Engineering*, 20:549–553, 1984.
[16] J. Chung and G. M. Hulbert. A time integration algorithm for structural dynamics with improved numerical dissipation: The generalized-α method. *Journal of Applied Mechanics*, 60:371–375, 1993.
[17] P. G. Ciarlet and P-A. Raviart. Maximum principle and uniform convergence for the finite element method. *Computer Methods in Applied Methods and Engineering*, 2:17–31, 1973.
[18] J. Crank. *The Mathematics of Diffusion*. Oxford University Press, New York, second edition, 1980.
[19] J. Douglas and T. Dupont. Galerkin methods for parabolic equations. *SIAM Journal on Numerical Analysis*, 07:575–626, 1970.
[20] M. A. T. Elshebli. Discrete maximum principle for the finite element solution of linear nonstationary diffusionreaction problems. *Applied Mathematical Modeling*, 32:1530–1541, 1998.
[21] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, Rhode Island, 1998.
[22] I. Farago, R. Horvath, and S. Korotov. Discrete maximum principle for linear parabolic problems solved on hybrid meshes. *Applied Numerical Mathematics*, 53:249–264, 2005.
[23] M. Fiedler. *Special Matrices and Their Applications in Numerical Mathematics*. Martinus Nijhoff Publishers, Dordrecht, The Netherlands, 1986.
[24] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer, New York, 2001.

[25] M. E. Gurtin and A. C. Pipkin. A general theory of heat conduction with finite speed. *Archive for Rational Mechanics and Analysis*, 31:113–126, 1968.
[26] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer-Verlag, New York, 1996.
[27] E. Hairer, C. Lubich, and M. Roche. *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*. Lecture Notes in Mathematics. Springer-Verlag, New York, 1989.
[28] I. Harari. Stability of semidiscrete formulations for parabolic problems at small time steps. *Computer Methods in Applied Mechanics and Engineering*, 193:1491–1516, 2004.
[29] P. Herrera and A. Valocchi. Positive solution of two-dimensional solute transport in heterogeneous aquifers. *Ground Water*, 44:803–813, 2006.
[30] R. Horvath. Sufficient conditions of the discrete maximum-minimum principle for parabolic problems on rectangular meshes. *International Journal of Computers and Mathematics with Applications*, 55:2306–2317, 2008.
[31] W. Huang. Sign-preserving of principal eigenfunctions in P1 finite element approximation of eigenvalue problems of second-order elliptic operators. *Journal of Computational Physics*, 274:230–244, 2014.
[32] W. Huang, L. Kamenski, and J. Lang. Stability of explicit Runge–Kutta methods for finite element approximation of linear parabolic equations on anisotropic meshes. *WIAS Preprint No. 1869*, 2013.
[33] T. J. R. Hughes. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Prentice-Hall, Englewood Cliffs, New Jersey, 1987.
[34] J. Ignaczak and M. O. Starzewski. *Thermoelasticity with Finite Wave Speeds*. Oxford Science Publications, New York, 2009.
[35] F. Ilinca and J. F. Hetu. Galerkin gradient least-squares formulations for transient conduction heat transfer. *Computer Methods in Applied Mechanics and Engineering*, 191:3073–3097, 2002.
[36] K. E. Jansen, C. H. Whiting, and G. H. Hulbert. A generalized-α method for integrating the filtered Navier-Stokes equations with a stabilized finite element method. *Computer Methods in Applied Mechanics and Engineering*, 190:305–319, 2000.
[37] J. Lang and A. Walter. An adaptive Rothe method for nonlinear reaction-diffusion systems. *Applied Numerical Mathematics*, 13:135–146, 1993.
[38] E. E. Levi. Sull' equazione del calore. *Annali di Matematica Pura ed Applicata*, 14:187–264, 1908.
[39] X. Li and W. Huang. Maximum principle for the finite element solution of time-dependent anisotropic diffusion problems. *Numerical Methods for Partial Differential Equations*, 29:1963–1985, 2013.
[40] K. Lipnikov, M. Shashkov, D. Svyatskiy, and Y. Vassilevski. Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes. *Journal of Computational Physics*, 227:492–512, 2007.
[41] K. Lipnikov, D. Svyatskiy, and Y. Vassilevski. Interpolation-free monotone finite volume method for diffusion equations on polygonal meshes. *Journal of Computational Physics*, 228: 703–716, 2009.
[42] K. Lipnikov, D. Svyatskiy, and Y. Vassilevski. A monotone finite volume method for advectiondiffusion equations on unstructured polygonal meshes. *Journal of Computational Physics*, 229:4017–4032, 2010.
[43] K. Lipnikov, G. Manzini, and D. Svyatskiy. Analysis of the monotonicity conditions in the mimetic finite difference method for elliptic problems. *Journal of Computational Physics*, 230: 2620–2642, 2011.

[44] J. C. Maxwell. On the dynamical theory of gases. *Philosophical Transactions of Royal Society of London*, A157:26–78, 1866.
[45] A. Mizukami. Variable explicit finite element methods for unsteady heat conduction equations. *Computer Methods in Applied Mechanics and Engineering*, 59:101–109, 1986.
[46] M. K. Mudunuru and K. B. Nakshatrala. A framework for coupled deformation-diffusion analysis with application to degradation/healing. *International Journal for Numerical Methods in Engineering*, 89:1144–1170, 2012.
[47] COMSOL Multiphysics. Version 4.3 a. *COMSOL Inc, Burlington, MA*, 2012.
[48] T. Munson, J. Sarich, S. Wild, S. Benson, and L. C. McInnes. TAO 2.0 Users Manual. Technical Report ANL/MCS-TM-322, Mathematics and Computer Science Division, Argonne National Laboratory, 2012. http://www.mcs.anl.gov/tao.
[49] H. Nagarajan and K. B. Nakshatrala. Enforcing the non-negativity constraint and maximum principles for diffusion with decay on general computational grids. *International Journal for Numerical Methods in Fluids*, 67:820–847, 2011.
[50] K. B. Nakshatrala and A. J. Valocchi. Non-negative mixed finite element formulations for a tensorial diffusion equation. *Journal of Computational Physics*, 228:6726–6752, 2009.
[51] K. B. Nakshatrala, A. Prakash, and K. D. Hjelmstad. On dual Schur domain decomposition method for linear first-order transient problems. *Journal of Computational Physics*, 228:7957–7985, 2009.
[52] K. B. Nakshatrala, M. K. Mudunuru, and A. J. Valocchi. A numerical framework for diffusion-controlled bimolecular-reactive systems to enforce maximum principles and non-negative constraint. *Journal of Computational Physics*, 253:278–307, 2013.
[53] L. Nirenberg. A strong maximum principle for parabolic equations. *Communications on Pure and Applied Mathematics*, 6:167–177, 1953.
[54] M. N. Ozisik. *Heat Conduction*. John Wiley & Sons, Inc., New York, second edition, 1993.
[55] J.-S. Pang. Methods for quadratic programming: A survey. *Computers and Chemical Engineering*, 5:583–594, 1983.
[56] C. V. Pao. *Nonlinear Parabolic and Elliptic Equations*. Springer-Verlag, New York, 1993.
[57] G. S. Payette, K. B. Nakshatrala, and J. N. Reddy. On the performance of high-order finite elements with respect to maximum principles and the non-negative constraint for diffusion-type equations. *International Journal for Numerical Methods in Engineering*, 91:742–771, 2012.
[58] L. Petzold. Differential/algebraic equations are not ODEs. *SIAM Journal on Scientific and Statistical Computing*, 3:367–384, 1982.
[59] M. Picone. Maggiorazione degli integrali delle equazioni totalmente paraboliche alle derivate parziali del secondo ordine. *Annali di Matematica Pura ed Applicata*, 7:145–192, 1929.
[60] G. Porru and S. Serra. Maximum principles for parabolic equations. *Journal of the Australian Mathematical Society*, 56:41–52, 1994.
[61] C. Le Potier. Finite volume monotone scheme for highly anisotropic diffusion operators on unstructured triangular meshes. *Comptes Rendus Mathematique*, 341:787–792, 2005.
[62] M. H. Protter and H. F. Weinberger. *Maximum Principles in Differential Equations*. Springer-Verlag, New York, 1999.
[63] E. Rank, C. Katz, and H. Werner. On the importance of the discrete maximum principle in transient analysis using finite element methods. *International Journal for Numerical Methods in Engineering*, 19:1771–1782, 1983.
[64] J. N. Reddy. *An Introduction to the Finite Element Method*. McGraw-Hill, New York, third edition, 2005.
[65] E. Rothe. Zweidimensionale parabolische randwertaufgaben als grenzfall eindimensionaler

randwertaufgaben. *Mathematische Annalen*, 102:650–670, 1930.
[66] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, 2003.
[67] H. R. Thomas and Z. Zhou. An analysis of factors that govern the minimum time step size to be used in the finite element analysis of diffusion problems. *Communications in Numerical Methods in Engineering*, 14:809–819, 1998.
[68] Y. Ye and E. Tse. An extension of Karmarkar's projective algorithm for convex quadratic programming. *Mathematical Programming*, 44:157–179, 1989.
[69] O. C. Zienkiewicz and R. L. Taylor. *The Finite Element Method : Vol.1*. McGraw-Hill, New York, 1989.