

© Copyright by Maruti Kumar Mudunuru, August 2015

All Rights Reserved

ON ENFORCING MAXIMUM PRINCIPLES AND
ELEMENT-WISE SPECIES BALANCE FOR
ADVECTIVE-DIFFUSIVE-REACTIVE SYSTEMS

A Dissertation

Presented to

the Faculty of the Department of Civil and Environmental Engineering

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in Civil Engineering

by

Maruti Kumar Mudunuru

August 2015

ON ENFORCING MAXIMUM PRINCIPLES AND
ELEMENT-WISE SPECIES BALANCE FOR
ADVECTIVE-DIFFUSIVE-REACTIVE SYSTEMS

Maruti Kumar Mudunuru

Approved:

Chair of the Committee
Kalyana Nakshatrala, Assistant Professor
Civil & Environmental Engineering

Committee Members:

Kaspar Willam, Distinguished Professor
Civil & Environmental Engineering

Cumaraswamy Vipulanandan, Professor
Civil & Environmental Engineering

Keh-Han Wang, Professor
Civil & Environmental Engineering

Yashashree Kulkarni, Assistant Professor
Mechanical Engineering

Gino Lim, Associate Professor,
Industrial Engineering

Suresh K. Khator, Associate Dean
Cullen College of Engineering

Roberto Ballarini, Professor and Chair
Civil & Environmental Engineering

Acknowledgements

As a crucial phase of learning process comes to an end, first, I would like to acknowledge the Ph.D. ecosystem at the University of Houston. The variety in courses offered and interaction with various professors helped me in my overall development. However, experiencing such an environment would not be possible if I did not join Dr. Kalyana Babu Nakshatrala's research group at UH. Very rarely (quantitatively, close to machine precision) one finds an advisor who is available for help at all times (even on holidays) and during job search. My research philosophy has been molded based on Dr. Nakshatrala's constant reminder of a famous quote from Dr. Richard Feynman: "*People who wish to analyze nature without using mathematics must settle for a reduced understanding*". In a nutshell, his teaching philosophy, research commitment, mentoring, and guiding of students reminds me of NBA's Gregg Popovich player development program and play-making strategies. The freedom he gave me to collaborate and discuss with various people within his research group (CAML) and outside is immeasurable.

I would also like to thank the committee members of my dissertation Dr. Yashashree Kulkarni, Dr. Gino Lim, Dr. Cumaraswamy Vipulanandan, Dr. Keh-Han Wang, and Dr. Kaspar Willam for their valuable suggestions during my research proposal. I am particularly grateful to Dr. Cumaraswamy Vipulanandan and Dr. Junuthula N. Reddy for their constant interest in my professional development.

Research in our CAML group would not have been fruitful without stimulating discussions and arguments on variety of topics with Saeid Karmi, Justin Chang, Mohammad Shabouei, and Can Xu. They made my research life interesting at UH. Special thanks to Dr. Gregory Payette (Greg) and Dr. Venkat Vallala (Venkat) for numerous interesting technical discussions. The summer internship at ExxonMobil

URC had a profound impact on my professional and overall development.

I thank my friends Abhishek Velichala, Dr. Sireesh Dadi, Vivek Vandrasi, and Sandeep Ghorawat for making my stay at Houston memorable. Especially my roommate Abhishek Velichala, for introducing me to the NBA and some genre-defining Japanese graphic novels such as One Piece, Hunter x Hunter, and Slam Dunk. The influence of NBA and One Piece on my personal life and tangential thinking is tremendous.

My deepest thanks go to my parents and my younger brother, Rajesh Kumar Mudunuru. I am truly grateful and indebted to their valuable suggestions on improving my vocabulary and computer programming skills. Without their love, sacrifice and support, I would have never started this journey and realize my educational aspirations. This dissertation is dedicated to them.

To my parents and my younger brother, Rajesh Kumar Mudunuru.

ON ENFORCING MAXIMUM PRINCIPLES AND
ELEMENT-WISE SPECIES BALANCE FOR
ADVECTIVE-DIFFUSIVE-REACTIVE SYSTEMS

An Abstract

of a

Dissertation

Presented to

the Faculty of the Department of Civil and Environmental Engineering

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in Civil Engineering

by

Maruti Kumar Mudunuru

August 2015

Abstract

This dissertation aims at developing robust numerical methodologies to solve advective-diffusive-reactive systems that provide accurate and physical solutions for a wide range of input data (e.g., Péclet and Damköhler numbers) and for complicated geometries. It is well-known that physical quantities like concentration of chemical species and the absolute temperature naturally attain non-negative values. Moreover, the governing equations of an advective-diffusive-reactive system are either elliptic (in the case of steady-state response) or parabolic (in the case of transient response) partial differential equations, which possess important mathematical properties like comparison principles, maximum-minimum principles, non-negativity, and monotonicity of solution. It is desirable and in many situations necessary for a predictive numerical solver to meet important physical constraints. For example, a negative value for the concentration in a numerical simulation of reactive-transport will result in an algorithmic failure.

The objective of this dissertation is two fold. *First*, we show that many existing popular numerical formulations, open source scientific software packages, and commercial packages do *not* inherit or mimic fundamental properties of continuous advective-diffusive-reactive systems. For instance, the popular standard single-field Galerkin formulation produce *negative* values and spurious node-to-node oscillations for the primary variables in advection-dominated and reaction-dominated diffusion-type equations. Furthermore, the violation is not mere numerical noise and cannot be neglected. *Second*, we shall provide various numerical methodologies to overcome such difficulties. We critically evaluate their performance and computational cost for a wide range of Péclet and Damköhler numbers.

We first derive necessary and sufficient conditions on the finite element matrices to satisfy discrete comparison principle, discrete maximum principle, and non-negative constraint. Based on these conditions, we obtain restrictions on the computational mesh and generate physics-compatible meshes that satisfy discrete properties using open source mesh generators. We then show that imposing restrictions on computational grids may not always be a viable approach to achieve physically meaningful non-negative solutions for complex geometries and highly anisotropic media. We therefore develop a novel structure-preserving numerical methodology for advective-diffusive-reactive systems that satisfies local and global species balance, comparison principles, maximum principles, and the non-negative constraint on coarse general computational grids. This methodology can handle complex geometries and highly anisotropic media. The proposed framework can be an ideal candidate for predictive simulations in groundwater modeling, reactive transport, environmental fluid mechanics, and modeling of degradation of materials. The framework can also be utilized to numerically obtain scaling laws for complicated problems with non-trivial initial and boundary conditions.

Table of Contents

Acknowledgments	v
Abstract	ix
Table of Contents	xi
List of Figures	xvi
List of Tables	xxvii
Notation	xxix
1 INTRODUCTION	1
2 A PRIMER ON PARTIAL DIFFERENTIAL EQUATIONS	4
2.1 Definition and classification of PDEs	4
2.2 Notation about $C^m(\Omega)$, $C^m(\bar{\Omega})$ and Sobolev spaces	6
3 ON LOCAL AND GLOBAL SPECIES CONSERVATION ER- RORS FOR NONLINEAR ECOLOGICAL MODELS AND CHEM- ICALLY REACTING FLOWS	9
3.1 Introduction to transport-controlled oscillatory chemical reactions	10
3.1.1 Governing equations for mixing in oscillatory media	11
3.1.2 Problem statement and main contributions	12
3.1.3 Outline of this chapter	15
3.2 Reduced-order models: Qualitative nature of the solution	15
3.2.1 Hopf bifurcations and chemical reaction system stability	17
3.2.2 Stability analysis of CDIMA and BZ reduced-order models	19
3.3 Numerical formulations: Species balance errors	22
3.3.1 SUPG formulation	23
3.3.2 GLS formulation	23
3.3.3 Local and global species balance errors	24
3.4 Chemically reacting lid-driven cavity flows	25

3.5	Summary and conclusions	34
4	ON MESH RESTRICTIONS TO SATISFY COMPARISON PRINCIPLES, MAXIMUM PRINCIPLES, AND THE NON-NEGATIVE CONSTRAINT: RECENT DEVELOPMENTS AND NEW RESULTS	35
4.1	Introduction and motivation	36
4.1.1	Strategy I: Mesh restrictions	40
4.1.2	Strategy II: Non-negativity, monotone, and monotonicity preserving formulations	42
4.1.3	Strategy III: Post-processing methods	43
4.1.4	Main contributions and an outline of this chapter	45
4.2	Linear second-order elliptic equation and associated mathematical principles	47
4.2.1	Single-field Galerkin formulation	52
4.2.2	Discrete single-field Galerkin formulation	53
4.3	An in-depth look at continuous and discrete principles	61
4.3.1	Simply connected vs. multiple connected domains	62
4.3.2	Minimum principles, and non-negative and min-max constraints	63
4.3.3	High-order finite element methods	65
4.3.4	Relationship between various DCPs and DMPs	66
4.4	Mesh restrictions to satisfy discrete principles	72
4.4.1	Geometrical properties and finite element analysis of simplicial elements	75
4.4.2	Sufficient conditions for a three-node triangular element	78
4.4.3	Numerical examples based on different types of triangulations	93
4.4.4	Sufficient conditions for a rectangular element	118

4.5	Concluding remarks and open questions	121
5	ON ENFORCING MAXIMUM PRINCIPLES AND ACHIEVING ELEMENT-WISE SPECIES BALANCE FOR ADVECTION-DIFFUSION-REACTION EQUATIONS UNDER FINITE ELEMENT METHOD	125
5.1	Introduction and motivation	126
5.2	Governing equations: Advective-diffusive-reactive systems	131
5.2.1	Weak formulations	134
5.2.2	Maximum principles and the non-negative constraint	135
5.2.3	On appropriate Neumann BCs	140
5.3	Plausible approaches and their shortcomings	142
5.3.1	Approach #1: Clipping/cut-off methods	143
5.3.2	Approach #2: Mesh restrictions	143
5.3.3	Approach #3: Using non-negative methodologies for diffusion equations	144
5.3.4	Approach #4: Posing the discrete equations as a P -LCP	145
5.3.5	Approach #5: Posing the discrete problem as constrained normal equations	149
5.4	Proposed computational framework	151
5.4.1	Design synopsis of the proposed numerical methodology	153
5.4.2	Weighted primitive LSFEM	155
5.4.3	Weighted negatively stabilized streamline diffusion LSFEM	156
5.4.4	Discrete equations	159
5.5	Coercivity, error estimates, and stabilization parameters	162
5.6	Numerical h -Convergence and benchmark problems	172
5.6.1	Convergence analysis for $D(x, y) = 10^{-2}$	174

5.6.2	Thermal boundary layer problem	180
5.7	Transport-controlled bimolecular reactions	181
5.7.1	One-dimensional steady-state analysis of product formation in fast reactions	185
5.7.2	Steady-state plume formation from boundary in a reaction tank	188
5.7.3	Transient analysis of vortex stirred mixing in a reaction tank .	192
5.7.4	Transient analysis of species mixing in cellular flows	195
5.8	Summary and concluding remarks	204
6	NUMERICAL FORMULATIONS FOR STEADY-STATE AND TRANSIENT SEMI-LINEAR REACTION-DIFFUSION EQUA- TIONS, AND THEIR STRUCTURE PRESERVING PROPER- TIES	207
6.1	Introduction and motivation	208
6.1.1	Main contributions and outline of this chapter	209
6.2	Governing equations and mathematical principles for reaction-diffusion equations	211
6.2.1	Mathematical principles and relevant notation	212
6.2.2	Mesh and time-step restrictions for linear second-order elliptic and parabolic PDEs	215
6.3	Nonlinear techniques to solve reaction-diffusion equations	220
6.3.1	Pao's method	220
6.3.2	Picard's method	227
6.3.3	Traditional Newton-Raphson method	227
6.3.4	Modification of traditional Newton-Raphson method	229
6.3.5	Consistent linearization method	229
6.3.6	General comments on various nonlinear methods	231

6.4	Physics-based chemical reaction models of monotone-type	232
6.4.1	Enzyme kinetics (Michaelis-Menton type) chemical reaction model	233
6.4.2	Population growth (Fisher type) and genetics model	236
6.4.3	Nuclear reactor dynamics model	238
6.4.4	Chemical reactor dynamics model	240
6.4.5	Moderate/slow bimolecular chemical reaction model	241
6.5	Representative numerical examples	243
6.5.1	Autocatalytic reaction-diffusion problem	244
6.6	Concluding remarks	248
7	CONCLUSIONS AND FUTURE WORK	259
7.1	Future work	261
	REFERENCES	263
	Appendices	288

List of Figures

Figure 3.1	Transport-controlled chemically reacting lid-driven cavity flows: A pictorial description of the initial boundary value problem for CDIMA and BZ reaction schemes. Herein, $L_x = L_y = 1$ and non-reactive volumetric sources are zero.	27
Figure 3.2	Transport-controlled CDIMA reaction (Isotropic diffusivity): This figure shows the concentration profiles of I^- and ClO_2^- ions at various time levels.	28
Figure 3.3	Transport-controlled CDIMA reaction: This figure shows the concentration profiles of I^- and ClO_2^- ions at $t = 1$. We also see spurious node-to-node oscillations and negative values are as high as -0.46.	29
Figure 3.4	Transport-controlled BZ reaction (Isotropic diffusivity): This figure shows the concentration profiles of $HBrO_2$, Br^- , and Ce^{4+} at various times. Analysis is performed based on the parameters given by equations (3.4.2a) and (3.4.3a).	30
Figure 3.5	Transport-controlled BZ reaction (Anisotropic diffusivity): This figure shows the concentration profiles of $HBrO_2$, Br^- , and Ce^{4+} at various times. Analysis is performed based on the parameters given by (3.4.2a), (3.4.3c), and (3.4.4a).	31
Figure 3.6	Transport-controlled BZ reaction: This figure shows the concentration profiles of $HBrO_2$, Br^- , and Ce^{4+} at $t = 1$. We see spurious node-to-node oscillations in various parts of the domain and the negative values are as high as -0.35.	32

Figure 3.7	Transport-controlled CDIMA reaction: This figure shows the local species balance errors for I^- and ClO_2^- ions at $t = 1$. Analysis is performed based on the parameter set given by equations (3.4.2a), (3.4.3a), (3.4.3c), and (3.4.4a).	33
Figure 3.8	Transport-controlled BZ reaction: This figure shows the local species balance errors for $HBrO_2$, Br^- , and Ce^{4+} at $t = 1$. Analysis is performed based on the parameter set given by equations (3.4.2a), (3.4.3a), (3.4.3c), and (3.4.4a).	33
Figure 4.1	ABAQUS unstructured meshes for an L-shaped domain with multiple holes: The left and right figures show an instance of three-node triangular and four-node quadrilateral meshes employed in the numerical simulation of a pure anisotropic diffusion problem using ABAQUS.	39
Figure 4.2	ABAQUS numerical simulation for an L-shaped domain with multiple holes: The contours of concentration obtained using ABAQUS are based on three-node triangular mesh.	39
Figure 4.3	Minimum and maximum values for concentration in an L-shaped domain with multiple holes: The left and right figures show the minimum and maximum values attained in the computational domain.	40
Figure 4.4	Percentage of violation in minimum and maximum constraints for concentration in an L-shaped domain with multiple holes: The left and right figures show the percentage of nodes that have violated the constraints.	40

Figure 4.5	Venn diagram for the space of solutions based on mesh restrictions: A pictorial description of the space of numerical solutions satisfying various DMPs and DCPs based on equation (4.2.23) and Theorem 4.2.6.	67
Figure 4.6	Venn diagram for the space of solutions based on various numerical formulations: A pictorial description of the space of numerical solutions satisfying various DMPs, DCPs, and NC.	69
Figure 4.7	Geometrical properties of an arbitrary simplex in 2D: A pictorial description of simplicial mesh element properties.	74
Figure 4.8	T3 element for heterogeneous isotropic diffusivity: A pictorial description of the feasible region is shown in light blue color.	85
Figure 4.9	T3 element for anisotropic diffusivity when $\widetilde{D}_{xy} = 0$: A pictorial description of the feasible region (left figure) for the coordinates (a, b) is indicated in light blue color.	88
Figure 4.10	T3 element for anisotropic diffusivity when $\widetilde{D}_{xy} < 0$: The left figure indicates the feasible region for the coordinates (a, b) in light blue color. The right figure indicates that the T3 element can be acute/right/obtuse-angled.	89
Figure 4.11	T3 element for anisotropic diffusivity when $\widetilde{D}_{xy} > 0$: The left figure indicates the feasible region for the coordinates (a, b) in light blue color. The right figure indicates that the T3 element can be acute/right/obtuse-angled.	89
Figure 4.12	T3 element for fixed η and varying ϵ : A pictorial description of the feasible region (light blue color) for a fixed η and varying ϵ . Analysis is performed for $\eta = -1$ and $\epsilon = \{2, 10, 50, 100, 200, 500\}$. . .	90

Figure 4.13	T3 element for fixed ϵ and varying η : A pictorial description of the feasible region (light blue color) for a fixed ϵ and varying η . Analysis is performed for $\epsilon = 100$ and $\eta = \{-8, -4, -2, 2, 4, 8\}$	91
Figure 4.14	DMP-based T3 elements for heterogeneous isotropic and anisotropic diffusivity: A pictorial description of a mesh generation procedure to obtain a new triangulation using a given background mesh. . .	95
Figure 4.15	Test problem #1: The top left figure shows a coarse triangulation employed in the numerical study, <i>which is to the scale</i> . The top right figure and the bottom two figures show the concentration profiles obtained using this mesh.	104
Figure 4.16	Test problem #1: The top left figure shows the maximum angle possible in each element of the mesh. The top right figure and the bottom two figures show the <i>element maximum generalized Delaunay-type condition</i>	105
Figure 4.17	Test problem #2: The computational domain under consideration is a bi-unit square with one of its vertices at origin $O = (0, 0)$. . .	106
Figure 4.18	Test problem #2: The left figure shows the background mesh on which BAMG operates to give an anisotropic triangulation, which is shown in the right figure.	107
Figure 4.19	Test problem #2: The left figure shows the concentration profile based on the background mesh, while the right figure shows the concentration profile using the anisotropic triangulation.	107
Figure 4.20	Test problem #3: The left figure shows the background mesh and the right figure shows the anisotropic triangulation obtained using BAMG for all the four cases.	109

Figure 4.21	Test problem #3: This figure shows the concentration profiles for four different cases based on the background mesh and anisotropic meshes shown in Figure 4.20.	110
Figure 4.22	Issues with traditional mesh refinement: Concentration profiles for the fracture domain when $\mathbf{v} = (0.1, 1.0)$ and $\alpha = 1.0$. The white region in the figures shows the area in which the numerical simulation has violated the NC and maximum constraint.	114
Figure 4.23	Issues with traditional mesh refinement: The left figure shows the anisotropic mesh obtained using the traditional mesh refinement procedure on the anisotropic triangulation given in Figure 4.18. The right figure shows the concentration profile obtained using this refined mesh.	115
Figure 4.24	Issues with traditional mesh refinement: The left figure shows the anisotropic mesh obtained using the traditional mesh refinement procedure on the anisotropic triangulation given in Figure 4.20. The right figure shows the concentration profile obtained using this refined mesh.	115
Figure 4.25	Issues with non-traditional mesh refinement: The left figure shows a refined anisotropic mesh obtained using the non-traditional approach. The right figure shows the concentration profile obtained using this refined mesh (did not converge in <code>MaxIters = 100</code>).	116
Figure 4.26	Local species balance errors: The figures show the errors incurred in satisfying local species balance for various test problems on coarse meshes.	119
Figure 5.1	This figure illustrates concentration and flux boundary conditions.	133

Figure 5.2	Academic problem: This figure compares the numerical solution with the exact solution.	152
Figure 5.3	Numerical h -convergence study: A pictorial description of the two-dimensional boundary value problem.	173
Figure 5.4	Numerical h -convergence study: This figure shows the typical computational meshes used in the numerical convergence analysis.	175
Figure 5.5	Numerical h -convergence study: This figure shows the convergence rates for the concentration and flux vector in L_2 -norm and H^1 -semi-norm with and without LSB constraints.	176
Figure 5.6	Numerical h -convergence study: The top and bottom left figures show the contours of error incurred in satisfying LSB for unconstrained LSFEM. The right set of figures show the contours of Lagrange multiplier enforcing LSB constraint using the proposed LSFEM.	177
Figure 5.7	Numerical h -convergence study: These figures show the decrease of $\epsilon_{\text{MaxAbsLSB}}$ and ϵ_{AbsGSB} with respect to XSeed for a series of hierarchical three-node triangular and four-node quadrilateral meshes.	178
Figure 5.8	Numerical h -convergence study: This figure shows the CPU time (in seconds) of the proposed computational framework for unconstrained primitive and unconstrained negatively stabilized streamline diffusion LSFEMs.	179
Figure 5.9	Numerical h -convergence study: This figure shows the computational overhead incurred in satisfying LSB as compared to that of the corresponding unconstrained formulations.	179

Figure 5.10 Thermal boundary layer problem: This figure shows a pictorial description of the boundary value problem.	181
Figure 5.11 Thermal boundary layer problem: This figure shows the contours of concentration obtained for both unconstrained and constrained LS-FEMs based on Q4 finite element mesh.	181
Figure 5.12 Thermal boundary problem: This figure shows the contours of the error incurred in satisfying LSB for various unconstrained LSFEM formulations using Q4 meshes.	182
Figure 5.13 1D irreversible bimolecular fast reaction problem: A pictorial description of the boundary value problem.	186
Figure 5.14 1D irreversible bimolecular fast reaction problem (Case #1): This figure compares the concentration profile of the reactants and the product with the analytical solution.	187
Figure 5.15 1D irreversible bimolecular fast reaction problem (Case #2): This figure compares the concentration profile of the chemical species A , B , and C to that of the analytical solution.	189
Figure 5.16 Plume development from boundary in a reaction tank: The top figure provides a pictorial description of the boundary value problem. The bottom figure shows the contours of the stream function corresponding to the advection velocity vector field.	193
Figure 5.17 Plume development from boundary in a reaction tank (Type #1): This figure shows the concentration profiles of the product C based on unconstrained primitive LSFEM.	194

Figure 5.18	Plume development from boundary in a reaction tank (Type #1): This figure shows the concentration profiles of the product C based on unconstrained and constrained negatively stabilized streamline diffusion LSFEM.	194
Figure 5.19	Plume development from boundary in a reaction tank (Type #2): This figure shows the concentration profiles of the product C based on unconstrained primitive LSFEM.	195
Figure 5.20	Plume development from boundary in a reaction tank (Type #2): This figure shows the concentration profiles of the product C based on unconstrained and constrained negatively stabilized streamline diffusion LSFEM.	196
Figure 5.21	Plume development from boundary in a reaction tank (Type #1): This figure shows the variation Θ_C^2 with mesh refinement under the weighted negatively stabilized streamline diffusion LSFEM.	197
Figure 5.22	Plume development from boundary in a reaction tank (Type #1): This figure shows the variation $\log(\Theta_C^2)$ with respect to $\sqrt{\text{Pe}_L}$ for isotropic diffusivity under the weighted negatively stabilized streamline diffusion LSFEM with LSB and DMP constraints.	197
Figure 5.23	Vortex-stirred mixing in a reaction tank: A pictorial description of the initial boundary value problem	198
Figure 5.24	Vortex-stirred mixing in a reaction tank: This figure shows the concentration profiles of the product C after the first time-step using the unconstrained weighted negatively stabilized streamline diffusion LSFEM.	199

Figure 5.25	Vortex-stirred mixing in a reaction tank: This figure shows the concentration profiles of the product C at $y = 0.5$ after the first time-step using the unconstrained weighted negatively stabilized streamline diffusion LSFEM.	200
Figure 5.26	Vortex-stirred mixing in a reaction tank: This figure shows the concentration profiles of the product C at various time levels using the weighted negatively stabilized streamline diffusion LSFEM with and without constraints.	201
Figure 5.27	Transport-controlled mixing in cellular flows: A pictorial description of the initial boundary value problem and associated advection velocity field for the cellular flow.	202
Figure 5.28	Transport-controlled mixing in cellular flows: This figure shows the concentration profiles of the product C at various time levels using the unconstrained and constrained weighted negatively stabilized streamline diffusion LSFEM.	203
Figure 6.1	Anisotropic diffusivity tensor test problem: The left figure provides a pictorial description of the problem with the relevant boundary and initial conditions. The right figure shows the anisotropic \mathcal{M} -uniform mesh employed in the computational study.	219
Figure 6.2	Anisotropic diffusivity tensor test problem (mesh and time-step restrictions): The above figures show the concentration profiles for steady-state and transient cases.	221
Figure 6.3	Autocatalytic reaction-diffusion test problem: The left figure provides a pictorial description of the test problem with the relevant reaction source term, boundary conditions, and initial conditions.	251

Figure 6.4	Autocatalytic reaction-diffusion test problem (steady-state and case # 1): This figure shows the concentration profile obtained using a coarse mesh (21×21) at various iteration levels based on Pao's method.	252
Figure 6.5	Autocatalytic reaction-diffusion test problem (transient and case # 1): This figure shows the concentration profile obtained using a coarse mesh (21×21) at different time levels based on Pao's method for the case when $\Delta t = 10^{-4}$	253
Figure 6.6	Autocatalytic reaction-diffusion test problem (transient and case # 1): This figure shows the concentration profile obtained using a h -refined mesh (81×81) at different time levels based on Pao's method for the case when $\Delta t = 10^{-4}$	253
Figure 6.7	Autocatalytic reaction-diffusion test problem (transient and case # 2): This figure shows the converged concentration profiles obtained using a coarse and a h -refined mesh at the first time-step using different Δt 's based on Pao's method.	254
Figure 6.8	Autocatalytic reaction-diffusion test problem (transient and case # 2): This figure shows the converged concentration profiles obtained using a h -refined mesh (81×81) at various time levels using $\Delta t = 10^{-3}$ based on Pao's method.	255
Figure 6.9	Autocatalytic reaction-diffusion test problem (case # 3): This figure shows the converged concentration profiles obtained using a h -refined mesh (161×161) at first time step using different Δt 's based on Pao's method.	256

Figure 6.10 Autocatalytic reaction-diffusion test problem (transient and case # 4): This figure shows the converged concentration profiles obtained using a h -refined mesh (161×161) at first time step using different Δt 's based on Pao's method.	257
Figure 6.11 Autocatalytic reaction-diffusion test problem (transient and case # 4): This figure shows the converged concentration profiles obtained using a h -refined mesh at various time levels using $\Delta t = 10^{-1}$ based on Pao's method.	257

List of Tables

Table 3.1	Global species balance error for CDIMA reaction scheme for various h -refined meshes at $t = 1$. Analysis is performed for anisotropic diffusivity. Negative values for concentration of chemical species are clipped.	27
Table 3.2	Global species balance error for BZ reaction scheme for various h -refined meshes at $t = 1$. Analysis is performed for anisotropic diffusivity. Negative values are clipped.	29
Table 4.1	Fractured domain with isotropic diffusivity: For AD equation	116
Table 4.2	Fractured domain with isotropic diffusivity: For ADR equation	116
Table 4.3	Errors in local and global species balance: For pure isotropic and anisotropic diffusion equation.	118
Table 6.1	Anisotropic diffusivity tensor test problem: Quantitative results for minimum concentration and % of nodes that have violated the non-negative constraint in the computational domain.	217
Table 6.2	Autocatalytic reaction-diffusion test problem (steady-state and case # 1): Quantitative results for steady-state analysis using Pao's method on coarse mesh (21×21) at various iteration levels.	248
Table 6.3	Autocatalytic reaction-diffusion test problem (transient and case # 1): Quantitative results for transient analysis using Pao's method on coarse mesh (21×21) at various iteration levels.	249
Table 6.4	Autocatalytic reaction-diffusion test problem (transient and case # 2): Quantitative (converged) results for transient analysis based on Pao's method. Numerical simulations are performed using coarse and h -refined meshes at first time-step using different Δt 's.	249

Table 6.5	Autocatalytic reaction-diffusion test problem (steady-state and case # 3): Quantitative results for steady-state analysis using Pao's method on a h -refined mesh (161×161) at various iteration levels.	249
Table 6.6	Autocatalytic reaction-diffusion test problem (transient, case # 3, and case # 4): Quantitative (converged) results for transient analysis based on Pao's method. Numerical simulations are performed using h -refined meshes at first time-step using different Δt 's.	250
Table 6.7	Autocatalytic reaction-diffusion test problem (steady-state, case # 1, and case # 3): Quantitative results for steady-state analysis based on traditional NR method at various iteration levels. Numerical simulations are performed using various h -refined meshes.	250

Notation

The following are some key and frequently used symbols in the dissertation:

- a/τ : Lowercase English/Greek alphabet denotes scalars
- \mathbf{a} : Lowercase boldface normal letters denotes the continuum vectors
- \mathbf{A} : Uppercase boldface normal letters denotes the second-order tensors
- \mathbf{a} : Lowercase boldface italic letters denotes the finite element vectors
- \mathbf{A} : Uppercase boldface italic letters denotes the finite element matrices
- \mathbb{I} : Fourth-order identity tensor
- \mathbb{T} : Transposer, a fourth-order tensor
- \mathbb{S} : Symmetrizer, a fourth-order tensor
- \mathbb{R} : The set of real numbers
- \mathcal{O} : The Big-Oh notation
- nd : The number of spatial dimensions
- Ω : Open bounded domain
- $\partial\Omega$: Boundary of the domain
- t : Time
- \mathcal{I} : Total time of interest
- \mathbf{x} : A spatial point
- $c(\mathbf{x}, t)$: Concentration field of single chemical species
- $c_i(\mathbf{x}, t)$: Concentration field of i -th chemical species
- $\mathbf{q}(\mathbf{x}, t)$: Total flux vector
- $\alpha(\mathbf{x}, t)$: Decay/Linear-reaction coefficient
- $\mathbf{v}(\mathbf{x}, t)$: Advection velocity vector field
- $\mathbf{D}(\mathbf{x}, t)$: Diffusivity tensor
- $\mathbf{f}(\mathbf{x}, t)$: Non-reactive component of volumetric source
- $r_i(\mathbf{x}, t, c_i)$: Reactive component of volumetric source
- L2 : Two node linear element
- T3 : Three node triangular element
- Q4 : Four node quadrilateral element
- LSB : Local species balance
- GSB : Global species balance

AD	:	Advection-diffusion
ADR	:	Advection-diffusion-reaction
PDE	:	Partial differential equation
DAE	:	Differential-algebraic equations
DAI	:	Differential-algebraic inequalities
KKT	:	Karush-Kuhn Tucker conditions
DGF	:	Discrete Green's function
WCT	:	Well-centered triangular mesh
NN/NC	:	Non-negative constraint
DMP	:	Discrete maximum principle
DCP	:	Discrete comparison principle
wMP	:	Weak maximum principle
WMP	:	Strict weak maximum principle
sMP	:	Strong maximum principle
SMP	:	Strict strong maximum principle
wCP	:	Weak comparison principle
sCP	:	Strong comparison principle
DwMP/DwMP \mathcal{K}	:	Discrete weak maximum principle
DWMP/DWMP \mathcal{K}	:	Discrete strict weak maximum principle
DsMP/DsMP \mathcal{K}	:	Discrete strong maximum principle
DSMP/DSMP \mathcal{K}	:	Discrete strict strong maximum principle
DwCP/DwCP \mathcal{K}	:	Discrete weak comparison principle
DsCP/DsCP \mathcal{K}	:	Discrete strong comparison principle
XSeed	:	Number of (finite element) nodes in a mesh along x-direction
YSeed	:	Number of (finite element) nodes in a mesh along y-direction
N_{ele}	:	Total number of non-overlapping open sub-domains
N_v	:	Total number of vertices
N_{iv}	:	Total number of interior vertices
\mathcal{T}_h/Ω_h	:	Finite element mesh
h	:	Finite element mesh size
$\mathcal{M}(\mathbf{x})$:	Metric tensor to generate a DMP-based mesh
MaxIters	:	Maximum number of iterations
StopCrit	:	A stopping criteria

MOHL	:	Method of horizontal lines
MOVL	:	Method of vertical lines
FEM	:	Finite element method
FDS	:	Finite difference schemes
FVM	:	Finite volume method
MFDM	:	Mimetic finite difference methods
PP	:	Post-processing methods
LSFEM	:	Least-squares finite element methods
CDIMA	:	Chlorine dioxide-iodine-malonic acid reaction
BZ	:	Belousov-Zhabotinsky reaction
SG	:	Standard single-field Galerkin finite element method
SUPG	:	Streamline Upwind Petrov Galerkin finite element method
GLS	:	Galerkin Least-Squares finite element method
$\delta_o, \delta_1, \delta_2, \tau_o, \tau_1, \tau_2$:	Non-negative user-defined element stabilization parameters
$\text{grad}[\bullet]$:	The gradient operator with respect to \mathbf{x}
$\text{div}[\bullet]$:	The divergence operator with respect to \mathbf{x}
\boxtimes	:	Box product
\odot	:	Kronecker product
$(\bullet; \bullet)_K$:	The standard L_2 inner product over set K
$\langle \bullet; \bullet \rangle$:	The standard inner-product in Euclidean spaces
$\ \bullet\ $:	Standard Euclidean norm
$\text{vec}[\bullet]$:	Reshaping operator of a matrix
$\text{mat}[\bullet]$:	Reshaping operator of a four-dimensional array
$\text{mat}_1[\bullet], \text{mat}_2[\bullet]$:	Reshaping operators of a three-dimensional array
ϵ_{mach}	:	The machine precision for a 64-bit machine
Pe_h	:	An element Péclet number
Da_h	:	An element Damköhler number
Pe_D/Pe_D	:	Physics-based Péclet number
Da_I	:	Physics-based Damköhler number of first kind
$\text{Da}_{II,D}/\text{Da}_{II,D}$:	Physics-based Damköhler number of second kind
$\mathfrak{F}_{\text{Prim}}$:	Weighted primitive LSFEM
$\mathfrak{F}_{\text{NgStb}}$:	Weighted negatively stabilized streamline diffusion LSFEM

- $C(\Omega)/C^0(\Omega)$: The set of continuous functions on Ω
- $C^k(\Omega)$: The set of functions having derivatives up to the order k continuous
- $C^\infty(\Omega)$: The set of infinitely differentiable functions
- $C_0^\infty(\Omega)$: The set of infinitely differentiable functions with compact support
- $\mathcal{C}/\mathcal{W}/\mathcal{Q}$: Function spaces for weak formulations
- $\mathcal{C}_h/\mathcal{W}_h/\mathcal{Q}_h$: Finite element function spaces
- $H^1(\Omega)$: Standard Sobolev space
- $H^m(\Omega)$: Standard Sobolev space for non-negative integer m

Chapter 1

INTRODUCTION

“Not ignorance, but ignorance of
ignorance, is the death of
knowledge.”

Alfred North Whitehead

Advection-diffusion-reaction partial differential equations are pervasive in engineering, sciences, and economics. Obtaining stable and accurate numerical solutions for these equations can be challenging. This is because the underlying equations are coupled, nonlinear, and non-self-adjoint. Currently, there is neither a robust computational framework available nor a reliable commercial package known that can handle various complex situations. Hence, this dissertation aims at developed robust numerical methodologies to solve advective- diffusive-reactive systems that provide accurate and physical solutions for a wide range of input data, Péclet numbers, Damköhler numbers, and for complicated geometries.

In continuous setting, the governing equations for advective-diffusive-reactive systems possess various important properties (such as non-negativity, maximum principles, comparison principles, monotonicity, monotone property, local species balance, and global species balance). All these properties are *not* inherited during finite difference, finite volume, and finite element discretizations. In this dissertation, we *unequivocally* demonstrate that many existing numerical and commercial packages

do not provide physically meaningful values for the concentration of the chemical species for various realistic benchmark problems. We shall also show that the popular stabilized numerical formulations violate non-negative constraint and various other discrete principles. Furthermore, through representative numerical simulations, we show that *unphysical* values for concentration of chemical species due to violation of non-negative constraint and spurious node-to-node oscillations will result in *large errors* in local and global species balance. In discrete setting, there are various ways to satisfy different discrete principles, local species balance, and global species balance. In this dissertation, we shall provide a comprehensive analysis of various versions of comparison principles, maximum principles, and the non-negative constraint and their influence on meeting local and global species balance.

Each chapter is self contained. It has its own introduction, literature survey, and design philosophy of the proposed numerical methods. Chapter 2 provides a quick summary of the mathematical properties of underlying partial differential equations. In Chapter 3, we show the deficiencies of some popular stabilized finite element formulations such as SUPG (Streamline Upwind Petrov Galerkin) and GLS (Galerkin Least-Squares) for nonlinear ecological models and chemical reacting flows. We also quantify the errors incurred in satisfying the local and global species balance for various realistic benchmark problems. This demonstrate the need and importance of developing locally conservative non-negative numerical formulations for chemically reacting systems.

Chapter 4 discusses a methodology based on mesh restrictions to overcome such limitations. We proposed a nonlinear iterative method to generate meshes using various open source mesh generators such as Gmsh, BAMG, and FreeFem++ to satisfy different discrete principles. Using representative numerical examples and mathematical analysis, pros and cons of mesh restriction approach is critically analyzed. This research work forms the basis for Chapter 5. In this Chapter, we propose a

new and novel methodology based on least-squares finite element framework combining the principles of constrained optimization methods to satisfy local and global species balance, comparison principles, maximum principles, and the non-negative constraint on coarse general computational grids. Using this framework, we also obtained numerically a scaling law for a transport-controlled bimolecular reaction. Chapter 6 discusses nonlinear techniques and numerical formulations for steady-state and transient semi-linear reaction-diffusion equations, and their structure preserving properties. Numerical experiments for a popular physics-based chemical reaction model are studied. Finally, in Chapter 7, conclusions are drawn and future directions are outlined.

Chapter 2

A PRIMER ON PARTIAL DIFFERENTIAL EQUATIONS

“Every kind of science, if it has only reached a certain degree of maturity, automatically becomes a part of mathematics.”

David Hilbert

This chapter is devoted to a quick introduction to the mathematical analysis and some theoretical issues of some popular partial differential equations (PDEs). Herein, our research focuses on various particular PDEs that are important for applications within the context of diffusion-type equations Evans (1998).

2.1 DEFINITION AND CLASSIFICATION OF PDES

In general, a PDE for a function $u(x_1, \dots, x_n)$ is of the form as

$$F(x_1, \dots, x_n, u, u_{x_1}, \dots, u_{x_n}, u_{x_1x_1}, u_{x_1x_2}, \dots) = 0. \quad (2.1.1)$$

The *order* of the above equation is the highest derivative occurring in the equation. A PDE is called *linear* if it depends linearly on u and its derivatives. If all derivatives of

u occur linearly with coefficients depending only on \mathbf{x} , then the equation is *semilinear*. If all highest-order derivatives of u occur linearly with coefficients depending only on \mathbf{x} , u and lower-order derivatives of u , then the equation is *quasilinear*. Otherwise, the equation is *nonlinear*.

In dealing with partial differential equations, it is useful to differentiate between several types. In particular, we classify PDEs of second order as *elliptic*, *hyperbolic*, and *parabolic*. Both the theoretical and numerical treatment differ considerably for the three types. The general linear PDE of second order in n variables $\mathbf{x} = (x_1, \dots, x_n)$ has the form as

$$\sum_{i,j=1}^n a_{ij}(\mathbf{x})u_{x_i x_j} + \sum_{i=1}^n b_i(\mathbf{x})u_{x_i} + c(\mathbf{x})u = f(\mathbf{x}). \quad (2.1.2)$$

Since $u_{x_i x_j} = u_{x_j x_i}$ for any function which is twice continuously differentiable, without loss of generality we can assume the symmetry $a_{ij}(\mathbf{x}) = a_{ji}(\mathbf{x})$. The corresponding $n \times n$ matrix is denoted as $\mathbf{A} = (a_{ij}(\mathbf{x}))$.

Definition 2.1.1. *The equation (2.1.2) is called elliptic at the point \mathbf{x} provided the matrix $\mathbf{A}(\mathbf{x})$ is positive definite. The equation (2.1.2) is called hyperbolic at the point \mathbf{x} provided the matrix $\mathbf{A}(\mathbf{x})$ has one negative eigenvalue and $n-1$ positive eigenvalues. The equation (2.1.2) is called parabolic at the point \mathbf{x} provided the matrix $\mathbf{A}(\mathbf{x})$ is positive semidefinite but not positive definite. An equation is called elliptic, hyperbolic or parabolic provided it has the corresponding property for all points of the domain.*

We now specifically consider second-order linear partial differential equation. A general form of such an equation can be written as follows:

$$a_{11}u_{xx} + a_{12}u_{xy} + a_{22}u_{yy} + b_1u_x + b_2u_y + cu = f, \quad (2.1.3)$$

where

- Elliptic equation: $a_{12}^2 < 4a_{11}a_{22}$
- Hyperbolic equation: $a_{12}^2 > 4a_{11}a_{22}$
- Parabolic equation: $a_{12}^2 = 4a_{11}a_{22}$.

Definition 2.1.2. *A boundary (or initial) value problem is called well-posed (in the sense of Hadamard) if a solution exists, is unique, and depends continuously on the given data.*

2.2 NOTATION ABOUT $C^M(\Omega)$, $C^M(\bar{\Omega})$ AND SOBOLEV SPACES

Let $\Omega \subset \mathbb{R}^n$ be an open set. Let $\bar{\Omega}$ denote the closure (all limit points) of Ω . The boundary $\partial\Omega$ of a domain Ω is the set of all limit points of Ω . That is, $\partial\Omega = \bar{\Omega} - \Omega$. If $\Omega = \mathbb{R}^n$ then $\partial\Omega = \emptyset$. We shall denote the continuous functions on Ω as $C(\Omega)$, and those whose first-order derivatives are also all continuous functions by $C^1(\Omega)$. Similarly, for $k \in \mathbb{N}$, $C^k(\Omega)$ denotes the functions having all derivatives up to the order k continuous on Ω . Let $C(\bar{\Omega})$ denote the space of all continuous functions on Ω that can be continuously extended to the boundary $\partial\Omega$. Similarly, one can define $C^k(\bar{\Omega})$. The set of infinitely differentiable (or sometimes called smooth) functions is denoted by $C^\infty(\Omega)$. The set of all infinitely differential functions with compact support is denoted by $C_0^\infty(\Omega)$ or $C_c^\infty(\Omega)$. It should be noted that even if Ω is bounded, a function $u \in C^k(\Omega)$ may not be bounded as it may grow near the boundary.

Definition 2.2.1. *The support of a function $f(\mathbf{x})$ defined on $\Omega \subset \mathbb{R}^n$ is the closure of the set of points where $f(\mathbf{x})$ is nonzero: $\text{supp } f = \overline{\{\mathbf{x} \in \Omega : f(\mathbf{x}) \neq 0\}}$.*

Definition 2.2.2. *A function $f(\mathbf{x})$ defined on a domain Ω is integrable if $\int_\Omega |f(\mathbf{x})| \, d\Omega$ is defined and bounded. We shall denote all such functions by $L^1(\Omega)$.*

Definition 2.2.3. *A function $f(\mathbf{x})$ defined on a domain Ω is square integrable if $\int_\Omega f^2 \, d\Omega$ is defined and bounded. We shall denote all such functions by $L_2(\Omega)$. That*

is,

$$L_2(\Omega) := \{f(\mathbf{x}) : \Omega \rightarrow \mathbb{R} \mid \int_{\Omega} f^2 \, d\Omega < +\infty\}. \quad (2.2.1)$$

Definition 2.2.4. The space $L_{loc}^1(\Omega)$ is set of all functions that are “locally” integrable, i.e., integrable on any compact subset of Ω but not necessarily integrable at the boundary of Ω or at infinity.

Definition 2.2.5. $L^p(\Omega)$ is the set of all (measurable) functions $u(\mathbf{x})$ defined on Ω such that the norm

$$\|u\|_{p,\Omega} := \left(\int_{\Omega} |u(\mathbf{x})|^p \, d\Omega \right)^{1/p}$$

is finite.

Remark 2.2.6. Note that functions in $L^1(\Omega)$ or $L_{loc}^1(\Omega)$ may have discontinuities (including singularities) provided they are not too severe for the integral to converge.

From the above definitions, we have the following inclusions:

1. $C(\overline{\Omega}) \subset C(\Omega)$
2. $C^\infty(\Omega) \subset \dots \subset C^1(\Omega) \subset C(\Omega) \subset L^1(\Omega) \subset L_{loc}^1(\Omega)$

Definition 2.2.7. The $H^1(\Omega)$ is the set of all functions that belong to $L_2(\Omega)$ and their partial derivatives are also square integrable. That is,

$$H^1(\Omega) := \left\{ f(\mathbf{x}) \in L_2(\Omega) \mid \frac{\partial f}{\partial x_i} \in L_2(\Omega), i = 1, \dots, n \right\}. \quad (2.2.2)$$

Similarly, $H^2(\Omega)$ is the set of all functions that belong to $H^1(\Omega)$ and each component of the Hessian belongs to $L_2(\Omega)$. That is,

$$H^2(\Omega) := \left\{ f(\mathbf{x}) \in H^1(\Omega) \mid \frac{\partial^2 f}{\partial x_i \partial x_j} \in L_2(\Omega); i, j = 1, \dots, n \right\}. \quad (2.2.3)$$

The spaces $H^1(\Omega), H^2(\Omega), \dots$ are called Sobolev spaces.

Theorem 2.2.8. *The spaces $H^k(\Omega)$ ($k = 1, 2, \dots$) are all Hilbert spaces.*

Chapter 3

ON LOCAL AND GLOBAL SPECIES CONSERVATION ERRORS FOR NONLINEAR ECOLOGICAL MODELS AND CHEMICALLY REACTING FLOWS

“A theory with mathematical
beauty is more likely to be correct
than an ugly one that fits some
experimental data.”

Paul Dirac

Advection-controlled and diffusion-controlled oscillatory chemical reactions appear in various areas of life sciences, hydrogeological systems, and contaminant transport. These types of reactions are everywhere in nature, ranging from large-scale atmospheric and ocean currents to flow past micro-organisms and plankton in porous media. *In this chapter*, we analyze whether the existing numerical formulations and commercial packages provide physically meaningful values for concentration of the chemical species for two popular oscillatory chemical kinetic schemes. The first one corresponds to the CDIMA (chlorine dioxide-iodine-malonic acid) reaction while the

second one is a simplified version of BZ (Belousov-Zhabotinsky) reaction of a non-linear chemical oscillator. The governing equations for species balance are presented based on the theory of interacting continua. This results in a set of coupled non-linear partial differential equations. Obtaining analytical solutions is not practically viable. Moreover, it is well-known in literature that if the local dynamics becomes complex, the range of possible dynamic behavior in the presence of diffusion and advection becomes practically unlimited. Hence, we resort to numerical solutions. These solutions are constructed using popular finite element formulations such as SUPG (Streamline Upwind Petrov Galerkin) and GLS (Galerkin Least-Squares). The resulting non-linear system of equations are solved using the Newton-Raphson method. However, it should be noted that the numerical solution behavior is dependent on certain stoichiometric coefficients and chemical reaction rates (which are not constants). In order to make the computational analysis tractable, an estimate on the range of these system-dependent parameters is obtained based on model reduction performed on the strong-form of the governing equations. Finally, we quantify the errors incurred in satisfying the local and global species balance for various realistic benchmark problems. Through these representative numerical examples, we shall demonstrate the need and importance of developing locally conservative non-negative numerical formulations for chaotic and oscillatory chemically reacting systems.

3.1 INTRODUCTION TO TRANSPORT-CONTROLLED OSCILLATORY CHEMICAL REACTIONS

Chemical reactions involving a number of interacting chemical species with non-linear kinetics are ubiquitous in air pollution modeling Neufeld and H.-García (2010), plankton population dynamics Epstein and Pojman (1998), and contaminant transport Yong and Thomas (1997). It has been reported in the literature that some components of a non-linear chemical kinetics system can exhibit oscillatory dynamics

in certain range of system parameters Poppe and Lustfeld (1997). Typically, when the period of these oscillations is within the range of characteristic timescales of advection/diffusion processes, there will be a significant change in the global dynamics of the chemically reacting system. For example, oceanic plankton, phytoplankton–zooplankton, and other more complicated plankton population models exhibit oscillatory solutions Edwards and Yool (2000). Moreover, oscillations and chaotic fluctuations generated by the plankton population dynamics can provide a mechanism for the coexistence of enormous number of plankton species, which are competing for limited key resources Huisman and Weissing (1999). Some popular oscillatory chemical reaction schemes, which are commonly studied by chemists, physicists, and mathematicians, are Lotka-Volterra, Briggs-Rauscher, Bray-Liebhafsky, Belousov-Zhabotinsky, CIMA, and CDIMA Neufeld and H.-García (2010); Epstein and Pojman (1998).

3.1.1 Governing equations for mixing in oscillatory media

The interplay between mixing and oscillations in chemically reacting flows and biological systems have been studied in different contexts. A popular example is the study of chaotic dynamics of oscillatory chemical reactions in a stirred reactor. Stronger stirring leads to more uniform concentrations of chemical species within the reactor. Hence, one expects that such a system should be well approximated by a set of differential equations that describes the temporal dynamics of the mean concentrations (independently of the stirring rate). However, it is well-known from experiments that significant non-uniformities in the concentration field persist even at high stirring rates, resulting in *stirring effects* Menzinger and Dutt (1990); Noszticzius et al. (1991). Such effects cannot be captured by simple models that assume spatially uniform concentrations Neufeld and H.-García (2010).

The framework offered by the theory of interacting continua Bowen (1976) can provide complex models, which can describe non-uniformities in the concentration

field and other intricate effects resulting from higher stirring rates Neufeld and H.-García (2010). This framework is based on continuum description and is ideal to study advection-controlled or diffusion-controlled oscillatory chemical reactions. The governing equations for the fate of the chemical species are given as follows:

$$\frac{\partial c_i}{\partial t} + \text{div}[\mathbf{v}(\mathbf{x}, t)c_i - \mathbf{D}(\mathbf{x}, t) \text{grad}[c_i]] = f_i(\mathbf{x}, t) + r_i(\mathbf{x}, t, c_1, \dots, c_n) \quad \text{in } \Omega \times]0, \mathcal{I}[, \quad (3.1.1a)$$

$$c_i(\mathbf{x}, t) = c_i^{\text{P}}(\mathbf{x}, t) \quad \text{on } \Gamma_i^{\text{D}} \times]0, \mathcal{I}[, \quad (3.1.1b)$$

$$(\mathbf{v}(\mathbf{x}, t)c_i(\mathbf{x}, t) - \mathbf{D}(\mathbf{x}, t) \text{grad}[c_i(\mathbf{x}, t)]) \bullet \hat{\mathbf{n}}(\mathbf{x}) = h_i^{\text{P}}(\mathbf{x}, t) \quad \text{on } \Gamma_i^{\text{N}} \times]0, \mathcal{I}[, \quad \text{and} \quad (3.1.1c)$$

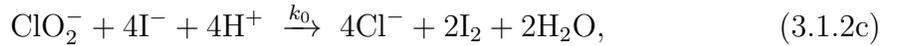
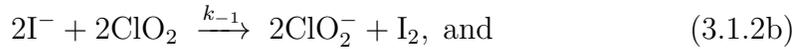
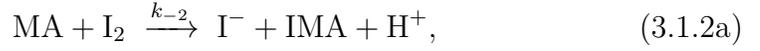
$$c_i(\mathbf{x}, t = 0) = c_i^{\text{O}}(\mathbf{x}) \quad \text{in } \Omega, \quad (3.1.1d)$$

where $c_i(\mathbf{x}, t)$ is the molar concentration of the i -th chemical species ($i = 1, \dots, n$), $\mathbf{v}(\mathbf{x}, t)$ is the advection velocity vector field, $\mathbf{D}(\mathbf{x}, t)$ is the diffusivity tensor. $f_i(\mathbf{x}, t)$ and $r_i(\mathbf{x}, t, c_1, \dots, c_n)$ are, respectively, the non-reactive and reactive components of the volumetric source. $c_i^{\text{O}}(\mathbf{x})$ is the initial condition of i -th chemical species. Correspondingly, $c_i^{\text{P}}(\mathbf{x}, t)$ and $h_i^{\text{P}}(\mathbf{x}, t)$ are the prescribed concentration and normal flux on the boundary. Ω is the domain in which the chemical reaction takes place, which is typically assumed to be a bounded open set in a mathematical setting. Γ_i^{D} and Γ_i^{N} are respectively the Dirichlet and Neumann boundaries of the domain. The time is denoted by $t \in]0, \mathcal{I}[$, where \mathcal{I} is the total time of interest.

3.1.2 Problem statement and main contributions

In continuous setting, the governing equations given by (3.1.1a)–(6.4.25e) possess various important properties such as non-negativity, maximum principles, comparison principles, monotonicity, monotone property, local species balance, and global species balance Nakshatrala and Valocchi (2009); Nagarajan and Nakshatrala (2011); Mudunuru and Nakshatrala (2012); Nakshatrala et al. (2013), Nakshatrala et al.

(2013); Karimi and Nakshatrala (2015); Mudunuru and Nakshatrala (2015). In general, all these properties are *not inherited* during finite difference, finite volume, and finite element discretizations. *In this chapter*, we analyze whether the existing numerical and commercial packages provide physically meaningful values for the concentration of the chemical species for various realistic benchmark problems. Furthermore, we also quantify the errors incurred in satisfying the local and global species balance for two popular chemical kinetics schemes Neufeld and H.-García (2010). First one being an important and versatile group of oscillatory reactions involving the chlorite ion and iodine-containing reactants. Examples include CIMA (chlorite-iodine-malonic acid) and the CDIMA (chlorine dioxide-iodine-malonic acid) reactions schemes Epstein and Pojman (1998). A simplified reaction scheme capturing the essential features of the chemical kinetics is as follows:



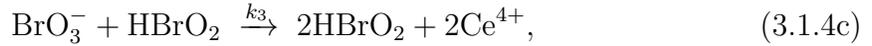
where MA and IMA are the malonic acid and the iodomalonic acid, respectively. Based on the usual experimental conditions, for the CDIMA case, it is appropriate to assume that c_{I^-} and $c_{\text{ClO}_2^-}$ are the only dynamical variables. The concentration of other chemical species stay essentially as constants. With these approximations, an appropriate reactive-component volumetric source model for the CDIMA reaction is given as

$$r_{\text{I}^-} = \alpha_1 - c_{\text{I}^-} - \frac{4c_{\text{I}^-}c_{\text{ClO}_2^-}}{1 + c_{\text{I}^-}^2} \text{ and} \quad (3.1.3a)$$

$$r_{\text{ClO}_2^-} = \alpha_2 \left(c_{\text{I}^-} - \frac{c_{\text{I}^-}c_{\text{ClO}_2^-}}{1 + c_{\text{I}^-}^2} \right), \quad (3.1.3b)$$

where α_1 and α_2 are parameters, which depend on the reaction rate k_0 .

Second one corresponds to a (three-species) Oregonator model, which is a simplified version of BZ (Belousov-Zhabotinsky) reaction scheme of a non-linear chemical oscillator. This is given as follows Neufeld and H.-García (2010):



where X = oxidizable organic species (for example, MA or IMA) and γ is a stoichiometric parameter. It should be noted that there are several simplified versions of the BZ reaction as the chemical mechanism is very complex. But the most common version involves the oxidation of malonic acid (MA) by bromate ions (BrO_3^-) in acid medium and catalyzed by cerium, which oscillates during the reaction between the Ce^{4+} and Ce^{3+} states. The first two reactions given by equations describe the depletion of bromide (Br^-). While the last three reactions build the concentration of HBrO_2 and Ce^{4+} that finally leads to Br^- recovery, and then to a new cycle. Assuming that $c_{\text{BrO}_3^-}$ and c_X remain constant, and noticing that HOBr enters the chemical reaction as a passive product, an appropriate reactive-component volumetric source model for the BZ reaction is given as follows Field and Noyes (1974):

$$r_{\text{HBrO}_2} = \frac{1}{\epsilon_1} \left(\beta c_{\text{Br}^-} - c_{\text{HBrO}_2} c_{\text{Br}^-} + c_{\text{HBrO}_2} (1 - c_{\text{HBrO}_2}) \right), \quad (3.1.5a)$$

$$r_{\text{Br}^-} = \frac{1}{\epsilon_2} \left(\gamma c_{\text{Ce}^{4+}} - \beta c_{\text{Br}^-} - c_{\text{HBrO}_2} c_{\text{Br}^-} \right), \text{ and} \quad (3.1.5b)$$

$$r_{\text{Ce}^{4+}} = c_{\text{HBrO}_2} - c_{\text{Ce}^{4+}}, \quad (3.1.5c)$$

where the constants $\epsilon_1 > 0$, $\epsilon_2 > 0$, and $\beta > 0$ are given as

$$\epsilon_1 = \frac{k_5 c_X}{k_3 c_{\text{BrO}_3^-}}, \quad \epsilon_2 = \frac{2k_4 k_5 c_X}{k_2 k_3 c_{\text{BrO}_3^-}}, \quad \beta = \frac{2k_1 k_4}{k_2 k_3}. \quad (3.1.6)$$

It should be noted that as γ is a stoichimetric coefficient, we need to have $\gamma > 0$.

3.1.3 Outline of this chapter

The remainder of this chapter is organized as follows. Section 3.2 presents a model reduction method based on dynamical systems approach to obtain qualitative nature of the solution for CDIMA and BZ reaction schemes. In Section 3.3, we present a non-linear finite element formulation based on SUPG and GLS to obtain numerical solutions for our transport-controlled chemically reacting flows. The resulting system of equations are solved using the Newton-Raphson method. In Section 3.4, representative numerical examples are presented to illustrate the performance of the proposed computational framework with respect to non-negativity, local species balance, and global species balance. Conclusions are drawn in Section 3.5.

3.2 REDUCED-ORDER MODELS: QUALITATIVE NATURE OF THE SOLUTION

The complex behavior in the oscillatory media arises from the diffusive and advection coupling of the local dynamics. The range of possible dynamic behavior in the presence of advection/diffusion becomes practically unlimited when the local dynamics becomes more complex. This is because coupling chaotic subsystems could produce an extremely rich dynamics (which is the case for even periodic local dynamics) Pikovsky and Popovych (2003). Advectively/diffusively coupled chemical and biological oscillators may become synchronized and spatial heterogeneity may lead to additional instabilities Pikovsky et al. (2001). This may result in target waves, spiral

patterns, front instabilities, and several different types of spatio-temporal chaos Kuramoto (2003). It is extremely difficult to simulate or capture all of these complex dynamic phenomena without properly understanding the qualitative behavior of the coupled advective-diffusive-reactive systems.

It should be noted that various model reduction frameworks exist in literature Antoulas et al. (2001) to understand the qualitative nature of the solution for the equations given by (3.1.1a)–(6.4.25e). Herein, we shall consider model reduction based on a simplified dynamical systems approach Neufeld and H.-García (2010). The dynamical system framework provides rich tools of mathematical results to analyze the solution behaviour for various interesting experimental scenarios Kinoshita (2013). This framework has many advantages, one of which is to obtain qualitative and quantitative information on the parameters such as α_1 , α_2 , β , ϵ_1 , and ϵ_2 (which is used in our finite element simulation to perform a parametric study).

To summarize, using dynamical systems framework, we reduce the non-linear coupled partial differential equations to an autonomous system of first-order coupled non-linear ordinary differential equations. The ROM equations for the CDIMA reaction scheme is given as

$$\frac{dc_{I^-}}{dt} = \alpha_1 - c_{I^-} - \frac{4c_{I^-}c_{ClO_2^-}}{1 + c_{I^-}^2} \text{ and} \quad (3.2.1a)$$

$$\frac{dc_{ClO_2^-}}{dt} = \alpha_2 \left(c_{I^-} - \frac{c_{I^-}c_{ClO_2^-}}{1 + c_{I^-}^2} \right). \quad (3.2.1b)$$

Correspondingly, the ROM equations for the BZ reaction based on the three-species

Oregonator model is given as

$$\frac{dc_{\text{HBrO}_2}}{dt} = \frac{1}{\epsilon_1} \left(\beta c_{\text{Br}^-} - c_{\text{HBrO}_2} c_{\text{Br}^-} + c_{\text{HBrO}_2} (1 - c_{\text{HBrO}_2}) \right), \quad (3.2.2a)$$

$$\frac{dc_{\text{Br}^-}}{dt} = \frac{1}{\epsilon_2} \left(\gamma c_{\text{Ce}^{4+}} - \beta c_{\text{Br}^-} - c_{\text{HBrO}_2} c_{\text{Br}^-} \right), \text{ and} \quad (3.2.2b)$$

$$\frac{dc_{\text{Ce}^{4+}}}{dt} = c_{\text{HBrO}_2} - c_{\text{Ce}^{4+}}. \quad (3.2.2c)$$

From continuous maximum principle Evans (1998), we need to have c_{I^-} , $c_{\text{ClO}_2^-}$, c_{HBrO_2} , c_{Br^-} , and $c_{\text{Ce}^{4+}}$ to be non-negative in the entire domain and time interval of interest.

3.2.1 Hopf bifurcations and chemical reaction system stability

In general, the numerical solution for the CDIMA and BZ reaction scheme ROM models are obtained by solving a systems of non-linear Differential-Algebraic Inequalities (DAI). In discrete setting, these set of non-linear DAIs are given as

$$\frac{d\mathbf{c}}{dt} = \mathbf{r}(\mathbf{c}), \quad (3.2.3a)$$

$$\mathbf{c}(t = 0) = \mathbf{c}_0 \succeq \mathbf{0}, \text{ and} \quad (3.2.3b)$$

$$\mathbf{c} \succeq \mathbf{0} \quad \forall t \in]0, \mathcal{I}[, \quad (3.2.3c)$$

where \mathbf{c} is the vector of unknown concentrations of chemical species for a given reaction scheme and \mathbf{r} is the reactive component of the volumetric source. For example, in-case of CDIMA scheme, $\mathbf{c} = [c_{\text{I}^-}, c_{\text{ClO}_2^-}]^T$. Similar inference can be drawn on BZ reaction scheme. It should be noted that the non-linear DAI system given by (3.2.3a)–(3.2.3c) may have multiple solutions. This is because in complex chemically reacting systems, many times it is difficult to estimate the values of the parameters related to stoichiometric coefficients and chemical reaction rates. Moreover, some parameters are frequently not fully identifiable from the experimental data at all Er-rami et al. (2015). This parametric uncertainty with large potential variations of

parameters by several orders of magnitudes often leads to severe limitations of existing numerical techniques to obtain physically meaningful values for concentrations of chemical species even for reduced-order models (see the following subsection 3.2.2). Constrained optimization-based numerical continuation methods can be used to obtain all the set of solutions. However, it should be noted that not all the solutions obtained are stable.

Stability or *multistability* of a given set of numerical solutions depends on the nature of *Hopf bifurcation fixed points*. Recently, efficient algorithmic methods using convex coordinates and tropical geometry are proposed for parametric detection of Hopf bifurcations fixed points Errami et al. (2015). This computational framework is for a class of chemical reaction networks with symbolic rate constants that obey *generalized mass action kinetics* such as equations (3.2.3a)–(3.2.3c). However, constructing such effective numerical methods within the context of transport-controlled chemically reacting systems has been rarely studied. This is because obtaining Hopf bifurcation fixed points and subsequently analyzing their stability nature for the following non-linear DAI system:

$$\mathbf{M} \frac{d\mathbf{c}_i}{dt} + \mathbf{K} \mathbf{c}_i = \mathbf{f}(t) + \mathbf{r}(\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_i, \dots, \mathbf{c}_{n-1}, \mathbf{c}_n) \text{ and} \quad (3.2.4a)$$

$$\mathbf{c}_i \succeq \mathbf{0} \quad \forall t \in]0, \mathcal{I}[\text{ and } \forall i = 1, 2, \dots, n \quad (3.2.4b)$$

is challenging due to the high computational costs. \mathbf{M} and \mathbf{K} are the mass and stiffness matrices of a given finite element discretization. \mathbf{c}_i is the nodal concentration of the i^{th} -chemical species. \mathbf{f} and \mathbf{r} are the corresponding discretized non-reactive and reactive components of the volumetric source. n is the total number of chemical species involved in the CDIMA/BZ reaction scheme. It should be noted that \mathbf{M} is symmetric and positive definite. While \mathbf{K} is neither symmetric nor positive definite.

In case of advection-dominated/reaction-dominated scenarios, this nature of the

matrix \mathbf{K} poses numerical difficulties in obtaining physically meaningful values for Hopf bifurcation fixed points (which should be non-negative). Detailed analysis of these fixed points and their stability within the context of finite element method is beyond the scope of this dissertation. Herein, for CDIMA and BZ reduced-order models, we have calculated the Hopf bifurcation fixed points and analyzed their stability analytically. The following subsection describe the nature of these fixed points and the qualitative aspects of the solution near them for all $t \in]0, \mathcal{I}[$. From this subsection, it will be evident that obtaining *non-negative parametric Hopf bifurcation fixed points* and examining their stability even for reduced-order models is highly demanding.

3.2.2 Stability analysis of CDIMA and BZ reduced-order models

The fixed points of the CDIMA reaction scheme are given as

$$c_{\text{I}^-}^* = \frac{\alpha_1}{5} \quad c_{\text{ClO}_2^-}^* = 1 + \frac{\alpha_1^2}{25}. \quad (3.2.5)$$

From equation (3.2.5), it is clear that $\alpha_1 \geq 0$ (as the concentration of chemical species has to be non-negative). The characteristic polynomial of the CDIMA reaction scheme based on linearizing the system of equations given by (3.2.1a)–(3.2.1b) near these fixed points is given as

$$\lambda_{\text{CDIMA}}^2 - \left(\frac{4\alpha_1^2 - 5\alpha_1\alpha_2 - 125}{\alpha_1^2 + 25} \right) \lambda_{\text{CDIMA}} + \frac{5\alpha_1\alpha_2(4\alpha_1^2 + 125)}{(\alpha_1^2 + 25)^2} = 0. \quad (3.2.6)$$

Based on Routh-Hurwitz stability criterion for second-order polynomial Wood (1990), it is evident that the fixed points given by equation (3.2.5) are stable if the following conditions are satisfied:

$$4\alpha_1^2 - 5\alpha_1\alpha_2 - 125 < 0 \quad \alpha_1 > 0 \quad \alpha_2 > 0. \quad (3.2.7)$$

Otherwise they are unstable. Hence, for a certain parameter space (α_1, α_2) there is a curve of Hopf bifurcation so that oscillations in the chemical concentrations are expected at one side of that curve. This Hopf bifurcation curve is the set of all parameters (α_1, α_2) , which *does not satisfy* the following inequality:

$$0 < \alpha_1 < \frac{5}{8} \left(\alpha_2 + \sqrt{\alpha_2^2 + 80} \right). \quad (3.2.8)$$

In a similar fashion, the fixed points for the BZ reaction scheme (for $\beta \neq 0$) are given as

$$c_{\text{HBrO}_2}^* = c_{\text{Br}^-}^* = c_{\text{Ce}^{4+}}^* = 0, \quad (3.2.9a)$$

$$c_{\text{HBrO}_2}^* = c_{\text{Ce}^{4+}}^* = \frac{(1 - \beta - \gamma) + \sqrt{(1 - \beta - \gamma)^2 + 4\beta(1 + \gamma)}}{2} \quad c_{\text{Br}^-}^* = \frac{\gamma c_{\text{HBrO}_2}^*}{c_{\text{HBrO}_2}^* + \beta}, \text{ and} \quad (3.2.9b)$$

$$c_{\text{HBrO}_2}^* = c_{\text{Ce}^{4+}}^* = \frac{(1 - \beta - \gamma) - \sqrt{(1 - \beta - \gamma)^2 + 4\beta(1 + \gamma)}}{2} \quad c_{\text{Br}^-}^* = \frac{\gamma c_{\text{HBrO}_2}^*}{c_{\text{HBrO}_2}^* + \beta}. \quad (3.2.9c)$$

Unlike the CDIMA scheme, BZ scheme has a set of three fixed points. Based on the *non-negativity of the chemical species* and *positivity of BZ chemical reaction parameters*, it is apparent that only *certain* set of fixed points are allowable. In order to delineate physics-compatible fixed points for the BZ scheme, various conditions on the combination of parameters β and γ are investigated. These are given as follows:

- For any value of $1 - \beta - \gamma$, the fixed point given by equation (3.2.9c) is not allowed as $c_{\text{HBrO}_2}^* = c_{\text{Ce}^{4+}}^* < 0$.

The characteristic polynomial of the BZ reaction scheme based on linearizing the

system of equations given by (3.2.2a)–(3.2.2c) near these fixed points is given as

$$\begin{aligned} & \lambda_{\text{BZ}}^3 + \left(1 + \frac{c_{\text{HBrO}_2}^* + \beta}{\epsilon_2} + \frac{2c_{\text{HBrO}_2}^* + c_{\text{Br}^-}^* - 1}{\epsilon_1} \right) \lambda_{\text{BZ}}^2 \\ & + \left(\frac{c_{\text{Br}^-}^* (\beta - c_{\text{HBrO}_2}^*)}{\epsilon_1 \epsilon_2} + \frac{c_{\text{HBrO}_2}^* + \beta}{\epsilon_2} + \frac{2c_{\text{HBrO}_2}^* + c_{\text{Br}^-}^* - 1}{\epsilon_1} \right) \lambda_{\text{BZ}} \\ & + \frac{(c_{\text{HBrO}_2}^* + \beta) (2c_{\text{HBrO}_2}^* + c_{\text{Br}^-}^* - 1) + (\beta - c_{\text{HBrO}_2}^*) (c_{\text{Br}^-}^* - \gamma)}{\epsilon_1 \epsilon_2} = 0, \end{aligned} \quad (3.2.10)$$

where $c_{\text{HBrO}_2}^*$ and $c_{\text{Br}^-}^*$ are the fixed points given by equations (3.2.9a)–(3.2.9c). Based on Routh-Hurwitz stability criterion for third-order polynomial, it is evident that the fixed points given by equation (3.2.9a)–(3.2.9c) are stable if the following conditions are satisfied:

$$1 + \frac{c_{\text{HBrO}_2}^* + \beta}{\epsilon_2} + \frac{2c_{\text{HBrO}_2}^* + c_{\text{Br}^-}^* - 1}{\epsilon_1} > 0, \quad (3.2.11a)$$

$$\frac{c_{\text{Br}^-}^* (\beta - c_{\text{HBrO}_2}^*)}{\epsilon_1 \epsilon_2} + \frac{c_{\text{HBrO}_2}^* + \beta}{\epsilon_2} + \frac{2c_{\text{HBrO}_2}^* + c_{\text{Br}^-}^* - 1}{\epsilon_1} > 0, \quad (3.2.11b)$$

$$\frac{(c_{\text{HBrO}_2}^* + \beta) (2c_{\text{HBrO}_2}^* + c_{\text{Br}^-}^* - 1) + (\beta - c_{\text{HBrO}_2}^*) (c_{\text{Br}^-}^* - \gamma)}{\epsilon_1 \epsilon_2} > 0, \text{ and} \quad (3.2.11c)$$

$$\begin{aligned} & \left(1 + \frac{c_{\text{HBrO}_2}^* + \beta}{\epsilon_2} + \frac{2c_{\text{HBrO}_2}^* + c_{\text{Br}^-}^* - 1}{\epsilon_1} \right) \times \\ & \left(\frac{c_{\text{Br}^-}^* (\beta - c_{\text{HBrO}_2}^*)}{\epsilon_1 \epsilon_2} + \frac{c_{\text{HBrO}_2}^* + \beta}{\epsilon_2} + \frac{2c_{\text{HBrO}_2}^* + c_{\text{Br}^-}^* - 1}{\epsilon_1} \right) > \\ & \frac{(c_{\text{HBrO}_2}^* + \beta) (2c_{\text{HBrO}_2}^* + c_{\text{Br}^-}^* - 1) + (\beta - c_{\text{HBrO}_2}^*) (c_{\text{Br}^-}^* - \gamma)}{\epsilon_1 \epsilon_2}. \end{aligned} \quad (3.2.11d)$$

From equations (3.2.11a)–(3.2.11d), it is clear that for a certain parameter space $(\beta, \gamma, \epsilon_1, \epsilon_2)$ there is a region of Hopf bifurcation so that oscillations in the chemical concentrations are expected towards the opposite side of that region. For instance, if the fixed point is given by equation (3.2.9a), the Hopf bifurcation region is independent of stoichiometric coefficient γ . It is characterized by the set of all parameters $(\beta, \epsilon_1, \epsilon_2)$, which *satisfies* the following inequality: $\beta \epsilon_1 < \epsilon_2$.

3.3 NUMERICAL FORMULATIONS: SPECIES BALANCE ERRORS

On coarse meshes, it is well-known in literature that the standard single-field Galerkin formulation exhibits spurious node-to-node oscillations for advection-dominated advection-diffusion-reaction equations Gresho and Sani (2000). This is due to the presence of characteristic layers such as interior and/or boundary layers in the solution (when advection processes is predominant than the diffusion and reaction processes). Hence, we resort to stabilized numerical formulations. *However, it should be noted that spurious node-to-node oscillations or numerical instabilities in the solution based on a certain numerical formulation should not be confused with the (physically meaningful) oscillatory behaviour of a chemically reacting system.* These two are entirely different.

Herein, we shall present two popular stabilized weak formulations for the initial boundary value problem given by equations (3.1.1a)–(6.4.25e). These are used to obtain numerical solutions for the concentration of the chemical species in Section 3.4. First one being the streamline upwind Petrov-Galerkin (SUPG) formulation and second one is the Galerkin least-squares (GLS) formulation. Before we present these two popular stabilized weak formulations, we will introduce the following function spaces:

$$\mathcal{C}_i^t := \left\{ c_i(\mathbf{x}, \bullet) \in H^1(\Omega) \mid c_i(\mathbf{x}, t) = c_i^p(\mathbf{x}, t) \text{ on } \Gamma_i^D \times]0, \mathcal{I}[\right\} \text{ and} \quad (3.3.1a)$$

$$\mathcal{W}_i := \left\{ w_i(\mathbf{x}) \in H^1(\Omega) \mid w_i(\mathbf{x}) = 0 \text{ on } \Gamma_i^D \right\}, \quad (3.3.1b)$$

where $H^1(\Omega)$ is a standard Sobolev space Evans (1998). For convenience, we shall denote the standard L_2 inner-product for a given two fields $a(\mathbf{x}, t)$ and $b(\mathbf{x}, t)$ over a

set \mathcal{D} as

$$(a; b)_{\mathcal{D}} = \int_{\mathcal{D}} a(\mathbf{x}) \bullet b(\mathbf{x}) \, d\mathcal{D}. \quad (3.3.2)$$

The subscript on the inner-product will be dropped if $\mathcal{D} = \Omega$

3.3.1 SUPG formulation

Find $c_i(\mathbf{x}, t) \in \mathcal{C}_i^t$ such that we have

$$\begin{aligned} & (w_i; \frac{\partial c_i}{\partial t}) - (\text{grad}[w_i] \bullet \mathbf{v}; c_i) + (\text{grad}[w_i]; \mathbf{D}(\mathbf{x}, t) \text{grad}[c_i]) \\ & + \sum_{e=1}^{Nele} \left(\tau_{\text{SUPG}} \mathbf{v} \bullet \text{grad}[w_i]; \frac{\partial c_i}{\partial t} + \text{div} [\mathbf{v}c_i - \mathbf{D}(\mathbf{x}, t) \text{grad}[c_i]] - f_i - r_i \right)_{\Omega_e} \\ & = (w_i; f_i) - (w_i; q_i^{\text{P}})_{\Gamma_i^{\text{N}}} \quad \forall w_i(\mathbf{x}) \in \mathcal{W}_i, \end{aligned} \quad (3.3.3)$$

where $\bar{\Omega} = \bigcup_{e=1}^{Nele} \bar{\Omega}^e$ and $Nele$ is the total number of mesh elements. The superposed bar denotes the set closure. The boundary of Ω^e is denoted as $\partial\Omega^e := \bar{\Omega}^e - \Omega^e$. τ_{SUPG} is the stabilization parameter within the context of the SUPG formulation. Herein, we shall use the stabilization parameter proposed by John and Knobloch John and Knobloch (2007), which is given as

$$\tau_{\text{SUPG}} = \frac{h_{\Omega_e}}{2\|\mathbf{v}\|} \xi_0(\text{Pe}_h), \quad \xi_0(\chi) = \coth(\chi) - \frac{1}{\chi}, \quad (3.3.4)$$

where h_{Ω_e} is the maximum element length, ξ_0 is known as the upwind function, and $\text{Pe}_h = \frac{h_{\Omega_e} \|\mathbf{v}\|}{2\lambda_{\min}}$ is the local (element) Péclet number. λ_{\min} is the minimum eigenvalue of the anisotropic diffusivity.

3.3.2 GLS formulation

Find $c_i(\mathbf{x}, t) \in \mathcal{C}_i^t$ such that we have

$$\begin{aligned}
& (w_i; \frac{\partial c_i}{\partial t}) - (\text{grad}[w_i] \bullet \mathbf{v}; c_i) + (\text{grad}[w_i]; \mathbf{D}(\mathbf{x}, t)\text{grad}[c_i]) \\
& + \sum_{e=1}^{N_{ele}} \left(\frac{w_i}{\Delta t} + \text{div}[\mathbf{v}w_i - \mathbf{D}(\mathbf{x}, t)\text{grad}[w_i]]; \tau_{\text{GLS}} \left(\frac{\partial c_i}{\partial t} + \text{div}[\mathbf{v}c_i - \mathbf{D}(\mathbf{x}, t)\text{grad}[c_i]] - f_i - r_i \right) \right)_{\Omega_e} \\
& = (w_i; f_i) - (w_i; q_i^{\text{p}})_{\Gamma_i^{\text{N}}} \quad \forall w_i(\mathbf{x}) \in \mathcal{W}_i, \tag{3.3.5}
\end{aligned}$$

where τ_{GLS} is the stabilization parameter under the GLS formulation and Δt is the time-step. Herein, we shall take $\tau_{\text{GLS}} = \tau_{\text{SUPG}}$, which is a prevailing practice. For these two stabilized finite element formulations, we shall now present the errors incurred in satisfying local species balance and global species balance.

3.3.3 Local and global species balance errors

Given $c_i(\mathbf{x}, t)$ and for any arbitrary $\Omega_e \in \Omega$, the error incurred in satisfying the local species balance of i^{th} chemical species is given as

$$\epsilon_{i, \Omega_e} = \int_{\Omega_e} \frac{\partial c_i}{\partial t} d\Omega_e + \int_{\partial\Omega_e} \mathbf{q}_i \bullet \hat{\mathbf{n}} d\Gamma_e - \int_{\Omega_e} (f_i + r_i) d\Omega_e, \tag{3.3.6}$$

where $\mathbf{q}_i = \mathbf{v}c_i - \mathbf{D}(\mathbf{x}, t)\text{grad}[c_i]$ is the total flux of i^{th} chemical species. Corresponding, the error incurred in satisfying the global species balance of i^{th} chemical species for the entire Ω is given as

$$\epsilon_i = \sum_{e=1}^{N_{ele}} \epsilon_{i, \Omega_e}. \tag{3.3.7}$$

In the next section, we shall quantify ϵ_i . Furthermore, we study the influence of non-negativity of chemical species on the performance of SUPG and GLS numerical formulations with respect to species balance errors.

3.4 CHEMICALLY REACTING LID-DRIVEN CAVITY FLOWS

In this section, we shall present a benchmark problem Erturk (2009) to investigate the accuracy of SUPG and GLS for transport-controlled CDIMA and BZ reaction models. A pictorial description of the initial boundary value problem is shown in Figure 3.1. The stream function and the corresponding advection velocity field to model lid-driven cavity flows is given as follows Adrover et al. (2002):

$$\Psi(x, y) = \sin^2\left(\frac{\pi x}{L_x}\right) \sin\left(\frac{\pi y^2}{L_y^2}\right), \quad (3.4.1a)$$

$$v_x(x, y) = -\frac{\partial \Psi}{\partial y} = -\frac{2\pi y}{L_y^2} \sin^2\left(\frac{\pi x}{L_x}\right) \cos\left(\frac{\pi y^2}{L_y^2}\right), \text{ and} \quad (3.4.1b)$$

$$v_y(x, y) = \frac{\partial \Psi}{\partial x} = \frac{2\pi}{L_x} \sin\left(\frac{\pi x}{L_x}\right) \cos\left(\frac{\pi x}{L_x}\right) \sin\left(\frac{\pi y^2}{L_y^2}\right). \quad (3.4.1c)$$

Numerical simulations are performed for various sets of reaction scheme parameters are taken as follows Neufeld and H.-García (2010):

$$\alpha_1 = 1 \quad \alpha_2 = 10^{-1} \quad \gamma = 1 \quad \beta = 10^{-2} \quad \epsilon_1 = 4 \times 10^{-2} \quad \epsilon_2 = 8 \times 10^{-1} \text{ and} \quad (3.4.2a)$$

$$\alpha_1 = 10^{-3} \quad \alpha_2 = 10^{-3} \quad \gamma = 1 \quad \beta = 10^{-4} \quad \epsilon_1 = 1 \quad \epsilon_2 = 1. \quad (3.4.2b)$$

The total time of interest \mathcal{I} is taken to be equal to 3. The discrete non-linear system of equations are solved using backward Euler and Newton-Raphson method with tolerance being equal to 10^{-4} . The time-step is equal to 0.1. Four-node structured quadrilateral meshes are used for the numerical simulation. The mesh size is taken as $h = \frac{1}{160}$. Analysis is performed for different types of diffusivities, which are given

as follows:

$$D(\mathbf{x}) = 10^{-4}, \quad (3.4.3a)$$

$$D(\mathbf{x}) = 10^{-6}, \text{ and} \quad (3.4.3b)$$

$$\mathbf{D}(\mathbf{x}) = \kappa_0 \begin{pmatrix} (y + \kappa_1)^2 + \kappa_2(x + \kappa_1)^2 & -(1 - \kappa_2)(x + \kappa_1)(y + \kappa_1) \\ -(1 - \kappa_2)(x + \kappa_1)(y + \kappa_1) & \kappa_2(y + \kappa_1)^2 + (x + \kappa_1)^2 \end{pmatrix}. \quad (3.4.3c)$$

The parameters κ_0 , κ_1 , and κ_2 are taken as

$$\kappa_0 = 10^{-2}, \kappa_1 = 10^{-2}, \kappa_2 = 10^{-3} \text{ and} \quad (3.4.4a)$$

$$\kappa_0 = 10^{-4}, \kappa_1 = 10^{-2}, \kappa_2 = 10^{-3}. \quad (3.4.4b)$$

It should be noted that for these set of parameters and mesh size, the element Péclet number $\mathbb{P}e_h \gg 1$. However, the stabilized numerical formulations (such as SUPG and GLS) fail to prevent spurious node-to-node oscillations. Furthermore, in most of these cases they produce non-physical values for the concentration of the chemical species involved in CDIMA and BZ reaction schemes. The corresponding concentration profiles and local species balance errors are shown in Figures 3.2–3.8. The white region represents the area in which concentration has violated the non-negative constraint. For $c_{\text{ClO}_2^-}$ at $t = 1$, this negative value is as high as -0.16. Analysis is performed based on the parameter set given by equations (3.4.2a) and (3.4.3a).

The errors incurred in satisfying global species balance is quantified in Table 3.1 and Table 3.2. From these figures and tables, it is clear that violating non-negative constraint and having spurious node-to-node oscillations in the numerical solution has a profound impact on species balance errors. Moreover, the negative values obtained and the species balance errors quantified are not close to machine precision, which is $\epsilon_{\text{mach}} \approx 2.22 \times 10^{-16}$ for a 64-bit machine.

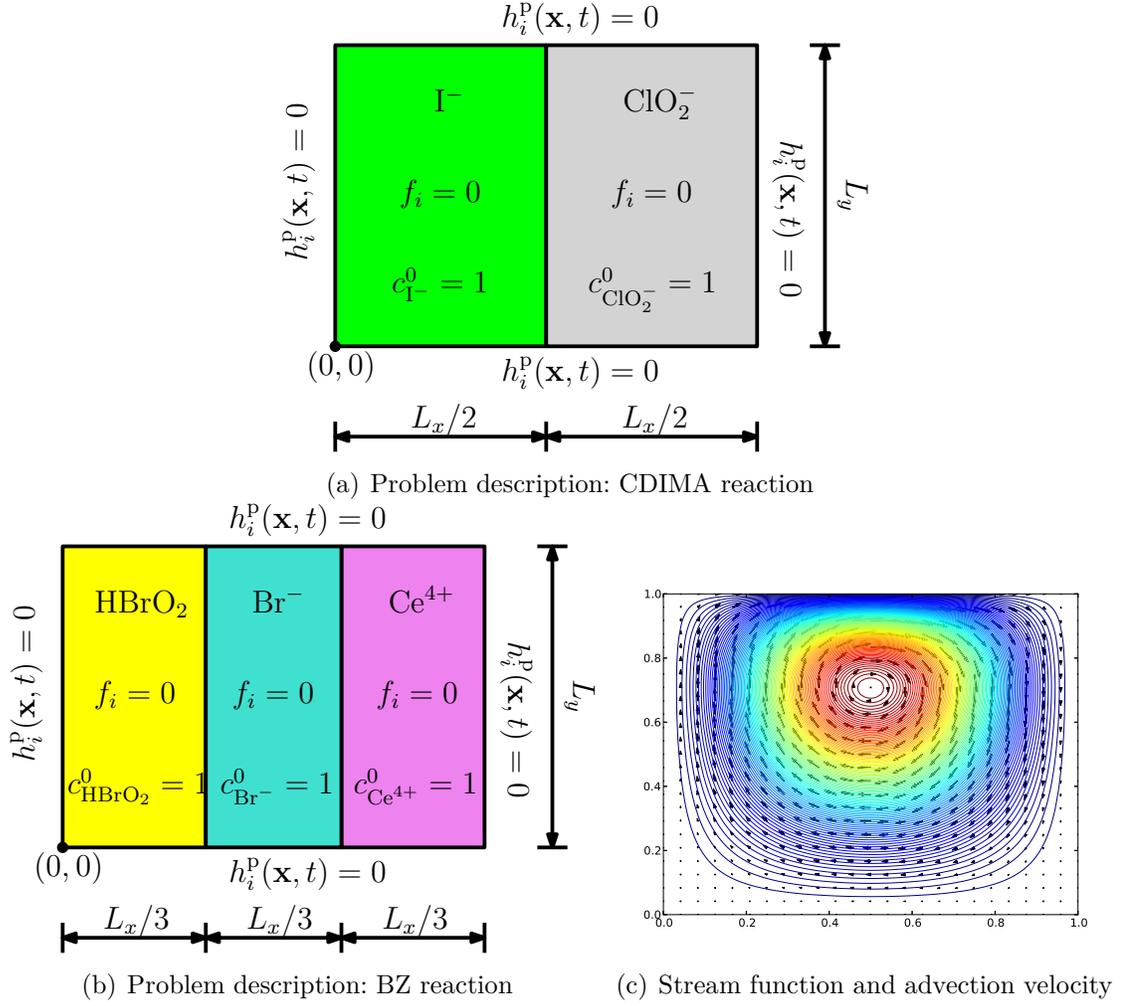


Figure 3.1: Transport-controlled chemically reacting lid-driven cavity flows: A pictorial description of the initial boundary value problem for CDIMA and BZ reaction schemes. Herein, $L_x = L_y = 1$ and non-reactive volumetric sources are zero.

Table 3.1: Global species balance error for CDIMA reaction scheme for various h -refined meshes at $t = 1$. Analysis is performed for anisotropic diffusivity. Negative values for concentration of chemical species are clipped.

Mesh	$ \epsilon_{I^-} $	$ \epsilon_{ClO_2^-} $
21×21	2.58×10^{-4}	5.22×10^{-2}
41×41	4.76×10^{-4}	5.20×10^{-2}
81×81	2.76×10^{-5}	4.83×10^{-2}
161×161	7.79×10^{-5}	4.28×10^{-2}

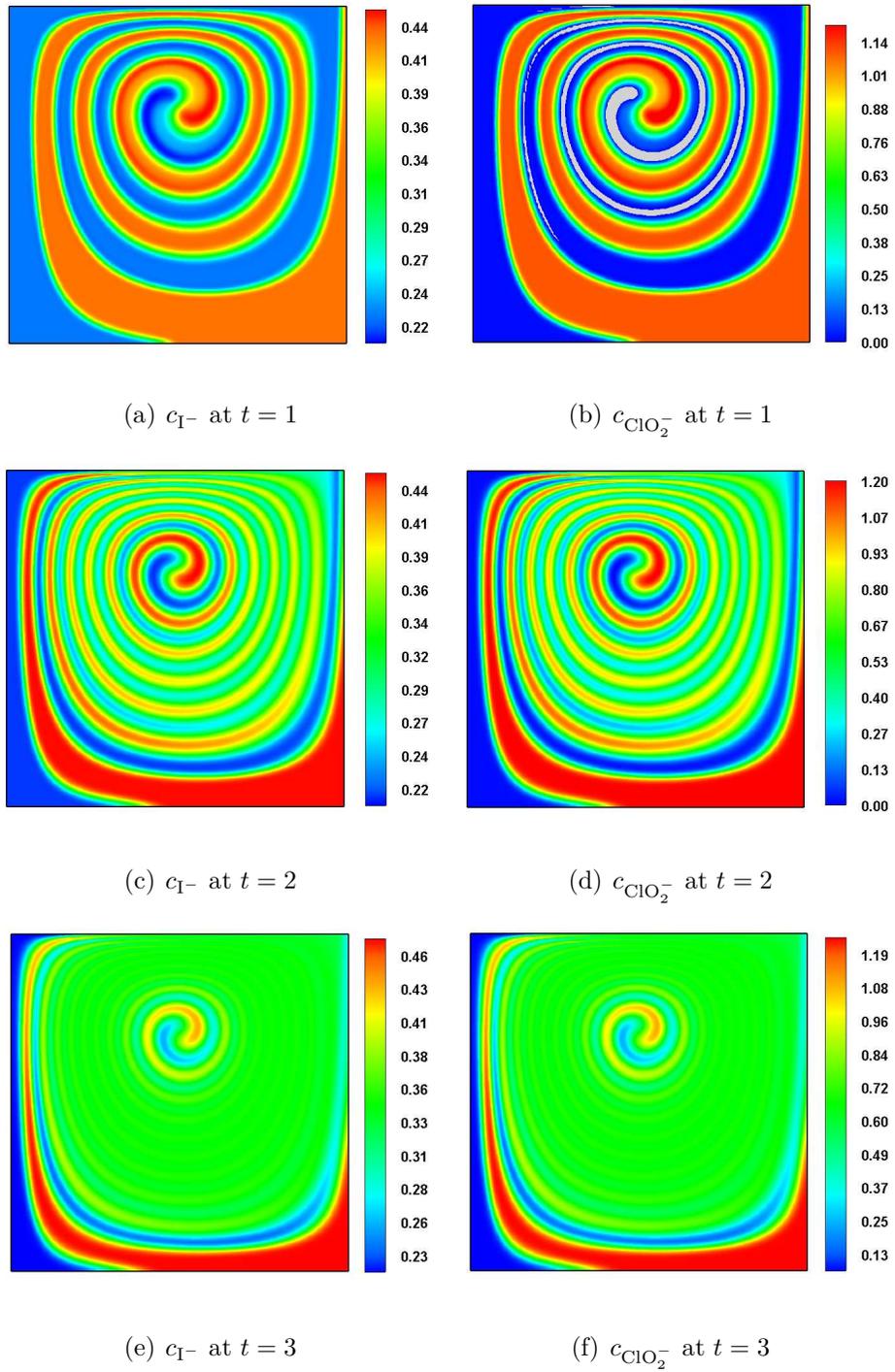


Figure 3.2: Transport-controlled CDIMA reaction (Isotropic diffusivity): This figure shows the concentration profiles of I^- and ClO_2^- ions at various time levels.

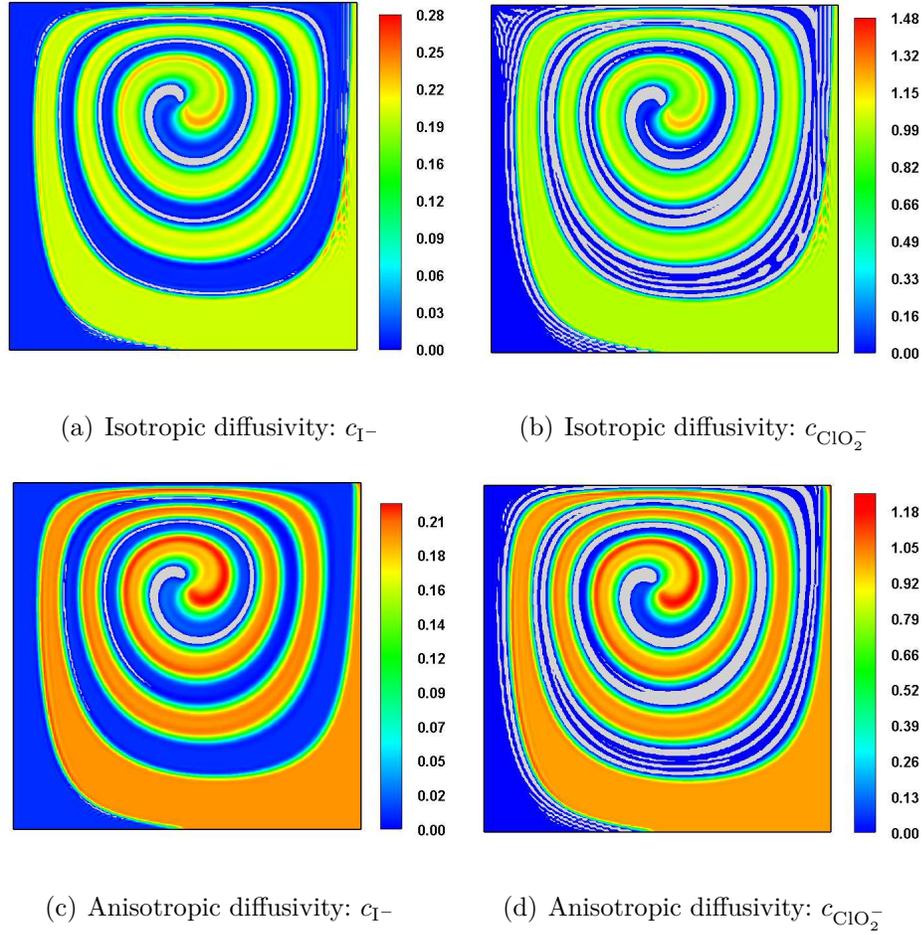


Figure 3.3: Transport-controlled CDIMA reaction: This figure shows the concentration profiles of I^- and ClO_2^- ions at $t = 1$. We also see spurious node-to-node oscillations and negative values are as high as -0.46.

Table 3.2: Global species balance error for BZ reaction scheme for various h -refined meshes at $t = 1$. Analysis is performed for anisotropic diffusivity. Negative values are clipped.

Mesh	$ \epsilon_{HBrO_2} $	$ \epsilon_{Br^-} $	$ \epsilon_{Ce^{4+}} $
21×21	6.68×10^{-3}	3.82	8.33×10^{-3}
41×41	6.25×10^{-3}	3.80	9.83×10^{-3}
81×81	4.52×10^{-3}	3.71	1.02×10^{-3}
161×161	2.63×10^{-3}	3.64	1.00×10^{-3}

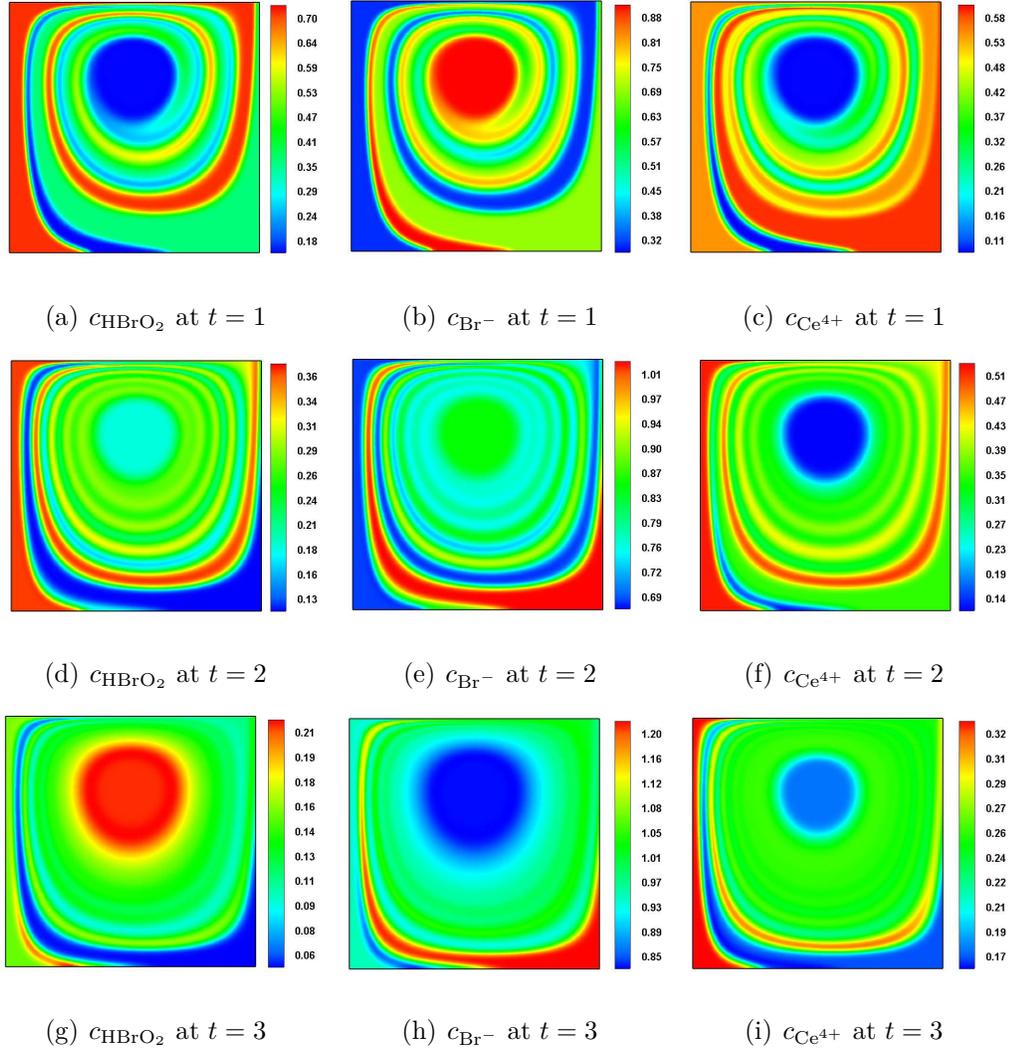


Figure 3.4: Transport-controlled BZ reaction (Isotropic diffusivity): This figure shows the concentration profiles of HBrO_2 , Br^- , and Ce^{4+} at various times. Analysis is performed based on the parameters given by equations (3.4.2a) and (3.4.3a).

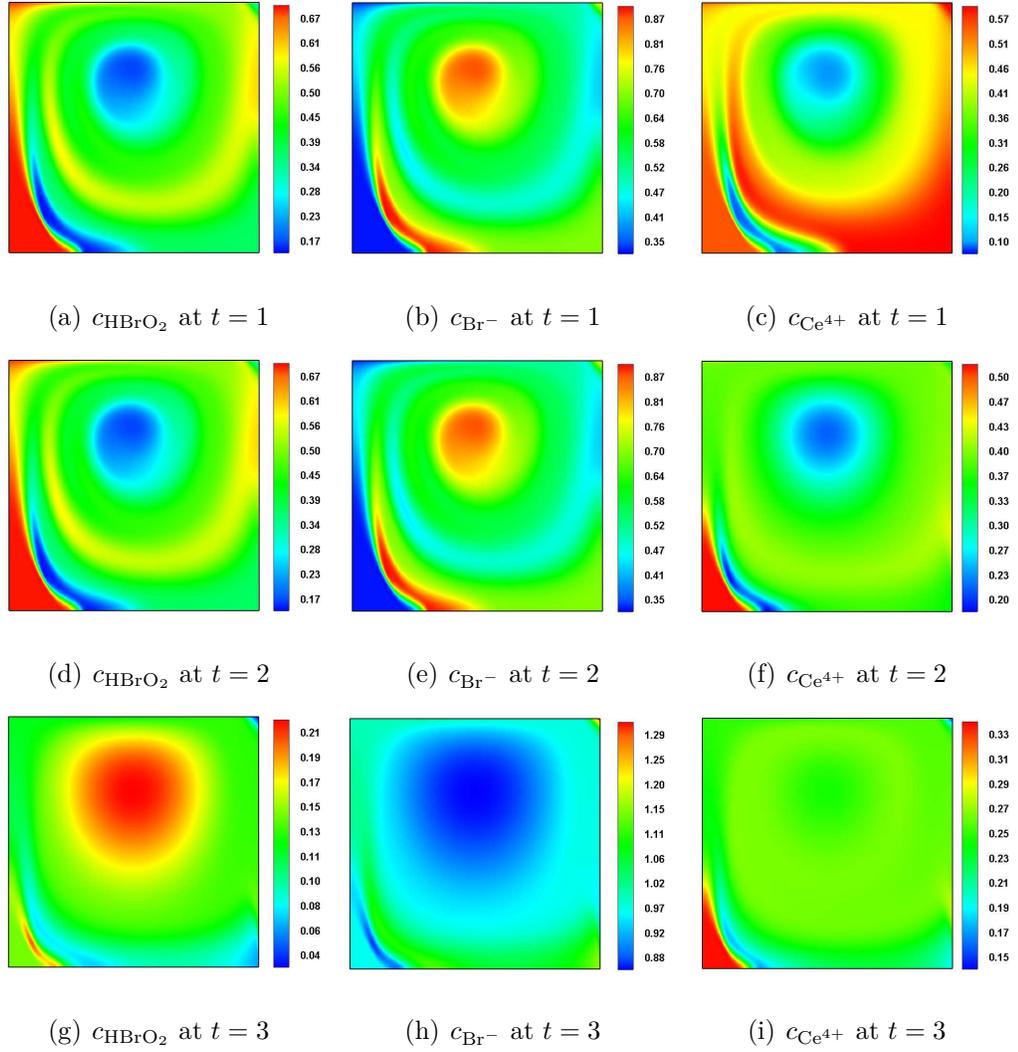


Figure 3.5: Transport-controlled BZ reaction (Anisotropic diffusivity): This figure shows the concentration profiles of HBrO_2 , Br^- , and Ce^{4+} at various times. Analysis is performed based on the parameters given by (3.4.2a), (3.4.3c), and (3.4.4a).

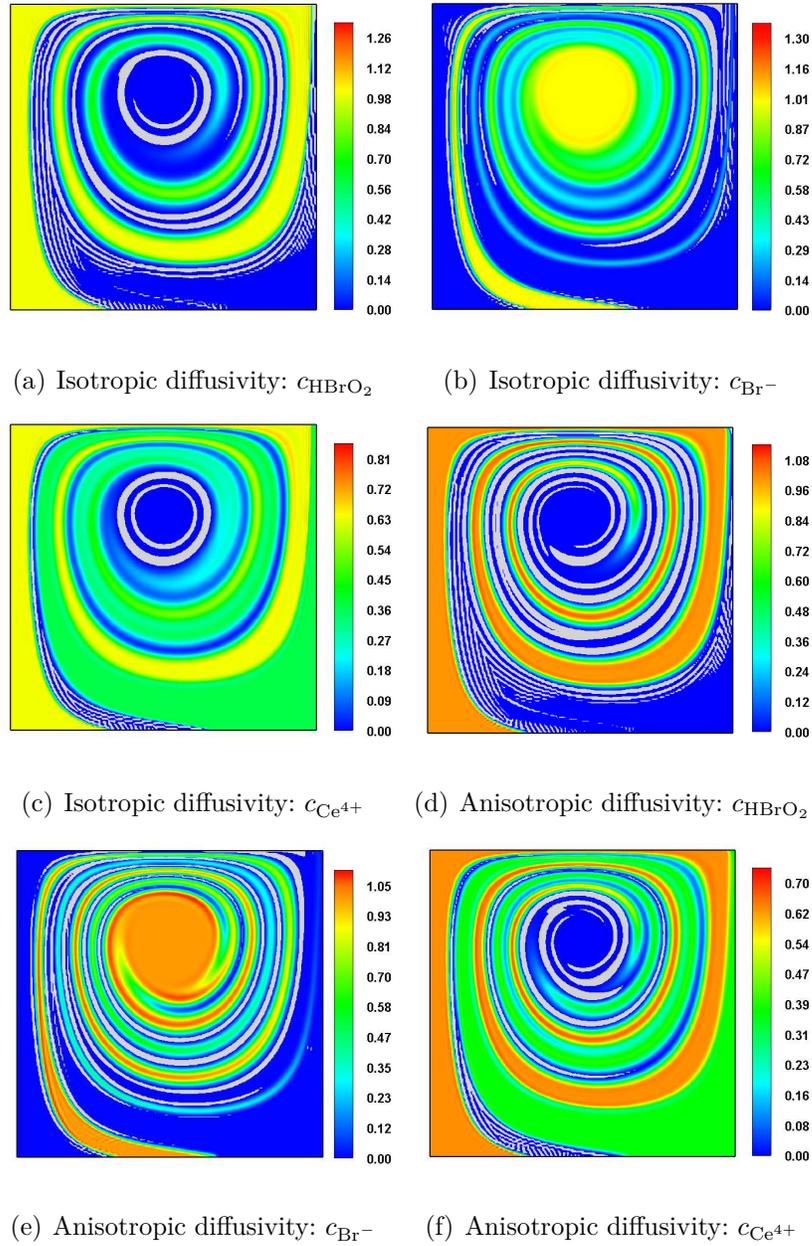


Figure 3.6: Transport-controlled BZ reaction: This figure shows the concentration profiles of HBrO_2 , Br^- , and Ce^{4+} at $t = 1$. We see spurious node-to-node oscillations in various parts of the domain and the negative values are as high as -0.35 .

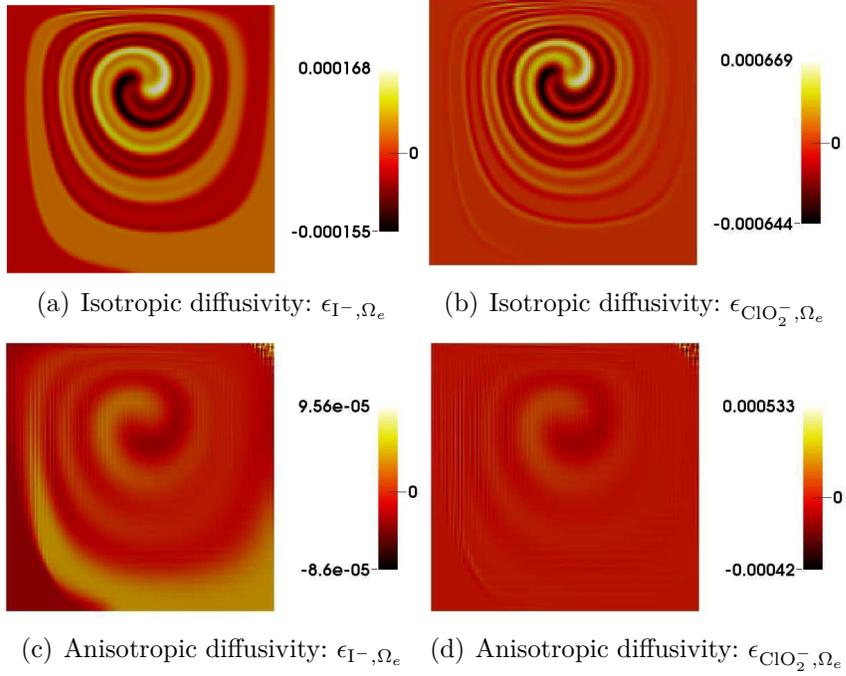


Figure 3.7: Transport-controlled CDIMA reaction: This figure shows the local species balance errors for I^- and ClO_2^- ions at $t = 1$. Analysis is performed based on the parameter set given by equations (3.4.2a), (3.4.3a), (3.4.3c), and (3.4.4a).

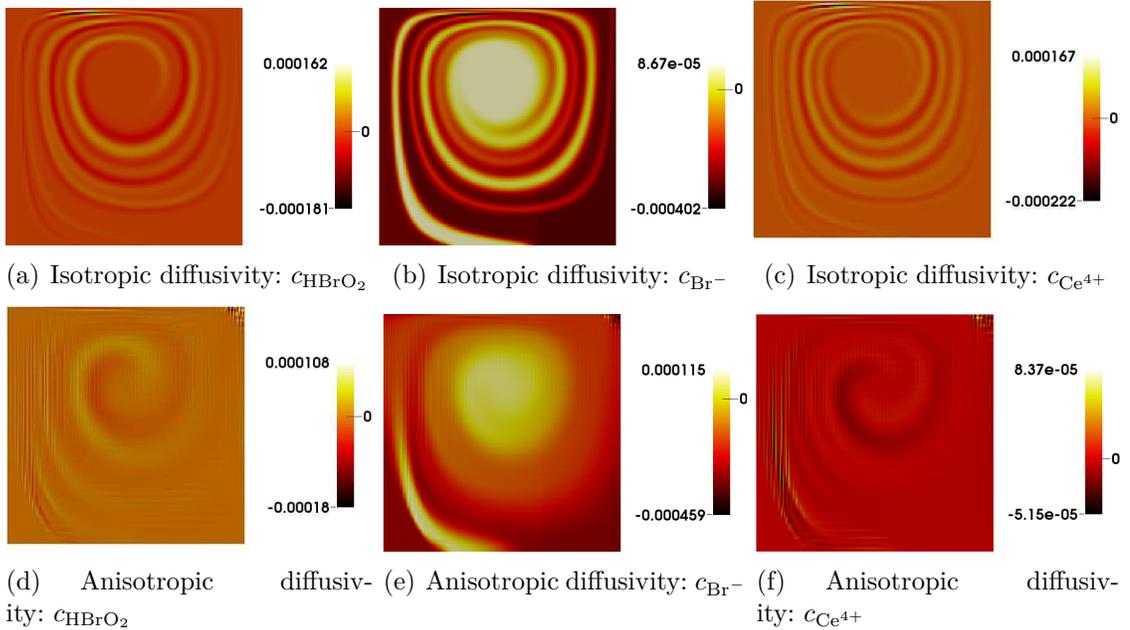


Figure 3.8: Transport-controlled BZ reaction: This figure shows the local species balance errors for $HBrO_2$, Br^- , and Ce^{4+} at $t = 1$. Analysis is performed based on the parameter set given by equations (3.4.2a), (3.4.3a), (3.4.3c), and (3.4.4a).

3.5 SUMMARY AND CONCLUSIONS

In this chapter, we have shown that not satisfying local and global species balance and meeting maximum principles and non-negative constraint has dire consequences in transport-controlled chemical reactions. We have quantified the errors incurred in satisfying species balance errors for CDIMA and BZ reaction schemes using stabilized numerical formulations such as SUPG and GLS. First, we have written the governing equations for the advective-diffusive-reactive systems based on the framework offered by the theory of interacting continua. Qualitative nature of the solution for CDIMA and BZ reaction schemes is presented using non-linear dynamical systems approach. Next, we presented a non-linear stabilized finite element formulations to obtain numerical solutions for our transport-controlled CDIMA and BZ advective-diffusive-reactive systems. The resulting non-linear system of equations are solved using the Newton-Raphson method. Lastly, we have performed numerical simulation to quantify species balance errors using chemically reacting lid-driven cavity flow benchmark problem. *Based on these numerical simulations, we have observed that unphysical values for concentration of chemical species due to violation of non-negative constraint and spurious node-to-node oscillations can result in large errors in local and global species balance.* Hence, this shows the need and importance of developing locally conservative non-negative numerical formulations for advection-dominated and reaction-dominated advective-diffusive-reactive systems.

Chapter 4

ON MESH RESTRICTIONS TO SATISFY COMPARISON PRINCIPLES, MAXIMUM PRINCIPLES, AND THE NON-NEGATIVE CONSTRAINT: RECENT DEVELOPMENTS AND NEW RESULTS

“It is an error to believe that rigor is the enemy of simplicity. On the contrary, we find it confirmed by numerous examples that the rigorous method is at the same time simpler and more easily comprehended. The very effort for rigor forces us to find out simpler methods of proof.”

David Hilbert

This chapter concerns with mesh restrictions that are needed to satisfy several important mathematical properties – maximum principles, comparison principles, and

the non-negative constraint – for a general linear second-order elliptic partial differential equation. We critically review some recent developments in the field of discrete maximum principles, derive new results, and discuss some possible future research directions in this area. In particular, we derive restrictions for a three-node triangular (T3) element and a four-node quadrilateral (Q4) element to satisfy comparison principles, maximum principles, and the non-negative constraint under the standard single-field Galerkin formulation. Analysis is restricted to uniformly elliptic linear differential operators in divergence form with Dirichlet boundary conditions specified on the entire boundary of the domain. Various versions of maximum principles and comparison principles are discussed in both continuous and discrete settings. In the literature, it is well-known that an acute-angled triangle is sufficient to satisfy the discrete weak maximum principle for pure *isotropic* diffusion. Herein, we show that this condition can be either too restrictive or not sufficient to satisfy various discrete principles when one considers anisotropic diffusivity, advection velocity field, or linear reaction coefficient. Subsequently, we derive appropriate restrictions on the mesh for simplicial (e.g., T3 element) and non-simplicial (e.g., Q4 element) elements. Based on these conditions, an iterative algorithm is developed to construct simplicial meshes that preserves discrete maximum principles using existing open source mesh generators. Various numerical examples based on different types of triangulations are presented to show the pros and cons of placing restrictions on a computational mesh. We also quantify local and global mass conservation errors using representative numerical examples and illustrate the performance of metric-based meshes with respect to mass conservation.

4.1 INTRODUCTION AND MOTIVATION

Diffusion-type equations are commonly encountered in various branches of engineering, sciences, and even in economics Crank (1975); Mei (2000); Aoki (2004).

These equations have been well-studied in Applied Mathematics, and several properties and *a priori* estimates have been derived Pao (1993). Numerous numerical formulations have been proposed and their performance has been analyzed both theoretically and numerically Gresho and Sani (2000). Several sophisticated software packages, such as ABAQUS Abaqus (2014), ANSYS Ansys (2015), COMSOL Comsol (2014), and MATLAB's PDE Toolbox Mat (2015), have been developed to solve these types of equations. Special solvers for solving the resulting discrete equations have also been proposed and studied adequately Gresho and Sani (2000).

It should, however, be noted that a numerical solution always loses some mathematical properties that the exact solution possesses. In particular, the aforementioned software packages and popular numerical formulations do not satisfy the so-called discrete comparison principles (DCPs), discrete maximum principles (DMPs), and the non-negative constraint (NC). This chapter is concerned with numerical solutions for anisotropic advection-diffusion-reaction equations.

To provide a motivation for the present work, we solve a pure anisotropic diffusion equation in an L-shaped domain with multiple holes using the commercial software package ABAQUS Abaqus (2014). Numerical simulations are performed using various unstructured finite element meshes (see Figure 4.1, which is to the scale), and using the following popular anisotropic diffusivity tensor from hydrogeological and subsurface flow literature Nakshatrala et al. (2013):

$$\mathbf{D}(\mathbf{x}) = \mathbf{R}\mathbf{D}_{\text{eigen}}\mathbf{R}^T, \quad (4.1.1)$$

where $\mathbf{D}_{\text{eigen}}$ is a diagonal matrix comprised of the eigenvalues of $\mathbf{D}(\mathbf{x})$. The corresponding principal eigenvectors are the column entries in the orthogonal matrix \mathbf{R} .

The expressions for $\mathbf{D}_{\text{eigen}}$ and \mathbf{R} are assumed as

$$\mathbf{R} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \quad \mathbf{D}_{\text{eigen}} = \begin{pmatrix} d_{\text{max}} & 0 \\ 0 & d_{\text{min}} \end{pmatrix} \quad \text{and} \quad (4.1.2a)$$

$$d_{\text{max}} = 10^3 \quad d_{\text{min}} = 1 \quad \theta = \pi/3 \quad \mathbf{v}(\mathbf{x}) = \mathbf{0} \quad \alpha(\mathbf{x}) = 0 \quad f(\mathbf{x}) = 0. \quad (4.1.2b)$$

Herein, d_{max} and d_{min} correspond to the maximum and minimum eigenvalues. θ corresponds to the angle of orientation of the eigenvector coordinate system. It should be noted that these eigenvalues have physical significance and are related to the *transverse and longitudinal diffusivities* in the eigenvector coordinate system. Diffusion process is simulated based on equations (5.2.1a)–(5.2.1b). The prescribed concentration on the sides of the L-shaped domain is equal to zero. Correspondingly, the concentration on the perimeter of the holes are set to be equal to one. Very fine triangular (where the total number of nodes and mesh elements are equal to 86326 and 169453) and quadrilateral (where the total number of nodes and mesh elements are equal to 91778 and 90625) meshes are used to perform ABAQUS numerical simulations.

The concentration profile obtained using ABAQUS numerical simulations is shown in Figure 4.2. In this figure, we have not shown the concentration contour using four-node quadrilateral mesh, as it is almost identical to that of the contour obtained by employing three-node triangular mesh. The white area within the L-shaped domain with multiple holes represents the regions in which the obtained numerical nodal concentrations are *negative* and at the same time exceeded the *maximum* value provided by the maximum principle. Quantitatively, more than 6% of the nodes have unphysical negative values for the concentration. To be precise, in the case of triangular mesh, 2.22% and 3.92% of the nodes have violated the non-negative and maximum constraints. Correspondingly, the minimum and maximum values for concentration

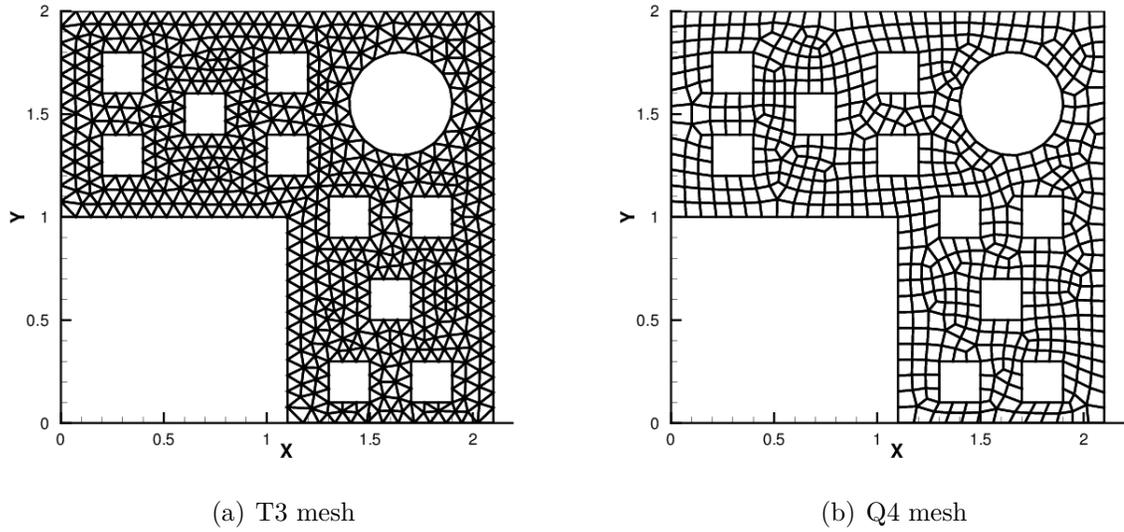


Figure 4.1: ABAQUS unstructured meshes for an L-shaped domain with multiple holes: The left and right figures show an instance of three-node triangular and four-node quadrilateral meshes employed in the numerical simulation of a pure anisotropic diffusion problem using ABAQUS.

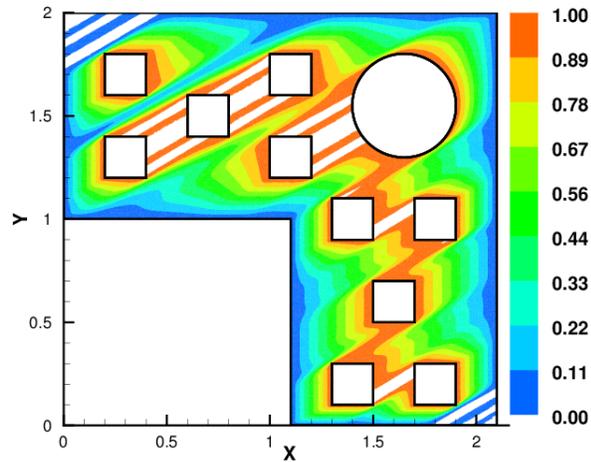
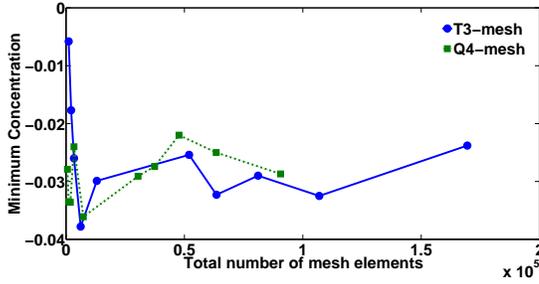
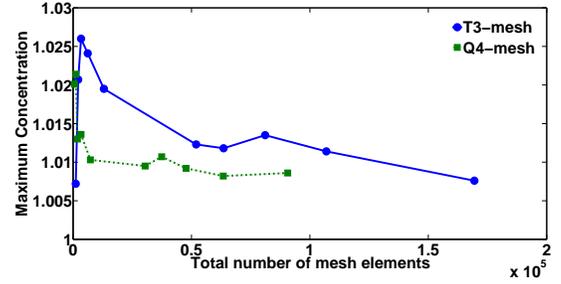


Figure 4.2: ABAQUS numerical simulation for an L-shaped domain with multiple holes: The contours of concentration obtained using ABAQUS are based on three-node triangular mesh.

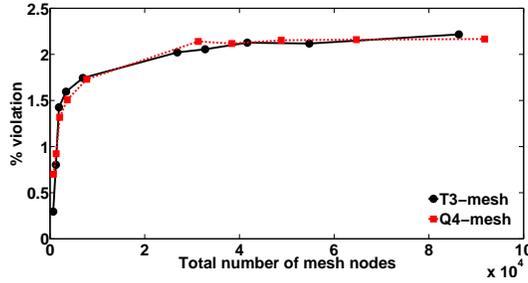


(a) Non-negative constraint violation

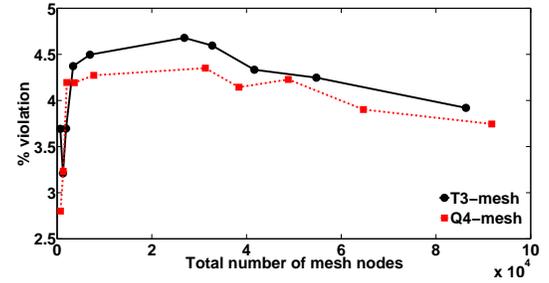


(b) Maximum constraint violation

Figure 4.3: Minimum and maximum values for concentration in an L-shaped domain with multiple holes: The left and right figures show the minimum and maximum values attained in the computational domain.



(a) % violation: Non-negative constraint



(b) % violation: Maximum constraint

Figure 4.4: Percentage of violation in minimum and maximum constraints for concentration in an L-shaped domain with multiple holes: The left and right figures show the percentage of nodes that have violated the constraints.

obtained are -0.0238 and 1.0076 . These are considerably far away from the possible values, which are between 0 and 1 . Similarly, for quadrilateral mesh, these values are slightly lower. Quantitatively, these are around 2.17% and 3.75% . But the minimum and maximum values of concentration (-0.0287 and 1.0086) are slightly higher than that of the triangular mesh. Figures 4.3 and 4.4 show that these negative values do *not* decrease with mesh refinement. There are three possible routes to overcome such limitations and satisfy DCPs, DMPs, and NC; which we shall describe below.

4.1.1 Strategy I: Mesh restrictions

The first strategy is to place restrictions on the mesh to meet maximum principles and the non-negative constraint. For isotropic homogeneous diffusivity, Ciarlet and

Raviart Ciarlet and Raviart (1973) have shown that numerical solutions based on the single-field Galerkin finite element formulation, in general, does not converge *uniformly*. However, the single-field Galerkin formulation is a converging scheme. Ciarlet and Raviart have also shown that a sufficient condition for single-field Galerkin formulation to converge uniformly for pure *isotropic* diffusion is to employ a well-centered three-node triangular element mesh with low-order interpolation.

The obvious advantage is that one can use the single-field Galerkin formulation without any modification. The drawback is that an appropriate computational mesh may not exist because of the required restrictions on the shape and size of the finite element. For example, it is not an easy task (sometimes it is *not* possible) to generate a well-centered triangular mesh for any given two-dimensional domain Vanderzee et al. (2010). Note that requiring a mesh to be well-centered is a more stringent than requiring the mesh to be Delaunay. In fact, a well-centered mesh is Delaunay but the converse need not be true.

In scientific literature, there are numerous commercial and non-commercial mesh generators that produce premium quality structured and unstructured meshes for various complicated domains. For instance, the survey paper by Owen Owen (1998) accounts for more than 70 unstructured mesh generation software products. But, it needs to be emphasized that Owen Owen (1998) rarely mentions about non-obtuse, acute, and anisotropic \mathcal{M} -uniform mesh generators. However, it is evident from the above discussion that these types of meshes have a profound impact on solving various important physical problems related to diffusion-type equations. In recent years, there has been considerable effort in developing such types of mesh generators. For example, some open source meshing software packages which are relevant to mesh restrictions methodology are

- Non-obtuse and acute triangulations in 2D: aCute Erten and Üngör (2007, 2009a,b) (a meshing software, which is based on Triangle Shewchuk (1996))

- Anisotropic \mathcal{M} -uniform triangulations in 2D: BAMG Hecht (2006) in FreeFem++ Hecht et al. (2014); Hecht (2012), BL2D Laug and Borouchaki (1996)
- Anisotropic \mathcal{M} -uniform triangulations in 3D: Mmg3d Dobrzynski (2012)
- Locally uniform anisotropic Delaunay meshes (surface, 2D, and 3D): CGALmesh Alliez et al. (2003); Boissonnat et al. (2008, 2009, 2011)

However, the use of these mesh generators in the area of numerical analysis and engineering, in particular, to construct mesh restrictions for diffusion-type equations to satisfy DCPs, DMPs, and NC is hardly known. Recently, Huang and co-workers Li and Huang (2010); Lu et al. (2012); Huang (2013) used BAMG to generate anisotropic simplicial meshes to satisfy various discrete properties for linear advection-diffusion-reaction equations. But in their research works, the computational domains under consideration are not complicated. In addition, they did not study the role played by mesh restrictions on meeting local and global species balance.

4.1.2 Strategy II: Non-negativity, monotone, and monotonicity preserving formulations

The second strategy is mainly concerned with developing new numerical formulations based on physical and variational principles as to satisfy DCPs, DMPs, and NC. These formulations can be broadly classified into three categories:

- *Non-negative formulations*: A numerical formulation is said to be non-negative if the resulting numerical solution satisfies certain DMPs and NC. The formulation need not satisfy DCPs or stronger versions of DMPs.
- *Monotone formulations*: A numerical formulation is said to be monotone if the resulting numerical solution satisfies certain DMPs, DCPs, and NC. The formulation is not required to satisfy stronger versions of any given discrete principle. In addition, the solution may contain spurious oscillations.

- *Monotonicity preserving formulations:* A numerical formulation is said to be monotonicity preserving if the resulting numerical solution does not exhibit spurious oscillations within itself. There is no restriction on the numerical solution to satisfy DMPs and DCPs.

A non-negative formulation need not satisfy monotone conditions, a monotone numerical formulation need not be monotonicity preserving, and vice-versa. It is still an open research problem to develop a numerical formulation that meets all the aforementioned properties. Some notable research works, which take Strategy II, are references Ciarlet (1970a); Varga (1966) for finite difference schemes (FDS), Brezzi et al. (2005); Lipnikov et al. (2011) for mimetic finite difference methods (MFDM), Potier (2009); Nordbotten et al. (2007); Droniou and Potier (2011) for finite volume methods (FVM), and Burman and Ern (2005); Drăgănescu et al. (2005); Liska and Shashkov (2008); Nakshatrala and Valocchi (2009) for finite element methods (FEM). It needs to be emphasized that most of these techniques involve non-linear solution procedures. For example, the optimization-based finite element formulations proposed in Nakshatrala and Valocchi (2009); Nagarajan and Nakshatrala (2011), Nakshatrala et al. (2013), Nakshatrala et al. (2013) enforce the desired properties as explicit constraints under variationally consistent constrained minimization problems. This, of course, comes at an expense of additional computational cost.

4.1.3 Strategy III: Post-processing methods

The third strategy is to employ a post-processing (PP) method to recover various discrete properties. In the literature, there exists various types of PP methods for diffusion-type equations. Some notable research works in this direction include:

- Local and global remapping/repair methods Kucharik et al. (2003); Garimella et al. (2007)

- Constrained monotonic regression based methods Burdakov et al. (2012)
- Cutoff methods (also known as the clipping methods) Kreuzer (2014); Lu et al. (2013)
- A combination of remapping/repair methods and cutoff methods Wang et al. (2012); Zhao et al. (2013)

We now briefly describe the pros and cons of these methods. Note that it is difficult to apply these techniques to recover DCPs, DMPs, and NC for higher-order FEM methods, as the shape functions can change their sign within the element. In addition, most of the above methods do *not* have a variational basis.

The remapping/repair techniques proposed by Shashkov and co-workers are designed to improve the quality of numerical solutions by satisfying certain mathematical properties in the discrete setting. Even though these are efficient, conservative, and linearity- and bound-preserving interpolation algorithms, they are mesh-dependent. Moreover, a systematic application of such algorithms for anisotropic advection-diffusion-reaction equations to satisfy DCPs, DMPs, and NC has not been done yet Wang et al. (2012); Zhao et al. (2013).

The post-processing procedure proposed by Burdakov *et al.* Burdakov et al. (2012) is based on a constrained monotonic regression problem. The procedure is a locally conservative, bound-preserving, monotonicity-recovering, and constrained optimization-based PP method. It is applicable to FDS, FVM, and FEM. But in the case of FEM, this PP method is valid only for linear and multi-linear shape functions. In order to construct appropriate constraints for the optimization problem, one needs to know *a priori* information on the lower bounds, upper bounds, and monotonicity of the numerical solution for a given physical problem. In general, obtaining the qualitative and quantitative nature of the solution is not always possible. If such information on the monotonicity, and lower and upper bounds for the numerical

solution is not known *a priori*, then this method reduces to the standard clipping procedure. In addition, one should note that it is not always possible to satisfy DCPs using this constrained monotonic regression algorithm. For example, one can construct a counterexample similar to the one presented in (Nakshatrala et al., 2013, Section 4)) to show that it does not satisfy DCP.

Finally, we would like to emphasize that *a posteriori cutoff method* is a variational crime. In general, this approach is neither conservative nor satisfies DMPs and DCPs. The primary objective is to chop-off the values of a numerical solution if it is less than a given number. In the case of highly anisotropic diffusion problems and for distorted meshes, this method predicts erroneous numerical results Nagarajan and Nakshatrala (2011); Nakshatrala et al. (2013). By specifying the cutoff value to be zero, it is always guaranteed to satisfy NC through this methodology. In addition, if the nature of the solution is known *a priori*, then one can also prevent undershooting and overshooting of the numerical solution by chopping off those values.

4.1.4 Main contributions and an outline of this chapter

Herein, we focus on the first approach of placing restrictions on the computational mesh to meet desired mathematical properties. We derive sufficient conditions on the restrictions to be placed on the three-node triangular and four-node quadrilateral finite elements to meet comparison principles, maximum principles, and the non-negative constraint in the case of heterogeneous anisotropic advection-diffusion-reaction (ADR) equations. The notable contributions of this chapter are:

- (i) We provide an in-depth review of various versions of comparison principles, maximum principles, and the non-negative constraint in the continuous setting.
- (ii) We derive necessary and sufficient conditions on the coefficient (i.e., the "stiffness") matrix to satisfy discrete weak and strong comparison principles.

- (iii) A relationship between various discrete principles within the context of mesh restrictions, numerical formulations, and post-processing methods is presented.
- (iv) We propose an iterative method to generate simplicial meshes that satisfy discrete properties using open source mesh generators such as **BAMG** Hecht (2006), **FreeFem++** Hecht et al. (2014); Hecht (2012), and **Gmsh** Geuzaine and Remacle (2015).
- (v) Different types of non-dimensional quantities are proposed for anisotropic diffusivity, which are variants of the standard Péclet and Damköhler numbers. These quantities are extremely useful in numerical simulations and have not been discussed in the literature.
- (vi) Lastly, several realistic numerical examples are presented to corroborate the theoretical findings as well as to show the importance of preserving discrete principles.

The remainder of this chapter is organized as follows. In Section 4.2, we present the governing equations for a general linear second-order elliptic equation and discuss associated mathematical principles: comparison principles, maximum principles, and the non-negative constraint. Section 4.3 provides several important remarks on the continuous and discrete properties of elliptic equations. In Section 4.4, we shall derive mesh restrictions for the three-node triangular element and the rectangular element to meet the discrete versions of maximum principles, comparison principles, and the non-negative constraint. Finally, conclusions are drawn in Section 4.5.

We will denote scalars by lower case English alphabet or lower case Greek alphabet (e.g., concentration c and density ρ). We will make a distinction between vectors in the continuum and finite element settings. Similarly, a distinction will be made between second-order tensors in the continuum setting versus matrices in the discrete setting. The continuum vectors are denoted by lower case boldface normal

letters, and the second-order tensors will be denoted using upper case boldface normal letters (e.g., 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 (e.g., vector \mathbf{f} and matrix \mathbf{K}). Other notational conventions are introduced as needed.

4.2 LINEAR SECOND-ORDER ELLIPTIC EQUATION AND ASSOCIATED MATHEMATICAL PRINCIPLES

Let $\Omega \subset \mathbb{R}^{nd}$ be a open bounded domain, where “ nd ” denotes the number of spatial dimensions. The boundary of the domain is denoted by $\partial\Omega$, which is assumed to be piecewise smooth. Mathematically, $\partial\Omega := \overline{\Omega} - \Omega$, where a superposed bar denotes the set closure. A spatial point is denoted by $\mathbf{x} \in \overline{\Omega}$. The gradient and divergence operators with respect to \mathbf{x} are, respectively, denoted by $\text{grad}[\bullet]$ and $\text{div}[\bullet]$. Let $c(\mathbf{x})$ denote the concentration field. We assume that Dirichlet boundary condition (i.e., the concentration) is prescribed on the entire boundary. The rest of this chapters deals with the following boundary value problem, which is written in *divergence form*:

$$\mathcal{L}[c] := -\text{div}[\mathbf{D}(\mathbf{x})\text{grad}[c(\mathbf{x})]] + \mathbf{v}(\mathbf{x}) \bullet \text{grad}[c(\mathbf{x})] + \alpha(\mathbf{x})c(\mathbf{x}) = f(\mathbf{x}) \quad \text{in } \Omega \text{ and} \quad (4.2.1a)$$

$$c(\mathbf{x}) = c^p(\mathbf{x}) \quad \text{on } \partial\Omega, \quad (4.2.1b)$$

where \mathcal{L} denotes the second-order linear differential operator, $f(\mathbf{x})$ is the prescribed volumetric source, $\alpha(\mathbf{x})$ is the linear reaction coefficient, $\mathbf{v}(\mathbf{x})$ is the velocity vector field, $\mathbf{D}(\mathbf{x})$ is the anisotropic diffusivity tensor, and $c^p(\mathbf{x})$ is the prescribed concentration. Physics of the problem demands that the diffusivity tensor (which is a

second-order tensor) be symmetric:

$$\mathbf{D}^T(\mathbf{x}) = \mathbf{D}(\mathbf{x}) \quad \forall \mathbf{x} \in \Omega. \quad (4.2.2)$$

Remark 4.2.1. *In mathematical analysis, the divergence form is a suitable setting for the application of energy methods. However, some studies on maximum principles do employ the following non-divergence form:*

$$\mathcal{L}[c] = \sum_{i,j=1}^{nd} (\mathbf{P})_{ij} \frac{\partial^2 c}{\partial x_i \partial x_j} + \sum_{i=1}^{nd} (\mathbf{q})_i \frac{\partial c}{\partial x_i} + r(\mathbf{x})c, \quad (4.2.3)$$

where the coefficient $(\mathbf{P})_{ij}$, $(\mathbf{q})_i$, and $r(\mathbf{x})$, can be related to the physical quantities such as the diffusivity tensor, velocity field, and linear reaction coefficient. It should however be noted that the non-divergence form exists irrespective of differentiability of the diffusivity tensor. If $\mathbf{D}(\mathbf{x})$ is continuously differentiable, then there exists a one-to-one correspondence between the divergence form and the non-divergence form. In such cases, the operator \mathcal{L} in the divergence form given by equation (5.2.1a) can be put into the following non-divergence form:

$$\mathcal{L}[c] = -\mathbf{D}(\mathbf{x}) \bullet \text{grad} [\text{grad}[c(\mathbf{x})]] + (\mathbf{v}(\mathbf{x}) - \text{div} [\mathbf{D}(\mathbf{x})]) \bullet \text{grad}[c(\mathbf{x})] + \alpha(\mathbf{x})c(\mathbf{x}). \quad (4.2.4)$$

Based on the nature of the coefficients and connectedness of the physical domain, different versions of maximum and comparison principles exist in the mathematical literature Evans (1998); Gilbarg and Trudinger (2001). As stated earlier, we shall restrict our study to Dirichlet boundary conditions on the entire boundary. Analysis pertaining to Neumann and mixed boundary conditions in the context of maximum principles, comparison principles, and the non-negative constraint is beyond the scope of this chapter, and one can consult references Karátson and Korotov (2005); Borsuk

and Kondratiev (2006); Pao (1993).

We shall say that the operator \mathcal{L} is *elliptic* at a point $\mathbf{x} \in \Omega$ if

$$0 < \lambda_{\min}(\mathbf{x})\boldsymbol{\xi} \bullet \boldsymbol{\xi} \leq \boldsymbol{\xi} \bullet \mathbf{D}(\mathbf{x})\boldsymbol{\xi} \leq \lambda_{\max}(\mathbf{x})\boldsymbol{\xi} \bullet \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in \mathbb{R}^{nd} \setminus \{\mathbf{0}\}, \quad (4.2.5)$$

where $\lambda_{\min}(\mathbf{x})$ and $\lambda_{\max}(\mathbf{x})$ are, respectively, the minimum and maximum eigenvalues of $\mathbf{D}(\mathbf{x})$. The operator \mathcal{L} is said to be *strictly elliptic* if there exists a constant λ_0 such that

$$0 < \lambda_0 \leq \lambda_{\min}(\mathbf{x}) \quad \forall \mathbf{x} \in \Omega \quad (4.2.6)$$

and *uniformly elliptic* if

$$0 < \frac{\lambda_{\max}(\mathbf{x})}{\lambda_{\min}(\mathbf{x})} < +\infty \quad \forall \mathbf{x} \in \Omega. \quad (4.2.7)$$

In the studies on maximum principles, it is common to impose the following restrictions on the velocity field $\mathbf{v}(\mathbf{x})$ and the linear reaction coefficient $\alpha(\mathbf{x})$:

$$\alpha(\mathbf{x}) \geq 0 \quad \forall \mathbf{x} \in \Omega, \quad (4.2.8a)$$

$$\alpha(\mathbf{x}) - \frac{1}{2} \operatorname{div} [\mathbf{v}(\mathbf{x})] \geq 0 \quad \forall \mathbf{x} \in \Omega, \text{ and} \quad (4.2.8b)$$

$$0 \leq \frac{|(\mathbf{v}(\mathbf{x}))_i|}{\lambda_{\min}(\mathbf{x})} \leq \beta_0 < +\infty \quad \forall \mathbf{x} \in \Omega \quad \text{and} \quad \forall i = 1, \dots, nd, \quad (4.2.8c)$$

where β_0 is a bounded non-negative constant. If $(\mathbf{D})_{ij}$ and $(\mathbf{v})_i$ are continuous in Ω , then the operator \mathcal{L} is uniformly elliptic for any bounded subdomain $\Omega' \subset\subset \Omega$ (which means that Ω' is *compactly embedded* in Ω) and the condition given in equation (4.2.8c) holds. The restrictions given in equation (4.2.8b) can be relaxed in some situations (e.g., see references Lu et al. (2012); Huang (2013)). But the constraint on $\alpha(\mathbf{x})$ given by equation (4.2.8a) cannot be relaxed. If $\alpha(\mathbf{x}) < 0$, then equation (5.2.1a) is referred

to as an Helmholtz-type equation, which does not possess a maximum principle. From the theory of partial differential equations, it is well-known that the aforementioned boundary value problem given by equations (5.2.1a)–(5.2.1b) satisfies the so-called (weak and strong) comparison principles, (weak and strong) maximum principles, and the non-negative constraint. For future reference and for completeness, we shall briefly outline the main results. For a more detailed mathematical treatment, one could consult references Evans (1998); Pao (1993); Gilbarg and Trudinger (2001).

Theorem 4.2.2 (Continuous weak and strict weak maximum principles). *Let \mathcal{L} be a uniformly elliptic operator satisfying the conditions given by equations (4.2.8a)–(4.2.8c). In addition, let $\mathbf{D}(\mathbf{x})$ be continuously differentiable. Suppose that $c(\mathbf{x}) \in C^2(\Omega) \cap C^0(\bar{\Omega})$ satisfies the differential inequality $\mathcal{L}[c] \leq 0$ in Ω , then the maximum of $c(\mathbf{x})$ in $\bar{\Omega}$ is obtained on $\partial\Omega$. That is, $c(\mathbf{x})$ possesses the weak maximum principle (wMP), which can be written as*

$$\max_{\mathbf{x} \in \bar{\Omega}} [c(\mathbf{x})] \leq \max \left[0, \max_{\mathbf{x} \in \partial\Omega} [c(\mathbf{x})] \right]. \quad (4.2.9)$$

Moreover, if $\alpha(\mathbf{x}) = 0$, then we have the strict weak maximum principle (WMP):

$$\max_{\mathbf{x} \in \bar{\Omega}} [c(\mathbf{x})] = \max_{\mathbf{x} \in \partial\Omega} [c(\mathbf{x})]. \quad (4.2.10)$$

Theorem 4.2.3 (Continuous strong and strict strong maximum principles). *Let the domain Ω be simply connected. Given that $c(\mathbf{x})$ satisfies wMP and the conditions given in Theorem 6.2.1, then $c(\mathbf{x})$ cannot attain an interior non-negative maximum in $\bar{\Omega}$ unless it is a constant. This means that, $c(\mathbf{x})$ possesses the strong maximum principle (sMP) if the following hold:*

$$\max_{\mathbf{x} \in \bar{\Omega}} [c(\mathbf{x})] = \max_{\mathbf{x} \in \bar{\Omega}} [c(\mathbf{x})] = m \geq 0 \quad \Rightarrow \quad c(\mathbf{x}) \equiv m \quad \text{in } \bar{\Omega}. \quad (4.2.11)$$

Moreover, if $\alpha(\mathbf{x}) = 0$ and $c(\mathbf{x})$ satisfies WMP, then we have the strict strong maximum principle (SMP) given as

$$\max_{\mathbf{x} \in \Omega} [c(\mathbf{x})] = \max_{\mathbf{x} \in \overline{\Omega}} [c(\mathbf{x})] = m \quad \Rightarrow \quad c(\mathbf{x}) \equiv m \quad \text{in } \overline{\Omega}. \quad (4.2.12)$$

Theorem 4.2.4 (Continuous weak and strong comparison principles). *Let $c_1(\mathbf{x})$ and $c_2(\mathbf{x}) \in C^2(\Omega) \cap C^0(\overline{\Omega})$. Suppose \mathcal{L} be a uniformly elliptic operator satisfying the conditions given by the equations (4.2.8a)–(4.2.8c). Then \mathcal{L} is said to possess*

- *the weak comparison principle (wCP) if $c_1(\mathbf{x})$ and $c_2(\mathbf{x})$ satisfies wMP, $\mathcal{L}[c_1] \leq \mathcal{L}[c_2]$ in Ω , and $c_1(\mathbf{x}) \leq c_2(\mathbf{x})$ on $\partial\Omega$, then the following holds:*

$$c_1(\mathbf{x}) \leq c_2(\mathbf{x}) \quad \forall \mathbf{x} \in \overline{\Omega}. \quad (4.2.13)$$

- *the strong comparison principle (sCP) if $c_1(\mathbf{x})$ and $c_2(\mathbf{x})$ satisfies sMP, $\mathcal{L}[c_1] < \mathcal{L}[c_2]$ in Ω , and $c_1(\mathbf{x}) \leq c_2(\mathbf{x})$ on $\partial\Omega$, then the following holds:*

$$c_1(\mathbf{x}) < c_2(\mathbf{x}) \quad \forall \mathbf{x} \in \Omega. \quad (4.2.14)$$

For mathematical proofs to Theorems 6.2.1–4.2.4, see reference Gilbarg and Trudinger (2001). Numerical formulations based on the finite element method, finite volume method, and finite difference method exist to solve the boundary value problem (5.2.1a)–(5.2.1b). It is well-known that the framework offered by the finite element method is particularly attractive in obtaining accurate numerical results for elliptic partial differential equations. In particular, the single-field Galerkin formulation is a very popular finite element formulation. In this chapter, we shall use the single-field Galerkin formulation to derive mesh restrictions. It should be, however, noted that restrictions imposed on a mesh may alter if an alternate numerical formulation is employed. But the overall procedure presented in this chapter can be

employed to derive mesh restrictions for other numerical formulations.

4.2.1 Single-field Galerkin formulation

Let us define the following function spaces:

$$\mathcal{C} := \left\{ c(\mathbf{x}) \in H^1(\Omega) \mid c(\mathbf{x}) = c^p(\mathbf{x}) \text{ on } \partial\Omega \right\} \text{ and} \quad (4.2.15a)$$

$$\mathcal{W} := \left\{ w(\mathbf{x}) \in H^1(\Omega) \mid w(\mathbf{x}) = 0 \text{ on } \partial\Omega, \right\} \quad (4.2.15b)$$

where $H^1(\Omega)$ is a standard Sobolev space Evans (1998). Given two fields $a(\mathbf{x})$ and $b(\mathbf{x})$ on a set \mathcal{D} , the standard L_2 inner-product over \mathcal{D} will be denoted as

$$(a; b)_{\mathcal{D}} = \int_{\mathcal{D}} a(\mathbf{x}) \bullet b(\mathbf{x}) \, d\mathcal{D}. \quad (4.2.16)$$

The subscript on the inner-product will be dropped if $\mathcal{D} = \Omega$. The single-field Galerkin formulation for the boundary value problem (5.2.1a)–(5.2.1b) can be written as: Find $c(\mathbf{x}) \in \mathcal{C}$ such that we have

$$\mathcal{B}(w; c) = L(w) \quad \forall w(\mathbf{x}) \in \mathcal{W}, \quad (4.2.17)$$

where the bilinear form and the linear functional are, respectively, defined as

$$\mathcal{B}(w; c) := (w; \alpha(\mathbf{x})c) + (w; \mathbf{v}(\mathbf{x}) \bullet \text{grad}[c]) + (\text{grad}[w]; \mathbf{D}(\mathbf{x})\text{grad}[c]) \text{ and} \quad (4.2.18a)$$

$$L(w) := (w; f(\mathbf{x})). \quad (4.2.18b)$$

4.2.2 Discrete single-field Galerkin formulation

Let the computational domain Ω be decomposed into “*Nele*” non-overlapping open sub-domains, which in the finite element context will be elements. That is,

$$\bar{\Omega} = \bigcup_{e=1}^{Nele} \bar{\Omega}^e. \quad (4.2.19)$$

The boundary of Ω^e is denoted as $\partial\Omega^e := \bar{\Omega}^e - \Omega^e$. Let $\mathbb{P}^1(\Omega^e)$ denote the vector space spanned by linear polynomials on the sub-domain Ω^e . We shall define the following finite dimensional subsets of \mathcal{C} and \mathcal{W} :

$$\mathcal{C}^h := \left\{ c^h(\mathbf{x}) \in \mathcal{C} \mid c^h(\mathbf{x}) \in C^0(\bar{\Omega}); c^h(\mathbf{x})|_{\Omega^e} \in \mathbb{P}^1(\Omega^e); e = 1, \dots, Nele \right\} \text{ and} \quad (4.2.20a)$$

$$\mathcal{W}^h := \left\{ w^h(\mathbf{x}) \in \mathcal{W} \mid w^h(\mathbf{x}) \in C^0(\bar{\Omega}); w^h(\mathbf{x})|_{\Omega^e} \in \mathbb{P}^1(\Omega^e); e = 1, \dots, Nele \right\}. \quad (4.2.20b)$$

A corresponding finite element formulation can be written as follows: Find $c^h(\mathbf{x}) \in \mathcal{C}^h$, such that we have

$$\mathcal{B}(w^h; c^h) = L(w^h) \quad \forall w^h(\mathbf{x}) \in \mathcal{W}^h, \quad (4.2.21)$$

where $\mathcal{B}(w^h; c^h)$ and $L(w^h)$ are, respectively, given as

$$\mathcal{B}(w^h; c^h) := (w^h; \alpha(\mathbf{x})c^h) + (w^h; \mathbf{v}(\mathbf{x}) \bullet \text{grad}[c^h]) + (\text{grad}[w^h]; \mathbf{D}(\mathbf{x})\text{grad}[c^h]) \text{ and} \quad (4.2.22a)$$

$$L(w^h) := (w^h; f(\mathbf{x})). \quad (4.2.22b)$$

Let “ n_t ” denote the total number of degrees-of-freedom, “ n_f ” denote the free degrees-of-freedom, and “ n_p ” be the prescribed degrees-of-freedom for the concentration vector. Obviously, we have $n_t = n_f + n_p$. We assume that $n_t, n_p \geq 2$. After

finite element discretization, the discrete equations for the boundary value problem take the following form:

$$\mathbf{K}\mathbf{c} = \mathbf{r}, \quad (4.2.23)$$

where $\mathbf{K} \equiv [\mathbf{K}_{ff} | \mathbf{K}_{fp}]$ is the stiffness matrix, $\mathbf{c} \equiv [\mathbf{c}_f^T | \mathbf{c}_p^T]^T$ is the vector containing nodal concentration, and $\mathbf{r} = [\mathbf{r}_f]^T$ is the corresponding nodal volumetric source vector. The stiffness matrices \mathbf{K} , \mathbf{K}_{ff} , and \mathbf{K}_{fp} are, respectively, of size $n_f \times n_t$, $n_f \times n_f$, and $n_f \times n_p$. Correspondingly, the nodal concentration vectors \mathbf{c} , \mathbf{c}_f , and \mathbf{c}_p are of sizes $n_t \times 1$, $n_f \times 1$, and $n_p \times 1$. Similar inference is applicable to the load vector \mathbf{r} .

Before we state a discrete version of (weak and strong) maximum and comparison principles, we introduce the required notation. The symbols \preceq and \succeq shall denote component-wise inequalities for vectors and matrices. That is, given two (finite dimensional) vectors \mathbf{a} and \mathbf{b}

$$\mathbf{a} \preceq \mathbf{b} \quad \text{means that} \quad a_i \leq b_i \quad \forall i. \quad (4.2.24)$$

Correspondingly, given two matrices \mathbf{A} and \mathbf{B}

$$\mathbf{A} \preceq \mathbf{B} \quad \text{means that} \quad (\mathbf{A})_{ij} \leq (\mathbf{B})_{ij} \quad \forall i, j. \quad (4.2.25)$$

Similarly, one can define the symbol \succeq , \prec , and \succ . In the remainder of this chapter, we will be frequently using the symbols $\mathbf{0}$ and \mathbf{O} , which, respectively, denote a zero vector and a zero matrix.

We shall now briefly outline the main results corresponding to the discrete weak and strong maximum principles in the form of definitions and theorems. Using these results, we shall discuss in detail about discrete comparison principles. However,

it should be noted that Theorem 4.2.8 and its proof are new and have not been discussed elsewhere. We shall also present the necessary and sufficient conditions on the stiffness matrices \mathbf{K}_{ff} and \mathbf{K}_{fp} to satisfy different versions of discrete maximum principles and the non-negative constraint. For more details, see references Ishihara (1987); Mincsovcics and Hórvath (2012).

Definition 4.2.5 (Discrete maximum principles Mincsovcics and Hórvath (2012)).

A numerical formulation is said to possess

- *the discrete weak maximum principle (DwMP) if*

$$\mathbf{r} \preceq \mathbf{0} \quad \text{implies} \quad \max[\mathbf{c}] \leq \max[0, \max[\mathbf{c}_p]] \quad (4.2.26)$$

- *the discrete strict weak maximum principle (DWMP) if*

$$\mathbf{r} \preceq \mathbf{0} \quad \text{implies} \quad \max[\mathbf{c}] = \max[\mathbf{c}_p] \quad (4.2.27)$$

- *the discrete strong maximum principle (DsMP) if it possesses DwMP and satisfies the following condition:*

$$\mathbf{r} \preceq \mathbf{0}, \quad \text{and} \quad \max[\mathbf{c}] = \max[\mathbf{c}_f] = m \geq 0 \quad \text{implies} \quad \mathbf{c} = m\mathbf{1} \quad (4.2.28)$$

- *the discrete strict strong maximum principle (DSMP) if it possesses DWMP and satisfies the following condition:*

$$\mathbf{r} \preceq \mathbf{0}, \quad \text{and} \quad \max[\mathbf{c}] = \max[\mathbf{c}_f] = m \quad \text{implies} \quad \mathbf{c} = m\mathbf{1} \quad (4.2.29)$$

where $\max[\bullet]$ denotes the maximal element of a vector and the symbol $\mathbf{1}$ is the vector whose components are all equal to 1.

Theorem 4.2.6 (Necessary and sufficient conditions to satisfy DMPs). *The stiffness matrix \mathbf{K} given by equation (4.2.23) is said to possess*

- *the discrete weak maximum principle (DwMP $_{\mathbf{K}}$) if and only if all of the following conditions are satisfied:*

$$(a) \mathbf{K}_{ff}^{-1} \succeq \mathbf{O} \quad (b) -\mathbf{K}_{ff}^{-1} \mathbf{K}_{fp} \succeq \mathbf{O} \quad (c) -\mathbf{K}_{ff}^{-1} \mathbf{K}_{fp} \mathbf{1} \preceq \mathbf{1} \quad (4.2.30)$$

- *the discrete strict weak maximum principle (DWMP $_{\mathbf{K}}$) if and only if all of the following conditions are satisfied:*

$$(a) \mathbf{K}_{ff}^{-1} \succ \mathbf{O} \quad (b) -\mathbf{K}_{ff}^{-1} \mathbf{K}_{fp} \succeq \mathbf{O} \quad (c) -\mathbf{K}_{ff}^{-1} \mathbf{K}_{fp} \mathbf{1} = \mathbf{1} \quad (4.2.31)$$

- *the discrete strong maximum principle (DsMP $_{\mathbf{K}}$) if and only if all of the following conditions are satisfied:*

$$(a) \mathbf{K}_{ff}^{-1} \succ \mathbf{O} \quad (b) -\mathbf{K}_{ff}^{-1} \mathbf{K}_{fp} \succ \mathbf{O} \quad (c) -\mathbf{K}_{ff}^{-1} \mathbf{K}_{fp} \mathbf{1} \prec \mathbf{1} \quad \text{or} \quad -\mathbf{K}_{ff}^{-1} \mathbf{K}_{fp} \mathbf{1} = \mathbf{1} \quad (4.2.32)$$

- *the discrete strict strong maximum principle (DSMP $_{\mathbf{K}}$) if and only if all of the following conditions are satisfied:*

$$(a) \mathbf{K}_{ff}^{-1} \succ \mathbf{O} \quad (b) -\mathbf{K}_{ff}^{-1} \mathbf{K}_{fp} \succ \mathbf{O} \quad (c) -\mathbf{K}_{ff}^{-1} \mathbf{K}_{fp} \mathbf{1} = \mathbf{1} \quad (4.2.33)$$

Proof. For a proof, see reference Mincsovcics and Hórvath (2012). □

Definition 4.2.7 (Discrete weak and strong comparison principles). *A numerical formulation is said to possess*

- the discrete weak comparison principle (DwCP) if it satisfies DwMP, and

$$\mathbf{c}_1 \preceq \mathbf{c}_2 \text{ on } \partial\Omega \quad \text{and} \quad \mathbf{r}_1 \preceq \mathbf{r}_2 \text{ in } \Omega \quad \text{implies} \quad \mathbf{c}_1 \preceq \mathbf{c}_2 \text{ in } \bar{\Omega} \quad (4.2.34)$$

- the discrete strong comparison principle (DsCP) if it satisfies DsMP, and

$$\mathbf{c}_1 \preceq \mathbf{c}_2 \text{ on } \partial\Omega \quad \text{and} \quad \mathbf{r}_1 \prec \mathbf{r}_2 \text{ in } \Omega \quad \text{implies} \quad \mathbf{c}_1 \prec \mathbf{c}_2 \text{ in } \Omega \quad (4.2.35)$$

Theorem 4.2.8 (Necessary and sufficient conditions to satisfy DCPs). *Let \mathbf{c}_1 and \mathbf{c}_2 be two nodal concentration vectors corresponding to the volumetric source vectors \mathbf{r}_1 and \mathbf{r}_2 based on the equation (4.2.23). If \mathbf{c}_1 and \mathbf{c}_2 satisfy DwMP and the hypothesis of DwCP (i.e., $\mathbf{c}_1 \preceq \mathbf{c}_2$ on $\partial\Omega$ and $\mathbf{r}_1 \preceq \mathbf{r}_2$ in Ω), then a necessary and sufficient condition to satisfy the discrete weak comparison principle (DwCP $_{\mathbf{K}}$) (which means that $\mathbf{c}_1 \preceq \mathbf{c}_2$ in $\bar{\Omega}$) is that the stiffness matrix \mathbf{K} possess DwMP $_{\mathbf{K}}$ (which is given by equation (6.3.9) in Theorem 4.2.6).*

If \mathbf{c}_1 and \mathbf{c}_2 satisfy DsMP and the hypothesis of DsCP, (i.e., $\mathbf{c}_1 \preceq \mathbf{c}_2$ on $\partial\Omega$ and $\mathbf{r}_1 \prec \mathbf{r}_2$ in Ω), then a necessary and sufficient condition to satisfy the discrete strong comparison principle (DsCP $_{\mathbf{K}}$) (which means that $\mathbf{c}_1 \prec \mathbf{c}_2$ in Ω) is that the stiffness matrix \mathbf{K} possess DsMP $_{\mathbf{K}}$ (which is given by equation (4.2.32) in Theorem 4.2.6).

Proof. For convenience, let us define the following

$$\mathbf{c}_3 := \mathbf{c}_1 - \mathbf{c}_2 \text{ and} \quad (4.2.36a)$$

$$\mathbf{r}_3 := \mathbf{r}_1 - \mathbf{r}_2. \quad (4.2.36b)$$

Clearly, \mathbf{c}_3 and \mathbf{r}_3 satisfy the following:

$$\mathbf{K}\mathbf{c}_3 = \mathbf{r}_3. \quad (4.2.37)$$

Necessary condition to satisfy DwCP \mathbf{K} : Let $\mathbf{c}_1 \preceq \mathbf{c}_2$ in $\overline{\Omega}$, which implies that $\mathbf{c}_3 \preceq \mathbf{0}$ in $\overline{\Omega}$. The hypothesis of DwCP \mathbf{K} and the fact that $\mathbf{c}_3 \preceq \mathbf{0}$ in $\overline{\Omega}$ imply the following:

$$\mathbf{r}_3 \preceq \mathbf{0} \quad \text{in } \Omega, \quad (4.2.38a)$$

$$\mathbf{c}_3 \preceq \mathbf{0} \quad \text{on } \partial\Omega, \quad (4.2.38b)$$

$$\max[\mathbf{c}_3] \leq 0 \quad \text{on } \partial\Omega, \quad \text{and} \quad (4.2.38c)$$

$$\max_{\overline{\Omega}}[\mathbf{c}_3] \leq \max\left[0, \max_{\partial\Omega}[\mathbf{c}_3]\right] = 0, \quad (4.2.38d)$$

which implies that \mathbf{c}_3 satisfies DwMP (based on equation (4.2.26) in Definition 4.2.5). But vector \mathbf{c}_3 also satisfies equation (4.2.37). Hence, according to equation (6.3.9) and the hypothesis of Theorem 4.2.6, it is evident that \mathbf{K} must possess DwMP \mathbf{K} . This completes the proof for the necessary condition to satisfy DwCP \mathbf{K} .

Sufficient condition to satisfy DwCP \mathbf{K} : It is given that \mathbf{c}_1 and \mathbf{c}_2 satisfy DwMP. Equations (4.2.36a)–(4.2.36b) and DwCP \mathbf{K} imply that

$$\mathbf{c}_3 \preceq \mathbf{0} \quad \text{on } \partial\Omega \quad \text{and} \quad (4.2.39a)$$

$$\mathbf{r}_3 \preceq \mathbf{0} \quad \text{in } \Omega. \quad (4.2.39b)$$

If the stiffness matrix \mathbf{K} possess DwMP \mathbf{K} , it is evident from Theorem 4.2.6 and equations (4.2.37), (4.2.39a)–(4.2.39b) that vector \mathbf{c}_3 satisfies DwMP. Hence, from Definition 4.2.5 and equations (4.2.26), (4.2.39a)–(4.2.39b), we have the following result:

$$\max_{\overline{\Omega}}[\mathbf{c}_3] \leq \max\left[0, \max_{\partial\Omega}[\mathbf{c}_3]\right] = 0. \quad (4.2.40)$$

From equation (4.2.40), it is evident that the least upper bound for any component of vector \mathbf{c}_3 is equal to zero. Hence, we have $\mathbf{c}_3 \preceq \mathbf{0}$ in $\overline{\Omega}$. This implies that $\mathbf{c}_1 \preceq \mathbf{c}_2$

on $\overline{\Omega}$, which completes the proof for the sufficient condition to satisfy DwCP $_{\mathbf{K}}$.

Necessary condition to satisfy DsCP $_{\mathbf{K}}$: Following the arguments about the proof for the necessary condition to satisfy the DwCP $_{\mathbf{K}}$ property, it is evident that \mathbf{c}_3 satisfies DwMP. In addition, we are given that $\mathbf{c}_1 \prec \mathbf{c}_2$ in Ω . This implies $\mathbf{c}_3 \prec \mathbf{0}$ in Ω . Based on the hypothesis of DsCP $_{\mathbf{K}}$ and utilizing the fact that $\mathbf{c}_3 \prec \mathbf{0}$ in Ω yields the following

$$\max [\mathbf{c}_3] \leq 0 \quad \text{on } \partial\Omega \text{ and} \quad (4.2.41a)$$

$$\max [\mathbf{c}_3] < 0 \quad \text{in } \Omega. \quad (4.2.41b)$$

This means that vector $\mathbf{0}$ is the least upper bound for \mathbf{c}_3 in $\overline{\Omega}$, and any component of \mathbf{c}_3 is *strictly* less than zero in the interior of the domain Ω . From equation (4.2.41a) and (4.2.41b), it is clear that the *non-negative* maximum value for vector \mathbf{c}_3 occurs on the boundary $\partial\Omega$. From equation (4.2.28) in Definition 4.2.5, it follows that vector \mathbf{c}_3 satisfies DsMP. Hence, according to conditions specified by equation (4.2.32) and the hypothesis of Theorem 4.2.6, it is evident that \mathbf{K} must possess the DsMP $_{\mathbf{K}}$ property. This completes the proof for the necessary condition to satisfy DsCP $_{\mathbf{K}}$.

Sufficient condition to satisfy DsCP $_{\mathbf{K}}$: Given that \mathbf{c}_1 and \mathbf{c}_2 satisfy DsMP. Under the assumptions of DsCP $_{\mathbf{K}}$ and from equations (4.2.36a)–(4.2.36b), we have the following relations:

$$\mathbf{r}_3 \prec \mathbf{0} \quad \text{in } \Omega \text{ and} \quad (4.2.42a)$$

$$\mathbf{c}_3 \preceq \mathbf{0} \quad \text{on } \partial\Omega. \quad (4.2.42b)$$

If the stiffness matrix \mathbf{K} possess DsMP $_{\mathbf{K}}$, it is evident from Theorem 4.2.6 and equations (4.2.37), (4.2.42a)–(4.2.42b) that vector \mathbf{c}_3 satisfies DsMP. Hence, by appealing to Definition 4.2.5 and equation (4.2.28), if vector \mathbf{c}_3 *does not* attain a non-negative

maximum value at an interior point of Ω , then we have the following result:

$$\max_{\Omega} [\mathbf{c}_3] < \max \left[0, \max_{\partial\Omega} [\mathbf{c}_3] \right] = 0, \quad (4.2.43)$$

which implies that each component of vector \mathbf{c}_3 is less than zero. Hence, we have $\mathbf{c}_1 \prec \mathbf{c}_2$ in Ω . Suppose, if vector \mathbf{c}_3 attains a non-negative maximum value at an interior point of Ω , then according to the surmise of DsMP, we first need to satisfy DwMP. So from equation (4.2.26), we have the following relation:

$$\max_{\Omega} [\mathbf{c}_3] \leq \max \left[0, \max_{\partial\Omega} [\mathbf{c}_3] \right] = 0. \quad (4.2.44)$$

Secondly, according to DsMP, we also need to satisfy the equation (4.2.28). These conditions in terms of vector \mathbf{c}_3 are given as follows:

$$\max_{\overline{\Omega}} [\mathbf{c}_3] = \max_{\Omega} [\mathbf{c}_3] = m \geq 0 \text{ and} \quad (4.2.45a)$$

$$\mathbf{c}_3 = m\mathbf{1} \quad \text{in } \overline{\Omega}. \quad (4.2.45b)$$

From equations (4.2.44) and (4.2.45a)–(4.2.45b), it is evident that $m = 0$; which implies that $\mathbf{c}_3 = \mathbf{0}$. Thus, we have $\mathbf{c}_1 = \mathbf{c}_2$ in $\overline{\Omega}$. But from equation (4.2.37), it is obvious that $\mathbf{r}_3 = \mathbf{0}$ in Ω , which contradicts the hypothesis of DsCP $_{\mathbf{K}}$ given by the equation (4.2.42a). Hence, we have the final result $\mathbf{c}_1 \prec \mathbf{c}_2$ in Ω , which completes the proof for the sufficient condition to satisfy DsCP $_{\mathbf{K}}$. \square

In the next section, we shall discuss the various factors that influence the satisfaction of discrete versions of maximum principles, comparison principles, and the non-negative constraint. These factors include, mesh restrictions, numerical formulations, and post-processing methods.

4.3 AN IN-DEPTH LOOK AT CONTINUOUS AND DISCRETE PRINCIPLES

Based on the finite element methodology outlined in subsection 4.2.2, we shall analyze the properties that the stiffness matrix \mathbf{K} inherits from the continuous problem. An important attribute that the discrete system needs to have in order to mimic the mathematical properties that the continuous system possesses is that the stiffness matrix \mathbf{K}_{ff} has to be a (*reducible or irreducible*) *monotone matrix*. The part (a) in all the equations (6.3.9)–(4.2.33) of Theorem 4.2.6 corresponds to reducibility or irreducibility of \mathbf{K}_{ff} .

On general computational grids, it is well-known that the stiffness matrix \mathbf{K}_{ff} obtained via low-order finite element discretization *might not* be a monotone matrix Ciarlet and Raviart (1973); Drăgănescu et al. (2005); Mincsovcics and Hórvath (2012). So, the discrete single-field Galerkin formulation might (or shall) violate the non-negative constraint, discrete maximum principles, and discrete comparison principles on unstructured computational meshes Ciarlet and Raviart (1973); Ishihara (1987); Liska and Shashkov (2008); Nakshatrala and Valocchi (2009); Nagarajan and Nakshatrala (2011); Mincsovcics and Hórvath (2012). The violation is more severe if the diffusion tensor is anisotropic. One of the ways to overcome such unphysical values for concentration and preserve the discrete properties is to restrict the element shape and size in a computational mesh. This can be achieved by developing sufficient mesh conditions under which \mathbf{K}_{ff} is ensured to be a reducibly or irreducibly diagonally dominant matrix Berman and Plemmons (1979). Before we discuss such a class of monotone matrices, which are easily amenable for deriving mesh restrictions, some important remarks on various DMPs and their relationship to DCPs and NC are in order. We would like to emphasize that such a comprehensive discussion is not reported elsewhere in the literature.

4.3.1 Simply connected vs. multiple connected domains

For many applications in mathematics, sciences, and engineering, it is necessary to at least satisfy the weak or strict weak maximum principle. But there are numerous cases where in it is required to satisfy a strong version of the maximum principle Ishihara (1987); Drăgănescu et al. (2005). In such scenarios, geometry and topology of the domain play a vital role. According to the hypothesis of Theorem 4.2.3, it is evident that a strong maximum principle exists if the domain is simply connected (see reference (Gilbarg and Trudinger, 2001, Chapter 3)). However, one should not immediately conclude that if a domain is not simply connected, then a strong maximum principle will not exist Mincsovcics and Hórvath (2012).

In a discrete setting, Ishihara Ishihara (1987), Drăgănescu et.al. Drăgănescu et al. (2005), and Mincsovcics and Hovárth Mincsovcics and Hórvath (2012) have conducted various numerical experiments related to discrete strong maximum principles for multiple connected domains. They performed analysis related to satisfaction of $DsMP_{\mathbf{K}}$ and $DSMP_{\mathbf{K}}$ for various non-obtuse and acute triangulations for multiple connected domains. In particular, Mincsovcics and Hovárth discuss various interesting examples related to the irreducibility property of the stiffness matrix \mathbf{K}_{ff} when the domain is not simply connected. In all of their examples, they solve the following equations:

$$\alpha c - \Delta c = 0 \quad \text{in } \Omega \text{ and} \tag{4.3.1a}$$

$$c(\mathbf{x}) = c^p(\mathbf{x}) \quad \text{on } \partial\Omega, \tag{4.3.1b}$$

where the linear decay $\alpha = 0$ or $\alpha = 128$. Through numerical experiments, the authors demonstrate that even though the triangulation satisfies the non-obtuse or acute angled mesh condition (proposed by Ciarlet and Raviart Ciarlet and Raviart

(1973)), it is not guaranteed to fulfill either DsMP or DSMP. This means that non-obtuse Erten and Üngör (2007) and well-centered triangulation Vanderzee et al. (2010) of any given domain will always satisfy the weak DMPs, but need not satisfy the strong DMPs.

Within the context of directed graphs Berman and Plemmons (1979); Drăgănescu et al. (2005), there is a one-to-one correspondence between irreducibility of the stiffness matrix \mathbf{K}_{ff} and the interior vertices of the computational mesh Huang (2013). In order to satisfy the discrete (strong and strictly strong) maximum principle, the mesh has to be *interiorly connected*, which in turn implies that \mathbf{K}_{ff} has to be irreducible Varga (2009). By interiorly connected mesh, we mean that any pair of interior vertices of the mesh are connected at least by an interior edge path Drăgănescu et al. (2005). Hence, $\mathbf{K}_{ff}^{-1} \succ \mathbf{0}$ and $-\mathbf{K}_{ff}^{-1}\mathbf{K}_{fp} \succ \mathbf{0}$ in Theorem 4.2.6 correspond to this discrete connectedness property of the computational mesh Drăgănescu et al. (2005); Mincsovcics and Hórvath (2012). However, it should be noted that irreducibility is a necessary condition, but not sufficient. For other details on numerical aspects related to mesh connectivity, see references (Mincsovcics and Hórvath, 2012, Section 4, Figures 1–4), Drăgănescu et al. (2005), and Huang (2013).

4.3.2 Minimum principles, and non-negative and min-max constraints

Due to linearity of the operator \mathcal{L} , similar theorems corresponding to minimum principles and the non-negative constraint for equations (5.2.1a)–(5.2.1b) can be derived. To obtain the non-negative solution and corresponding min-max constraint on $c(\mathbf{x})$, we shall appeal to the continuous weak minimum/minimum-maximum principle, which can be written as follows Evans (1998); Gilbarg and Trudinger (2001):

Lemma 4.3.1 (Continuous weak minimum/minimum-maximum principle). *Let \mathcal{L} be a uniformly elliptic operator satisfying the conditions given by (4.2.8a)–(4.2.8c) and $\mathbf{D}(\mathbf{x})$ be continuously differentiable. Given that $\mathcal{L}[c] \geq 0$, $c^p(\mathbf{x}) \geq 0$, and $c(\mathbf{x}) \in$*

$C^2(\Omega) \cap C^0(\bar{\Omega})$, then $c(\mathbf{x})$ possess a continuous weak minimum principle, which is given as

$$\min_{\mathbf{x} \in \bar{\Omega}} [c(\mathbf{x})] \geq \min \left[0, \min_{\mathbf{x} \in \partial\Omega} [c(\mathbf{x})] \right]. \quad (4.3.2)$$

Moreover, if $\mathcal{L}[c] = 0$, then we obtain the classical weak minimum-maximum principle for $c(\mathbf{x})$ in $\bar{\Omega}$, which is given as

$$\min \left[0, \min_{\mathbf{x} \in \partial\Omega} [c(\mathbf{x})] \right] \leq c(\mathbf{x}) \leq \max \left[0, \max_{\mathbf{x} \in \partial\Omega} [c(\mathbf{x})] \right]. \quad (4.3.3)$$

Proof. For a proof, see references Gilbarg and Trudinger (2001); Evans (1998). \square

It is evident from equation (4.3.2) that for $f(\mathbf{x}) \geq 0$ and $c^p(\mathbf{x}) \geq 0$, we have $c(\mathbf{x}) \geq 0$ for any $\mathbf{x} \in \bar{\Omega}$. Correspondingly, a discrete version of continuous weak minimum principle and weak minimum-maximum principle is given as follows:

Definition 4.3.2 (Discrete weak minimum/minimum-maximum principle). *A numerical formulation is said to possess*

- *the discrete weak minimum principle if*

$$\mathbf{r} \succeq \mathbf{0} \quad \text{implies} \quad \min [\mathbf{c}] \geq \min [0, \min [\mathbf{c}_p]] \quad (4.3.4)$$

- *the discrete weak minimum-maximum principle if*

$$\mathbf{r} = \mathbf{0} \quad \text{implies} \quad c_{\min} \mathbf{1} \preceq \mathbf{c} \preceq c_{\max} \mathbf{1}, \quad (4.3.5)$$

where $c_{\min} := \min [\mathbf{c}_p]$, $c_{\max} := \max [\mathbf{c}_p]$, and $\min [\bullet]$ denotes the minimal element of a vector.

4.3.3 High-order finite element methods

An attribute of low-order finite elements, which plays a central role in designing non-negative formulations for diffusion-type equations, is that the shape functions for these elements are monotonic and do not change their sign within the element Nagarajan and Nakshatrala (2011). Moreover, they are convenient to generate computational meshes for complex geometries, to perform error analysis, and for adaptive local mesh refinement George and Frey (2010). High-order finite elements are widely used for solving smooth problems, as one can obtain exponential convergence under high-order interpolations for these problems. But the shape functions of high-order finite elements change the sign within an element, which makes them not suitable under most of the current non-negative formulations (e.g., see reference Nagarajan and Nakshatrala (2011); Payette et al. (2012)). The conditions presented in this chapter will also not be applicable to high-order finite elements for the same reason of change in sign of interpolation functions within an element.

As compared to low-order finite element methods, the discrete counterparts of continuous weak and strong maximum principles for high-order finite element methods is not well understood yet. This is because of the complicated task related to the test of non-negativity of a multivariate polynomial Vejchodský (2010). It should be noted that construction of non-negative high-order shape functions for finite element methods is still an unsolved problem and its roots can be traced back to the famous Hilbert's 17th problem Prestel and Delzell (2001); Reznick (2000). Within the context of variational methods, probably, the research works by Ciarlet Ciarlet (1970b); Ciarlet and Varga (1970) are the first attempt to develop high-order non-negative shape functions to satisfy a discrete maximum principle. This study is based on a general theory of discrete Green's function (DGF) for uniformly elliptic linear partial differential operators. Later, various attempts were made by different researchers to develop shape functions and derive mesh restrictions based on the DGF approach. Most of

them are pertinent to one-dimensional problems or particular cases of isotropic diffusion. The conditions to be met for high-order elements to satisfy DMPs based on DGF methodology are much more stringent and have a less broad scope for general applications. Furthermore, one should be aware that the discrete analogues of continuous Green's functions are applicable only for linear problems, and cannot be extended to non-linear problems (such as semi-linear and quasi-linear elliptic partial differential equations). Hence, such a method will have limited scope. For more details, one can consult the following references Höhn and Mittelman (1981); Šolín and Vejchodský (2007); Vejchodský (2010).

4.3.4 Relationship between various DCPs and DMPs

It is evident from Definitions 4.2.5, 4.2.7, and 4.3.2; Lemma 4.3.1; and Theorems 4.2.6 and 4.2.8 that if a numerical formulation satisfies either DwCP/DwCP $_{\mathbf{K}}$ or DsCP/DsCP $_{\mathbf{K}}$, then it automatically obeys DwMP/DwMP $_{\mathbf{K}}$ and NC. Figure 4.5 illustrates a graphical representation among various numerical solution spaces that satisfy different DMPs and DCPs within the context of mesh restrictions. By a (finite dimensional) numerical solution space $\mathcal{V}_{\mathbf{p}}$, we mean a set of numerical solutions $\{\mathbf{c}_i\}_{i=1}^n$, which satisfy a given discrete property given by \mathbf{p} . For example, if a concentration vector \mathbf{c}_i corresponding to a given volumetric source vector \mathbf{r}_i satisfies the discrete property DwMP, then $\mathbf{c}_i \in \mathcal{V}_{\text{DwMP}}$. It should be noted that within the context of DMPs, DCPs, and NC; $\mathcal{V}_{\mathbf{p}}$ is *not* a vector space. This is because if $\mathbf{c}_i \in \mathcal{V}_{\mathbf{p}}$, then according to non-negative property $-\mathbf{c}_i \notin \mathcal{V}_{\mathbf{p}}$, which is one of the properties needed for $\mathcal{V}_{\mathbf{p}}$ to be a vector space. Moreover, it is evident from the above figure that $\mathcal{V}_{\text{DSMP}_{\mathbf{K}}} \subset \mathcal{V}_{\text{DsMP}_{\mathbf{K}}} \subset \mathcal{V}_{\text{DWMP}_{\mathbf{K}}} \subset \mathcal{V}_{\text{DwMP}_{\mathbf{K}}}$ and $\mathcal{V}_{\text{DsCP}_{\mathbf{K}}} \subset \mathcal{V}_{\text{DwCP}_{\mathbf{K}}}$. But we would like to emphasize that we *do not* have the following enclosures: $\mathcal{V}_{\text{DwCP}_{\mathbf{K}}} \subset \mathcal{V}_{\text{DWMP}_{\mathbf{K}}}$ and $\mathcal{V}_{\text{DsCP}_{\mathbf{K}}} \subset \mathcal{V}_{\text{DSMP}_{\mathbf{K}}}$.

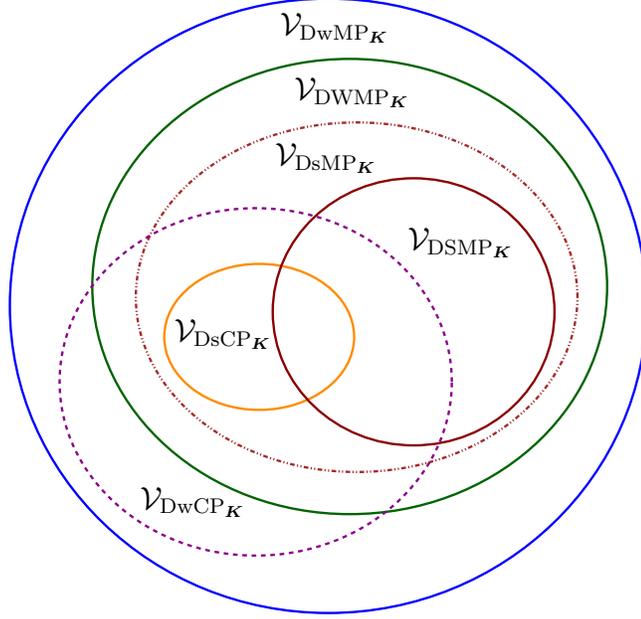


Figure 4.5: Venn diagram for the space of solutions based on mesh restrictions: A pictorial description of the space of numerical solutions satisfying various DMPs and DCPs based on equation (4.2.23) and Theorem 4.2.6.

In numerical literature, most of the numerical methods that exist to satisfy various DMPs are mainly non-linear. In the past decade, considerable advancements have been made to fulfill various version of DMPs for a certain class of linear elliptic and parabolic partial differential equations. But it should be noted that there is seldom research progress related to satisfaction of different DCPs (see (Nakshatrala et al., 2013, Section 4)). *Hence, we would like to highlight that developing a general and variationally consistent numerical technique to encompass all these discrete principles is still an open problem.*

Herein, we would like to emphasize that the route taken to satisfy \mathbf{p} is very important. One can fulfill a discrete property \mathbf{p} in numerous ways. In general, this is achieved by either placing mesh restrictions or developing a new (non-negative or monotone or monotonicity based) numerical formulation or through various post-processing methods. Couple of these techniques are developed based along the lines similar to Theorem 4.2.6 and others based on Definition 4.2.5. But it should be noted

that developing numerical formulations accordant to Theorem 4.2.6 is much more difficult than that of Definition 4.2.5. This is because in order to satisfy Theorem 4.2.6, we need to place restrictions on the stiffness matrices \mathbf{K}_{ff} and \mathbf{K}_{fp} . On the other hand, the hypothesis of Definition 4.2.5 does not assume any particular constraints on \mathbf{K} . Hence, we would like to differentiate between the set of discrete properties given by DwMP $_{\mathbf{K}}$, DWMP $_{\mathbf{K}}$, DsMP $_{\mathbf{K}}$, DSMP $_{\mathbf{K}}$, DwCP $_{\mathbf{K}}$, and DsCP $_{\mathbf{K}}$ to that of DwMP, DWMP, DsMP, DSMP, DwCP, and DsCP.

In spite of the fact that there are several numerical methods available to satisfy a given discrete property \mathfrak{p} , from the characterization of $\mathcal{V}_{\mathfrak{p}}$, it is evident that the resulting numerical solution spaces will be the same (for example, we have $\mathcal{V}_{\text{DwCP}_{\mathbf{K}}} \equiv \mathcal{V}_{\text{DwCP}}$). From Theorems 4.2.6 and 4.2.8, it is evident that among various DMPs and DCPs, we have the following set inclusions:

$$\mathcal{V}_{\text{DSMP}_{\mathbf{K}}} \subset \mathcal{V}_{\text{DsMP}_{\mathbf{K}}} \subset \mathcal{V}_{\text{DWMP}_{\mathbf{K}}} \subset \mathcal{V}_{\text{DwMP}_{\mathbf{K}}} \text{ and} \quad (4.3.6a)$$

$$\mathcal{V}_{\text{DsCP}_{\mathbf{K}}} \subset \mathcal{V}_{\text{DwCP}_{\mathbf{K}}}. \quad (4.3.6b)$$

But it should be noted that a similar type of enclosure for numerical solution spaces between DMPs and DCPs does not hold:

$$\mathcal{V}_{\text{DwCP}_{\mathbf{K}}} \not\subset \mathcal{V}_{\text{DWMP}_{\mathbf{K}}}, \quad (4.3.7a)$$

$$\mathcal{V}_{\text{DwCP}_{\mathbf{K}}} \not\subset \mathcal{V}_{\text{DsMP}_{\mathbf{K}}}, \quad (4.3.7b)$$

$$\mathcal{V}_{\text{DwCP}_{\mathbf{K}}} \not\subset \mathcal{V}_{\text{DSMP}_{\mathbf{K}}}, \text{ and} \quad (4.3.7c)$$

$$\mathcal{V}_{\text{DsCP}_{\mathbf{K}}} \not\subset \mathcal{V}_{\text{DSMP}_{\mathbf{K}}}. \quad (4.3.7d)$$

The reason for such a non-enclosure stems from the hypothesis of Theorems 4.2.6 and 4.2.8, wherein we only need to satisfy DwMP $_{\mathbf{K}}$ for DwCP $_{\mathbf{K}}$ and DsMP $_{\mathbf{K}}$ for DsCP $_{\mathbf{K}}$. In a discrete setting, a numerical methodology may inherit one or more than one

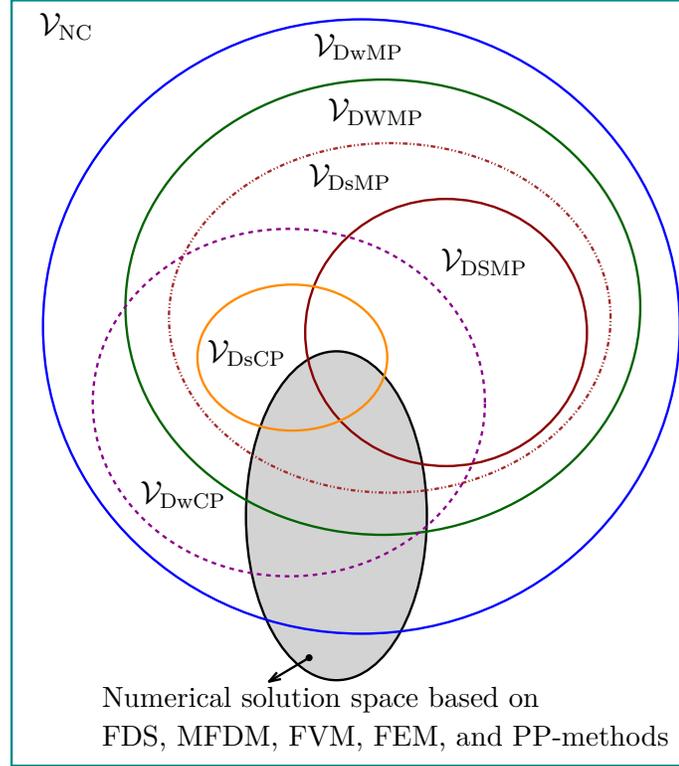


Figure 4.6: Venn diagram for the space of solutions based on various numerical formulations: A pictorial description of the space of numerical solutions satisfying various DMPs, DCPs, and NC.

of these discrete principles and in some cases none. Now, we shall discuss in detail a class of numerical formulations, which satisfy a certain discrete property \mathbf{p} . We shall epitomize our findings based on various popular research works in literature, which span across different disciplines, such as computational geometry, optimization theory, numerical linear algebra, and partial differential equations.

- (a) *Isotropic and anisotropic non-obtuse angle conditions:* Recently, Huang and co-workers Li and Huang (2010); Lu et al. (2012); Huang (2013) were able to satisfy certain discrete properties through mesh restrictions, which are based on anisotropic \mathcal{M} -uniform mesh generation techniques. However, their theoretical investigation is mainly restricted to linear simplicial elements, specifically, the three-node triangular element and four-node tetrahedral element. But one should note that it is difficult to extend the procedure outlined by Huang and co-workers

to multi-linear elements, such as the four-node quadrilateral element, six-node wedge element, and eight-node brick element. This is because the partial derivatives of the shape functions for multi-linear finite elements are not constant (for more details, see subsection 4.4.4 of this chapter, which discusses mesh restrictions for a rectangular element). Nevertheless, constructing a WCT mesh Vanderzee et al. (2010) or an anisotropic \mathcal{M} -uniform triangular mesh Huang (2005); Schneider (2013) that satisfies various DMPs for an *arbitrary domain* is still an open problem Křížek and Qun (1995). The key-concept we would like to emphasize is that the numerical solutions obtained for isotropic diffusion-type equations using WCT meshes and anisotropic diffusion-type equations using the diffusivity tensor based anisotropic \mathcal{M} -uniform meshes satisfy all versions of discrete maximum principles. In addition, if the hypothesis of DwCP $_{\mathbf{K}}$ and DsCP $_{\mathbf{K}}$ is satisfied, then these meshes also satisfy all versions of discrete comparison principles.

- (b) *Non-linear finite volume and mimetic finite difference methods:* Le Potier’s method Potier (2009) and Lipnikov et al. Lipnikov et al. (2007) are some of the noteworthy works in the direction of FVM that satisfy the non-negative constraint, but do not possess a discrete version of the comparison principles and the maximum principles. However, it should be noted that recently these authors have developed techniques based on non-linear finite volume methods Potier (2009); Droniou and Potier (2011) and mimetic finite difference methods Lipnikov et al. (2011) to satisfy various versions of DMPs for a certain *specific* class of linear *self-adjoint* elliptic operators. But it should be noted that there is no discussion on satisfying various DCPs.
- (c) *Optimization-based finite element methods:* Based on the works by Liska and Shashkov Liska and Shashkov (2008) and Nakshatrala and co-workers Nakshatrala and Valocchi (2009); Nagarajan and Nakshatrala (2011), Nakshatrala et al.

(2013), Nakshatrala et al. (2013), the optimization-based low-order finite element methods, under certain conditions (when \mathbf{K}_{ff} is symmetric and positive definite), can be written as follows:

$$\mathbf{K}_{ff}\mathbf{c}_f = \mathbf{r} - \mathbf{K}_{fp}\mathbf{c}_p + \boldsymbol{\lambda}_{\min} - \boldsymbol{\lambda}_{\max}, \quad (4.3.8a)$$

$$c_{\min}^* \mathbf{1} \preceq \mathbf{c}_f \preceq c_{\max}^* \mathbf{1}, \quad (4.3.8b)$$

$$\boldsymbol{\lambda}_{\min} \succeq \mathbf{0}, \quad (4.3.8c)$$

$$\boldsymbol{\lambda}_{\max} \preceq \mathbf{0}, \quad (4.3.8d)$$

$$(\mathbf{c}_f - c_{\min}^* \mathbf{1}) \bullet \boldsymbol{\lambda}_{\min} = 0, \text{ and} \quad (4.3.8e)$$

$$(c_{\max}^* \mathbf{1} - \mathbf{c}_f) \bullet \boldsymbol{\lambda}_{\max} = 0, \quad (4.3.8f)$$

where c_{\min}^* and c_{\max}^* are the minimum and maximum concentration values possible in $\overline{\Omega}$. These values can be obtained based on the boundary conditions and a prior knowledge about the solution. $\boldsymbol{\lambda}_{\min}$ is the vector of Lagrange multipliers corresponding to the constraint $c_{\min}^* \mathbf{1} \preceq \mathbf{c}_f$ and similarly $\boldsymbol{\lambda}_{\max}$ is the vector of Lagrange multipliers corresponding to the constraint $\mathbf{c}_f \preceq c_{\max}^* \mathbf{1}$.

Based on the nature of constraints, one can satisfy different discrete principles and it should be emphasized that DsMP, DSMP, DwCP, and DsCP can be fulfilled only under certain conditions. If either $\mathbf{r} \succ \mathbf{0}$ or $\mathbf{r} \prec \mathbf{0}$, then the non-negative constraint and the weaker versions of discrete minimum/maximum principles can be satisfied by specifying either c_{\min}^* or c_{\max}^* . But in the case of $\mathbf{r} = \mathbf{0}$, both c_{\min}^* and c_{\max}^* can be prescribed based on the Dirichlet boundary conditions; moreover, according to Definition 4.3.2 and from equation (4.3.5), we have $c_{\min}^* = c_{\min}$ and $c_{\max}^* = c_{\max}$. However, one should note that if the qualitative and quantitative nature of the solution is known a priori, then one can satisfy *all* of the discrete versions of (weak and strong) maximum principles by specifying c_{\min}^* and c_{\max}^* in the Karush-Kuhn-Tucker conditions given by equations (4.3.8a)–(4.3.8f). In

general, these methods do not inherit a discrete strong maximum principle and a discrete comparison principle. For more details, a counter example is shown in the reference (Nakshatrala et al., 2013, Section 4, Figure 1)). Nevertheless, satisfying DsMP, DSMP, DwCP, and DsCP is still an open problem and are interesting topics to investigate in future endeavors.

- (d) *Variationally inconsistent methods*: In literature, there are various post-processing methods Kreuzer (2014); Burdakov et al. (2012); Lu et al. (2013) available that can recover certain discrete properties if a prior information about the numerical solution is known. However, one should note that such methods are variationally inconsistent. A summary of the above discussion between various discrete principles within the context of FDS, MFD, FVM, FEM and PP based methods is pictorially described in Figure 4.6.

In the next section, we shall derive sufficient conditions on the three-node triangular element and four-node quadrilateral element to satisfy discrete versions of comparison principles, maximum principles, and the non-negative constraint.

4.4 MESH RESTRICTIONS TO SATISFY DISCRETE PRINCIPLES

In this section, we shall utilize and build upon the research works of Huang and co-workers Huang (2005); Li and Huang (2010); Lu et al. (2012); Huang (2013) for linear second-order elliptic equations. We first present, without proofs, relevant mathematical and geometrical results required to obtain mesh restrictions for simplicial elements. We will then use these results to construct mesh restriction theorems and generate various types of triangulations (see Algorithm 1 and the discussion in subsection 4.4.3) using the open source mesh generators, such as `Gmsh` Geuzaine and Remacle (2015) and `BAMG` Hecht (2006) available in the `FreeFem++` software package Hecht et al. (2014); Hecht (2012). It should be emphasized that these results

cannot be extended to Q4 element, as this element is not simplicial. For more details on mesh restrictions for Q4 element, see subsection 4.4.4 of this chapter.

Let $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_{nd+1}$ denote the vertices of an arbitrary simplex $\Omega_e \in \mathcal{T}_h$, where \mathcal{T}_h is a simplicial triangulation of the domain Ω . The subscript ‘ h ’ in the triangulation \mathcal{T}_h corresponds to the maximum element size (which will be described later in this section, see equation (4.4.20)). Based on various values of h , we have an affine family of such simplicial meshes denoted by $\{\mathcal{T}_h\}$. Designate the total number of vertices and the corresponding interior vertices of \mathcal{T}_h by ‘ Nv ’ and ‘ Niv ’. The edge matrix of Ω_e , which is denoted by \mathbf{E}_{Ω_e} , is defined as

$$\mathbf{E}_{\Omega_e} := [\hat{\mathbf{x}}_2 - \hat{\mathbf{x}}_1, \hat{\mathbf{x}}_3 - \hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_{nd+1} - \hat{\mathbf{x}}_1] \quad \forall \Omega_e \in \mathcal{T}_h, \quad (4.4.1)$$

then an edge connecting vertices $\hat{\mathbf{x}}_p$ and $\hat{\mathbf{x}}_q$ of Ω_e is denoted as e_{pq} . Correspondingly, the edge vector $\mathbf{e}_{pq, \Omega_e}$ (which can be expressed as a linear combination of the elements corresponding to the edge matrix \mathbf{E}_{Ω_e}) and the element boundary $\partial\Omega_e$ in-terms of $\hat{\mathbf{x}}_p, \hat{\mathbf{x}}_q$, and e_{pq} are given by

$$\partial\Omega_e = \bigcup_{\substack{p,q=1 \\ p \neq q}}^{nd+1} e_{pq} \quad \mathbf{e}_{pq, \Omega_e} = \hat{\mathbf{x}}_q - \hat{\mathbf{x}}_p \quad \forall p, q = 1, 2, \dots, nd+1 \quad \text{and} \quad p \neq q. \quad (4.4.2)$$

Following references Brandts et al. (2008); Li and Huang (2010), a set of \mathbf{q} -vectors corresponding to this edge matrix \mathbf{E}_{Ω_e} are defined as

$$\mathbf{E}_{\Omega_e}^{-T} := [\mathbf{q}_2, \mathbf{q}_3, \dots, \mathbf{q}_{nd+1}] \quad \mathbf{q}_1 + \sum_{p=2}^{nd+1} \mathbf{q}_p = \mathbf{0} \quad \forall \Omega_e \in \mathcal{T}_h. \quad (4.4.3)$$

Let φ_{p_g} denote the linear basis function associated with the p_g -th global vertex in the triangulation \mathcal{T}_h . Then, $\{\varphi_{p_g}\}_{p_g=1}^{Nv}$ and $\{\varphi_{p_g}\}_{p_g=1}^{Niv}$ span the respective finite dimensional subsets \mathcal{C}^h and \mathcal{W}^h given by equations (4.2.20a)–(4.2.20b). Denote the face

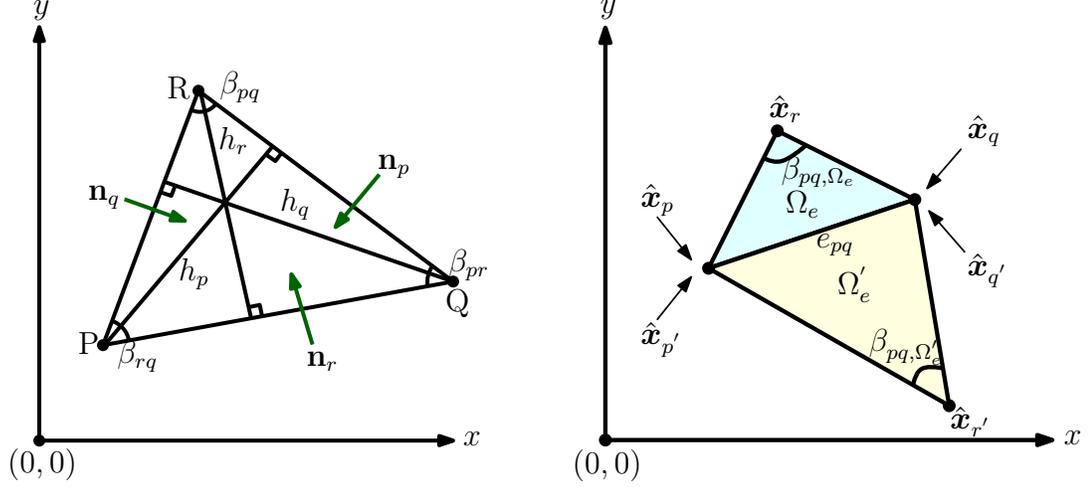


Figure 4.7: Geometrical properties of an arbitrary simplex in 2D: A pictorial description of simplicial mesh element properties.

opposite to vertex $\hat{\mathbf{x}}_p$ by F_p and the corresponding unit inward normal pointing towards the vertex $\hat{\mathbf{x}}_p$ by \mathbf{n}_p . The perpendicular distance (or the height) from vertex $\hat{\mathbf{x}}_p$ to face F_p is denoted by h_p . In the case of 2D, the above set of geometrical properties are pictorially described in Figure 4.7. The left figure shows a pictorial description of various geometrical properties, such as unit inward normals (\mathbf{n}_p , \mathbf{n}_q , and \mathbf{n}_r), dihedral angles in Euclidean metric (β_{pq} , β_{pr} , and β_{rq}), and heights (h_p , h_q , and h_r) of an arbitrary element $\Omega_e \in \mathcal{T}_h$. Correspondingly, the vertices of this triangle PQR are given by $\hat{\mathbf{x}}_p$, $\hat{\mathbf{x}}_q$, and $\hat{\mathbf{x}}_r$. The right figure shows an arbitrary patch of elements Ω_e and Ω'_e , (which belong to the triangulation \mathcal{T}_h) sharing a common edge e_{pq} . The edge e_{pq} connects the coordinates $\hat{\mathbf{x}}_p (= \hat{\mathbf{x}}_{p'})$ and $\hat{\mathbf{x}}_q (= \hat{\mathbf{x}}_{q'})$. The dihedral angles in Euclidean metric opposite to edge e_{pq} are denoted by β_{pq, Ω_e} and β_{pq, Ω'_e} .

Definition 4.4.1 (Positive linear maps). *Let $\mathbb{M}_n := \mathbb{M}_{n \times n}(\mathbb{R})$ be the set of all real matrices of size $n \times n$, which forms a vector space over the field \mathbb{R} . A linear map $\Phi : \mathbb{M}_n \rightarrow \mathbb{M}_k$ is called positive if $\Phi(\mathbf{A})$ is positive semi-definite whenever \mathbf{A} is positive semi-definite. Similarly, Φ is called strictly positive if $\Phi(\mathbf{A})$ is positive definite whenever \mathbf{A} is positive definite.*

Theorem 4.4.2 (Strictly positive linear mapping of anisotropic diffusivity).

Let $\Phi[\bullet] := \int_{\Omega_e} [\bullet] d\Omega$. Show that $\Phi[\bullet]$ is a linear map and $\Phi[\mathbf{D}(\mathbf{x})]$ is symmetric, uniformly elliptic, and bounded above.

Proof. From Definition 4.4.1, it is evident that $\Phi[\bullet]$ is a linear map and $\Phi[\mathbf{D}(\mathbf{x})]$ is symmetric. From equations (6.2.2) and (4.2.19), we have $\boldsymbol{\xi} \bullet \mathbf{D}(\mathbf{x})\boldsymbol{\xi} > 0$ and $\text{meas}(\Omega_e) > 0$. It is well known that Lebesgue integration of a scalar for a strictly positive measure is always greater than zero. Hence, $\int_{\Omega_e} \boldsymbol{\xi} \bullet \mathbf{D}(\mathbf{x})\boldsymbol{\xi} d\Omega > 0 \quad \forall \mathbf{x} \in \Omega_e$. Now, integrating equation (6.2.2) over Ω_e results in the following relation:

$$0 < \gamma_{\min} \boldsymbol{\xi} \bullet \boldsymbol{\xi} \leq \frac{1}{\text{meas}(\Omega_e)} \int_{\Omega_e} \boldsymbol{\xi} \bullet \mathbf{D}(\mathbf{x})\boldsymbol{\xi} d\Omega \leq \gamma_{\max} \boldsymbol{\xi} \bullet \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in \mathbb{R}^{nd} \setminus \{\mathbf{0}\} \text{ and } \forall \mathbf{x} \in \Omega_e, \quad (4.4.4)$$

where the positive constants γ_{\min} and γ_{\max} are respectively the integral average of minimum and maximum eigenvalues of $\mathbf{D}(\mathbf{x})$ over Ω_e . These are given as

$$\gamma_{\min} := \frac{1}{\text{meas}(\Omega_e)} \int_{\Omega_e} \lambda_{\min}(\mathbf{x}) d\Omega \quad \gamma_{\max} := \frac{1}{\text{meas}(\Omega_e)} \int_{\Omega_e} \lambda_{\max}(\mathbf{x}) d\Omega. \quad (4.4.5)$$

Since the vector $\boldsymbol{\xi}$ is independent of \mathbf{x} and Ω_e , we can interchange the order of integration. This gives us the following equation:

$$0 < \gamma_{\min} \boldsymbol{\xi} \bullet \boldsymbol{\xi} \leq \frac{1}{\text{meas}(\Omega_e)} \boldsymbol{\xi} \bullet \Phi[\mathbf{D}(\mathbf{x})]\boldsymbol{\xi} \leq \gamma_{\max} \boldsymbol{\xi} \bullet \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in \mathbb{R}^{nd} \setminus \{\mathbf{0}\} \text{ and } \forall \mathbf{x} \in \Omega_e, \quad (4.4.6)$$

which shows that $\Phi[\mathbf{D}(\mathbf{x})]$ is a strictly positive linear map of $\mathbf{D}(\mathbf{x})$ and indeed preserves its properties. \square

4.4.1 Geometrical properties and finite element analysis of simplicial elements

Based on the above notation, we have the following important mathematical results relating the linear basis functions, finite element matrices, and geometrical

properties of simplicial elements.

1. The integral element average anisotropic diffusivity $\widetilde{\mathbf{D}}_{\Omega_e}$ is given by

$$\widetilde{\mathbf{D}}_{\Omega_e} := \frac{1}{\text{meas}(\Omega_e)} \int_{\Omega_e} \mathbf{D}(\mathbf{x}) d\Omega \quad \forall \Omega_e \in \mathcal{T}_h \quad (4.4.7)$$

is a strictly positive linear map (see Definition 4.4.1 and Theorem 4.4.2).

2. For any arbitrary simplicial element $\Omega_e \in \mathcal{T}_h$, the vector \mathbf{q}_p associated with the face F_p , the gradient of the linear basis function $\text{grad}[\varphi_{p_g}]$, the unit inward normal \mathbf{n}_p , and the height h_p are related as follows Křížek and Qun (1995); Brandts et al. (2008):

$$\mathbf{q}_p = \text{grad}[\varphi_{p_g}] \Big|_{\Omega_e} = \frac{\mathbf{n}_p}{h_p} \quad \forall p = 1, 2, \dots, nd + 1. \quad (4.4.8)$$

3. The dihedral angle β_{pq} (measured in the Euclidean metric) between any two faces F_p and F_q is related to \mathbf{q} -vectors (\mathbf{q}_p and \mathbf{q}_q) and unit inward normals (\mathbf{n}_p and \mathbf{n}_q) by the following equation Li and Huang (2010); Lu et al. (2012):

$$\cos(\beta_{pq}) = -\mathbf{n}_p \bullet \mathbf{n}_q = -\frac{\mathbf{q}_p \bullet \mathbf{q}_q}{\|\mathbf{q}_p\| \|\mathbf{q}_q\|} \quad \forall p, q = 1, 2, \dots, nd + 1 \quad \text{and} \quad p \neq q, \quad (4.4.9)$$

where $\|\bullet\|$ is the standard Euclidean norm. Similarly, the dihedral angle $\beta_{pq, \widetilde{\mathbf{D}}_{\Omega_e}^{-1}}$ measured in $\widetilde{\mathbf{D}}_{\Omega_e}^{-1}$ metric is given as

$$\cos(\beta_{pq, \widetilde{\mathbf{D}}_{\Omega_e}^{-1}}) = -\frac{\mathbf{q}_p \bullet \widetilde{\mathbf{D}}_{\Omega_e} \mathbf{q}_q}{\|\mathbf{q}_p\|_{\widetilde{\mathbf{D}}_{\Omega_e}} \|\mathbf{q}_q\|_{\widetilde{\mathbf{D}}_{\Omega_e}}} \quad \forall p, q = 1, 2, \dots, nd + 1 \quad \text{and} \quad p \neq q, \quad (4.4.10)$$

where $\|\bullet\|_{\widetilde{\mathbf{D}}_{\Omega_e}}$ denotes the norm in $\widetilde{\mathbf{D}}_{\Omega_e}$ metric. For example, $\|\mathbf{q}_p\|_{\widetilde{\mathbf{D}}_{\Omega_e}} = \sqrt{\mathbf{q}_p \bullet \widetilde{\mathbf{D}}_{\Omega_e} \mathbf{q}_p}$.

4. For any simplex $\Omega_e \in \mathcal{T}_h$, the gradient of the linear basis functions ($\text{grad}[\varphi_{p_g}]$ and $\text{grad}[\varphi_{q_g}]$), the dihedral angle β_{pq} , and heights (h_p and h_q) are related as follows Li and Huang (2010); Lu et al. (2012):

$$\text{meas}(\Omega_e) \left(\text{grad}[\varphi_{p_g}] \Big|_{\Omega_e} \bullet \text{grad}[\varphi_{q_g}] \Big|_{\Omega_e} \right) = - \frac{\text{meas}(\Omega_e) \cos(\beta_{pq})}{h_p h_q}$$

$$\forall p, q = 1, 2, \dots, nd + 1 \quad \text{and} \quad p \neq q. \quad (4.4.11)$$

and in 2D, equation (4.4.11) reduces to:

$$\text{meas}(\Omega_e) \left(\text{grad}[\varphi_{p_g}] \Big|_{\Omega_e} \bullet \text{grad}[\varphi_{q_g}] \Big|_{\Omega_e} \right) = - \frac{\cot(\beta_{pq})}{2}$$

$$\forall p, q = 1, 2, \dots, nd + 1 \quad \text{and} \quad p \neq q \quad (4.4.12)$$

5. For any arbitrary simplicial element $\Omega_e \in \mathcal{T}_h$, the integral element average anisotropic diffusivity $\widetilde{\mathbf{D}}_{\Omega_e}$, the gradient of the linear basis functions ($\text{grad}[\varphi_{p_g}]$ and $\text{grad}[\varphi_{q_g}]$), the dihedral angle $\beta_{pq, \widetilde{\mathbf{D}}_{\Omega_e}^{-1}}$ measured in $\widetilde{\mathbf{D}}_{\Omega_e}^{-1}$ metric, and heights (h_p and h_q) are related as follows Lu et al. (2012):

$$\text{meas}(\Omega_e) \left(\text{grad}[\varphi_{p_g}] \Big|_{\Omega_e} \bullet \widetilde{\mathbf{D}}_{\Omega_e} \text{grad}[\varphi_{q_g}] \Big|_{\Omega_e} \right) = - \frac{\text{meas}(\Omega_e) \cos(\beta_{pq, \widetilde{\mathbf{D}}_{\Omega_e}^{-1}})}{\|\mathbf{q}_p\|_{\widetilde{\mathbf{D}}_{\Omega_e}^{-1}} \|\mathbf{q}_q\|_{\widetilde{\mathbf{D}}_{\Omega_e}^{-1}}}$$

$$\forall p, q = 1, 2, \dots, nd + 1 \quad \text{and} \quad p \neq q \quad (4.4.13)$$

and in 2D, equation (4.4.13) reduces to:

$$\text{meas}(\Omega_e) \left(\text{grad}[\varphi_{p_g}] \Big|_{\Omega_e} \bullet \widetilde{\mathbf{D}}_{\Omega_e} \text{grad}[\varphi_{q_g}] \Big|_{\Omega_e} \right) = - \frac{\det[\widetilde{\mathbf{D}}_{\Omega_e}]^{\frac{1}{2}}}{2} \cot(\beta_{pq, \widetilde{\mathbf{D}}_{\Omega_e}^{-1}})$$

$$\forall p, q = 1, 2, \dots, nd + 1 \quad \text{and} \quad p \neq q. \quad (4.4.14)$$

4.4.2 Sufficient conditions for a three-node triangular element

Using the mathematical results outlined in subsection 4.4.1, we shall present various sufficient conditions on the T3 element to satisfy different types of discrete properties. In general, there are *two different approaches* to obtain sufficient conditions. The first approach, which shall be called *global stiffness restriction method*, involves manipulating the entries of the global stiffness matrix, so that it is either weakly or strictly diagonally dominant (see Theorems 4.4.3 and 4.4.4) based on the nature of $\alpha(\mathbf{x})$. This means that \mathbf{K}_{ff}^{-1} exists and $\mathbf{K}_{ff}^{-1} \succeq \mathbf{0}$. The component wise entries of \mathbf{K}_{ff} satisfy the following conditions:

- (a) Positive diagonal entries: $(\mathbf{K}_{ff})_{ii} > 0$,
- (b) Non-positive off-diagonal entries: $(\mathbf{K}_{ff})_{ij} \leq 0 \quad \forall i \neq j$, and

one of the following two conditions:

- (c) Strict diagonal dominance of rows: $|(\mathbf{K}_{ff})_{ii}| > \sum_{i \neq j} |(\mathbf{K}_{ff})_{ij}| \quad \forall i, j$
- (c) Weak diagonal dominance of rows: $|(\mathbf{K}_{ff})_{ii}| \geq \sum_{i \neq j} |(\mathbf{K}_{ff})_{ij}| \quad \forall i, j$

The second method, which shall be called *local stiffness restriction method*, engineers on stiffness matrices at the local level, so that they are weakly diagonally dominant. Once we ascertain that all of the local stiffness matrices are weakly diagonally dominant, then the standard finite element assembly process Wathen (1989) guarantees that the global stiffness matrix \mathbf{K}_{ff} is monotone and weakly diagonally dominant. In particular, if $\alpha(\mathbf{x}) > 0$, then \mathbf{K}_{ff} is strictly diagonally dominant. It should be noted that the above three conditions are sufficient, but not necessary, for the global stiffness matrix \mathbf{K}_{ff} to be monotone.

4.4.2.1 Global and local stiffness restriction methods

In general, to get an explicit analytical formula for \mathbf{K}_{ff}^{-1} is extremely difficult and not practically viable. Hence, it is not feasible to find mesh restrictions based on the condition that $\mathbf{K}_{ff}^{-1} \succeq \mathbf{0}$. So an expedient route to obtain monotone stiffness matrices through mesh restrictions is by means of weakly or strictly diagonally dominant matrices, which form a subset to the class of monotone matrices (Varga, 2009, Section 3, Corollary 3.20 and Corollary 3.21). The obvious edge being that there is no need to compute $(\mathbf{K}_{ff}^{-1})_{ij}$ explicitly. Based on the global stiffness restriction method, we shall now present *stronger and weaker mesh restriction theorems*. These mesh restriction theorems shall be used in constructing triangular meshes to satisfy different discrete principles (see subsection 4.4.3).

Theorem 4.4.3 (Anisotropic non-obtuse angle condition). *If any nd -simplicial mesh satisfies the following anisotropic non-obtuse angle condition:*

$$0 < \frac{h_p \|\mathbf{v}\|_{\infty, \bar{\Omega}_e}}{(nd+1) \Lambda_{\min, \tilde{\mathbf{D}}_{\Omega_e}}} + \frac{h_p h_q \|\alpha\|_{\infty, \bar{\Omega}_e}}{(nd+1)(nd+2) \Lambda_{\min, \tilde{\mathbf{D}}_{\Omega_e}}} \leq \cos(\beta_{pq, \tilde{\mathbf{D}}_{\Omega_e}^{-1}})$$

$$\forall p, q = 1, 2, \dots, nd+1, p \neq q, \Omega_e \in \mathcal{T}_h \quad (4.4.15)$$

then we have the following three results:

- The global stiffness matrix \mathbf{K}_{ff} is (reducibly/irreducibly) weakly diagonally dominant if $\alpha(\mathbf{x}) \geq 0$ and is (reducibly/irreducibly) strictly diagonally dominant if $\alpha(\mathbf{x}) > 0$ for all $\mathbf{x} \in \bar{\Omega}$.
- The discrete single-field Galerkin formulation given by equations (4.2.22a)–(4.2.22b) in combination with equations (4.4.7)–(4.4.14) satisfies DwMP $_{\mathbf{K}}$ /DWMP $_{\mathbf{K}}$, where $\|\mathbf{v}\|_{\infty, \bar{\Omega}_e}$ and $\|\alpha\|_{\infty, \bar{\Omega}_e}$ are defined as

$$\|\mathbf{v}\|_{\infty, \bar{\Omega}_e} := \underset{\mathbf{x} \in \bar{\Omega}_e}{\text{maximize}} \|\mathbf{v}(\mathbf{x})\| \quad \|\alpha\|_{\infty, \bar{\Omega}_e} := \underset{\mathbf{x} \in \bar{\Omega}_e}{\text{maximize}} \alpha(\mathbf{x}) \quad (4.4.16)$$

and $\Lambda_{\min, \widetilde{\mathbf{D}}_{\Omega_e}}$ denotes the minimum eigenvalue of $\widetilde{\mathbf{D}}_{\Omega_e}$; h_p , h_q , and $\beta_{pq, \widetilde{\mathbf{D}}_{\Omega_e}^{-1}}$ are respectively the heights and metric based dihedral angle opposite to the face F_r of element Ω_e .

- Moreover, if the triangulation \mathcal{T}_h is interiorly connected, then the global stiffness matrix \mathbf{K}_{ff} is irreducibly weakly or strictly diagonally dominant based on the nature of $\alpha(\mathbf{x})$ and the discrete single-field Galerkin formulation given by equations (4.2.22a)–(4.2.22b) satisfies $\text{DsMP}_{\mathbf{K}}/\text{DSMP}_{\mathbf{K}}$.

Proof. For proof, see References Lu et al. (2012); Huang (2013). \square

Theorem 4.4.4 (Generalized Delaunay-type angle condition). *In 2D, if a simplicial mesh satisfies the following generalized Delaunay-type angle condition (which is much weaker than that of equation (4.4.15)):*

$$0 < \frac{1}{2} \left[\beta_{pq, \widetilde{\mathbf{D}}_{\Omega_e}^{-1}} + \beta_{pq, \widetilde{\mathbf{D}}_{\Omega'_e}^{-1}} \right] + \frac{1}{2} \text{arccot} \left(\sqrt{\frac{\det[\widetilde{\mathbf{D}}_{\Omega'_e}]}{\det[\widetilde{\mathbf{D}}_{\Omega_e}]} \cot(\beta_{pq, \widetilde{\mathbf{D}}_{\Omega'_e}^{-1}})} - \frac{2 \mathfrak{C}_{q, \Omega_e, \Omega'_e}}{\sqrt{\det[\widetilde{\mathbf{D}}_{\Omega_e}]}} \right) + \frac{1}{2} \text{arccot} \left(\sqrt{\frac{\det[\widetilde{\mathbf{D}}_{\Omega_e}]}{\det[\widetilde{\mathbf{D}}_{\Omega'_e}]} \cot(\beta_{pq, \widetilde{\mathbf{D}}_{\Omega_e}^{-1}})} - \frac{2 \mathfrak{C}_{q, \Omega_e, \Omega'_e}}{\sqrt{\det[\widetilde{\mathbf{D}}_{\Omega'_e}]}} \right) \leq \pi \quad (4.4.17)$$

for every internal edge e_{pq} connecting the p -th and q -th vertices of the elements Ω_e and Ω'_e that share this common edge (see Figure 4.7), then we have the following three results:

- The global stiffness matrix \mathbf{K}_{ff} is (reducibly/irreducibly) weakly diagonally dominant if $\alpha(\mathbf{x}) \geq 0$ and is (reducibly/irreducibly) strictly diagonally dominant if $\alpha(\mathbf{x}) > 0$ for all $\mathbf{x} \in \overline{\Omega}$.
- The discrete single-field Galerkin formulation given by equations (4.2.22a)–(4.2.22b) in association with equations (4.4.7)–(4.4.14) satisfies $\text{DwMP}_{\mathbf{K}}/\text{DWMP}_{\mathbf{K}}$. The quantities h_{p, Ω_e} and h_{q, Ω_e} are the heights of Ω_e , h_{p, Ω'_e} and h_{q, Ω'_e} are the

heights of Ω'_e , $\beta_{pq, \tilde{\mathcal{D}}_{\Omega_e}^{-1}}$ and $\beta_{pq, \tilde{\mathcal{D}}_{\Omega'_e}^{-1}}$ are the relevant metric based dihedral angles in the elements Ω_e and Ω'_e that face the edge e_{pq} , and the parameters $\|\mathbf{v}\|_{\infty, \bar{\Omega}_e}$ and $\|\alpha\|_{\infty, \bar{\Omega}_e}$ are evaluated in Ω_e based on the equations (4.4.16). Similarly, $\|\mathbf{v}\|_{\infty, \bar{\Omega}'_e}$ and $\|\alpha\|_{\infty, \bar{\Omega}'_e}$ are evaluated in Ω'_e . The constant $\mathfrak{C}_{q, \Omega_e, \Omega'_e}$ in equation (4.4.17) is given as follows:

$$\mathfrak{C}_{q, \Omega_e, \Omega'_e} := \text{meas}(\Omega_e) \left(\frac{\|\mathbf{v}\|_{\infty, \bar{\Omega}_e}}{3h_{q, \Omega_e}} + \frac{\|\alpha\|_{\infty, \bar{\Omega}_e}}{12} \right) + \text{meas}(\Omega'_e) \left(\frac{\|\mathbf{v}\|_{\infty, \bar{\Omega}'_e}}{3h_{q, \Omega'_e}} + \frac{\|\alpha\|_{\infty, \bar{\Omega}'_e}}{12} \right) \quad (4.4.18)$$

- Additionally, if the triangulation \mathcal{T}_h is interiorly connected, then the global stiffness matrix \mathbf{K}_{ff} is irreducibly weakly or strictly diagonally dominant based on the nature of $\alpha(\mathbf{x})$ and the discrete single-field Galerkin formulation given by equations (4.2.22a)–(4.2.22b) satisfies $\text{DsMP}_{\mathbf{K}}/\text{DSMP}_{\mathbf{K}}$.

Proof. For proof, see References Lu et al. (2012); Huang (2013). □

Each method (global stiffness restriction method or local stiffness restriction method) has its own advantages and disadvantages. The advantage of the global stiffness restriction method is that we can operate at a global level. This gives us different types of relationships between various mesh parameters, which can be used in generating different types of triangulations, such as Delaunay-Voronoi, non-obtuse, well-centered, and anisotropic \mathcal{M} -uniform finite element meshes. For instance, Taylor series expansion of equations (4.4.15) and (4.4.17) gives the following restrictions on

the metric based dihedral angles for all simplicial elements in a triangulation:

$$0 < \beta_{pq, \widetilde{\mathcal{D}}_{\Omega_e}^{-1}} \leq \frac{\pi}{2} - \mathcal{O}\left(h\|\mathbf{v}\|_{\infty, \mathcal{T}_h} + h^2\|\alpha\|_{\infty, \mathcal{T}_h}\right) \text{ and} \quad (4.4.19a)$$

$$\begin{aligned} 0 < \frac{1}{2} \left[\beta_{pq, \widetilde{\mathcal{D}}_{\Omega_e}^{-1}} + \beta_{pq, \widetilde{\mathcal{D}}_{\Omega'_e}^{-1}} \right] + \frac{1}{2} \operatorname{arccot} \left(\sqrt{\frac{\det[\widetilde{\mathcal{D}}_{\Omega'_e}]}{\det[\widetilde{\mathcal{D}}_{\Omega_e}]} \cot(\beta_{pq, \widetilde{\mathcal{D}}_{\Omega_e}^{-1})} \right) \\ + \frac{1}{2} \operatorname{arccot} \left(\sqrt{\frac{\det[\widetilde{\mathcal{D}}_{\Omega_e}]}{\det[\widetilde{\mathcal{D}}_{\Omega'_e}]} \cot(\beta_{pq, \widetilde{\mathcal{D}}_{\Omega_e}^{-1})} \right) \leq \pi - \mathcal{O}\left(h\|\mathbf{v}\|_{\infty, \mathcal{T}_h} + h^2\|\alpha\|_{\infty, \mathcal{T}_h}\right), \end{aligned} \quad (4.4.19b)$$

where the maximum element size h , maximum element normed velocity $\|\mathbf{v}\|_{\infty, \mathcal{T}_h}$, and maximum element normed linear reaction coefficient $\|\alpha\|_{\infty, \mathcal{T}_h}$ are given as

$$h := \max_{\Omega_e \in \mathcal{T}_h} [h_{\max, \Omega_e}] \quad \|\mathbf{v}\|_{\infty, \mathcal{T}_h} := \max_{\Omega_e \in \mathcal{T}_h} [\|\mathbf{v}\|_{\infty, \overline{\Omega_e}}] \quad \|\alpha\|_{\infty, \mathcal{T}_h} := \max_{\Omega_e \in \mathcal{T}_h} [\|\alpha\|_{\infty, \overline{\Omega_e}}] \quad (4.4.20)$$

where h_{\max, Ω_e} is the maximum possible height in a given simplicial element Ω_e , and $\mathcal{O}(\bullet)$ is the standard “big-oh” notation. Specifically, on a h -refined triangular mesh, which conforms to Theorems 4.4.3 and 4.4.4, then equation (4.4.19a) implies that all the dihedral angles when measured in the metric of $\widetilde{\mathcal{D}}_{\Omega_e}^{-1}$ have to be $\mathcal{O}(h\|\mathbf{v}\|_{\infty, \mathcal{T}_h} + h^2\|\alpha\|_{\infty, \mathcal{T}_h})$ acute/non-obtuse and equation (4.4.19b) indicates that the triangulation needs to be $\mathcal{O}(h\|\mathbf{v}\|_{\infty, \mathcal{T}_h} + h^2\|\alpha\|_{\infty, \mathcal{T}_h})$ Delaunay.

One downside of the global stiffness restriction approach is that obtaining mesh conditions is mathematically cumbersome. Moreover, extending it to non-simplicial low-order finite elements is extremely hard and is not straightforward. This is because the basis functions φ_{p_g} spanning the finite dimensional subsets \mathcal{C}^h and \mathcal{W}^h are multilinear, which makes $\operatorname{grad}[\varphi_{p_g}]$ on any arbitrary element Ω_e to be non-constant. So most of properties given by equations (4.4.7)–(4.4.14) are not valid for low-order finite elements such as the Q4 element and its corresponding elements in higher dimensions.

Now we shall describe the local stiffness restriction method and highlight the pros

and cons of using this methodology. For the sake of illustration, we shall consider a pure anisotropic diffusion equation and assume that $\hat{\mathbf{x}}_p = (0, 0)$, $\hat{\mathbf{x}}_q = (1, 0)$, and $\hat{\mathbf{x}}_r = (a, b)$. Our objective is to find the coordinates (a, b) , such that the local stiffness matrix is weakly diagonally dominant for any given type of diffusivity tensor. The local stiffness matrix for an anisotropic diffusion equation based on discrete single-field Galerkin formulation is given by

$$\mathbf{K}_e = \int_{\Omega_e} \mathbf{B} \mathbf{D}(\mathbf{x}) \mathbf{B}^t d\Omega \quad \mathbf{B} = \frac{1}{b} \begin{pmatrix} -b & (a-1) \\ b & -a \\ 0 & 1 \end{pmatrix}. \quad (4.4.21)$$

In the subsequent subsections, we present various sufficient conditions through which we can find these coordinates (a, b) and glean information on the possible angles and corresponding shape and size of the triangle PQR.

4.4.2.2 T3 element for heterogeneous isotropic diffusivity

In this subsection, we consider the case where the diffusivity is isotropic and heterogeneous in the total domain. For this case, we show that the diffusivity *does not* have any influence on determining the coordinates (a, b) . This means that the restrictions we obtain on the coordinates and the angles of the triangle PQR is independent of how the diffusivity is varying across the domain. The following is the local stiffness matrix for scalar heterogeneous isotropic diffusion:

$$\mathbf{K}_e = \frac{\widetilde{D}}{2b} \begin{pmatrix} b^2 + (a-1)^2 & a - a^2 - b^2 & (a-1) \\ a - a^2 - b^2 & a^2 + b^2 & -a \\ (a-1) & -a & 1 \end{pmatrix}, \quad (4.4.22)$$

where \widetilde{D} is the integral average of the diffusivity $D(\mathbf{x})$ over the actual T3 element Ω_e (triangle PQR). We shall now present the sufficient conditions so that the matrix

\mathbf{K}_e is weakly diagonally dominant:

Condition #1

Positive diagonal entries: $(\mathbf{K}_e)_{ii} > 0 \quad \forall i = 1, 2, 3$. This restriction gives us the following inequalities:

$$\frac{\widetilde{D}}{2b} (b^2 + (a-1)^2) > 0 \quad \frac{\widetilde{D}}{2b} (a^2 + b^2) > 0 \quad \frac{\widetilde{D}}{2b} > 0. \quad (4.4.23)$$

As $\widetilde{D} > 0$ and $b > 0$, it is evident that all of the inequalities given by equations (4.4.23) are trivially satisfied. Hence, this condition has no effect on obtaining restrictions on coordinates (a, b) .

Condition #2

Weak diagonal dominance of rows: $|(\mathbf{K}_e)_{ii}| \geq \sum_{i \neq j} |(\mathbf{K}_e)_{ij}| \quad \forall i, j$, where $i = 1, 2, 3$ and $j = 1, 2, 3$. This restriction gives the following inequalities:

$$b^2 + (a-1)^2 \geq (a^2 + b^2 - a) + (1-a) \quad a^2 + b^2 \geq a + (a^2 + b^2 - a) \quad 1 \geq (1-a) + a. \quad (4.4.24)$$

Note that these inequalities (4.4.24) are trivially satisfied. Hence, this condition has no influence on obtaining restrictions on triangle PQR.

Condition #3

Non-positive off-diagonal entries: $(\mathbf{K}_e)_{ij} \leq 0 \quad \forall i \neq j$, where $i = 1, 2, 3$ and $j = 1, 2, 3$. As $\widetilde{D} > 0$ and $b > 0$, we get the following inequalities:

$$\left(a - \frac{1}{2}\right)^2 + b^2 \geq \left(\frac{1}{2}\right)^2 \quad a \leq 1 \quad a \geq 0. \quad (4.4.25)$$

The region in which the coordinates (a, b) satisfy the above inequalities given by

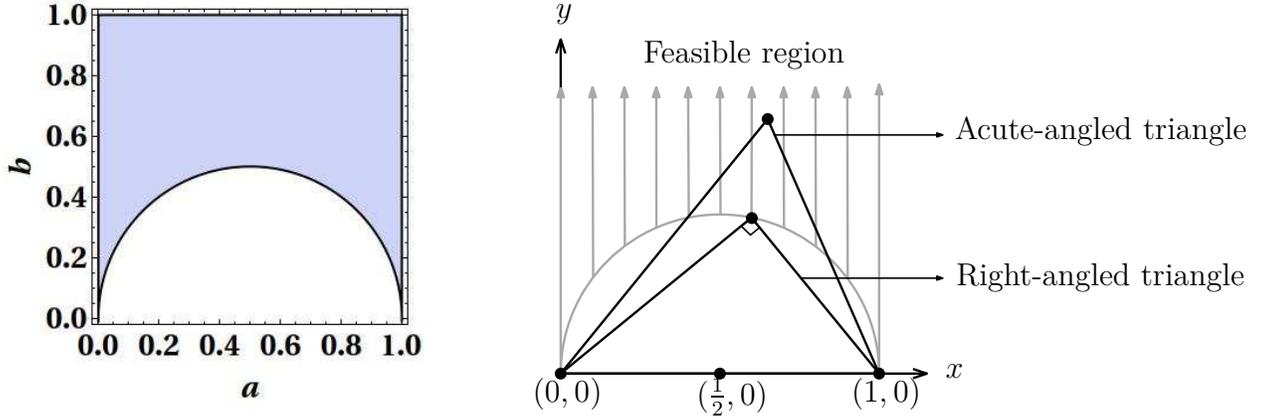


Figure 4.8: T3 element for heterogeneous isotropic diffusivity: A pictorial description of the feasible region is shown in light blue color.

the equation (4.4.25) is shown in Figure 4.8. According to these inequalities (4.4.25), heterogeneity of the scalar diffusivity has *no role* in obtaining the feasible region for the coordinates (a, b) . It is evident from Figure 4.8 that the interior angles of the triangle PQR are either acute or at most right-angle. Based on the sufficient conditions, one can also notice that an obtuse-angled triangle is not possible. So in order to satisfy discrete comparison principles, discrete maximum principles, and non-negative constraint, the triangulation of a given computational domain must contain acute-angled triangles or right-angled triangles. These three sufficient conditions show that non-obtuse or well-centered triangulations inherit all the three discrete versions of continuous properties of scalar heterogeneous isotropic diffusion equations.

4.4.2.3 T3 element for heterogeneous anisotropic diffusivity

In this subsection, we consider the case where the diffusivity $\mathbf{D}(\mathbf{x}) = \begin{pmatrix} D_{xx}(\mathbf{x}) & D_{xy}(\mathbf{x}) \\ D_{xy}(\mathbf{x}) & D_{yy}(\mathbf{x}) \end{pmatrix}$ is anisotropic and heterogeneous across the domain. For the sake of brevity and ease of manipulations, we shall drop the symbol (\mathbf{x}) in the components of the diffusivity tensor. Note that the symbol ‘ \mathbf{x} ’ in the components of $\mathbf{D}(\mathbf{x})$ is dropped for the sake of convenience and *should not* be interpreted as though the diffusivity tensor is constant. As discussed in Section 4.2, the diffusivity tensor needs to satisfy certain properties.

Based on equations (4.2.2) and (6.2.2), we derive various results related to $\mathbf{D}(\mathbf{x})$ that will be used in deriving mesh restrictions.

Remark 4.4.5. *In 2D, it is trivial to show that, if the matrix $\mathbf{D}(\mathbf{x})$ is symmetric, uniformly elliptic, and bounded above, then its components satisfy the following relations:*

$$D_{xx} > 0 \quad D_{yy} > 0 \quad D_{xx}D_{yy} > D_{xy}^2. \quad (4.4.26)$$

Let us denote $\epsilon := \frac{\tilde{D}_{yy}}{D_{xx}}$ and $\eta := \frac{\tilde{D}_{xy}}{D_{xx}}$, where \tilde{D}_{xx} , \tilde{D}_{xy} , and \tilde{D}_{yy} are the components of matrix $\tilde{\mathbf{D}}_{\Omega_e}$ given by equation (4.4.7). From Theorem 4.4.2, it is evident that $\tilde{D}_{xx} > 0$, $\tilde{D}_{yy} > 0$, and $\tilde{D}_{xx}\tilde{D}_{yy} > \tilde{D}_{xy}^2$. So from equation (4.4.26), we have $\eta \in (-\sqrt{\epsilon}, \sqrt{\epsilon})$. These two non-dimensional quantities ϵ and η govern the mesh restrictions that we impose on the coordinates (a, b) . From equation (4.4.21), the stiffness matrix for any given anisotropic diffusivity tensor is given as

$$\mathbf{K}_e = \begin{pmatrix} \frac{\tilde{D}_{xx}b^2 - 2\tilde{D}_{xy}b(a-1) + \tilde{D}_{yy}(a-1)^2}{2b} & -\frac{\tilde{D}_{xx}b^2 + \tilde{D}_{xy}(b-2ab) + \tilde{D}_{yy}a(a-1)}{2b} & \frac{-\tilde{D}_{xy}b + \tilde{D}_{yy}(a-1)}{2b} \\ -\frac{\tilde{D}_{xx}b^2 + \tilde{D}_{xy}(b-2ab) + \tilde{D}_{yy}a(a-1)}{2b} & \frac{\tilde{D}_{xx}b^2 - 2\tilde{D}_{xy}ab + \tilde{D}_{yy}a^2}{2b} & \frac{\tilde{D}_{xy}b - \tilde{D}_{yy}a}{2b} \\ \frac{-\tilde{D}_{xy}b + \tilde{D}_{yy}(a-1)}{2b} & \frac{\tilde{D}_{xy}b - \tilde{D}_{yy}a}{2b} & \frac{\tilde{D}_{yy}}{2b} \end{pmatrix}. \quad (4.4.27)$$

We now present sufficient conditions so that the matrix \mathbf{K}_e is weakly diagonally dominant.

Condition #4

Positive diagonal entries: $(\mathbf{K}_e)_{ii} > 0 \quad \forall i = 1, 2, 3$, gives the following relations:

$$\left(b\sqrt{\tilde{D}_{xx}} - |a-1|\sqrt{\tilde{D}_{yy}} \right)^2 + 2b|a-1| \left(\sqrt{\tilde{D}_{xx}}\sqrt{\tilde{D}_{yy}} - \text{Sgn}[|a-1|]\tilde{D}_{xy} \right) > 0 \quad \text{and} \quad (4.4.28a)$$

$$\left(b\sqrt{\tilde{D}_{xx}} - |a|\sqrt{\tilde{D}_{yy}} \right)^2 + 2b|a| \left(\sqrt{\tilde{D}_{xx}}\sqrt{\tilde{D}_{yy}} - \text{Sgn}[|a|]\tilde{D}_{xy} \right) > 0, \quad (4.4.28b)$$

where $\text{Sgn}[\bullet]$ is the standard signum function (which provides the sign of the real number). It is evident that $\sqrt{\widetilde{D}_{xx}}\sqrt{\widetilde{D}_{yy}} > \widetilde{D}_{xy}$. Hence, equations (4.4.28a)-(4.4.28b) are trivially satisfied for any abscissa a .

Condition #5

Non-positive off-diagonal entries: $(\mathbf{K}_e)_{ij} \leq 0 \quad \forall i \neq j$, where $i = 1, 2, 3$, and $j = 1, 2, 3$. This restriction gives the following relations:

$$\left(a - \frac{1}{2}\right)^2 + \left(\frac{b}{\sqrt{\epsilon}}\right)^2 - 2b\left(\frac{\eta}{\epsilon}\right)\left(a - \frac{1}{2}\right) \geq \left(\frac{1}{2}\right)^2 \quad \frac{a-1}{b} \leq \frac{\eta}{\epsilon} \quad \frac{a}{b} \geq \frac{\eta}{\epsilon}, \quad (4.4.29)$$

which dictate the feasible region for coordinates (a, b) . For a given ϵ and by varying η , which lies between $-\sqrt{\epsilon}$ and $\sqrt{\epsilon}$, we get different feasible regions for (a, b) . Herein, we have chosen $\epsilon = 10$ and $\eta \in \{-1, 0, 1\}$. For these values, we have plotted the feasible region based on the inequalities (4.4.29). From Figures 4.9–4.13, the following can be inferred based on the feasible region:

- If $\eta = 0$, the possible T3 elements are either acute-angled or right-angled triangles.
- If either $\eta < 0$ or $\eta > 0$, then obtuse-angled triangles are also possible. Moreover, the resulting triangles can be *skinny* or *skewed*.

In Figure 4.9, the numerical values for the two parameters, which decide the feasible region, are chosen to be $\epsilon = 10$ and $\eta = 0$. In this case, the right figure indicates that acute-angled and right-angled triangles are possible. As ϵ increases, the coordinate b has to increase proportionally to satisfy the inequality given by the equation (4.4.29). In Figure 4.10, we have chosen $\epsilon = 10$ and $\eta = -1$. For a fixed η as ϵ increases, the value of coordinate b also increases. So it is a daunting task to find a viable T3 element. One can also notice that the feasible region is *not* contiguous. In Figure 4.11, we have chosen $\epsilon = 10$ and $\eta = 1$. For a fixed η as ϵ increases, the

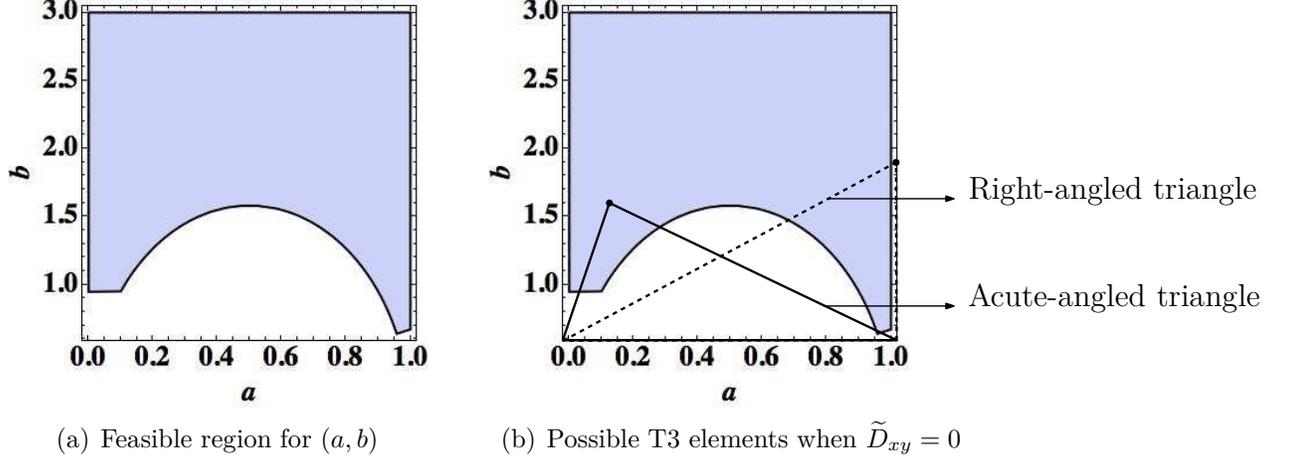


Figure 4.9: T3 element for anisotropic diffusivity when $\tilde{D}_{xy} = 0$: A pictorial description of the feasible region (left figure) for the coordinates (a, b) is indicated in light blue color.

value of coordinate b also increases. For higher values of ϵ , it is very difficult to find a suitable T3 element, which can mesh any given computational domain.

Condition #6

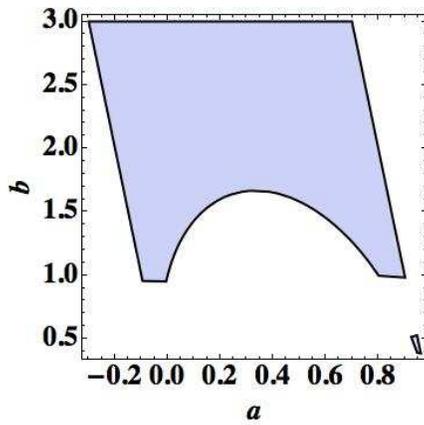
Weak diagonal dominance of rows: $|(\mathbf{K}_e)_{ii}| \geq \sum_{i \neq j} |(\mathbf{K}_e)_{ij}| \quad \forall i, j$, where $i = 1, 2, 3$, and $j = 1, 2, 3$. This gives the following relations:

$$(b^2 - 2\eta b(a - 1) + \epsilon(a - 1)^2) \geq (b^2 + \eta(b - 2ab) + \epsilon a(a - 1)) + (\eta b - \epsilon(a - 1)), \quad (4.4.30a)$$

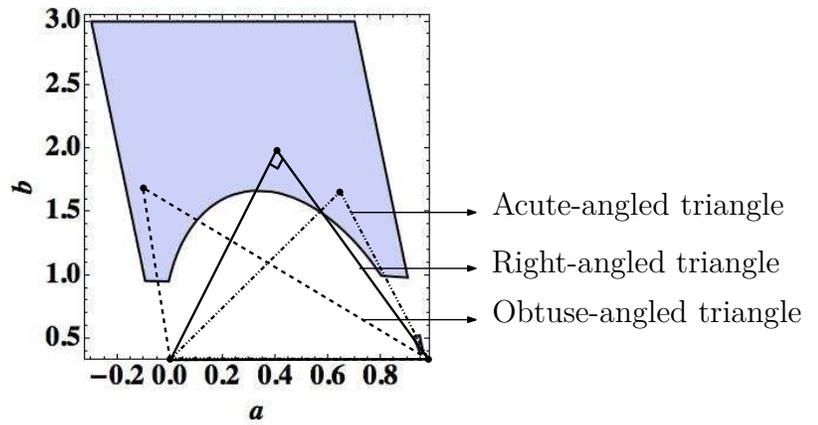
$$(b^2 - 2\eta ab + \epsilon a^2) \geq (b^2 + \eta(b - 2ab) + \epsilon a(a - 1)) (\epsilon a - \eta b), \text{ and} \quad (4.4.30b)$$

$$\epsilon \geq (\eta b - \epsilon(a - 1)) + (\epsilon a - \eta b), \quad (4.4.30c)$$

if Condition #4 and Condition #5 are satisfied, then this condition is trivially satisfied. In a similar fashion, for a general case, where $\hat{\mathbf{x}}_p = (x_1, y_1)$, $\hat{\mathbf{x}}_q = (x_2, y_2)$, and $\hat{\mathbf{x}}_r = (x_3, y_3)$, we have the following conditions based on the local stiffness restriction

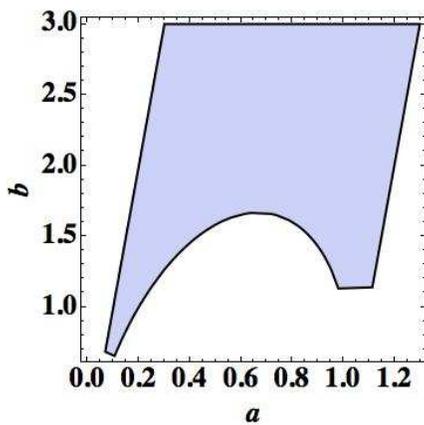


(a) Feasible region for (a, b)

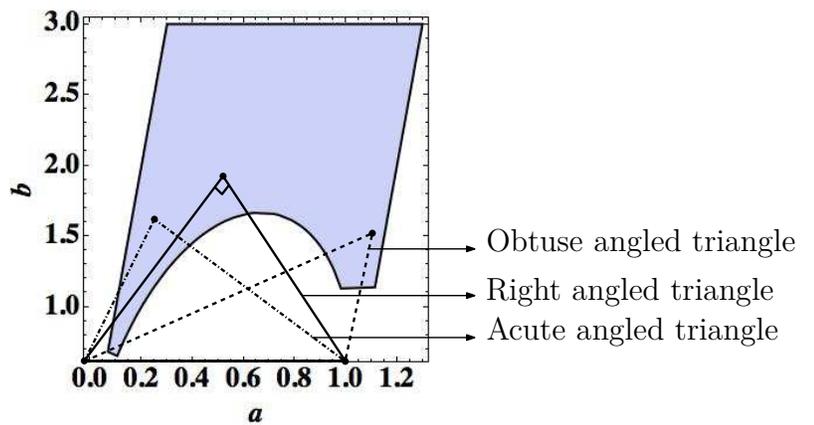


(b) Possible T3 elements when $\tilde{D}_{xy} < 0$

Figure 4.10: T3 element for anisotropic diffusivity when $\tilde{D}_{xy} < 0$: The left figure indicates the feasible region for the coordinates (a, b) in light blue color. The right figure indicates that the T3 element can be acute/right/obtuse-angled.

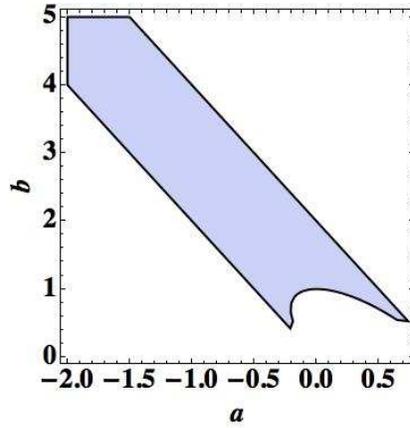


(a) Feasible region for (a, b)

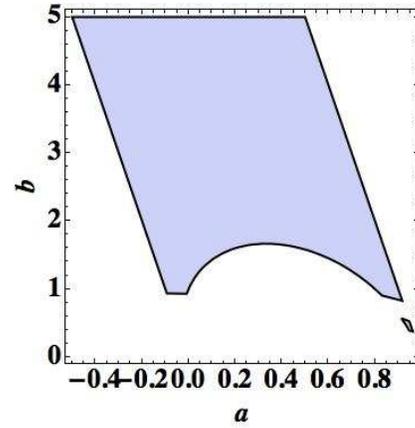


(b) Possible T3 elements when $\tilde{D}_{xy} > 0$

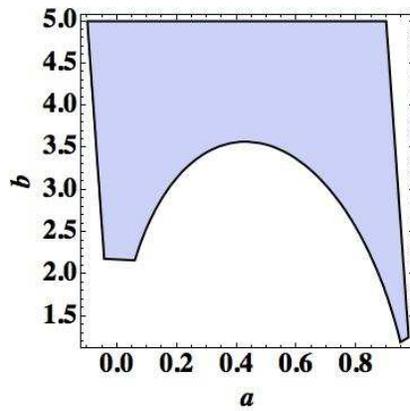
Figure 4.11: T3 element for anisotropic diffusivity when $\tilde{D}_{xy} > 0$: The left figure indicates the feasible region for the coordinates (a, b) in light blue color. The right figure indicates that the T3 element can be acute/right/obtuse-angled.



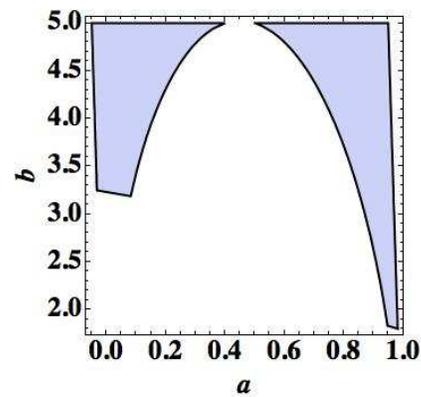
(a) $\epsilon = 2$ and $\eta = -1$



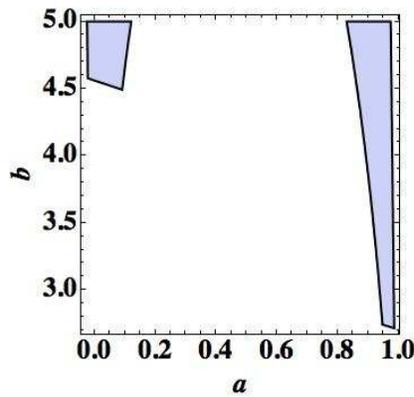
(b) $\epsilon = 10$ and $\eta = -1$



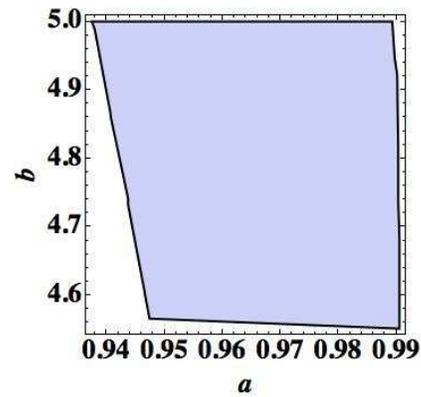
(c) $\epsilon = 50$ and $\eta = -1$



(d) $\epsilon = 100$ and $\eta = -1$



(e) $\epsilon = 200$ and $\eta = -1$



(f) $\epsilon = 500$ and $\eta = -1$

Figure 4.12: T3 element for fixed η and varying ϵ : A pictorial description of the feasible region (light blue color) for a fixed η and varying ϵ . Analysis is performed for $\eta = -1$ and $\epsilon = \{2, 10, 50, 100, 200, 500\}$.

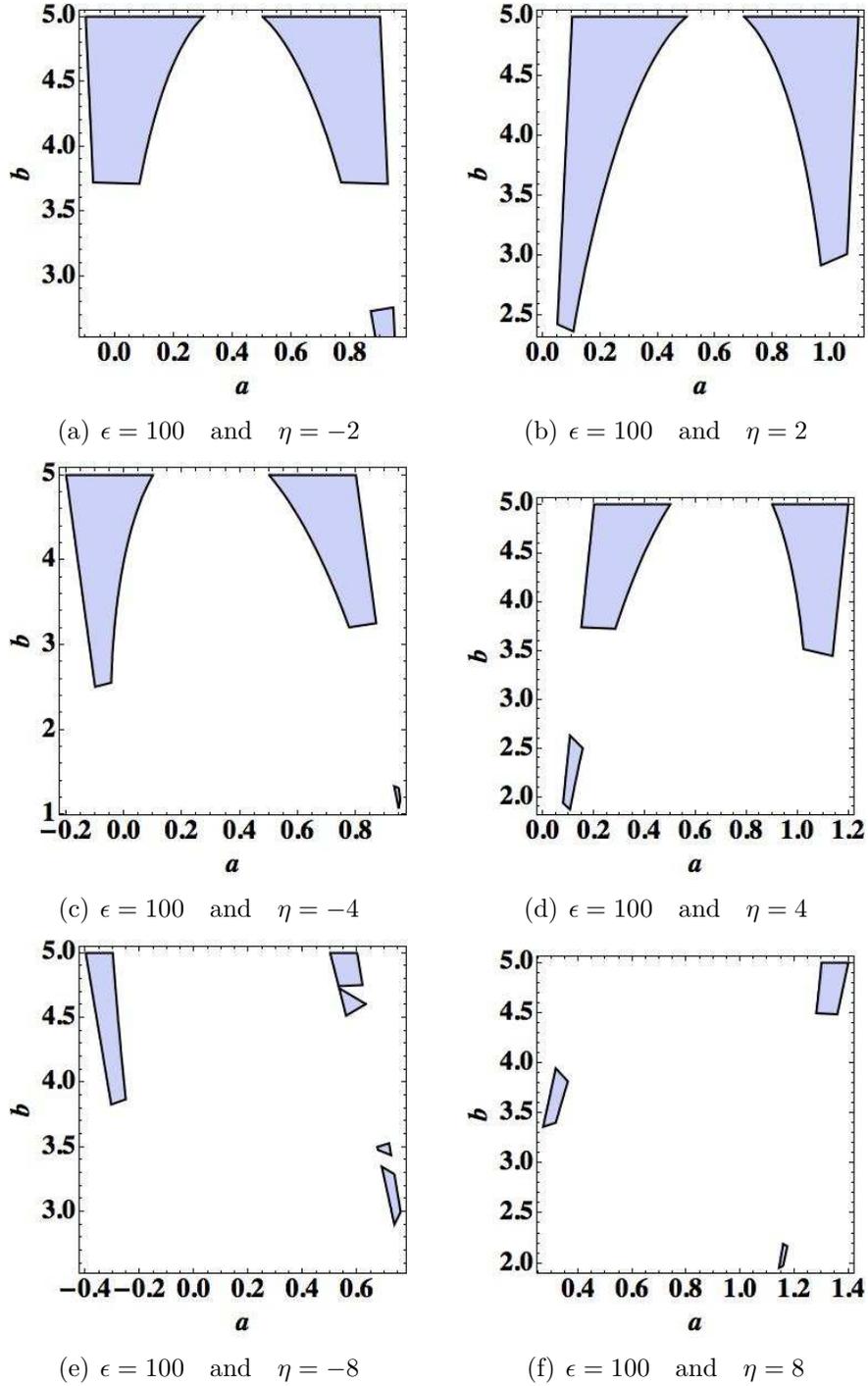


Figure 4.13: T3 element for fixed ϵ and varying η : A pictorial description of the feasible region (light blue color) for a fixed ϵ and varying η . Analysis is performed for $\epsilon = 100$ and $\eta = \{-8, -4, -2, 2, 4, 8\}$.

method:

$$(y_1 - y_3)(y_3 - y_2) - \eta(x_1 - x_3)(y_3 - y_2) - \eta(x_3 - x_2)(y_1 - y_3) + \epsilon(x_1 - x_3)(x_3 - x_2) \leq 0, \quad (4.4.31a)$$

$$(y_2 - y_1)(y_3 - y_2) - \eta(x_3 - x_2)(y_2 - y_1) - \eta(x_2 - x_1)(y_3 - y_2) + \epsilon(x_2 - x_1)(x_3 - x_2) \leq 0, \quad (4.4.31b)$$

$$(y_1 - y_3)(y_2 - y_1) - \eta(x_1 - x_3)(y_2 - y_1) - \eta(x_2 - x_1)(y_1 - y_3) + \epsilon(x_1 - x_3)(x_2 - x_1) \leq 0, \text{ and} \quad (4.4.31c)$$

$$(x_1 - x_3)(y_2 - y_1) - (x_1 - x_2)(y_3 - y_1) > 0, \quad (4.4.31d)$$

where the first three inequalities given by equations (4.4.31a)–(4.4.31c) are obtained based on the condition that $(\mathbf{K}_e)_{ij} \leq 0$. The last inequality given by equation (4.4.31d) is the result of the condition that $\text{meas}(\Omega_e) > 0$.

The benefit (attractive feature) of the local stiffness restriction method is that the local stiffness matrix for the discrete Galerkin formulation given by equations (4.2.22a)–(4.2.22b) can be calculated quite easily and could be extended to even non-simplicial elements (see subsection 4.4.4). Using this approach, we can obtain general restrictions and analytical expressions relating various coordinates of an arbitrary mesh element, using popular symbolic packages like **Mathematica Wolfram** (2013). But a flip-side of this procedure is that incorporating the inequalities given by equations (4.4.31a)–(4.4.31d) in a mesh generator is very difficult and needs further detailed investigation. Additionally, the conditions obtained using this method are stringent and similar to that of Theorem 4.4.3 given by the global stiffness restriction method (which will be evident based on the numerical examples discussed in the following subsection). Finally, it should be noted that extending the local stiffness restriction method to include advection and linear reaction is straightforward and shall not be dealt with to save space. We shall now present various numerical examples and respective triangular meshes corresponding to different types of $\mathbf{D}(\mathbf{x})$. Using these meshes, we shall analyze and study in detail, which kind of DMPs and DCPs are preserved.

4.4.3 Numerical examples based on different types of triangulations

In this subsection, we shall first briefly discuss on a metric tensor $\mathcal{M}(\mathbf{x})$ to satisfy DCPs, DMPs, and NC. Based on this metric tensor, we shall describe an algorithm to generate various types of DMP-based triangulations (mainly utilizing open source mesh generators, such as **Gmsh** and **BAMG**). Simplicial meshes constructed based on $\mathcal{M}(\mathbf{x})$ (where the metric tensor $\mathcal{M}(\mathbf{x})$ is *not equal* to a scalar multiple of identity tensor) are called anisotropic \mathcal{M} -uniform simplicial meshes. They are uniform in the metric specified by $\mathcal{M}(\mathbf{x})$ Huang (2005); George and Frey (2010) and are of primal importance in satisfying various important discrete properties in the areas of transport of chemical species, fluid mechanics, and porous media applications Castro-Diaz et al. (1997); Frey and Alauzet (2005). Given a $\mathcal{M}(\mathbf{x})$, there are different approaches to generate anisotropic \mathcal{M} -uniform simplicial meshes. Some of the notable research works in this direction include blue refinement, bubble packing, Delaunay-type triangulation, directional refinement, front advancing, local refinement and modification, and variational mesh generation George and Frey (2010); Schneider (2013). In general, the metric tensor $\mathcal{M}(\mathbf{x})$ is symmetric and positive definite, and gives relevant information on the shape, size, and orientation of mesh elements in the computational domain.

Let χ be an affine mapping from the reference element Ω_{ref} to background mesh element Ω_e . Denote its Jacobian by \mathbf{J}_{Ω_e} . The affine mapping χ and its Jacobian \mathbf{J}_{Ω_e} are given as

$$\mathbf{x} = \chi(\hat{\mathbf{X}}, \mathbf{N}) = \hat{\mathbf{X}}^T \mathbf{N}^T \quad \mathbf{J}_{\Omega_e} = \hat{\mathbf{X}}^T \mathbf{D}\mathbf{N}, \quad (4.4.32)$$

where $\hat{\mathbf{X}}$ is the nodal matrix, which comprises of nodal vertices $(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_{nd+1})$ of an arbitrary simplex Ω_e . The elements of vector \mathbf{N} consists of shape functions corresponding to Ω_{ref} . The entries of matrix $\mathbf{D}\mathbf{N}$ (which are constants for simplicial

elements) correspond to the derivatives of shape functions. Straightforward manipulations on equations (4.4.8) and (4.4.32) give the following result:

$$\mathbf{q}_{p,\text{ref}} = \mathbf{J}_{\Omega_e} \mathbf{q}_p, \quad (4.4.33)$$

where $\mathbf{q}_{p,\text{ref}}$ is the corresponding \mathbf{q} -vector of the reference element Ω_{ref} . Huang Huang (2005) has shown that an anisotropic \mathcal{M} -uniform simplicial mesh satisfies the following two conditions:

$$\rho_{\Omega_e} \text{meas}(\Omega_e) = \frac{\sigma_h}{N_{ele}} \quad \forall \Omega_e \in \mathcal{T}_h \text{ and} \quad (4.4.34a)$$

$$\frac{1}{nd} \text{tr} [\mathbf{J}_{\Omega_e}^T \mathcal{M}_{\Omega_e} \mathbf{J}_{\Omega_e}] = \det [\mathbf{J}_{\Omega_e}^T \mathcal{M}_{\Omega_e} \mathbf{J}_{\Omega_e}]^{\frac{1}{nd}} \quad \forall \Omega_e \in \mathcal{T}_h. \quad (4.4.34b)$$

The quantities \mathcal{M}_{Ω_e} , ρ_{Ω_e} , and σ_h are given as follows:

$$\mathcal{M}_{\Omega_e} := \frac{1}{\text{meas}(\Omega_e)} \int_{\Omega_e} \mathcal{M}(\mathbf{x}) d\Omega \quad \forall \Omega_e \in \mathcal{T}_h, \quad (4.4.35a)$$

$$\rho_{\Omega_e} := \sqrt{\det [\mathcal{M}_{\Omega_e}]} \quad \forall \Omega_e \in \mathcal{T}_h, \text{ and} \quad (4.4.35b)$$

$$\sigma_h := \sum_{\Omega_e \in \mathcal{T}_h} \rho_{\Omega_e} \text{meas}(\Omega_e). \quad (4.4.35c)$$

The condition given by the equation (4.4.34a) is called *equidistribution condition* and that by equation (4.4.34b) is called *alignment condition*. Equidistribution condition decides the size of Ω_e , while alignment condition characterizes the shape and orientation of Ω_e . From AM-GM inequality, equation (4.4.34b) implies

$$\mathbf{J}_{\Omega_e}^T \mathcal{M}_{\Omega_e} \mathbf{J}_{\Omega_e} = \left(\frac{\sigma_h}{N_{ele}} \right)^{\frac{2}{nd}} \mathbf{I} \quad \forall \Omega_e \in \mathcal{T}_h. \quad (4.4.36)$$

Now, through trivial manipulations on equations (4.4.33), (4.4.10), and (4.4.36), the

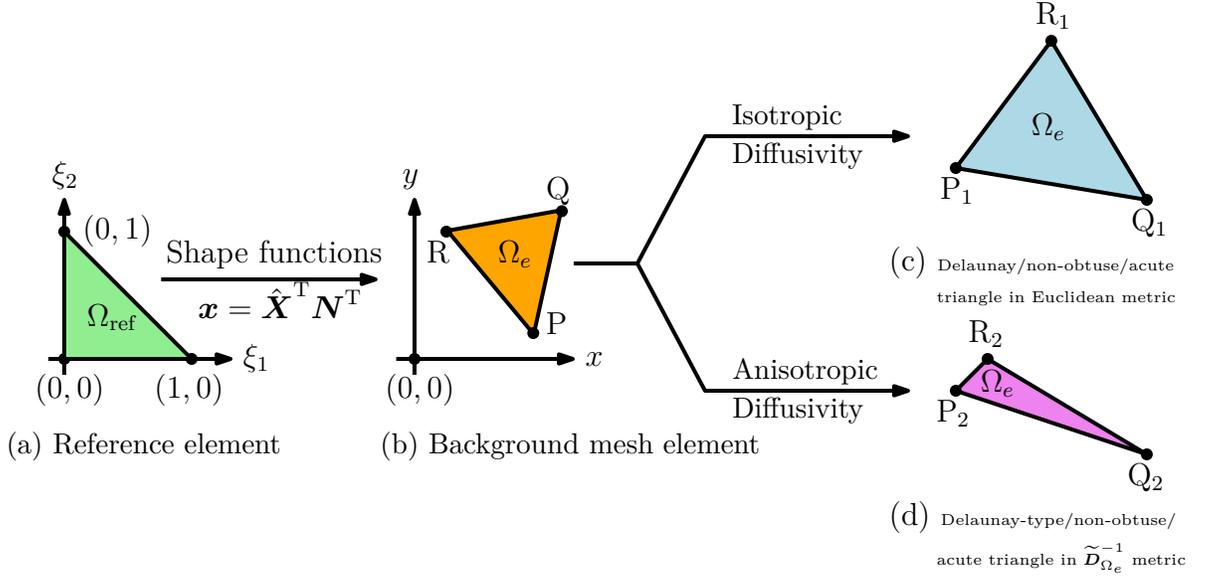


Figure 4.14: DMP-based T3 elements for heterogeneous isotropic and anisotropic diffusivity: A pictorial description of a mesh generation procedure to obtain a new triangulation using a given background mesh.

metric tensor \mathcal{M}_{Ω_e} has to satisfy the following equation in order to meet various DMPs as

$$\mathcal{M}_{\Omega_e} = \Theta_{\Omega_e} \tilde{\mathcal{D}}_{\Omega_e}^{-1}, \quad (4.4.37)$$

where Θ_{Ω_e} is an arbitrary piecewise positive scalar constant, which in general is a user-define parameters. Loosely speaking, an anisotropic \mathcal{M} -uniform simplicial mesh satisfying (4.4.37) satisfies weak DMPs. Furthermore, if the resulting simplicial mesh is *interiorly connected*, then it satisfies strong DMPs.

For a given type of metric tensor (for example, $\mathcal{M}(\mathbf{x})$ given by equation (4.4.37)), open source mesh generators (such as BAMG, BL2D, and Mmg3d) take this as an input and operate on a background mesh to produce an anisotropic \mathcal{M} -uniform simplicial mesh. Algorithm 1 provides a methodology to develop an anisotropic \mathcal{M} -uniform simplicial mesh based on a background mesh, and Figure 4.14 highlights the salient aspects of this algorithm. Nevertheless, it should be noted that Algorithm 1 is a general algorithm to generate DMP-based triangulations (is not limited to either Gmsh

or BAMG) for any type of mesh generator, which operates on a background mesh. On the other hand, there are certain open source and commercially available software packages, such as CGAL cga (2015) and Simmetrix Mes (2015), which create an anisotropic \mathcal{M} -uniform simplicial mesh directly based on the metric tensor Nguyen et al. (2009); Huang et al. (2013) without the need of background meshes. Investigation of such mesh generators is beyond the scope of this research and is neither critical nor central to the ideas discussed here. Before we discuss various numerical examples based on different types of triangulations, we shall now present certain important (mesh-based) non-dimensional numbers relevant to our numerical study. Such a discussion is first of its kind and is not discussed elsewhere.

Remark 4.4.6. *It should be noted that STEP-7 in Algorithm 1 might be computationally intensive, as we need to solve a series of small constrained optimization problem for each ' $\Omega_{e,i} \in \tilde{\mathcal{T}}_{h,i}$ ' and at every iteration level ' i '. The corresponding constrained optimization problems are given as*

$$\|\mathbf{v}\|_{e,i} := \underset{\mathbf{x} \in \bar{\Omega}_{e,i}}{\text{maximize}} \quad \|\mathbf{v}(\mathbf{x})\| \quad \forall 1 \leq i \leq \text{MaxIters} \text{ and} \quad (4.4.38a)$$

$$\|\alpha\|_{e,i} := \underset{\mathbf{x} \in \bar{\Omega}_{e,i}}{\text{maximize}} \quad \alpha(\mathbf{x}) \quad \forall 1 \leq i \leq \text{MaxIters}. \quad (4.4.38b)$$

As $\bar{\Omega}_{e,i}$ is convex, triangular, and closed, by half-space representation theorem for convex polytopes (Boyd and Vandenberghe, 2004, Section-2.2.4) equation (4.4.38a) can be written as

$$\underset{\mathbf{x} \in \mathbb{R}^2}{\text{maximize}} \quad \|\mathbf{v}(\mathbf{x})\| \text{ and} \quad (4.4.39a)$$

$$\text{subject to} \quad \mathbf{A}_{e,i}\mathbf{x} \preceq \mathbf{b}_{e,i} \quad \forall 1 \leq i \leq \text{MaxIters}, \quad (4.4.39b)$$

Algorithm 1 An iterative method to generate an anisotropic \mathcal{M} -uniform mesh satisfying discrete principles

- 1: INPUT: Background mesh ($\mathcal{T}_{h,0}$, N_{ele_0} , Nv_0 , and Nbv_0); anisotropic diffusivity tensor ($\mathbf{D}(\mathbf{x})$); velocity vector field ($\mathbf{v}(\mathbf{x})$); linear reaction coefficient ($\alpha(\mathbf{x})$); maximum number of iterations (MaxIters); piecewise positive scalar element metric constants ($\{\Theta_e\}_{e=1}^{N_{ele_0}}$); and a stopping criteria (StopCrit)
 - $\mathcal{T}_{h,0}$ is the initial background triangulation on which an anisotropic mesh generator operates
 - Nv_0 and Nbv_0 are correspondingly the total number of vertices and boundary vertices
 - 2: Set the iteration number: $i = 0$
 - 3: **while** (True) **do**
 - 4: Compute the element average anisotropic diffusivity tensor using a quadrature rule (for example, see Reference Zienkiewicz et al. (2013))
 - $\tilde{\mathbf{D}}_{e,i} := \frac{1}{\text{meas}(\Omega_e)} \int_{\Omega_e} \mathbf{D}(\mathbf{x}) d\Omega \quad \forall e = 1, 2, \dots, N_{ele_i}$
 - 5: Compute the element metric tensor by explicitly inverting $\tilde{\mathbf{D}}_{e,i}$
 - $\mathcal{M}_{e,i} := \Theta_e (\tilde{\mathbf{D}}_{e,i})^{-1} \quad \forall e = 1, 2, \dots, N_{ele_i}$
 - 6: Based on the set of metric tensors $\{\mathcal{M}_{e,i}\}_{e=1}^{N_{ele_i}}$, compute a new triangulation $\tilde{\mathcal{T}}_{h,i}$
 - Output the new triangulation $\tilde{\mathcal{T}}_{h,i}$. Corresponding to this $\tilde{\mathcal{T}}_{h,i}$, we have \tilde{N}_{ele_i} , $\tilde{N}v_i$, and $\tilde{N}bv_i$
 - 7: Compute the following quantities: $\forall e = 1, 2, \dots, \tilde{N}_{ele_i}$
 - $\tilde{\mathbf{D}}_{e,i}$; $\Lambda_{\min, \tilde{\mathbf{D}}_{e,i}}$; $\|\mathbf{v}\|_{e,i} := \|\mathbf{v}(\mathbf{x})\|_{\infty, \tilde{\Omega}_{e,i}}$; and $\|\alpha\|_{e,i} := \|\alpha(\mathbf{x})\|_{\infty, \tilde{\Omega}_{e,i}}$
 - Need to use a constrained optimization methodology to calculate $\|\mathbf{v}\|_{e,i}$ and $\|\alpha\|_{e,i}$ (see Remark 4.4.6 for more details)
 - 8: **if** ($\text{StopCrit} = \text{Anisotropic non-obtuse angle condition}$) **then**
 - 9: Check the inequality given by equation (4.4.15) in Theorem 4.4.3 $\forall e = 1, 2, \dots, \tilde{N}_{ele_i}$
 - 10: **if** (true) **then**
 - 11: OUTPUT: The triangulation $\tilde{\mathcal{T}}_{h,i}$ and corresponding $\{\mathcal{M}_{e,i}\}_{e=1}^{N_{ele_i}}$. EXIT
 - 12: **else**
 - 13: Update $\mathcal{T}_{h,i} \leftarrow \tilde{\mathcal{T}}_{h,i}$, $N_{ele_i} \leftarrow \tilde{N}_{ele_i}$, $Nv_i \leftarrow \tilde{N}v_i$, $Nbv_i \leftarrow \tilde{N}bv_i$, and $i \leftarrow (i + 1)$
 - 14: **end if**
 - 15: **end if**
 - 16: **if** ($\text{StopCrit} = \text{Generalized Delaunay-type angle condition}$) **then**
 - 17: Check the inequality given by equation (4.4.17) in Theorem 4.4.4 $\forall e = 1, 2, \dots, \tilde{N}_{ele_i}$
 - 18: **if** (true) **then**
 - 19: OUTPUT: The triangulation $\tilde{\mathcal{T}}_{h,i}$ and corresponding $\{\mathcal{M}_{e,i}\}_{e=1}^{N_{ele_i}}$. EXIT
 - 20: **else**
 - 21: Update $\mathcal{T}_{h,i} \leftarrow \tilde{\mathcal{T}}_{h,i}$, $N_{ele_i} \leftarrow \tilde{N}_{ele_i}$, $Nv_i \leftarrow \tilde{N}v_i$, $Nbv_i \leftarrow \tilde{N}bv_i$, and $i \leftarrow (i + 1)$
 - 22: **end if**
 - 23: **end if**
 - 24: **if** ($i > \text{MaxIters}$) **then**
 - 25: OUTPUT: The *existing* triangulation $\tilde{\mathcal{T}}_{h,i}$ and corresponding $\{\mathcal{M}_{e,i}\}_{e=1}^{N_{ele_i}}$.
 - 26: Anisotropic \mathcal{M} -uniform triangulation not found in MaxIters . EXIT
 - 27: **end if**
 - 28: **end while**
-

where $\mathbf{A}_{e,i}$ is a 3×2 matrix and $\mathbf{b}_{e,i}$ is a 3×1 vector, whose coefficients correspond to the linear inequalities defining the relevant half-spaces and supporting hyperplanes of the triangle $\bar{\Omega}_{e,i}$. If the element maximum value $\|\mathbf{v}\|_{e,i}$ is known a priori (through analytically or by means of a rigorous mathematical analysis), then one can use such information in STEP-7 of Algorithm 1. Otherwise, we need to solve equations (4.4.39a)–(4.4.39b) using the standard constrained optimization algorithms for small-scale problems Boyd and Vandenberghe (2004). Similarly, equation (4.4.38b) can be reformulated based on the lines of equations (4.4.39a)–(4.4.39b).

4.4.3.1 Péclet and Damköhler numbers for simplicial meshes

Herein, we shall describe three types of Péclet and Damköhler numbers for simplicial meshes. They are devised based on Theorems 4.4.3 and 4.4.4. However, it should be noted that extending it to non-simplicial elements, such as Q4, is not straightforward. This is because in order to construct Péclet and Damköhler numbers for non-simplicial meshes, one needs to obtain mesh restrictions using the global stiffness restriction method. This is beyond the scope of the current chapter.

- *Element Péclet and Damköhler numbers:* Based on Theorem 4.4.3, one can define the following mesh-based non-dimensional element Péclet ($\mathbb{P}e_{\Omega_e}$) and Damköhler ($\mathbb{D}a_{\Omega_e}$) numbers:

$$\mathbb{P}e_{\Omega_e} := \frac{h_{\max, \Omega_e} \|\mathbf{v}\|_{\infty, \bar{\Omega}_e}}{\Lambda_{\min, \tilde{\mathcal{D}}_{\Omega_e}}} \quad \mathbb{D}a_{\Omega_e} := \frac{h_{\max, \Omega_e} h_{\text{pumax}, \Omega_e} \|\alpha\|_{\infty, \bar{\Omega}_e}}{\Lambda_{\min, \tilde{\mathcal{D}}_{\Omega_e}}}, \quad (4.4.40)$$

where the height $h_{\text{pumax}, \Omega_e}$ is given as follows:

$$0 < h_1 \leq h_2 \leq \dots \leq h_i \leq \dots \leq h_{\text{pumax}, \Omega_e} \leq h_{\max, \Omega_e} \\ \forall i = 1, 2, \dots, nd + 1, \quad \Omega_e \in \mathcal{T}_h. \quad (4.4.41)$$

Correspondingly, using equations (4.4.40) in equation (4.4.15) gives the following (stronger) mesh restriction condition based on Theorem 4.4.3:

$$0 < \frac{\mathbb{P}e_{\Omega_e}}{(nd+1) \cos(\beta_{ij, \tilde{\mathcal{D}}_{\Omega_e}^{-1}})} + \frac{\mathbb{D}a_{\Omega_e}}{(nd+1)(nd+2) \cos(\beta_{ij, \tilde{\mathcal{D}}_{\Omega_e}^{-1}})} \leq 1$$

$$i = \max \text{ and } j = \text{pumax}, i \neq j, \forall \Omega_e \in \mathcal{T}_h. \quad (4.4.42)$$

- *Edge Péclet and Damköhler numbers:* In a similar fashion, utilizing Theorem 4.4.4, one can define the following mesh-based non-dimensional edge Péclet ($\mathbb{P}e_{\Omega_e, e_{pq}}$) and Damköhler ($\mathbb{D}a_{\Omega_e, e_{pq}}$) numbers:

$$\mathbb{P}e_{\Omega_e, e_{pq}} := \frac{\text{meas}(\Omega_e) \|\mathbf{v}\|_{\infty, \bar{\Omega}_e}}{h_{q, \Omega_e} \sqrt{\det[\tilde{\mathcal{D}}_{\Omega_e}]}} \quad \mathbb{D}a_{\Omega_e, e_{pq}} := \frac{\text{meas}(\Omega_e) \|\alpha\|_{\infty, \bar{\Omega}_e}}{\sqrt{\det[\tilde{\mathcal{D}}_{\Omega_e}]}}. \quad (4.4.43)$$

Correspondingly, using equations (4.4.43) in equation (4.4.17) gives the following (weaker) mesh restriction condition based on Theorem 4.4.4:

$$0 < \frac{1}{2\pi} \left[\beta_{pq, \tilde{\mathcal{D}}_{\Omega_e}^{-1}} + \beta_{pq, \tilde{\mathcal{D}}_{\Omega'_e}^{-1}} \right]$$

$$+ \frac{1}{2\pi} \text{arccot} \left(\sqrt{\frac{\det[\tilde{\mathcal{D}}_{\Omega'_e}]}{\det[\tilde{\mathcal{D}}_{\Omega_e}]}} \left(\cot(\beta_{pq, \tilde{\mathcal{D}}_{\Omega'_e}^{-1}}) - \frac{2\mathbb{P}e_{\Omega'_e, e_{pq}}}{3} - \frac{\mathbb{D}a_{\Omega'_e, e_{pq}}}{6} \right) - \frac{2\mathbb{P}e_{\Omega_e, e_{pq}}}{3} - \frac{\mathbb{D}a_{\Omega_e, e_{pq}}}{6} \right)$$

$$+ \frac{1}{2\pi} \text{arccot} \left(\sqrt{\frac{\det[\tilde{\mathcal{D}}_{\Omega_e}]}{\det[\tilde{\mathcal{D}}_{\Omega'_e}]} \left(\cot(\beta_{pq, \tilde{\mathcal{D}}_{\Omega_e}^{-1}}) - \frac{2\mathbb{P}e_{\Omega_e, e_{pq}}}{3} - \frac{\mathbb{D}a_{\Omega_e, e_{pq}}}{6} \right) - \frac{2\mathbb{P}e_{\Omega'_e, e_{pq}}}{3} - \frac{\mathbb{D}a_{\Omega'_e, e_{pq}}}{6} \right)$$

$$\leq 1. \quad (4.4.44)$$

- *Global mesh Péclet and Damköhler numbers:* On sufficiently fine h -refined simplicial meshes (which confirm to Theorem 4.4.3), one can define a global mesh Péclet ($\mathbb{P}e_h$) and Damköhler ($\mathbb{D}a_h$) numbers by modifying equations (4.4.40) as

$$\mathbb{P}e_h := \frac{h \max_{\Omega_e \in \mathcal{T}_h} [\|\mathbf{v}\|_{\infty, \bar{\Omega}_e}]}{\min_{\Omega_e \in \mathcal{T}_h} [\Lambda_{\min, \tilde{\mathcal{D}}_{\Omega_e}]}} \quad \mathbb{D}a_h := \frac{h^2 \max_{\Omega_e \in \mathcal{T}_h} [\|\alpha\|_{\infty, \bar{\Omega}_e}]}{\min_{\Omega_e \in \mathcal{T}_h} [\Lambda_{\min, \tilde{\mathcal{D}}_{\Omega_e}]}}. \quad (4.4.45)$$

Conservatively, equation (4.4.42) can be modified to give a (stronger) global mesh restriction condition based on Theorem 4.4.3:

$$0 < \frac{\mathbb{P}e_h}{(nd+1) \min_{\Omega_e \in \mathcal{T}_h} \left[\cos(\beta_{ij, \tilde{D}_{\Omega_e}^{-1}}) \right]} + \frac{\mathbb{D}a_h}{(nd+1)(nd+2) \min_{\Omega_e \in \mathcal{T}_h} \left[\cos(\beta_{ij, \tilde{D}_{\Omega_e}^{-1}}) \right]} \leq 1. \quad (4.4.46)$$

$i = \max$ and $j = \text{pmax}$, $i \neq j$.

One should note that equation (4.4.46) can provide a useful a priori (conservative) estimate on h for constructing highly refined simplicial meshes.

4.4.3.2 *Physics-based Péclet and Damköhler numbers*

For isotropic diffusivity, it is well-known that the following three *physics-based* non-dimensional numbers can be used to understand the qualitative nature of the solutions Gresho and Sani (2000); Donea and Huerta (2003):

$$\mathbb{P}e_D := \frac{\|\mathbf{v}(\mathbf{x})\|_\infty L}{\min_{x \in \Omega} [D(\mathbf{x})]} \quad \text{Péclet number,} \quad (4.4.47a)$$

$$\mathbb{D}a_I := \frac{\|\alpha(\mathbf{x})\|_\infty L}{\|\mathbf{v}(\mathbf{x})\|_\infty} \quad \text{Damköhler number of first kind, and} \quad (4.4.47b)$$

$$\mathbb{D}a_{II,D} := \frac{\|\alpha(\mathbf{x})\|_\infty L^2}{\min_{x \in \Omega} [D(\mathbf{x})]} \quad \text{Damköhler number of second kind,} \quad (4.4.47c)$$

where L is the characteristic length of the domain and $\|\bullet\|_\infty$ is the standard max-norm/infinity-norm for vectors. However, it should be noted that the above three non-dimensional numbers are not independent, as they satisfy the relation $\mathbb{D}a_{II,D} = \mathbb{P}e_D \mathbb{D}a_I$. Physically, the non-dimensional Péclet number characterizes the relative dominance of advection as compared to diffusion processes. For larger Péclet numbers, the advection process dominates and for smaller Péclet numbers, the diffusion process dominates.

In the literature on chemically reacting systems (e.g., see (Chung, 2010, Table

22.2.1)), there exists various Damköhler numbers that relate progress of chemical reactions with respect to mixing, diffusivity, reaction coefficient, thermal effects, and advection. The Damköhler number of first-kind, $\mathbb{D}a_I$, gives information about the relative influence of a linear reaction coefficient to that of advection. For small values of $\mathbb{D}a_I$, advection progresses much faster than decay of the chemical species and has the opposite effect for large values of the number. The Damköhler number of second kind, $\mathbb{D}a_{II,D}$, gives information related to the progress of chemical reaction with respect to diffusion. Based on the chemical system under consideration, these three non-dimensional numbers dictate how the reaction, advection, and diffusion interact with each other. On the other hand, one should note that extending it to anisotropic diffusivity is not straightforward. Motivated by equations (4.4.45), a way to define *physics-based* non-dimensional numbers for anisotropic diffusivity tensor $\mathbf{D}(\mathbf{x})$ is as

$$\mathbb{P}e_{\mathbf{D}} := \frac{\|\mathbf{v}(\mathbf{x})\|_{\infty} L}{\min_{\mathbf{x} \in \bar{\Omega}} [\Lambda_{min, \mathbf{D}(\mathbf{x})}]} \quad \text{Péclet number and} \quad (4.4.48a)$$

$$\mathbb{D}a_{II, \mathbf{D}} := \frac{\|\alpha(\mathbf{x})\|_{\infty} L^2}{\min_{\mathbf{x} \in \bar{\Omega}} [\Lambda_{min, \mathbf{D}(\mathbf{x})}]} \quad \text{Damköhler number of second kind,} \quad (4.4.48b)$$

where $\Lambda_{min, \mathbf{D}(\mathbf{x})}$ is the minimum eigenvalue of $\mathbf{D}(\mathbf{x})$ at a given point \mathbf{x} . The Damköhler number of first kind, $\mathbb{D}a_I$, given by equation (4.4.47b) remains the same as it does not depend on diffusivity tensor $\mathbf{D}(\mathbf{x})$. Alternatively, inspired by equations (4.4.43), one can define a different set of physics-based Péclet and Damköhler numbers. These

are given as

$$\mathbb{P}e_{\mathbf{D}} := \frac{\|\mathbf{v}(\mathbf{x})\|_{\infty} L}{\sqrt{\min_{\mathbf{x} \in \Omega} [\Lambda_{min, \mathbf{D}(\mathbf{x})}] \max_{\mathbf{x} \in \Omega} [\Lambda_{max, \mathbf{D}(\mathbf{x})}]}} \quad \text{Péclet number and} \quad (4.4.49a)$$

$$\mathbb{D}a_{II, \mathbf{D}} := \frac{\|\alpha(\mathbf{x})\|_{\infty} L^2}{\sqrt{\min_{\mathbf{x} \in \Omega} [\Lambda_{min, \mathbf{D}(\mathbf{x})}] \max_{\mathbf{x} \in \Omega} [\Lambda_{max, \mathbf{D}(\mathbf{x})}]}} \quad \text{Damköhler number of second kind,} \quad (4.4.49b)$$

where $\Lambda_{max, \mathbf{D}(\mathbf{x})}$ is the maximum eigenvalue of $\mathbf{D}(\mathbf{x})$ at a given point \mathbf{x} . One should note that both these sets of non-dimensional numbers (given by equations (4.4.48a)–(4.4.48b) and (4.4.49a)–(4.4.49b)) are perfectly valid, as anisotropy in diffusivity tensor can introduce multiple ways of defining physics-based Péclet and Damköhler numbers. Now, we shall present various numerical examples to demonstrate the *pros and cons* of the mesh restrictions approach.

4.4.3.3 Test problem #1: Transport in fractured media

This test problem has profound impact in simulating the transport of chemical species in fractured media Therrien and Sudicky (1996). The numerical simulations, using the Delaunay-Voronoi triangulation (with `MaxIters` = 50) for different cases of diffusivity, velocity field, and linear reaction coefficient, are presented in Figure 4.15. Homogeneous Dirichlet boundary conditions are prescribed on the sides of the fractured domain; $c^p(\mathbf{x}) = 1.0$ on the left set of fracture lines and $c^p(\mathbf{x}) = 2.5$ on the right set of fracture lines. The volumetric source $f(\mathbf{x})$ is zero inside the fractured domain. Diffusivity is assumed to isotropic and scalar, whose value is given by $D(\mathbf{x}) = 10^{-3}$. We perform numerical simulations for three different cases of velocity field and linear reaction coefficient, which are given by $\mathbf{v}(\mathbf{x}) = (0.0, 0.0)$ and $\alpha(\mathbf{x}) = 0.0$, $\mathbf{v}(\mathbf{x}) = (0.1, 1.0)$ and $\alpha(\mathbf{x}) = 0.0$, and $\mathbf{v}(\mathbf{x}) = (0.1, 1.0)$ and $\alpha(\mathbf{x}) = 1.0$.

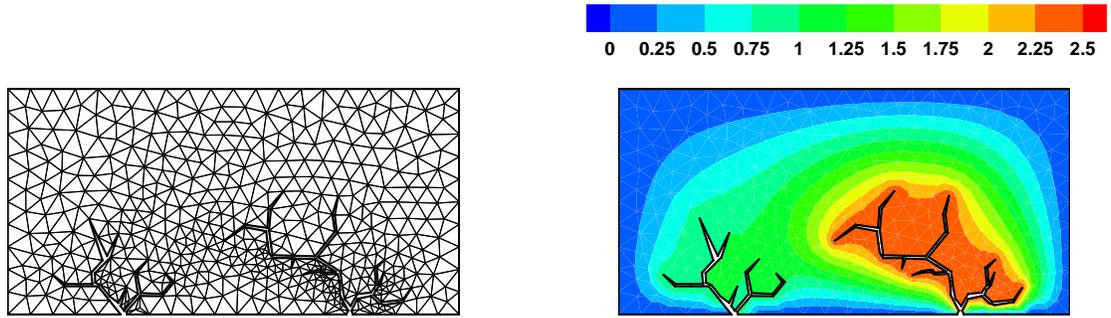
From Figure 4.15, it can be inferred that we need highly refined DMP-based

triangular meshes to obtain physically meaningful values for concentration for large values of edge Péclet and Damköhler numbers. The white region in the figures indicates the area in which the value of concentration is negative and also violated the maximum constraint. The coarse Delaunay-Voronoi mesh obtained using the open source mesh generator `Gmsh` satisfies NC and DMPs in the case of pure diffusion. But, this is not true for AD and ADR cases. In such scenarios, it produces unphysical values for the concentration field. Moreover, the percentage of nodes that have violated NC and maximum constraint is also very high.

Quantitatively, Tables 4.1 and 4.2 provide more details pertinent to the violations in NC and DMPs for AD and ADR cases. It should be noted that the decrease in unphysical values of concentration is not monotonic. This is because the **generalized Delaunay-type angle condition**, in general, *does not ensure uniform convergence* for diffusion-type equations in L^∞ norm (for example, in case of pure isotropic diffusion, see the mesh restriction result by Ciarlet and Raviart Ciarlet and Raviart (1973)). Figure 4.16 shows that the weak DMP-based condition is satisfied only for pure diffusion equation. But in all other cases, **generalized Delaunay-type condition** is violated. In the case of pure diffusion, it should be noted that DMP-based mesh given in Figure 4.16 is *not* interiorly connected. Hence, it only satisfies $DWMP_K$, but not $DSMP_K$.

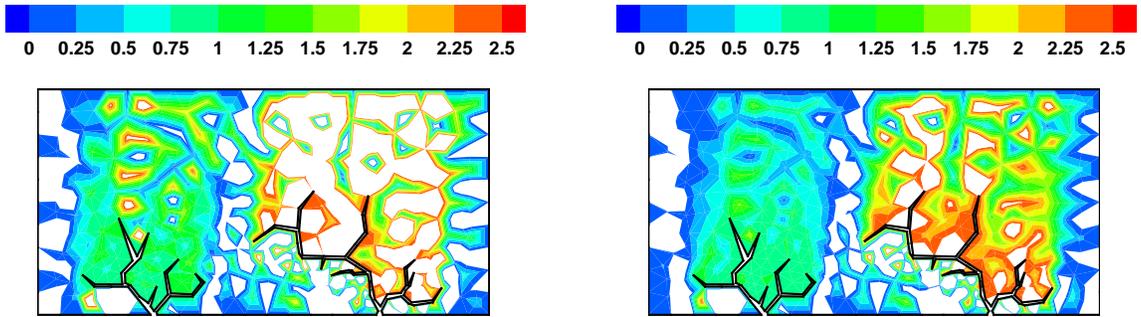
4.4.3.4 *Test problem #2: Species dispersion in subsurface flows*

A pictorial description of the boundary value problem with various parameters is shown in Figure 4.17. Homogeneous Dirichlet boundary conditions are prescribed on all sides of the square. The volumetric source $f(\mathbf{x})$ is zero inside the domain, except for the square region (including the boundaries) located at vertex $H = (0.375, 0.375)$. In this region, $f(\mathbf{x})$ is equal to unity. Herein, we assume that the velocity vector field and linear reaction coefficient are equal to zero everywhere in the computational



(a) Delaunay-Voronoi mesh: $Nv = 539$ and $Nele = 906$

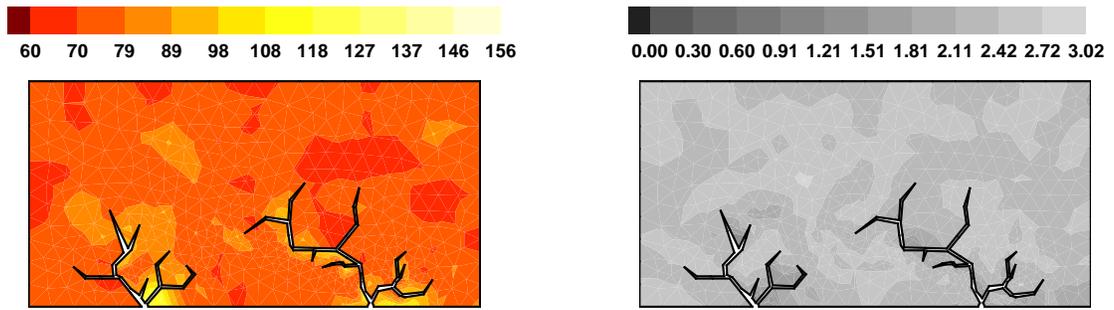
(b) $\mathbf{v} = (0, 0)$ and $\alpha = 0$



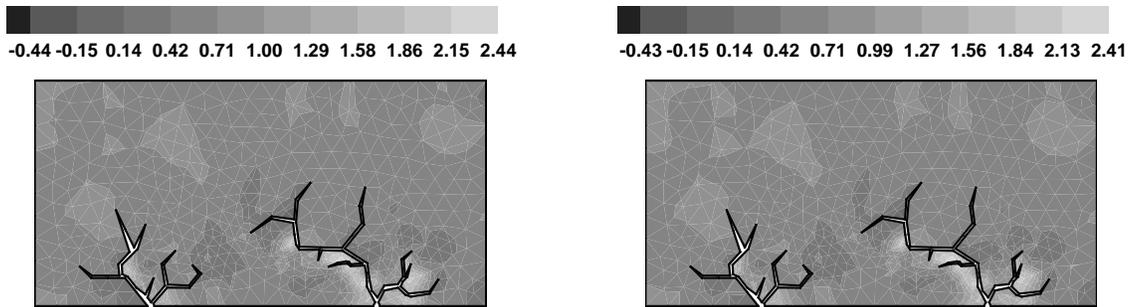
(c) $\mathbf{v} = (0.1, 1.0)$ and $\alpha = 0$

(d) $\mathbf{v} = (0.1, 1.0)$ and $\alpha = 1.0$

Figure 4.15: Test problem #1: The top left figure shows a coarse triangulation employed in the numerical study, *which is to the scale*. The top right figure and the bottom two figures show the concentration profiles obtained using this mesh.



(a) Element maximum angles: $Nv = 539$ and (b) Delaunay-type condition: $\mathbf{v} = (0, 0)$ and $\alpha = 0$
 $N_{ele} = 906$



(c) Delaunay-type condition: $\mathbf{v} = (0.1, 1.0)$ and $\alpha = 0$ (d) Delaunay-type condition: $\mathbf{v} = (0.1, 1.0)$ and $\alpha = 1.0$

Figure 4.16: Test problem #1: The top left figure shows the maximum angle possible in each element of the mesh. The top right figure and the bottom two figures show the *element maximum generalized Delaunay-type condition*.

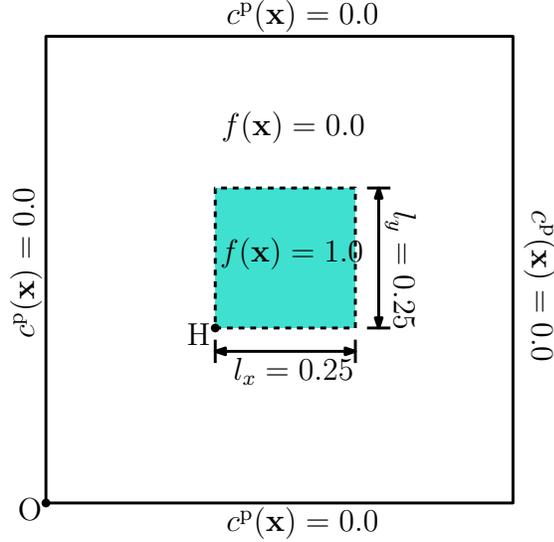


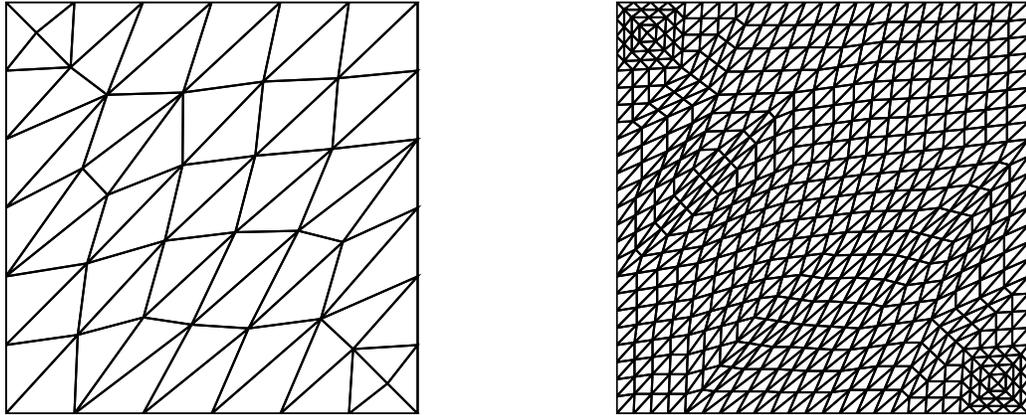
Figure 4.17: Test problem #2: The computational domain under consideration is a bi-unit square with one of its vertices at origin $O = (0, 0)$.

domain. Relevant (coarse) background mesh used and the corresponding DMP-based mesh (with `MaxIters` = 50) obtained using `BAMG` are shown in Figure 4.18. The diffusivity tensor for this problem is taken from the subsurface hydrology literature Pinder and Celia (2006) and is given as

$$\mathbf{D}_{\text{subsurface}}(\mathbf{x}) = \alpha_T \|\mathbf{v}\| \mathbf{I} + \frac{\alpha_L - \alpha_T}{\|\mathbf{v}\|} \mathbf{v} \otimes \mathbf{v}, \quad (4.4.50)$$

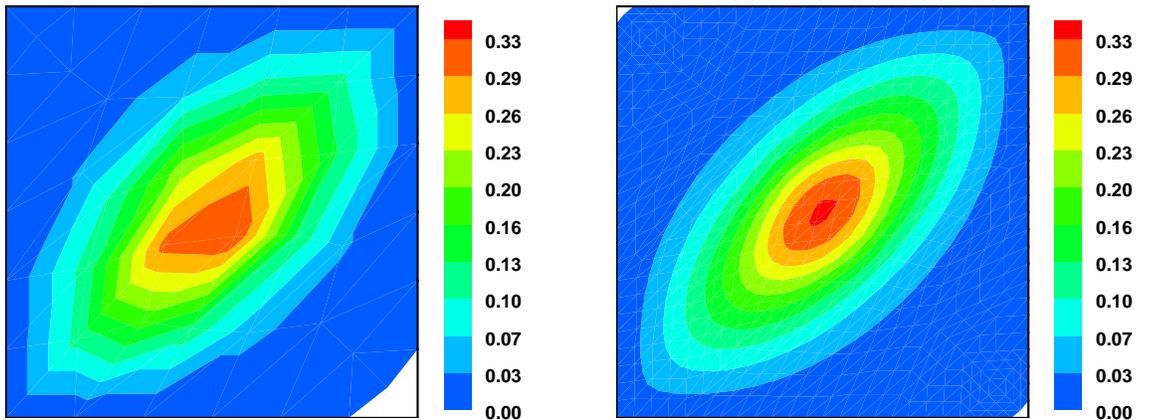
where \otimes is the tensor product, \mathbf{I} is the identity tensor, \mathbf{v} is velocity vector field of the subsurface flow, and α_T and α_L are, respectively, transverse and longitudinal diffusivity coefficients with $\alpha_T = 0.01$ and $\alpha_L = 0.1$. It should be emphasized that we have neglected advection. Correspondingly, the numerical values for the velocity vector field used to define the diffusion tensor is given by $\mathbf{v}(\mathbf{x}) = (1.0, 1.0)$. This test problem has importance in simulating diffusion of chemical species in subsurface flows of hydrogeological systems Dentz et al. (2011).

Numerical simulations using these meshes are shown in Figure 4.19. The white region in this figure depicts the area in which the value of concentration is negative. From Figure 4.19, it is apparent that the coarse anisotropic triangulation violates



(a) Background mesh: $Nv = 47$ and $Nele = 68$ (b) Anisotropic mesh: $Nv = 593$ and $Nele = 1088$

Figure 4.18: Test problem #2: The left figure shows the background mesh on which BAMG operates to give an anisotropic triangulation, which is shown in the right figure.



(a) Background mesh: $\mathbf{v} = (0, 0)$ and $\alpha = 0$ (b) Anisotropic mesh: $\mathbf{v} = (0, 0)$ and $\alpha = 0$

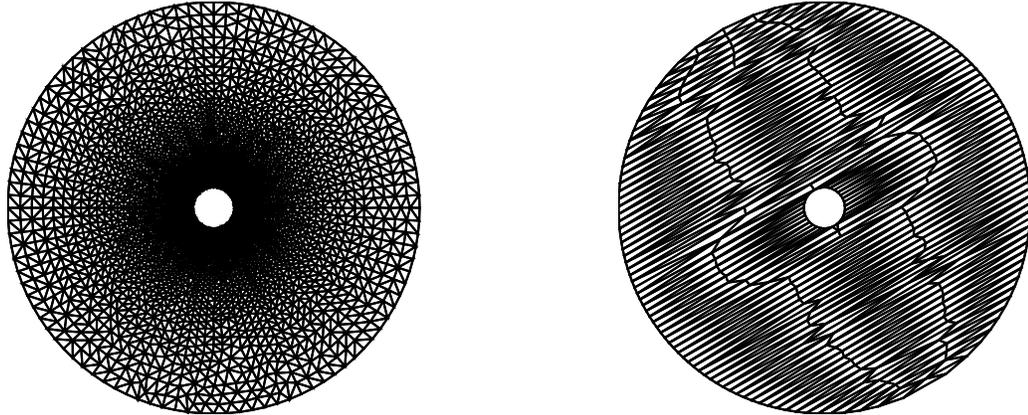
Figure 4.19: Test problem #2: The left figure shows the concentration profile based on the background mesh, while the right figure shows the concentration profile using the anisotropic triangulation.

the DMPs and NC. This is because the Algorithm 1 did not converge in `MaxIters`. However, quantitatively, this violation in NC is low as compared to background mesh. Specifically, the minimum concentration and the percentage of nodes that have violated the non-negative constraint on the background mesh is about -4.8×10^{-5} and 2.13%, while these values on the anisotropic triangulation are around -1.35×10^{-8} and 0.34%. Additionally, from Figure 4.19, it is evident that we need a highly refined DMP-based anisotropic mesh to avoid negative values for concentration. Nevertheless, it should be noted that there is a considerable decrease in the negative values for concentration if a traditionally h -refined anisotropic triangulation is used (see Figure 4.23 and subsection 4.4.3.6 for more details).

4.4.3.5 Test problem #3: Contaminant transport in leaky wells

The computational domain is a circle with a hole centered at origin $(0, 0)$. The radius of the circular hole and the circular domain are 0.1 and 1.0. Numerical simulations are performed for four different cases of the velocity vector field and linear reaction coefficient, which are given by $\mathbf{v}(\mathbf{x}) = (0.0, 0.0)$ and $\alpha(\mathbf{x}) = 0.0$, $\mathbf{v}(\mathbf{x}) = (1.5, 1.0)$ and $\alpha(\mathbf{x}) = 1.0$, $\mathbf{v}(\mathbf{x}) = (5.0, 0.5)$ and $\alpha(\mathbf{x}) = 1.0$, and $\mathbf{v}(\mathbf{x}) = (0.0, 0.0)$ and $\alpha(\mathbf{x}) = 1000$. Each case is designed to test a particular aspect. For example, $\mathbf{v}(\mathbf{x}) = (5.0, 0.5)$ and $\alpha(\mathbf{x}) = 1.0$ corresponds to advection-dominated ADR problems, while $\mathbf{v}(\mathbf{x}) = (0.0, 0.0)$ and $\alpha(\mathbf{x}) = 1000$ corresponds to reaction-dominated diffusion-reaction problems. The diffusivity tensor for this problem is given by equations (4.1.1)–(4.1.2a). Correspondingly, the values for the parameters d_{\max} , d_{\min} , and θ are equal to 1, 0.001, and $\frac{\pi}{3}$.

The computational domain in this test problem has various practical applications related to well design in multi-aquifers and groundwater distribution systems Konikow and Hornberger (2006), understanding emanation of gaseous hydrocarbons through bore-holes in oil and gas reservoirs Myers (2012), and to study leakage of contaminants

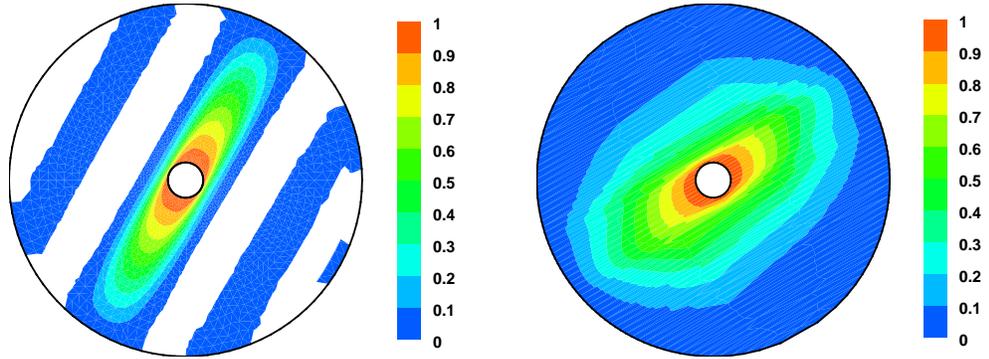


(a) Background mesh: $Nv = 5079$ and $Nele = 9918$ (b) Anisotropic mesh: $Nv = 297$ and $Nele = 436$

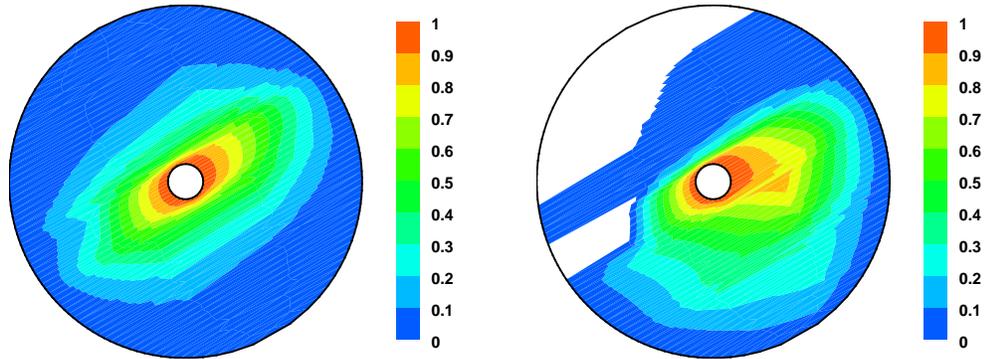
Figure 4.20: Test problem #3: The left figure shows the background mesh and the right figure shows the anisotropic triangulation obtained using BAMG for all the four cases.

(such as CO_2 , salts, and nitrates) through abandoned wells Avci (1994); Lacombe et al. (1995); Ebigbo et al. (2007); Nakshatrala and Turner (2013). The background mesh used and the corresponding DMP-based (coarse) mesh obtained using BAMG are shown in Figure 4.20 ($\text{MaxIters} = 50$). For the cases when $\mathbf{v}(\mathbf{x}) = (0.0, 0.0)$ and $\alpha(\mathbf{x}) = 0.0$, $\mathbf{v}(\mathbf{x}) = (1.5, 1.0)$ and $\alpha(\mathbf{x}) = 1.0$, Algorithm 1 converged in MaxIters . But for $\mathbf{v}(\mathbf{x}) = (5.0, 0.5)$ and $\alpha(\mathbf{x}) = 1.0$, $\mathbf{v}(\mathbf{x}) = (0.0, 0.0)$ and $\alpha(\mathbf{x}) = 1000$, Algorithm 1 did not converge in MaxIters . Herein, the DMP-based coarse mesh is composed of *needle-type* triangles. This is because the ratio of the minimum eigenvalue of $\mathbf{D}(\mathbf{x})$ to its maximum is 0.001 (which is related to the aspect ratio of the sides of the triangle in the DMP-based anisotropic mesh). Moreover, it is evident that the triangles in the mesh are aligned and oriented along the principal axis of the eigenvectors of the diffusivity tensor.

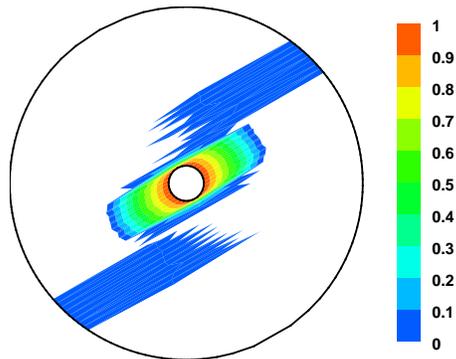
Numerical simulations using these meshes are shown in Figure 4.21. The white region in the figures (circular annulus) shows the area in which the value of concentration is negative. The minimum concentration and the percentage of nodes that have



(a) Background mesh: $\mathbf{v} = (0, 0)$ and $\alpha = 0$ (b) Anisotropic mesh: $\mathbf{v} = (0, 0)$ and $\alpha = 0$



(c) Anisotropic mesh: $\mathbf{v} = (1.5, 1.0)$ and $\alpha = 1.0$ (d) Anisotropic mesh: $\mathbf{v} = (5.0, 0.5)$ and $\alpha = 1.0$



(e) Anisotropic mesh: $\mathbf{v} = (0, 0)$ and $\alpha = 1000$

Figure 4.21: Test problem #3: This figure shows the concentration profiles for four different cases based on the background mesh and anisotropic meshes shown in Figure 4.20.

violated the non-negative constraint for the background mesh are about -1.67×10^{-2} and 30.28%. As the anisotropic mesh is coarse and the Algorithm 1 did not converge in `MaxIters` for advection-dominated ADR problems and reaction-dominated diffusion-reaction problems, the resulting mesh not only violates the non-negative constraint, but also produces spurious oscillations. The minimum concentration and the percentage of nodes that have violated the non-negative constraint for the case when $\mathbf{v} = (5.0, 0.5)$ and $\alpha = 1.0$ are about -1.78×10^{-1} and 13.47%, whereas for the case when $\mathbf{v} = (0.0, 0.0)$ and $\alpha = 1000$ are around -2.79×10^{-1} and 20.54%. Hence in both these cases, we need a highly refined DMP-based mesh to avoid negative values for concentration (see subsection 4.4.3.6 for more details). For scenarios when $\mathbf{v}(\mathbf{x}) = (0.0, 0.0)$ and $\alpha(\mathbf{x}) = 0.0$, $\mathbf{v}(\mathbf{x}) = (1.5, 1.0)$ and $\alpha(\mathbf{x}) = 1.0$, the coarse DMP-based mesh is interiorly connected and satisfies $\text{DsMP}_{\mathbf{K}}$ (or $\text{DSMP}_{\mathbf{K}}$, when $\alpha(\mathbf{x}) = 0$).

4.4.3.6 Issues with DMP-based h -refinement

From the above set of test problems, it is apparent that mesh refinement is needed to avoid spurious node-to-node oscillations and satisfaction of various discrete principles (mainly for the cases when the values of the velocity vector field and linear reaction coefficient are predominant as compared to minimum eigenvalue of diffusivity tensor). In general, within the context of computational geometry and mesh generation literature, there are various methods to generate different types of h -refined meshes George and Frey (2010); Schneider (2013). However, it is evident from these test problems that we are interested in h -refined meshes that conform to either `anisotropic non-obtuse angle condition` or `generalized Delaunay-type angle condition`. Herein, we shall study two different and important methodologies that are popular in mesh generation literature and implemented in the open source

mesh generators, such as Gmsh and BAMG in FreeFem++. Our objective is to understand whether the h -refined meshes generated using these mesh generators satisfy DMPs, DCPs, and NC or not.

Traditional h -refinement

In this method, an h -refined mesh is obtained by splitting each triangle in the coarse mesh into multiple triangles. Based on this methodology, numerical simulations are performed using the traditional h -refined meshes (which are derived based on the DMP-based coarse meshes presented in Figures 4.15, 4.18, and 4.20). The resulting concentration profiles for test problems #1, #2, and #3, are shown in Figures 4.22–4.24. Qualitatively and quantitatively, the following inferences can be drawn from these numerical results:

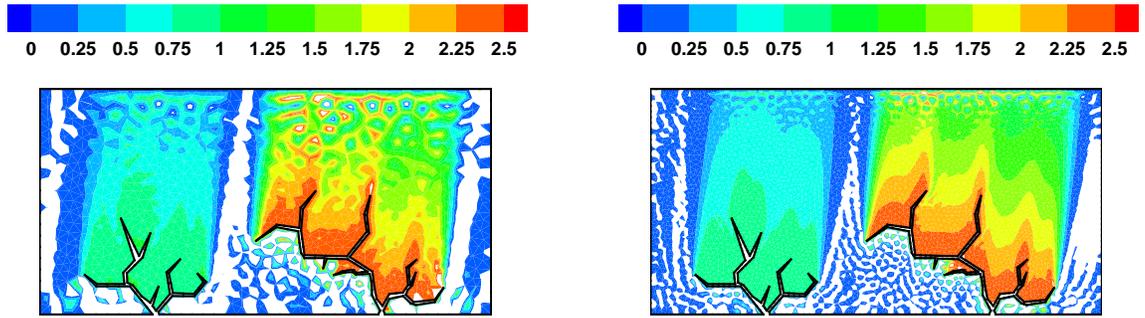
- Test problem #1: For isotropic diffusivity, from Figure 4.22, Table 4.1, and Table 4.2, it is apparent that there is a decrease in negative values for concentration and reduction in spurious oscillations.
- Test problem #2: Qualitatively, based on Figure 4.23, it can be concluded that there is a considerable decrease in the violation of non-negative constraint. Quantitatively, for this h -refined anisotropic triangulation, the minimum concentration and the percentage of nodes that have violated the non-negative constraint is about -1.15×10^{-11} and 0.02% (which is significantly close to machine epsilon $\epsilon_{\text{mach}} \approx 2.22 \times 10^{-16}$).
- Test problem #3: For the pure anisotropic diffusion case, the *coarse* anisotropic DMP-based mesh given in Figure 4.20 satisfies NC and all of the DMPs. However, contrary to this, on traditional h -refinement, the resulting anisotropic triangulation produces unphysical negative values and violates the DMPs. Quantitatively, this violation is far from machine epsilon, ϵ_{mach} . Correspondingly,

the minimum concentration and the percentage of nodes that have violated the non-negative constraint is about -2.7×10^{-1} and 28.51%, which is way higher as compared the numerical simulations based on the background mesh.

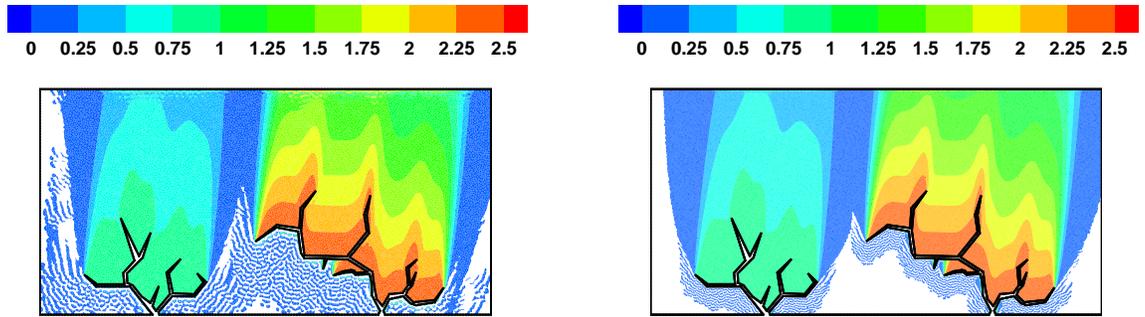
To summarize, *it is clear that traditional h -refinement does not always reduce the unphysical numerical values.* Certainly, there is a decrease in negative values for concentration for test problem #1 and #2. But, this is not the case for test problem #3. This is because the methodology to obtain traditional h -refinement meshes from a given coarse mesh *need not conform* to the conditions outlined in either Theorem 4.4.3 or Theorem 4.4.4.

Non-traditional h -refinement

In this method, a h -refined mesh is obtained *directly* from the background mesh (using Algorithm 1) by change certain parameters related to metric tensor, geometry of the domain, and number of nodes on the boundary of the domain (Hecht et al., 2014, Chapter-5). It should be noted that this methodology is completely different from the traditional h -refinement procedure, as we never generate a coarse DMP-based triangulation and then split the respective mesh elements. However, even this non-traditional h -refined mesh is not guaranteed to produce physics-compatible numerical values for concentration (as the mesh generation procedure need not converge in **MaxIters**). This is evident from the numerical simulations performed on the non-traditional h -refined mesh for test problem #3. The corresponding concentration profile is shown in Figure 4.25. Quantitatively, the minimum concentration and the percentage of nodes that have violated the non-negative constraint are about -9.69×10^{-2} and 34.11%, which is comparatively similar to that of the traditional h -refinement methodology.

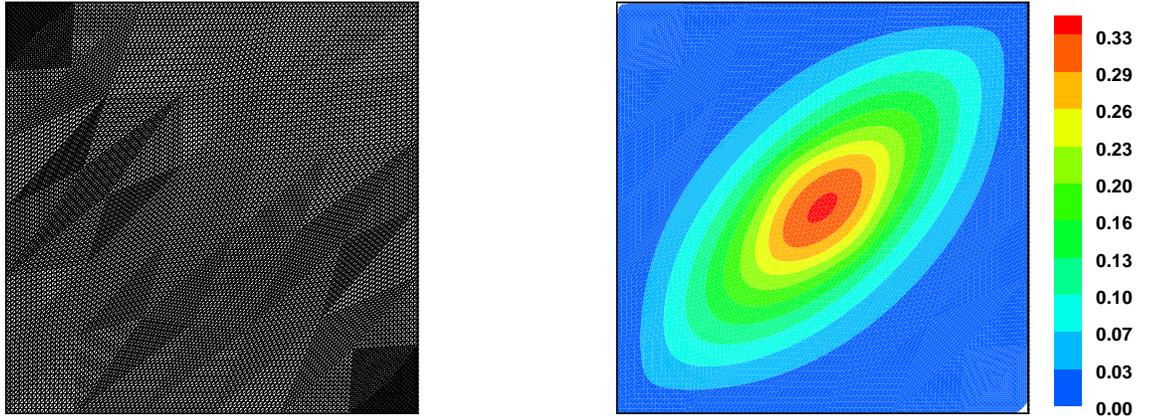


(a) Delaunay mesh: $Nv = 1564$ and $Nele = 2826$ (b) Delaunay mesh: $Nv = 5090$ and $Nele = 9620$



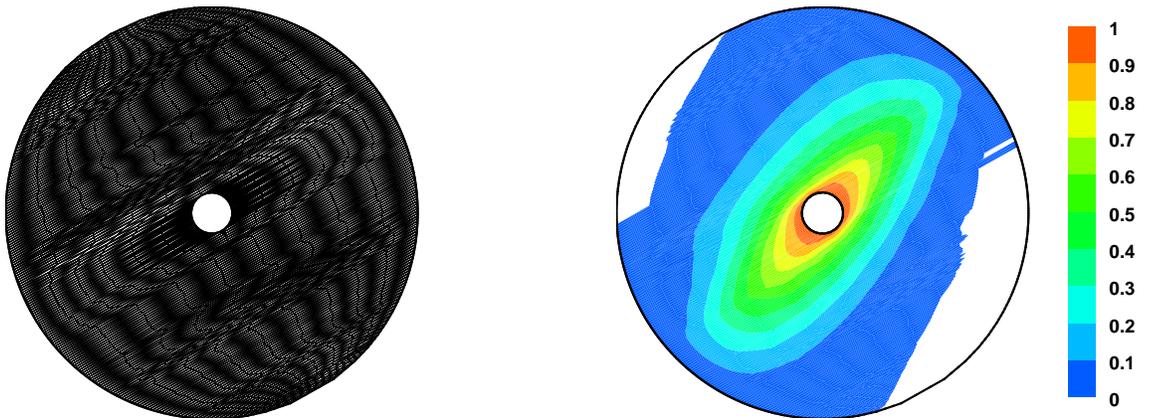
(c) Delaunay mesh: $Nv = 18372$ and $Nele = 35665$ (d) Delaunay mesh: $Nv = 69995$ and $Nele = 137881$

Figure 4.22: Issues with traditional mesh refinement: Concentration profiles for the fracture domain when $\mathbf{v} = (0.1, 1.0)$ and $\alpha = 1.0$. The white region in the figures shows the area in which the numerical simulation has violated the NC and maximum constraint.



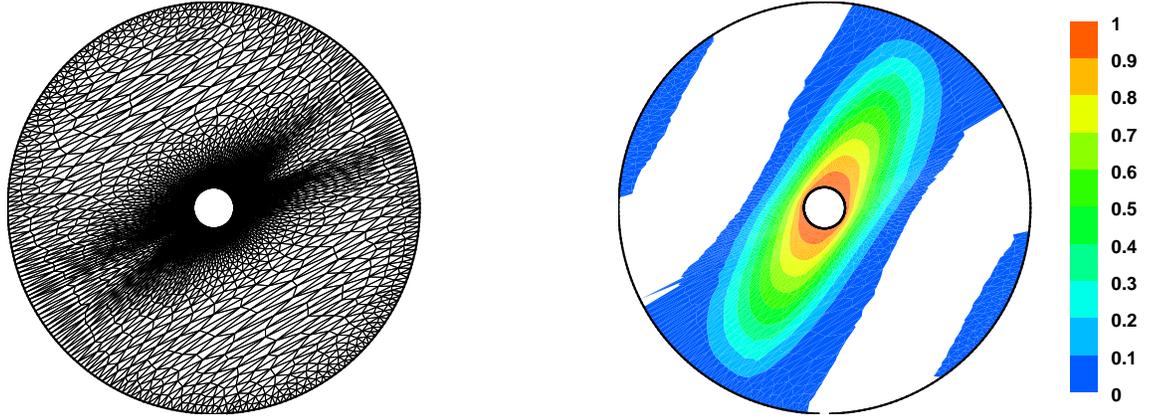
(a) Anisotropic mesh: $Nv = 8897$ and $Nele = 17408$ (b) Pure anisotropic diffusion: Concentration profile

Figure 4.23: Issues with traditional mesh refinement: The left figure shows the anisotropic mesh obtained using the traditional mesh refinement procedure on the anisotropic triangulation given in Figure 4.18. The right figure shows the concentration profile obtained using this refined mesh.



(a) Anisotropic mesh: $Nv = 2199$ and $Nele = 3924$ (b) Pure anisotropic diffusion: Concentration profile

Figure 4.24: Issues with traditional mesh refinement: The left figure shows the anisotropic mesh obtained using the traditional mesh refinement procedure on the anisotropic triangulation given in Figure 4.20. The right figure shows the concentration profile obtained using this refined mesh.



(a) Anisotropic mesh: $Nv = 2647$ and $Nele = 4789$ (b) Pure anisotropic diffusion: Concentration profile

Figure 4.25: Issues with non-traditional mesh refinement: The left figure shows a refined anisotropic mesh obtained using the non-traditional approach. The right figure shows the concentration profile obtained using this refined mesh (did not converge in $MaxIters = 100$).

Table 4.1: Fractured domain with isotropic diffusivity: For AD equation

Delaunay triangulation ($Nv, Nele, h$)	Concentration		% of nodes violated	
	Minimum	Maximum	(NC)	(DMP)
(539, 906, 0.1)	-10.51	9.85	17.44	13.73
(1564, 2826, 0.05)	-18.26	5.86	18.22	14.19
(5090, 9620, 0.025)	-4.65	5.71	16.09	13.77
(18372, 35665, 0.0125)	-2.97	5.17	12.82	12.70
(69995, 137881, 0.00625)	-1.62	3.89	8.08	8.47

Table 4.2: Fractured domain with isotropic diffusivity: For ADR equation

Delaunay triangulation ($Nv, Nele, h$)	Concentration		% of nodes violated	
	Minimum	Maximum	(NC)	(DMP)
(539, 906, 0.1)	-9.64	4.53	14.29	8.16
(1564, 2826, 0.05)	-1.54	3.57	18.03	1.53
(5090, 9620, 0.025)	-4.47	3.30	16.01	0.29
(18372, 35665, 0.0125)	-2.95	2.56	12.89	0.04
(69995, 137881, 0.00625)	-1.62	2.50	8.06	0.00

4.4.3.7 *Errors in local and global species balance*

It is well-known that without using a post-processing method, the discrete standard single-field Galerkin formulation *does not possess* local and global conservation properties Hughes et al. (2000); Burdakov et al. (2012). In finite element literature, there are various post-processing methods to quantify the errors incurred in satisfying local and global species balance Gresho and Sani (2000); Donea and Huerta (2003); Zienkiewicz et al. (2013). Herein, we shall use Herrmann’s method of optimal sampling (Zienkiewicz et al., 2013, Subsection 15.2.2), which is a popular post-processing technique to obtain derived quantities from primary variables (such as the concentration variable in single-field finite element formulations). In the context of recovery-based error estimators, this post-processing method is also known as traditional global smoothing method. To summarize, this method involves minimizing a unconstrained least-squares flux functional to obtain nodal flux vectors. Then, using these flux vectors, local and global species balance errors are computed.

Qualitatively, the contours corresponding to local species balance errors for test problems #1, #2, and #3 on *coarse* DMP-based meshes are shown in Figure 4.26. From this figure, it is clear that test problem #3 exhibits considerable errors in satisfying local species balance. Quantitatively, Table 4.3 provides local and global species balance errors on traditional *h*-refined meshes for test problems #1 and #2. From this table, it can be concluded that the decrease in these errors is slow. Hence, there is need for locally conservative DMP-preserving finite element methods.

Remark 4.4.7. *It should be noted that there are various ways in which one can develop locally conservative global smoothing methods. For example, this can be achieved by constructing a constrained monotonic regression based method (following the procedure outlined by Burdakov et al. Burdakov et al. (2012)) to obtain a conservative flux vector. However, as discussed in Section 4.1 and Section 4.3, it should be note that such type of post-processing methods are not variationally consistent and refutes*

Table 4.3: Errors in local and global species balance: For pure isotropic and anisotropic diffusion equation.

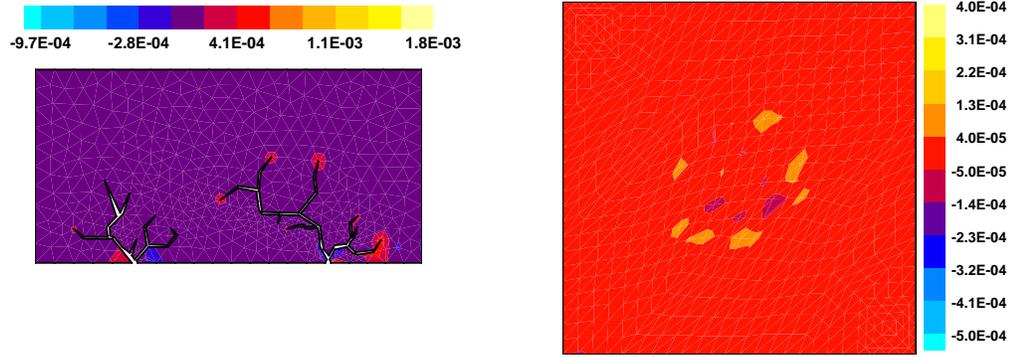
h	Test problem #1		h	Test problem #2	
	LSB (abs. max.)	GSB		LSB (abs. max.)	GSB
0.1	1.80×10^{-3}	1.53×10^{-2}	0.08	5.75×10^{-4}	-1.44×10^{-4}
0.05	4.66×10^{-4}	1.23×10^{-2}	0.042	1.47×10^{-4}	-1.38×10^{-4}
0.025	3.48×10^{-4}	8.48×10^{-3}	0.028	7.34×10^{-5}	-1.12×10^{-4}
0.0125	7.14×10^{-4}	7.82×10^{-3}	0.0135	1.59×10^{-5}	-6.83×10^{-5}
0.00625	6.75×10^{-4}	5.38×10^{-3}	0.0075	5.46×10^{-6}	-4.36×10^{-5}

the purpose of developing physics-compatible DMP-based meshes.

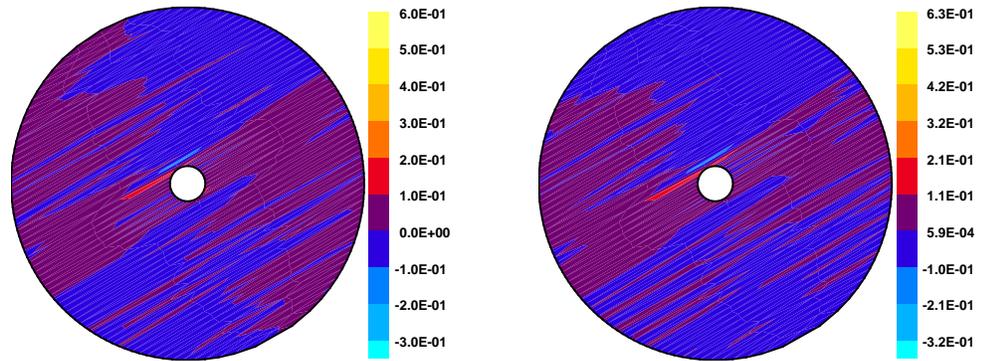
4.4.4 Sufficient conditions for a rectangular element

For the sake of illustration, consider a rectangular element whose vertices are located at $(0, 0)$, $(a, 0)$, (a, b) , and $(0, b)$. Our objective is to derive conditions on a and b , such that the local stiffness matrix is weakly diagonally dominant. Herein, we consider a pure anisotropic diffusion equation and derive restrictions on the coordinates of the rectangular element. Based on the lines of the *local stiffness restriction method* outlined in subsection 4.4.2.1, the local stiffness matrix for the case when diffusivity $\mathbf{D}(\mathbf{x}) = \begin{pmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{pmatrix}$ is anisotropic and constant in Ω is given as

$$\mathbf{K}_e = \begin{pmatrix} \frac{bD_{xx}}{3a} + \frac{D_{xy}}{2} + \frac{aD_{yy}}{3b} & -\frac{bD_{xx}}{3a} + \frac{aD_{yy}}{6b} & \frac{bD_{xx}}{6a} - \frac{aD_{yy}}{3b} & -\frac{bD_{xx}}{6a} - \frac{D_{xy}}{2} - \frac{aD_{yy}}{6b} \\ -\frac{bD_{xx}}{3a} + \frac{aD_{yy}}{6b} & \frac{bD_{xx}}{3a} - \frac{D_{xy}}{2} + \frac{aD_{yy}}{3b} & -\frac{bD_{xx}}{6a} + \frac{D_{xy}}{2} - \frac{aD_{yy}}{6b} & \frac{bD_{xx}}{6a} - \frac{aD_{yy}}{3b} \\ \frac{bD_{xx}}{6a} - \frac{aD_{yy}}{3b} & -\frac{bD_{xx}}{6a} + \frac{D_{xy}}{2} - \frac{aD_{yy}}{6b} & \frac{bD_{xx}}{3a} - \frac{D_{xy}}{2} + \frac{aD_{yy}}{3b} & -\frac{bD_{xx}}{3a} + \frac{aD_{yy}}{6b} \\ -\frac{bD_{xx}}{6a} - \frac{D_{xy}}{2} - \frac{aD_{yy}}{6b} & \frac{bD_{xx}}{6a} - \frac{aD_{yy}}{3b} & -\frac{bD_{xx}}{3a} + \frac{aD_{yy}}{6b} & \frac{bD_{xx}}{3a} + \frac{D_{xy}}{2} + \frac{aD_{yy}}{3b} \end{pmatrix}. \quad (4.4.51)$$



(a) Test problem #1: $\mathbf{v} = (0, 0)$ and $\alpha = 0$ (b) Test problem #2: $\mathbf{v} = (0, 0)$ and $\alpha = 0$



(c) Test problem #3: $\mathbf{v} = (0, 0)$ and $\alpha = 0$ (d) Test problem #3: $\mathbf{v} = (1.5, 1.0)$ and $\alpha = 1.0$

Figure 4.26: Local species balance errors: The figures show the errors incurred in satisfying local species balance for various test problems on coarse meshes.

Condition #7

From AM-GM inequality, we have the following relation:

$$\frac{bD_{xx}}{3a} + \frac{aD_{yy}}{3b} \geq \frac{2}{3}\sqrt{D_{xx}D_{yy}} > \frac{2}{3}|D_{xy}| \geq \frac{1}{2}|D_{xy}|, \quad (4.4.52)$$

which implies that the positive diagonal entries: $(\mathbf{K}_e)_{ii} > 0 \quad \forall i = 1, 2, 3, 4$, is trivially satisfied.

Condition #8

Non-positive off-diagonal entries: $(\mathbf{K}_e)_{ij} \leq 0 \quad \forall i \neq j$ where $i = 1, 2, 3, 4$, and $j = 1, 2, 3, 4$, needs to be satisfied. For instance, when $i = 1$ and $j = 2, 3, 4$, this restriction gives the following relations:

$$\sqrt{\frac{D_{xx}}{2D_{yy}}} \leq \frac{a}{b} \leq \sqrt{\frac{2D_{xx}}{D_{yy}}} \quad -\frac{bD_{xx}}{6a} - \frac{D_{xy}}{2} - \frac{aD_{yy}}{6b} \leq 0. \quad (4.4.53)$$

Additionally, we should also satisfy the restrictions imposed by equation (4.4.26) on diffusivity tensor. In a similar fashion, one can derive restrictions for other combinations of i and j .

Condition #9

Weak diagonal dominance of rows: $|(\mathbf{K}_e)_{ii}| \geq \sum_{i \neq j} |(\mathbf{K}_e)_{ij}| \quad \forall i, j$ where $i = 1, 2, 3, 4$, and $j = 1, 2, 3, 4$, is trivially satisfied if Condition #7 and Condition #8 are met. This is because from equation (4.4.51), it is evident that $(\mathbf{K}_e)_{ii} + \sum_{i \neq j} (\mathbf{K}_e)_{ij} = 0 \quad \forall i, j$ where $i = 1, 2, 3, 4$, and $j = 1, 2, 3, 4$.

Finally, as explained in subsection 4.4.2.1, extending the local stiffness restriction approach to incorporate advection and linear reaction is straightforward and shall not be dealt with to save space. Moreover, it is easy to construct mesh restrictions for any set of arbitrary coordinates of a quadrilateral element using symbolic

packages like `Mathematica`. But the resulting inequalities will be more complex to analyze mathematically and visualize graphically.

4.5 CONCLUDING REMARKS AND OPEN QUESTIONS

We outlined a general procedure to obtain the restrictions that are needed for a computational grid to satisfy various mathematical principles – comparison principles, maximum principles, and the non-negative constraint. We illustrated the workings of this procedure by obtaining the mesh restrictions for T3 and Q4 finite elements. The procedure is, however, equally applicable to other low-order finite elements.

First, we critiqued three different approaches to satisfy maximum principles, comparison principles, and non-negative constraint for a general linear second-order elliptic equation. A pictorial description of a generic relationship between DMPs, DCPs, and NC based on a Venn diagram is proposed. This sketch helps to easily relate the space of solutions obtained using mesh restrictions, non-negative numerical formulations, and post-processing methods. We then presented necessary and sufficient conditions on the stiffness matrix \mathbf{K} to meet the mathematical properties. Using these conditions, we derived stronger and weaker mesh restrictions for T3 element. The stronger mesh restriction corresponds to the **anisotropic non-obtuse angle condition** while the weaker one corresponds to the **generalized Delaunay-type angle condition**. Motivated by these mesh restriction conditions, different kinds of Péclet and Damköhler numbers are proposed for advective-diffusive-reactive systems when diffusivity is anisotropic.

For isotropic diffusivity, we established that acute-angled or right-angled triangle is sufficient to satisfy discrete principles. However, for anisotropic diffusivity, we showed that in order to satisfy DMPs, DCPs, and NC, all the dihedral angles of a simplex measured in the metric of $\widetilde{\mathbf{D}}_{\Omega_e}^{-1}$ have to be either $\mathcal{O}(h\|\mathbf{v}\|_{\infty, \mathcal{T}_h} + h^2\|\alpha\|_{\infty, \mathcal{T}_h})$ acute/non-obtuse or $\mathcal{O}(h\|\mathbf{v}\|_{\infty, \mathcal{T}_h} + h^2\|\alpha\|_{\infty, \mathcal{T}_h})$ Delaunay. Pictorially, this means

that the feasible region for T3 and Q4 elements to satisfy various discrete principles is based on a metric tensor whose components are a function of anisotropic diffusivity with respect to a suitable coordinate system. Then, an anisotropic metric tensor and an iterative algorithm to generate various types of DMP-based triangulations are described. Different numerical examples and respective DMP-based triangular meshes are presented for different types of $\mathbf{D}(\mathbf{x})$ to demonstrate the pros and cons of imposing mesh restrictions. Furthermore, the errors incurred in satisfying local and global species balance are documented. Based on these numerical experiments, the following inferences can be drawn:

- (C1) For pure isotropic or anisotropic diffusion equation, a coarse DMP-based triangulation is sufficient to satisfy various discrete principles. However, for advection-dominated and reaction-dominated scenarios, we need a highly refined DMP-preserving computational mesh to obtain non-negative solutions.
- (C2) Existing traditional and non-traditional methods of h -refinement may not guarantee the satisfaction of DMPs, DCPs, and NC always.
- (C3) On coarse DMP-based meshes, errors incurred in satisfying local and global species balance for highly anisotropic diffusion-type problems is considerable due to the skewed nature of the mesh elements. Moreover, the decrease in local and global species balance errors upon h -refinement is slow.
- (C4) DMP-based meshes change as one alters the underlying anisotropic diffusivity tensor.

In the light of the recent developments and motivated by the above discussions, we have chosen to emphasize on the following *four open problems* that we consider particularly interesting in view of their mathematical richness, numerical challenges, and potential applications:

- (OP1) In this chapter, all the meshes used in the numerical examples are of Delaunay-type. This is because most of the existing open source mesh generators such as BAMG, Gmsh, Triangle, BL2D, Mmg3d, and CGALmesh are Delaunay. Recently, Erten and Üngör Erten and Üngör (2007, 2009a,b) have developed a non-obtuse/acute angled mesh generator called aCute by modifying Triangle. However, aCute is restricted to 2D and Euclidian metric tensors. Hence, such a software can only be used to satisfy discrete principles for problems involving heterogeneous isotropic diffusivity. Having an anisotropic non-obtuse/acute \mathcal{M} -uniform meshing software would be of great importance, as the numerical solutions obtained from these meshes not only satisfy DMPs, DCPs, and NC, but also converge uniformly (an attractive aspect in finite element analysis Ciarlet and Raviart (1973)). *To date, there is no such mesh generator.* Developing such a software will have a profound impact on obtaining physically meaningful numerical solutions for diffusion-type equations.
- (OP2) For advection-dominated and reaction-dominated advection-diffusion-reaction problems, mesh refinement that adheres to DMPs is needed to obtain stable and sufficiently accurate numerical solutions. As described in (C2), not every method of h -refinement is DMP-preserving. Hence, a consistent way of generating a DMP-based h -refined mesh (that satisfies **Generalized Delaunay-type angle condition**) is still unresolved.
- (OP3) From (C3), it is apparent that local and global mass conservation property is needed. An approach to preserve such a property without violating DMPs, DCPs, and NC, is to obtain mesh restrictions for mixed Galerkin formulation based on lowest-order Raviart-Thomas spaces. Recently, Huang and Wang Huang and Wang (2014) have developed a methodology to satisfy

DMPs for a class of locally conservative weak Galerkin methods using lowest-order Raviart-Thomas spaces. However, this methodology is limited to pure anisotropic steady-state diffusion equation in two-dimensions. Hence, a mesh restriction based method to satisfy different discrete principles, local species balance, and global species balance for anisotropic advection-diffusion-reaction equations thus far is unsolved.

(OP4) In order to construct DMP-based meshes for low-order non-simplicial finite elements such as Q4, from subsection 4.4.4, it is evident that stronger and weaker mesh conditions are needed. *So far, there are no mesh restriction theorems analogous to simplicial meshes that can provide a general framework to construct non-simplicial meshes for anisotropic diffusivity tensors.* Theoretically and numerically, it would be very interesting and informative to have a comparative study on the performance of simplicial vs. non-simplicial DMP-based meshes for various benchmark problems discussed in Section 4.4.

Nevertheless, due to enormous research activity in the field of advection-diffusion-reaction equations, it is impossible to list every open question on preserving DCPs, DMPs, and NC. To conclude, the research findings in this chapter will be invaluable to the research community and finite element practitioners in two respects. *First*, it will guide the existing users on the restrictions to be placed on the computational mesh to meet important mathematical properties like maximum principles, comparison principles, and the non-negative constraint. *Second*, for complex geometries and highly anisotropic media, this study has clearly shown that placing restrictions on computational grids may not always be a viable approach to achieve physically meaningful non-negative solutions. *We hope that this research work will motivate researchers to develop new methodologies for advective-diffusive-reactive systems that satisfy local and global species balance, comparison principles, maximum principles, and the non-negative constraint on coarse general computational grids.*

Chapter 5

ON ENFORCING MAXIMUM PRINCIPLES AND ACHIEVING ELEMENT-WISE SPECIES BALANCE FOR ADVECTION-DIFFUSION-REACTION EQUATIONS UNDER FINITE ELEMENT METHOD

“As time goes on, it becomes increasingly evident that the rules which the mathematician finds interesting are the same as those which nature has chosen.”

Paul Dirac

We present a robust computational framework for advective-diffusive-reactive systems that satisfies maximum principles, the non-negative constraint, and element-wise species balance property. The proposed methodology is valid on general computational grids, can handle heterogeneous anisotropic media, and provides accurate numerical solutions even for very high Péclet numbers. The significant contribution

of this chapter is to incorporate advection (which makes the spatial part of the differential operator non-self-adjoint) into the non-negative computational framework, and overcome numerical challenges associated with advection. We employ low-order mixed finite element formulations based on least-squares formalism, and enforce explicit constraints on the discrete problem to meet the desired properties. The resulting constrained discrete problem belongs to convex quadratic programming for which a unique solution exists. Maximum principles and the non-negative constraint give rise to bound constraints while element-wise species balance gives rise to equality constraints. The resulting convex quadratic programming problems are solved using an interior-point algorithm. Several numerical results pertaining to advection-dominated problems are presented to illustrate the robustness, convergence, and the overall performance of the proposed computational framework.

5.1 INTRODUCTION AND MOTIVATION

Advection-diffusion-reaction (ADR) equations are pervasive in the mathematical modeling of several important phenomena in mathematical physics and engineering sciences. Some examples include degradation/healing of materials under extreme environmental conditions Chatterjee et al. (2001), coupled chemo-thermo-mechano-diffusion problems in composites Sih et al. (1986), contaminant transport Bear et al. (1993), turbulent mixing in atmospheric sciences Cant and Mastorakos (2008), diffusion-controlled biochemical reactions Saltzman (2001), tracer modeling in hydrogeology Leibundgut et al. (2009), and ionic mobility in biological systems Keener and Sneyd (2009). Additionally, ADR equation serves as a good mathematical model in numerical analysis, as it offers various unique challenges in obtaining stable and accurate numerical solutions Morton (1996).

The primary variables in these mathematical models are typically the concentration and/or the (absolute) temperature. These quantities naturally attain non-negative values. Under some popular constitutive models (such as Fourier model and Fickian model, and their generalizations), these physical quantities satisfy diffusion-type equations, which are elliptic/parabolic partial differential equations (PDEs) and can be non-self-adjoint. These PDEs are known to satisfy important mathematical properties like maximum principles and the non-negative constraint (e.g., see Gilbarg and Trudinger (2001)). A predictive numerical formulation needs to satisfy these mathematical properties and physical laws like the (local and global) species balance. It is well-documented in the literature that traditional numerical methods perform poorly for advection-dominated ADR equations (e.g., see Donea and Huerta (2003); Morton (1996)). In the past few decades, considerable progress has been made to obtain sufficiently accurate numerical solutions for ADR equations on coarse computational grids Codina (2000). It is then natural to ask: “*why there is a need for yet another numerical formulation for ADR equation?*”. We now outline several reasons for such a need.

- (a) *Localized phenomena and node-to-node spurious oscillations:* For advection-dominated problems, it is well-known that the standard single-field Galerkin finite element formulation gives node-to-node spurious oscillations on coarse computational grids Morton (1996). Moreover, it cannot capture steep gradients such as interior and boundary layers. Various alternate numerical techniques have been proposed to avoid these spurious oscillations Gresho and Sani (2000). Some methods seem to capture steep gradients in interior layers while others capture boundary layers. However, most of them do not seem to capture both interior and boundary layers, and at the same time avoid node-to-node spurious oscillations Augustin et al. (2011). A notable work towards this direction is by Hsieh and Yang Hsieh

and Yang (2009), which can capture both interior and boundary layers under adequate mesh refinement. However, this formulation has several other deficiencies some of which are discussed below and illustrated using numerical examples in subsequent sections.

- (b) *Violation of the non-negative constraint and maximum principles:* As mentioned earlier, physical quantities such as concentration and temperature naturally attain non-negative values. It is highly desirable for a numerical formulation to respect these physical constraints. This is particularly important in a numerical simulation of reactive transport, as a negative value for concentration will result in an algorithmic failure. However, it is clearly documented in the literature that many existing formulations based on finite element Liska and Shashkov (2008); Nakshatrala and Valocchi (2009); Nagarajan and Nakshatrala (2011), finite volume Potier (2005), and finite difference Brezzi et al. (2005) do not satisfy the non-negative constraint and maximum principles in the discrete setting. They also discuss various methodologies to satisfy such properties. But most of these methodologies are for pure diffusion equations, which are self-adjoint. For example, in Reference Nakshatrala and Valocchi (2009), two mixed formulations based on RT0 spaces and variational multiscale formalism have been modified to meet the non-negative constraint for diffusion equations. This approach is not directly applicable to ADR equations for two reasons. First, these formulations do not cure the node-to-node spurious oscillations. Second, they do not possess a variational structure for ADR equations. Some numerical formulations are constructed to satisfy the non-negative constraint and maximum principles by taking advection into account (e.g., Burman and Hansbo (2004); Burman and Ern (2005)). However, they do not satisfy local and global species balance, and are restricted to *isotropic* diffusion. Conservative post-processing methods exist in the literature to recover certain desired properties such as species balance. But

such formulations are not variationally consistent Burdakov et al. (2012).

- (c) *Satisfying local and global species balance:* In transport problems, the balance of species is an important physical law that needs to be met. It is therefore desirable for a numerical formulation to satisfy local and global species balance, say, up to machine precision (which is approximately 10^{-16} on a 64-bit machine). However, many finite element formulations do not satisfy local and global species balance (see Codina (1998, 2000); Hsieh and Yang (2009)). The main focus of the methods outlined in these papers is to capture the localized phenomena such as boundary and interior layers. Moreover, these works did not quantify the errors incurred in satisfying local species balance and global species balance. It needs to be emphasized that many finite element methods do exist that inherently satisfy local and global species balance, for example, Raviart-Thomas spaces Raviart and Thomas (1977) and BDM spaces Brezzi et al. (1987). But these approaches do not fix other issues discussed herein such as the node-to-node spurious oscillations or meeting maximum principles for ADR equations.
- (d) *Other influential factors:* Some other important factors that can affect the performance of a numerical formulation are the advection velocity field and its divergence, anisotropy of the medium, reaction coefficients, topology of the medium, computational mesh, multiple spatial-scales arising due to the heterogeneity of the medium, and multiple time-scales involved in various physical processes. Another aspect that brings tremendous numerical challenges is chemical reactions involving multiple species.

It is a herculean task to address all the aforementioned deficiencies, and we strongly believe that it may take a series of papers to overcome all the deficiencies. A similar sentiment is shared in the review article by Stynes entitled “*Numerical methods for convection-diffusion problems or the 30 years war*” Stynes (2013). We

therefore take motivation from George Pólya’s quote Pólya (2009): “*If you can’t solve a problem, then there is an easier problem you can solve: find it.*” In this chapter, we shall pose a problem that is simpler than the grand challenge of overcoming all the aforementioned numerical deficiencies but still make a significant advancement with respect to the current state-of-the-art. We then provide a solution to this simpler problem. To state it more precisely, the main contribution of this chapter is developing a least-squares-based finite element framework for ADR equations that possesses the following properties on general computational grids:

- (P1) No spurious node-to-node oscillations in the entire domain.
- (P2) Captures interior and boundary layers for advection-dominated problems.
- (P3) Satisfies discrete maximum principles and the non-negative constraint.
- (P4) Satisfies local and global species balance.
- (P5) Gives sufficiently accurate solutions even on coarse computational grids¹.

To the best of authors’ knowledge, there exists no finite element methodology for advective-diffusive-reactive systems that possesses the desirable properties (P1)–(P5).

The rest of the chapter is organized as follows. Section 5.2 presents the governing equations for an advective-diffusive-reactive system, and discusses the associated mathematical properties. Section 5.3 outlines several plausible approaches, and discusses their drawbacks in meeting the mentioned mathematical properties. In Section 5.4, we propose a constrained optimization-based low-order mixed finite element method to satisfy maximum principles, the non-negative constraint, local species balance, and global species balance. In Section 5.6, we perform a numerical h -convergence study using a benchmark problem. In Section 5.7, we specialize to

¹One may expect some subjectivity in calling a mesh to be coarse. But in this chapter, we will define precisely what is meant by a “coarse mesh” for advection-diffusion-reaction equations in terms of M -matrices.

transport-limited bimolecular reactions to solve problems related to plume development and mixing in isotropic/anisotropic heterogeneous media. Finally, conclusions are drawn in Section 5.8.

We shall denote scalars by lowercase English alphabet or lowercase Greek alphabet (e.g., concentration c and stabilization parameter τ). The continuum vectors are denoted by lowercase boldface normal letters, and the second-order tensors will be denoted using uppercase boldface normal letters (e.g., vector \mathbf{x} and second-order tensor \mathbf{D}). In the finite element context, the vectors are denoted using lowercase boldface italic letters, and the matrices are denoted using uppercase boldface italic letters (e.g., vector \mathbf{v} and matrix \mathbf{K}). We shall use NN to denote non-negative, DMP denotes discrete maximum principle, LSB to denote local species balance, and GSB to denote global species balance. We shall use XSeed to denote the number of (finite element) nodes in a mesh along x-direction. Likewise for YSeed. Other notational conventions adopted in this chapter are introduced as needed.

5.2 GOVERNING EQUATIONS: ADVECTIVE-DIFFUSIVE-REACTIVE SYSTEMS

Let $\Omega \subset \mathbb{R}^{nd}$ be a bounded open domain, where “ nd ” denotes the number of spatial dimensions. The boundary of the domain is denoted by $\partial\Omega$, which is assumed to be piecewise smooth. Mathematically, $\partial\Omega := \overline{\Omega} - \Omega$, where a superposed bar denotes the set closure. A spatial point is denoted by $\mathbf{x} \in \overline{\Omega}$. The gradient and divergence operators with respect to \mathbf{x} are, respectively, denoted by $\text{grad}[\bullet]$ and $\text{div}[\bullet]$. The unit outward normal to the boundary is denoted by $\hat{\mathbf{n}}(\mathbf{x})$. Let $c(\mathbf{x})$ denote the concentration field. The boundary is divided into two parts: Γ^c and Γ^q such that $\text{meas}(\Gamma^c) > 0$, $\overline{\Gamma^c} \cup \overline{\Gamma^q} = \partial\Omega$, and $\Gamma^c \cap \Gamma^q = \emptyset$. Γ^c is the part of the boundary on which the concentration is prescribed and Γ^q is the other part of the boundary on which the diffusive/total flux is prescribed.

The governing equations for steady-state response of an ADR system take the following form:

$$\alpha(\mathbf{x})c(\mathbf{x}) + \operatorname{div} [c(\mathbf{x})\mathbf{v}(\mathbf{x}) - \mathbf{D}(\mathbf{x})\operatorname{grad}[c(\mathbf{x})]] = f(\mathbf{x}) \quad \text{in } \Omega, \quad (5.2.1a)$$

$$c(\mathbf{x}) = c^p(\mathbf{x}) \quad \text{on } \Gamma^c, \text{ and} \quad (5.2.1b)$$

$$\left(\left(\frac{1 - \operatorname{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) \mathbf{v}(\mathbf{x})c(\mathbf{x}) - \mathbf{D}(\mathbf{x})\operatorname{grad}[c(\mathbf{x})] \right) \bullet \hat{\mathbf{n}}(\mathbf{x}) = q^p(\mathbf{x}) \quad \text{on } \Gamma^q. \quad (5.2.1c)$$

$\mathbf{v}(\mathbf{x})$ is the known advection velocity field, $f(\mathbf{x})$ is the prescribed volumetric source, $\mathbf{D}(\mathbf{x})$ is the anisotropic diffusivity tensor, $\alpha(\mathbf{x})$ is the linear reaction coefficient, $c^p(\mathbf{x})$ is the prescribed concentration, $q^p(\mathbf{x})$ is the prescribed diffusive/total flux, and the sign function is defined as

$$\operatorname{Sign}[\varphi] := \begin{cases} -1 & \text{if } \varphi < 0 \\ 0 & \text{if } \varphi = 0 \\ +1 & \text{if } \varphi > 0 \end{cases}. \quad (5.2.2)$$

The advection velocity need not be solenoidal in our treatment (i.e., $\operatorname{div}[\mathbf{v}(\mathbf{x})]$ need not be zero). The Neumann boundary condition given in equation (5.2.1c) can be interpreted as

$$\hat{\mathbf{n}}(\mathbf{x}) \bullet (\mathbf{v}(\mathbf{x})c(\mathbf{x}) - \mathbf{D}(\mathbf{x})\operatorname{grad}[c(\mathbf{x})]) = q^p(\mathbf{x}) \quad \text{on } \Gamma_-^q \text{ (total flux) and} \quad (5.2.3a)$$

$$-\hat{\mathbf{n}}(\mathbf{x}) \bullet \mathbf{D}(\mathbf{x})\operatorname{grad}[c(\mathbf{x})] = q^p(\mathbf{x}) \quad \text{on } \Gamma_+^q \text{ (diffusive flux),} \quad (5.2.3b)$$

where Γ_+^q and Γ_-^q are, respectively, defined as (see Figure 5.1)

$$\Gamma_-^q := \{ \mathbf{x} \in \Gamma^q \mid \mathbf{v}(\mathbf{x}) \bullet \hat{\mathbf{n}}(\mathbf{x}) < 0 \} \quad \text{(inflow boundary) and} \quad (5.2.4a)$$

$$\Gamma_+^q := \{ \mathbf{x} \in \Gamma^q \mid \mathbf{v}(\mathbf{x}) \bullet \hat{\mathbf{n}}(\mathbf{x}) \geq 0 \} \quad \text{(outflow boundary).} \quad (5.2.4b)$$

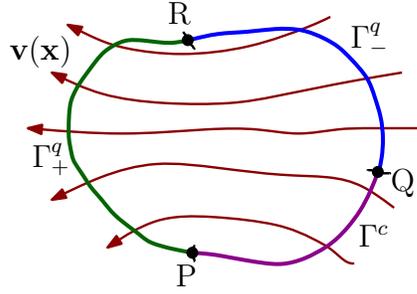


Figure 5.1: This figure illustrates concentration and flux boundary conditions.

In Figure 5.1, Γ_-^q corresponds to the inflow boundary while Γ_+^q corresponds to the outflow boundary. Total flux is prescribed on Γ_-^q , diffusive flux is prescribed on Γ_+^q , and concentration is prescribed on Γ^c . $P = \bar{\Gamma}^c \cap \bar{\Gamma}_+^q$, $Q = \bar{\Gamma}^c \cap \bar{\Gamma}_-^q$, and $R = \bar{\Gamma}_+^q \cap \bar{\Gamma}_-^q$. For well-posedness, we have $\bar{\Gamma}^c \cup \bar{\Gamma}_+^q \cup \bar{\Gamma}_-^q = \partial\Omega$, $\Gamma^c \cap \Gamma_+^q = \emptyset$, $\Gamma^c \cap \Gamma_-^q = \emptyset$, and $\Gamma_+^q \cap \Gamma_-^q = \emptyset$.

Remark 5.2.1. *In the literature, more predominantly in the numerical literature, the term advection is often used synonymously with convection. It should, however, be noted that these two terms describe different physical phenomena, and there is a need to clarify the terminology here. An ADR equation arises from the balance of mass of a given species. In 1D, an ADR equation takes the following form:*

$$\alpha(x)c(x) + \frac{d(vc)}{dx} - \frac{d}{dx} \left(D(x) \frac{dc}{dx} \right) = f(x), \quad (5.2.5)$$

which is mathematically equivalent to the following equation:

$$\left(\alpha(x) + \frac{dv}{dx} \right) c(x) + v(x) \frac{dc}{dx} - \frac{d}{dx} \left(D(x) \frac{dc}{dx} \right) = f(x). \quad (5.2.6)$$

One can obtain a “similar” mathematical equation by linearizing the incompressible Navier-Stokes equation, and an appropriate name for this linearized equation is the convection-dissipation-reaction (CDR) equation. The CDR equation in 1D has the

following mathematical form:

$$\frac{dv_0}{dx}\tilde{v}(x) + v_0(x)\frac{d\tilde{v}}{dx} - \frac{d}{dx}\left(\mu(x)\frac{d\tilde{v}}{dx}\right) = b(x, p_0(x)) + 2v_0(x)\frac{dv_0}{dx}, \quad (5.2.7)$$

where $\tilde{v}(x)$ is the velocity of the fluid, and $p_0(x)$ and $v_0(x)$ are known pressure and velocity fields about which the Navier-Stokes equation is linearized. From equations (5.2.6) and (5.2.7), it is evident that 1D ADR equation and 1D CDR equation have similar mathematical forms. However, their physical underpinnings are completely different, as the Navier-Stokes equation is obtained by substituting a specific constitutive model into the balance of linear momentum.

5.2.1 Weak formulations

The following function spaces will be used in the rest of this chapter:

$$\mathcal{C} := \left\{ c(\mathbf{x}) \in H^1(\Omega) \mid c(\mathbf{x}) = c^p(\mathbf{x}) \text{ on } \Gamma^c \right\}, \quad (5.2.8a)$$

$$\mathcal{W} := \left\{ w(\mathbf{x}) \in H^1(\Omega) \mid w(\mathbf{x}) = 0 \text{ on } \Gamma^c \right\}, \text{ and} \quad (5.2.8b)$$

$$\mathcal{Q} := \left\{ \mathbf{q}(\mathbf{x}) \in (L_2(\Omega))^{nd} \mid \text{div}[\mathbf{q}(\mathbf{x})] \in L_2(\Omega) \right\}, \quad (5.2.8c)$$

where $\mathbf{q}(\mathbf{x}) = c(\mathbf{x})\mathbf{v}(\mathbf{x}) - \mathbf{D}(\mathbf{x})\text{grad}[c(\mathbf{x})]$ and $H^1(\Omega)$ is a standard Sobolev space Evans (1998). Given two vector fields $a(\mathbf{x})$ and $b(\mathbf{x})$ on a set K , the standard L_2 inner product over K is denoted as:

$$(a; b)_K = \int_K a(\mathbf{x}) \bullet b(\mathbf{x}) \, dK. \quad (5.2.9)$$

The subscript will be dropped if $K = \Omega$. The most popular way to construct a weak formulation is to employ the Galerkin formalism. Based on the manner in which one applies the divergence theorem, the single-field Galerkin formulation for equations (5.2.1a)–(5.2.1c) can be posed in two different ways.

Single-field Galerkin formulation #1 (SG₁): Find $c(\mathbf{x}) \in \mathcal{C}$ such that we have

$$\begin{aligned}
& (w; \alpha c) - (\text{grad}[w] \bullet \mathbf{v}; c) + (\text{grad}[w]; \mathbf{D}(\mathbf{x})\text{grad}[c]) \\
& + \left(w; \left(\frac{1 + \text{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) (\mathbf{v} \bullet \hat{\mathbf{n}}) c \right)_{\Gamma^q} \\
& = (w; f) - (w; q^p)_{\Gamma^q} \quad \forall w(\mathbf{x}) \in \mathcal{W}.
\end{aligned} \tag{5.2.10}$$

Single-field Galerkin formulation #2 (SG₂): Find $c(\mathbf{x}) \in \mathcal{C}$ such that we have

$$\begin{aligned}
& (w; (\alpha + \text{div}[\mathbf{v}]) c) + (w; \text{grad}[c] \bullet \mathbf{v}) + (\text{grad}[w]; \mathbf{D}(\mathbf{x})\text{grad}[c]) \\
& - \left(w; \left(\frac{1 - \text{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) (\mathbf{v} \bullet \hat{\mathbf{n}}) c \right)_{\Gamma^q} \\
& = (w; f) - (w; q^p)_{\Gamma^q} \quad \forall w(\mathbf{x}) \in \mathcal{W}.
\end{aligned} \tag{5.2.11}$$

Note that the solution obtained will be the same regardless whether we use either SG₁ or SG₂. However, this is not true if one uses total/diffusive flux on Neumann boundary without giving due consideration to inflow and/or outflow Neumann boundary conditions. For more details, see subsection 5.2.3.

5.2.2 Maximum principles and the non-negative constraint

A basic qualitative property of elliptic boundary value problems is the maximum principle. This property gives a priori estimate for $c(\mathbf{x})$ in Ω through its values on Γ^c . The following assumptions will be made to present a continuous maximum principle for ADR equations with both Dirichlet and Neumann boundary conditions:

(A1) Ω is piecewise smooth domain with Lipschitz continuous boundary $\partial\Omega$.

(A2) The scalar functions $\alpha : \overline{\Omega} \rightarrow \mathbb{R}$, $(\mathbf{v})_i : \overline{\Omega} \rightarrow \mathbb{R}$, and $(\mathbf{D})_{ij} : \overline{\Omega} \rightarrow \mathbb{R}$ are continuously differentiable in their respective domains for $i = 1, \dots, nd$. Furthermore, $f \in L_2(\Omega)$, $q^p \in L_2(\Gamma^q)$, and $c^p = g^*|_{\Gamma^c}$ with $g^* \in H^1(\Omega)$.

(A3) The diffusivity tensor is assumed to be symmetric, uniformly elliptic, and bounded above. That is, there exists two constants (i.e., independent of \mathbf{x}), $0 < \gamma_{lb} \leq \gamma_{ub} < +\infty$, such that we have

$$0 < \gamma_{lb} \mathbf{y} \bullet \mathbf{y} \leq \mathbf{y} \bullet \mathbf{D}(\mathbf{x}) \mathbf{y} \leq \gamma_{ub} \mathbf{y} \bullet \mathbf{y} \quad \forall \mathbf{y} \in \mathbb{R}^{nd} \setminus \{\mathbf{0}\}. \quad (5.2.12)$$

(A4) The advection velocity field $\mathbf{v}(\mathbf{x})$ and the reaction coefficient $\alpha(\mathbf{x})$ are restricted as:

$$0 \leq \alpha(\mathbf{x}) + \operatorname{div} [\mathbf{v}(\mathbf{x})] \leq \beta_1(\mathbf{x}) \quad \forall \mathbf{x} \in \Omega, \quad (5.2.13a)$$

$$0 \leq \alpha(\mathbf{x}) + \frac{1}{2} \operatorname{div} [\mathbf{v}(\mathbf{x})] \leq \beta_2(\mathbf{x}) \quad \forall \mathbf{x} \in \Omega, \text{ and} \quad (5.2.13b)$$

$$0 \leq |\mathbf{v}(\mathbf{x}) \bullet \hat{\mathbf{n}}(\mathbf{x})| \leq \beta_3(\mathbf{x}) \quad \forall \mathbf{x} \in \Gamma^q, \quad (5.2.13c)$$

where $\beta_1(\mathbf{x}) \in L_{nd/2}(\Omega)$, $\beta_2(\mathbf{x}) \in L_{nd/2}(\Omega)$, and $\beta_3(\mathbf{x}) \in L_{nd-1}(\Gamma^q)$. It is assumed that functions $\beta_1(\mathbf{x})$, $\beta_2(\mathbf{x})$, and $\beta_3(\mathbf{x})$ are bounded above for a unique weak solution to exist based on the Lax-Milgram theorem.

(A5) The part of the boundary on which Dirichlet boundary condition is enforced is not empty (i.e., $\Gamma^c \neq \emptyset$).

We shall use the standard abbreviation of a.e. for almost everywhere Evans (1998).

Theorem 5.2.2 (A continuous maximum principle). *Let assumptions (A1)–(A5) hold and let the unique weak solution $c(\mathbf{x})$ of equations (5.2.1a)–(5.2.1c) belong to*

$C^1(\Omega) \cap C^0(\overline{\Omega})$. If $f(\mathbf{x}) \in L_2(\Omega)$ and $q^p(\mathbf{x}) \in L_2(\Gamma^q)$ satisfy:

$$f(\mathbf{x}) \leq 0 \quad \text{a.e. in } \Omega \text{ and} \quad (5.2.14a)$$

$$q^p(\mathbf{x}) \geq 0 \quad \text{a.e. on } \Gamma_+^q \cup \Gamma_-^q \quad (5.2.14b)$$

then $c(\mathbf{x})$ satisfies a continuous maximum principle of the following form:

$$\max_{\mathbf{x} \in \overline{\Omega}} [c(\mathbf{x})] \leq \max \left[0, \max_{\mathbf{x} \in \Gamma^c} [c^p(\mathbf{x})] \right]. \quad (5.2.15)$$

In particular, if $c^p(\mathbf{x}) \geq 0$ then

$$\max_{\mathbf{x} \in \overline{\Omega}} [c(\mathbf{x})] = \max_{\mathbf{x} \in \Gamma^c} [c^p(\mathbf{x})] \quad (5.2.16)$$

If $c^p(\mathbf{x}) \leq 0$ then we have the following non-positive property:

$$\max_{\mathbf{x} \in \overline{\Omega}} [c(\mathbf{x})] \leq 0 \quad (5.2.17)$$

Proof. Let Φ_{\max} and $m(\mathbf{x})$ are, respectively, defined as

$$\Phi_{\max} := \max \left[0, \max_{\mathbf{x} \in \Gamma^c} [c^p(\mathbf{x})] \right] \text{ and} \quad (5.2.18)$$

$$m(\mathbf{x}) := \max [0, c(\mathbf{x}) - \Phi_{\max}] \quad (5.2.19)$$

It is easy to check that $m(\mathbf{x})$ is a non-negative, continuous, and piecewise $C^1(\Omega)$ function. From equation (5.2.19), it is evident that $m(\mathbf{x})|_{\Gamma^c} = 0$ and $c(\mathbf{x}) = m(\mathbf{x}) + \Phi_{\max}$ for any $\mathbf{x} \in \overline{\Omega}$ unless $m(\mathbf{x}) = 0$. Moreover, $m(\mathbf{x}) \in \mathcal{W}$. By choosing $w(\mathbf{x}) =$

$m(\mathbf{x})$, the weak formulation given by equation (5.2.10) becomes:

$$\begin{aligned}
& (m; \alpha(m + \Phi_{\max})) - (\text{grad}[m] \bullet \mathbf{v}; (m + \Phi_{\max})) + (\text{grad}[m]; \mathbf{D} \text{grad}[m]) \\
& + \left(m; \left(\frac{1 + \text{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) \mathbf{v} \bullet \hat{\mathbf{n}} (m + \Phi_{\max}) \right)_{\Gamma^q} \\
& = (m; f) - (m; q^p)_{\Gamma^q}. \tag{5.2.20}
\end{aligned}$$

It is easy to establish the following identities:

$$\begin{aligned}
(m; \mathbf{v} \bullet \hat{\mathbf{n}} (m + \Phi_{\max}))_{\Gamma^q} &= (\text{grad}[m] \bullet \mathbf{v}; (m + \Phi_{\max})) + (m; \text{div}[\mathbf{v}] (m + \Phi_{\max})) \\
& + (m; \text{grad}[m] \bullet \mathbf{v}), \tag{5.2.21a}
\end{aligned}$$

$$\begin{aligned}
2(\text{grad}[m] \bullet \mathbf{v}; (m + \Phi_{\max})) &= (m; \mathbf{v} \bullet \hat{\mathbf{n}} (m + \Phi_{\max}))_{\Gamma^q} - (m; \text{div}[\mathbf{v}] (m + \Phi_{\max})) \\
& - (\Phi_{\max}; \text{grad}[m] \bullet \mathbf{v}), \tag{5.2.21b}
\end{aligned}$$

$$(\Phi_{\max}; \text{grad}[m] \bullet \mathbf{v}) = (\Phi_{\max}; \mathbf{v} \bullet \hat{\mathbf{n}} m)_{\Gamma^q} - (\Phi_{\max}; \text{div}[\mathbf{v}] m), \text{ and} \tag{5.2.21c}$$

$$\begin{aligned}
(\text{grad}[m] \bullet \mathbf{v}; (m + \Phi_{\max})) &= \left(m; \mathbf{v} \bullet \hat{\mathbf{n}} \left(\Phi_{\max} + \frac{1}{2} m \right) \right)_{\Gamma^q} \\
& - \left(m; \text{div}[\mathbf{v}] \left(\Phi_{\max} + \frac{1}{2} m \right) \right). \tag{5.2.21d}
\end{aligned}$$

Using the above identities, equation (5.2.20) can be written as:

$$\begin{aligned}
& \left(m; \left(\alpha + \frac{1}{2} \text{div}[\mathbf{v}] \right) m \right) + (m; (\alpha + \text{div}[\mathbf{v}]) \Phi_{\max}) + (\text{grad}[m]; \mathbf{D} \text{grad}[m]) \\
& + \left(m; \frac{|\mathbf{v} \bullet \hat{\mathbf{n}}|}{2} m \right)_{\Gamma^q} - \left(m; \left(\frac{1 - \text{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) (\mathbf{v} \bullet \hat{\mathbf{n}}) \Phi_{\max} \right)_{\Gamma^q} \\
& = (m; f) - (m; q^p)_{\Gamma^q}. \tag{5.2.22}
\end{aligned}$$

From equations (5.2.12) and (5.2.13a)–(5.2.13c), it is evident that

$$\begin{aligned} & \left(m; \left(\alpha + \frac{1}{2} \operatorname{div}[\mathbf{v}] \right) m \right) + (m; (\alpha + \operatorname{div}[\mathbf{v}]) \Phi_{\max}) + (\operatorname{grad}[m]; \mathbf{D} \operatorname{grad}[m]) \\ & + \left(m; \frac{|\mathbf{v} \bullet \hat{\mathbf{n}}|}{2} m \right)_{\Gamma^q} - \left(m; \left(\frac{1 - \operatorname{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) \mathbf{v} \bullet \hat{\mathbf{n}} \Phi_{\max} \right)_{\Gamma^q} \geq 0. \end{aligned} \quad (5.2.23)$$

From equation (5.2.14) we have:

$$(m; f) - (m; q^p)_{\Gamma^q} \leq 0. \quad (5.2.24)$$

Therefore, one can conclude that

$$\begin{aligned} & \left(m; \left(\alpha + \frac{1}{2} \operatorname{div}[\mathbf{v}] \right) m \right) + (m; (\alpha + \operatorname{div}[\mathbf{v}]) \Phi_{\max}) + (\operatorname{grad}[m]; \mathbf{D} \operatorname{grad}[m]) \\ & + \left(m; \frac{|\mathbf{v} \bullet \hat{\mathbf{n}}|}{2} m \right)_{\Gamma^q} - \left(m; \left(\frac{1 - \operatorname{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) (\mathbf{v} \bullet \hat{\mathbf{n}}) \Phi_{\max} \right)_{\Gamma^q} = 0. \end{aligned} \quad (5.2.25)$$

In the light of assumption (A3) and equation (5.2.25), we need to have $\operatorname{grad}[m] = 0$ (as $\mathbf{D}(\mathbf{x})$ is bounded below by a constant γ_{lb}). This further implies:

$$m(\mathbf{x}) \equiv \phi_0 \geq 0 \quad \forall \mathbf{x} \in \overline{\Omega}, \quad (5.2.26)$$

where ϕ_0 is a non-negative constant. Since $m(\mathbf{x})|_{\Gamma^c} = 0$ and $\operatorname{meas}(\Gamma^c) > 0$, we have $\phi_0 = 0$. This implies that $c(\mathbf{x}) \leq \Phi_{\max}$, which further implies the validity of the inequality given by equation (5.2.15). Finally, equations (5.2.16) and (5.2.17) are trivial consequences of equation (5.2.15). \square

We have employed the SG_1 formulation in the proof of Theorem 5.2.2. One will come to the same conclusion even under the SG_2 formulation. By reversing the signs in equation (5.2.14), one can easily obtain the following continuous minimum principle.

Corollary 5.2.3 (A continuous minimum principle). *Let assumptions (A1)–(A5) hold and let the unique weak solution $c(\mathbf{x})$ of equations (5.2.1a)–(5.2.1c) belong to $C^1(\Omega) \cap C^0(\overline{\Omega})$. If $f(\mathbf{x}) \in L_2(\Omega)$ and $q^p(\mathbf{x}) \in L_2(\Gamma^q)$ satisfy*

$$f(\mathbf{x}) \geq 0 \quad \text{a.e. in } \Omega \text{ and} \quad (5.2.27a)$$

$$q^p(\mathbf{x}) \leq 0 \quad \text{a.e. on } \Gamma_+^q \cup \Gamma_-^q \quad (5.2.27b)$$

then $c(\mathbf{x})$ satisfies a continuous minimum principle of the following form:

$$\min_{\mathbf{x} \in \overline{\Omega}} [c(\mathbf{x})] \geq \min \left[0, \min_{\mathbf{x} \in \Gamma^c} [c^p(\mathbf{x})] \right]. \quad (5.2.28)$$

In particular, if $c^p(\mathbf{x}) \leq 0$ then

$$\min_{\mathbf{x} \in \overline{\Omega}} [c(\mathbf{x})] = \min_{\mathbf{x} \in \Gamma^c} [c^p(\mathbf{x})]. \quad (5.2.29)$$

If $c^p(\mathbf{x}) \geq 0$ then we have the following non-negative property:

$$\min_{\mathbf{x} \in \overline{\Omega}} [c(\mathbf{x})] \geq 0. \quad (5.2.30)$$

This chapter also deals with transient analysis, and the details are provided in Sections 5.4 and 5.7.

5.2.3 On appropriate Neumann BCs

Many existing numerical formulations Ayub and Masud (2003) and packages such as ABAQUS Aba (2014), ANSYS Ans (2015), COMSOL Com (2014), and MATLAB's PDE Toolbox MAT (2015) do not pose the Neumann BCs in correct form for advection-diffusion equations. These formulations and packages either use the diffusive flux or the total flux on the entire Neumann boundary without discerning the

following situations:

- Do we have inflow (i.e., $\mathbf{v} \bullet \hat{\mathbf{n}} \leq 0$) on the entire Neumann boundary?
- Do we have outflow (i.e., $\mathbf{v} \bullet \hat{\mathbf{n}} \geq 0$) on the entire Neumann boundary?
- Or do we have both inflow and outflow on the Neumann boundary?

These conditions will dictate whether the resulting boundary value problem is well-posed or not. If a numerical formulation does not take into account these conditions, the numerical solution can exhibit instabilities, which will have dire consequences in mixing problems. To illustrate, consider the following 1D boundary value problem:

$$\frac{d}{dx} \left(vc - D \frac{dc}{dx} \right) = 0 \quad \forall x \in (0, L) \text{ and} \quad (5.2.31a)$$

$$c(x=0) = c_0, \quad (5.2.31b)$$

where v , D and c_0 are constants, and L is the length of the domain. We now consider two different cases for the Neumann BC:

$$\left(vc - D \frac{dc}{dx} \right) = q_0 \quad \text{at } x = L \text{ and} \quad (5.2.32a)$$

$$-D \frac{dc}{dx} = q_0 \quad \text{at, } x = L, \quad (5.2.32b)$$

where q_0 is a constant. Equation (5.2.32a) corresponds to the total flux BC while equation (5.2.32b) is the diffusive flux BC. The analytical solutions for these two different Neumann BCs are, respectively, given as:

$$c_1(x) = \frac{1}{v} \left(q_0 + (vc_0 - q_0) e^{\frac{vx}{D}} \right) \text{ and} \quad (5.2.33a)$$

$$c_2(x) = \frac{1}{v} \left(vc_0 + q_0 e^{\frac{-vL}{D}} - q_0 e^{\frac{v(x-L)}{D}} \right). \quad (5.2.33b)$$

The solution $c_1(x)$ blows if $v > 0$, and $c_2(x)$ blows if $v < 0$. On the other hand, the exact solution based on the Neumann BC given in equation (5.2.1c) is well-posed for both inflow and outflow cases.

To summarize, the boundary value problem is well-posed under the prescribed diffusive flux on the entire Neumann boundary if the flow is outflow on the entire Γ^q . The boundary value problem is well-posed under the prescribed total flux on the entire Neumann boundary if the flow is inflow on the entire Γ^q . The Neumann BC given by equation (5.2.1c) is more general, and the boundary value problem under this BC is well-posed even if the Neumann boundary is composed of both inflow and outflow.

5.3 PLAUSIBLE APPROACHES AND THEIR SHORTCOMINGS

There are numerous numerical formulations available in the literature for advective-diffusive-reactive systems. A cavalier look at these formulations can be deceptive, as one may expect more than what these formulations can actually provide. We now discuss some approaches that seem plausible to satisfy the maximum principle and the non-negative constraint for an advective-diffusive-reactive system, and illustrate their shortcomings. This discussion is helpful in two ways. *First*, it sheds light on the complexity of the problem, and can bring out the main contributions made in this chapter. *Second*, the discussion can provide a rationale behind the approach taken in this chapter in order to develop the proposed computational framework. To start with, it is well-known that the single-field Galerkin formulation does not perform well, as it produces spurious node-to-node oscillations on coarse grids Donea and Huerta (2003). The formulation also violates the non-negative constraint and maximum principles for anisotropic medium, and does not possess element-wise species balance property Nakshatrala and Valocchi (2009); Nagarajan and Nakshatrala (2011).

5.3.1 Approach #1: Clipping/cut-off methods

There are various post-processing procedures such as clipping/cut-off methods Burdakov et al. (2012); Kreuzer (2014) to ensure a certain numerical formulation satisfies the non-negative constraint. The idea of these methods is to cut-off the negative values in a numerical solution. This approach predicts erroneous numerical results for highly anisotropic diffusion or advection-diffusion problems Nakshatrala et al. (2013); Karimi and Nakshatrala (2015). Moreover, such a post-processing procedure is a variational crime. Furthermore, clipping/cut-off methods satisfy neither maximum principles nor element-wise local species balance property.

5.3.2 Approach #2: Mesh restrictions

Recently, there has been a surge on the study of constructing meshes to satisfy various discrete maximum principles both within the context of single-field and mixed finite element formulations Huang (2014); Huang and Wang (2014); Mudunuru and Nakshatrala (2015). The primary objective of these methods is to develop restrictions on the computational meshes to meet the underlying principles. However, it should be noted that there are various drawbacks for these methods. The important ones are described as follows:

- (i) Most of these mesh restriction methods are for simplicial meshes (such as three-node triangular element and four-node tetrahedral element). Extending these results to non-simplicial elements is not trivial or may not be possible.
- (ii) The boundary conditions are restricted to only Dirichlet on the entire boundary of the domain. Incorporating mixed boundary conditions or a general Neumann BC given by equation (5.2.1c) has not been addressed.
- (iii) Generating a DMP-based mesh for complex domains is extremely difficult and sometimes impossible.

- (iv) For highly advection-dominated and reaction-dominated problems, we need a highly refined DMP-based meshes. Constructing such meshes is computationally intensive.
- (v) Even though the mesh restriction conditions put forth for the weak Galerkin method by Huang and Wang Huang and Wang (2014) is locally conservative, it is restricted to pure anisotropic diffusion equations. Generalizing it to obtain locally conservative DMP-based meshes for anisotropic ADR equations is not apparent. Moreover, it still suffers from the above set of drawbacks.

5.3.3 Approach #3: Using non-negative methodologies for diffusion equations

Recently, optimization-based finite element methods Liska and Shashkov (2008); Nakshatrala and Valocchi (2009); Nagarajan and Nakshatrala (2011); Nakshatrala et al. (2013) are proposed to satisfy the non-negative constraint and maximum principles for diffusion-type equations. These non-negative methodologies are for self-adjoint operators and are constructed by invoking Vainberg's theorem Vainberg (1964). That is, they utilize the fact that there exists a scalar functional such that the Gâteaux variation of this functional provides the weak formulation and the Euler-Lagrange equations provide the corresponding governing equations for the diffusion problem. Corresponding to this continuous variational/minimization functional, a discrete non-negative constrained optimization-based finite element method is developed. Unfortunately, such a variational principle based on Vainberg's theorem does not exist for the Galerkin weak formulation for an ADR equation, as the spatial operator is non-self-adjoint Nakshatrala and Valocchi (2010).

5.3.4 Approach #4: Posing the discrete equations as a P -LCP

Let h be the maximum element size, $\|\mathbf{v}\|_{\infty,\Omega}$ be the maximum value for advection velocity field, $\alpha_{\infty,\Omega}$ be the maximum value for linear reaction coefficient, and λ_{\min} be the minimum eigenvalue of $\mathbf{D}(\mathbf{x})$ in the entire domain. Mathematically, these quantities are defined as:

$$h := \max_{\Omega_e \in \Omega_h} [h_{\Omega_e}], \quad (5.3.1a)$$

$$\|\mathbf{v}\|_{\infty,\Omega} := \max_{1 \leq i \leq nd} [|(\mathbf{v}(\mathbf{x}))_i|] \quad \forall \mathbf{x} \in \Omega, \quad (5.3.1b)$$

$$\alpha_{\infty,\Omega} := \max_{\mathbf{x} \in \Omega} [\alpha(\mathbf{x})], \quad (5.3.1c)$$

$$\lambda_{\min} := \min_{\mathbf{x} \in \Omega} [\lambda_{\min,\mathbf{D}(\mathbf{x})}], \text{ and} \quad (5.3.1d)$$

$$\lambda_{\max} := \max_{\mathbf{x} \in \Omega} [\lambda_{\max,\mathbf{D}(\mathbf{x})}], \quad (5.3.1e)$$

where Ω_h is a regular linear finite element partition of the domain Ω such that $\overline{\Omega}_h = \bigcup_{e=1}^{N_{ele}} \overline{\Omega}_e$. “ N_{ele} ” is the total number of discrete non-overlapping open sub-domains. The boundary of Ω_e is denoted as $\partial\Omega_e := \overline{\Omega}_e - \Omega_e$. h_{Ω_e} is the diameter of element Ω_e . $\lambda_{\min,\mathbf{D}(\mathbf{x})}$ and $\lambda_{\max,\mathbf{D}(\mathbf{x})}$ are, respectively, the minimum and maximum eigenvalue of $\mathbf{D}(\mathbf{x})$ at a given point $\mathbf{x} \in \Omega$. Correspondingly, the element Péclet number $\mathbb{P}e_h$ and the element Damköhler number $\mathbb{D}a_h$ are defined as:

$$\mathbb{P}e_h := \frac{\|\mathbf{v}\|_{\infty,\Omega} h}{2\lambda_{\min}} \text{ and} \quad (5.3.2a)$$

$$\mathbb{D}a_h := \frac{\alpha_{\infty,\Omega} h^2}{\lambda_{\min}}. \quad (5.3.2b)$$

Herein, $\mathbb{D}a_h$ is defined based on linear reaction coefficient and diffusivity. However, it should be noted that there are various ways to construct different types of element Damköhler numbers (for instance, see Reference Chung (2010) for isotropic diffusivity).

After low-order finite element discretization of either SG_1 or SG_2 , the discrete equations for the ADR boundary value problem take the following form:

$$\mathbf{K}\mathbf{c} = \mathbf{f}, \quad (5.3.3)$$

where \mathbf{K} is the stiffness matrix (which is neither symmetric nor positive definite), \mathbf{c} is the vector containing nodal concentrations, and \mathbf{f} is the volumetric source vector. The matrix \mathbf{K} is of size $ncdofs \times ncdofs$, where “ $ncdofs$ ” denotes the number of free degrees-of-freedom for the concentration. The vectors \mathbf{c} and \mathbf{f} are of size $ncdofs \times 1$.

In the rest of this chapter, the symbols \succeq and \preceq will be used to denote the component-wise comparison of vectors and matrices. That is, given two vectors \mathbf{a} and \mathbf{b} , $\mathbf{a} \preceq \mathbf{b}$ means that $(\mathbf{a})_i \leq (\mathbf{b})_i$ for all i . Likewise, given two matrices \mathbf{A} and \mathbf{B} , $\mathbf{A} \preceq \mathbf{B}$ means that $(\mathbf{A})_{ij} \leq (\mathbf{B})_{ij}$ for all i and j . The mathematical means of the symbols \succeq , \prec and \succ should now be obvious. We shall use $\mathbf{0}$ and \mathbf{O} to denote zero vector and zero matrix, respectively.

Definition 5.3.1 (*P-matrix, Z-matrix, and M-matrix*). A matrix $\mathbf{A} \in \mathbb{R}^{nd \times nd}$ is a *P-matrix* if $\frac{1}{2}(\mathbf{A} + \mathbf{A}^T)$ is positive-definite. The matrix is a *Z-matrix* if $(\mathbf{A})_{ij} \leq 0$, where $i \neq j$ and $i, j = 1, \dots, nd$. The matrix is an *M-matrix* if \mathbf{A} is a *P-matrix* and a *Z-matrix*.

Definition 5.3.2 (*Coarse mesh demarcation for anisotropic ADR equations*). A regular low-order finite element computational mesh Ω_h is said to be coarse with respect to

- (a) spurious oscillations if $\text{Pe}_h > 1$
- (b) spurious oscillations and large linear reaction coefficient if $\text{Pe}_h > 1$ and $\text{Da}_h > 1$
- (c) spurious oscillations, large linear reaction coefficient, and a discrete maximum principle if the stiffness matrix \mathbf{K} associated with either SG_1 or SG_2 is not an

M-matrix

It can be easily shown through counterexamples that the stiffness matrix \mathbf{K} for ADR equation will not always be a Z -matrix. We shall now provide two such counterexamples. The first counterexample is the low-order finite element discretization based on two-node linear element for the following 1D ADR equation (with constant velocity, diffusivity, and linear reaction coefficients):

$$\alpha c + v \frac{dc}{dx} - D \frac{d^2c}{dx^2} = f(x) \quad \forall x \in \Omega := (0, 1) \text{ and} \quad (5.3.4a)$$

$$c(x) = c^p(x) \quad \forall x \in \partial\Omega := \{0, 1\} \quad (5.3.4b)$$

with $\alpha \geq 0$, $D > 0$, and $v \in \mathbb{R}$. The entries of stiffness matrix \mathbf{K} for an i^{th} intermediate node (using equal-sized two-node linear finite element) is given as follows:

$$\frac{\alpha h}{6} \begin{bmatrix} 1 & 4 & 1 \end{bmatrix} \begin{Bmatrix} c_{i-1} \\ c_i \\ c_{i+1} \end{Bmatrix} + \frac{v}{2} \begin{bmatrix} -1 & 0 & 1 \end{bmatrix} \begin{Bmatrix} c_{i-1} \\ c_i \\ c_{i+1} \end{Bmatrix} + \frac{D}{h} \begin{bmatrix} -1 & 2 & -1 \end{bmatrix} \begin{Bmatrix} c_{i-1} \\ c_i \\ c_{i+1} \end{Bmatrix}. \quad (5.3.5)$$

On trivial manipulations on equation (5.3.5), it is evident that the stiffness matrix is a Z -matrix if and only if the following condition is satisfied:

$$h \leq h_{\max} := \frac{12D}{3|v| + \sqrt{9v^2 + 24\alpha D}} \quad (5.3.6)$$

which is not always the case. The second counterexample is based on a simplicial finite element discretization (e.g., three-node triangular/four-node tetrahedron element) of ADR equation with Dirichlet BCs on the entire boundary. If any nd -simplicial mesh does not satisfy the following condition then \mathbf{K} is not a Z -matrix (Mudunuru and

Nakshatrala, 2015, Theorem 4.3):

$$0 < \frac{h_p \|\mathbf{v}\|_{\infty, \bar{\Omega}_e}}{(nd+1) \Lambda_{\min, \tilde{\mathbf{D}}_{\Omega_e}}} + \frac{h_p h_q \alpha_{\infty, \bar{\Omega}_e}}{(nd+1)(nd+2) \Lambda_{\min, \tilde{\mathbf{D}}_{\Omega_e}}} \leq \cos(\beta_{pq, \tilde{\mathbf{D}}_{\Omega_e}^{-1}})$$

$$\forall p, q = 1, 2, \dots, nd+1, p \neq q, \Omega_e \in \Omega_h, \quad (5.3.7)$$

where p and q are the any two given arbitrary vertices of Ω_e . $\tilde{\mathbf{D}}_{\Omega_e}$ is the integral element average anisotropic diffusivity. $\Lambda_{\min, \tilde{\mathbf{D}}_{\Omega_e}}$ denotes the minimum eigenvalue of $\tilde{\mathbf{D}}_{\Omega_e}$. h_p and h_q are the perpendicular distance (or the height) from vertices p and q to their respective opposite faces. $\beta_{pq, \tilde{\mathbf{D}}_{\Omega_e}^{-1}}$ is the dihedral angle measured in $\tilde{\mathbf{D}}_{\Omega_e}^{-1}$ metric between two faces opposite to vertices p and q of an element Ω_e .

Proposition 5.3.3 (*P*-matrix linear complementarity problem for ADR equations).

Given that assumptions (A1)–(A5) hold, then the stiffness matrix \mathbf{K} associated with low-order finite element discretization of either SG₁ or SG₂ is a P-matrix. Furthermore, if \mathbf{c} has to be DMP-preserving on any arbitrary coarse mesh, then the resulting constrained discrete equations of single-field Galerkin formulation can be posed as a P-LCP.

Proof. We prove only for SG₁ formulation and extending it to SG₂ is a trivial manipulation. The symmetric part of the element stiffness matrix \mathbf{K}_e is given as:

$$\begin{aligned} \frac{1}{2} (\mathbf{K}_e + \mathbf{K}_e^T) &= \int_{\tilde{\Omega}_e} \left(\alpha(\mathbf{x}) + \frac{1}{2} \operatorname{div}[\mathbf{v}(\mathbf{x})] \right) \mathbf{N}^T \mathbf{N} \, d\Omega_e + \int_{\tilde{\Omega}_e} \mathbf{B} \mathbf{D}(\mathbf{x}) \mathbf{B}^T \, d\Omega_e \\ &\quad + \int_{\Gamma_e^q} \frac{1}{2} |\mathbf{v} \cdot \hat{\mathbf{n}}| \mathbf{N}^T \mathbf{N} \, d\Gamma_e^q, \end{aligned} \quad (5.3.8)$$

where \mathbf{N} is row vector containing shape functions and $\mathbf{B} = (\mathbf{D}\mathbf{N})\mathbf{J}^{-1}$ (see Appendix 7.1 for details on $\mathbf{D}\mathbf{N}$ and \mathbf{J}). From equation (5.3.8) and assumptions (A1)–(A5), it is evident $\frac{1}{2} (\mathbf{K}_e + \mathbf{K}_e^T)$ is positive semi-definite. Furthermore, the assembly procedure ensures that the global stiffness matrix \mathbf{K} is positive definite (Wathen, 1989, Section

2 and Section 3). As the mesh is coarse, \mathbf{K} is not an M -matrix. But we want \mathbf{c} to satisfy the DMP constraints. Hence, this results in the following constrained discrete system of equations:

$$\mathbf{K}\mathbf{c} = \mathbf{f} + \boldsymbol{\lambda}, \quad (5.3.9a)$$

$$\boldsymbol{\lambda} \succeq \mathbf{0}, \quad (5.3.9b)$$

$$\mathbf{c} \succeq \mathbf{0}, \text{ and} \quad (5.3.9c)$$

$$\boldsymbol{\lambda} \bullet \mathbf{c} = 0. \quad (5.3.9d)$$

As \mathbf{K} is a P -matrix, the above system is a P -matrix linear complementarity problem. This completes the proof. \square

It needs to be emphasized that solving a LCP problem with P -matrix is, in general, NP-hard Rüst (2007). Therefore, posing the discrete equations as a LCP problem and numerically solving it is not a viable approach, especially for large-scale ADR problems with high $\mathbb{P}e_h$. Moreover, it is not always feasible to find a DMP-based h -refined mesh that will produce accurate results for ADR equation for sufficiently high element Péclet number and element Damköhler number.

5.3.5 Approach #5: Posing the discrete problem as constrained normal equations

One way of constructing an optimization-based approach to meet the non-negative constraint is to rewrite the discrete problem as the following constrained normal equations:

$$\underset{\mathbf{c} \in \mathbb{R}^{ncdofs}}{\text{minimize}} \quad \frac{1}{2} \langle \mathbf{c}; \mathbf{K}^T \mathbf{K} \mathbf{c} \rangle - \langle \mathbf{c}; \mathbf{K}^T \mathbf{f} \rangle \text{ and} \quad (5.3.10a)$$

$$\text{subject to} \quad \mathbf{c} \succeq \mathbf{0}, \quad (5.3.10b)$$

where $\langle \bullet; \bullet \rangle$ denotes the standard inner-product in Euclidean spaces. The corresponding first-order optimality conditions can be written as follows:

$$\mathbf{K}^T \mathbf{K} \mathbf{c} = \mathbf{K}^T \mathbf{f} + \boldsymbol{\lambda}, \quad (5.3.11a)$$

$$\mathbf{c} \succeq \mathbf{0}, \quad (5.3.11b)$$

$$\boldsymbol{\lambda} \succeq \mathbf{0}, \text{ and} \quad (5.3.11c)$$

$$\lambda_i c_i = 0 \quad \forall i = 1, \dots, ncdofs. \quad (5.3.11d)$$

If there are no constraints, the optimization problem becomes:

$$\underset{\mathbf{c} \in \mathbb{R}^{ncdofs}}{\text{minimize}} \quad \frac{1}{2} \langle \mathbf{c}; \mathbf{K}^T \mathbf{K} \mathbf{c} \rangle - \langle \mathbf{c}; \mathbf{K}^T \mathbf{f} \rangle. \quad (5.3.12)$$

The first-order optimality condition for the unconstrained discrete optimization problem is:

$$\mathbf{K}^T \mathbf{K} \mathbf{c} = \mathbf{K}^T \mathbf{f}. \quad (5.3.13)$$

In the numerical mathematics literature (e.g., see Demmel (1997)), the above system of equations (5.3.13) is referred to as normal equations. The three main deficiencies of this approach are:

- (i) The constrained optimization-based normal equations method does not avoid node-to-node spurious oscillations. In addition, there is no obvious way of fixing the method to avoid this type of unphysical solutions.
- (ii) It is well-known that the condition number of $\mathbf{K}^T \mathbf{K}$ will be much worse than \mathbf{K} . So the numerical solution will be less reliable, less accurate, and numerically not stable Demmel (1997).

(iii) The discrete optimization problem given by equation (5.3.12) on which non-negative constraints are enforced does not have a corresponding continuous variational/minimization problem.

We shall use the following academic problem to illustrate the aforementioned deficiencies:

$$v \frac{dc}{dx} - D \frac{d^2c}{dx^2} = f \quad \forall x \in (0, 1) \text{ and} \quad (5.3.14a)$$

$$c(x = 0) = c(x = 1) = 0, \quad (5.3.14b)$$

where v , D , and f are assumed to be constants. In our numerical experiment, we have taken the number of mesh elements to be 11, $v/D = 150$, and $f = 1$. The element Péclet number ($\mathbb{P}e_h = \frac{vh}{2D}$) is approximately 6.82. Since it is greater than unity, there will be spurious node-to-node oscillations under the standard single-field Galerkin formulation. From Figure 5.2, it is evident that the normal equations approach does not eliminate the spurious node-to-node oscillations. The condition number of the stiffness matrix under the standard single-field Galerkin formulation is 8.41, whereas the condition number of the stiffness matrix under the normal equations approach is 70.69. For small element Péclet numbers, deficiency (i) can be avoided. But deficiencies (ii) and (iii) will still be present and cannot be circumvented. *Hence, posing the discrete problem as constrained normal equations is not a viable approach to meet maximum principles and the non-negative constraint.*

5.4 PROPOSED COMPUTATIONAL FRAMEWORK

We employ least-squares formalism to develop a class of structure-preserving numerical formulations whose solutions satisfy DMP, LSB, and GSB. The formulations are built based on minimization of unconstrained/constrained quadratic least-squares

functionals. In a least-squares-based finite element formulation, a non-physical least-squares functional is constructed in terms of the sum of the squares of the residuals in an appropriate norm. These residuals are based on the underlying governing equations. However, it should be noted that LSFEMs are different from the Galerkin least-squares or stabilized mixed methods, where least-squares terms are added locally or globally to variational problems.

The success of LSFEM is due to the rich mathematical foundations that influence both the analysis and the algorithmic development. LSFEM offers several attractive features. The resulting weak formulations are coercive. Hence, a unique global minimizer exists for the least-squares functional and this minimizer coincides with the exact solution. Conforming finite element discretizations of least-squares functionals leads to stable and (eventually) optimally accurate numerical solutions. For mixed LSFEM-based formulations, equal order interpolation can be used for all the unknowns, which is computationally the most convenient. The resulting algebraic system is symmetric and positive definite. Thus, the discrete system can be solved using standard and robust iterative numerical methods. For more details on LSFEM for various applications, see Bochev and Gunzberger Bochev and Gunzburger (2009)

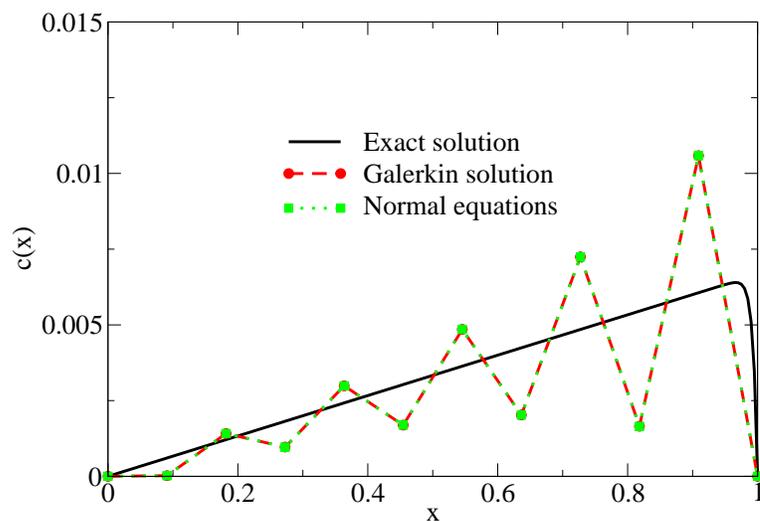


Figure 5.2: Academic problem: This figure compares the numerical solution with the exact solution.

and Jiang Jiang (1998).

5.4.1 Design synopsis of the proposed numerical methodology

The central idea of the proposed computational framework is to constrain a least-squares functional with LSB and non-negative constraints. The main steps involved in the design of the proposed computational framework are:

- (i) The governing equations of the ADR problem are written in first-order mixed form.
- (ii) We construct a stabilized least-squares functional for these first-order governing equations.
- (iii) We construct algebraic equality constraints to enforce element-wise/local species balance (LSB).
- (iv) We enforce bound constraints to the constructed LSFEM to meet maximum principles and the non-negative constraint in the discrete setting. In order to achieve this, we shall use low-order finite element interpolation for $c(\mathbf{x})$.

The first-order mixed form of the governing equations can be written as:

$$\mathbf{q}(\mathbf{x}) - \mathbf{v}(\mathbf{x})c(\mathbf{x}) + \mathbf{D}(\mathbf{x})\text{grad}[c(\mathbf{x})] = \mathbf{0} \quad \text{in } \Omega, \quad (5.4.1a)$$

$$\text{div}[\mathbf{q}(\mathbf{x})] = f(\mathbf{x}) - \alpha(\mathbf{x})c(\mathbf{x}) \quad \text{in } \Omega, \quad (5.4.1b)$$

$$c(\mathbf{x}) = c^p(\mathbf{x}) \quad \text{on } \Gamma^c, \text{ and} \quad (5.4.1c)$$

$$\left(\mathbf{q}(\mathbf{x}) - \left(\frac{1 + \text{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) \mathbf{v}(\mathbf{x})c(\mathbf{x}) \right) \bullet \hat{\mathbf{n}}(\mathbf{x}) = q^p(\mathbf{x}) \quad \text{on } \Gamma^q. \quad (5.4.1d)$$

The bound constraints to meet discrete maximum principles take the following form:

$$c_{\min} \mathbf{1} \preceq \mathbf{c} \preceq c_{\max} \mathbf{1} \quad \text{in } \bar{\Omega}_h, \quad (5.4.2)$$

where c_{\min} and c_{\max} are the minimum and maximum concentration values possible in $\overline{\Omega}$. The LSB equality constraints can be constructed in two different ways. The first approach is based on the integral statement of the balance of species on an element, and takes the following mathematical form:

$$\int_{\Omega_e} \alpha(\mathbf{x})c(\mathbf{x}) \, d\Omega_e + \int_{\partial\Omega_e} \mathbf{q}(\mathbf{x}) \bullet \hat{\mathbf{n}}(\mathbf{x}) \, d\Gamma_e = \int_{\Omega_e} f(\mathbf{x}) \, d\Omega_e. \quad (5.4.3)$$

The second approach is to enforce equation (5.4.1b) in each mesh element $\overline{\Omega}_e$ in an integral form:

$$\int_{\Omega_e} \alpha(\mathbf{x})c(\mathbf{x}) \, d\Omega_e + \int_{\Omega_e} \operatorname{div}[\mathbf{q}(\mathbf{x})] \, d\Omega_e = \int_{\Omega_e} f(\mathbf{x}) \, d\Omega_e. \quad (5.4.4)$$

One can obtain equation (5.4.3) by applying the divergence theorem to equation (5.4.4), which means that these two approaches are equivalent in the continuous setting. This will not always be the case in the discrete setting. In the case of simplicial and non-simplicial low-order finite elements, these approaches are equivalent. However, these two approaches can be different in the case of higher-order finite elements. This is because in certain higher-order finite elements (e.g., nine-node quadrilateral element), not all the nodes are on the boundary of the element. The flux at an interior node contributes to the second integral in equation (5.4.4) but not to the corresponding term in equation (5.4.3). These issues are beyond the scope of this chapter. Herein, we take the first approach given by equation (5.4.3).

We next construct two different least-squares functionals and analyze the influence of various constraints on the performance of these LSFEMs. It should be noted that Hsieh and Yang Hsieh and Yang (2009) have proposed similar least-squares functionals, but they considered homogeneous isotropic steady-state advection-diffusion equations. Moreover, even in the simple setting of isotropic diffusivity, they did

not consider general Neumann BCs, spatially varying velocity fields, simplicial vs. non-simplicial elements, or the effects of DMPs and LSB on the performance of the least-squares functionals. This chapter investigates all the mentioned aspects: we incorporate anisotropy, heterogeneity, transient effects, linear reaction terms, non-solenoidal spatially varying velocity fields, and DMP and LSB constraints.

5.4.2 Weighted primitive LSFEM

The weighted primitive LSFEM is the standard way of constructing a LSFEM-based formulation. It does not contain any additional stabilization terms. The weighted primitive least-squares functional $\mathfrak{F}_{\text{Prim}}(c, \mathbf{q}) : \mathcal{C} \times \mathcal{Q} \rightarrow \mathbb{R}$ in L_2 -norm can be written as

$$\begin{aligned} \mathfrak{F}_{\text{Prim}}(c, \mathbf{q}) := & \frac{1}{2} \left\| \mathbf{A}(\mathbf{x}) (\mathbf{q} - c\mathbf{v} + \mathbf{D}\text{grad}[c]) \right\|_{\Omega}^2 \\ & + \frac{1}{2} \left\| \beta(\mathbf{x}) (\alpha c + \text{div}[\mathbf{q}] - f) \right\|_{\Omega}^2 \\ & + \frac{1}{2} \left\| \left(\mathbf{q} - \left(\frac{1 + \text{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) c\mathbf{v} \right) \bullet \hat{\mathbf{n}} - q^{\text{P}} \right\|_{\Gamma^q}^2, \end{aligned} \quad (5.4.5)$$

where the second-order tensor $\mathbf{A}(\mathbf{x})$ and the scalar function $\beta(\mathbf{x})$ are the weights, which are defined as follows:

$$\mathbf{A}(\mathbf{x}) = \begin{cases} \mathbf{I} & \text{LS Type-1} \\ \mathbf{D}^{-1/2}(\mathbf{x}) & \text{LS Type-2} \end{cases} \quad \text{and} \quad (5.4.6a)$$

$$\beta(\mathbf{x}) = \begin{cases} 1 & \text{LS Type-1} \\ 1 & \text{if } \alpha(\mathbf{x}) = 0 \\ \alpha^{-1/2}(\mathbf{x}) & \text{if } \alpha(\mathbf{x}) \neq 0 \end{cases} \quad \text{LS Type-2.} \quad (5.4.6b)$$

A corresponding weak form can be obtained by setting the Gâteaux variation of the functional (5.4.5) to zero. We shall show in Sections 5.6 and 5.7 that a naive way of

constructing LSFEM formulation, just like the weighted primitive LSFEM, does not perform well for advection-dominated ADR problems. Moreover, enforcing LSB and DMP constraints do not seem to have a profound effect. In order to adequately capture steep boundary and interior layers, we introduce an alternate stabilized LSFEM formulation, which will be referred to as the weighted negatively stabilized streamline diffusion LSFEM. This stabilized LSFEM formulation will be able to handle a wide spectrum of ADR problems ranging from advection-dominated to reaction-dominated problems.

5.4.3 Weighted negatively stabilized streamline diffusion LSFEM

The underlying idea of the proposed stabilized LSFEM formulation is to combine the streamline diffusion and stabilized Galerkin formulations. This is motivated by the prior studies that combining these two formulations exhibit enhanced stability (for example, see Lazarov et al. (1997); Hsieh and Yang (2009)). In this formulation, we introduce a small element-wise stabilization parameter δ_{Ω_e} to correct $\mathbf{q}(\mathbf{x})$ in the streamline direction by adding second-order derivatives of $c(\mathbf{x})$. The modified flux along the streamline direction takes the following form:

$$\mathbf{q} = c\mathbf{v} - \mathbf{D}\text{grad}[c] + \delta_{\Omega_e}\mathbf{v}(\text{div}[c\mathbf{v} - \mathbf{D}\text{grad}[c]]). \quad (5.4.7)$$

Correspondingly, the species balance equation accounting for these corrections will be:

$$\alpha c + \text{div}[\mathbf{q}] = f + f_{\delta_{\Omega_e}}, \quad (5.4.8)$$

where

$$f_{\delta_{\Omega_e}} := \delta_{\Omega_e}(\text{grad}[f - \alpha c] \bullet \mathbf{v} + \text{div}[\mathbf{v}](f - \alpha c)). \quad (5.4.9)$$

The modification to the flux (given by equations (5.4.7)–(5.4.9)) will present two different ways of constructing Neumann BCs.

The first way utilizes the quantities $\mathbf{q}(\mathbf{x})$, $c(\mathbf{x})$, $\alpha(\mathbf{x})$, and $f(\mathbf{x})$, and takes the following form:

$$\left(\mathbf{q} - \left(\frac{1 + \text{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) c\mathbf{v} - \delta_{\Omega_e} (f - \alpha c) \mathbf{v} \right) \bullet \hat{\mathbf{n}}(\mathbf{x}) = q^p(\mathbf{x}) \quad \text{on } \Gamma^q. \quad (5.4.10)$$

The second way utilizes $\mathbf{q}(\mathbf{x})$, $c(\mathbf{x})$, and the first and second derivatives of $c(\mathbf{x})$. The corresponding expression for Neumann BCs takes the following form:

$$\left(\mathbf{q} - \left(\frac{1 + \text{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) c\mathbf{v} - \delta_{\Omega_e} (\text{div}[c\mathbf{v} - \mathbf{D}\text{grad}[c]] \mathbf{v}) \right) \bullet \hat{\mathbf{n}}(\mathbf{x}) = q^p(\mathbf{x}) \quad \text{on } \Gamma^q. \quad (5.4.11)$$

In the continuous setting, equations (5.4.10) and (5.4.11) are equivalent. However, in the discrete setting, the performance of these equations can be different based on the kind of (finite) element being employed. For example, for simplicial elements (such as three-node triangular (T3) element and four-node tetrahedral (T4) element) and four-node quadrilateral (Q4) element, both $\text{div}[\text{grad}[c(\mathbf{x})]]$ and $\text{grad}[\text{grad}[c(\mathbf{x})]]$ are zero for $\bar{\Omega}_e \in \Gamma^q$. This is because the Hessian of \mathbf{N} , which is \mathbf{DDN} , is a zero matrix for both two-node linear (L2) and three-node triangular elements. For more details, see Appendix 7.1. But, this is not the case with non-simplicial linear finite elements for $nd = 3$ and higher-order finite elements (in any dimension). Hence, the Neumann BCs based on equation (5.4.11) are not always physically consistent. However, Neumann BCs based on equation (5.4.10) are always consistent irrespective of the finite element used. Herein, we have chosen Neumann BCs given by equation (5.4.10).

Based on the above set of equations (5.4.7)–(5.4.10), we construct a L_2 -norm based least-squares functional. Additionally, as in the Galerkin least-squares method,

we add a stabilization term to this functional. This stabilization term is:

$$\frac{1}{2} \sum_{\Omega_e \in \Omega_h} \tau_{\Omega_e} \left\| \operatorname{div}[\mathbf{c}\mathbf{v} - \mathbf{D}\operatorname{grad}[c]] + \alpha c - f \right\|_{\Omega_e}^2. \quad (5.4.12)$$

The least-squares functional for the weighted negatively stabilized streamline diffusion formulation $\mathfrak{F}_{\text{NgStb}}(c, \mathbf{q}) : \mathcal{C} \times \mathcal{Q} \rightarrow \mathbb{R}$ in L_2 -norm takes the following form:

$$\begin{aligned} \mathfrak{F}_{\text{NgStb}}(c, \mathbf{q}) := & \frac{1}{2} \sum_{\Omega_e \in \Omega_h} \left\| \mathbf{A}(\mathbf{x}) \left(\mathbf{q} - \mathbf{c}\mathbf{v} + \mathbf{D}\operatorname{grad}[c] - \delta_{\Omega_e} \mathbf{v} (\operatorname{div}[\mathbf{c}\mathbf{v} - \mathbf{D}\operatorname{grad}[c]]) \right) \right\|_{\Omega_e}^2 \\ & + \frac{1}{2} \sum_{\Omega_e \in \Omega_h} \left\| \beta(\mathbf{x}) (\alpha c + \operatorname{div}[\mathbf{q}] - f - f_{\delta_{\Omega_e}}) \right\|_{\Omega_e}^2 \\ & + \frac{1}{2} \sum_{\Omega_e \in \Gamma^q} \left\| \left(\mathbf{q} - \left(\frac{1 + \operatorname{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) \mathbf{c}\mathbf{v} - \delta_{\Omega_e} (f - \alpha c) \mathbf{v} \right) \bullet \hat{\mathbf{n}} - q^p \right\|_{\Omega_e}^2 \\ & + \frac{1}{2} \sum_{\Omega_e \in \Omega_h} \tau_{\Omega_e} \left\| \operatorname{div}[\mathbf{c}\mathbf{v} - \mathbf{D}\operatorname{grad}[c]] + \alpha c - f \right\|_{\Omega_e}^2. \end{aligned} \quad (5.4.13)$$

The element dependent parameters $\tau_{\Omega_e} \leq 0$ and $\delta_{\Omega_e} \leq 0$ are given as

$$\delta_{\Omega_e} = - \frac{\delta_o \lambda_{\min} h_{\Omega_e}^2}{\left(\lambda_{\max}^2 + \delta_1 \max_{\mathbf{x} \in \overline{\Omega}} [(\alpha + \operatorname{div}[\mathbf{v}])^2] h^2 + \delta_2 \max_{\mathbf{x} \in \overline{\Omega}} [\|\operatorname{div}[\mathbf{D}]\|^2] h^2 \right)} \quad \text{and} \quad (5.4.14a)$$

$$\tau_{\Omega_e} = - \frac{\tau_o \lambda_{\min}^2 h_{\Omega_e}^2}{\left(\lambda_{\max}^2 + \tau_1 \max_{\mathbf{x} \in \overline{\Omega}} [(\alpha + \operatorname{div}[\mathbf{v}])^2] h^2 + \tau_2 \max_{\mathbf{x} \in \overline{\Omega}} [\|\operatorname{div}[\mathbf{D}]\|^2] h^2 \right)}, \quad (5.4.14b)$$

where δ_o , δ_1 , δ_2 , τ_o , τ_1 , and τ_2 are non-negative constants. Section 5.5 provides a thorough mathematical justification behind the above stabilization parameters.

For unconstrained LSFEMs, the errors incurred in satisfying LSB and GSB can be calculated as

$$\epsilon_{\text{LSB}}^{(e)} = \int_{\Omega_e} \alpha(\mathbf{x}) c(\mathbf{x}) \, d\Omega + \int_{\partial\Omega_e} \mathbf{q}(\mathbf{x}) \bullet \hat{\mathbf{n}}(\mathbf{x}) \, d\Gamma - \int_{\Omega_e} f(\mathbf{x}) \, d\Omega \quad \text{and} \quad (5.4.15a)$$

$$\epsilon_{\text{GSB}} = \sum_{e=1}^{N_{\text{ele}}} \epsilon_{\text{LSB}}^{(e)}, \quad (5.4.15b)$$

where $c(\mathbf{x})$ and $\mathbf{q}(\mathbf{x})$ are the solutions obtained by solving a given unconstrained LS-FEM. In numerical h -convergence study, we are interested in the following quantities with respect to h -refinement:

$$\epsilon_{\text{MaxAbsLSB}} = \max_{\Omega_e \in \Omega_h} [|\epsilon_{\text{LSB}}^{(e)}|] \quad \text{and} \quad (5.4.16)$$

$$\epsilon_{\text{AbsGSB}} = |\epsilon_{\text{GSB}}|. \quad (5.4.17)$$

Few remarks about the species balance are in order. In writing equation (5.4.17), we have assumed that the mesh is conforming, and the test and trial functions belong to $C^0(\Omega)$ (i.e., there is inter-element continuity of the functions). Under the proposed computational framework, we place explicit (equality) constraints to meet $\epsilon_{\text{LSB}}^{(e)} = 0 \quad \forall e = 1, \dots, N_{ele}$. By meeting the local species balance, the global species balance is trivially met.

5.4.4 Discrete equations

Let \mathbf{K}_{cc} denote the stiffness matrix obtained by lower-order finite element discretization of the LSFEM terms involving $c(\mathbf{x})$ and $w(\mathbf{x})$. Similarly, we can define the stiffness matrices $\mathbf{K}_{c\mathbf{q}}$, $\mathbf{K}_{\mathbf{q}c}$, and $\mathbf{K}_{\mathbf{q}\mathbf{q}}$. The load vectors are denoted by \mathbf{r}_c and $\mathbf{r}_{\mathbf{q}}$, respectively. These vectors are obtained from the finite element discretization of the LSFEM terms involving $w(\mathbf{x})$ and $\mathbf{p}(\mathbf{x})$. It should be noted that the stiffness matrices \mathbf{K}_{cc} and $\mathbf{K}_{\mathbf{q}\mathbf{q}}$ are symmetric and positive definite. Furthermore, $\mathbf{K}_{\mathbf{q}c} = \mathbf{K}_{c\mathbf{q}}^T$.

The corresponding constrained optimization problem in the discrete setting for

the proposed locally conservative DMP-preserving LSFEMs can be written as follows:

$$\underset{\substack{\mathbf{c} \in \mathbb{R}^{ncdofs} \\ \mathbf{q} \in \mathbb{R}^{nqdots}}}{\text{minimize}} \quad \frac{1}{2} \langle \mathbf{c}; \mathbf{K}_{cc} \mathbf{c} \rangle + \langle \mathbf{c}; \mathbf{K}_{cq} \mathbf{q} \rangle + \frac{1}{2} \langle \mathbf{q}; \mathbf{K}_{qq} \mathbf{q} \rangle - \langle \mathbf{c}; \mathbf{r}_c \rangle - \langle \mathbf{q}; \mathbf{r}_q \rangle \text{ and} \quad (5.4.18a)$$

$$\text{subject to} \quad \begin{cases} \mathbf{A}_c \mathbf{c} + \mathbf{A}_q \mathbf{q} = \mathbf{b}_f \\ c_{\min} \mathbf{1} \preceq \mathbf{c} \preceq c_{\max} \mathbf{1} \end{cases}, \quad (5.4.18b)$$

where “*nqdots*” denotes the number of degrees-of-freedom for the flux vector, and “*ncdofs*” denotes the number of degrees-of-freedom for the concentration. The vector of size $ncdofs \times 1$ with all entries to be unity is denoted as $\mathbf{1}$. Recall that $\langle \bullet; \bullet \rangle$ denotes the standard inner-product on the Euclidean spaces. The finite element discretization of the local species balance equation gives rise to the global LSB matrices \mathbf{A}_c and \mathbf{A}_q , and the global LSB vector \mathbf{b}_f . The matrices \mathbf{A}_c and \mathbf{A}_q are of sizes $Nele \times ncdofs$ and $Nele \times nqdots$, respectively. Similar inference can be drawn on the sizes of \mathbf{b}_f , \mathbf{r}_c , \mathbf{r}_q , \mathbf{K}_{cq} , and \mathbf{K}_{qq} . Since $\mathbf{K}_{qc} = \mathbf{K}_{cq}^T$ and the matrices \mathbf{K}_{cc} and \mathbf{K}_{qq} are symmetric and positive definite, the constrained optimization problem (5.4.18a)–(5.4.18b) belongs to *convex quadratic programming* and has a unique global minimizer (Boyd and Vandenberghe (2004)). The corresponding first-order optimality conditions – popularly known as the Karush-Kuhn-Tucker (KKT) conditions – for this discrete

optimization problem take the following form:

$$\mathbf{K}_{cc}\mathbf{c} + \mathbf{K}_{cq}\mathbf{q} = \mathbf{r}_c - \mathbf{A}_c^T\boldsymbol{\lambda}_c + \boldsymbol{\mu}_{\min} - \boldsymbol{\mu}_{\max}, \quad (5.4.19a)$$

$$\mathbf{K}_{cq}^T\mathbf{c} + \mathbf{K}_{qq}\mathbf{q} = \mathbf{r}_q - \mathbf{A}_q^T\boldsymbol{\lambda}_q, \quad (5.4.19b)$$

$$\mathbf{A}_c\mathbf{c} + \mathbf{A}_q\mathbf{q} = \mathbf{b}_f, \quad (5.4.19c)$$

$$\boldsymbol{\mu}_{\min} \succeq \mathbf{0}, \quad (5.4.19d)$$

$$\boldsymbol{\mu}_{\max} \succeq \mathbf{0}, \quad (5.4.19e)$$

$$(\mathbf{c} - c_{\min}\mathbf{1}) \bullet \boldsymbol{\mu}_{\min} = 0, \text{ and} \quad (5.4.19f)$$

$$(c_{\max}\mathbf{1} - \mathbf{c}) \bullet \boldsymbol{\mu}_{\max} = 0, \quad (5.4.19g)$$

where $\boldsymbol{\lambda}_c$ and $\boldsymbol{\lambda}_q$ are the Lagrange multipliers enforcing the LSB equality constraints, which stem from equation (5.4.19c). $\boldsymbol{\mu}_{\min}$ and $\boldsymbol{\mu}_{\max}$ are the KKT multipliers enforcing the DMP inequality constraints given by $c_{\min}\mathbf{1} \preceq \mathbf{c}$ and $\mathbf{c} \preceq c_{\max}\mathbf{1}$. Note that the non-negative constraint is a subset of the DMP inequality constraints. To wit, setting $c_{\min} = 0$ and $c_{\max} = +\infty$ will result in explicit non-negative constraints on the nodal concentrations.

In the next two sections, we illustrate the performance of the proposed computational framework for advection-dominated ADR problems and transport-controlled irreversible fast bimolecular reactions. In all the numerical simulations reported in this chapter, the constrained optimization problem is solved using the MATLAB's MAT (2015) built-in function handler `quadprog`, which has a robust solver based on an interior-point numerical algorithm presented in References Gould and Toint (2004); Mehrotra (1992); Gondzio (1996). One can alternatively employ the optimization solvers from SciPy Jones et al. (2014). The tolerance in the stopping criterion for solving convex quadratic programming problems is taken as $100\epsilon_{\text{mach}}$, where $\epsilon_{\text{mach}} \approx 2.22 \times 10^{-16}$ is the machine precision for a 64-bit machine.

There are various approaches to numerically solve transient diffusion-type systems. It is desirable to have a numerical strategy that converts and utilizes the solvers for steady-state diffusion-type equations to solve transient systems. It has been recently shown that the method of horizontal lines using the backward Euler time-stepping scheme is one of the viable approaches to respect maximum principles and the non-negative constraint in the discrete setting Nakshatrala et al. (2013). The method of horizontal lines discretizes the time domain first, and thereby converts the transient ADR equations at each time-level into a system of governing equations similar to (5.2.1a)–(5.2.1c). This methodology, thus, helps us to use the computational framework provided in Section 5.4. One can employ a numerical procedure similar to Algorithm 1 provided in Reference Nakshatrala et al. (2013) to advance the numerical solution over the time. Numerical results for transient systems are presented in Section 5.7.

5.5 COERCIVITY, ERROR ESTIMATES, AND STABILIZATION PARAMETERS

Herein, we shall establish coercivity and error estimates for the homogeneous weighted LSFEMs (i.e., $c^p(\mathbf{x}) = 0$ on $\partial\Omega$). Based on this mathematical analysis, we obtain the stabilization parameters that are used in this chapter for the weighted negatively stabilized streamline diffusion LSFEM. To this end, let $H^m(\Omega)$ denote the standard Sobolev space for a given non-negative integer m Evans (1998). The associated standard inner product and norm are denoted by $(\bullet; \bullet)_m$ and $\|\bullet\|_m$, respectively. On the function space \mathcal{Q} , the inner product $(\bullet; \bullet)_{\text{div}}$ and norm $\|\bullet\|_{\text{div}}$ are defined as:

$$(\mathbf{p}; \mathbf{q})_{\text{div}} := (\mathbf{p}; \mathbf{q})_0 + (\text{div}[\mathbf{p}]; \text{div}[\mathbf{q}])_0 \quad \forall \mathbf{p}, \mathbf{q} \in \mathcal{Q} \text{ and} \quad (5.5.1a)$$

$$\|\mathbf{p}\|_{\text{div}}^2 := \|\mathbf{p}\|_0^2 + \|\text{div}[\mathbf{p}]\|_0^2 \quad \forall \mathbf{p} \in \mathcal{Q}. \quad (5.5.1b)$$

The Poincaré-Friedrichs inequality takes the following form Bochev and Gunzburger (2009): there exists a constant $C_{pf} > 0$ such that we have

$$\|u\|_0 \leq C_{pf} \|\text{grad}[u]\|_0 \quad \forall u \in H_0^1(\Omega), \quad (5.5.2)$$

Consider the classical weighted primitive least-squares functional $\mathfrak{F}_{\text{Prim}}((c, \mathbf{q}), f)$ given by equation (5.4.5) with $c(\mathbf{x}) = 0$ on $\partial\Omega$. If $(c, \mathbf{q}) \in \mathcal{C} \times \mathcal{Q}$ is an exact solution of the equations (5.2.1a)–(5.2.1c), then (c, \mathbf{q}) must be a unique zero minimizer of $\mathfrak{F}_{\text{Prim}}((c, \mathbf{q}), f)$ on $\mathcal{C} \times \mathcal{Q}$. Hence, for any $\epsilon \in \mathbb{R}$, we have

$$\left. \frac{d}{d\epsilon} \mathfrak{F}_{\text{Prim}}((c, \mathbf{q}) + \epsilon(w, \mathbf{p}), f) \right|_{\epsilon=0} = 0 \quad \forall (w, \mathbf{p}) \in \mathcal{W} \times \mathcal{Q}, \quad (5.5.3)$$

which is identical to

$$\mathfrak{B}_{\text{Prim}}((c, \mathbf{q}); (w, \mathbf{p})) = \mathfrak{L}_{\text{Prim}}((w, \mathbf{p})) \quad \forall (w, \mathbf{p}) \in \mathcal{W} \times \mathcal{Q}, \quad (5.5.4)$$

where $\mathfrak{B}_{\text{Prim}}(\bullet; \bullet)$ and $\mathfrak{L}_{\text{Prim}}(\bullet)$ are the corresponding bilinear and linear forms for the weighted primitive least-squares functional $\mathfrak{F}_{\text{Prim}}$. It should be noted that

$$\mathfrak{B}_{\text{Prim}}((w, \mathbf{p}); (w, \mathbf{p})) = \mathfrak{F}_{\text{Prim}}((w, \mathbf{p}), f = 0) \quad \forall (w, \mathbf{p}) \in \mathcal{W} \times \mathcal{Q}. \quad (5.5.5)$$

Equation (5.5.5) is used to prove coercivity and boundedness estimates for the bilinear form $\mathfrak{B}_{\text{Prim}}$. Now consider the finite element discretization of the equation (5.5.4). Let $\mathcal{C}_h \subseteq \mathcal{C}$, $\mathcal{W}_h \subseteq \mathcal{W}$, and $\mathcal{Q}_h \subseteq \mathcal{Q}$ be the finite element function spaces spanned by piecewise polynomials of degree less than or equal to r over Ω_h . It should be noted that r is an integer and $r \geq 1$. Then, the discrete weighted primitive LSFEM can be

written as follows: Find $(c_h, \mathbf{q}_h) \in \mathcal{C}_h \times \mathcal{Q}_h$ such that

$$\mathfrak{B}_{\text{Prim}}((c_h, \mathbf{q}_h); (w_h, \mathbf{p}_h)) = \mathfrak{L}_{\text{Prim}}((w_h, \mathbf{p}_h)) \quad \forall (w_h, \mathbf{p}_h) \in \mathcal{W}_h \times \mathcal{Q}_h, \quad (5.5.6)$$

where (c_h, \mathbf{q}_h) is the finite element solution with respect to the chosen basis functions spanning the finite element space $\mathcal{C}_h \times \mathcal{Q}_h$. Similar inference holds for $\mathfrak{F}_{\text{NgStb}}((c, \mathbf{q}), f)$, $\mathfrak{B}_{\text{NgStb}}$, and $\mathfrak{L}_{\text{NgStb}}$.

We assume that Ω_h is quasi-uniform Jiang (1998); Bochev and Gunzburger (2009). That is, there exists a constant $\hat{C} > 0$ independent of h such that $h \leq \hat{C}h_{\Omega_e}$ for all $\Omega_e \in \Omega_h$. Additionally, we assume that the following inverse inequality holds on these quasi-uniform meshes. There exists a constant $\tilde{C} > 0$ independent of h such that

$$\tilde{C} \sum_{\Omega_e \in \Omega_h} h_{\Omega_e}^2 \left\| \text{div}[\text{grad}[c_h]] \right\|_{0, \Omega_e}^2 \leq \|\text{grad}[c_h]\|_0^2 \quad \forall c_h \in \mathcal{C}_h \text{ and} \quad (5.5.7a)$$

$$\left\| \mathbf{D} \bullet \text{grad}[\text{grad}[c_h]] \right\|_0 \leq \left\| \text{tr}[\mathbf{D}] \text{tr}[\text{grad}[\text{grad}[c_h]]] \right\|_0 = \text{tr}[\mathbf{D}] \left\| \text{div}[\text{grad}[c_h]] \right\|_0 \quad \forall c_h \in \mathcal{C}_h, \quad (5.5.7b)$$

where $\text{tr}[\bullet]$ is the trace of a matrix. In proposing equation (5.5.7b), we assumed that the Hessian of c_h , $\text{grad}[\text{grad}[c_h]]$, is positive semi-definite.

All the results presented here are applicable for a general anisotropic diffusivity tensor, advection velocity vector field, and linear reaction coefficient. One can obtain simplified results for isotropy by taking $\mathbf{D}(\mathbf{x}) = D(\mathbf{x})\mathbf{I}$, where \mathbf{I} is an identity tensor.

Theorem 5.5.1 (Coercivity for weighted primitive LSFEM). *There exist constants $C_{\text{Prim1}} > 0$ and $C_{\text{Prim2}} > 0$ independent of \mathbf{D} and h such that for all $(w_h, \mathbf{p}_h) \in$*

$\mathcal{W}_h \times \mathcal{Q}_h$:

$$\tilde{\mathfrak{F}}_{\text{Prim}}((w_h, \mathbf{p}_h), f = 0) \geq C_{\text{Prim1}} \gamma_{\min}^2 \lambda_{\min}^2 \|\text{grad}[w_h]\|_0^2 \text{ and} \quad (5.5.8a)$$

$$\tilde{\mathfrak{F}}_{\text{Prim}}((w_h, \mathbf{p}_h), f = 0) \geq C_{\text{Prim2}} \gamma_{\min}^2 \lambda_{\min}^2 \left(\|w_h\|_1^2 + \frac{\|\mathbf{p}_h\|_{\text{div}}^2}{1 + \lambda_{\min}^2 + \lambda_{\max}^2} \right) \quad (5.5.8b)$$

where the positive constant γ_{\min} is:

$$\gamma_{\min} := \min \left[1, \min_{\mathbf{x} \in \bar{\Omega}} [\beta(\mathbf{x})], \min_{\mathbf{x} \in \bar{\Omega}} [\lambda_{\min, \mathbf{A}(\mathbf{x})}] \right], \quad (5.5.9)$$

where $\lambda_{\min, \mathbf{A}(\mathbf{x})}$ is the minimum eigenvalue of $\mathbf{A}(\mathbf{x})$ at a given point $\mathbf{x} \in \bar{\Omega}$.

Proof. Consider the weighted primitive least-squares functional (5.4.5) with $f = 0$. Equation (5.5.9) implies:

$$\begin{aligned} \frac{2\tilde{\mathfrak{F}}_{\text{Prim}}}{\gamma_{\min}^2} &\geq \left\| \mathbf{p}_h - w_h \mathbf{v} + \mathbf{D} \text{grad}[w_h] - \mu \text{grad}[w_h] \right\|_{0, \Omega}^2 + \left\| \alpha w_h + \text{div}[\mathbf{p}_h] - \mu w_h \right\|_{0, \Omega}^2 \\ &\quad + 2\mu (\mathbf{p}_h - w_h \mathbf{v} + \mathbf{D} \text{grad}[w_h]; \text{grad}[w_h])_{0, \Omega} + 2\mu (\alpha w_h + \text{div}[\mathbf{p}_h]; w_h)_{0, \Omega} \\ &\quad - \mu^2 \|w_h\|_{0, \Omega}^2 - \mu^2 \|\text{grad}[w_h]\|_{0, \Omega}^2, \end{aligned} \quad (5.5.10)$$

where μ is a positive constant, which will be determined later. Using Poincaré-Friedrichs inequality and Green's formulae, equation (5.5.10) can be written as

$$\frac{2\tilde{\mathfrak{F}}_{\text{Prim}}}{\gamma_{\min}^2} \geq \mu \left(2\lambda_{\min} - \mu (1 + C_{pf}^2) \right) \|\text{grad}[w_h]\|_{0, \Omega}^2. \quad (5.5.11)$$

We obtain equation (5.5.8a) by choosing

$$\mu = \frac{\lambda_{\min}}{1 + C_{pf}^2}. \quad (5.5.12)$$

There exist two non-negative constants $C_{\mathbf{v}}$ and C_{α} (for instance, $C_{\mathbf{v}} = \max_{\mathbf{x} \in \bar{\Omega}} [\|\mathbf{v}\|^2]$)

and $C_\alpha = \max_{\mathbf{x} \in \bar{\Omega}} [\alpha^2]$) such that

$$\|w_h\|_1^2 = \|w_h\|_0^2 + \|\text{grad}[w_h]\|_0^2 \leq \frac{2(1 + C_{pf}^2)^2}{\gamma_{\min}^2 \lambda_{\min}^2} \mathfrak{F}_{\text{Prim}}, \quad (5.5.13a)$$

$$\begin{aligned} \|\mathbf{p}_h\|_0^2 &\leq 2\|\mathbf{p}_h - w_h \mathbf{v} + \mathbf{D} \text{grad}[w_h]\|_{0,\Omega}^2 + 2\| -w_h \mathbf{v} + \mathbf{D} \text{grad}[w_h]\|_{0,\Omega}^2 \\ &\leq \left(1 + \frac{2C_{\mathbf{v}} C_{pf}^2 (1 + C_{pf}^2)}{\lambda_{\min}^2} + 2(1 + C_{pf}^2) \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \right) \frac{4\mathfrak{F}_{\text{Prim}}}{\gamma_{\min}^2}, \end{aligned} \quad (5.5.13b)$$

$$\begin{aligned} \|\text{div}[\mathbf{p}_h]\|_0^2 &\leq 2\|\alpha w_h + \text{div}[\mathbf{p}_h]\|_{0,\Omega}^2 + 2\|\alpha w_h\|_{0,\Omega}^2 \\ &\leq \left(1 + \frac{C_\alpha C_{pf}^2 (1 + C_{pf}^2)}{\lambda_{\min}^2} \right) \frac{4\mathfrak{F}_{\text{Prim}}}{\gamma_{\min}^2}. \end{aligned} \quad (5.5.13c)$$

It is easy to check that inequalities (5.5.13a)–(5.5.13c) imply inequality (5.5.8b). \square

Theorem 5.5.2 (Coercivity and boundedness estimate for NSSD LSFEM). *Given that equations (5.5.7a)–(5.5.7b) hold. If for each $\Omega_e \in \Omega_h$ we take*

$$\delta_{\Omega_e} = -\frac{\tilde{C} \lambda_{\min} h_{\Omega_e}^2}{4(nd^2 \lambda_{\max}^2 + \tilde{C} C_{pf}^2 \delta_{\alpha \mathbf{v}} h^2 + \tilde{C} \delta_{\mathbf{D}} h^2)} \text{ and} \quad (5.5.14a)$$

$$\tau_{\Omega_e} = -\frac{\tilde{C} \lambda_{\min}^2 h_{\Omega_e}^2}{32(1 + C_{pf}^2)(nd^2 \lambda_{\max}^2 + \tilde{C} C_{pf}^2 \delta_{\alpha \mathbf{v}} h^2 + \tilde{C} \delta_{\mathbf{D}} h^2)} \quad (5.5.14b)$$

then for all $(w_h, \mathbf{p}_h) \in \mathcal{W}_h \times \mathcal{Q}_h$ there exist two constants $C_{\text{NgStb}0} > 0$ and $C_{\text{NgStb}4} > 0$ independent of \mathbf{D} and h such that we have:

Coercivity

$$\begin{aligned} \mathfrak{F}_{\text{NgStb}}((w_h, \mathbf{p}_h), f = 0) &\geq \frac{11\gamma_{\min}^2 \lambda_{\min}^2 \|\text{grad}[w_h]\|_0^2}{32(1 + C_{pf}^2)} \\ &+ \sum_{\Omega_e \in \Omega_h} \frac{\tilde{C} \gamma_{\min}^2 \lambda_{\min}^2 h_{\Omega_e}^2 \|\mathbf{v} \bullet \text{grad}[w_h]\|_{0,\Omega_e}^2}{32(1 + C_{pf}^2)(nd^2 \lambda_{\max}^2 + \tilde{C} C_{pf}^2 \delta_{\alpha \mathbf{v}} h^2 + \tilde{C} \delta_{\mathbf{D}} h^2)}. \end{aligned} \quad (5.5.15)$$

Boundedness estimate

$$\begin{aligned} C_{\text{NgStb1}} \|w_h\|_1^2 + C_{\text{NgStb2}} \|\mathbf{p}_h\|_{\text{div}}^2 + C_{\text{NgStb3}} \|\mathbf{v} \bullet \text{grad}[w_h]\|_0^2 &\leq \mathfrak{F}_{\text{NgStb}}((w_h, \mathbf{p}_h), f = 0) \\ &\leq C_{\text{NgStb4}} \left(\|w_h\|_1^2 + \|\mathbf{p}_h\|_{\text{div}}^2 + \|\mathbf{v} \bullet \text{grad}[w_h]\|_0^2 \right), \end{aligned} \quad (5.5.16)$$

where the constant γ_{\min} is given by equation (5.5.9). The constants $\delta_{\alpha\mathbf{v}}$, $\delta_{\mathbf{D}}$, C_{NgStb1} , C_{NgStb2} , and C_{NgStb3} are given as follows:

$$\delta_{\alpha\mathbf{v}} = \max_{\mathbf{x} \in \Omega} [(\alpha + \text{div}[\mathbf{v}])^2], \quad (5.5.17a)$$

$$\delta_{\mathbf{D}} = \max_{\mathbf{x} \in \Omega} [\|\text{div}[\mathbf{D}]\|^2], \quad (5.5.17b)$$

$$C_{\text{NgStb1}} = C_{\text{NgStb0}} \gamma_{\min}^2 \lambda_{\min}^2, \quad (5.5.17c)$$

$$C_{\text{NgStb2}} = \frac{C_{\text{NgStb0}} \gamma_{\min}^2 \lambda_{\min}^2 \delta_{\alpha\mathbf{v}\mathbf{D}}^2}{(1 + \lambda_{\min}^2 + \lambda_{\max}^2) \delta_{\alpha\mathbf{v}\mathbf{D}}^2 + \delta_{\alpha\mathbf{v}\mathbf{D}} \lambda_{\min}^2 h^2 + \delta_{\alpha\mathbf{v}1} \lambda_{\min}^2 h^4}, \text{ and} \quad (5.5.17d)$$

$$C_{\text{NgStb3}} = \frac{C_{\text{NgStb0}} \gamma_{\min}^2 \lambda_{\min}^2 h^2}{\delta_{\alpha\mathbf{v}\mathbf{D}}}. \quad (5.5.17e)$$

The constants $\delta_{\alpha\mathbf{v}1}$ and $\delta_{\alpha\mathbf{v}\mathbf{D}}$ in the above equations are defined as follows:

$$\delta_{\alpha\mathbf{v}1} = \max_{\mathbf{x} \in \Omega} [(\text{grad}[\alpha] \bullet \mathbf{v} + \alpha \text{div}[\mathbf{v}])^2] \text{ and} \quad (5.5.18a)$$

$$\delta_{\alpha\mathbf{v}\mathbf{D}} = \lambda_{\max}^2 + \delta_{\alpha\mathbf{v}} h^2 + \delta_{\mathbf{D}} h^2 \quad (5.5.18b)$$

Proof. The boundedness estimate is a direct consequence of the triangle inequality. Herein, we shall proceed to show the validity of coercivity estimates, specifically, equation (5.5.15) and the left hand side of (5.5.16). Let $\mu > 0$ be a constant, which will be determined later. Using equation (5.5.9) and (5.4.13) with $f = 0$, we have

$$\begin{aligned}
\frac{2\mathfrak{F}_{\text{NgStb}}}{\gamma_{\min}^2} &\geq \sum_{\Omega_e \in \Omega_h} \left\| \mathbf{p}_h - w_h \mathbf{v} + \mathbf{D} \text{grad}[w_h] - \delta_{\Omega_e} \mathbf{v} (\text{div}[w_h \mathbf{v} - \mathbf{D} \text{grad}[w_h]]) - \mu \text{grad}[w_h] \right\|_{0, \Omega_e}^2 \\
&+ \sum_{\Omega_e \in \Omega_h} \left\| \alpha w_h + \text{div}[\mathbf{p}_h] + \delta_{\Omega_e} \text{div}[\alpha w_h \mathbf{v}] - \mu w_h \right\|_{0, \Omega_e}^2 - \mu^2 \|w_h\|_{0, \Omega}^2 - \mu^2 \|\text{grad}[w_h]\|_{0, \Omega}^2 \\
&+ \sum_{\Omega_e \in \Omega_h} \tau_{\Omega_e} \left\| \alpha w_h + \text{div}[w_h \mathbf{v} - \mathbf{D} \text{grad}[w_h]] \right\|_{0, \Omega_e}^2 \\
&+ \sum_{\Omega_e \in \Omega_h} 2\mu (\alpha w_h + \text{div}[\mathbf{p}_h + \delta_{\Omega_e} \alpha w_h \mathbf{v}]; w_h)_{0, \Omega} \\
&+ \sum_{\Omega_e \in \Omega_h} 2\mu (\mathbf{p}_h - w_h \mathbf{v} + \mathbf{D} \text{grad}[w_h] - \delta_{\Omega_e} \mathbf{v} (\text{div}[w_h \mathbf{v} - \mathbf{D} \text{grad}[w_h]]); \text{grad}[w_h])_{0, \Omega}. \quad (5.5.19)
\end{aligned}$$

Using Theorem 5.5.1, equation (5.5.14a)–(5.5.14b), Cauchy-Schwartz inequality, Poincaré-Friedrichs inequality, Green's formulae, and following inequalities

$$\begin{aligned}
2\tau_{\Omega_e} ((\alpha + \text{div}[\mathbf{v}]) w_h; \mathbf{v} \bullet \text{grad}[w_h])_{0, \Omega_e} &\geq \tau_{\Omega_e} \|(\alpha + \text{div}[\mathbf{v}]) w_h\|_{0, \Omega_e}^2 \\
&+ \tau_{\Omega_e} \|\mathbf{v} \bullet \text{grad}[w_h]\|_{0, \Omega_e}^2, \quad (5.5.20a)
\end{aligned}$$

$$\begin{aligned}
-2\tau_{\Omega_e} ((\alpha + \text{div}[\mathbf{v}]) w_h; \mathbf{D} \bullet \text{grad}[\text{grad}[w_h]])_{0, \Omega_e} &\geq \tau_{\Omega_e} \|(\alpha + \text{div}[\mathbf{v}]) w_h\|_{0, \Omega_e}^2 \\
&+ \tau_{\Omega_e} \|\mathbf{D} \bullet \text{grad}[\text{grad}[w_h]]\|_{0, \Omega_e}^2, \quad (5.5.20b)
\end{aligned}$$

$$\begin{aligned}
-2\tau_{\Omega_e} (\mathbf{v} \bullet \text{grad}[w_h]; \mathbf{D} \bullet \text{grad}[\text{grad}[w_h]])_{0, \Omega_e} &\geq \tau_{\Omega_e} \|\mathbf{v} \bullet \text{grad}[w_h]\|_{0, \Omega_e}^2 \\
&+ \tau_{\Omega_e} \|\mathbf{D} \bullet \text{grad}[\text{grad}[w_h]]\|_{0, \Omega_e}^2, \quad (5.5.20c)
\end{aligned}$$

$$\begin{aligned}
2\tau_{\Omega_e} ((\alpha + \text{div}[\mathbf{v}]) w_h; \text{div}[\mathbf{D}] \bullet \text{grad}[w_h])_{0, \Omega_e} &\geq \tau_{\Omega_e} \|(\alpha + \text{div}[\mathbf{v}]) w_h\|_{0, \Omega_e}^2 \\
&+ \tau_{\Omega_e} \|\text{div}[\mathbf{D}] \bullet \text{grad}[w_h]\|_{0, \Omega_e}^2, \quad (5.5.20d)
\end{aligned}$$

$$\begin{aligned}
-2\tau_{\Omega_e} (\text{div}[\mathbf{D}] \bullet \text{grad}[w_h]; \mathbf{D} \bullet \text{grad}[\text{grad}[w_h]])_{0, \Omega_e} &\geq \tau_{\Omega_e} \|\text{div}[\mathbf{D}] \bullet \text{grad}[w_h]\|_{0, \Omega_e}^2 \\
&+ \tau_{\Omega_e} \|\mathbf{D} \bullet \text{grad}[\text{grad}[w_h]]\|_{0, \Omega_e}^2, \quad \text{and} \quad (5.5.20e)
\end{aligned}$$

$$\begin{aligned}
2\tau_{\Omega_e} (\mathbf{v} \bullet \text{grad}[w_h]; \text{div}[\mathbf{D}] \bullet \text{grad}[w_h])_{0, \Omega_e} &\geq \tau_{\Omega_e} \|\mathbf{v} \bullet \text{grad}[w_h]\|_{0, \Omega_e}^2 \\
&+ \tau_{\Omega_e} \|\text{div}[\mathbf{D}] \bullet \text{grad}[w_h]\|_{0, \Omega_e}^2 \quad (5.5.20f)
\end{aligned}$$

we have the following inequality:

$$\begin{aligned}
\sum_{\Omega_e \in \Omega_h} \tau_{\Omega_e} \left\| \alpha w_h + \operatorname{div}[w_h \mathbf{v} - \mathbf{D} \operatorname{grad}[w_h]] \right\|_{0, \Omega_e}^2 &\geq -\frac{\lambda_{\min}^2}{16(1 + C_{pf}^2)} \\
&- \sum_{\Omega_e \in \Omega_h} \frac{\tilde{C} \lambda_{\min}^2 h_{\Omega_e}^2 \|\mathbf{v} \bullet \operatorname{grad}[w_h]\|_{0, \Omega_e}^2}{16(1 + C_{pf}^2) (nd^2 \lambda_{\max}^2 + \tilde{C} C_{pf}^2 \delta_{\alpha \mathbf{v}} h^2 + \tilde{C} \delta_{\mathbf{D}} h^2)}. \tag{5.5.21}
\end{aligned}$$

Similarly, using the following equality:

$$2\mu \delta_{\Omega_e} (\operatorname{div}[\alpha w_h \mathbf{v}]; w_h)_{0, \Omega_e} = -2\mu \delta_{\Omega_e} (\alpha w_h; \mathbf{v} \bullet \operatorname{grad}[w_h])_{0, \Omega_e} = \mu \delta_{\Omega_e} (\operatorname{div}[\alpha \mathbf{v}] w_h; w_h)_{0, \Omega_e} \tag{5.5.22}$$

in combination with the following inequalities

$$\begin{aligned}
-2\mu \delta_{\Omega_e} ((\alpha + \operatorname{div}[\mathbf{v}]) w_h; \mathbf{v} \bullet \operatorname{grad}[w_h])_{0, \Omega_e} &\geq 2\mu \delta_{\Omega_e} \|(\alpha + \operatorname{div}[\mathbf{v}]) w_h\|_{0, \Omega_e}^2 \\
&+ \frac{\mu \delta_{\Omega_e}}{2} \|\mathbf{v} \bullet \operatorname{grad}[w_h]\|_{0, \Omega_e}^2, \tag{5.5.23a}
\end{aligned}$$

$$\begin{aligned}
2\mu \delta_{\Omega_e} (\operatorname{div}[\mathbf{D}] \bullet \operatorname{grad}[w_h]; \mathbf{v} \bullet \operatorname{grad}[w_h])_{0, \Omega_e} &\geq 2\mu \delta_{\Omega_e} \|\operatorname{div}[\mathbf{D}] \bullet \operatorname{grad}[w_h]\|_{0, \Omega_e}^2 \\
&+ \frac{\mu \delta_{\Omega_e}}{2} \|\mathbf{v} \bullet \operatorname{grad}[w_h]\|_{0, \Omega_e}^2, \tag{5.5.23b}
\end{aligned}$$

$$\begin{aligned}
2\mu \delta_{\Omega_e} (\mathbf{D} \bullet \operatorname{grad}[\operatorname{grad}[w_h]]; \mathbf{v} \bullet \operatorname{grad}[w_h])_{0, \Omega_e} &\geq 2\mu \delta_{\Omega_e} \|\mathbf{D} \bullet \operatorname{grad}[\operatorname{grad}[w_h]]\|_{0, \Omega_e}^2 \\
&+ \frac{\mu \delta_{\Omega_e}}{2} \|\mathbf{v} \bullet \operatorname{grad}[w_h]\|_{0, \Omega_e}^2 \tag{5.5.23c}
\end{aligned}$$

and choosing $\mu = \frac{\lambda_{\min}}{1 + C_{pf}^2}$, equation (5.5.19) reduces to the following:

$$\begin{aligned}
\frac{2\mathfrak{F}_{\text{NgStb}}}{\gamma_{\min}^2} &\geq \frac{3\lambda_{\min}^2}{4(1 + C_{pf}^2)} + \sum_{\Omega_e \in \Omega_h} \frac{\tilde{C} \lambda_{\min}^2 h_{\Omega_e}^2 \|\mathbf{v} \bullet \operatorname{grad}[w_h]\|_{0, \Omega_e}^2}{8(1 + C_{pf}^2) (nd^2 \lambda_{\max}^2 + \tilde{C} C_{pf}^2 \delta_{\alpha \mathbf{v}} h^2 + \tilde{C} \delta_{\mathbf{D}} h^2)} \\
&+ \sum_{\Omega_e \in \Omega_h} \tau_{\Omega_e} \left\| \alpha w_h + \operatorname{div}[w_h \mathbf{v} - \mathbf{D} \operatorname{grad}[w_h]] \right\|_{0, \Omega_e}^2. \tag{5.5.24}
\end{aligned}$$

From equations (5.5.21) and (5.5.25a), we get the desired result given by equation (5.5.15). The second part of the proof is similar to Theorem 5.5.1. These exist a

constant $C_{\alpha\mathbf{vD}} > 0$ (for instance, $C_{\alpha\mathbf{vD}} = \max [nd^2, \tilde{C}, \tilde{C}C_{pf}^2]$) such that

$$nd^2\lambda_{\max}^2 + \tilde{C}C_{pf}^2\delta_{\alpha\mathbf{v}}h^2 + \tilde{C}\delta_{\mathbf{D}}h^2 \leq C_{\alpha\mathbf{vD}}\delta_{\alpha\mathbf{vD}}, \quad (5.5.25a)$$

$$\|\text{grad}[w_h]\|_0^2 \leq \frac{32(1 + C_{pf}^2)\mathfrak{F}_{\text{NgStb}}}{11\gamma_{\min}^2\lambda_{\min}^2}, \text{ and} \quad (5.5.25b)$$

$$\|\mathbf{v} \bullet \text{grad}[w_h]\|_0^2 \leq \frac{32C_{\alpha\mathbf{vD}}\delta_{\alpha\mathbf{vD}}(1 + C_{pf}^2)\tilde{C}^2\mathfrak{F}_{\text{NgStb}}}{\tilde{C}\gamma_{\min}^2\lambda_{\min}^2h^2}. \quad (5.5.25c)$$

Using Cauchy-Schwartz inequality on $\|\mathbf{v} \bullet \text{grad}[w_h]\|_0$ and (5.5.25b) gives

$$\|\mathbf{v} \bullet \text{grad}[w_h]\|_0^2 \leq \|\mathbf{v}\|_0^2\|\text{grad}[w_h]\|_0^2 \leq \frac{32C_{\mathbf{v}}(1 + C_{pf}^2)\mathfrak{F}_{\text{NgStb}}}{11\gamma_{\min}^2\lambda_{\min}^2}. \quad (5.5.26)$$

Now, consider the terms $\|w_h\|_1^2$ and $\|\mathbf{p}_h\|_{\text{div}}^2$:

$$\|w_h\|_1^2 = \|w_h\|_0^2 + \|\text{grad}[w_h]\|_0^2 \leq \frac{32(1 + C_{pf}^2)^2\mathfrak{F}_{\text{NgStb}}}{11\gamma_{\min}^2\lambda_{\min}^2}, \quad (5.5.27a)$$

$$\begin{aligned} \|\mathbf{p}_h\|_0^2 &\leq 2 \sum_{\Omega_e \in \Omega_h} \left\| \mathbf{p}_h - w_h \mathbf{v} + \mathbf{D}\text{grad}[w_h] - \delta_{\Omega_e} \mathbf{v} (\text{div}[w_h \mathbf{v} - \mathbf{D}\text{grad}[w_h]]) \right\|_{0, \Omega_e}^2 \\ &\quad + 2 \sum_{\Omega_e \in \Omega_h} \left\| -w_h \mathbf{v} + \mathbf{D}\text{grad}[w_h] - \delta_{\Omega_e} \mathbf{v} (\text{div}[w_h \mathbf{v} - \mathbf{D}\text{grad}[w_h]]) \right\|_{0, \Omega_e}^2, \end{aligned} \quad (5.5.27b)$$

$$\begin{aligned} \|\text{div}[\mathbf{p}_h]\|_0^2 &\leq 2 \sum_{\Omega_e \in \Omega_h} \left\| \alpha w_h + \text{div}[\mathbf{p}_h] + \delta_{\Omega_e} \text{div}[\alpha w_h \mathbf{v}] \right\|_{0, \Omega_e}^2 \\ &\quad + 2 \sum_{\Omega_e \in \Omega_h} \left\| \alpha w_h + \delta_{\Omega_e} \text{div}[\alpha w_h \mathbf{v}] \right\|_{0, \Omega_e}^2. \end{aligned} \quad (5.5.27c)$$

Using (5.5.25a)–(5.5.26) and repeated use of triangle inequality on (5.5.27b) and (5.5.27c) gives the boundedness estimate (5.5.16). \square

Theorem 5.5.3 (Error estimate for proposed LSFEM). *Given that equations (5.2.1a)–(5.2.1c) have a sufficiently smooth solution $(c, \mathbf{q}) \in (\mathcal{C} \times \mathcal{Q}) \cap (H^{r+1}(\Omega))^3$. Then the finite element solution (c_h, \mathbf{q}_h) of the unconstrained weighted negatively stabilized*

streamline diffusion LSFEM satisfies the following error estimate:

$$\begin{aligned} \sqrt{C_{\text{NgStb1}}} \|c - c_h\|_1 + \sqrt{C_{\text{NgStb2}}} \|\mathbf{q} - \mathbf{q}_h\|_{\text{div}} + \sqrt{C_{\text{NgStb3}}} \|\mathbf{v} \bullet \text{grad}[c - c_h]\|_0 \\ \leq C_{\text{NgStb}} h^r (\|c\|_{r+1} + \|\mathbf{q}\|_{r+1}), \end{aligned} \quad (5.5.28)$$

where $C_{\text{NgStb}} > 0$ is a constant independent of \mathbf{D} and h .

Proof. Let $c_I \in \mathcal{C}_h$ and $\mathbf{q}_I \in \mathcal{Q}_h$ be the standard finite element interpolants of c and \mathbf{q} , respectively. From the interpolation theory Bochev and Gunzburger (2009), we have

$$\|c - c_I\|_1 \leq Ch^r \|c\|_{r+1} \text{ and} \quad (5.5.29a)$$

$$\|\mathbf{q} - \mathbf{q}_I\|_{\text{div}} \leq Ch^r \|\mathbf{q}\|_{r+1} \quad (5.5.29b)$$

for some positive constant C independent of \mathbf{D} and h . The error $(c - c_h, \mathbf{q} - \mathbf{q}_h)$ satisfies the following orthogonality property:

$$\mathfrak{B}_{\text{NgStb}}((c_h - c, \mathbf{q}_h - \mathbf{q}); (w_h, \mathbf{p}_h)) = 0 \quad \forall (w_h, \mathbf{p}_h) \in \mathcal{W}_h \times \mathcal{Q}_h. \quad (5.5.30)$$

Cauchy-Schwartz inequality implies:

$$\begin{aligned} \mathfrak{B}_{\text{NgStb}}^{1/2}((c_h - c_I, \mathbf{q}_h - \mathbf{q}_I); (c_h - c_I, \mathbf{q}_h - \mathbf{q}_I)) \leq \\ \mathfrak{B}_{\text{NgStb}}^{1/2}((c - c_I, \mathbf{q} - \mathbf{q}_I); (c - c_I, \mathbf{q} - \mathbf{q}_I)). \end{aligned} \quad (5.5.31)$$

From Theorem 5.5.2 and interpolation estimates (5.5.29a)–(5.5.29b), one can obtain the desired error estimate (5.5.28). \square

From the above mathematical analysis, it is evident that the element-dependent

stabilization parameters $\tau_{\Omega_e} \leq 0$ and $\delta_{\Omega_e} \leq 0$ can be taken as

$$\delta_{\Omega_e} = -\frac{\delta_o \lambda_{\min} h_{\Omega_e}^2}{\left(\lambda_{\max}^2 + \delta_1 \max_{\mathbf{x} \in \Omega} [(\alpha + \operatorname{div}[\mathbf{v}])^2] h^2 + \delta_2 \max_{\mathbf{x} \in \Omega} [\|\operatorname{div}[\mathbf{D}]\|^2] h^2 \right)} \quad \text{and} \quad (5.5.32a)$$

$$\tau_{\Omega_e} = -\frac{\tau_o \lambda_{\min}^2 h_{\Omega_e}^2}{\left(\lambda_{\max}^2 + \tau_1 \max_{\mathbf{x} \in \Omega} [(\alpha + \operatorname{div}[\mathbf{v}])^2] h^2 + \tau_2 \max_{\mathbf{x} \in \Omega} [\|\operatorname{div}[\mathbf{D}]\|^2] h^2 \right)}, \quad (5.5.32b)$$

where δ_o , δ_1 , δ_2 , τ_o , τ_1 , and τ_2 are non-negative constants.

Remark 5.5.4. *The mathematical analysis provided by Hsieh and Yang Hsieh and Yang (2009) can be obtained as a special case of the mathematical analysis presented above. Specifically, take $\alpha = 0$, $\mathbf{D}(\mathbf{x})$ to be homogeneous and isotropic, and $\mathbf{v}(\mathbf{x})$ to be solenoidal and constant.*

5.6 NUMERICAL H -CONVERGENCE AND BENCHMARK PROBLEMS

We shall employ a popular problem from the literature, which is commonly used to assess the accuracy of numerical formulations for advective-diffusive systems (e.g., see Franca et al. (1998); Hsieh and Yang (2009)). The test problem is constructed using the method of manufactured solutions. The computational domain is a bi-unit square: $\Omega = (0, 1) \times (0, 1)$. The advection velocity vector field is taken as $\mathbf{v}(\mathbf{x}) = \hat{\mathbf{e}}_y$, where $\hat{\mathbf{e}}_y$ is the unit vector along the y -direction. The scalar diffusivity is denoted by $D(\mathbf{x})$. The concentration field is taken as follows:

$$c(x, y) = \frac{\sin(\pi x)}{e^{m_2 - m_1} - 1} \left(e^{m_2 - m_1} e^{m_1 y} - e^{m_2 y} \right), \quad (5.6.1)$$

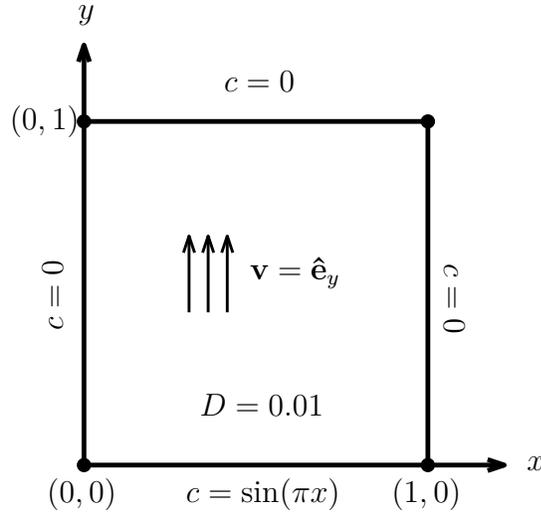


Figure 5.3: Numerical h -convergence study: A pictorial description of the two-dimensional boundary value problem.

where the constants m_1 and m_2 are given in terms of the scalar diffusivity:

$$m_1 = \frac{1 - \sqrt{1 + 4\pi^2 D^2}}{2D} \text{ and} \quad (5.6.2a)$$

$$m_2 = \frac{1 + \sqrt{1 + 4\pi^2 D^2}}{2D}. \quad (5.6.2b)$$

We have taken $D(\mathbf{x}) = 10^{-2}$ in our numerical simulations. This choice is arbitrary, and is primarily motivated to check whether the proposed framework gives stable, reliable, and accurate numerical results for advection-dominated problems. For the chosen value of the diffusivity, the solution (5.6.1) exhibits steep gradients near the boundary of the domain. A pictorial description of the boundary value problem is provided by Figure 5.3. Dirichlet boundary conditions are prescribed on the entire boundary.

Numerical solutions for the concentration and the flux vector are obtained by prescribing Dirichlet boundary conditions on all the four sides of the computational

domain. These conditions are enforced strongly and are given as:

$$c(\mathbf{x}) = \begin{cases} \sin(\pi x) & \text{for } y = 0 \\ 0 & \text{for } x = 0 \text{ or } x = 1 \text{ or } y = 1. \end{cases} \quad (5.6.3)$$

Using equation (5.6.1), one can calculate the corresponding flux vector and volumetric source needed for the convergence analysis.

5.6.1 Convergence analysis for $D(x, y) = 10^{-2}$

Herein, we will discuss the performance of negatively stabilized streamline diffusion LSFEM with and without LSB constraints. In case of unconstrained setting, we also quantify the errors incurred in satisfying LSB and GSB. Numerical simulations are performed using a series of hierarchical structured meshes based on three-node triangular (T3) and four-node quadrilateral (Q4) elements with XSeed and YSeed ranging from 11 to 81. Figure 5.4 provides the typical computational meshes used in the numerical h -convergence analysis. The meshes shown in this figure have 21 nodes along each side of the computational domain (i.e., XSeed = YSeed = 21). A series of hierarchical computational meshes are employed in the study with 11×11 , 21×21 , 41×41 and 81×81 nodes.

The weights for the primitive and negatively stabilized streamline diffusion LS-FEMs are taken to be of LS Type-1 (i.e., $\mathbf{A}(\mathbf{x}) = \mathbf{I}$ and $\beta(\mathbf{x}) = 1$). The element stabilization parameters for negatively stabilized streamline diffusion LSFEM are taken as $\delta_o = 0.01$ and $\tau_o = 0.01$. The convergence of the proposed computational framework with respect to L_2 -norm and H^1 -semi-norm is illustrated in Figure 5.5. Convergence studies are performed using T3- and Q4-based meshes under the negatively stabilized streamline diffusion LSFEM. It is evident that the Q4 element slightly outperforms the T3 element in terms of rates of convergence. From this figure, one can notice that

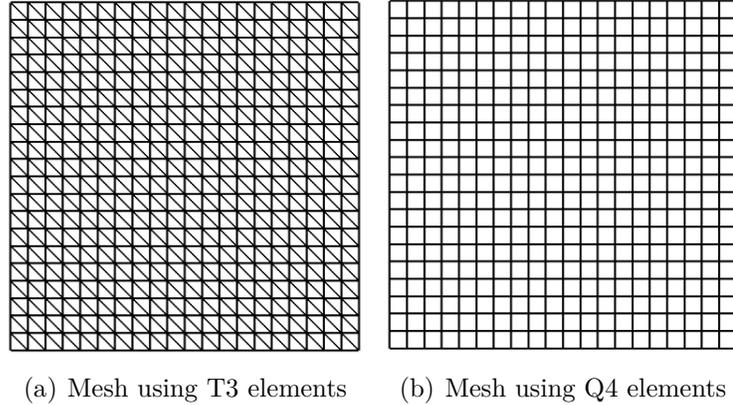


Figure 5.4: Numerical h -convergence study: This figure shows the typical computational meshes used in the numerical convergence analysis.

near optimal convergence rates are achieved for the concentration field in both L_2 -norm and H^1 -semi-norm for unconstrained negatively stabilized streamline diffusion formulation. For the flux vector, near optimal convergence rate is obtained in L_2 -norm but not in H^1 -semi-norm. This is because of the steep gradients in the concentration field at the boundary $y = 1$, which is due to the small value for the diffusivity. Enforcing LSB constraints considerably improves the H^1 -semi-norm convergence rate for the flux vector. However, for the flux variables, there is a slight decrease in L_2 -norm convergence rate as compared to the unconstrained negatively stabilized streamline diffusion LSFEM. Similar decrease in convergence rates of L_2 -norm and H^1 -semi-norm for the concentration has been observed. This can be attributed to the fact that LSB constraints improve the accuracy of the flux vector inside the boundary layers but has little effect away from it.

Remark 5.6.1. *It should be noted that the convergence rates reported in Figure 5.5 for the unconstrained negatively stabilized streamline diffusion LSFEM are in accordance with the mathematical analysis provided by Kopteva Kopteva (2004) and Stynes Stynes (2013). These results are obtained for singularly perturbed advection-diffusion*

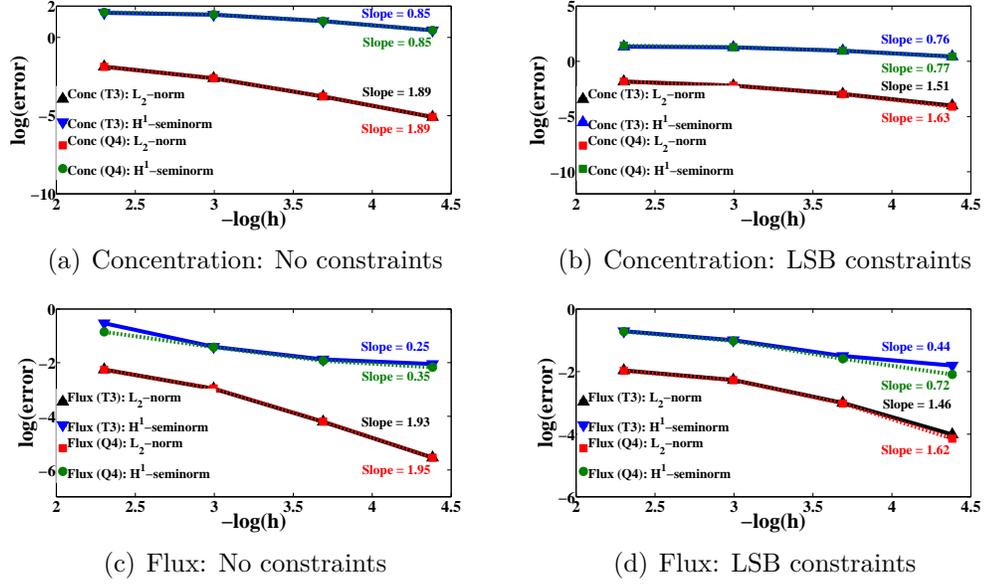
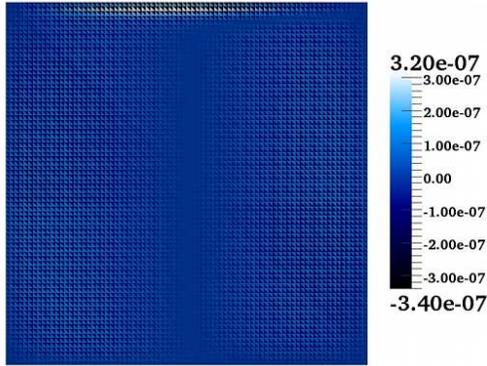


Figure 5.5: Numerical h -convergence study: This figure shows the convergence rates for the concentration and flux vector in L_2 -norm and H^1 -semi-norm with and without LSB constraints.

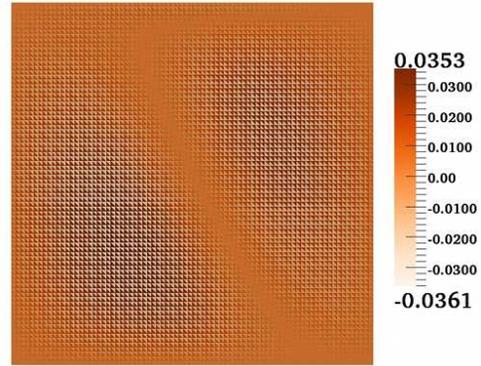
equation based on a class of unconstrained streamline diffusion finite element formulations. Kopteva Kopteva (2004) shows that one can get at best first-order convergence inside boundary and characteristic layers even on special meshes.

From Figure 5.5 one can also conclude that the Q4 element performs better than the T3 element. These trends in the convergence rates for different meshes are due to the fact that higher-order derivatives (e.g., $\text{div}[\text{grad}[c(\mathbf{x})]]$) in the stabilization terms for negatively stabilized streamline diffusion LSFEM vanish for T3 element. But these stabilization terms are non-zero for a Q4 element. The reason is that the shape functions for a T3 element are affine while that of a Q4 element are bilinear.

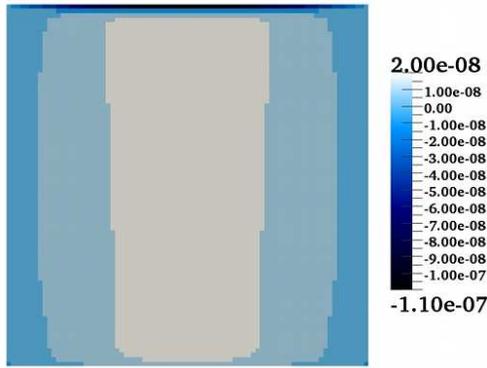
Another important aspect of this numerical h -convergence study is to quantify the errors incurred in satisfying LSB and GSB for unconstrained LSFEMs. The contours of the error distribution in LSB and the Lagrange multipliers enforcing the LSB constraints are shown in Figure 5.6. It is apparent that errors incurred in satisfying LSB are smaller under Q4 meshes than under T3 meshes. Note that



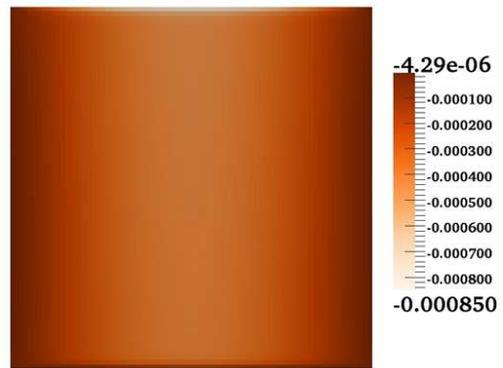
(a) T3 mesh: Error in LSB



(b) T3 mesh: Lagrange multiplier enforcing LSB



(c) Q4 mesh: Error in LSB



(d) Q4 mesh: Lagrange multiplier enforcing LSB

Figure 5.6: Numerical h -convergence study: The top and bottom left figures show the contours of error incurred in satisfying LSB for unconstrained LSFEM. The right set of figures show the contours of Lagrange multiplier enforcing LSB constraint using the proposed LSFEM.

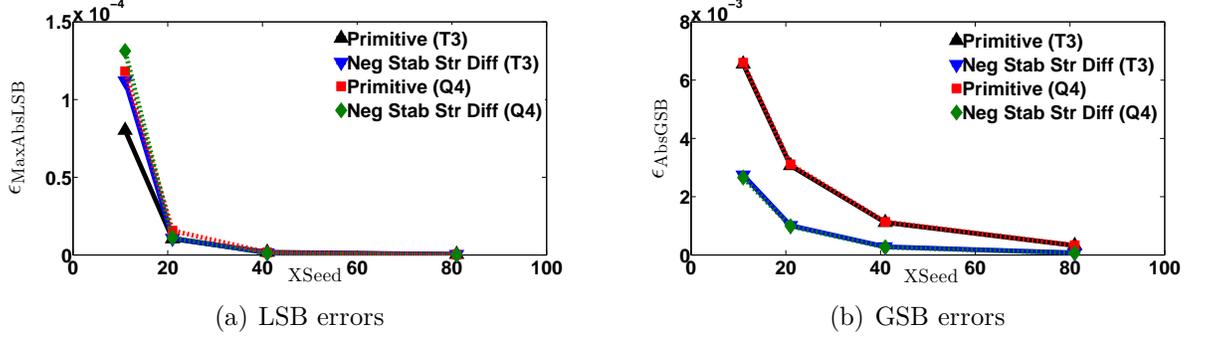


Figure 5.7: Numerical h -convergence study: These figures show the decrease of $\epsilon_{\text{MaxAbsLSB}}$ and ϵ_{AbsGSB} with respect to XSeed for a series of hierarchical three-node triangular and four-node quadrilateral meshes.

the Lagrange multipliers enforcing the LSB constraint can have negative value as opposed to KKT multipliers. Numerical simulations are performed based on three-node triangular mesh and four-node quadrilateral mesh with 81 nodes on each side of the domain. The decrease in $\epsilon_{\text{MaxAbsLSB}}$ and ϵ_{AbsGSB} on h -refinement is shown in Figure 5.7. From this figure, one can notice that the errors in LSB and GSB for a Q4 mesh are lesser than that of a T3 mesh. On h -refinement, the decrease in $\epsilon_{\text{MaxAbsLSB}}$ and ϵ_{AbsGSB} is slow and not close to machine precision (See equations (5.4.15)–(5.4.17) for the definitions of $\epsilon_{\text{MaxAbsLSB}}$ and ϵ_{AbsGSB}). Numerical simulations are performed using the unconstrained primitive and negatively stabilized streamline diffusion LSFEMs. For $X\text{Seed} = 81$, $\epsilon_{\text{MaxAbsLSB}}$ and ϵ_{AbsGSB} are in $\mathcal{O}(10^{-6})$. In addition, the decrease in LSB and GSB errors with respect to h -refinement is slow, and the values are not close to the machine precision.

Finally, the computational cost of the unconstrained and constrained LSFEMs for both T3 and Q4 meshes are shown in Figures 5.8 and 5.9. It is clear that the computational cost associated with a Q4 mesh is higher than that of a T3 mesh. This can be again be attributed to the non-vanishing stabilization terms (e.g., $\text{div}[\text{grad}[c(\mathbf{x})]]$) in the negatively stabilized streamline diffusion LSFEM for Q4 meshes. For Q4 mesh, as $\text{div}[\text{grad}[c]] \neq 0$, the computational cost is higher than that of the T3 mesh.

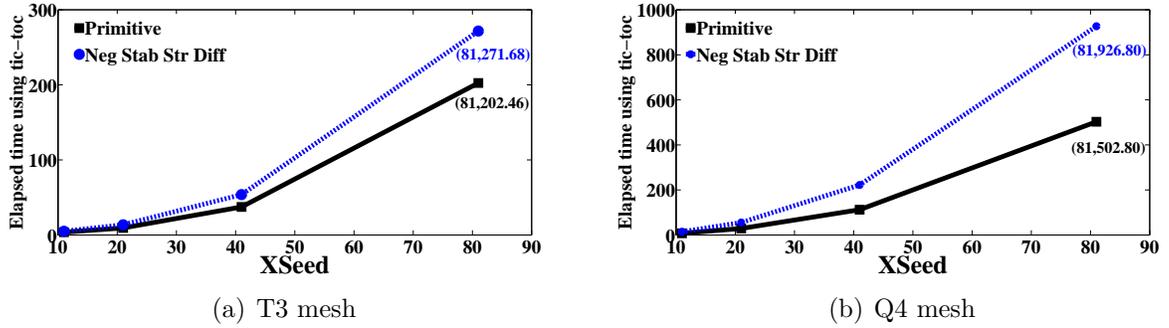


Figure 5.8: Numerical h -convergence study: This figure shows the CPU time (in seconds) of the proposed computational framework for unconstrained primitive and unconstrained negatively stabilized streamline diffusion LSFEMs.

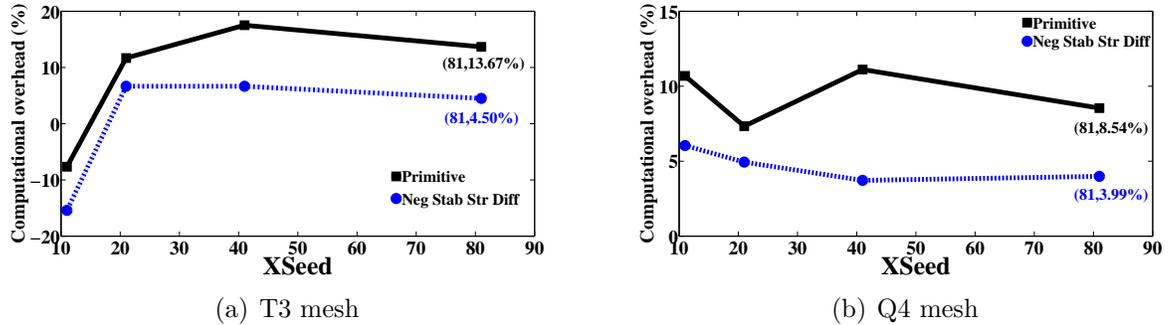


Figure 5.9: Numerical h -convergence study: This figure shows the computational overhead incurred in satisfying LSB as compared to that of the corresponding unconstrained formulations.

For constrained LSFEMs, the maximum additional computational cost (for both LSFEMs) did not exceed 15%, which has been tested on a hierarchy of meshes. In addition, for $XSeed = 11$, we obtained negative value for the computational overhead. This is because the interior point convex algorithm used in MATLAB optimization solver MAT (2015) pre-processes the constrained convex quadratic programming problem simplifies the given LSB constraints by removing redundancies. Hence, for very low number of unknowns, the computational cost associated with interior point convex algorithm is much faster than the LU solver for the unconstrained optimization problem.

5.6.2 Thermal boundary layer problem

This benchmark problem has wide practical applications in the areas of heat and mass transfer. Herein, we shall use this benchmark problem to study the performance of unconstrained and constrained LSFEM formulations in capturing steep gradients near the boundary for advection-dominated scenarios. Consider a rectangular domain $\Omega = \{(x, y) \in [0, 1] \times [0, 0.5]\}$ with velocity field $\mathbf{v}(x, y) = 2y\hat{\mathbf{e}}_x$, where $\hat{\mathbf{e}}_x$ is the unit vector along the x -direction. The volumetric source is assumed to be homogeneous (i.e., $f(x, y) = 0$), and the scalar diffusivity is taken to be $D(x, y) = 10^{-4}$. The boundary conditions are:

$$c(x, y) = \begin{cases} 0 & \text{for } 0 < x \leq 1 \text{ and } y = 0 \\ 2y & \text{for } x = 1 \text{ and } 0 \leq y \leq 0.5 \\ 1 & \text{for } 0 \leq x \leq 1 \text{ and } y = 1 \\ 1 & \text{for } x = 0 \text{ and } 0 \leq y \leq 0.5 \end{cases}. \quad (5.6.4)$$

A pictorial description of the boundary value problem is provided in Figure 5.10. Dirichlet boundary conditions are prescribed on all four sides of the computational domain. We have taken $c(\mathbf{x}) = 1$ at $\mathbf{x} = (0, 0)$. The weights are taken to be that of LS Type-1 (see equations (5.4.6a) and (5.4.6b)). The element-level stabilization parameters for negatively stabilized streamline diffusion LSFEM are taken to be $\delta_o = 0.01$ and $\tau_o = 0.001$. Numerical simulations are performed using four-node quadrilateral mesh with XSeed = 41 and YSeed = 21. The element Péclet number will then be $\mathbb{P}e_h = 125$. The obtained concentration contours are shown in Figure 5.11. It is evident from these figures that numerical solution obtained from the primitive LSFEM contains node-to-node spurious oscillations. These oscillations did not reduce even after enforcing the LSB and NN constraints. But the negatively stabilized streamline diffusion LSFEM is able to capture the steep gradients near the boundary without

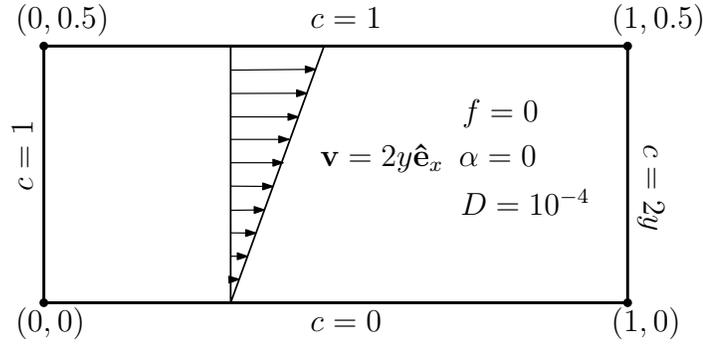


Figure 5.10: Thermal boundary layer problem: This figure shows a pictorial description of the boundary value problem.

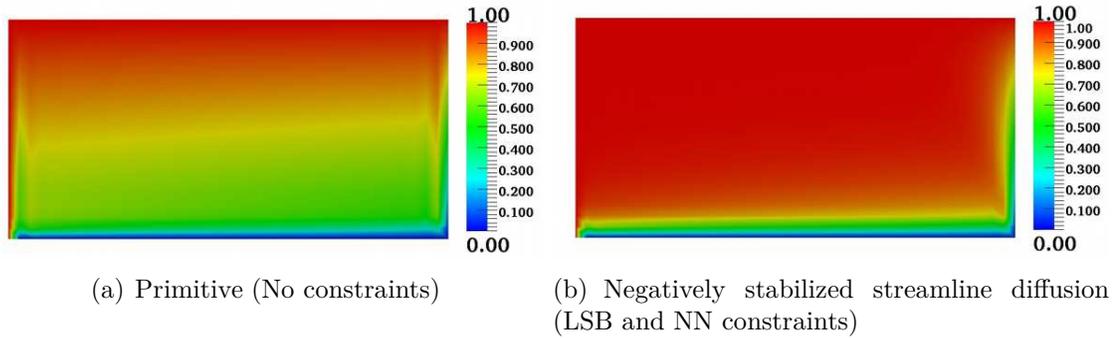
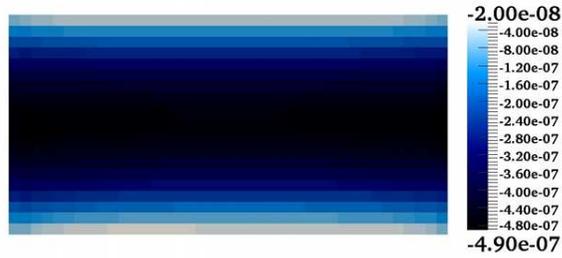


Figure 5.11: Thermal boundary layer problem: This figure shows the contours of concentration obtained for both unconstrained and constrained LSFEMs based on Q4 finite element mesh.

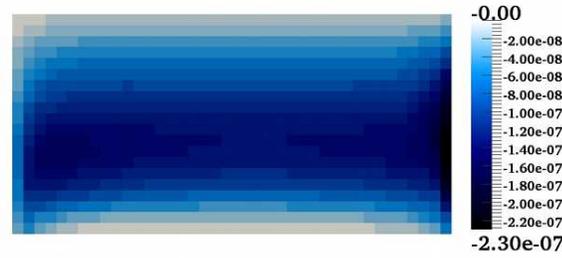
producing spurious oscillations. The errors incurred in satisfying LSB for *unconstrained* LSFEM formulations are shown in Figure 5.12. One can notice that the error is more dominant in the interior of the domain under the primitive LSFEM, whereas the error is dominant at the boundary $x = 1$ under the negatively stabilized streamline diffusion formulation.

5.7 TRANSPORT-CONTROLLED BIMOLECULAR REACTIONS

In this section, we shall apply the proposed mixed LSFEM-based computational framework to study transport-controlled bimolecular chemical reactions. Specifically, we are interested in the spatial distribution, plume formation, and chaotic mixing of



(a) Primitive



(b) Negatively stabilized streamline diffusion

Figure 5.12: Thermal boundary problem: This figure shows the contours of the error incurred in satisfying LSB for various unconstrained LSFEM formulations using Q4 meshes.

chemical species at high Péclet numbers. To this end, consider the following irreversible bimolecular chemical reaction:



where A , B , and C are the species involved in the chemical reaction; n_A , n_B , and n_C are their respective (positive) stoichiometric coefficients. The fate of these chemical

species are governed by the following coupled advective-diffusive-reactive system:

$$\frac{\partial c_A}{\partial t} + \operatorname{div}[\mathbf{v}c_A - \mathbf{D}(\mathbf{x}, t) \operatorname{grad}[c_A]] = f_A(\mathbf{x}, t) - n_A r(\mathbf{x}, t, c_A, c_B, c_C) \quad \text{in } \Omega \times]0, \mathcal{I}[,$$
(5.7.2a)

$$\frac{\partial c_B}{\partial t} + \operatorname{div}[\mathbf{v}c_B - \mathbf{D}(\mathbf{x}, t) \operatorname{grad}[c_B]] = f_B(\mathbf{x}, t) - n_B r(\mathbf{x}, t, c_A, c_B, c_C) \quad \text{in } \Omega \times]0, \mathcal{I}[,$$
(5.7.2b)

$$\frac{\partial c_C}{\partial t} + \operatorname{div}[\mathbf{v}c_C - \mathbf{D}(\mathbf{x}, t) \operatorname{grad}[c_C]] = f_C(\mathbf{x}, t) + n_C r(\mathbf{x}, t, c_A, c_B, c_C) \quad \text{in } \Omega \times]0, \mathcal{I}[,$$
(5.7.2c)

$$c_i(\mathbf{x}, t) = c_i^{\text{p}}(\mathbf{x}, t) \quad \text{on } \Gamma_i^c \times]0, \mathcal{I}[,$$
(5.7.2d)

$$\left(\left(\frac{1 - \operatorname{Sign}[\mathbf{v} \bullet \hat{\mathbf{n}}]}{2} \right) \mathbf{v}(\mathbf{x}, t) c_i(\mathbf{x}, t) - \mathbf{D}(\mathbf{x}, t) \operatorname{grad}[c_i(\mathbf{x}, t)] \right) \bullet \hat{\mathbf{n}}(\mathbf{x}) = h_i^{\text{p}}(\mathbf{x}, t) \quad \text{on } \Gamma_i^q \times]0, \mathcal{I}[, \text{ and}$$
(5.7.2e)

$$c_i(\mathbf{x}, t = 0) = c_i^0(\mathbf{x}) \quad \text{in } \Omega,$$
(5.7.2f)

where $i = A, B,$ and C . $\mathbf{v}(\mathbf{x}, t)$ is the advection velocity vector field, $f_i(\mathbf{x}, t)$ constitutes the non-reactive volumetric source, $c_i^{\text{p}}(\mathbf{x}, t)$ is the Dirichlet boundary condition, and $h_i^{\text{p}}(\mathbf{x}, t)$ is the Neumann boundary condition of the i -th chemical species. $r(\mathbf{x}, t, c_A, c_B, c_C)$ is the bimolecular chemical reaction rate, which is a non-linear function of the concentrations of the chemical species involved in the reaction. $c_i^0(\mathbf{x})$ is the initial condition of i -th chemical species. $t \in [0, \mathcal{I}]$ denote the time, where \mathcal{I} is the total time of interest. The coupled governing equations (5.7.2a)–(5.7.2e) can be converted to a set of uncoupled advection-diffusion equations using the following linear algebraic transformation:

$$c_F := c_A + \left(\frac{n_A}{n_C} \right) c_C \quad \text{and} \quad (5.7.3a)$$

$$c_G := c_B + \left(\frac{n_B}{n_C} \right) c_C. \quad (5.7.3b)$$

As we are interested in *fast* bimolecular chemical reactions, it is acceptable to assume that the chemical species A and B cannot co-exist at any given location \mathbf{x} and time t . Hence, c_A , c_B , and c_C can be evaluated as follows:

$$c_A(\mathbf{x}, t) = \max \left[c_F(\mathbf{x}, t) - \left(\frac{n_A}{n_B} \right) c_G(\mathbf{x}, t), 0 \right], \quad (5.7.4a)$$

$$c_B(\mathbf{x}, t) = \max \left[c_G(\mathbf{x}, t) - \left(\frac{n_B}{n_A} \right) c_F(\mathbf{x}, t), 0 \right], \text{ and} \quad (5.7.4b)$$

$$c_C(\mathbf{x}, t) = \left(\frac{n_C}{n_A} \right) (c_F(\mathbf{x}, t) - c_A(\mathbf{x}, t)). \quad (5.7.4c)$$

In Reference Nakshatrala et al. (2013), a similar mathematical model has been studied in the context of maximum principles and the non-negative constraint. However, the study has neglected the advection, and did not address local and global species balance. These aspects are very important and cannot be neglected in the numerical simulations of chemically reacting systems. In particular, advection can play a predominant role in the study of bioremediation Borden and Bedient (1986), transverse mixing-controlled chemical reactions in hydro-geological media Willingham et al. (2008), and contaminant degradation problems Dentz et al. (2011). This chapter precisely addresses such problems in which advection is dominant, and satisfying species balance at both local and global levels is extremely important.

Herein, we perform numerical simulations for highly spatially varying advection velocity fields and time-periodic flows. See Reference Neufeld and H.-García (2010) for a discussion on time-periodic flows. For such problems in 2D, the following quantity is of considerable importance, which is referred as the position weighted second moment of the product C concentration:

$$\Theta_C^2(t) = \frac{\int_{\Omega} (y - y_0)^2 c_C(\mathbf{x}, t) \, d\Omega}{\int_{\Omega} c_C(\mathbf{x}, t) \, d\Omega}, \quad (5.7.5)$$

where y_0 is the location of a convenient reference horizontal line. In our numerical simulations, we have taken y_0 to be the y -coordinate of the start of the formation of product C . Since $c_C(\mathbf{x}, t) \geq 0$, $\Theta_C^2(t)$ is a non-negative quantity. In subsequent sections, we study the utility of this quantity as *a posteriori* criterion to assess numerical accuracy. We also analyze the variation of Θ_C^2 with respect to $\mathbb{P}e_h$. We also present the numerical results that shed light on the impact of advection on the formation of the product C . In all our numerical simulations, we have taken the weights in primitive and negatively stabilized streamline diffusion LSFEMs to be that of LS Type-1 (i.e., $\mathbf{A}(\mathbf{x}) = \mathbf{I}$ and $\beta(\mathbf{x}) = 1$).

Remark 5.7.1. *In the literature, to study mixing processes due to advection, spectral methods Adrover et al. (2002), pseudospectral methods Tsang (2009), and model reduction methods Neufeld and H.-García (2010) are commonly employed. However, such methods are limited to time-periodic flows, periodic initial and boundary conditions, simple geometries, and homogeneous isotropic diffusivity. Extending these methods to complicated geometries, general initial and boundary conditions, complicated advection velocity fields, and heterogeneous isotropic and anisotropic diffusivity is not trivial and may not even be possible. Moreover, these methods do not guarantee the satisfaction of non-negativity, DMPs, LSB, and GSB. The proposed computational framework is aimed at filling this lacuna.*

5.7.1 One-dimensional steady-state analysis of product formation in fast reactions

Analysis is performed for two different advection velocities: $v = 0.25$ and $v = 1.0$. Diffusivity is assumed to be 2.5×10^{-3} . The stoichiometric coefficients are assumed to be: $n_A = 2$, $n_B = 1$, and $n_C = 1$. Numerical simulations are performed for two different cases as described below.

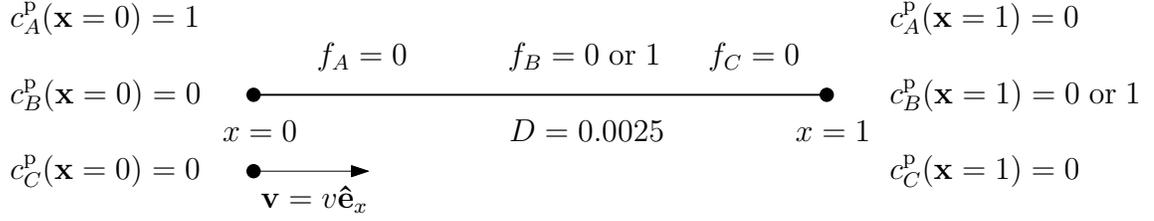


Figure 5.13: 1D irreversible bimolecular fast reaction problem: A pictorial description of the boundary value problem.

5.7.1.1 Case #1

A pictorial description of the boundary value problem is shown in Figure 5.13. For Case #1: $f_B(\mathbf{x}) = 1$ and $c_B^p(\mathbf{x} = 1) = 0$, and for Case #2: $f_B(\mathbf{x}) = 0$ and $c_B^p(\mathbf{x} = 1) = 1$. The objective of this case study is to analyze whether the proposed negatively stabilized streamline diffusion LSFEM can produce physically meaningful values for $c_i(x)$ on coarse meshes. Based on the linear algebraic transformation given by equations (5.7.3a)–(5.7.3b), the analytical solution for invariants F and G can be written as:

$$c_F(x) = \left(1 - \frac{1 - \exp(vx/D)}{1 - \exp(v/D)} \right) \text{ and} \quad (5.7.6a)$$

$$c_G(x) = \frac{f_G}{v} \left(x - \frac{1 - \exp(vx/D)}{1 - \exp(v/D)} \right). \quad (5.7.6b)$$

Using equations (5.7.4a)–(5.7.4c), one can obtain the analytical solution for product C .

For the numerical solution, we have taken XSeed = 11. The element stabilization parameters for negatively stabilized streamline diffusion LSFEM are taken as $\delta_o = 0.08$ and $\tau_o = 0.04$ when $v = 0.25$. For $v = 1.0$, δ_o and τ_o , are assumed to equal to 0.083 and 0.0121, respectively. The analytical and numerical solutions are compared in Figure 5.14. As per this figure, the primitive LSFEM produces node-to-node oscillations near the boundaries of the domain. Furthermore, its numerical solution considerably deviates from the analytical solution in the entire domain. For $\mathbb{P}e_h = 5$

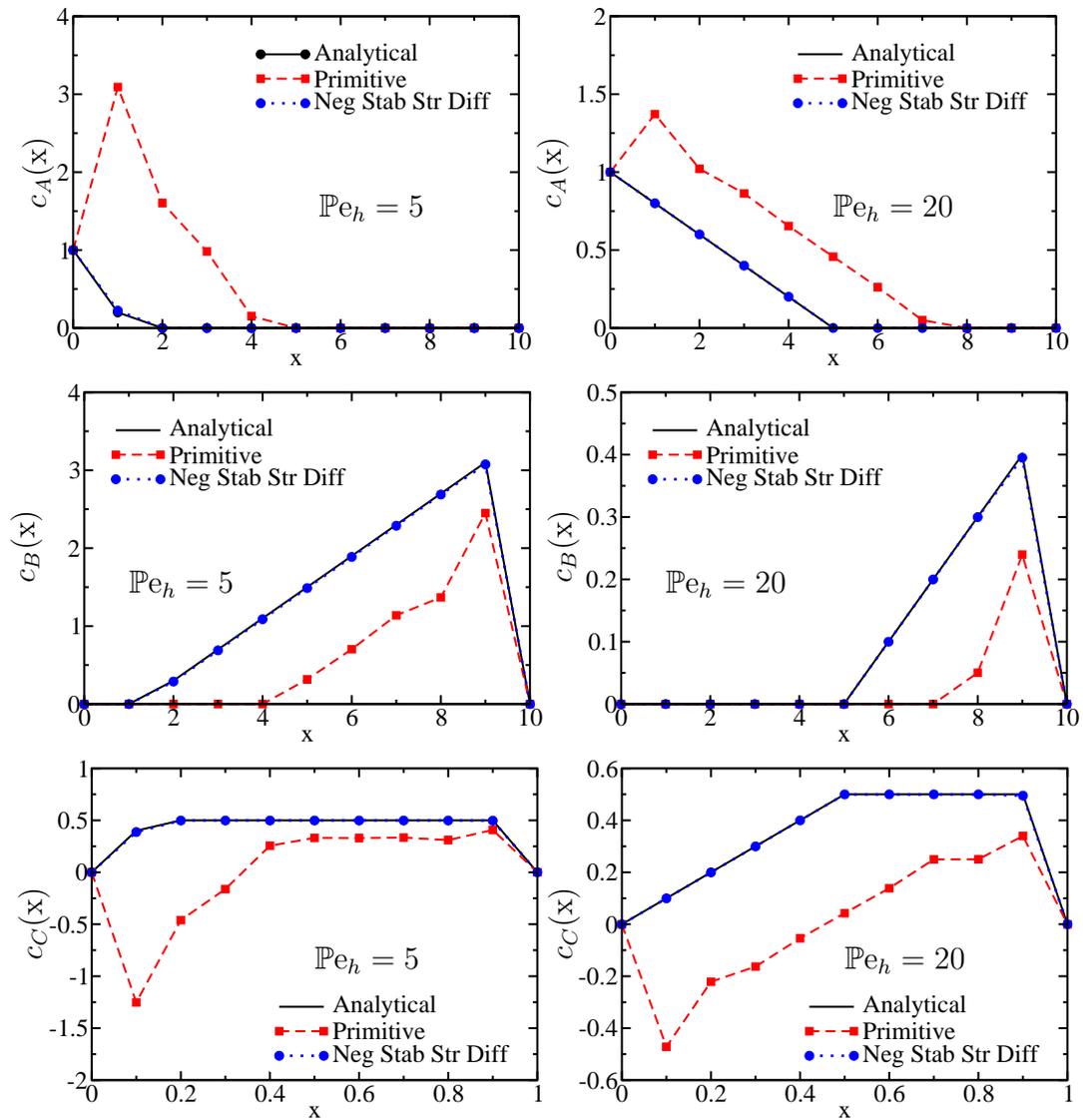


Figure 5.14: 1D irreversible bimolecular fast reaction problem (Case #1): This figure compares the concentration profile of the reactants and the product with the analytical solution.

and $\mathbb{P}e_h = 20$, the negative value for the concentration is as low as -1.25 and -0.47 . On the other hand, the negatively stabilized streamline diffusion LSFEM is able to capture the analytical solution profile in the entire domain without producing negative values in the concentration field.

5.7.1.2 Case #2

A pictorial description of the boundary value problem is provided in Figure 5.13. The objective of this case study is to examine whether the proposed LSFEM can capture steep gradients in the solution near the boundary. The analytical solution for the invariants F and G can be written as:

$$c_F(x) = \left(1 - \frac{1 - \exp(vx/D)}{1 - \exp(v/D)} \right) \text{ and} \quad (5.7.7a)$$

$$c_G(x) = \left(\frac{1 - \exp(vx/D)}{1 - \exp(v/D)} \right). \quad (5.7.7b)$$

Figure 5.15 compares the obtained the numerical solution with the analytical solution. The negatively stabilized streamline diffusion LSFEM is able to accurately capture the steep gradients near the boundary.

5.7.2 Steady-state plume formation from boundary in a reaction tank

A pictorial description of the boundary value problem is provided in Figure 5.16. The computational domain is a rectangle with $L_x = 2$ and $L_y = 1$. Dirichlet boundary conditions with $c_A^p = c_B^p = 1$ are specified on the left side of the domain. Elsewhere, $c_i^p(\mathbf{x})$ is taken to be zero for all the chemical species involved in the bimolecular reaction. The non-reactive volumetric source is assumed to be zero in the entire domain for all the chemical species. The stoichiometric coefficients are taken as $n_A = 1$, $n_B = 1$ and $n_C = 1$. The advection velocity field is defined through the

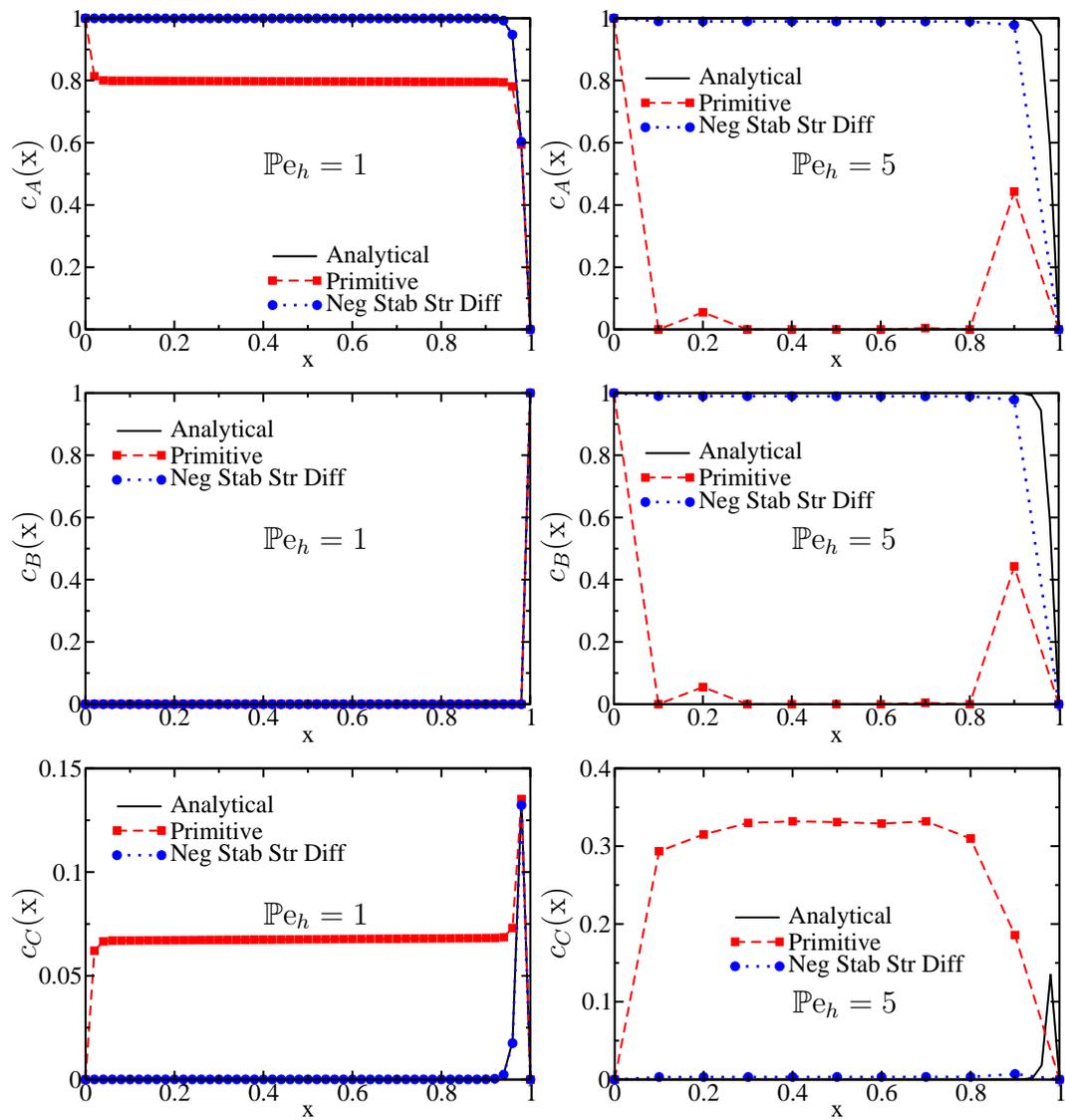


Figure 5.15: 1D irreversible bimolecular fast reaction problem (Case #2): This figure compares the concentration profile of the chemical species A , B , and C to that of the analytical solution.

following multi-mode stream function Nakshatrala et al. (2013):

$$\psi(\mathbf{x}) = -y - \sum_{k=1}^3 A_k \cos\left(\frac{p_k \pi x}{L_x} - \frac{\pi}{2}\right) \sin\left(\frac{q_k \pi y}{L_y}\right), \quad (5.7.8)$$

where $\mathbf{x} = (x, y)$, $(p_1, p_2, p_3) = (4, 5, 10)$, $(q_1, q_2, q_3) = (1, 5, 10)$, and $(A_1, A_2, A_3) = (0.08, 0.02, 0.01)$. The corresponding components of the advection velocity can be written as follows:

$$v_x(\mathbf{x}) = -\frac{\partial \psi}{\partial y} = 1 + \sum_{k=1}^3 A_k \frac{q_k \pi}{L_y} \cos\left(\frac{p_k \pi x}{L_x} - \frac{\pi}{2}\right) \cos\left(\frac{q_k \pi y}{L_y}\right) \quad \text{and} \quad (5.7.9a)$$

$$v_y(\mathbf{x}) = +\frac{\partial \psi}{\partial x} = \sum_{k=1}^3 A_k \frac{p_k \pi}{L_x} \sin\left(\frac{p_k \pi x}{L_x} - \frac{\pi}{2}\right) \sin\left(\frac{q_k \pi y}{L_y}\right). \quad (5.7.9b)$$

It is easy to check that $\text{div}[\mathbf{v}(\mathbf{x})] = 0$. The contours of the stream function and the corresponding advection velocity vector field are shown in Figure 5.16. Numerical simulations are performed using the following two different types of diffusivities:

- *Type #1*: $D(\mathbf{x}) = 10^{-2}$
- *Type #2*: $\mathbf{D}(\mathbf{x}) = \mathbf{R}\mathbf{D}_0\mathbf{R}^T$, where \mathbf{R} and \mathbf{D}_0 are given as:

$$\mathbf{R} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \quad \text{and} \quad (5.7.10a)$$

$$\mathbf{D}_0(\mathbf{x}) = \omega_0 \begin{pmatrix} y_*^2 + \omega_2 x_*^2 & -(1 - \omega_2)x_* y_* \\ -(1 - \omega_2)x_* y_* & \omega_2 y_*^2 + x_*^2 \end{pmatrix}, \quad (5.7.10b)$$

where $x_* = x + \omega_1$ and $y_* = y + \omega_1$. The parameters θ , ω_0 , ω_1 , and ω_2 are equal to $\frac{\pi}{6}$, 1.0, 10^{-3} , and 10^{-3} . Correspondingly, the eigenvalues of $\mathbf{D}(\mathbf{x})$ are $\omega_0(x_*^2 + y_*^2)$ and $\omega_0\omega_2(x_*^2 + y_*^2)$. The contrast/anisotropic ratio of the media (which is the ratio of maximum to minimum eigenvalue) is as high as 10^3 .

Herein, we employed a structured mesh based on Q4 elements. Numerical simulations are performed with varying mesh sizes and polynomial orders ($p = 1, 2, 3$) to demonstrate the pros and cons of various unconstrained and constrained LSFEMs. The stabilization parameters are taken as $\delta_o = \tau_o = 10^{-3}$ and $\delta_2 = \tau_2 = 10^{-4}$. The contours of the concentration of the product C are shown in Figures 5.17–5.20 for both the primitive and negatively stabilized streamline diffusion LSFEMs. The white patches in the figures denote the regions in which the non-negative constraint has been violated. The variation of Θ_C^2 with respect to XSeed and Pe_L are shown in Figures 5.21–5.22. From these figures, the following inferences can be drawn:

- (i) It is clear that both low-order and higher-order polynomials violate the non-negative constraint and DMPs under unconstrained formulations. Moreover, mesh refinement and polynomial refinement do not seem to reduce the amount of violated region for DMP constraints. The negative values are in the range $\mathcal{O}(10^{-2})$ to $\mathcal{O}(10^{-4})$, which are not close to the machine precision $\epsilon_{\text{mach}} = \mathcal{O}(10^{-16})$.
- (ii) The proposed framework based on $p = 1$ is able to satisfy all the desired properties, and is able to predict physically meaningful values for the concentration and the flux.
- (iii) The primitive LSFEM and the unconstrained negatively stabilized streamline diffusion LSFEM give unphysical values for the position weighted second moment of the product C (i.e., Θ_C^2). On the other hand, the proposed computational framework is able to accurately describe the variation of Θ_C^2 with respect to mesh refinement. In addition, the numerical values for Θ_C^2 reaches a plateau on h -refinement, which indicates convergence. However, this is not observed with the unconstrained primitive and negatively stabilized streamline diffusion LSFEMs. Herein, analysis is performed using XSeed = YSeed = 201. Through

numerical simulations, we observed that $\log(\Theta_C^2) \propto \sqrt{\text{Pe}_L}$.

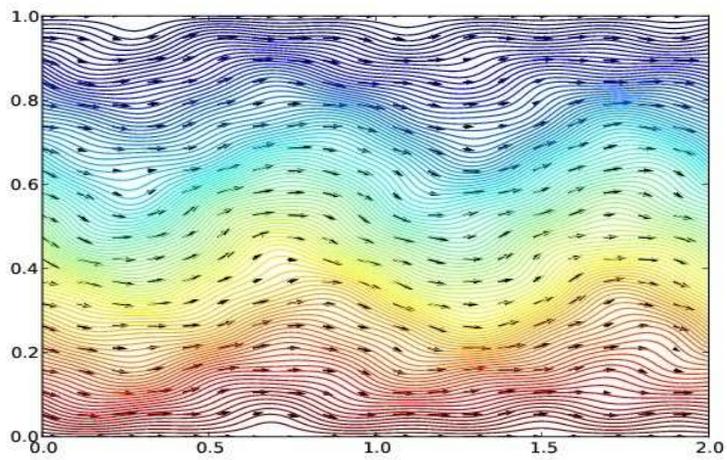
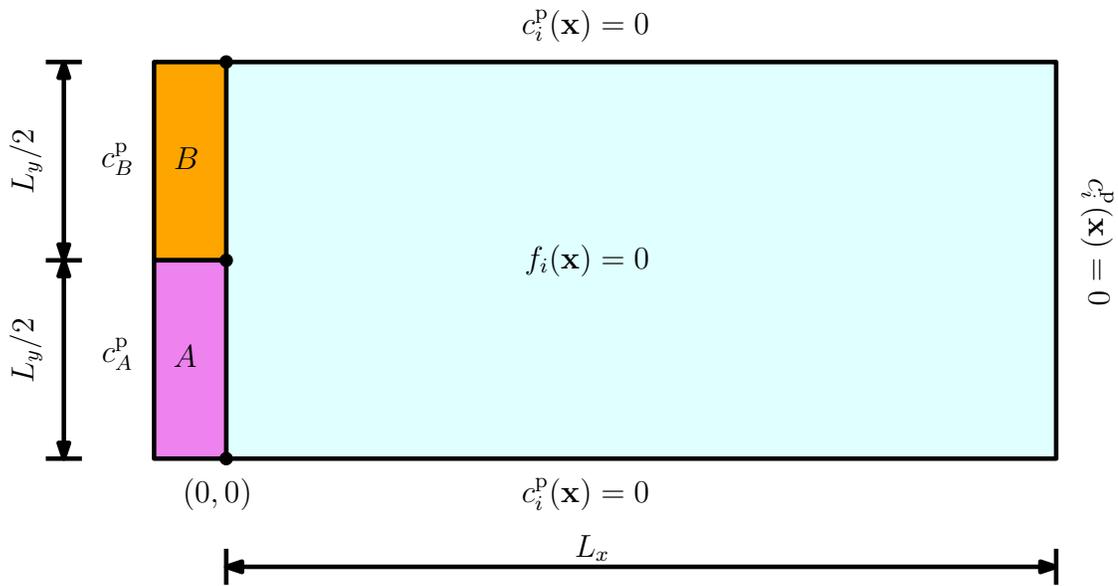
Finally, it should be emphasized that placing explicit non-negative constraints on the nodal concentrations does not ensure non-negativity of the concentration in the entire computational domain. This is due to the fact that higher-order shape functions change their sign within an element Payette et al. (2012).

5.7.3 Transient analysis of vortex stirred mixing in a reaction tank

Figure 5.23 provides a description of the problem with appropriate initial and boundary conditions. The top-left figure provides a pictorial description of the initial boundary value problem. The top-right figure shows the contours of the stream function corresponding to the advection velocity field. The bottom figures show the initial conditions for the reactants A and B such that $\langle c_A(\mathbf{x}, t = 0) \rangle = \langle c_B(\mathbf{x}, t = 0) \rangle = 1$. The computational domain is a square with $L_x = L_y = 1$. For all chemical species, zero flux boundary condition is prescribed on the entire boundary. The non-reactive volumetric source is zero in the entire domain for all the chemical species A , B , and C . The stoichiometric coefficients are taken as $n_A = 1$, $n_B = 1$, and $n_C = 1$. For advection velocity, we employ the following vortex-based flow field:

$$\mathbf{v}(\mathbf{x}) = \cos(2\pi y)\hat{\mathbf{e}}_x + \cos(2\pi x)\hat{\mathbf{e}}_y. \quad (5.7.11)$$

The total time of interest is taken as $\mathcal{I} = 5$. We assume scalar diffusivity to be $D = 10^{-2}$. The stabilization parameters are taken as $\delta_o = \tau_o = 10^{-3}$ and $\delta_1 = \tau_1 = 10^{-4}$. Figures 5.24–5.26 provide the concentration profiles of unconstrained and constrained negatively stabilized streamline diffusion LSFEM with NN constraints. We have taken $X\text{Seed} = Y\text{Seed} = 121$. If constraints are not enforced, one gets unphysical negative values for the concentration of product C . This will be particularly true in the early times of a numerical simulation. Note that the violations of the non-negative



(b) Stream function and advection velocity vector field

Figure 5.16: Plume development from boundary in a reaction tank: The top figure provides a pictorial description of the boundary value problem. The bottom figure shows the contours of the stream function corresponding to the advection velocity vector field.

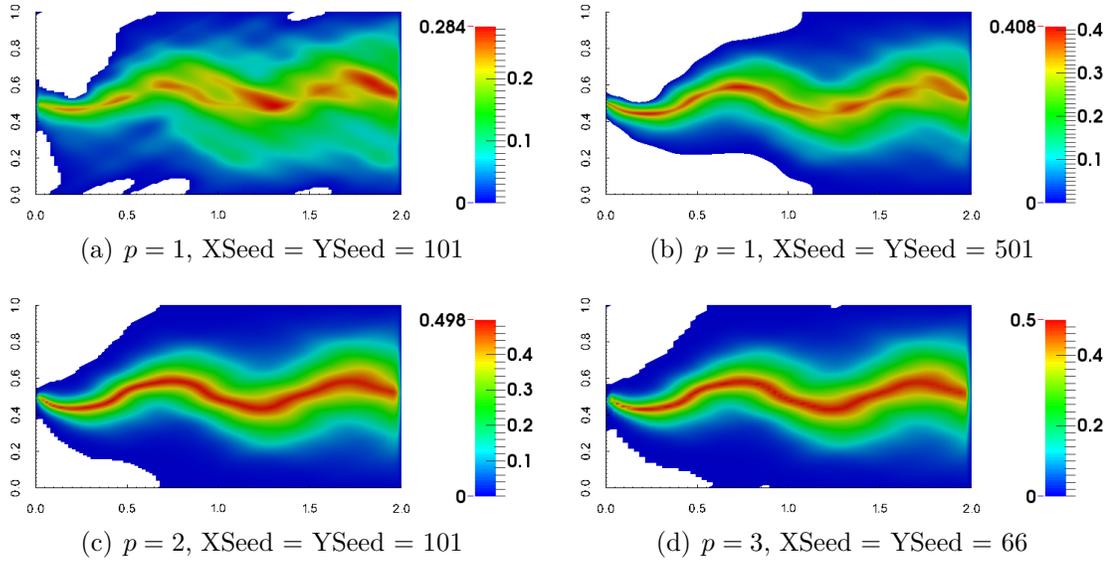


Figure 5.17: Plume development from boundary in a reaction tank (Type #1): This figure shows the concentration profiles of the product C based on unconstrained primitive LSFEM.

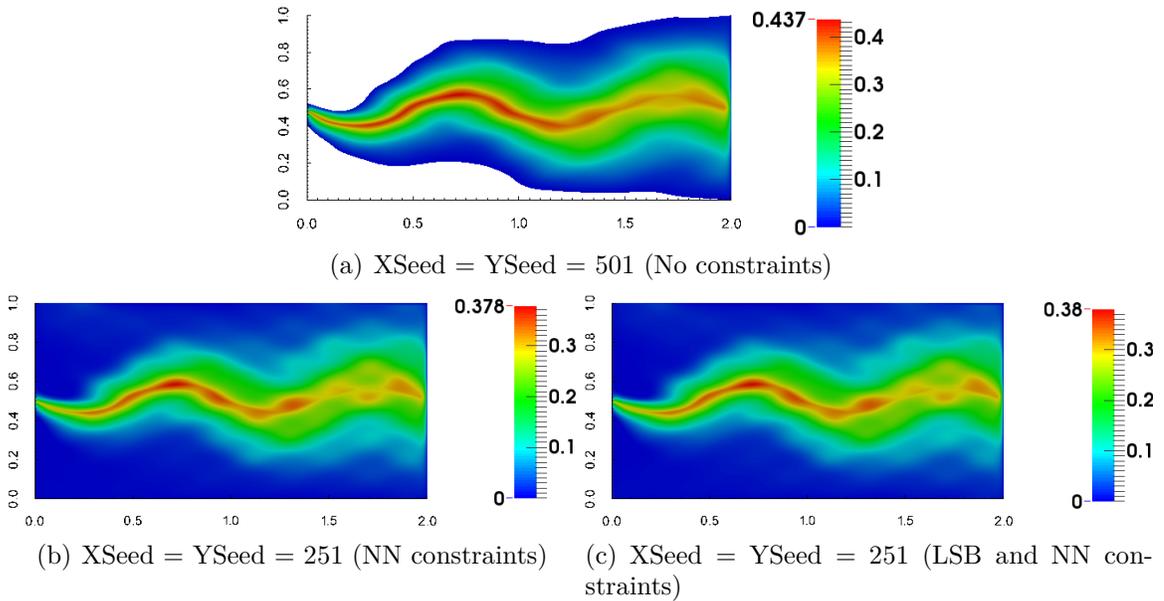


Figure 5.18: Plume development from boundary in a reaction tank (Type #1): This figure shows the concentration profiles of the product C based on unconstrained and constrained negatively stabilized streamline diffusion LSFEM.

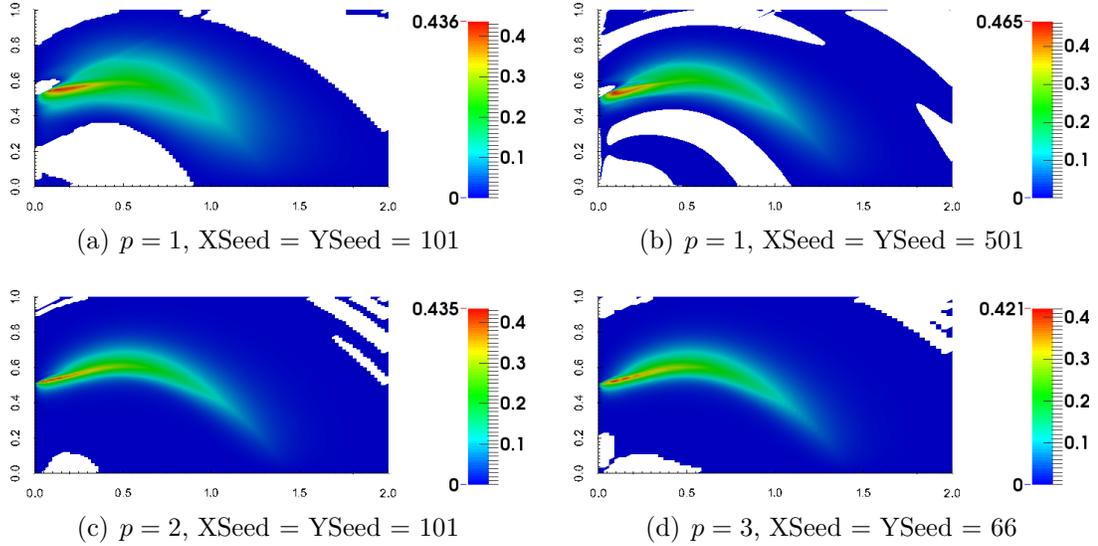
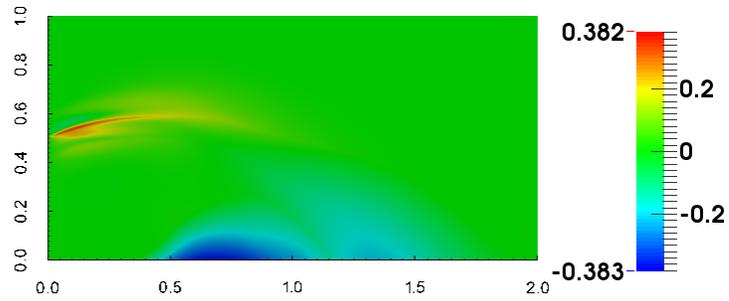


Figure 5.19: Plume development from boundary in a reaction tank (Type #2): This figure shows the concentration profiles of the product C based on unconstrained primitive LSFEM.

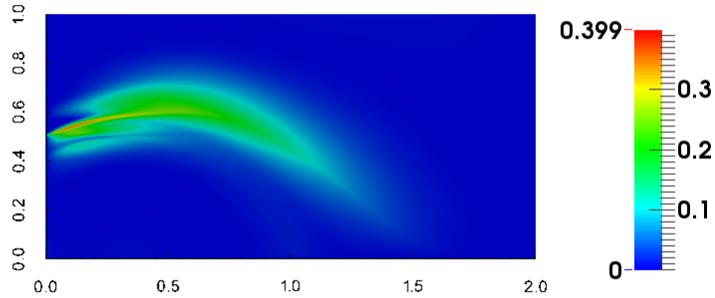
constraint are significant, and are present for various choices of the time-step. In Figure 5.26, the time-step is taken as $\Delta t = 0.1$. Herein, XSeed = YSeed = 121. As t increases, the product C should accumulate near the center of the two vortices. The proposed computational framework is able to accurately capture such features, and the obtained solutions are physical at all times. From these figures it is evident that existing numerical formulations do not provide accurate information on the fate of reactants and products for all times. On the other hand, the proposed methodology predicts results accurately for both early and late times.

5.7.4 Transient analysis of species mixing in cellular flows

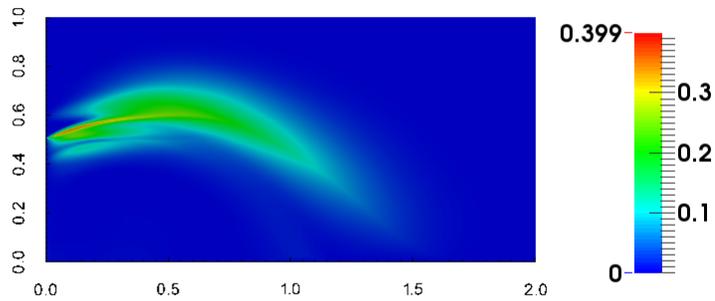
A pictorial description of the initial boundary value problem is provided in Figure 5.27. We have taken $L_x = 1$, $L_y = 0.5$, XSeed = 61 and YSeed = 241. The stoichiometric coefficients are taken as $n_A = 1$, $n_B = 1$ and $n_C = 1$. The time-step is taken as $\Delta t = 0.1$. The total time of interest is taken as $\mathcal{I} = 5$. The scalar diffusivity is taken as $D = 5 \times 10^{-3}$. The stabilization parameters are taken as $\delta_o = \tau_o = 10^{-3}$



(a) $X_{Seed} = Y_{Seed} = 251$ (No constraints)

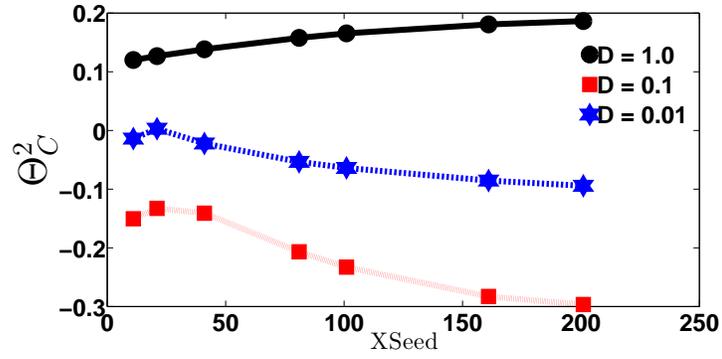


(b) $X_{Seed} = Y_{Seed} = 251$ (NN constraints)

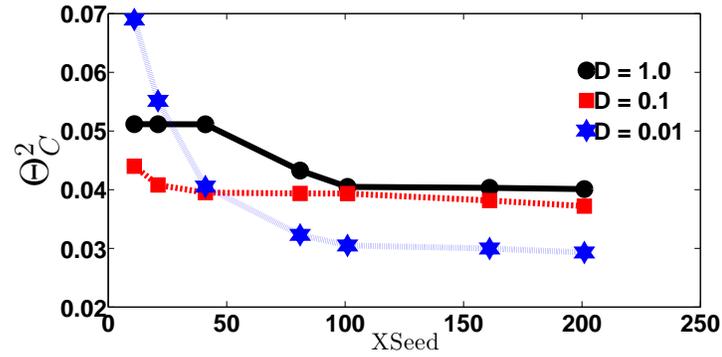


(c) $X_{Seed} = Y_{Seed} = 251$ (LSB and NN constraints)

Figure 5.20: Plume development from boundary in a reaction tank (Type #2): This figure shows the concentration profiles of the product C based on unconstrained and constrained negatively stabilized streamline diffusion LSFEM.



(a) No constraints



(b) With LSB and DMP constraints

Figure 5.21: Plume development from boundary in a reaction tank (Type #1): This figure shows the variation Θ_C^2 with mesh refinement under the weighted negatively stabilized streamline diffusion LSFEM.

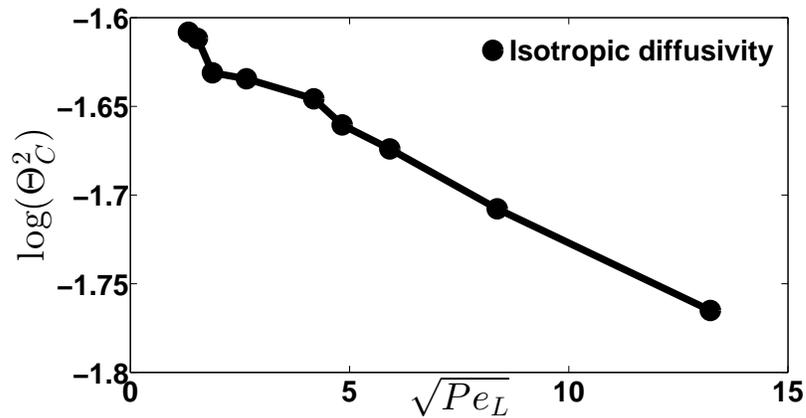
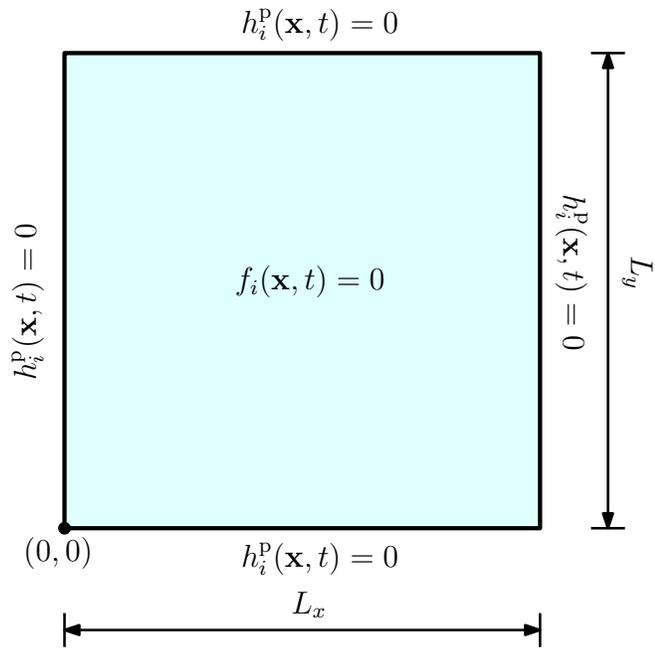
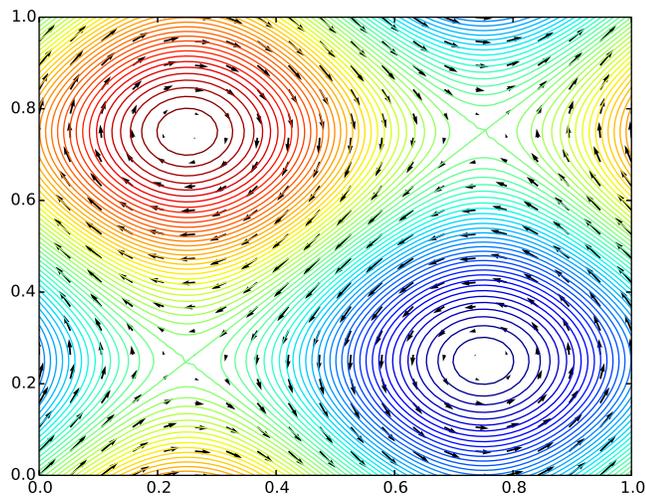


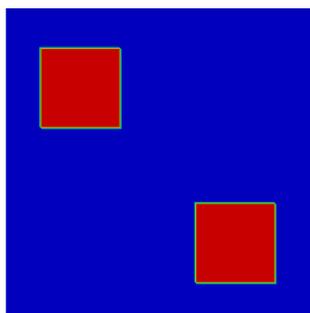
Figure 5.22: Plume development from boundary in a reaction tank (Type #1): This figure shows the variation $\log(\Theta_C^2)$ with respect to $\sqrt{Pe_L}$ for isotropic diffusivity under the weighted negatively stabilized streamline diffusion LSFEM with LSB and DMP constraints.



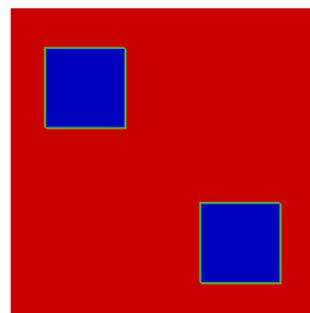
(a) Problem description



(b) Stream function and associated advection velocity field



(c) Reactant A: Initial condition



(d) Reactant B: Initial condition

Figure 5.23: Vortex-stirred mixing in a reaction tank: A pictorial description of the initial boundary value problem

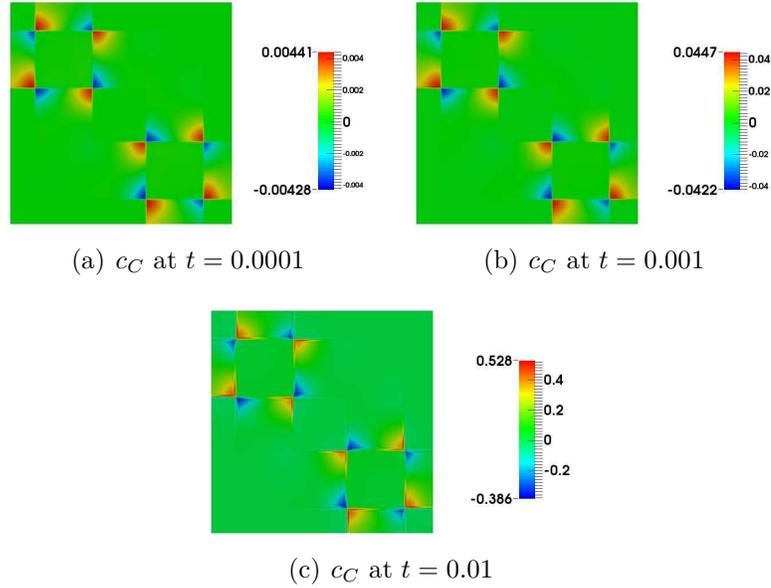
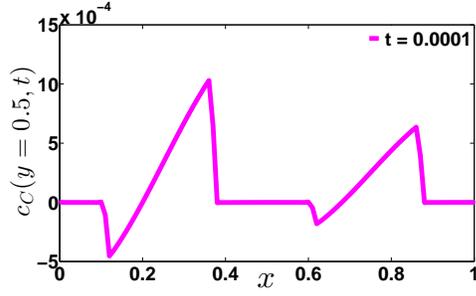


Figure 5.24: Vortex-stirred mixing in a reaction tank: This figure shows the concentration profiles of the product C after the first time-step using the unconstrained weighted negatively stabilized streamline diffusion LSFEM.

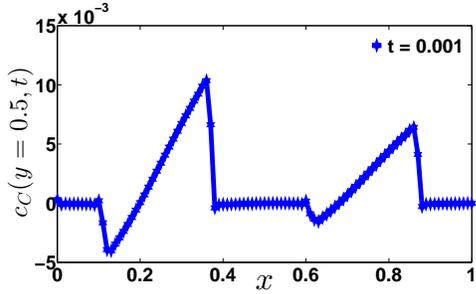
and $\delta_1 = \tau_1 = 10^{-4}$. The advection velocity vector field for the cellular flow is given by

$$\mathbf{v}(\mathbf{x}) = -\sin\left(\frac{2\pi x}{L_{\text{Cell}}}\right)\cos\left(\frac{2\pi y}{L_{\text{Cell}}}\right)\hat{\mathbf{e}}_x + \cos\left(\frac{2\pi x}{L_{\text{Cell}}}\right)\sin\left(\frac{2\pi y}{L_{\text{Cell}}}\right)\hat{\mathbf{e}}_y, \quad (5.7.12)$$

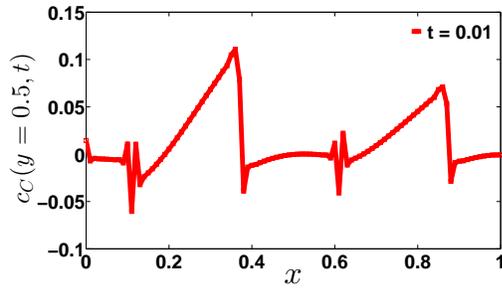
where $L_{\text{Cell}} = 0.5$. It should be noted that the concentration of the product C should be between 0 and 1. Figure 5.28 shows the concentration profiles of the product C under the unconstrained and constrained negatively stabilized streamline diffusion. It is evident that the *unconstrained* LSFEM violates both the non-negative and maximum constraints. On the other hand, the proposed computational framework with LSB and DMP constraints provides physically meaningful profiles for the concentration of the product.



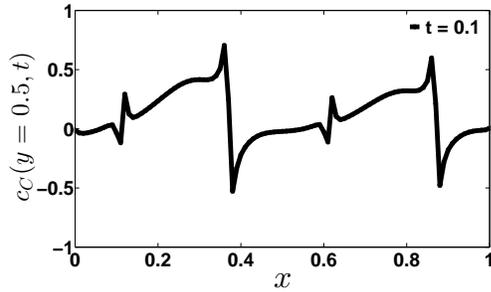
(a) $c_C(y = 0.5, t = 0.0001)$ with $\Delta t = 0.0001$



(b) $c_C(y = 0.5, t = 0.001)$ with $\Delta t = 0.001$



(c) $c_C(y = 0.5, t = 0.01)$ with $\Delta t = 0.01$



(d) $c_C(y = 0.5, t = 0.1)$ with $\Delta t = 0.1$

Figure 5.25: Vortex-stirred mixing in a reaction tank: This figure shows the concentration profiles of the product C at $y = 0.5$ after the first time-step using the unconstrained weighted negatively stabilized streamline diffusion LSFEM.

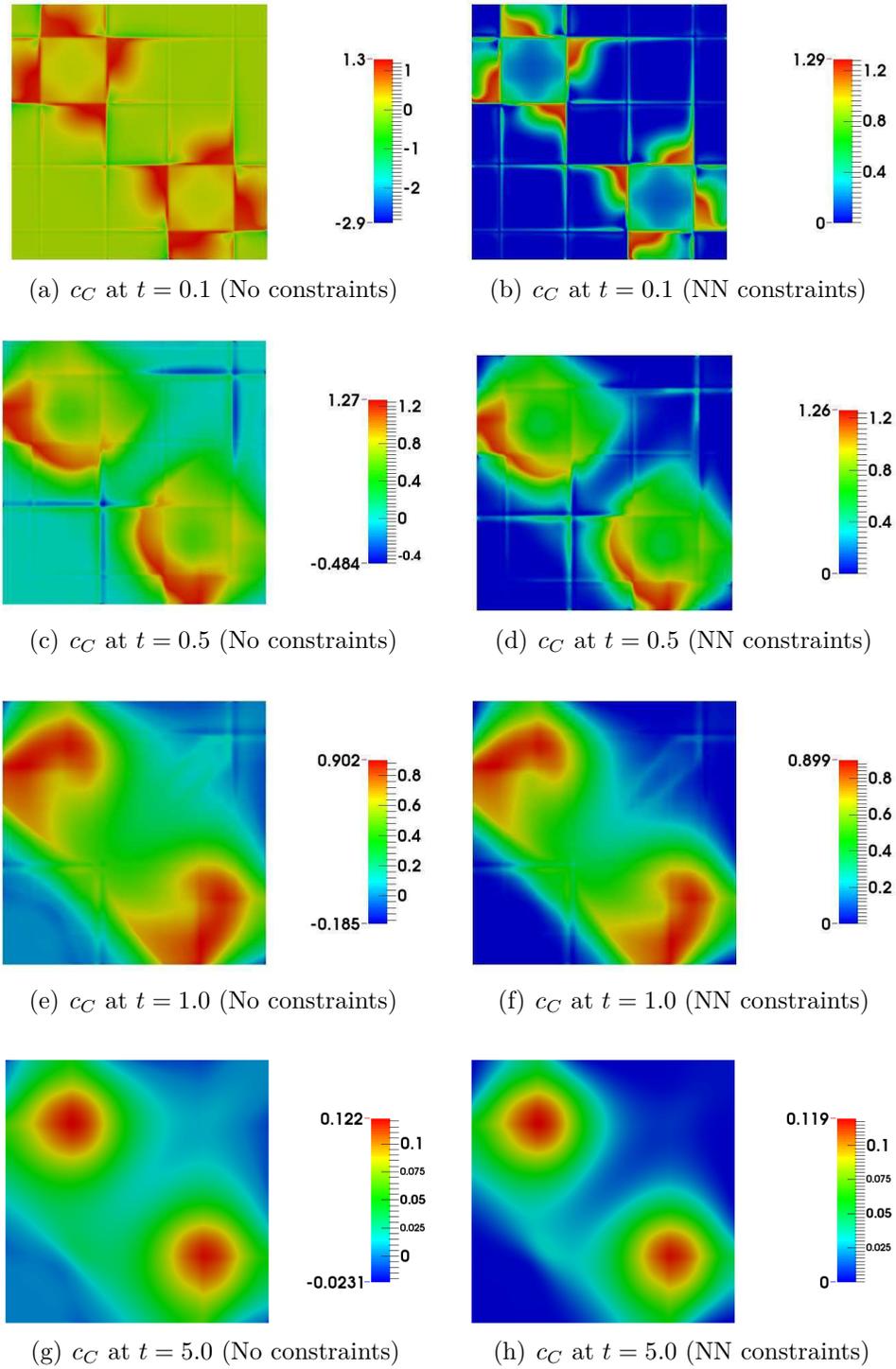
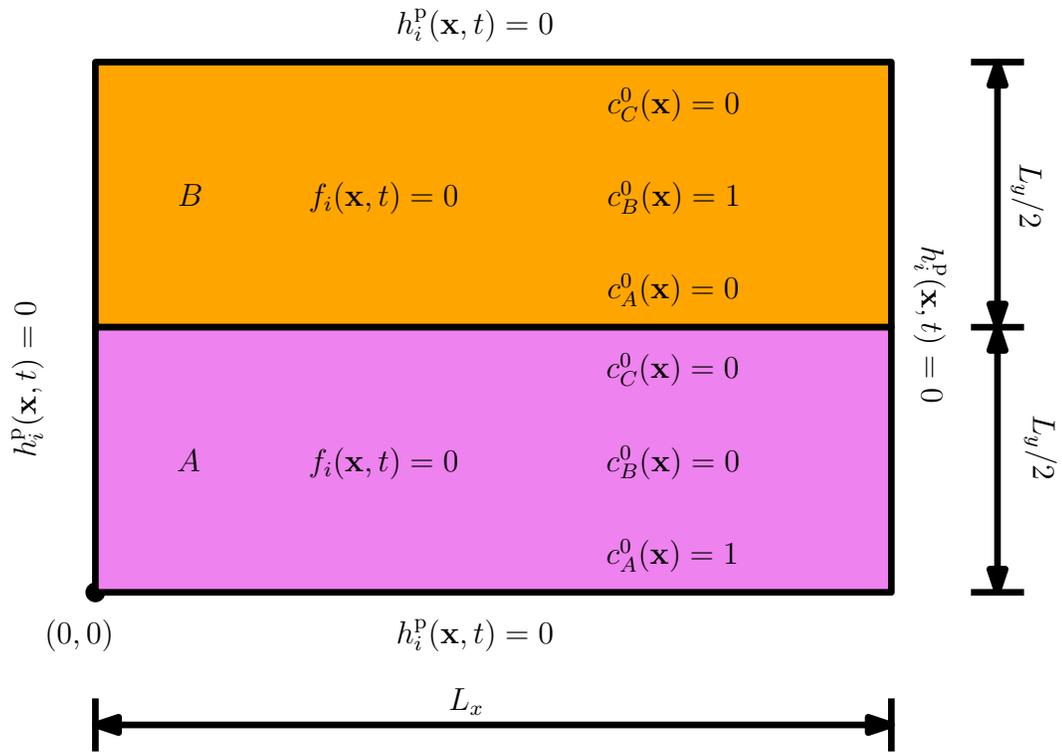
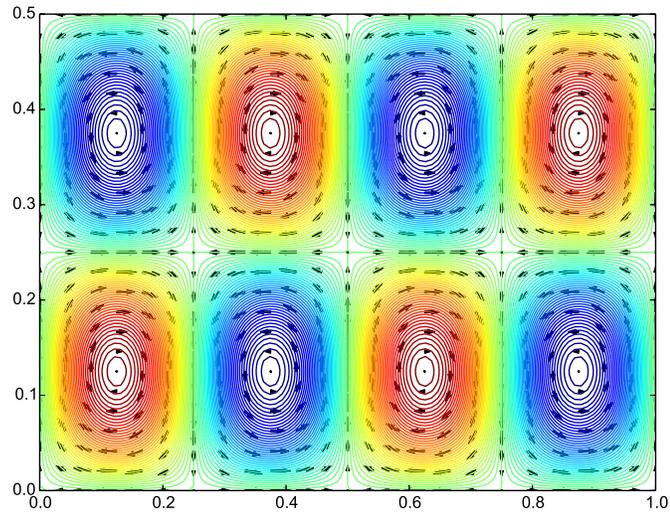


Figure 5.26: Vortex-stirred mixing in a reaction tank: This figure shows the concentration profiles of the product C at various time levels using the weighted negatively stabilized streamline diffusion LSFEM with and without constraints.



(a) Problem description



(b) Stream function and advection velocity

Figure 5.27: Transport-controlled mixing in cellular flows: A pictorial description of the initial boundary value problem and associated advection velocity field for the cellular flow.

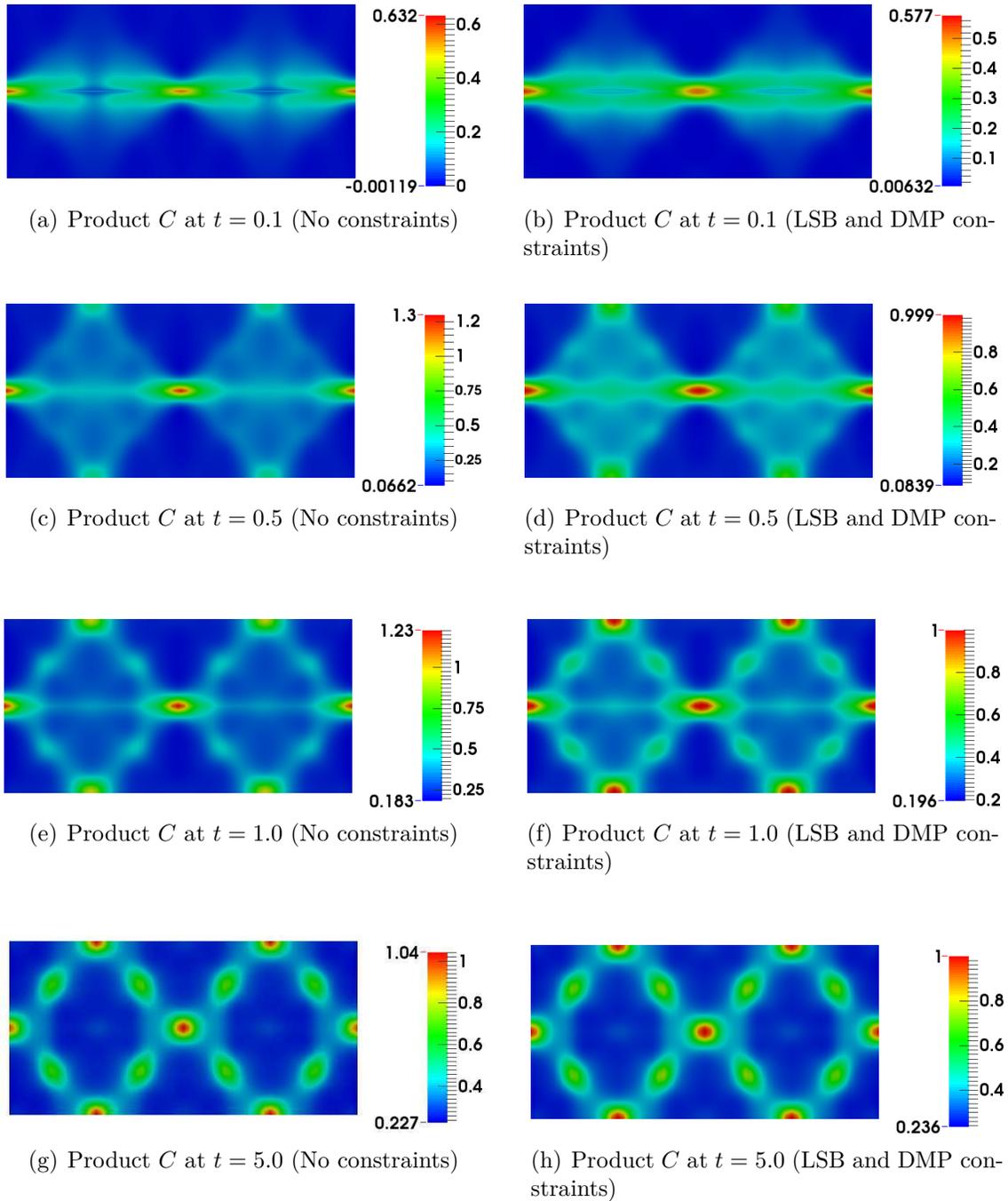


Figure 5.28: Transport-controlled mixing in cellular flows: This figure shows the concentration profiles of the product C at various time levels using the unconstrained and constrained weighted negatively stabilized streamline diffusion LSFEM.

5.8 SUMMARY AND CONCLUDING REMARKS

We presented a robust computational framework for (steady-state and transient) advection-diffusion-reaction equations that satisfies the non-negative constraint, maximum principles, local species balance, and global species balance. The framework can handle general computational grids, anisotropic diffusivity, highly heterogeneous velocity fields, and provides physically meaningful numerical solutions without node-to-node spurious oscillations even on coarse computational meshes. The main *contributions* of the chapter can be summarized as follows:

- (C1) We constructed and proved a continuous maximum principle that includes both Dirichlet and Neumann boundary conditions. It also takes into account the inflow and outflow Neumann boundary conditions in establishing the maximum principle.
- (C2) We described in detail the shortcomings of several plausible numerical approaches to satisfy the maximum principle, the non-negative constraint, and species balance.
- (C3) We proposed a locally conservative DMP-preserving computational framework and constructed element stabilization parameters that are valid for a general reaction coefficient, advection velocity, and diffusivity. The framework has been carefully constructed using the least-squares finite element method (LSFEM). It is also shown that a naive implementation of LSFEM will not meet the desired properties.
- (C4) The discrete problem under the proposed framework is well-posed, and it can be shown that a unique solution exists.
- (C5) We performed numerical convergence studies on the computational framework.

We also systematically analyzed and documented the performance of the proposed framework with various benchmark problems and realistic examples.

- (C6) We obtained numerically a scaling law for a transport-controlled bimolecular reaction.
- (C7) In chemically reactive systems, it is important to predict the fate of reactants and products during the early times. We have shown that the existing formulations may not provide accurate information for such scenarios. On the other hand, using numerical experiments, we have shown that the proposed framework predicts accurate results for both early and late times.

The *salient features and performance* of the proposed computational framework can be summarized as follows:

- (S1) The rate of decrease of errors in LSB and GSB for *unconstrained* negatively stabilized streamline diffusion LSFEM with h -refinement is slow and is about $\mathcal{O}(h)$. Furthermore, this numerical formulation violates various discrete principles and the non-negative constraint for both isotropic and anisotropic diffusivities. On the other hand, the proposed non-negative computational framework is able to satisfy LSB and GSB up to machine precision on an arbitrary computational mesh.
- (S2) The proposed computational framework with NN and LSB constraints eliminates the spurious node-to-node oscillations and provides physically meaningful values for concentration. Furthermore, it is able to furnish reasonable answers with various time-steps and at various time levels even on coarse computational grids.
- (S3) It has been shown that existing formulations may *fail* to give acceptable results for non-negative statistical quantities such as Θ_C^2 , which is defined in Section

5.7.2. However, the proposed methodology always provides non-negative values for Θ_C^2 . The quantity Θ_C^2 can be used as *a posteriori* criteria to assess accuracy of numerical solutions for complex initial and boundary value problems for which non-negativity and species conservation are important.

(S4) Due to the aforementioned desired properties, our proposed computational framework can be an ideal candidate to numerically obtain scaling laws for complicated problems with non-trivial initial and boundary conditions. Therefore, the proposed framework will be vital for predictive simulations in groundwater modeling, reactive transport, environmental fluid mechanics, and modeling of degradation of materials.

A possible future research work is to implement and analyze the performance of the proposed numerical methodology in a parallel environment. A related research is to design tailored iterative solvers and associated pre-conditioners for our proposed numerical methodology.

Chapter 6

NUMERICAL FORMULATIONS FOR STEADY-STATE AND TRANSIENT SEMI-LINEAR REACTION-DIFFUSION EQUATIONS, AND THEIR STRUCTURE PRESERVING PROPERTIES

“To those who do not know
mathematics it is difficult to get
across a real feeling as to the
beauty, the deepest beauty, of
nature ... If you want to learn
about nature, to appreciate nature,
it is necessary to understand the
language that she speaks in.”

Richard Feynman

In this chapter, we shall first briefly discuss about the pros and cons of imposing mesh and time-step restrictions to satisfy non-negative constraint and discrete maximum principles for linear elliptic and parabolic partial differential equations. Then, for semilinear elliptic and parabolic partial differential equations, we shall perform

various numerical experiments to investigate which properties (such as non-negative constraint, discrete maximum principles, discrete comparison principles, and monotone property) are inherited from continuous to discrete setting. Analysis is carried out based on the standard single-field Galerkin formulation and a non-negative formulation using nonlinear iterative techniques such as Pao's method, traditional Newton-Raphson method, and modification of traditional Newton-Raphson method. Through a representative numerical example we shall demonstrate that under discrete setting not all continuous properties are inherited. Specifically, traditional Newton-Raphson method and its modifications *do not* preserve the monotone property. Additionally, they need not satisfy the non-negative constraint, discrete maximum principles, and discrete comparison principles under the standard single-field Galerkin formulation. However, Pao's method *preserves* all the discrete properties under certain conditions on the mesh and time-step. Finally, through numerical examples we shall show that due to mesh and time-step restrictions Pao's method might be computationally expensive as compared to the traditional Newton-Raphson method and its modifications. Moreover, from the numerical experiments we observe that the terminal rate of convergence for Pao's method is almost linear.

6.1 INTRODUCTION AND MOTIVATION

This chapter deals with numerical formulations and their ability to preserve the underlying mathematical properties of semilinear elliptic and parabolic partial differential equations. Reaction-Diffusion equations arise in various areas of life sciences Rice (1985); Farkas (2001); Murray (1993), porous media applications Bowen (1976); Dentz et al. (2011); Hornung (1996), and chemically reacting systems Erdi and Toth (1989); McCarty and Criddle (2012); Kotomin and Kuzovkov (1996). The governing

equations for a general reaction-diffusion system hinges on various physical parameters such as scale of physical domain and its geometric effects, anisotropy and heterogeneity of the diffusivity tensor, temporal scales involved in the chemical reactions, influence of catalyst, diffusion rate, and most importantly the temperature of the system. In many scenarios, it is well-known in literature that the governing equations for such type of reaction-diffusion systems are modeled as either semilinear second-order elliptic or parabolic partial differential equations based on the steady-state or transient response Pao (1993); Mei (2000); Leung (2009). In continuous setting, it has been shown that these equations satisfy various important mathematical principles such as maximum principles, comparison principles, and monotone property Pao (1993); Leung (2009). But in the discrete setting, even for the linear case, it is recognized that many popular numerical formulations and commercially available packages violate these important properties Liska and Shashkov (2008); Ciarlet and Raviart (1973); Nagarajan and Nakshatrala (2011); Nakshatrala and Valocchi (2009); Nakshatrala et al. (2013). To our knowledge such type of analysis is not performed extensively within the context of reaction-diffusion equations. Herein, our objective is to analyze whether such discrete properties are preserved or lost within the context of standard single-field Galerkin and non-negative formulations.

6.1.1 Main contributions and outline of this chapter

In this chapter, we shall briefly discuss on the following three important numerical aspects related to linear and semilinear elliptic and parabolic partial differential equations:

- Pros and cons of mesh and time-step restrictions to satisfy different discrete properties such as non-negative constraint, discrete maximum principles, discrete comparison principles, and monotone property within the context of standard single-field Galerkin formulation.

- Pros and cons of various nonlinear techniques such as Pao’s method, traditional Newton-Raphson method and its modifications in satisfying different discrete properties both under standard single-field Galerkin and non-negative formulations.
- Need for *robust physics-compatible numerical formulations and time-stepping schemes* to obtain stable and accurate solutions for reaction-diffusion equations which exhibit bifurcations, blowups, and most importantly *fast and slow* decay/growth of solutions based on the nature of (reaction) volumetric source.

The remainder of this chapter is as follows: In Section 6.2, we discuss on governing equations for reaction-diffusion systems, various important mathematical principles, and mesh and time-step restrictions for linear second-order elliptic and parabolic partial differential equations. In Section 6.3, we will illustrate Pao’s method to obtain numerical solutions to semilinear elliptic and parabolic second-order partial differential equations within the context of both standard single-field Galerkin formulation and non-negative methodology. Furthermore, we shall describe the traditional Newton-Raphson method and its modifications to obtain numerical solutions and examine their differences with respect to the Pao’s method in satisfying various discrete principles. In Section 6.4, we discuss various reaction models of monotone-type. Representative numerical examples are presented in Section 6.5. Finally, conclusions are drawn in Section 6.6.

The standard symbolic notation is adopted in this chapter. We shall denote scalars by lower-case English alphabet or lower-case Greek alphabet (e.g., concentration c and density ρ). We shall make a distinction between vectors in the continuum and finite element settings. Similarly, a distinction shall be made between second-order tensors in the continuum setting versus matrices in the context of the finite element method. The continuum vectors are denoted by lower case boldface normal

letters, and the second-order tensors will be denoted using upper case boldface normal letters (e.g., 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 (e.g., vector \mathbf{v} and matrix \mathbf{K}). It should be noted that repeated indices do not imply summation. (That is, Einstein’s summation convention is not employed in this chapter.) Other notational conventions adopted in this chapter are introduced as needed.

6.2 GOVERNING EQUATIONS AND MATHEMATICAL PRINCIPLES FOR REACTION-DIFFUSION EQUATIONS

Herein, we are interested in the numerical solution to the following reaction-diffusion equation, which mainly arises in autocatalytic chemical reactions Leung (2009); Mei (2000); Karátson and Korotov (2005) and diffusion-kinetic enzyme problems Murray (1968a,b):

$$\mathcal{L}[c] := \frac{\partial c(\mathbf{x}, t)}{\partial t} - \operatorname{div} [\mathbf{D}(\mathbf{x}) \operatorname{grad}[c(\mathbf{x}, t)]] = f(\mathbf{x}, t, c(\mathbf{x}, t)) \quad \text{in } \Omega \times (0, \mathcal{I}), \quad (6.2.1a)$$

$$c(\mathbf{x}, t) = c_p(\mathbf{x}, t) \quad \text{on } \partial\Omega \times (0, \mathcal{I}), \quad \text{and} \quad (6.2.1b)$$

$$c(\mathbf{x}, t = 0) = c_0(\mathbf{x}) \quad \text{in } \Omega, \quad (6.2.1c)$$

where $\Omega \subset \mathbb{R}^{nd}$ is a open bounded domain and “ nd ” denotes the number of spatial dimensions. The boundary of the domain is denoted by $\partial\Omega$, which is assumed to be piecewise smooth. Mathematically, $\partial\Omega := \overline{\Omega} - \Omega$, where a superposed bar denotes the set closure. A spatial point is denoted by $\mathbf{x} \in \overline{\Omega}$. The gradient and divergence operators with respect to \mathbf{x} are, respectively, denoted by $\operatorname{grad}[\bullet]$ and $\operatorname{div}[\cdot]$. The variable $t \in [0, \mathcal{I}]$ denotes the time and \mathcal{I} denotes the length of time interval. The quantity of interest, $c(\mathbf{x}, t)$, denotes the concentration of a reacting chemical species. Herein, it is assumed that Dirichlet boundary condition ‘ $c_p(\mathbf{x}, t)$ ’ is prescribed on

the entire boundary. The prescribed volumetric source $f(\mathbf{x}, t, c(\mathbf{x}, t))$ represents the reaction of the chemical species and $\mathbf{D}(\mathbf{x})$ is the anisotropic diffusivity tensor. Physics of the problem demands that the diffusivity tensor be symmetric, uniformly elliptic, and bounded above. This means, that there exists two constants $0 < \xi_1 \leq \xi_2 < +\infty$ such that

$$0 < \xi_1 \mathbf{y} \cdot \mathbf{y} \leq \mathbf{y} \cdot \mathbf{D}(\mathbf{x}) \mathbf{y} \leq \xi_2 \mathbf{y} \cdot \mathbf{y} \quad \forall \mathbf{y} \in \mathbb{R}^{nd} \setminus \{\mathbf{0}\} \quad \text{and} \quad \mathbf{x} \in \Omega. \quad (6.2.2)$$

In the literature on partial differential equations, it is well known that above equations given by (6.2.1a)–(6.2.1c) possess various types of maximum principles and comparison principles under certain hypothesis on the input data and domain regularity Evans (1998); Gilbarg and Trudinger (2001); Pucci and Serrin (2007). We shall now introduce some notation required to describe a continuous weak maximum and comparison principle.

6.2.1 Mathematical principles and relevant notation

Let $\Omega_{\mathcal{I}} := \Omega \times (0, \mathcal{I})$ denote a parabolic cylinder and the corresponding parabolic boundary is defined as:

$$\Gamma_{\mathcal{I}} := \left\{ (\mathbf{x}, t) \in \overline{\Omega_{\mathcal{I}}} \mid \mathbf{x} \in \partial\Omega \text{ or } t = 0 \right\}. \quad (6.2.3)$$

Let $C^m(\Omega)$ denote the set of functions defined on Ω that are continuously differentiable up to m -th order. Let $C_1^2(\Omega_{\mathcal{I}})$ be a function space with differing smoothness in variables \mathbf{x} and t defined as:

$$C_1^2(\Omega_{\mathcal{I}}) := \left\{ c(\mathbf{x}, t) : \Omega_{\mathcal{I}} \rightarrow \mathbb{R} \mid c, \frac{\partial c}{\partial x_i}, \frac{\partial^2 c}{\partial x_i \partial x_j}, \frac{\partial c}{\partial t} \in C(\Omega_{\mathcal{I}}); i, j = 1, 2, \dots, nd \right\}. \quad (6.2.4)$$

Based on the above notation, we shall now present a continuous weak maximum principle, continuous weak comparison principle, and monotone property related to semilinear second-order elliptic and parabolic partial differential equations in form of theorems without proofs. For proofs and further information, see References Pao (1993); Gilbarg and Trudinger (2001); Karátson and Korotov (2005).

Theorem 6.2.1 (Continuous weak maximum principle). *Let $c(\mathbf{x}, t) \in C_1^2(\Omega_{\mathcal{I}}) \cap C^0(\overline{\Omega}_{\mathcal{I}})$ and $\mathcal{L}[c] \leq 0$ in $\Omega_{\mathcal{I}}$. Then $c(\mathbf{x}, t)$ achieves its maximum on the parabolic boundary, which is given as:*

$$\max_{(\mathbf{x}, t) \in \overline{\Omega}_{\mathcal{I}}} [c(\mathbf{x}, t)] = \max_{(\mathbf{x}, t) \in \Gamma_{\mathcal{I}}} [c(\mathbf{x}, t)]. \quad (6.2.5)$$

Specifically, if $\mathcal{L}[c] := -\operatorname{div}[\mathbf{D}(\mathbf{x})\operatorname{grad}[c(\mathbf{x})]] \leq 0$, then we have the following result for $c(\mathbf{x})$:

$$\max_{\mathbf{x} \in \overline{\Omega}} [c(\mathbf{x})] = \max_{\mathbf{x} \in \partial\Omega} [c(\mathbf{x})]. \quad (6.2.6)$$

Theorem 6.2.2 (Continuous weak comparison principle). *Given that $c_1(\mathbf{x}, t)$ and $c_2(\mathbf{x}, t) \in C_1^2(\Omega_{\mathcal{I}}) \cap C^0(\overline{\Omega}_{\mathcal{I}})$. If $c_1(\mathbf{x}, t)$ and $c_2(\mathbf{x}, t)$ satisfy Theorem 6.2.1, $\mathcal{L}[c_1] \leq \mathcal{L}[c_2]$ in $\Omega_{\mathcal{I}}$, and $c_1(\mathbf{x}, t) \leq c_2(\mathbf{x}, t)$ on $\Gamma_{\mathcal{I}}$, then we have the following result:*

$$c_1(\mathbf{x}, t) \leq c_2(\mathbf{x}, t) \quad \forall (\mathbf{x}, t) \in \overline{\Omega}_{\mathcal{I}}. \quad (6.2.7)$$

Especially, if $\mathcal{L}[c] := -\operatorname{div}[\mathbf{D}(\mathbf{x})\operatorname{grad}[c(\mathbf{x})]]$, then we have the following result:

$$c_1(\mathbf{x}) \leq c_2(\mathbf{x}) \quad \forall \mathbf{x} \in \overline{\Omega}. \quad (6.2.8)$$

Definition 6.2.3 (Ordered upper and lower solutions). *Given that $c_{\text{us}}(\mathbf{x}, t) \in$*

$C_1^2(\Omega_{\mathcal{I}}) \cap C^0(\overline{\Omega_{\mathcal{I}}})$ and if it satisfies the following inequalities:

$$\mathcal{L}[c_{\text{us}}] \geq f(\mathbf{x}, t, c_{\text{us}}(\mathbf{x}, t)) \quad \text{in } \Omega \times (0, \mathcal{I}), \quad (6.2.9a)$$

$$c_{\text{us}}(\mathbf{x}, t) \geq c_{\text{p}}(\mathbf{x}, t) \quad \text{on } \partial\Omega \times (0, \mathcal{I}), \text{ and} \quad (6.2.9b)$$

$$c_{\text{us}}(\mathbf{x}, t = 0) \geq c_0(\mathbf{x}) \quad \text{in } \Omega, \quad (6.2.9c)$$

then $c_{\text{us}}(\mathbf{x}, t)$ is called an upper solution to the equations (6.2.1a)–(6.2.1c). Correspondingly, $c_{\text{ls}}(\mathbf{x}, t) \in C_1^2(\Omega_{\mathcal{I}}) \cap C^0(\overline{\Omega_{\mathcal{I}}})$ is called a lower solution if it satisfies the reverse inequalities given by equations (6.2.10)–(6.2.9c). The variables $c_{\text{us}}(\mathbf{x}, t)$ and $c_{\text{ls}}(\mathbf{x}, t)$ are called ordered upper and lower solutions if they satisfy Theorem (6.2.2). That is,

$$c_{\text{ls}}(\mathbf{x}, t) \leq c_{\text{us}}(\mathbf{x}, t) \quad \forall (\mathbf{x}, t) \in \overline{\Omega_{\mathcal{I}}}. \quad (6.2.10)$$

Similarly, one can define ordered upper and lower solutions if $\mathcal{L}[c] := -\text{div}[\mathbf{D}(\mathbf{x})\text{grad}[c(\mathbf{x})]]$.

Theorem 6.2.4 (Monotone property). *Given that $c_{\text{us}}(\mathbf{x}, t)$ and $c_{\text{ls}}(\mathbf{x}, t)$ are an ordered upper and lower solutions to the governing equations given by (6.2.1a)–(6.2.1c). Let a sequence of solutions $\{c_k(\mathbf{x}, t)\}_{k \in \mathbb{N}}$ satisfy the following set of inequalities for every $k \in \mathbb{N}$:*

$$\mathcal{L}[c_k] \geq f(\mathbf{x}, t, c_k(\mathbf{x}, t)) \quad \text{in } \Omega \times (0, \mathcal{I}), \quad (6.2.11a)$$

$$c_k(\mathbf{x}, t) \geq c_{\text{p}}(\mathbf{x}, t) \quad \text{on } \partial\Omega \times (0, \mathcal{I}), \text{ and} \quad (6.2.11b)$$

$$c_k(\mathbf{x}, t = 0) \geq c_0(\mathbf{x}) \quad \text{in } \Omega. \quad (6.2.11c)$$

Furthermore, if the sequence $\{c_k(\mathbf{x}, t)\}_{k \in \mathbb{N}}$ satisfies the following inequality:

$$f(\mathbf{x}, t, c_{\text{ls}}(\mathbf{x}, t)) \leq \cdots \leq f(\mathbf{x}, t, c_k(\mathbf{x}, t)) \leq \cdots \leq f(\mathbf{x}, t, c_{\text{us}}(\mathbf{x}, t)) \\ \forall(\mathbf{x}, t) \in \Omega_{\mathcal{I}} \quad \text{and} \quad k \in \mathbb{N}. \quad (6.2.12)$$

then $\{c_k(\mathbf{x}, t)\}_{k \in \mathbb{N}}$ possess the monotone property given as:

$$c_{\text{ls}}(\mathbf{x}, t) \leq \cdots \leq c_k(\mathbf{x}, t) \leq \cdots \leq c_{\text{us}}(\mathbf{x}, t) \\ \forall(\mathbf{x}, t) \in \overline{\Omega}_{\mathcal{I}} \quad \text{and} \quad k \in \mathbb{N}. \quad (6.2.13)$$

Specifically, if $\mathcal{L}[c] := -\text{div}[\mathbf{D}(\mathbf{x})\text{grad}[c(\mathbf{x})]] \leq 0$, then we have the following result for $\{c_k(\mathbf{x})\}_{k \in \mathbb{N}}$:

$$c_{\text{ls}}(\mathbf{x}) \leq \cdots \leq c_k(\mathbf{x}) \leq \cdots \leq c_{\text{us}}(\mathbf{x}) \forall \mathbf{x} \in \overline{\Omega} \quad \text{and} \quad k \in \mathbb{N}. \quad (6.2.14)$$

6.2.2 Mesh and time-step restrictions for linear second-order elliptic and parabolic PDEs

In the literature on discrete maximum principles, there are several papers which discuss about restrictions on mesh and time-step to satisfy different types of discrete properties for transient diffusion-type equations. Some of the notable works in this direction include Horváth (2008); Berzins (2001); Elshebli (2008); Faragó et al. (2005); Ilinca and Héту (2002); Mizukami (1986); Porru and Serra (1994); Rank et al. (1983); Thomas and Zhou (1998). However, most of these works are concerned with either one-dimensional problems or isotropic media. Most importantly, they did not consider effects of anisotropy and heterogeneity of the medium in to consideration. Recently, Huang and co-workers have proposed mesh restrictions to satisfy discrete maximum principles for steady-state anisotropic diffusion-type equations Li and Huang (2010);

Huang and Li (2010); Huang (2010, 2013); Lu et al. (2012). This is based on a non-linear and iterative anisotropic \mathcal{M} -uniform mesh generation technique Huang (2001, 2005, 2006), wherein the metric tensor $\mathcal{M}(\mathbf{x})$ is evaluated based on the inverse of integral average of the diffusivity tensor. Employing this methodology in-combination with the method of vertical lines, they derived time-step restrictions to satisfy discrete maximum principles for time stepping schemes that fall under generalized α -method Li and Huang (2013) and explicit Runge-Kutta method Huang et al. (2013). Alternatively, one can discretize the linear parabolic second-order partial differential equation apriorly using any type of time integrator (which can be explicit or implicit) and then apply mesh restrictions to satisfy discrete maximum principles. Each methodology has its own advantages and disadvantages. The following are the main pros and cons of each methodology:

- (1) **Method of vertical lines (MOVL):** Given an anisotropic \mathcal{M} -uniform mesh measured in a metric specified by inverse of integral average of the diffusivity tensor, method of vertical lines provides the following bounds on the time-step:

$$0 < \zeta_{\text{DMP}} h^2 \leq \Delta t \leq \zeta_{\text{Stab}} h^2, \quad (6.2.15)$$

where h is the maximum element diameter of the given mesh. ζ_{DMP} is the lower bound needed to satisfy the discrete maximum principle. Correspondingly, ζ_{Stab} is the upper bound required for the sake of stability. These bounds are dependent on the time-stepping scheme, integral average of the diffusivity tensor, minimum and maximum eigenvalues of the inverse of integral average of the diffusivity tensor, and other mesh related parameters Li and Huang (2013); Huang et al. (2013). The advantage of this methodology is that one can satisfy discrete principles even for coarse anisotropic \mathcal{M} -uniform meshes if the condition given on Δt (equation (6.2.15)) is met. But on the other hand accuracy is compromised if a coarse mesh

Table 6.1: Anisotropic diffusivity tensor test problem: Quantitative results for minimum concentration and % of nodes that have violated the non-negative constraint in the computational domain.

Δt	Min. Conc. Value	% of nodes violated
10^{-3}	-2.16×10^{-1}	7.25
10^{-4}	-2.64×10^{-1}	22.45
10^{-5}	-2.97×10^{-1}	11.05

is used. Moreover, it should be noted that the condition on Δt given by equation (6.2.15) might be stringent if $\mathbf{D}(\mathbf{x})$ is highly heterogeneous and anisotropic, which is the case in many real-life applications Pinder and Celia (2006); Zheng and Bennett (2002).

- (2) **Method of horizontal lines (MOHL):** In this method, as time discretization is performed apriorly, this gives rise to a linear second-order diffusion with decay-type equation. The decay constant is positive and inversely proportional to Δt . The main advantage of MOHL as compared to MOVL is that there is no need to impose restrictions on Δt . However, the mesh restrictions become much more severe compared to MOVL approach as the decay constant is involved. Smaller time-steps results in higher values of decay constant. This means that we need fine anisotropic \mathcal{M} -uniform meshes to satisfy discrete maximum principles, which increases the computational cost dramatically.

Herein, we shall take the MOHL approach. Through a representative numerical example, we shall demonstrate the salient aspects involved in satisfying various discrete principles based on such anisotropic \mathcal{M} -uniform meshes.

6.2.2.1 *Anisotropic diffusivity tensor test problem*

The computational domain Ω is a bi-unit square with a square hole of length equal to 0.1. The diffusivity tensor is given by equation (4.1.1) and (4.1.2a). Herein, we assume $d_{\max} = 10^4$, $d_{\min} = 1$, and $\theta = \pi/6$. A pictorial description of the boundary

value problem and an anisotropic \mathcal{M} -uniform mesh corresponding to the diffusivity tensor is shown in the Figure 6.1. Vertex O is located at $(0,0)$ while vertex H is located at $(0.45,0.45)$. Dirichlet boundary conditions are prescribed on the square hole and on the bi-unit square domain.

This mesh is created using BAMG in FreeFem++ based on the nonlinear and iterative anisotropic \mathcal{M} -uniform mesh generation technique given by Algorithm 1. Correspondingly, the concentration profiles obtained using this mesh is shown in Figure 6.2. The white region in the subfigures of Figure 6.2 represents the area in which concentration is negative. For transient case, analysis is performed using various values of Δt . Correspondingly, the contours represent the concentration profiles at the first time-step. From this figure, it is evident that the anisotropic \mathcal{M} -uniform mesh (given in Figure 6.2) does not violate the non-negative constraint and discrete maximum principles in-case of steady-state and for transient analysis when Δt is equal to either 0.1 or 0.01. But when Δt is either 10^{-3} or 10^{-4} or 10^{-5} , there is considerable violation in the non-negative constraint. Accordingly, the minimum value for concentration and the % of nodes that have violated the non-negative constraint (below the machine epsilon) is quantified in Table 6.1. Hence, in order to avoid violation of non-negative constraint and discrete maximum principles for smaller time-steps, one needs to highly refine the mesh. Nevertheless, this increases the computational cost as one needs to generate a h -refined anisotropic \mathcal{M} -uniform mesh and then solve the diffusion-type equation on this mesh. It should be noted that generating such h -refined mesh (which need to adhere to certain constraints given by various discrete principles) might be difficult and sometimes impossible Schneider (2013); George and Frey (2010).

In the next section, we shall discuss on various nonlinear methods to obtain numerical solutions to semilinear elliptic and parabolic partial differential equations within the context of mesh and time-step restrictions. Moreover, we shall address

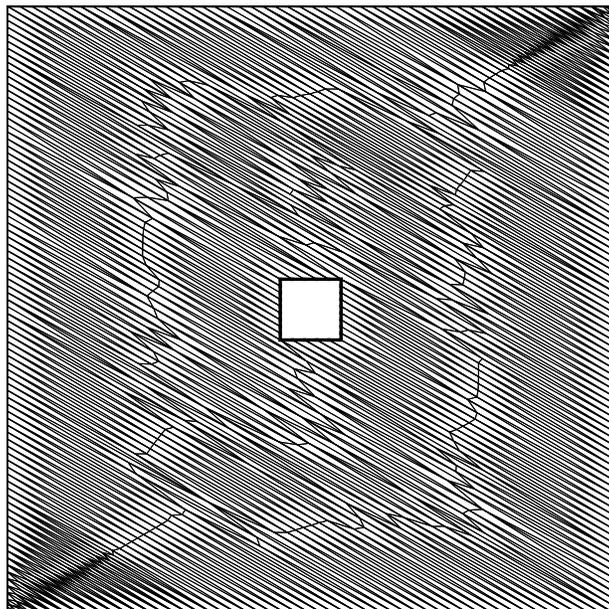
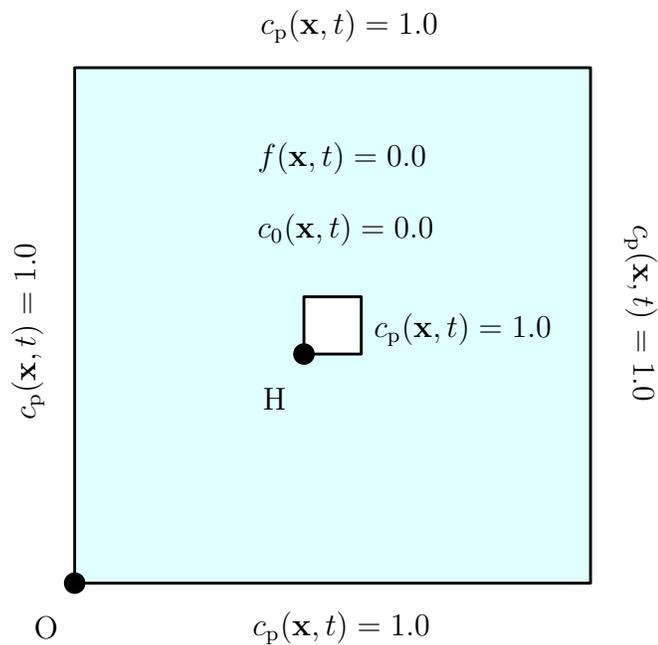


Figure 6.1: Anisotropic diffusivity tensor test problem: The left figure provides a pictorial description of the problem with the relevant boundary and initial conditions. The right figure shows the anisotropic \mathcal{M} -uniform mesh employed in the computational study.

which discrete properties are *preserved* or *lost* on low-order finite element discretization of standard single-field Galerkin and non-negative formulations for reaction-diffusion systems.

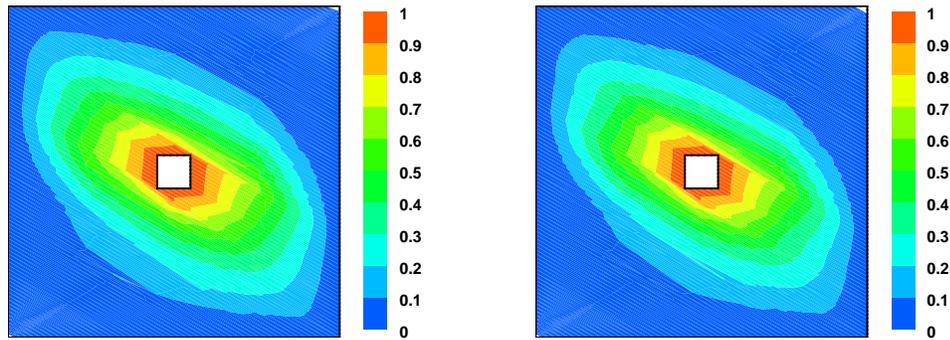
6.3 NONLINEAR TECHNIQUES TO SOLVE REACTION-DIFFUSION EQUATIONS

In general, for semilinear second-order elliptic and parabolic PDEs, the number and stability of numerical solutions depend upon the attribute of the volumetric (reaction) source term. A small variation in input parameters may cause abrupt change in the nature of the solution and may violate various discrete properties. It should be noted that these parameters are not only specific to the chemical reaction source term but also depend on the mesh size, time-step value, and numerical method employed. Herein, using various representative numerical examples, we analyze whether the existing and popular numerical schemes such as Pao's method, Picard's method, consistent linearization method, traditional Newton-Raphson method, and modification of traditional Newton-Raphson method can handle such scenarios.

6.3.1 Pao's method

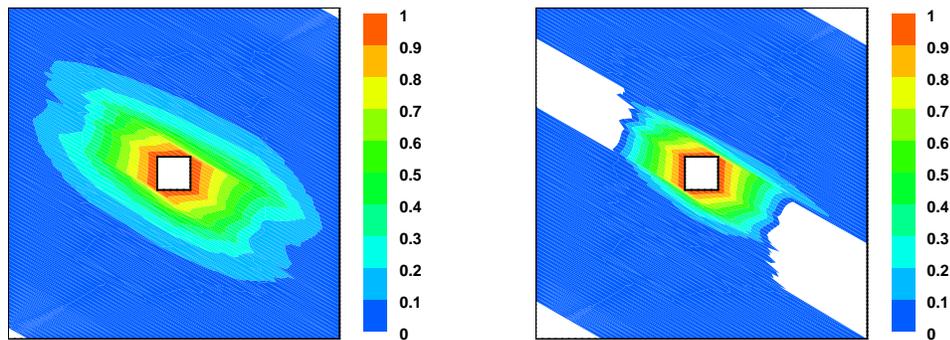
This method is also called as traditional monotone iterative method. In continuous setting, this method offers an existence-comparison theorem for both steady-state and transient reaction-diffusion equations of type given by equations (6.2.1a)–(6.2.1c). Pao's methodology is based on constructing ordered lower and upper solutions using monotone iteration.

Definition 6.3.1 (Sector of ordered upper and lower solutions). *Let $c_{ls}(\mathbf{x}, t)$ and $c_{us}(\mathbf{x}, t)$ be any given ordered lower and upper solutions. Then the sector of*



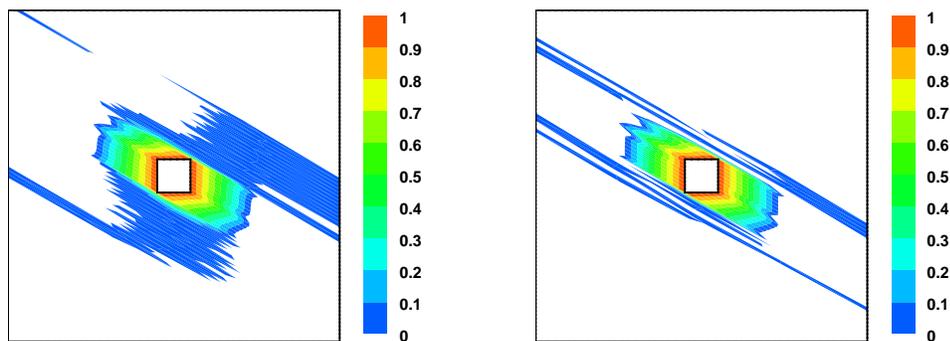
(a) Steady-state

(b) Transient: $\Delta t = 10^{-1}$



(c) Transient: $\Delta t = 10^{-2}$

(d) Transient: $\Delta t = 10^{-3}$



(e) Transient: $\Delta t = 10^{-4}$

(f) Transient: $\Delta t = 10^{-5}$

Figure 6.2: Anisotropic diffusivity tensor test problem (mesh and time-step restrictions): The above figures show the concentration profiles for steady-state and transient cases.

ordered upper and lower solutions $\langle c_{\text{ls}}, c_{\text{us}} \rangle$ is defined as a functional interval given as:

$$\langle c_{\text{ls}}, c_{\text{us}} \rangle := \left\{ c(\mathbf{x}, t) \in C^0(\overline{\Omega}_{\mathcal{I}}) \mid c_{\text{ls}} \leq c \leq c_{\text{us}} \right\}. \quad (6.3.1)$$

Theorem 6.3.2 (Continuous Pao's method). *Given that $f(\mathbf{x}, t, c(\mathbf{x}, t))$ satisfies the following one-sided Lipschitz condition:*

$$f(\mathbf{x}, t, c_i) - f(\mathbf{x}, t, c_j) \geq -\beta_{g\text{Lip}}(c_i - c_j) \quad \text{for any } c_i \text{ and } c_j \text{ satisfying } c_{\text{ls}} \leq c_j \leq c_i \leq c_{\text{us}}, \quad (6.3.2)$$

where $\beta_{g\text{Lip}}$ is a constant. Then the following iterative scheme (also called as Pao's method):

$$\frac{\partial c_k}{\partial t} - \text{div}[\mathbf{D}(\mathbf{x})\text{grad}[c_k]] + \beta_{g\text{Lip}}c_k = \beta_{g\text{Lip}}c_{k-1} + f(\mathbf{x}, t, c_{k-1}) \quad \text{in } \Omega \times (0, \mathcal{I}), \quad (6.3.3a)$$

$$c_k(\mathbf{x}, t) = c_p(\mathbf{x}, t) \quad \text{on } \partial\Omega \times (0, \mathcal{I}), \text{ and} \quad (6.3.3b)$$

$$c_k(\mathbf{x}, t = 0) = c_0(\mathbf{x}) \quad \text{in } \Omega \quad (6.3.3c)$$

to construct a sequence $\{c_k\}_{k \in \mathbb{N}}$ successively based on a suitable given initial iteration c_1 satisfies the monotone property. Moreover, the lower sequence $\{c_{k,\text{ls}}\}_{k \in \mathbb{N}}$ constructed with the initial iteration $c_1 = c_{\text{ls}}$ and the upper sequence $\{c_{k,\text{us}}\}_{k \in \mathbb{N}}$ created with $c_1 = c_{\text{us}}$ also satisfy the monotone property and are related as:

$$c_{\text{ls}} \leq c_{k,\text{ls}} \leq c_{k+1,\text{ls}} \leq c_{k+1,\text{us}} \leq c_{k,\text{us}} \leq c_{\text{us}} \quad \text{in } \overline{\Omega}_{\mathcal{I}} \quad \forall k = 1, 2, \dots \quad (6.3.4)$$

Proof. A proof can be found in Pao (1993). □

Theorem 6.3.3 (Existence of a transient ordered lower and upper solution).

Let $c_{\text{ls}}(\mathbf{x}, t)$ and $c_{\text{us}}(\mathbf{x}, t)$ be a nonnegative ordered lower and upper solution. If the reaction volumetric source $f(\mathbf{x}, t, c(\mathbf{x}, t))$, Dirichlet boundary condition $c_p(\mathbf{x}, t)$, and

initial condition $c_0(\mathbf{x})$, satisfy the following inequalities:

$$f(\mathbf{x}, t, c(\mathbf{x}, t) = 0) \geq 0, \quad c_p(\mathbf{x}, t) \geq 0, \quad c_0(\mathbf{x}) \geq 0 \quad (6.3.5)$$

then there exists a unique solution $c(\mathbf{x}, t)$ to the system of equations given by (6.2.1a)–(6.2.1c) and

$$c_{ls}(\mathbf{x}, t) = 0 \leq c(\mathbf{x}, t) \leq c_{us}(\mathbf{x}, t) \quad \text{in } \bar{\Omega}_{\mathcal{I}}. \quad (6.3.6)$$

Furthermore, let $f(\mathbf{x}, t, c(\mathbf{x}, t))$ be a C^1 -functional in the sector $\langle 0, c_{us} \rangle$ and for some positive constant ρ , if

$$f(\mathbf{x}, t, c(\mathbf{x}, t) = \rho) \leq 0, \quad c_p(\mathbf{x}, t) \leq \rho, \quad c_0(\mathbf{x}) \leq \rho \quad (6.3.7)$$

then $c_{us}(\mathbf{x}, t) = \rho$ is a positive upper solution and $0 \leq c(\mathbf{x}, t) \leq \rho$ in $\bar{\Omega}_{\mathcal{I}}$.

Proof. A proof can be found in Pao (1993). □

Theorem 6.3.4 (Existence of a steady-state ordered lower and upper solution).

Let $c_{ls}(\mathbf{x})$ and $c_{us}(\mathbf{x})$ be a nonnegative ordered lower and upper solution with $c_p(\mathbf{x}) \geq 0$. Given that $f(\mathbf{x}, c(\mathbf{x})) \geq 0$ or $f(\mathbf{x}, c(\mathbf{x})) \leq 0$ be a C^1 -functional for $c(\mathbf{x}) \in \langle 0, c_{us} \rangle$ such that either $f(\mathbf{x}, c(\mathbf{x}) = 0)$ or $c_p(\mathbf{x})$ is not identically zero. Then there exists at least one positive solution $c(\mathbf{x}) \in \langle 0, c_{us} \rangle$ and $c(\mathbf{x})$ is unique if $\frac{\partial f}{\partial c} \leq 0$ in $\langle 0, c_{us} \rangle$. Moreover, for some positive constant ρ , if

$$f(\mathbf{x}, c(\mathbf{x}) = \rho) \leq 0, \quad c_p(\mathbf{x}) \leq \rho, \quad (6.3.8)$$

then $c_{ls}(\mathbf{x}) = 0 \leq c(\mathbf{x}) \leq c_{us}(\mathbf{x}) = \rho$ in $\bar{\Omega}$ and the above conclusions hold in the sector $\langle 0, \rho \rangle$.

Proof. A proof can be found in Pao (1993). □

Theorem 6.3.5 (Discrete Pao's method). *Let $F_{k-1} := \beta_{gLip}c_{k-1} + f(\mathbf{x}, t, c_{k-1})$ and correspondingly the lower-order finite element discretization of F_{k-1} at any given time level $0 \leq t = t_n \leq \mathcal{I}$ be denoted as $\mathbf{F}_{k-1}^{(n)}$. It is given that method of horizontal lines based on the time-stepping schemes that fall under generalized α -method is applied to the equations (6.3.3a)–(6.3.3c). Furthermore, denote the stiffness matrix obtained using the standard Galerkin lower-order finite element discretization of equations (6.3.3a)–(6.3.3c) as \mathbf{K} . If \mathbf{K} satisfies the following conditions (properties of M -matrix Varga (2009)):*

$$(a) \mathbf{K}_{ff}^{-1} \succeq \mathbf{O} \quad (b) -\mathbf{K}_{ff}^{-1} \mathbf{K}_{fp} \succeq \mathbf{O} \quad (c) -\mathbf{K}_{ff}^{-1} \mathbf{K}_{fp} \mathbf{1} \preceq \mathbf{1}, \quad (6.3.9)$$

then the iterative scheme given in Theorem 6.3.2 to construct a sequence $\{\mathbf{c}_k^{(n)}\}_{k \in \mathbb{N}}$ successively based on a suitable given initial iteration $\mathbf{c}_{1,f}^{(n)}$ satisfies the monotone property. Additionally, we have the following monotone relation for the lower sequence $\{\mathbf{c}_{k,ls}^{(n)}\}_{k \in \mathbb{N}}$ and upper sequence $\{\mathbf{c}_{k,us}^{(n)}\}_{k \in \mathbb{N}}$ constructed based on the initial iteration $\mathbf{c}_{1,f}^{(n)} = \mathbf{c}_{ls,f}$ and $\mathbf{c}_{1,f}^{(n)} = \mathbf{c}_{us,f}$:

$$\mathbf{c}_{ls} \preceq \mathbf{c}_{k,ls}^{(n)} \preceq \mathbf{c}_{k+1,ls}^{(n)} \preceq \mathbf{c}_{k+1,us}^{(n)} \preceq \mathbf{c}_{k,us}^{(n)} \preceq \mathbf{c}_{us} \quad \text{in } \bar{\Omega}_{\mathcal{I}} \quad \forall k = 1, 2, \dots, \quad (6.3.10)$$

where the stiffness matrix $\mathbf{K} \equiv [\mathbf{K}_{ff} | \mathbf{K}_{fp}]$, $\mathbf{c}_k^{(n)} \equiv [(\mathbf{c}_{k,f}^{(n)})^T | \mathbf{c}_p^T]^T$ denotes the concentration vector at $t = t_n$ and iteration level k , $\mathbf{c}_{k,ls}^{(n)} \equiv [(\mathbf{c}_{k,ls,f}^{(n)})^T | \mathbf{c}_p^T]^T$, $\mathbf{c}_{k,us}^{(n)} \equiv [(\mathbf{c}_{k,us,f}^{(n)})^T | \mathbf{c}_p^T]^T$, $\mathbf{c}_{ls} \equiv [\mathbf{c}_{ls,f}^T | \mathbf{c}_{ls,p}^T]^T$, $\mathbf{c}_{us} \equiv [\mathbf{c}_{us,f}^T | \mathbf{c}_{us,p}^T]^T$, the symbol \preceq represents the component-wise inequality for vectors and matrices, $\mathbf{0}$ denotes a zero vector, $\mathbf{1}$ denotes a vector of ones, and \mathbf{O} denotes the zero matrix. The stiffness matrices \mathbf{K} , \mathbf{K}_{ff} , and \mathbf{K}_{fp} are, respectively, of size $n_f \times n_t$, $n_f \times n_f$, and $n_f \times n_p$. The nodal volumetric source vector $\mathbf{F}_{k-1}^{(n)}$ is of size $n_f \times 1$. Correspondingly, the nodal concentration vectors $\mathbf{c}_k^{(n)}$, $\mathbf{c}_{k,f}^{(n)}$, and \mathbf{c}_p are of sizes $n_t \times 1$, $n_f \times 1$, and $n_p \times 1$, where “ n_t ” denote the total number of degrees-of-freedom, “ n_f ” denote the free degrees-of-freedom, “ n_p ”

be the prescribed degrees-of-freedom, and $n_t = n_f + n_p$.

Proof. According to the hypothesis of the Theorem 6.3.5 and the condition given by equation (6.3.2), it is evident that the discrete counter part of F_{k-1} , which is given by $\mathbf{F}_{k-1}^{(n)}$ satisfies the following inequality similar to equation (6.2.12):

$$\mathbf{F}_{\text{ls}} \preceq \cdots \cdots \preceq \mathbf{F}_{k-1}^{(n)} \preceq \mathbf{F}_k^{(n)} \preceq \cdots \cdots \preceq \mathbf{F}_{\text{us}} \\ \forall (\mathbf{x}, t) \in \Omega_{\mathcal{I}} \quad \text{and} \quad k \in \mathbb{N}. \quad (6.3.11)$$

From Definition 6.2.3, it is evident that $\mathbf{c}_{\text{ls},p} \preceq \mathbf{c}_p \preceq \mathbf{c}_{\text{us},p}$. On applying generalized α -method for temporal discretization (with time-stepping scheme parameters $\alpha_f = 1$ and $\alpha_m = \gamma \in (0, 1]$), the discrete equations based on lower-order finite element discretization of standard single-field Galerkin formulation are given as

$$\mathbf{K}_{ff} \mathbf{c}_{k,f}^{(n)} = \mathbf{F}_{k-1}^{(n)} + \frac{1}{\Delta t} \mathbf{c}_f^{(n)} - \mathbf{K}_{fp} \mathbf{c}_p, \quad (6.3.12)$$

where $\mathbf{c}_f^{(n)}$ is the converged solution at $t = t_n$ and \mathbf{c}_p is the discrete counter part of the Dirichlet boundary condition $c_p(\mathbf{x}, t)$. Assuming principle of mathematical induction and a suitable initial guess $\mathbf{c}_{1,f}^{(n)}$ (for example $\mathbf{c}_{1,f}^{(n)} = \mathbf{0}$), it is apparent that the sequence $\{\mathbf{c}_k^{(n)}\}_{k \in \mathbb{N}}$ satisfies the monotone property through trivial algebraic manipulations on equation (6.3.12) using the conditions provided by equations (6.3.9) and (6.3.11). Furthermore, the initial guess $\mathbf{c}_{1,f}^{(n)} = \mathbf{0}$ also happens to be the discrete ordered lower solution $\mathbf{c}_{\text{ls},f}$. Hence, we have the following additional result:

$$\mathbf{c}_{\text{ls}} \preceq \mathbf{c}_{k,\text{ls}}^{(n)} \preceq \mathbf{c}_{k+1,\text{ls}}^{(n)} \preceq \mathbf{c}_{\text{us}} \quad \text{in } \bar{\Omega}_{\mathcal{I}} \quad \forall k = 1, 2, \dots \quad (6.3.13)$$

Similarly, for initial guess $\mathbf{c}_{1,f}^{(n)} = \mathbf{c}_{\text{us},f}$, using equations (6.3.12), (6.3.9), and (6.3.11),

we have the following set of inequalities:

$$\mathbf{c}_{2,\text{us}}^{(n)} \preceq \mathbf{c}_{1,\text{us}}^{(n)} \text{ and} \quad (6.3.14\text{a})$$

$$\mathbf{c}_{1,\text{ls}}^{(n)} \preceq \mathbf{c}_{2,\text{ls}}^{(n)} \preceq \mathbf{c}_{2,\text{us}}^{(n)} \preceq \mathbf{c}_{1,\text{us}}^{(n)}. \quad (6.3.14\text{b})$$

By appealing to principle of mathematical induction, we have the final result:

$$\mathbf{c}_{\text{ls}} \preceq \mathbf{c}_{k,\text{ls}}^{(n)} \preceq \mathbf{c}_{k+1,\text{ls}}^{(n)} \preceq \mathbf{c}_{k+1,\text{us}}^{(n)} \preceq \mathbf{c}_{k,\text{us}}^{(n)} \preceq \mathbf{c}_{\text{us}} \quad \text{in } \overline{\Omega}_{\mathcal{I}} \quad \forall k = 1, 2, \dots \quad (6.3.15)$$

which completes the proof. \square

It should be noted that if there exists a constant $\beta_{lLip} \geq -\beta_{gLip}$ such that the following inequality holds:

$$f(\mathbf{x}, t, c_i) - f(\mathbf{x}, t, c_j) \leq \beta_{lLip}(c_i - c_j) \quad \text{for any } c_i \text{ and } c_j \text{ satisfying } c_{\text{ls}} \leq c_j \leq c_i \leq c_{\text{us}} \quad (6.3.16)$$

then the Pao's iterative scheme described in Theorem 6.3.2 has to be modified as described below:

$$\frac{\partial c_k}{\partial t} - \text{div} [\mathbf{D}(\mathbf{x})\text{grad}[c_k]] - \beta_{lLip}c_k = -\beta_{lLip}c_{k-1} + f(\mathbf{x}, t, c_{k-1}) \quad \text{in } \Omega \times (0, \mathcal{I}), \quad (6.3.17\text{a})$$

$$c_k(\mathbf{x}, t) = c_p(\mathbf{x}, t) \quad \text{on } \partial\Omega \times (0, \mathcal{I}), \text{ and} \quad (6.3.17\text{b})$$

$$c_k(\mathbf{x}, t = 0) = c_0(\mathbf{x}) \quad \text{in } \Omega. \quad (6.3.17\text{c})$$

In the case when the reaction volumetric source $f(\mathbf{x}, t, c(\mathbf{x}, t))$ satisfies a Lipschitz condition in the form:

$$|f(\mathbf{x}, t, c_i) - f(\mathbf{x}, t, c_j)| \leq \beta_{Lip}|(c_i - c_j)| \quad \text{for any } c_i \text{ and } c_j \in \langle c_{\text{ls}}, c_{\text{us}} \rangle \quad (6.3.18)$$

the constants β_{gLip} , β_{lLip} , and β_{Lip} are given by $\beta_{gLip} = \beta_{lLip} = \beta_{Lip}$. Continuous

and discrete conditions outlined in Theorems 6.3.2 and 6.3.5 can be constructed in a similar fashion for the above set of equations (6.3.17a)–(6.3.17c). Herein, we shall skip such details for sake of saving space. In subsection 6.5.1, based on the above discrete setting, we shall compute the numerical solution and show that the resulting sequence of iterations is monotone and converges to a solution of the given problem. Other important aspects such as rate of convergence of discrete Pao’s method and satisfaction of various discrete principles are also discussed. Correspondingly, the residual at iteration level k (until convergence) for the Pao’s method is plotted, which is given as

$$\begin{aligned} \mathcal{R}_{k,\text{Pao}} := & \left(w; \frac{c_k(\mathbf{x}, t = t_n) - c(\mathbf{x}, t = t_{n-1})}{\Delta t} \right) + (\text{grad}[w]; \mathbf{D}(\mathbf{x})\text{grad}[c_k(\mathbf{x}, t = t_n)]) \\ & - (w; f(\mathbf{x}, t = t_n, c_k(\mathbf{x}, t = t_n))), \end{aligned} \quad (6.3.19)$$

where $w \in H_0^1(\Omega)$ is the test function Ern and Guermond (2000) and $H_0^1(\Omega)$ is the standard Sobolev space in Ω Evans (1998).

6.3.2 Picard’s method

This method is also known in numerical analysis literature as traditional fixed-point method Heath (2005); Quarteroni et al. (2006); Atkinson and Han (2001). It should be noted that Picard’s numerical scheme is a subset to that of Pao’s method and can be obtained by setting the parameter β in Theorem 6.3.2 to be equal to zero. Hence, all the relevant mathematical analysis and corresponding restrictions on the stiffness matrices can be borrowed from the Pao’s method.

6.3.3 Traditional Newton-Raphson method

In numerical analysis Heath (2005) and nonlinear finite element method Belytschko et al. (2000), one of the most frequently applied scheme to solve a system

of nonlinear algebraic equations is the traditional Newton-Raphson method Kelley (1987). The objective of this scheme is evaluate $c_{k+1}(\mathbf{x}, t = t_n)$ based on the Taylor series expansion (or linearization) of the residual at iteration level $k + 1$, which is given as

$$\mathcal{R}_{k+1,\text{TNR}} = \mathcal{R}_{k,\text{TNR}} + \frac{D\mathcal{R}_{k,\text{TNR}}}{Dc_k} \Delta c_{k+1}, \quad (6.3.20)$$

where $\frac{D\mathcal{R}_{k,\text{TNR}}}{Dc_k}$ corresponds to the Fréchet derivative at $c_k(\mathbf{x}, t = t_n)$ and $\Delta c_{k+1} := c_{k+1}(\mathbf{x}, t = t_n) - c_k(\mathbf{x}, t = t_n)$. Correspondingly, the expression for the residual $\mathcal{R}_{k,\text{TNR}}$ is given as

$$\begin{aligned} \mathcal{R}_{k,\text{TNR}} := & \left(w; \frac{c_k(\mathbf{x}, t = t_n) - c(\mathbf{x}, t = t_{n-1})}{\Delta t} \right) + (\text{grad}[w]; \mathbf{D}(\mathbf{x})\text{grad}[c_k(\mathbf{x}, t = t_n)]) \\ & - (w; f(\mathbf{x}, t = t_n, c_k(\mathbf{x}, t = t_n))) \end{aligned} \quad (6.3.21)$$

and that of $\frac{D\mathcal{R}_{k,\text{TNR}}}{Dc_k} \Delta c_{k+1}$ is given as

$$\begin{aligned} \frac{D\mathcal{R}_{k,\text{TNR}}}{Dc_k} \Delta c_{k+1} := & \left(w; \frac{\Delta c_{k+1}}{\Delta t} \right) + (\text{grad}[w]; \mathbf{D}(\mathbf{x})\text{grad}[\Delta c_{k+1}]) \\ & - \left(w; \frac{df(\mathbf{x}, t = t_n, c_k + \epsilon \Delta c_{k+1})}{d\epsilon} \Big|_{\epsilon=0} \right). \end{aligned} \quad (6.3.22)$$

The solution variable $c_{k+1}(\mathbf{x}, t = t_n)$ is evaluated through the following two step process:

- First step is to evaluate Δc_{k+1} by equating the residual at $k + 1$ to be zero and solving the resulting system of linear equations.
- In second step, one obtains $c_{k+1}(\mathbf{x}, t = t_n)$ through the following update: $c_{k+1} = c_k + \Delta c_{k+1}$

6.3.4 Modification of traditional Newton-Raphson method

This scheme attempts to rectify one of the drawbacks of traditional Newton-Raphson method by reorganizing the expressions in equations (6.3.21) and (6.3.22). Corresponding, the modified residual of equation (6.3.20) in terms of the unknown variable $c_{k+1}(\mathbf{x}, t = t_n)$ is given as

$$\begin{aligned} \mathcal{R}_{k+1, \text{MTNR}} := & \left(w; \frac{c_{k+1}(\mathbf{x}, t = t_n)}{\Delta t} \right) + (\text{grad}[w]; \mathbf{D}(\mathbf{x})\text{grad}[c_{k+1}(\mathbf{x}, t = t_n)]) \\ & - \left(w; \frac{df(\mathbf{x}, t = t_n, c_k + \epsilon\Delta c_{k+1})}{d(c_k + \epsilon\Delta c_{k+1})} \Big|_{\epsilon=0} c_{k+1}(\mathbf{x}, t = t_n) \right) - \left(w; \frac{c(\mathbf{x}, t = t_n)}{\Delta t} \right) \\ & - \left(w; f(\mathbf{x}, t = t_n, c_k(\mathbf{x}, t = t_n)) - \frac{df(\mathbf{x}, t = t_n, c_k + \epsilon\Delta c_{k+1})}{d(c_k + \epsilon\Delta c_{k+1})} \Big|_{\epsilon=0} c_k(\mathbf{x}, t = t_n) \right). \end{aligned} \quad (6.3.23)$$

The solution variable $c_{k+1}(\mathbf{x}, t = t_n)$ is evaluated by designating $\mathcal{R}_{k+1, \text{MTNR}}$ to be equal to zero. Correspondingly, one can enforce the non-negative constraint to the resulting discrete equations obtained from the lower-order finite element discretization of equation (6.3.23). The resulting linear complementarity problem can be solved using the standard optimization algorithms (such as active-set methods, barrier methods, and interior-point methods) available in literature Boyd and Vandenberghe (2004); Nocedal and Wright (1999).

6.3.5 Consistent linearization method

Cornerstone in mathematical analysis of nonlinear partial differential equations is consistent linearization method. This methodology is widely used in obtaining numerical solutions based on nonlinear finite element method and assessing their stability. This technique is similar to that of traditional Newton-Raphson method and its modification. However, in-order to evaluate $c_{k+1}(\mathbf{x}, t)$ at any time level, the operator ‘ \mathcal{L} ’ is linearized before constructing the residual. This linearized operator

$\mathcal{L}_{\text{CL}}[c_{k+1}]$ is given as

$$\mathcal{L}_{\text{CL}}[c_{k+1}] = \mathcal{L}_{\text{CL}}[c_k] + \frac{\text{D}\mathcal{L}_{\text{CL}}[c_k]}{\text{D}c_k} \Delta c_{k+1}, \quad (6.3.24)$$

where $\frac{\text{D}\mathcal{L}_{\text{CL}}[c_k]}{\text{D}c_k}$ corresponds to the Fréchet derivative at $c_k(\mathbf{x}, t)$ and $\Delta c_{k+1} := c_{k+1}(\mathbf{x}, t) - c_k(\mathbf{x}, t)$. Correspondingly, the expression for $\mathcal{L}_{\text{CL}}[c_k]$ is given as

$$\mathcal{L}_{\text{CL}}[c_k] := \frac{\partial c_k}{\partial t} - \text{div} [\mathbf{D}(\mathbf{x}) \text{grad}[c_k]] = f(\mathbf{x}, t, c_k). \quad (6.3.25)$$

Following a similar type of procedure given by equations (6.3.22) and (6.3.23), the final expression for the linearized operator $\mathcal{L}_{\text{CL}}[c_{k+1}]$ on which the finite element method is applied is given as

$$\mathcal{L}_{\text{CL}}[c_{k+1}] := \frac{\partial c_{k+1}}{\partial t} + \mu_{\text{CL}} c_{k+1} - \text{div} [\mathbf{D}(\mathbf{x}) \text{grad}[c_{k+1}]] = f_{\text{CL}}(\mathbf{x}, t, c_k), \quad (6.3.26)$$

where there terms μ_{CL} and f_{CL} are given as:

$$\mu_{\text{CL}} = - \left. \frac{df(\mathbf{x}, t, c_k + \epsilon \Delta c_{k+1})}{d(c_k + \epsilon \Delta c_{k+1})} \right|_{\epsilon=0} \quad \text{and} \quad (6.3.27a)$$

$$f_{\text{CL}} = f(\mathbf{x}, t, c_k) - \left. \frac{df(\mathbf{x}, t, c_k + \epsilon \Delta c_{k+1})}{d(c_k + \epsilon \Delta c_{k+1})} \right|_{\epsilon=0} c_k. \quad (6.3.27b)$$

In case of steady-state, the above PDE, which is linearized about the base state ‘ c_k ’ satisfies a continuous maximum principle if $\mu_{\text{CL}} \geq 0$. This condition cannot be relaxed. However, for transient case, $\mu_{\text{CL}} \leq 0$ is possible but there will be restriction on time-step to satisfy a maximum principle.

6.3.6 General comments on various nonlinear methods

On careful mathematical analysis of Pao's method, Picard's method, consistent linearization method, traditional Newton-Raphson method, and modification of traditional Newton-Raphson method, the following are some of the important aspects one needs to extra vigilant in order to enforce various discrete properties.

1. One of the advantages of traditional Newton-Raphson method is the rapid rate of convergence as compared to the Pao's method. But it should noted that this scheme suffers from a major drawback, which is the (potential) dissatisfaction of monotone property, discrete comparison principles, discrete maximum principles, and non-negative constraint. This is because the scheme given by equation (6.3.20) does not apriorly (or a posteriorly) ensure satisfaction of these discrete properties.
2. Furthermore, the scheme given by equation 6.3.20 does not provide flexibility to construct a non-negative formulation based on the lines given in Reference Nagarajan and Nakshatrala (2011); Nakshatrala et al. (2013) to satisfy atleast certain discrete properties (such as non-negative constraint and discrete maximum principles). Nevertheless, one can derive conditions on the discrete counter parts of the expressions (6.3.21) and (6.3.22) similar to that of Theorem (6.3.5). But it should be noted that these conditions do not have any variational basis. Additionally, enforcing such mathematical conditions in a numerical simulation is very difficult. For more details, see the discussion of numerical results presented in the subsection 6.5.1.

6.4 PHYSICS-BASED CHEMICAL REACTION MODELS OF MONOTONE-TYPE

As noted previously in Section 6.1, the reaction-diffusion system given by equations (6.2.1a)–(6.2.1c) covers a wide number of real-life applications in various branches of physical, chemical, and biological sciences. Herein, we shall describe some popular and specific chemical reaction models of monotone-type, i.e., the reaction volumetric source term $f(\mathbf{x}, t, c(\mathbf{x}, t))$ satisfies the one-sided Lipschitz condition given by equation 6.3.2 or equation (6.3.16). The reaction volumetric source functionals that we discuss below are infinitely differentiable with respect to $c(\mathbf{x}, t) \geq 0$ and satisfy the condition $f(\mathbf{x}, t, c(\mathbf{x}, t) = 0) \geq 0$. Additionally, we assume that boundary and initial data are non-negative. Hence, from Theorem 6.3.3 each of these chemical reaction models has a unique non-negative time-dependent solution, if there exists a non-negative upper solution. Similar inference can be drawn from Theorem 6.3.4 for steady-state reaction-diffusion problems. However, it should be noted that *more than one* non-negative steady-state solution may exist. For uniqueness of solution, we need to have $\frac{\partial f}{\partial c} \leq 0$ in $\langle 0, c_{\text{us}} \rangle$ (see Theorem 6.3.4).

The parameter β_{gLip} or β_{lLip} can be evaluated by using the famous Taylor's theorem Rudin (1976); Lax (2002) as $f(\mathbf{x}, t, c(\mathbf{x}, t))$ is C^∞ -functional with respect to $c(\mathbf{x}, t)$. Correspondingly, the Taylor's theorem for $f(\mathbf{x}, t, c_i)$ at c_j for $c_{\text{ls}} \leq c_j \leq c_i \leq c_{\text{us}}$ is given as

$$f(\mathbf{x}, t, c_i) = f(\mathbf{x}, t, c_j) + \frac{\partial f}{\partial c} \Big|_{c=c_j} (c_i - c_j) + \frac{\partial^2 f}{\partial c^2} \Big|_{c=c_\kappa} (c_i - c_j)^2, \quad (6.4.1)$$

where c_κ lies between c_j and c_i . In the conditions given by equation 6.3.2 and equation (6.3.16), the constants β_{gLip} and β_{lLip} are not necessarily positive. From equation (6.4.1) and the fact that $f(\mathbf{x}, t, c(\mathbf{x}, t))$ is continuously differentiable in $c(\mathbf{x}, t)$ for

$c(\mathbf{x}, t) \in \langle c_{\text{ls}}, c_{\text{us}} \rangle$, these constants are given as

$$\beta_{gLip} = \max_{\substack{(\mathbf{x}, t) \in \bar{\Omega}_T \\ c_{\text{ls}} \leq c \leq c_{\text{us}}}} \left[-\frac{\partial f}{\partial c} \right] \text{ and} \quad (6.4.2a)$$

$$\beta_{lLip} = \max_{\substack{(\mathbf{x}, t) \in \bar{\Omega}_T \\ c_{\text{ls}} \leq c \leq c_{\text{us}}}} \left[\frac{\partial f}{\partial c} \right]. \quad (6.4.2b)$$

The values of these parameters β_{gLip} or β_{lLip} shall be used in our numerical simulations to obtain the required non-negative solutions. Additionally, we check if other discrete properties are preserved or not during this nonlinear iterative process. After the discussion of each popular chemical reaction model, we shall provide the values for the parameters β_{gLip} or β_{lLip} . Furthermore, we shall construct the ordered solutions $c_{\text{ls}}(\mathbf{x}, t)$ and $c_{\text{us}}(\mathbf{x}, t)$. Finally, we will briefly discuss about the existence and uniqueness of the non-negative solutions for these models for both transient and steady-state reaction-diffusion problems.

6.4.1 Enzyme kinetics (Michaelis-Menton type) chemical reaction model

Consider the following monoenzymatic irreversible class of enzyme-kinetics reaction schemes occurring in a biochemical system Murray (1968b,a):



where A is the free enzyme, B is the substrate, AB is the enzyme-substrate complex, and C is the reaction product. The constants k_0 , k_1 , and k_2 represent the rates of reaction. Based on Michaelis-Menton hypothesis and Briggs-Haldane approximation Murray (1968b,a), the reaction volumetric source $f(\mathbf{x}, t, c(\mathbf{x}, t))$ corresponding to the substrate B is given as

$$f(\mathbf{x}, t, c(\mathbf{x}, t)) = -\frac{\sigma_{ek_1}c}{1 + \sigma_{ek_2}c}, \quad (6.4.4)$$

where the positive constants σ_{ek_1} and σ_{ek_2} are related to reaction rates k_0 , k_1 , and k_2 as

$$\sigma_{ek_1} = \frac{\sigma_{A_o} k_0 k_2}{k_1 + k_2} \text{ and} \quad (6.4.5a)$$

$$\sigma_{ek_2} = \frac{k_0}{k_1 + k_2} \quad (6.4.5b)$$

with the positive constant σ_{A_o} corresponding to the total amount of enzyme A present in the reacting domain. From equation (6.4.4), it is evident that $f(\mathbf{x}, t, c(\mathbf{x}, t) = 0) = 0$ and $f(\mathbf{x}, t, c(\mathbf{x}, t)) \leq 0$ for any $c(\mathbf{x}, t) \geq 0$. Hence, any positive constant ρ satisfying the condition given by equation (6.3.7) is an upper solution. By Theorem (6.3.3), we have a unique solution $c(\mathbf{x}, t)$ for the governing equations (6.2.1a)–(6.2.1c) with the reaction volumetric source function given by equation (6.4.4). Herein, for this $f(\mathbf{x}, t, c(\mathbf{x}, t))$, the following inferences can be drawn based on its first and second partial derivatives:

$$\frac{\partial f}{\partial c} = -\frac{\sigma_{ek_1}}{(1 + \sigma_{ek_2}c)^2} \geq -\sigma_{ek_1} \quad \forall c(\mathbf{x}, t) \geq 0 \text{ and} \quad (6.4.6a)$$

$$\frac{\partial^2 f}{\partial c^2} = \frac{2\sigma_{ek_1}\sigma_{ek_2}}{(1 + \sigma_{ek_2}c)^3} > 0 \quad \forall c(\mathbf{x}, t) \geq 0. \quad (6.4.6b)$$

From equations (6.4.1), (6.4.6a), and (6.4.6b), it is evident that $f(\mathbf{x}, t, c_i) - f(\mathbf{x}, t, c_j) \geq -\sigma_{ek_1}(c_i - c_j)$. Furthermore, from equations (6.4.2a) and (6.4.6a), we have $\beta_{gLip} = \sigma_{ek_1}$. In case of steady state problem, any constant $\rho \geq c_p(\mathbf{x})$ can be taken as an upper solution and the value of β_{gLip} is same as that of transient state. Since $\frac{\partial f}{\partial c} \leq 0$ for any $c(\mathbf{x}) \geq 0$, from Theorem 6.3.4 it is apparent that we have a unique steady state solution. Additionally, as $f(\mathbf{x}, c(\mathbf{x}) = \rho) \leq 0$, we have $0 \leq c(\mathbf{x}) \leq \rho$.

In case, a competitive inhibitor, the so-called substrate inhibition, is present a

different reaction rate leads to the following reaction model:

$$f(\mathbf{x}, t, c(\mathbf{x}, t)) = -\frac{\sigma_{sk_1} c}{1 + \sigma_{sk_2} c + \sigma_{sk_3} c^2}, \quad (6.4.7)$$

where σ_{sk_1} , σ_{sk_2} , and σ_{sk_3} are positive constants. Similar to the volumetric source functional given by equation (6.4.4), $f(\mathbf{x}, t, c(\mathbf{x}, t) = 0) = 0$ and $f(\mathbf{x}, t, c(\mathbf{x}, t)) \leq 0$ for any $c(\mathbf{x}, t) \geq 0$ for equation (6.4.7). Therefore, any constant $\rho \geq c_p(\mathbf{x}, t)$ and $\rho \geq c_0(\mathbf{x})$ is an upper solution for the time-dependent reaction-diffusion problem. By Theorem (6.3.3), we have a unique solution $c(\mathbf{x}, t)$ for the transient case and $0 \leq c(\mathbf{x}, t) \leq \rho$. Correspondingly, the first and second partial derivatives of $f(\mathbf{x}, t, c(\mathbf{x}, t))$ are given as

$$\frac{\partial f}{\partial c} = -\frac{\sigma_{sk_1} - \sigma_{sk_1} \sigma_{sk_3} c^2}{(1 + \sigma_{sk_2} c + \sigma_{sk_3} c^2)^2} \text{ and} \quad (6.4.8a)$$

$$\frac{\partial^2 f}{\partial c^2} = \frac{3\sigma_{sk_1} + 2\sigma_{sk_1} \sigma_{sk_3} c + (2\sigma_{sk_1} \sigma_{sk_2} \sigma_{sk_3} - 3\sigma_{sk_1} \sigma_{sk_3}) c^2 + 2\sigma_{sk_1} \sigma_{sk_2} \sigma_{sk_3} c^3}{(1 + \sigma_{sk_2} c + \sigma_{sk_3} c^2)^3}. \quad (6.4.8b)$$

From equation (6.4.8a) and mean value theorem in Mathematical analysis Rudin (1976), it is evident that $f(\mathbf{x}, t, c(\mathbf{x}, t))$ is a Lipschitz function as $\frac{\partial f}{\partial c}$ is uniformly bounded by a positive constant ' $\sigma_{sk_1} \beta_{sk}$ ' in \mathbb{R} . This positive constant β_{sk} for the reaction model given by equation (6.4.7) is

$$\beta_{sk} = \sup_{0 \leq c \leq \rho} \left[\left| \frac{1 - \sigma_{sk_3} c^2}{(1 + \sigma_{sk_2} c + \sigma_{sk_3} c^2)^2} \right| \right], \quad (6.4.9)$$

where $\sup[\bullet]$ is the standard supremum in mathematical analysis Rudin (1976). Since, $\frac{\partial f}{\partial c}$ is continuously differentiable in $\langle 0, \rho \rangle$, supremum can be replaced by the maximum. Moreover, $|1 - \sigma_{sk_3} c^2| \leq (1 + \sigma_{sk_2} c + \sigma_{sk_3} c^2)^2$ for all $0 \leq c(\mathbf{x}, t) \leq \rho$. Hence, the positive constant $\beta_{sk} \leq 1$. Conservatively, the Lipschitz constant for $f(\mathbf{x}, t, c(\mathbf{x}, t))$ given by equation (6.4.7) can be taken as $\beta_{Lip} = \sigma_{sk_1}$.

For the steady-state reaction diffusion problem, $\rho \geq c_p(\mathbf{x})$ can be taken as an upper solution. By Theorem 6.3.4, atleast one steady-state solution $c(\mathbf{x})$ exists and $0 \leq c(\mathbf{x}) \leq \rho$. Nevertheless, for the current volumetric source functional, we have the following:

$$\frac{\partial f}{\partial c} \leq 0 \quad \forall c \leq \frac{1}{\sqrt{\sigma_{sk_3}}} \text{ and} \quad (6.4.10a)$$

$$\frac{\partial f}{\partial c} \geq 0 \quad \forall c \geq \frac{1}{\sqrt{\sigma_{sk_3}}}. \quad (6.4.10b)$$

As $\frac{\partial f}{\partial c} \not\leq 0$ for all $c(\mathbf{x}) \in \langle 0, \rho \rangle$, uniqueness cannot be concluded without additional restrictions on $c_p(\mathbf{x})$ and σ_{sk_3} . If $c_p(\mathbf{x}) \leq \frac{1}{\sqrt{\sigma_{sk_3}}}$, then from Theorem 6.3.4 it is evident that the reaction volumetric source functional given by equation (6.4.7) cannot have multiple steady-state solutions. However, if this condition is not met, multiple solutions do exist for certain range of input data and model parameters. Further relevant mathematical details and general discussion on existence of multiple solutions, see Pao (Pao, 1993, Chapter-3). Additionally, within the context of 1D steady-state reaction-diffusion problems, numerical investigation has been performed by Pao et.al. Pao et al. (1985) and Kernevez (Kernevez, 1980, Chapter-3) for finding multiple steady-state solutions. But it should noted that seldom research effort has been performed for 2D and 3D problems.

6.4.2 Population growth (Fisher type) and genetics model

In understanding birth and death rates of genotypes (which correspond to the population genetics of diploid individuals), fisher type models are employed to understand the dynamics of the biological system. The following is a specific and popular model used to study such a genetic process Fife (1979) for normalized concentration

$c(\mathbf{x}, t)$ governed by equations (6.2.1a)–(6.2.1c):

$$f(\mathbf{x}, t, c(\mathbf{x}, t)) = \sigma_{pg_1} c(c - \sigma_{pg_2})(1 - c), \quad (6.4.11)$$

where σ_{pg_1} and σ_{pg_2} are positive constants. Furthermore, we have $0 < \sigma_{pg_2} < 1$. Mathematically, the chemical reaction model given by equation (6.4.11) also describes bistable transmission lines in electric circuit theory Nagumo et al. (1965). From equation (6.4.11), we have $f(\mathbf{x}, t, c(\mathbf{x}, t) = 0) = 0$ and $f(\mathbf{x}, t, c(\mathbf{x}, t)) \geq 0$ for any $c(\mathbf{x}, t) \geq 0$. As $c(\mathbf{x}, t)$ is normalized, we require $0 \leq c_p(\mathbf{x}, t) \leq 1$ and $0 \leq c_0(\mathbf{x}) \leq 1$. Under this requirement, it is evident that $\rho = 1$ is an upper solution. By Theorem 6.3.3, we have a unique time-dependent solution and $0 \leq c(\mathbf{x}, t) \leq 1$. Accordingly, the first, second, and third partial derivatives of $f(\mathbf{x}, t, c(\mathbf{x}, t))$ are given as

$$\frac{\partial f}{\partial c} = -\sigma_{pg_1} (3c^2 - 2(1 + \sigma_{pg_2})c + \sigma_{pg_2}), \quad (6.4.12a)$$

$$\frac{\partial^2 f}{\partial c^2} = 2\sigma_{pg_1} ((1 + \sigma_{pg_2}) - 3c), \text{ and} \quad (6.4.12b)$$

$$\frac{\partial^3 f}{\partial c^3} = -6\sigma_{pg_1}. \quad (6.4.12c)$$

From equations (6.4.12a)–(6.4.12c), it is apparent that $f(\mathbf{x}, t, c(\mathbf{x}, t))$ is a Lipschitz function as $\frac{\partial f}{\partial c}$ is uniformly bounded by a positive constant ‘ $\sigma_{pg_1} \beta_{pg}$ ’ in \mathbb{R} , where β_{pg} is given as

$$\beta_{pg} = \sup_{0 \leq c \leq 1} \left[\left| 3c^2 - 2(1 + \sigma_{pg_2})c + \sigma_{pg_2} \right| \right]. \quad (6.4.13)$$

As $\frac{\partial f}{\partial c}$ is continuously differentiable in $\langle 0, 1 \rangle$, the maximum value of β_{pg} occurs at either $c = 0$ or $c = \frac{1 + \sigma_{pg_2}}{3}$ or $c = 1$ depending on the value of parameter σ_{pg_2} . It should be noted that $c = \frac{1 + \sigma_{pg_2}}{3}$ is a local maxima but not a global maxima of β_{pg} .

Correspondingly, the value for β_{pg} is given as

$$\beta_{pg} = \max \left[\left\{ \sigma_{pg2}, (1 - \sigma_{pg2}) \right\} \right] < 1. \quad (6.4.14)$$

The Lipschitz constant for $f(\mathbf{x}, t, c(\mathbf{x}, t))$ given by equation (6.4.11) can be taken as $\beta_{Lip} = \sigma_{pg1}\beta_{pg}$. In case of steady-state problem, from Theorem 6.3.4 it is apparent that there exists atleast one steady-state solution such that $0 \leq c(\mathbf{x}) \leq 1$. Moreover, we have:

$$\frac{\partial f}{\partial c} \leq 0 \quad \forall c \in \left[\frac{1 + \sigma_{pg2} - \sqrt{1 - \sigma_{pg2} + \sigma_{pg2}^2}}{3}, \frac{1 + \sigma_{pg2} + \sqrt{1 - \sigma_{pg2} + \sigma_{pg2}^2}}{3} \right] \text{ and} \quad (6.4.15a)$$

$$\frac{\partial f}{\partial c} \geq 0 \quad \forall c \in \left[0, \frac{1 + \sigma_{pg2} - \sqrt{1 - \sigma_{pg2} + \sigma_{pg2}^2}}{3} \right] \cup \left[\frac{1 + \sigma_{pg2} + \sqrt{1 - \sigma_{pg2} + \sigma_{pg2}^2}}{3}, 1 \right]. \quad (6.4.15b)$$

Hence, there is no uniqueness result for $c(\mathbf{x}) \in \langle 0, 1 \rangle$. Indeed, for a certain set of parameters σ_{pg1} and σ_{pg2} , Pao (Pao, 1993, Chapter-3) has developed mathematical results and existence theorems to show that Fisher's model possesses multiple positive steady-state solutions. But it should be noted that seldom parametric studies have been done for this popular model, even within the context 1D steady-state reaction-diffusion problem.

6.4.3 Nuclear reactor dynamics model

In the analysis of space-time dependent nuclear reactor dynamics, the chemical reaction model for the number of neutrons produced per fission is given by the following functional Kastenber and Chambré (1968); Duderstadt and Hamilton (1976); Henry (1975):

$$f(\mathbf{x}, t, c(\mathbf{x}, t)) = c(\sigma_{nr1} - \sigma_{nr2}c) + q(\mathbf{x}, t), \quad (6.4.16)$$

where σ_{nr_1} and σ_{nr_2} are positive constants. $q(\mathbf{x}, t) \geq 0$ constitutes the non-reactive spatially varying volumetric source. From equation (6.4.16), it is evident that $f(\mathbf{x}, t, c(\mathbf{x}, t) = 0) = q(\mathbf{x}, t) \geq 0$. Hence, Theorem 6.3.3, implies that we have unique time-dependent non-negative solution. Furthermore, from equations (6.4.16) and (6.3.7) the positive constant ρ , which is an upper solution can be evaluated as

$$\rho(\sigma_{nr_1} - \sigma_{nr_2}\rho) + q \leq 0, \quad c_p(\mathbf{x}, t) \leq \rho, \quad c_0(\mathbf{x}) \leq \rho. \quad (6.4.17)$$

Conservatively, based on equation (6.4.17) a concise expression for calculating ρ is given as

$$\rho = \max \left[\left\{ 1, \frac{\sigma_{nr_1} + q_{nr}}{\sigma_{nr_2}}, \sup_{(\mathbf{x}, t) \in \bar{\Omega}_I} [c_p(\mathbf{x}, t)], \sup_{\mathbf{x} \in \bar{\Omega}} [c_0(\mathbf{x})] \right\} \right] \geq 1, \quad (6.4.18)$$

where the non-negative constant q_{nr} is given as

$$q_{nr} = \sup_{(\mathbf{x}, t) \in \bar{\Omega}_I} [q(\mathbf{x}, t)]. \quad (6.4.19)$$

Correspondingly, the first and second partial derivatives of $f(\mathbf{x}, t, c(\mathbf{x}, t))$ are given as

$$\frac{\partial f}{\partial c} = \sigma_{nr_1} - 2\sigma_{nr_2}c \text{ and} \quad (6.4.20a)$$

$$\frac{\partial^2 f}{\partial c^2} = -2\sigma_{nr_2}. \quad (6.4.20b)$$

From equation (6.4.20a), it apparent that $f(\mathbf{x}, t, c(\mathbf{x}, t))$ is a Lipschitz function as it is bounded above by a positive constant ' $\beta_{Lip} = \sigma_{nr_1}\beta_{nr}$ ', where β_{nr} is given as

$$\beta_{nr} = \sup_{0 \leq c \leq \rho} \left[\left| 1 - \frac{2\sigma_{nr_2}c}{\sigma_{nr_1}} \right| \right] = \max \left[\left\{ 1, \left| \frac{2\sigma_{nr_2}\rho}{\sigma_{nr_1}} - 1 \right| \right\} \right]. \quad (6.4.21)$$

In case of steady-state problem, $\frac{\partial f}{\partial c} \not\leq 0$ for all $c(\mathbf{x}) \in \langle 0, \rho \rangle$ without additional restrictions on σ_{nr_1} and σ_{nr_2} . Hence, in general, steady-state solution is not unique based on Theorem 6.3.4.

6.4.4 Chemical reactor dynamics model

In an irreversible isothermal chemical reaction where the temperature is a known constant σ_{temp} , a general expression for the reaction volumetric source is given as follows Aris (1975b,a):

$$f(\mathbf{x}, t, c(\mathbf{x}, t)) = -\sigma_{cr}c^m, \quad (6.4.22)$$

where m is the order of the reaction, $\sigma_{cr} = \sigma_{tn} \exp(\sigma_{an} - \frac{\sigma_{an}}{\sigma_{temp}})$ is the rate constant, σ_{tn} is the Thiele number for the chemical reaction, σ_{an} is the Arrhenius number corresponding to the kinetics of the m -th order reaction. Specifically, when $m = 2$ the equation given by (6.4.22) describes the second-order recombination reaction of free atoms or ions in the dissociation or ionization processes of subsonic motion of gases in chemical reactors and astrophysics applications Joseph and Lundgren (1973). From equation (6.4.22), it is evident that $f(\mathbf{x}, t, c(\mathbf{x}, t) = 0) = 0$ and $f(\mathbf{x}, t, c(\mathbf{x}, t)) \leq 0$ for any $c(\mathbf{x}, t) \geq 0$. Hence, any positive constant ρ satisfying the condition provided by equation (6.3.7) is an upper solution. By Theorem (6.3.3), we have a unique time-dependent solution. In some parallel reaction schemes involving multiple chemical species, a general reaction volumetric source is given as follows Erdi and Toth (1989); Fogler (2006); Rice (1985):

$$f(\mathbf{x}, t, c(\mathbf{x}, t)) = -\sum_{i=1}^n \sigma_{cr_i} c^{m_i}, \quad (6.4.23)$$

where n is the number of chemical species involved in the parallel reaction schemes, m_i is the order of i -th chemical reaction, and σ_{cr_i} is the corresponding rate constant for

the i -th chemical reaction. It should be noted that m_i and σ_{cr_i} are positive numbers.

6.4.5 Moderate/slow bimolecular chemical reaction model

Bimolecular chemical reaction involves two chemical species A and B , which react irreversibly to produce product C based on the following stoichiometry Nakshatrala et al. (2013):



where n_A , n_B and n_C are positive stoichiometric coefficients. Correspondingly, the governing equations for the bimolecular reaction-diffusion system is given as follows:

$$\frac{\partial c_A}{\partial t} - \text{div}[\mathbf{D}(\mathbf{x}) \text{grad}[c_A]] = q_A(\mathbf{x}, t) - n_A r \quad \text{in } \Omega \times (0, \mathcal{I}), \quad (6.4.25a)$$

$$\frac{\partial c_B}{\partial t} - \text{div}[\mathbf{D}(\mathbf{x}) \text{grad}[c_B]] = q_B(\mathbf{x}, t) - n_B r \quad \text{in } \Omega \times (0, \mathcal{I}), \quad (6.4.25b)$$

$$\frac{\partial c_C}{\partial t} - \text{div}[\mathbf{D}(\mathbf{x}) \text{grad}[c_C]] = q_C(\mathbf{x}, t) + n_C r \quad \text{in } \Omega \times (0, \mathcal{I}), \quad (6.4.25c)$$

$$c_i(\mathbf{x}, t) = c_i^p(\mathbf{x}, t) \quad \text{on } \partial\Omega \times (0, \mathcal{I}) \quad (i = A, B, C), \quad \text{and} \quad (6.4.25d)$$

$$c_i(\mathbf{x}, t = 0) = c_i^0(\mathbf{x}) \quad \text{in } \Omega \quad (i = A, B, C), \quad (6.4.25e)$$

where $q_i(\mathbf{x}, t)$ constitutes the non-reactive spatially varying volumetric source, $c_i^p(\mathbf{x}, t)$ is the Dirichlet boundary condition, and $c_i^0(\mathbf{x})$ is the initial condition for i -th chemical species. The rate of the bimolecular chemical reaction ‘ r ’ is given by $r = k_{bio}c_Ac_B$, where k_{bio} is the bilinear reaction rate coefficient. This type of chemical reaction model for ‘ r ’ is used to study a kinetic reaction between the electron donor and acceptor in reactive transport problems McCarty and Criddle (2012).

Based on the following linear transformation, one can obtain two independent

non-negative invariants, which are unaffected by the chemical reaction:

$$c_F := c_A + \left(\frac{n_A}{n_C}\right) c_C \text{ and} \quad (6.4.26a)$$

$$c_G := c_B + \left(\frac{n_B}{n_C}\right) c_C. \quad (6.4.26b)$$

Correspondingly, using the non-negative invariant set given by equations (6.4.26a)–(6.4.26b), the governing equations for bimolecular system (6.4.25a)–(6.4.25e) can be converted to a set of uncoupled linear diffusion equations. The governing equations for the invariant F can be written as follows:

$$\frac{\partial c_F}{\partial t} - \text{div}[\mathbf{D}(\mathbf{x}) \text{grad}[c_F]] = q_A(\mathbf{x}, t) + \left(\frac{n_A}{n_C}\right) q_C(\mathbf{x}, t) \quad \text{in } \Omega \times (0, \mathcal{I}), \quad (6.4.27a)$$

$$c_F(\mathbf{x}, t) = c_F^p(\mathbf{x}, t) := c_A^p(\mathbf{x}, t) + \left(\frac{n_A}{n_C}\right) c_C^p(\mathbf{x}, t) \quad \text{on } \partial\Omega \times (0, \mathcal{I}), \text{ and} \quad (6.4.27b)$$

$$c_F(\mathbf{x}, t = 0) = c_F^0(\mathbf{x}) := c_A^0(\mathbf{x}) + \left(\frac{n_A}{n_C}\right) c_C^0(\mathbf{x}) \quad \text{in } \Omega. \quad (6.4.27c)$$

In a similar fashion, the governing equations for the invariant G are given as follows:

$$\frac{\partial c_G}{\partial t} - \text{div}[\mathbf{D}(\mathbf{x}) \text{grad}[c_G]] = q_B(\mathbf{x}, t) + \left(\frac{n_B}{n_C}\right) q_C(\mathbf{x}, t) \quad \text{in } \Omega \times (0, \mathcal{I}), \quad (6.4.28a)$$

$$c_G(\mathbf{x}, t) = c_G^p(\mathbf{x}, t) := c_B^p(\mathbf{x}, t) + \left(\frac{n_B}{n_C}\right) c_C^p(\mathbf{x}, t) \quad \text{on } \partial\Omega \times (0, \mathcal{I}), \text{ and} \quad (6.4.28b)$$

$$c_G(\mathbf{x}, t = 0) = c_G^0(\mathbf{x}) := c_B^0(\mathbf{x}) + \left(\frac{n_B}{n_C}\right) c_C^0(\mathbf{x}) \quad \text{in } \Omega. \quad (6.4.28c)$$

Once c_F and c_G are known, then the concentration of product C is obtained based on the governing equations given by (6.4.25c)–(6.4.25e), non-negative invariant set given by equations (6.4.26a)–(6.4.26b), and rate of bimolecular chemical reaction

' $r = k_{bio}c_Ac_B$ '. The final form of these equations are given as

$$\frac{\partial c_C}{\partial t} - \text{div}[\mathbf{D}(\mathbf{x}) \text{grad}[c_C]] = f_C(\mathbf{x}, t, c_C) \quad \text{in } \Omega \times (0, \mathcal{I}), \quad (6.4.29a)$$

$$c_C(\mathbf{x}, t) = c_C^p(\mathbf{x}, t) \quad \text{on } \partial\Omega \times (0, \mathcal{I}), \text{ and} \quad (6.4.29b)$$

$$c_C(\mathbf{x}, t = 0) = c_C^0(\mathbf{x}) \quad \text{in } \Omega, \quad (6.4.29c)$$

where the reaction volumetric source term $f_C(\mathbf{x}, t, c_C)$ for the product C is given as

$$f_C(\mathbf{x}, t, c_C) = q_C(\mathbf{x}, t) + n_C k_{bio} \left(c_F - \left(\frac{n_A}{n_C} \right) c_C \right) \left(c_G - \left(\frac{n_B}{n_C} \right) c_C \right). \quad (6.4.30)$$

As $c_A(\mathbf{x}, t)$ and $c_B(\mathbf{x}, t)$ are non-negative, this imposes the following constraint on $c_C(\mathbf{x}, t)$ when seeking a solution to the equations (6.4.29a)–(6.4.29c):

$$0 \leq c_C(\mathbf{x}, t) \leq \min_{(\mathbf{x}, t) \in \Omega_{\mathcal{I}}} \left[\left(\frac{n_C}{n_A} \right) c_F(\mathbf{x}, t), \left(\frac{n_C}{n_B} \right) c_G(\mathbf{x}, t) \right]. \quad (6.4.31)$$

After solving for $c_C(\mathbf{x}, t)$, using the non-negative invariant set given by equations (6.4.26a)–(6.4.26b) one can obtain $c_A(\mathbf{x}, t)$ and $c_B(\mathbf{x}, t)$.

It should be noted that similar type of mathematical analysis for more complex chemical reaction models involving speciation McCarty and Criddle (2012), equilibrium and non-equilibrium sorption in solute transport Zheng and Bennett (2002); Bahr and Rubin (1987); Valocchi (1989), and double-Monod/Michaelis-Menten enzyme kinetics of biomass of living organisms in contaminant transport Cirpka and Valocchi (2007); Baveye and Valocchi (1989) can be accomplished. This will be considered in our future works.

6.5 REPRESENTATIVE NUMERICAL EXAMPLES

In this section, we shall illustrate the performance of the proposed computational framework for reaction-diffusion equations given by (6.2.1a)–(6.2.1c). Herein, within

the context of standard single-field Galerkin formulation and non-negative methodology, we shall study in-detail a specific chemical reaction model which are discussed in Section 6.4. Analyzing all these models in general setting (which includes advection and parametric studies) is beyond the scope of this chapter and shall be dealt in our future works.

6.5.1 Autocatalytic reaction-diffusion problem

A pictorial description of the boundary value problem is shown in Figure 6.3(a). Vertex O is located at $(0, 0)$. The domain under consideration is a bi-unit square. The right figure shows a typical three-node triangular structured coarse mesh employed in the computational study. This mesh has 21 nodes along each side of the domain. Corresponding, Figure 6.3(b) shows the mesh employed in the computational study. The value for the parameter m in the reaction volumetric source term is taken to be equal to 2. Stopping criterion (which is the tolerance on numerical value of the residual) for Pao's method is taken as 10^{-10} and that of Newton-Raphson methods is taken as 10^{-12} . The initial guess for the concentration vector is taken as $\mathbf{c}_1 = \mathbf{0}$. It should be noted that the vector $\mathbf{c}_1 = \mathbf{0}$ corresponds to \mathbf{c}_{1s} . Numerical simulations are performed both on coarse and h -refined meshes (which are obtained in a hierarchical fashion using the coarse mesh). In order to illustrate the pros and cons of Pao and Newton-Raphson methods in satisfying different discrete properties, we shall consider four different scenarios. In each scenario, the values for diffusivity tensor, boundary conditions, and initial conditions are changed accordingly to illustrate various numerical aspects in satisfying discrete properties. These cases and their significance are described in detail as follows:

Case # 1 $c_0(\mathbf{x}, t) = 0$, $c_p(\mathbf{x}, t) = 1$, and $\mathbf{D}(\mathbf{x}) = 1$. Both steady-state and transient analysis are performed for this scenario. The number of iterations required for convergence (which is related to the computational cost), minimum value

corresponding to the change in the concentration vectors from two successive iteration levels (which is connected to the monotone property), euclidean norm change in the concentration vectors between successive iterations, and correspondingly the residual values for each iteration in case of Pao's method are tabulated in Table 6.2. Figure 6.4 shows the contours of concentration profiles at various iteration levels for steady-state reaction-diffusion problem on coarse mesh. For any given value of 'k' until convergence, if each component of the vector ' $\mathbf{c}_k - \mathbf{c}_{k-1}$ ' is either non-negative or close to machine precision $\epsilon_{\text{mach}} \approx 2.22 \times 10^{-16}$ then we are ensured that the employed numerical scheme to obtain \mathbf{c}_k satisfies the monotone property. As the numerical values in the second column of Table 6.2 ($\min[\mathbf{c}_k - \mathbf{c}_{k-1}]$) are non-negative, it is evident that Pao's method satisfies the monotone property as described in Theorem 6.3.5 in addition to other discrete principles such as non-negative constraint, discrete maximum principle, and discrete comparison principle. This is due to the fact that the coarse mesh employed in the numerical simulations is non-obtuse when measured in Euclidian metric. But this is not the case for transient reaction-diffusion problems. This is because we have decay term in addition to diffusion term due to MOHL approach. Because of this decay term the mesh is neither acute nor non-obtuse. However, this mesh is a Delaunay-mesh when measured in Euclidian metric for the transient-reaction-diffusion problem. Hence, in case of Delaunay-based meshes it is well-known that we need a highly refined mesh to satisfy various discrete properties. This is apparent from Figure 6.5 and Figure 6.6. The white region within the bi-unit square domain represents the area in which the *numerical value* of concentration is *negative*. Even though the numerical scheme converges this negative value for concentration does not decrease and is quite high as quantified in Table 6.3 at various time and iteration levels. These values are obtained at time levels $t = 10^{-4}$ and $t = 10^{-3}$

using the $\Delta t = 10^{-4}$. The minimum concentration at various iteration levels for $t = 10^{-4}$ is around -3.65×10^{-1} and correspondingly for $t = 10^{-3}$ is around -1.54×10^{-2} . Additionally, from this table it clear that Pao's method also violates other discrete principles if the mesh is not fine enough. For both steady-state and transient analysis, it is numerically observed that the terminal convergence rate for Pao's method is of first-order.

Case # 2 $c_0(\mathbf{x}, t) = 1$, $c_p(\mathbf{x}, t) = 0$, and $\mathbf{D}(\mathbf{x}) = 1$. Herein, we perform only transient analysis. As for the steady-state, the value of concentration is zero everywhere in the domain. From the continuous maximum principle, as $f(\mathbf{x}, t) \leq 0$, it is evident that $c(\mathbf{x})$ should lie in-between 0 and 1. From Figures 6.7 and 6.8 it is evident that Discrete Pao's method satisfies this condition on h -refined meshes. Furthermore, from Table 6.4 it is evident that the violation in the maximum value for concentration is quite high on coarse meshes. From the numerical simulations conducted, it is also observed that Discrete Pao's method satisfies the monotone property on h -refined meshes. The reason for such behaviour is described in the previous case study (see **Case # 1**).

Case # 3 $c_0(\mathbf{x}, t) = 0$, $c_p(\mathbf{x}, t) = 1$, and $\mathbf{D}(\mathbf{x}) = \begin{pmatrix} (\epsilon_1+y)^2+\epsilon_2(\epsilon_1+x)^2 & -(1-\epsilon_2)(\epsilon_1+x)(\epsilon_1+y) \\ -(1-\epsilon_2)(\epsilon_1+x)(\epsilon_1+y) & (\epsilon_1+x)^2+\epsilon_2(\epsilon_1+y)^2 \end{pmatrix}$ where $\epsilon_1 = 10^{-3}$ and $\epsilon_2 = 10^{-3}$ Potier (2005). The objective of this case is to check whether the coarse mesh or the h -refined mesh shown in the Figure 6.3(b) satisfies various discrete properties for steady-state and transient analysis. From Figure 6.9 and Table 6.6 it is apparent that this mesh violates the non-negative constraint and does not satisfy other discrete principles for transient analysis. However, this is not the case for steady-state problem. From Figure 6.9 and Table 6.5, it is evident that this h -refined mesh satisfies the non-negative constraint and discrete maximum principle but violates the monotone property and discrete comparison principle. This is because the condition $\mathbf{K}_{ff}^{-1} \succeq \mathbf{O}$ in

equation (6.3.9), which is necessary to satisfy non-negative constraint and discrete maximum principle is met but the other two conditions ($-\mathbf{K}_{ff}^{-1}\mathbf{K}_{fp} \succeq \mathbf{O}$ and $-\mathbf{K}_{ff}^{-1}\mathbf{K}_{fp}\mathbf{1} \preceq \mathbf{1}$) given in equation (6.3.9) are not fulfilled. Hence, this mesh violates the monotone property and discrete comparison principle for the steady-state analysis.

Case # 4 $c_0(\mathbf{x}, t) = 1$, $c_p(\mathbf{x}, t) = 0$, and $\mathbf{D}(\mathbf{x}) = \begin{pmatrix} (\epsilon_1+y)^2+\epsilon_2(\epsilon_1+x)^2 & -(1-\epsilon_2)(\epsilon_1+x)(\epsilon_1+y) \\ -(1-\epsilon_2)(\epsilon_1+x)(\epsilon_1+y) & (\epsilon_1+x)^2+\epsilon_2(\epsilon_1+y)^2 \end{pmatrix}$ where $\epsilon_1 = 10^{-3}$ and $\epsilon_2 = 10^{-3}$ Potier (2005). The objective of this case study is similar to that of **Case # 3**. Herein, we perform only transient analysis as for the steady-state, the value of concentration field is zero everywhere in the domain. Herein, we observe some interesting features as compared to the previous case study. From Figure 6.10, Figure 6.11, and Table 6.6, it is observed the h -refined mesh satisfies the non-negative constraint and discrete maximum principle for lower values of time-step (for example, when $\Delta t = 10^{-1}$). The reason being the same as the previous case study, (i.e), the stiffness matrix $\mathbf{K}_{ff}^{-1} \succeq \mathbf{O}$. However, similar to the **Case # 3** this mesh violates the monotone property and discrete comparison principle as the conditions $-\mathbf{K}_{ff}^{-1}\mathbf{K}_{fp} \succeq \mathbf{O}$ and $-\mathbf{K}_{ff}^{-1}\mathbf{K}_{fp}\mathbf{1} \preceq \mathbf{1}$ are not satisfied. Finally, on a closing note, in order to satisfy all the discrete properties for **Case # 3** and **Case # 4** within the context of Pao's method, we need a DMP-based mesh based on $\mathbf{D}(\mathbf{x})$. Generating such a mesh for this heterogeneous anisotropic diffusivity tensor is extremely difficult and is beyond the scope of the current dissertation.

Similar type of analysis is performed based on traditional Newton-Raphson method and its modifications. To summarize, for all the above case studies, it is observed that the final solution is almost identical to that of the Pao's method for both steady-state and transient reaction-diffusion problems. Furthermore, from Table 6.7 is evident that we have terminal quadratic convergence rate. However, based on the numerical values of the term ' $\min[\mathbf{c}_k - \mathbf{c}_{k-1}]$ ' quantified in Table 6.7 it is also clear that traditional

Table 6.2: Autocatalytic reaction-diffusion test problem (steady-state and case # 1): Quantitative results for steady-state analysis using Pao’s method on coarse mesh (21×21) at various iteration levels.

k	$\min[\mathbf{c}_k - \mathbf{c}_{k-1}]$	$\ \mathbf{c}_k - \mathbf{c}_{k-1}\ $	$\mathcal{R}_{k,\text{Pao}}$
2	8.68×10^{-1}	1.98×10^1	4.13×10^{-2}
3	0.00	7.42×10^{-1}	3.41×10^{-3}
4	0.00	5.15×10^{-3}	2.42×10^{-5}
5	0.00	2.48×10^{-5}	1.16×10^{-7}
6	0.00	1.19×10^{-7}	5.61×10^{-10}
7	0.00	5.75×10^{-10}	2.71×10^{-12}

Newton-Raphson method and its variants do not satisfy the monotone property and correspondingly the discrete comparison principle. In addition, as noted in previous subsection we performed numerical simulations to check whether we can satisfy various discrete principles based on non-negative formulation outlined in Reference Nagarajan and Nakshatrala (2011) for the above set of cases. From these numerical simulations, we observed that the non-negative formulation did not converge in `MaxIters` = 100. This is because of the nature of the boundary and initial conditions chosen in the autocatalytic reaction-diffusion problem. In case of transient analysis, as there is a sudden change in the concentration values at time $t = 0$, we have steep gradients in concentration near the boundaries of the domain (for example, see Figures 6.6, 6.7, and 6.10). This has to be resolved either by mesh refinement or by a novel physics-compatible numerical formulation that can capture steep gradients in concentration at boundaries even on coarse meshes. However, one should note that it is difficult to resolve such issues related to discontinuities in the input data on coarse meshes within the context of non-negative formulation. This may be one of the factors in non-convergence of the non-negative formulation under such conditions.

6.6 CONCLUDING REMARKS

We have presented three different popular nonlinear techniques to solve semi-linear elliptic and parabolic partial differential equations. First, we discussed the

Table 6.3: Autocatalytic reaction-diffusion test problem (transient and case # 1): Quantitative results for transient analysis using Pao’s method on coarse mesh (21×21) at various iteration levels.

k	$t = 10^{-4}$			$t = 10^{-3}$		
	$\min[\mathbf{c}_k - \mathbf{c}_{k-1}]$	$\ \mathbf{c}_k - \mathbf{c}_{k-1}\ $	$\mathcal{R}_{k,\text{Pao}}$	$\min[\mathbf{c}_k - \mathbf{c}_{k-1}]$	$\ \mathbf{c}_k - \mathbf{c}_{k-1}\ $	$\mathcal{R}_{k,\text{Pao}}$
2	-3.65×10^{-1}	9.08	1.54×10^{-3}	-1.54×10^{-2}	9.24	3.78×10^{-3}
3	-9.16×10^{-6}	7.20×10^{-5}	5.00×10^{-9}	-2.93×10^{-6}	6.09×10^{-4}	7.11×10^{-7}
4	-1.19×10^{-9}	1.10×10^{-8}	7.52×10^{-13}	-9.69×10^{-10}	7.04×10^{-8}	7.93×10^{-11}
5	-1.68×10^{-13}	1.73×10^{-12}	5.97×10^{-15}	-1.67×10^{-13}	8.76×10^{-12}	1.30×10^{-14}

Table 6.4: Autocatalytic reaction-diffusion test problem (transient and case # 2): Quantitative (converged) results for transient analysis based on Pao’s method. Numerical simulations are performed using coarse and h -refined meshes at first time-step using different Δt ’s.

Δt	Mesh	$\max[\mathbf{c}_k]$	$\ \mathbf{c}_k - \mathbf{c}_{k-1}\ $	$\mathcal{R}_{k,\text{Pao}}$
10^{-5}	21×21	1.58	2.82×10^{-15}	6.67×10^{-13}
10^{-5}	81×81	1.27	4.49×10^{-15}	1.65×10^{-13}
10^{-4}	21×21	1.36	1.39×10^{-14}	7.67×10^{-14}
10^{-4}	81×81	1.00	1.96×10^{-15}	5.27×10^{-14}

Table 6.5: Autocatalytic reaction-diffusion test problem (steady-state and case # 3): Quantitative results for steady-state analysis using Pao’s method on a h -refined mesh (161×161) at various iteration levels.

k	$\min[\mathbf{c}_k - \mathbf{c}_{k-1}]$	$\ \mathbf{c}_k - \mathbf{c}_{k-1}\ $	$\mathcal{R}_{k,\text{Pao}}$
2	5.94×10^{-1}	1.25×10^2	3.96×10^{-3}
3	0.00	1.92×10^1	1.11×10^{-3}
4	0.00	1.69	1.07×10^{-4}
5	-3.33×10^{-16}	1.07×10^{-1}	6.79×10^{-6}
6	-1.72×10^{-10}	6.48×10^{-3}	4.13×10^{-7}
7	-9.61×10^{-10}	3.94×10^{-4}	2.51×10^{-8}
8	-8.55×10^{-11}	2.39×10^{-5}	1.53×10^{-9}
9	-5.64×10^{-12}	1.46×10^{-6}	9.29×10^{-11}
10	-3.46×10^{-13}	8.88×10^{-8}	5.66×10^{-12}
11	-2.72×10^{-14}	5.41×10^{-9}	3.48×10^{-13}
12	-1.55×10^{-15}	3.29×10^{-10}	3.14×10^{-14}
13	-1.14×10^{-14}	2.01×10^{-11}	2.06×10^{-14}
14	-2.85×10^{-14}	2.08×10^{-12}	2.04×10^{-14}
15	-1.99×10^{-14}	8.87×10^{-13}	2.02×10^{-14}

Table 6.6: Autocatalytic reaction-diffusion test problem (transient, case # 3, and case # 4): Quantitative (converged) results for transient analysis based on Pao’s method. Numerical simulations are performed using h -refined meshes at first time-step using different Δt ’s.

Δt	Case # 3			
	k	$\min[\mathbf{c}_k]$	$\ \mathbf{c}_k - \mathbf{c}_{k-1}\ $	$\mathcal{R}_{k,\text{Pao}}$
10^{-5}	4	-6.08×10^{-1}	8.89×10^{-15}	3.06×10^{-15}
10^{-4}	5	-6.07×10^{-1}	2.79×10^{-15}	2.65×10^{-15}
10^{-3}	6	-5.99×10^{-1}	6.47×10^{-14}	5.02×10^{-15}
10^{-2}	9	-5.23×10^{-1}	7.32×10^{-14}	9.28×10^{-15}
10^{-1}	15	-3.07×10^{-2}	2.95×10^{-13}	1.54×10^{-14}

Δt	Case # 4			
	k	$\min[\mathbf{c}_k]$	$\ \mathbf{c}_k - \mathbf{c}_{k-1}\ $	$\mathcal{R}_{k,\text{Pao}}$
10^{-5}	4	1.61	7.12×10^{-15}	8.63×10^{-14}
10^{-4}	5	1.60	4.15×10^{-15}	2.61×10^{-14}
10^{-3}	6	1.59	2.58×10^{-14}	2.28×10^{-14}
10^{-2}	7	1.51	3.57×10^{-13}	1.79×10^{-14}
10^{-1}	12	9.69×10^{-1}	3.67×10^{-13}	1.08×10^{-14}

Table 6.7: Autocatalytic reaction-diffusion test problem (steady-state, case # 1, and case # 3): Quantitative results for steady-state analysis based on traditional NR method at various iteration levels. Numerical simulations are performed using various h -refined meshes.

k	Case # 1 (81×81)			Case # 3 (161×161)		
	$\min[\mathbf{c}_k - \mathbf{c}_{k-1}]$	$\ \mathbf{c}_k - \mathbf{c}_{k-1}\ $	$\mathcal{R}_{k,\text{TNR}}/\mathcal{R}_{k,\text{MTNR}}$	$\min[\mathbf{c}_k - \mathbf{c}_{k-1}]$	$\ \mathbf{c}_k - \mathbf{c}_{k-1}\ $	$\mathcal{R}_{k,\text{TNR}}/\mathcal{R}_{k,\text{MTNR}}$
2	9.99×10^{-1}	7.89×10^1	1.79×10^1	8.69×10^{-1}	1.59×10^2	1.13×10^1
3	-6.63×10^{-2}	2.99	1.22×10^{-2}	2.03×10^{-1}	2.16×10^1	6.18×10^{-3}
4	-1.82×10^{-4}	7.01×10^{-3}	2.39×10^{-5}	-6.75×10^{-3}	6.52×10^{-1}	1.47×10^{-4}
5	-1.15×10^{-9}	4.18×10^{-8}	1.48×10^{-10}	-6.87×10^{-6}	6.36×10^{-4}	1.45×10^{-7}
6	-1.22×10^{-15}	4.68×10^{-14}	2.47×10^{-14}	-6.97×10^{-12}	6.24×10^{-10}	1.43×10^{-13}

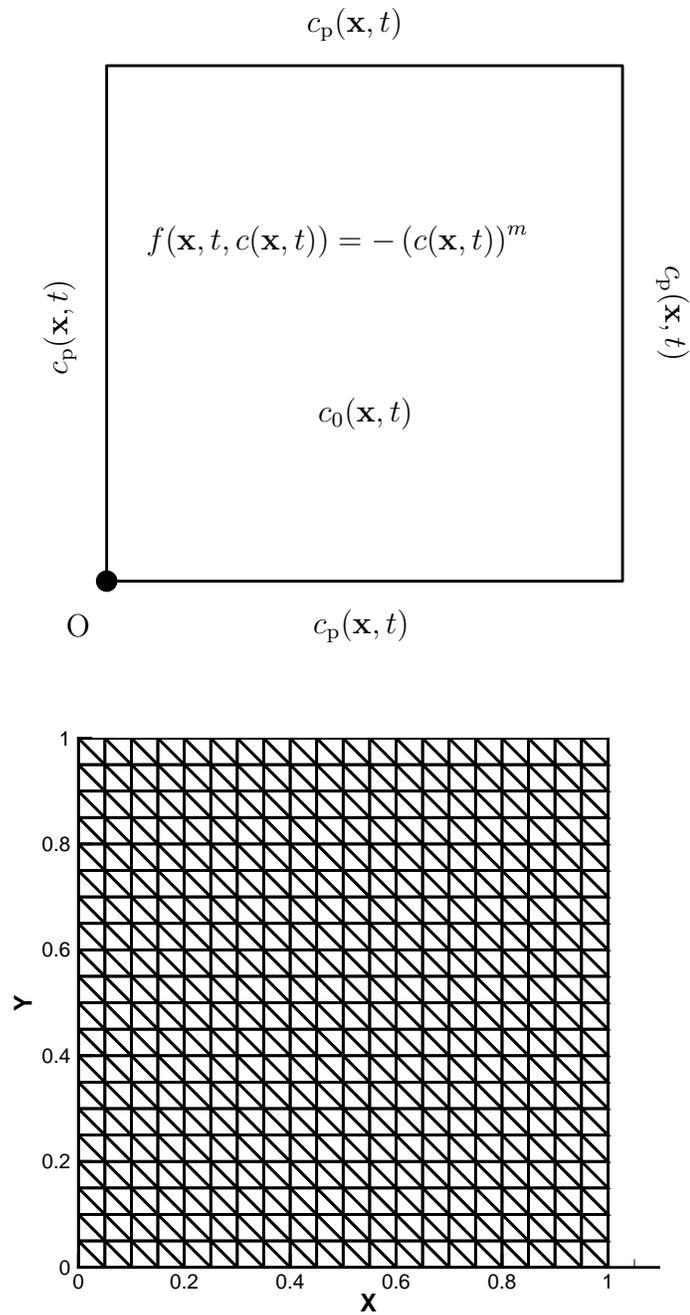


Figure 6.3: Autocatalytic reaction-diffusion test problem: The left figure provides a pictorial description of the test problem with the relevant reaction source term, boundary conditions, and initial conditions.

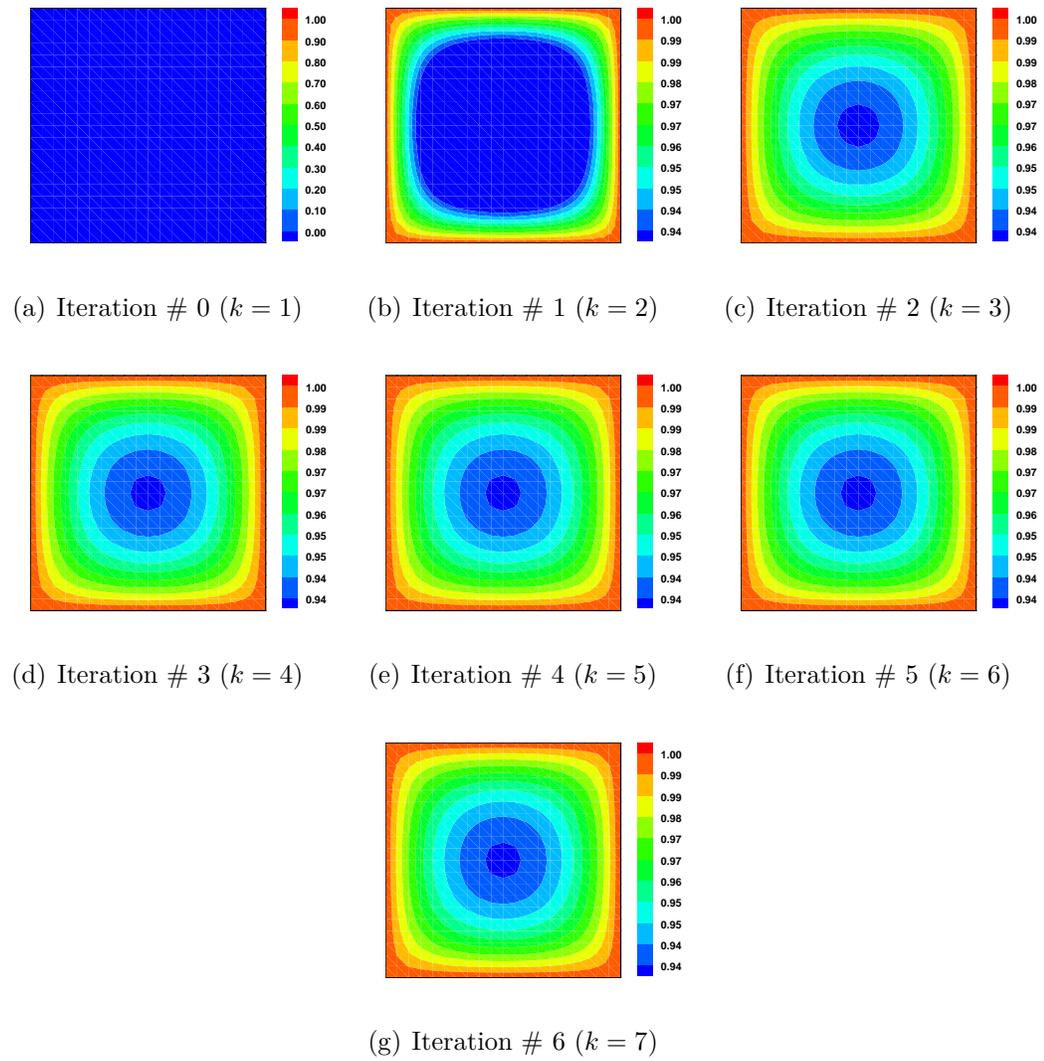
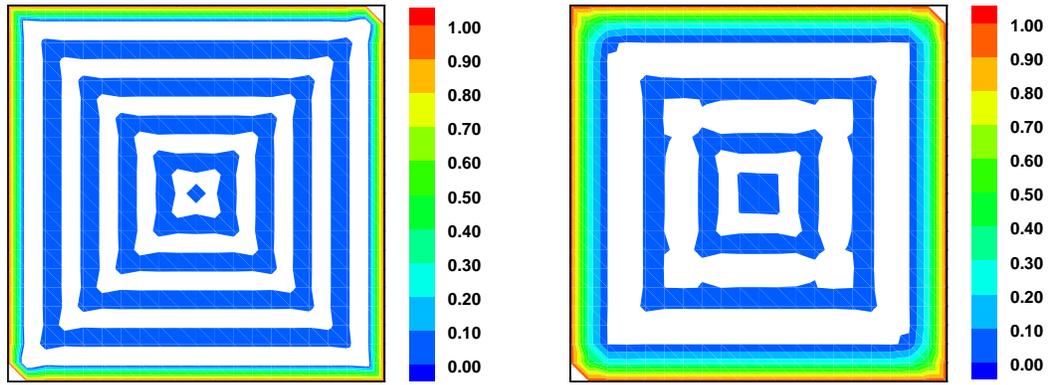


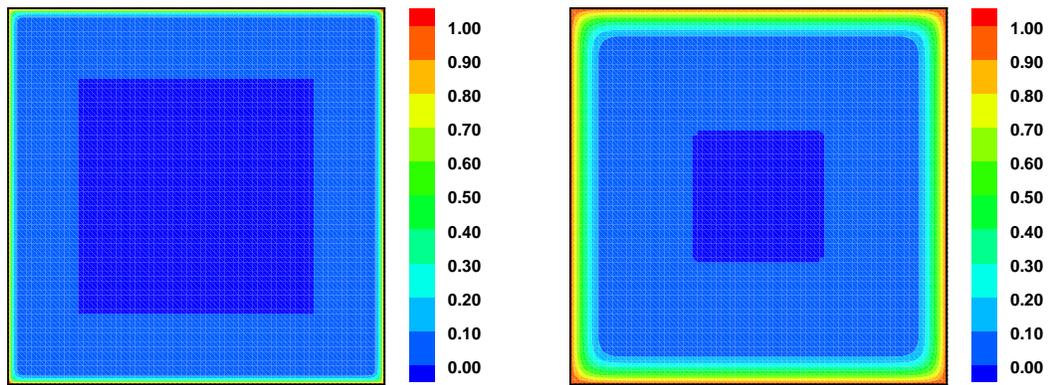
Figure 6.4: Autocatalytic reaction-diffusion test problem (steady-state and case # 1): This figure shows the concentration profile obtained using a coarse mesh (21×21) at various iteration levels based on Pao's method.



(a) $t = 10^{-4}$: Iteration # 4 ($k = 5$)

(b) $t = 10^{-3}$: Iteration # 4 ($k = 5$)

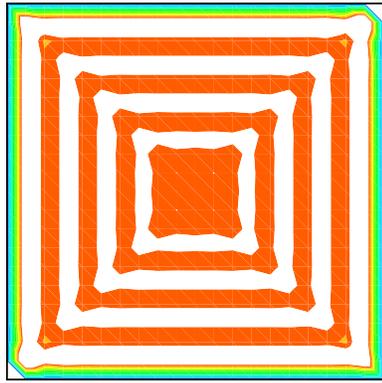
Figure 6.5: Autocatalytic reaction-diffusion test problem (transient and case # 1): This figure shows the concentration profile obtained using a coarse mesh (21×21) at different time levels based on Pao's method for the case when $\Delta t = 10^{-4}$.



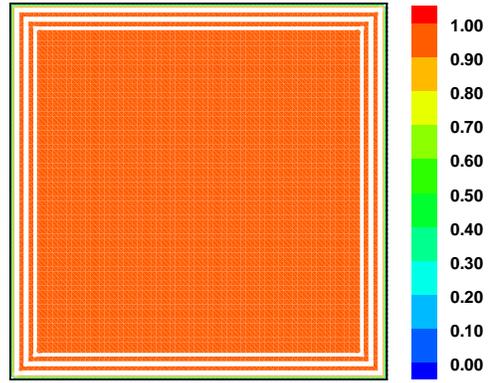
(a) $t = 10^{-4}$: Iteration # 5 ($k = 6$)

(b) $t = 10^{-3}$: Iteration # 5 ($k = 6$)

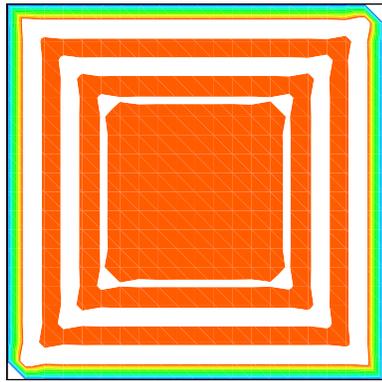
Figure 6.6: Autocatalytic reaction-diffusion test problem (transient and case # 1): This figure shows the concentration profile obtained using a h -refined mesh (81×81) at different time levels based on Pao's method for the case when $\Delta t = 10^{-4}$.



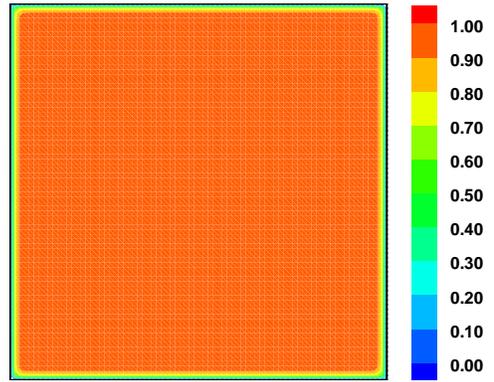
(a) $t = 10^{-5}$ and $\Delta t = 10^{-5}$: 21×21



(b) $t = 10^{-5}$ and $\Delta t = 10^{-5}$: 81×81

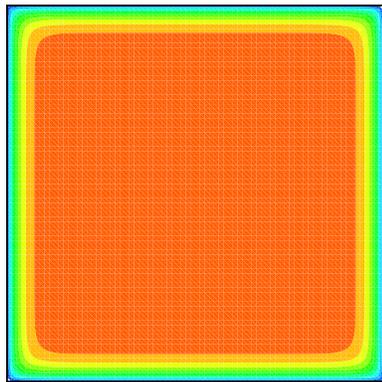


(c) $t = 10^{-4}$ and $\Delta t = 10^{-4}$: 21×21

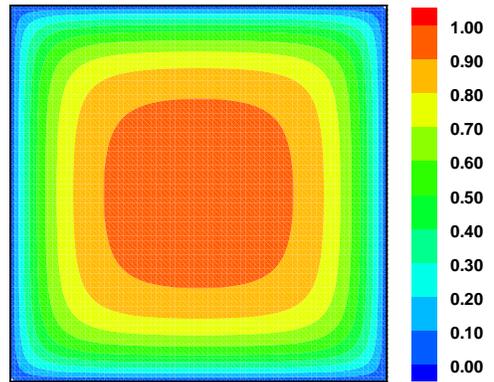


(d) $t = 10^{-4}$ and $\Delta t = 10^{-4}$: 81×81

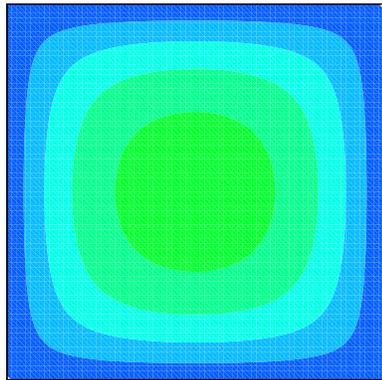
Figure 6.7: Autocatalytic reaction-diffusion test problem (transient and case # 2): This figure shows the converged concentration profiles obtained using a coarse and a h -refined mesh at the first time-step using different Δt 's based on Pao's method.



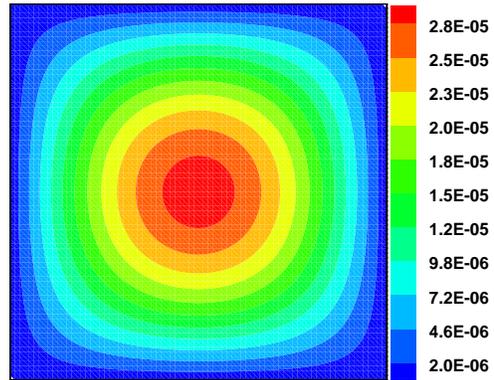
(a) $t = 10^{-3}$: Iteration # 5 ($k = 6$)



(b) $t = 10^{-2}$: Iteration # 7 ($k = 8$)

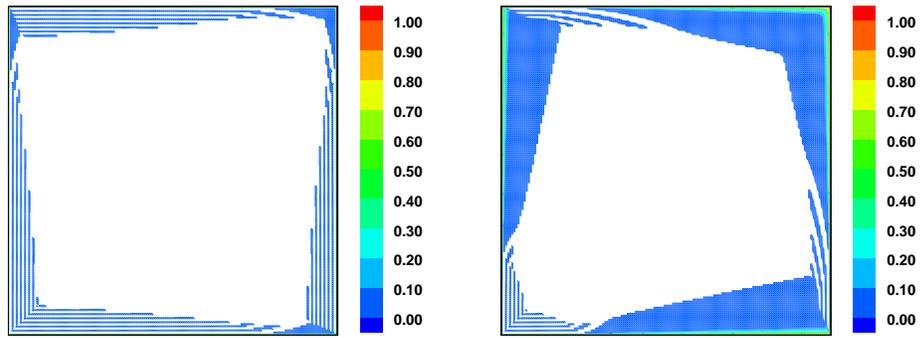


(c) $t = 10^{-1}$: Iteration # 11 ($k = 12$)

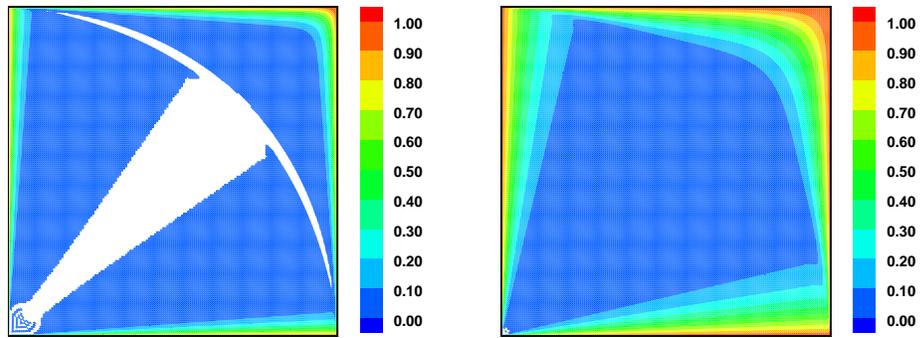


(d) $t = 1$: Iteration # 9 ($k = 10$)

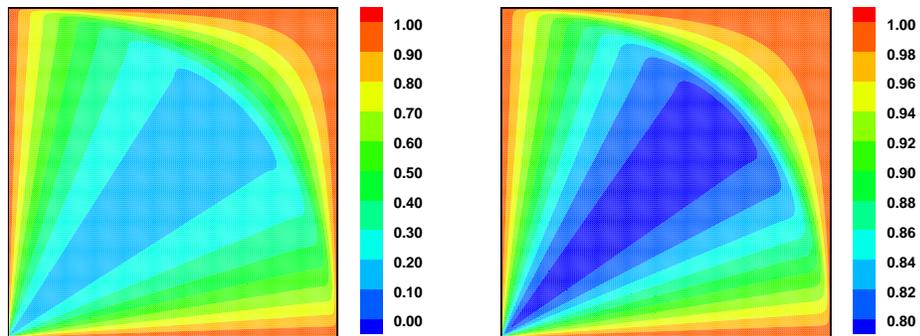
Figure 6.8: Autocatalytic reaction-diffusion test problem (transient and case # 2): This figure shows the converged concentration profiles obtained using a h -refined mesh (81×81) at various time levels using $\Delta t = 10^{-3}$ based on Pao's method.



(a) $\Delta t = 10^{-5}$: Iteration # 4 ($k = 5$) (b) $\Delta t = 10^{-4}$: Iteration # 5 ($k = 6$)

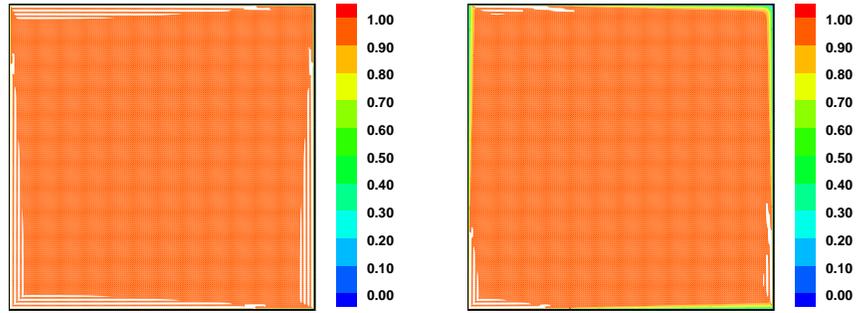


(c) $\Delta t = 10^{-3}$: Iteration # 6 ($k = 7$) (d) $\Delta t = 10^{-2}$: Iteration # 9 ($k = 10$)

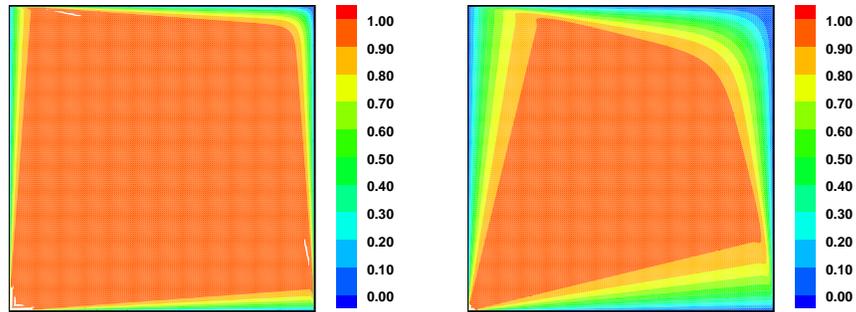


(e) $\Delta t = 10^{-1}$: Iteration # 15 ($k = 16$) (f) Steady-state: Iteration # 4 ($k = 5$)

Figure 6.9: Autocatalytic reaction-diffusion test problem (case # 3): This figure shows the converged concentration profiles obtained using a h -refined mesh (161×161) at first time step using different Δt 's based on Pao's method.

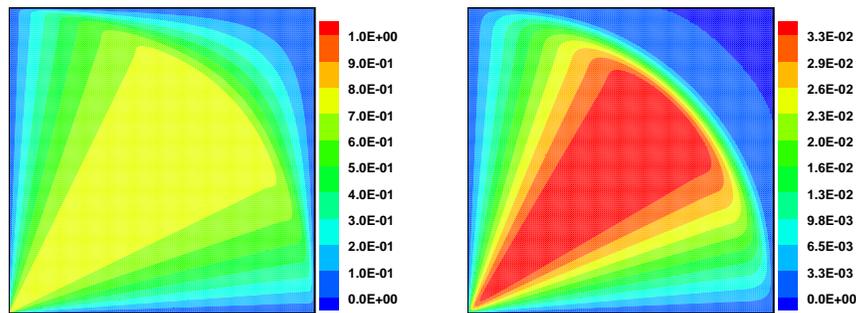


(a) $\Delta t = 10^{-5}$: Iteration # 4 ($k = 5$) (b) $\Delta t = 10^{-4}$: Iteration # 5 ($k = 6$)



(c) $\Delta t = 10^{-3}$: Iteration # 6 ($k = 7$) (d) $\Delta t = 10^{-2}$: Iteration # 7 ($k = 8$)

Figure 6.10: Autocatalytic reaction-diffusion test problem (transient and case # 4): This figure shows the converged concentration profiles obtained using a h -refined mesh (161×161) at first time step using different Δt 's based on Pao's method.



(a) $t = 10^{-1}$: Iteration # 12 ($k = 13$) (b) $t = 1$: Iteration # 15 ($k = 15$)

Figure 6.11: Autocatalytic reaction-diffusion test problem (transient and case # 4): This figure shows the converged concentration profiles obtained using a h -refined mesh at various time levels using $\Delta t = 10^{-1}$ based on Pao's method.

importance of mesh restrictions within the context of transient pure anisotropic diffusion equations. Then we outlined the importance of MOVL and MOHL approaches in solving the linear parabolic partial differential equations. Secondly, we discussed about the pros and cons of Pao's method, traditional Newton-Raphson method, and modification of traditional Newton-Raphson method both within the context of mesh restrictions and non-negative formulation. From numerical experiments we performed, it can be inferred that Pao's method satisfies various discrete properties under mesh restrictions. However, the terminal order of convergence observed was equal to 1. Traditional Newton-Raphson method and its modifications offer better terminal convergence rate but fail to satisfy various discrete properties both under standard single-field Galerkin formulation and non-negative formulation. Finally, from the numerical simulations performed it is evident that using the existing methods it is difficult to preserve all the discrete properties without alleviating the computational cost. Hence, this motivates us to develop new and novel numerical formulations that can capture various physical and mathematical aspects on coarse meshes.

Chapter 7

CONCLUSIONS AND FUTURE WORK

“Men pass away, but their deeds
abide.”

Augustin Cauchy

To conclude, in this dissertation we have proposed different innovative approaches to resolve various problems encountered in solving the advection-diffusion-reaction equations. The proposed methodology is developed on the foundations of constrained optimization least-squares-based low-order finite element framework. Through this framework we are able to satisfy various discrete principles, local species balance, and global species balance. Furthermore, it does not produce spurious node-to-node oscillations, captures interior and boundary layers, and provides sufficiently accurate solutions even on coarse computational grids for advection-dominated advection-diffusion-reaction problems. The *main contributions* of this dissertation can be summarized as follows:

(MC1) In this dissertation, we have unequivocally shown that the popular existing numerical formulations do not satisfy discrete comparison principles, discrete maximum principles, and non-negative constraint. Furthermore, from various representative numerical simulations we have observed that *unphysical* values for concentration of chemical species due to violation of non-negative

constraint and spurious node-to-node oscillations can result in large errors in local and global species balance.

(MC2) We have critiqued three different approaches to satisfy maximum principles, comparison principles, and non-negative constraint for a general linear second-order elliptic equation. We then developed necessary and sufficient conditions on the stiffness matrix \mathbf{K} to meet the mathematical properties. Using these conditions, we derived mesh conditions and constructed DMP-based meshes using various open source mesh generators. Through research findings in Chapter 4 we have shown that placing restrictions on computational grids may not always be a viable approach to achieve physically meaningful non-negative solutions for complex geometries and highly anisotropic media. This research work laid the foundations to develop new methodologies for advective-diffusive-reactive systems that satisfy local and global species balance, comparison principles, maximum principles, and the non-negative constraint on coarse general computational grids.

(MC3) In Chapter 5, we have presented one of a kind numerical methodology for advective-diffusive-reactive systems that satisfies various discrete principles, local species balance, and global species balance. To summarize, the framework has been carefully constructed using the least-squares finite element method (LSFEM). It is also shown that a naive implementation of LSFEM will not meet the desired discrete properties. Additionally, we obtained numerically a scaling law for a transport-controlled bimolecular reaction. In the final chapter, we have used the proposed non-negative methodology and proposed various non-linear techniques to solve slow/moderate reaction problems encountered in environmental and earth sciences.

7.1 FUTURE WORK

Based on the research conducted in this dissertation, we shall outline five promising future research topics and their potential impact on real life applications

(FW1) Recently, there has been surge in using discrete exterior calculus to construct numerical formulations for diffusion-type equations Hirani (2003); Hirani et al. (2015). This new field of emerging numerical techniques are called mimetic discretizations, which are based on the the calculus of differential forms Castillo and Miranda (2013). They are designed to preserve, as much as possible, properties of the continuous partial differential equation da Veiga et al. (2014). In these methods, the geometric nature of physics plays a crucial role and differential geometry is the necessary new language to model such physical problems. A future work can devoted to developing geometric structure-preserving mimetic discretization methods in-combination with mesh restrictions approach for advection-diffusion-reaction equations on surfaces.

(FW2) Many existing numerical formulations are limited for isotropic diffusivities, divergence-free velocity fields, and simple reaction coefficients. Using the mathematical stabilization techniques outlined in Chapter 5, one can construct stabilization parameters and bubble functions for various popular stabilized numerical formulations Gresho and Sani (2000); Donea and Huerta (2003) in-case of anisotropic diffusivity. In addition, many stabilized numerical formulations are catered towards linear non-self-adjoint elliptic operators. A future work can be to design, analyze, and critically review the performance of various stabilized numerical formulations for semilinear and quasilinear non-self-adjoint elliptic and parabolic operators with respect to various discrete principles, local species balance, and global species balance.

- (FW3) The discrete system consider in this dissertation are solved using direct solvers (such as LU solver) available in `MATLAB`. A future work can be to design M -matrix based pre-conditioners, multigrid pre-conditioning methods, and tailored iterative solvers for stabilized finite element formulations discussed in this dissertation.
- (FW4) Recently Chang et.al. Chang et al. (2015) have presented a large-scale non-negative computational framework for pure-diffusion type equations. They studied the performance of the non-negative optimization-based finite element methodology in the context of high performance computing (HPC). They constructed the computational framework utilizing the recently introduced robust `DMPLex` data structures in `PETSc` parallel environment and optimization solvers available in `TAO` libraries Balay et al. (2015); Munson et al. (2012). A future direction can be a systematic study of proposed physics-compatible constrained optimization-based LSFEM framework in parallel computing environment in the lines similar to Reference Chang et al. (2015).
- (FW5) Chemical reaction front propagation and species mixing is an important problem in a wide variety of chemical, biological, and environmental sciences Cencini et al. (2003); Méndez et al. (2010). Relevant applications of great interest include chemical reaction in liquids Mahoney et al. (2012), microfluidic devices Grigoriev and Schuster (2012), plankton population growth Weiss and Provenzale (2008), and reaction front dynamics in porous and random media Xin (2009); Tartakovsky et al. (2008). The proposed methodology can be modified to solve such problems and obtain relevant scaling laws (numerically) for advection-dominated and reaction-dominated flow and transport in anisotropic media.

References

- (2014). *ABAQUS/CAE/Standard, Version 6.14-1 (computer software)*. Simulia, Providence, Rhode Island, USA, www.simulia.com.
- (2014). *COMSOL Multiphysics User's Guide, Version 5.0-1 (computer software)*. COMSOL, Inc., Burlington, Massachusetts, USA, www.comsol.com.
- (2015). *ANSYS Multiphysics, Version 16.0 (computer software)*. ANSYS, Inc., Canonsburg, Pennsylvania, USA, www.ansys.com.
- (2015). *CGAL: Computational Geometry Algorithms Library (computer software)*. Website: <http://www.cgal.org>.
- (2015). *MATLAB 2015a (computer software)*. The MathWorks, Inc., Natick, Massachusetts, USA, www.mathworks.com.
- (2015). *Simmetrix-MeshSim, Automatic Mesh Generation Suite (computer software)*. Simmetrix, Inc., Website: <http://www.simmetrix.com/>, Clifton Park, New York, USA.
- Adrover, A., Cerbelli, S., and Giona, M. (2002). “A spectral approach to reaction/diffusion kinetics in chaotic flows.” *Computers & Chemical Engineering*, 26, 125–139.
- Alliez, P., C.-Steiner, D., Devillers, O., Lévy, B., and Desbrun, M. (2003). “Anisotropic polygonal remeshing.” *ACM Transactions on Graphics*, 22, 485–493.
- Antoulas, A. C., Sorensen, D. C., and Gugercin, S. (2001). “A survey of model reduction methods for large-scale systems.” *Contemporary Mathematics*, 280, 193–220.

- Aoki, M. (2004). *Modeling Aggregate Behavior and Fluctuations in Economics: Stochastic Views of Interacting Agents*. Cambridge University Press, Cambridge, UK.
- Aris, R. (1975a). *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts: Questions of Uniqueness, Stability, and Transient Behavior*, Vol. 2. Oxford: Clarendon Press, London, UK.
- Aris, R. (1975b). *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts: The Theory of the Steady State*, Vol. 1. Oxford: Clarendon Press, London, UK.
- Atkinson, K. and Han, W. (2001). *Theoretical Numerical Analysis: A Functional Analysis Framework*. Springer-Verlag, New York, USA.
- Augustin, M., Caiazzo, A., Fiebach, A., Fuhrmann, J., John, V., Linke, A., and Umla, R. (2011). “An assessment of discretizations for convection–dominated convection–diffusion equations.” *Computer Methods in Applied Mechanics and Engineering*, 200, 3395–3409.
- Avcı, C. B. (1994). “Evaluation of flow leakage through abandoned wells and boreholes.” *Water Resources Research*, 30, 2565–2578.
- Ayub, M. and Masud, A. (2003). “A new stabilized formulation for convective–diffusive heat transfer.” *Numerical Heat Transfer, Part B*, 44, 1–23.
- Bahr, J. M. and Rubin, J. (1987). “Direct comparison of kinetic and local equilibrium formulations for solute transport affected by surface reactions.” *Water Resources Research*, 23, 438–452.
- Balay, S., Abhyankar, S., Adams, M. F., Brown, J., Brune, P., Buschelman, K., Eijkhout, V., Gropp, W. D., Kaushik, M. G. K., McInnes, L. C., Rupp, K.,

- Smith, B. F., Zampini, S., and Zhang, H. (2015). “PETSc Users Manual.” *Report No. ANL-95/11 - Revision 3.6*, Argonne National Laboratory.
- Baveye, P. and Valocchi, A. J. (1989). “An evaluation of mathematical models of the transport of biologically reacting solutes in saturated soils and aquifers.” *Water Resources Research*, 25, 1413–1421.
- J. Bear, C. F. Tsang, and G. de Marsily, eds. (1993). *Flow and Contaminant Transport in Fractured Rock*. Academic Press Inc., San Diego, California, USA.
- Belytschko, T., Liu, W. K., and Moran, B. (2000). *Nonlinear Finite Elements for Continua and Structures*. John Wiley & Sons, Inc., New York, USA.
- Berman, A. and Plemmons, R. J. (1979). *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York, USA.
- Berzins, M. (2001). “Modified mass matrices and positivity preservation for hyperbolic and parabolic PDEs.” *Communications in Numerical Methods in Engineering*, 17, 659–666.
- Bochev, P. B. and Gunzburger, M. D. (2009). *Least-Squares Finite Element Methods*. Number 166 in Applied Mathematical Sciences. Springer, New York, USA.
- Boissonnat, J.-D., Pons, J.-P., and Yvinec, M. (2009). “From segmented images to good quality meshes using Delaunay refinement.” *Emerging Trends in Visual Computing*, F. Nielsen, ed., Vol. 5416 of *Lecture Notes in Computer Science*, Berlin, Heidelberg. Springer, 13–37.
- Boissonnat, J.-D., Wormser, C., and Yvinec, M. (2008). “Locally uniform anisotropic meshing.” *Proceedings of the twenty-fourth annual symposium on Computational geometry, ACM*. 270–277.

- Boissonnat, J.-D., Wormser, C., and Yvinec, M. (2011). “Anisotropic Delaunay mesh generation.” *HAL: inria-00615486, version 1*.
- Borden, R. C. and Bedient, P. B. (1986). “Transport of dissolved hydrocarbons influenced by Oxygen-limited biodegradation 2. Field application.” *Advances in Water Resources*, 22, 1983–1990.
- Borsuk, M. and Kondratiev, V. (2006). *Elliptic Boundary Value Problems of Second Order in Piecewise Smooth Domains*. Elsevier Science, San Diego, USA.
- Bowen, R. M. (1976). “Theory of mixtures.” *Continuum Physics*, A. C. Eringen, ed., Vol. III, Academic Press, New York.
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge, UK.
- Brandts, J. H., Korotov, S., and Křížek, M. (2008). “The discrete maximum principle for linear simplicial finite element approximations of a reaction–diffusion problem.” *Linear Algebra and its Applications*, 429, 2344–2357.
- Brezzi, F., Douglas, J., Durran, R., and Marini, L. D. (1987). “Mixed finite elements for second order elliptic problems in three variables.” *Numerische Mathematik*, 51, 237–250.
- Brezzi, F., Lipnikov, K., and Shashkov, M. (2005). “Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes.” *SIAM Journal on Numerical Analysis*, 43, 1872–1896.
- Burdakov, O., Kapyrin, I., and Vassilevski, Y. (2012). “Monotonicity recovering and accuracy preserving optimization methods for postprocessing finite element solutions.” *Journal of Computational Physics*, 231, 3126–3142.

- Burman, E. and Ern, A. (2005). “Stabilized Galerkin approximation of convection-diffusion-reaction equations: Discrete maximum principle and convergence.” *Mathematics of Computation*, 74, 1637–1652.
- Burman, E. and Hansbo, P. (2004). “Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems.” *Computer Methods in Applied Mechanics and Engineering*, 193, 1437–1453.
- Cant, R. S. and Mastorakos, E. (2008). *An Introduction to Turbulent Reacting Flows*. Imperial College Press, London, UK.
- Castillo, J. E. and Miranda, G. F. (2013). *Mimetic Discretization Methods*. CRC Press, Taylor & Francis Group, Boca Raton, Florida, USA.
- Castro-Diaz, M. J., Hecht, F., Mohammadi, B., and Pironneau, O. (1997). “Anisotropic unstructured mesh adaptation for flow simulations.” *International Journal for Numerical Methods in Fluids*, 25, 475–491.
- Cencini, M., Lopez, C., and Vergni, D. (2003). *The Kolmogorov Legacy in Physics*, Vol. 636 of *Lecture Notes in Physics*, chapter Reaction-Diffusion Systems: Front Propagation and Spatial Structures, 187–213. Springer-Verlag, Berlin, Heidelberg, Germany.
- Chang, J., Karra, S., and Nakshatrala, K. B. (2015). “Large-scale optimization-based non-negative computational framework for diffusion equations: Parallel implementation and performance studies.” *Available on arXiv:1506.08435*.
- U. K. Chatterjee, S. K. Bose, and S. K. Roy, eds. (2001). *Environmental Degradation of Metals: Corrosion Technology*. Marcel Dekker, Inc., New York, USA.
- Chung, T. J. (2010). *Computational Fluid Dynamics*. Cambridge University Press, New York, USA, second edition.

- Ciarlet, P. G. (1970a). “Discrete maximum principle for finite-difference operators.” *Aequationes Mathematicae*, 4, 338–352.
- Ciarlet, P. G. (1970b). “Discrete variational Green’s function – I.” *Aequationes Mathematicae*, 4, 74–82.
- Ciarlet, P. G. and Raviart, P.-A. (1973). “Maximum principle and uniform convergence for the finite element method.” *Computer Methods in Applied Methods and Engineering*, 2, 17–31.
- Ciarlet, P. G. and Varga, R. S. (1970). “Discrete variational Green’s function – II.” *Numerische Mathematik*, 16, 115–128.
- Cirpka, O. A. and Valocchi, A. J. (2007). “Two-dimensional concentration distribution for mixing controlled bioreactive transport in strata state.” *Advanced in Water Resources*, 30, 1668–1679.
- Codina, R. (1998). “Comparison of some finite element methods for solving the diffusion-convection-reaction equation.” *Computer Methods in Applied Mechanics and Engineering*, 156, 185–210.
- Codina, R. (2000). “On stabilized finite element methods for linear systems of convection-diffusion-reaction equations.” *Computer Methods in Applied Mechanics and Engineering*, 188, 61–82.
- Crank, J. (1975). *The Mathematics of Diffusion*. Clarendon press, Oxford, second edition.
- da Veiga, L. B., Lipnikov, K., and Manzini, G. (2014). *The Mimetic Finite Difference Method for Elliptic Problems*, Vol. 11 of *Modeling, Simulation & Applications*. Springer, Switzerland.

- Demmel, J. W. (1997). *Applied Numerical Linear Algebra*. SIAM, Philadelphia, Pennsylvania, USA.
- Dentz, M., Borgne, T. L., Englert, A., and Bijeljic, B. (2011). “Mixing, spreading and reaction in heterogeneous media: A brief review.” *Journal of Contaminant Hydrology*, 120-121, 1–17.
- Dobrzynski, C. (2012). “Mmg3d: User Guide.” *Technical Report, Website: <http://hal.inria.fr/docs/00/68/18/13/PDF/RT-422.pdf>*.
- Donea, J. and Huerta, A. (2003). *Finite Element Methods for Flow Problems*. John Wiley & Sons Inc., Chichester, UK.
- Droniou, J. and Potier, C. L. (2011). “Construction and convergence study of schemes preserving the elliptic local maximum principle.” *SIAM Journal of Numerical Analysis*, 49, 459–490.
- Drăgănescu, A., Dupont, T., and Scott, L. (2005). “Failure of the discrete maximum principle for an elliptic finite element problem.” *Mathematics of Computation*, 74, 1–23.
- Duderstadt, J. J. and Hamilton, L. J. (1976). *Nuclear reactor analysis*. John Wiley & Sons, Inc., New Jersey, USA.
- Ebigbo, A., Class, H., and Helmig, R. (2007). “CO₂ leakage through an abandoned well: Problem-oriented benchmarks.” *Computational Geosciences*, 11, 103–115.
- Edwards, A. M. and Yool, A. (2000). “The role of higher predation in plankton population models.” *Journal of Plankton Research*, 22, 1085–1112.
- Elshebli, M. A. T. (2008). “Discrete maximum principle for the finite element solution of linear non-stationary diffusion–reaction problems.” *Applied Mathematical Modelling*, 32, 1530–1541.

- Epstein, I. R. and Pojman, J. A. (1998). *An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns, and Chaos*, Vol. 10 of *Topics in Physical Chemistry*. Oxford University Press, New York, USA.
- Erdi, P. and Toth, J. (1989). *Mathematical Models of Chemical Reactions: Theory and Applications of Deterministic and Stochastic Models*. Manchester University Press, Manchester, UK.
- Ern, A. and Guermond, J.-L. (2000). *Theory and Practice of Finite Elements*. Springer-Verlag, New York, USA.
- Errami, H., Eiswirth, M., Grigoriev, D., Seiler, W. M., Sturm, T., and Weber, A. (2015). “Detection of Hopf bifurcations in chemical reaction networks using convex coordinates.” *Journal of Computational Physics*, 291, 279–302.
- Erten, H. and Üngör, A. (2007). “Computing acute and non-obtuse triangulations.” *19th Canadian Conference on Computational Geometry*, 205–208.
- Erten, H. and Üngör, A. (2009a). “Computing triangulations without small and large angles.” *Sixth International Symposium on Voronoi Diagrams, IEEE*. 192–201.
- Erten, H. and Üngör, A. (2009b). “Quality triangulations with locally optimal Steiner points.” *SIAM Journal on Scientific Computing*, 31, 2103–2130.
- Erturk, E. (2009). “Discussions on driven cavity flow.” *International Journal for Numerical Methods in Fluids*, 60, 275–294.
- Evans, L. C. (1998). *Partial Differential Equations*. American Mathematical Society, Providence, Rhode Island, USA.
- Faragó, I., Horváth, R., and Korotov, S. (2005). “Discrete maximum principle for linear parabolic problems solved on hybrid meshes.” *Applied Numerical Mathematics*, 53, 249–264.

- Farkas, M. (2001). *Dynamical Models in Biology*. Academic Press, San Diego, USA.
- Field, R. J. and Noyes, R. M. (1974). “Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction.” *The Journal of Chemical Physics*, 60, 1877–1884.
- Fife, P. C. (1979). *Mathematical Aspects of Reacting and Diffusing Systems*, Vol. 28 of *Lecture Notes in Biomathematics*. Springer-Verlag, New York, USA.
- Fogler, H. S. (2006). *Elements of Chemical Reaction Engineering*. Pearson Education, Inc., New Jersey, USA, fourth edition.
- Franca, L. P., Nesliturk, A., and Stynes, M. (1998). “On the stability of residual-free bubbles for convection-diffusion problems and their approximation by a two-level finite element method.” *Computer Methods in Applied Mechanics and Engineering*, 166, 35–49.
- Frey, P. J. and Alauzet, F. (2005). “Anisotropic mesh adaptation for CFD computations.” *Computer Methods in Applied Mechanics and Engineering*, 194, 5068–5082.
- Garimella, R., Kucharik, M., and Shashkov, M. (2007). “An efficient linearity and bound preserving conservative interpolation (remapping) on polyhedral meshes.” *Computers & fluids*, 36, 224–237.
- George, P.-L. and Frey, S. (2010). *Mesh Generation*, Vol. 32. John Wiley & Sons, Inc., New Jersey, USA, second edition.
- Geuzaine, C. and Remacle, J.-F. (2015). *Gmsh: A three-dimensional finite element mesh generator with pre- and post-processing facilities (computer software)*. URL: <http://www.geuz.org/gmsh/>.
- Gilbarg, D. and Trudinger, N. S. (2001). *Elliptic Partial Differential Equations of Second Order*. Springer, New York, USA.

- Gondzio, J. (1996). “Multiple centrality corrections in a primal-dual method for linear programming.” *Computational Optimization and Applications*, 6, 137–156.
- Gould, N. and Toint, P. L. (2004). “Preprocessing for quadratic programming.” *Mathematical Programming, Series B*, 100, 95–132.
- Graham, A. (1981). *Kronecker Products and Matrix Calculus: With Applications*. Halsted Press, Chichester, UK.
- Gresho, P. M. and Sani, R. L. (2000). *Incompressible Flow and the Finite Element Method: Advection-Diffusion*, Vol. 1. John Wiley & Sons, Inc., Chichester, UK.
- R. Grigoriev and H.-G. Schuster, eds. (2012). *Transport and Mixing in Laminar Flows: From Microfluidics to Oceanic Currents*. Reviews of Nonlinear Dynamics and Complexity. Wiley-VCH Verlag & Co., Weinheim, Germany.
- Heath, M. T. (2005). *Scientific Computing—An Introductory Survey*. McGraw-Hill, New York, USA, second edition.
- Hecht, F. (2006). “BAMG: Bidimensional Anisotropic Mesh Generator.” *Technical Report, INRIA, Rocquencourt, Website: <http://www.ann.jussieu.fr/hecht/ftp/bamg/bamg.pdf>*.
- Hecht, F. (2012). “New development in FreeFem++.” *Journal of Numerical Mathematics*, 20, 251–266.
- Hecht, F., Auliac, S., Pironneau, O., Morice, J., Hyaric, A. L., and Ohtsuka, K. (2014). *FreeFem++*. URL: <http://www.freefem.org/ff++/>, third edition. Version 3.26-2.
- Henry, A. F. (1975). *Nuclear reactor analysis*. MIT Press, Cambridge, MA, USA.
- Hirani, A. (2003). “Discrete exterior calculus,” PhD thesis, California Institute of Technology.

- Hirani, A. N., Nakshatrala, K. B., and Chaudhry, J. H. (2015). “A Numerical method for Darcy flow derived using Discrete Exterior Calculus.” *International Journal of Computational Methods in Engineering Science and Mechanics*, 16, 151–169.
- Höhn, D. W. and Mittelmann, D. H. D. (1981). “Some remarks on the discrete maximum-principle for finite elements in higher order.” *Computing*, 27, 145–154.
- Hornung, U. (1996). *Homogenization and Porous Media*. Springer-Verlag, New York, USA.
- Horváth, R. (2008). “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.
- Hsieh, P. W. and Yang, S. Y. (2009). “On efficient least-squares finite element methods for convection-dominated problems.” *Computer Methods in Applied Mechanics and Engineering*, 199, 183–196.
- Huang, W. (2001). “Variational mesh adaptation: Isotropy and equidistribution.” *Journal of Computational Physics*, 174, 903–924.
- Huang, W. (2005). “Metric tensors for anisotropic mesh generation.” *Journal of Computational Physics*, 204, 633–665.
- Huang, W. (2006). “Mathematical principles of anisotropic mesh adaptation.” *Communications in Computational Physics*, 1, 276–310.
- Huang, W. (2010). “Discrete maximum principle and a Delaunay-type mesh condition for linear finite element approximations of two-dimensional anisotropic diffusion problems.” *arXiv:1008.0562*.

- Huang, W. (2013). “Sign-preserving of principal eigenfunctions in P1 finite element approximation of eigenvalue problems of second-order elliptic operators.” *arXiv:1306.1987*.
- Huang, W. (2014). “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.
- Huang, W., Kamenski, L., and Lang, J. (2013). “Stability of explicit Runge–Kutta methods for finite element approximation of linear parabolic equations on anisotropic meshes.” *WIAS Preprint No. 1869*.
- Huang, W. and Li, X. (2010). “An anisotropic mesh adaptation method for the finite element solution of variational problems.” *Finite Elements in Analysis and Design*, 46, 61–73.
- Huang, W. and Wang, Y. (2014). “Discrete maximum principle for the weak Galerkin method for anisotropic diffusion problems.” *arXiv:1401.6232*.
- Huang, Y. Q., Su, Y. F., Wei, H. Y., and Yi, N. Y. (2013). “Anisotropic mesh generation methods based on ACVT and natural metric for anisotropic elliptic equation.” *Science China Mathematics*, 56, 2615–2630.
- Hughes, T. J. R., Engel, G., Mazzei, L., and Larson, M. G. (2000). “The continuous Galerkin method is locally conservative.” *Journal of Computational Physics*, 163, 467–488.
- Huisman, J. and Weissing, F. J. (1999). “Biodiversity of plankton by species oscillations and chaos.” *Nature*, 402, 407–410.
- Ilinca, F. and Héту, J.-F. (2002). “Galerkin gradient least-squares formulations for

- transient conduction heat transfer.” *Computer Methods in Applied Mechanics and Engineering*, 191, 3073–3097.
- Ishihara, K. (1987). “Strong and weak discrete maximum principles for matrices associated with elliptic problems.” *Linear Algebra and its Applications*, 88-89, 431–448.
- Jiang, B. (1998). *The Least-Squares Finite Element Method: Theory and Applications in Computational Fluid Dynamics and Electromagnetics*. Scientific Computation. Springer-Verlag, New York, USA.
- John, V. and Knobloch, P. (2007). “On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I—A review.” *Computer Methods in Applied Mechanics and Engineering*, 196, 2197–2215.
- Jones, E., Oliphant, T., and Peterson, P. (2014). “SciPy: Open source scientific tools for Python.
- Joseph, D. D. and Lundgren, T. S. (1973). “Quasilinear Dirichlet problems driven by positive sources.” *Archive for Rational Mechanics and Analysis*, 49, 241–269.
- Karátson, J. and Korotov, S. (2005). “Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions.” *Numerische Mathematik*, 99, 669–698.
- Karimi, S. and Nakshatrala, K. B. (2015). “A monolithic multi-time-step computational framework for first-order transient systems with disparate scales.” *Computer Methods in Applied Mechanics and Engineering*, 283, 419–453.
- Kastenbergh, W. E. and Chambré, P. L. (1968). “On the stability of nonlinear space–dependent reactor kinetics.” *Nuclear Science and Engineering*, 31, 67–79.

- Keener, J. and Sneyd, J. (2009). *Mathematical Physiology I: Cellular Physiology*. Springer, New York, USA.
- Kelley, C. T. (1987). *Solving Nonlinear Equations with Newton's Method*. Society for Industrial and Applied Mathematics, Philadelphia, USA.
- Kernevez, J.-P. (1980). *Enzyme Mathematics*, Vol. 10 of *Studies in Mathematics and its Applications*. North Holland Publishing Company, Elsevier, New York, USA.
- Kinoshita, S. (2013). *Pattern Formations & Oscillatory Phenomena*. Elsevier Publications, Oxford, UK.
- Konikow, L. F. and Hornberger, G. Z. (2006). “Modeling effects of multinode wells on solute transport.” *Groundwater*, 44, 648–660.
- Kopteva, N. (2004). “How accurate is the streamline–diffusion FEM inside characteristic (boundary and interior) layers?.” *Computer Methods in Applied Mechanics and Engineering*, 193, 4875–4889.
- Kotomin, E. and Kuzovkov, V. (1996). *Modern Aspects of Diffusion-Controlled Reactions: Cooperative Phenomena in Bimolecular Processes*, Vol. 34 of *Comprehensive Chemical Kinetics*, edited by R. G. Compton and G. Hancock. Elsevier Science Publishers B.V., Amsterdam, The Netherlands.
- Kreuzer, C. (2014). “A note on why enforcing discrete maximum principles by a simple a posteriori cutoff is a good idea.” *Numerical Methods for Partial Differential Equations*, 30, 994–1002.
- Kucharik, M., Shashkov, M., and Wendroff, B. (2003). “An efficient linearity-and-bound-preserving remapping method.” *Journal of Computational Physics*, 188, 462–471.

- Kuramoto, Y. (2003). *Chemical Oscillations, Waves, and Turbulence*. Dover Publications, New York, USA.
- Křížek, M. and Qun, L. (1995). “On diagonal dominance of stiffness matrices in 3D.” *East-West Journal of Numerical Mathematics*, 3, 59–69.
- Lacombe, S., Sudicky, E. A., Frapce, S. K., and Unger, A. J. A. (1995). “Influence of leaky boreholes on cross-formational groundwater flow and contaminant transport.” *Water Resources Research*, 31, 1871–1882.
- Laug, P. and Borouchaki, H. (1996). “The BL2D mesh generator: Beginner’s guide, user’s and programmer’s manual.” *Technical Report*, Website: <http://hal.archives-ouvertes.fr/docs/00/06/99/77/PDF/RT-0194.pdf>.
- Lax, P. D. (2002). *Functional Analysis*. John Wiley & Sons, Inc., New York, USA.
- Lazarov, R. D., Tobiska, L., and Vassilevski, P. S. (1997). “Streamline diffusion least-squares mixed finite element methods for convection–diffusion problems.” *East West Journal of Numerical Mathematics*, 5, 249–264.
- Leibundgut, C., Maloszewski, P., and Külls, C. (2009). *Tracers in Hydrology*. John Wiley & Sons Inc., West Sussex, UK.
- Leung, A. W. (2009). *Nonlinear Systems of Partial Differential Equations: Applications to Life and Physical Sciences*. World Scientific Publishing Ltd., Massachusetts, USA.
- Li, X. and Huang, W. (2010). “An anisotropic mesh adaptation method for the finite element solution of heterogeneous anisotropic diffusion problems.” *Journal of Computational Physics*, 229, 8072–8094.
- Li, X. and Huang, W. (2013). “Maximum principle for the finite element solution

- of time-dependent anisotropic diffusion problems.” *Numerical Methods for Partial Differential Equations*, 29, 1963–1985.
- Lipnikov, K., Manzini, G., and Svyatskiy, D. (2011). “Analysis of the monotonicity conditions in the mimetic finite difference method for elliptic problems.” *Journal of Computational Physics*, 230, 2620–2642.
- Lipnikov, K., Shashkov, M., Svyatskiy, D., and Vassilevski, Y. (2007). “Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes.” *Journal of Computational Physics*, 227, 492–512.
- Liska, R. and Shashkov, M. (2008). “Enforcing the discrete maximum principle for linear finite element solutions for elliptic problems.” *Communications in Computational Physics*, 3, 852–877.
- Lu, C., Huang, W., and Qiu, J. (2012). “Maximum principle in linear finite element approximations of anisotropic diffusion-convection-reaction problems.” *arXiv:1201.3564*.
- Lu, C., Huang, W., and Vleck, E. S. V. (2013). “The cutoff method for the numerical computation of nonnegative solutions of parabolic PDEs with application to anisotropic diffusion and lubrication-type equations.” *Journal of Computational Physics*, 242, 24–36.
- Mahoney, J., Bargteil, D., Kingsbury, M., Mitchell, K., and Solomon, T. (2012). “Invariant barriers to reactive front propagation in fluid flows.” *Europhysics Letters*, 98, 44005(6).
- McCarty, P. L. and Criddle, C. S. (2012). “Chemical and biological processes: The need for mixing.” *Delivery and Mixing in the Subsurface: Processes and Design Principles for In Situ Remediation*, P. K. Kitanidis and P. L. McCarty, eds., Springer, 7–52.

- Mehrotra, S. (1992). “On the implementation of a primal-dual interior point method.” *SIAM Journal on Optimization*, 2, 575–601.
- Mei, Z. (2000). *Numerical Bifurcation Analysis for Reaction-Diffusion Equations*. Springer-Verlag, New York, USA.
- Méndez, V., Fedotov, S., and Horsthemke, W. (2010). *Reaction–Transport Systems: Mesoscopic Foundations, Fronts, and Spatial Instabilities*. Springer Series in Synergetics. Springer-Verlag, Berlin, Heidelberg, Germany.
- Menzinger, M. and Dutt, A. K. (1990). “The myth of the well–stirred CSTR in chemical instability experiments: The chlorite/iodide reaction.” *The Journal of Physical Chemistry*, 94, 4510–4514.
- Mincsovcics, M. E. and Hórvath, T. L. (2012). “On the differences of the discrete weak and strong maximum principles for elliptic operators.” *Large-Scale Scientific Computing*, I. Lirkov, S. Margenov, and J. Waśniewski, eds., Vol. 7116 of *Lecture Notes in Computer Science*, Berlin, Heidelberg. Springer-Verlag, 614–621.
- Mizukami, A. (1986). “Variable explicit finite element methods for unsteady heat conduction equations.” *Computer Methods in Applied Mechanics and Engineering*, 59, 101–109.
- Morton, K. W. (1996). *Numerical Solution of Convection-Diffusion Problems*, Vol. 12 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall, London, UK.
- Mudunuru, M. K. and Nakshatrala, K. B. (2012). “A framework for coupled deformation-diffusion analysis with application to degradation / healing.” *International Journal for Numerical Methods in Engineering*, 89, 1144–1170.

- Mudunuru, M. K. and Nakshatrala, K. B. (2015). “On mesh restrictions to satisfy comparison principles, maximum principles, and the non-negative constraint: Recent developments and new results.” *Available on arXiv:1502.06164*.
- Munson, T., Sarich, J., Wild, S., Benson, S., and McInnes, L. C. (2012). “TAO 2.0 users manual.” *Report No. ANL/MCS-TM-322*, Mathematics and Computer Science Division, Argonne National Laboratory. <http://www.mcs.anl.gov/tao>.
- Murray, J. D. (1968a). “A simple method for obtaining approximate solutions for a class of diffusion–kinetics enzyme problems: I. General class and illustrative examples.” *Mathematical Biosciences*, 2, 379–411.
- Murray, J. D. (1968b). “A simple method for obtaining approximate solutions for a class of diffusion–kinetics enzyme problems: II. Further examples and nonsymmetric problems.” *Mathematical Biosciences*, 3, 115–133.
- Murray, J. D. (1993). *Mathematical Biology*. Springer-Verlag, New York, USA.
- Myers, T. (2012). “Potential contaminant pathways from hydraulically fractured shale to aquifers.” *Groundwater*, 50, 872–882.
- Nagarajan, H. and Nakshatrala, K. B. (2011). “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.
- Nagumo, J., Yoshizawa, S., and Arimoto, S. (1965). “Bistable transmission lines.” *Circuit Theory, IEEE Transactions*, 12, 400–412.
- Nakshatrala, K. B., Mudunuru, M. K., and Valocchi, A. J. (2013). “A numerical framework for diffusion-controlled bimolecular-reactive systems to enforce maximum principles and non-negative constraint.” *Journal of Computational Physics*, 253, 278–307.

- Nakshatrala, K. B., Nagarajan, H., and Shabouei, M. (2013). “A numerical methodology for enforcing maximum principles and the non-negative constraint for transient diffusion equations.” *Available on arXiv: 1206.0701v3*.
- Nakshatrala, K. B. and Turner, D. Z. (2013). “A mixed formulation for a modification to Darcy equation based on Picard linearization and numerical solutions to large-scale realistic problems.” *International Journal for Computational Methods in Engineering Science and Mechanics*, 14, 524–541.
- Nakshatrala, K. B. and Valocchi, A. J. (2009). “Non-negative mixed finite element formulations for a tensorial diffusion equation.” *Journal of Computational Physics*, 228, 6726–6752.
- Nakshatrala, K. B. and Valocchi, A. J. (2010). “Variational structure of the optimal artificial diffusion method for the advection-diffusion equation.” *International Journal of Computational Methods*, 7, 559–572.
- Neufeld, Z. and H.-García, E. (2010). *Chemical and Biological Processes in Fluid Flows: A Dynamical Systems Approach*. Imperial College Press, London, UK.
- Nguyen, H., Gunzburger, M., Ju, L., and Burkardt, J. (2009). “Adaptive anisotropic meshing for steady convection-dominated problems.” *Computer Methods in Applied Mechanics and Engineering*, 198, 2964–2981.
- Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. Springer Verlag, New York, USA.
- Nordbotten, J. M., Aavatsmark, I., and Eigestad, G. T. (2007). “Monotonicity of control volume methods.” *Numerische Mathematik*, 106, 255–288.

- Noszticzius, Z., Bodnar, Z., Garamszegi, L., and Wittmann, M. (1991). “Hydrodynamic turbulence and diffusion-controlled reactions: Simulation of the effect of stirring on the oscillating Belousov–Zhabotinskii reaction with the Radicalator model.” *The Journal of Physical Chemistry*, 95, 6575–6580.
- Owen, S. J. (1998). “A survey of unstructured mesh generation technology.” *7th International Meshing Roundtable, Sandia National Laboratory*. 239–267.
- Pao, C. V. (1993). *Nonlinear Parabolic and Elliptic Equations*. Springer-Verlag, New York, USA.
- Pao, C. V., Zhou, L., and Jin, X. J. (1985). “Multiple solutions of a boundary–value problem in enzyme kinetics.” *Advances in Applied Mathematics*, 6, 209–229.
- Payette, G. S., Nakshatrala, K. B., and Reddy, J. N. (2012). “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.
- Pikovsky, A. and Popovych, O. (2003). “Persistent patterns in deterministic mixing flows.” *EPL (Europhysics Letters)*, 61, 625–631.
- Pikovsky, A., Rosenblum, M., and Kurths, J. (2001). *Synchronization: A Universal Concept in Nonlinear Sciences*, Vol. 12 of *Cambridge Nonlinear Science Series*. Cambridge University Press, Cambridge, UK.
- Pinder, G. F. and Celia, M. A. (2006). *Subsurface Hydrology*. John Wiley & Sons, Inc., New Jersey, USA.
- Pólya, G. (2009). *Mathematical Discovery on Understanding, Learning, and Teaching Problem Solving*, Vol. 1. Ishi Press.

- Poppe, D. and Lustfeld, H. (1997). “Nonlinearities in the gas phase chemistry of the troposphere: Oscillating concentrations in a simplified mechanism.” *Journal of Geophysical Research: Atmospheres (1984–2012)*, 101, 14373–14380.
- Porru, G. and Serra, S. (1994). “Maximum principles for parabolic equations.” *Journal of the Australian Mathematical Society (Series A)*, 56, 41–52.
- Potier, C. L. (2005). “Finite volume monotone scheme for highly anisotropic diffusion operators on unstructured triangular meshes.” *Comptes Rendus Mathematique*, 341, 787–792.
- Potier, C. L. (2009). “A nonlinear finite volume scheme satisfying maximum and minimum principles for diffusion operators.” *International Journal of Finite Volumes*, 6(2).
- Prestel, A. and Delzell, C. N. (2001). *Positive Polynomials: From Hilbert’s 17th Problem to Real Algebra*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, Heidelberg, Germany.
- Pucci, P. and Serrin, J. (2007). *The Maximum Principle*. Birkhäuser Verlag, Basel, Switzerland.
- Quarteroni, A., Sacco, R., and Saleri, F. (2006). *Numerical Mathematics*. Springer-Verlag, New York, USA.
- Rank, E., Katz, C., and Werner, H. (1983). “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.
- Raviart, P. A. and Thomas, J. M. (1977). “A mixed finite element method for 2nd order elliptic problems.” *Mathematical Aspects of the Finite Element Method*, Springer-Verlag, New York, USA. 292–315.

- Reznick, B. (2000). “Some concrete aspects of Hilbert’s 17th problem.” *Contemporary Mathematics*, 253, 251–272.
- Rice, S. A. (1985). *Diffusion-Limited Reactions*, Vol. 25 of *Comprehensive Chemical Kinetics*, edited by C. H. Bamford and C. F. H. Tipper and R. G. Compton. Elsevier Science Publishers B.V., Amsterdam, The Netherlands.
- Rudin, W. (1976). *The Principles of Mathematical Analysis*. McGraw-Hill, New York, USA, third edition.
- Rüst, L. Y. (2007). “The P -Matrix Linear Complementarity Problem: Generalizations and Specializations,” PhD thesis, ETH Zürich, Switzerland.
- Saltzman, W. M. (2001). *Drug Delivery: Engineering Principles for Drug Therapy*. Oxford University Press, New York, USA.
- Schneider, R. (2013). “A review of anisotropic refinement methods for triangular meshes in FEM.” *Advanced Finite Element Methods and Applications*, T. Apel and O. Steinbach, eds., Vol. 66 of *Lecture Notes in Applied and Computational Mechanics*, Berlin, Heidelberg. Springer-Verlag, 133–152.
- Shewchuk, J. R. (1996). “Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator.” *Applied Computational Geometry: Towards Geometric Engineering*, M. C. Lin and D. Manocha, eds., Vol. 1148 of *Lecture Notes in Computer Science*, Springer-Verlag, 203–222. From the First ACM Workshop on Applied Computational Geometry.
- G. C. Sih, J. Michopoulos, and S. C. Chou, eds. (1986). *Hygrothermoelasticity*. Martinus Nijhoff Publishers, Dordrecht, The Netherlands.
- Stynes, M. (2013). “Numerical methods for convection–diffusion problems or the 30 years war.” *arXiv:1306.5172*.

- Tartakovsky, A. M., Tartakovsky, D. M., Scheibe, T. D., and Meakin, P. (2008). “Hybrid simulations of reaction–diffusion systems in porous media.” *SIAM Journal on Scientific Computing*, 30, 2799–2816.
- Therrien, R. and Sudicky, E. A. (1996). “Three-dimensional analysis of variably-saturated flow and solute transport in discretely-fractured porous media.” *Journal of Contaminant Hydrology*, 23, 1–44.
- Thomas, R. H. and Zhou, Z. (1998). “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.
- Tsang, Y. K. (2009). “Predicting the evolution of fast chemical reactions in chaotic flows.” *Physical Review E*, 80, 026305(8).
- Vainberg, M. M. (1964). *Variational Methods for the Study of Nonlinear Operators*. Holden-Day, Inc., San Francisco, USA.
- Valocchi, A. J. (1989). “Spatial moment analysis of the transport of kinetically adsorbing solutes through stratified aquifers.” *Water Resources Research*, 25, 273–279.
- Vanderzee, E., Hirani, A. N., Guoy, D., and Ramos, E. A. (2010). “Well-centered triangulation.” *SIAM Journal on Scientific Computing*, 31, 4497–4523.
- Varga, R. (1966). “On a discrete maximum principle.” *SIAM Journal on Numerical Analysis*, 3, 355–359.
- Varga, R. S. (2009). *Matrix Iterative Analysis*, Vol. 27 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Heidelberg, second revised and expanded edition.

- Vejchodský, T. (2010). *Angle conditions for discrete maximum principles in higher-order FEM*, 901–909. Numerical Mathematics and Advanced Applications 2009. Springer-Verlag.
- Šolín, P. and Vejchodský, T. (2007). “A weak discrete maximum principle for *hp*-FEM.” *Journal of Computational and Applied Mathematics*, 209, 54–65.
- Wang, S., Yuan, G., Li, Y., and Sheng, Z. (2012). “Discrete maximum principle based on repair technique for diamond type scheme of diffusion problems.” *International Journal for Numerical Methods in Fluids*, 70, 1188–1205.
- Wathen, A. J. (1989). “An analysis of some element-by-element techniques.” *Computer Methods in Applied Mechanics and Engineering*, 74, 271–287.
- J. B. Weiss and A. Provenzale, eds. (2008). *Transport and Mixing in Geophysical Flows*, Vol. 744 of *Lecture Notes in Physics*. Springer, Berlin, Heidelberg, Germany.
- Willingham, T. W., Werth, C. J., and Valocchi, A. J. (2008). “Evaluation of the effects of the porous media structure on mixing-controlled reactions using pore-scale modeling and micromodel experiments.” *Environmental Science & Technology*, 42, 3185–3193.
- Wolfram, S. (2013). *Mathematica, Version 9.0.1 (computer software)*. Wolfram Research, Inc., Champaign, Illinois, USA.
- Wood, W. L. (1990). *Practical Time-Stepping Schemes*. Oxford University Press, New York, USA.
- Xin, J. (2009). *An Introduction to Fronts in Random Media*, Vol. 5 of *Survey and Tutorials in the Applied Mathematical Sciences*. Springer Science+Business Media, LLC, New York, USA.

- R. N. Yong and H. R. Thomas, eds. (1997). *Geoenvironmental Engineering: Contaminated Ground: Fate of Pollutants and Remediation*. Thomas Telford Ltd, London, UK.
- Zhao, X., Chen, Y., Gao, Y., Yu, C., and Li, Y. (DOI: 10.1002/flid.3838, 2013). “Finite volume element methods for nonequilibrium radiation diffusion equations.” *International Journal for Numerical Methods in Fluids*.
- Zheng, C. and Bennett, G. D. (2002). *Applied Contaminant Transport Modeling*. John Wiley & Sons, Inc., New York, USA, second edition.
- Zienkiewicz, O. C., Taylor, R. L., and Zhu, J. Z. (2013). *The Finite Element Method: Its Basis and Fundamentals*. Butterworth-Heinemann, Elsevier Ltd., Oxford, UK, seventh edition.

APPENDIX: ELEMENT-LEVEL DISCRETIZATION OF STABILIZATION TERMS

The weighted negatively stabilized streamline diffusion least-squares formulation requires the evaluation of $\text{div}[\text{grad}[c(\mathbf{x})]]$ and $\text{grad}[\text{grad}[c(\mathbf{x})]]$ terms at the element-level. Since these terms are not typical, we present a compact way of discretizing these terms under the finite element method. It should be noted that one need not evaluate these terms for lowest-order simplicial elements (i.e., three-node triangular element and four-node tetrahedron element), as these terms will be identically zero. However, $\text{div}[\text{grad}[c(\mathbf{x})]]$ and $\text{grad}[\text{grad}[c(\mathbf{x})]]$ will be non-zero for high-order simplicial and non-simplicial elements (i.e., four-node quadrilateral element and eight-node brick element).

FOURTH-ORDER TENSORS

Let \mathbf{R} and \mathbf{S} be two second-order tensors. A fourth-order tensor product $\mathbf{R} \boxtimes \mathbf{S}$ is defined as

$$(\mathbf{R} \boxtimes \mathbf{S}) \mathbf{T} = \mathbf{R} \mathbf{T} \mathbf{S}^T \quad (.0.1)$$

for any second-order tensor \mathbf{T} . The fourth-order identity tensor can then be written as

$$\mathbb{I} = \mathbf{I} \boxtimes \mathbf{I}, \quad (.0.2)$$

where \mathbf{I} is the second-order identity tensor. The fourth-order transposer \mathbb{T} and symmetrizer \mathbb{S} tensors are defined as

$$\mathbb{T}\mathbf{A} = \mathbf{A}^T \text{ and} \quad (.0.3a)$$

$$\mathbb{S}\mathbf{A} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T), \quad (.0.3b)$$

where \mathbf{A} is a general second-order tensor.

PROPERTIES OF KRONECKER PRODUCTS RELEVANT TO FINITE ELEMENT DISCRETIZATION

Let \mathbf{A} and \mathbf{B} be matrices of sizes $n \times m$ and $p \times q$, respectively. Let a_{ij} and b_{ij} be, respectively, the ij^{th} components of \mathbf{A} and \mathbf{B} . The *Kronecker product* of these two matrices is an $np \times mq$ matrix that is defined as

$$\mathbf{A} \odot \mathbf{B} := \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1m}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & \dots & a_{nm}\mathbf{B} \end{bmatrix}. \quad (.0.4)$$

Another operator that will be useful is the $\text{vec}[\bullet]$ operator, which is defined as

$$\text{vec}[\mathbf{A}] := \begin{bmatrix} a_{11} \\ \vdots \\ a_{1m} \\ \vdots \\ a_{n1} \\ \vdots \\ a_{nm} \end{bmatrix}. \quad (.0.5)$$

The following standard properties can be established for $\text{mat}[\bullet]$ and $\text{vec}[\bullet]$ operators:

$$\text{vec}[\mathbf{A} + \mathbf{B}] = \text{vec}[\mathbf{A}] + \text{vec}[\mathbf{B}], \quad (.0.6)$$

$$(\mathbf{A} \odot \mathbf{B})(\mathbf{C} \odot \mathbf{D}) = (\mathbf{AC} \odot \mathbf{BD}), \text{ and} \quad (.0.7)$$

$$\text{vec}[\mathbf{ACB}] = (\mathbf{B}^T \odot \mathbf{A}) \text{vec}[\mathbf{C}]. \quad (.0.8)$$

For general properties of Kronecker products, see the book by Graham Graham (1981). However, this book does not contain many of the results presented below, which are useful for the current study.

For an effective computer implementation of LSFEM-based formulations, we shall represent four-dimensional and three-dimensional arrays as two-dimensional matrices. To this end, consider a four-dimensional array \mathbf{P} of size $m \times n \times p \times q$. The matrix representation of \mathbf{P} is denoted by $\text{mat}[\bullet]$, and is defined as

$$\text{mat}[\mathbf{P}] := \begin{bmatrix} P_{1111} & \dots & P_{11p1} & P_{1112} & \dots & P_{111q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ P_{mn11} & \dots & P_{mnp1} & P_{mn12} & \dots & P_{mnpq} \end{bmatrix}. \quad (.0.9)$$

The following properties can be established for the matrix representation of fourth-order tensors:

$$\text{vec}[\mathbf{PX}] = \text{mat}[\mathbf{P}]\text{vec}[\mathbf{X}], \quad (.0.10)$$

$$\text{mat}[\mathbf{A} \boxtimes \mathbf{B}] = \mathbf{B} \odot \mathbf{A}, \text{ and} \quad (.0.11)$$

$$\text{mat}[\mathbb{S}] = \frac{1}{2}(\mathbf{I} \odot \mathbf{I} + \text{mat}[\mathbb{T}]), \quad (.0.12)$$

where the matrix representation of \mathbb{T} takes the following form:

$$\text{mat}[\mathbb{T}] := \begin{bmatrix} \mathbf{I} \odot [1, 0, 0, \dots, 0]_n \\ \mathbf{I} \odot [0, 1, 0, \dots, 0]_n \\ \vdots \\ \mathbf{I} \odot [0, 0, 0, \dots, 1]_n \end{bmatrix}_{mn \times mn} . \quad (.0.13)$$

In the above equation, \mathbf{I} is an identity matrix of size $m \times m$. It can be shown that for a matrix \mathbf{Z} of size $m \times n$, $\text{vec}[\mathbf{Z}^T] = \text{mat}[\mathbb{T}]\text{vec}[\mathbf{Z}]$.

For a three-dimensional arrays, there are two useful matrix representations, which will be denoted as $\text{mat}_1[\bullet]$ and $\text{mat}_2[\bullet]$. Consider a three-dimensional array \mathbf{Q} of size $m \times n \times p$. The corresponding matrix representations of \mathbf{Q} are defined as

$$\text{mat}_1[\mathbf{Q}] := \begin{bmatrix} Q_{111} & \dots & Q_{1n1} & \dots & \dots & Q_{11p} & \dots & Q_{1np} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Q_{m11} & \dots & Q_{mn1} & \dots & \dots & Q_{m1p} & \dots & Q_{mnp} \end{bmatrix} \text{ and} \quad (.0.14)$$

$$\text{mat}_2[\mathbf{Q}] := \begin{bmatrix} Q_{111} & \dots & Q_{m11} \\ \vdots & \ddots & \vdots \\ Q_{11p} & \dots & Q_{m1p} \\ Q_{121} & \dots & Q_{m21} \\ \vdots & \ddots & \vdots \\ Q_{12p} & \dots & Q_{m2p} \\ \vdots & \ddots & \vdots \\ Q_{1np} & \dots & Q_{mnp} \end{bmatrix} . \quad (.0.15)$$

The following properties can be established for the matrix representations of three-dimensional arrays:

$$\text{vec}[\mathbf{QY}] = \text{mat}_1[\mathbf{Q}]\text{vec}[\mathbf{Y}] \text{ and} \quad (.0.16)$$

$$\text{vec}[\mathbf{Qz}] = \text{mat}_2[\mathbf{Q}]\text{vec}[\mathbf{z}]. \quad (.0.17)$$

FINITE ELEMENT DISCRETIZATION OF DIV[GRAD[C(**X**)] AND GRAD[GRAD[C(**X**)] TERMS

Let $\boldsymbol{\xi}$ denote the position vector in the reference finite element. The row vector containing the shape functions is denoted by \mathbf{N} , which is a function of $\boldsymbol{\xi}$. The derivatives and Hessian of \mathbf{N} with respect to $\boldsymbol{\xi}$ are, respectively, denoted as \mathbf{DN} and \mathbf{DDN} . That is, in indicial notation we have:

$$(\mathbf{DN})_{ij} = \frac{\partial N_i}{\partial \xi_j} \text{ and} \quad (.0.18a)$$

$$(\mathbf{DDN})_{ijk} = \frac{\partial^2 N_i}{\partial \xi_j \partial \xi_k}, \quad (.0.18b)$$

where N_i and ξ_i are, respectively, the i^{th} -component of the vectors \mathbf{N} and $\boldsymbol{\xi}$. The concentration field $c(\mathbf{x})$ and total flux vector $\mathbf{q}(\mathbf{x})$ are interpolated as

$$c(\mathbf{x}) = \hat{\mathbf{c}}^T \mathbf{N}^T(\boldsymbol{\xi}(\mathbf{x})) \text{ and} \quad (.0.19a)$$

$$\mathbf{q}(\mathbf{x}) = \hat{\mathbf{q}}^T \mathbf{N}^T(\boldsymbol{\xi}(\mathbf{x})), \quad (.0.19b)$$

where $\hat{\mathbf{c}}^T$ and $\hat{\mathbf{q}}^T$ denote the matrix containing nodal concentration and total flux vector. In negatively stabilized streamline diffusion LSFEM, $\text{div}[\text{grad}[c(\mathbf{x})]]$ and

$\text{grad}[\text{grad}[c(\mathbf{x})]]$ terms arise from the following expansions:

$$\text{div}[D(\mathbf{x})\text{grad}[c(\mathbf{x})]] = \text{grad}[D(\mathbf{x})] \bullet \text{grad}[c(\mathbf{x})] + D(\mathbf{x}) \text{div}[\text{grad}[c(\mathbf{x})]] \text{ and } \quad (.0.20a)$$

$$\text{div}[\mathbf{D}(\mathbf{x})\text{grad}[c(\mathbf{x})]] = \text{div}[\mathbf{D}(\mathbf{x})] \bullet \text{grad}[c(\mathbf{x})] + \mathbf{D}(\mathbf{x}) \bullet \text{grad}[\text{grad}[c(\mathbf{x})]]. \quad (.0.20b)$$

Based on the regularity of diffusivity, $\text{grad}[D(\mathbf{x})]$ and $\text{div}[\mathbf{D}(\mathbf{x})]$ can be calculated in multiple ways. If the diffusivity is *continuously differentiable*, then $\text{grad}[D(\mathbf{x})]$ and $\text{div}[\mathbf{D}(\mathbf{x})]$ can be directly evaluated analytically. This will considerably reduce the computational cost in the evaluation of local stiffness matrices. If $D(\mathbf{x})$ is not differentiable (but is square integrable), then $\text{grad}[D(\mathbf{x})]$ and $\text{div}[\mathbf{D}(\mathbf{x})]$ can be evaluated as

$$\text{grad}[D(\mathbf{x})] = \widehat{\mathbf{D}}^T (\mathbf{D}\mathbf{N})\mathbf{J}^{-1} \text{ and} \quad (.0.21a)$$

$$\text{div}[\mathbf{D}(\mathbf{x})] = \text{mat}_1[\widehat{\mathbf{D}}]\text{vec}[\text{mat}[\mathbb{T}](\mathbf{D}\mathbf{N})\mathbf{J}^{-1}], \quad (.0.21b)$$

where \mathbf{J} is the Jacobian matrix, and $\widehat{\mathbf{D}}$ is the nodal values for the diffusivity, whose size can be inferred based on the context (whether the diffusivity is either $D(\mathbf{x})$ or $\mathbf{D}(\mathbf{x})$). Using equations (.0.18) and (.0.19), the Laplacian and Hessian of $c(\mathbf{x})$ can be calculated as

$$\text{div}[\text{grad}[c(\mathbf{x})]] = \left(\left(\mathbf{I} - (\mathbf{D}\mathbf{N})\mathbf{J}^{-1}\widehat{\mathbf{x}}^T \right) \text{mat}_1[\mathbf{D}\mathbf{D}\mathbf{N}]\text{vec}[\mathbf{J}^{-1}\mathbf{J}^{-T}] \right)^T \text{vec}[\widehat{\mathbf{c}}^T] \text{ and} \quad (.0.22)$$

$$\text{vec}[\text{grad}[\text{grad}[c(\mathbf{x})]]] = \left(\left(\mathbf{J}^{-T} \odot \mathbf{J}^{-T} \right) \text{mat}_2[\mathbf{D}\mathbf{D}\mathbf{N}] \left(\mathbf{I} \odot \mathbf{I} - \widehat{\mathbf{x}}\mathbf{J}^T(\mathbf{D}\mathbf{N})^T \right) \right) \text{vec}[\widehat{\mathbf{c}}^T], \quad (.0.23)$$

where $\widehat{\mathbf{x}}$ is the matrix containing nodal coordinates.