Stabilized Finite Element Methods for Convection-Diffusion-Reaction, Helmholtz and Stokes Problems

P. Nadukandi
E. Oñate
J. García-Espinosa

Monograph CIMNE Nº-130, July 2012

Stabilized Finite Element Methods for Convection-Diffusion-Reaction, Helmholtz and Stokes Problems

P. Nadukandi
E. Oñate
J. García-Espinosa

Monograph CIMNE Nº-130, July 2012

International Center for Numerical Methods in Engineering
Gran Capitán s/n, 08034 Barcelona, Spain

INTERNATIONAL CENTER FOR NUMERICAL METHODS IN ENGINEERING
Edificio C1, Campus Norte UPC
Gran Capitán s/n
08034 Barcelona, Spain
www.cimne.com

First edition: March 2012

HIGH-PERFORMANCE MODEL REDUCTION PROCEDURES IN MULTISCALE SIMULATIONS
Monograph CIMNE M127

ISBN: 978-84-940243-3-7

Depósito legal: B-23420-2012

ABSTRACT

We present three new stabilized finite element (FE) based Petrov–Galerkin methods for the convection–diffusion–reaction (CDR), the Helmholtz and the Stokes problems, respectively. The work embarks upon a priori analysis of some consistency recovery procedures for some stabilization methods belonging to the Petrov–Galerkin framework. It was found that the use of some standard practices (e. g.M-Matrices theory) for the design of essentially non-oscillatory numerical methods is not appropriate when consistency recovery methods are employed. Hence, with respect to convective stabilization, such recovery methods are not preferred. Next, we present the design of a high-resolution Petrov–Galerkin (HRPG) method for the CDR problem. The structure of the method in 1D is identical to the consistent approximate upwind (CAU) Petrov–Galerkin method [68] except for the definitions of the stabilization parameters. Such a structure may also be attained via the Finite Calculus (FIC) procedure [141] by an appropriate definition of the characteristic length. The prefix *high-resolution* is used here in the sense popularized by Harten, i. e.second order accuracy for smooth/regular regimes and good shock-capturing in non-regular regimes. The design procedure in 1D embarks on the problem of circumventing the Gibbs phenomenon observed in L^2 projections. Next, we study the conditions on the stabilization parameters to circumvent the global oscillations due to the convective term. A conjuncture of the two results is made to deal with the problem at hand that is usually plagued by Gibbs, global and dispersive oscillations in the numerical solution. A multi dimensional extension of the HRPG method using multi-linear block finite elements is also presented.

Next, we propose a higher-order compact scheme (involving two parameters) on structured meshes for the Helmholtz equation. Making the parameters equal, we recover the alpha-interpolation of the Galerkin finite element method (FEM) and the classical central finite difference method. In 1D this scheme is identical to the alpha-interpolation method [140] and in 2D choosing the value 0.5 for both the parameters, we recover the generalized fourth-order compact Padé approximation [81, 168] (therein using the parameter $\gamma = 2$). We follow [10] for the analysis of this scheme and its performance on square meshes is compared with that of the quasi-stabilized FEM [10]. Generic expressions for the parameters are given that guarantees a dispersion accuracy of sixth-order should the parameters be distinct and fourth-order should they be equal. In the later case, an expression for the parameter is given that minimizes the maximum relative phase error in 2D. A Petrov–Galerkin formulation that yields the aforesaid scheme on structured meshes is also presented. Convergence studies of the error in the L^2 norm, the H^1 semi-norm and the l^∞ Euclidean norm is done and the pollution effect is found to be small.

Finally, we present a collection of stabilized FE methods derived via first-order and second-order FIC procedures for the Stokes problem. It is shown that several well known existing stabilized FE methods such as the penalty technique, the Galerkin Least Square (GLS) method [93], the Pressure Gradient Projection (PGP) method [35] and the orthogonal sub-scales (OSS) method [34] are recovered from the general

residual-based FIC stabilized form. A new family of Pressure Laplacian Stabilization (PLS) FE methods with consistent nonlinear forms of the stabilization parameters are derived. The distinct feature of the family of PLS methods is that they are nonlinear and residual-based, i. e.the stabilization terms depend on the discrete residuals of the momentum and/or the incompressibility equations. The advantages and disadvantages of these stabilization techniques are discussed and several examples of application are presented.

KEYWORDS : Convection–diffusion–reaction problem, Helmholtz equation, Stokes flow, Stabilized finite element methods, High-resolution Petrov–Galerkin method, linear interpolation of FEM and FDM stencils, Pressure Laplacian Stabilization, Dispersion analysis.

RESUMEN

Presentamos tres nuevos métodos estabilizados de tipo Petrov–Galerkin basado en elementos finitos (FE) para los problemas de convección–difusión–reacción (CDR), de Helmholtz y de Stokes, respectivamente. El trabajo comienza con un análisis a priori de un método de recuperación de la consistencia de algunos métodos de estabilización que pertenecen al marco de Petrov–Galerkin. Hallamos que el uso de algunas de las prácticas estándar (por ejemplo, la teoría de Matriz-M) para el diseño de métodos numéricos esencialmente no oscilatorios no es apropiado cuando utilizamos los métodos de recuperación de la consistencia. Por lo tanto, con respecto a la estabilización de convección, no preferimos tales métodos de recuperación. A continuación, presentamos el diseño de un método de Petrov–Galerkin de alta-resolución (HRPG) para el problema CDR. La estructura del método en 1D es idéntico al método CAU [68] excepto en la definición de los parámetros de estabilización. Esta estructura también se puede obtener a través de la formulación del cálculo finito (FIC) [141] usando una definición adecuada de la longitud característica. El prefijo de *alta-resolución* se utiliza aquí en el sentido popularizado por Harten, es decir, tener una solución con una precisión de segundo orden en los regímenes suaves y ser esencialmente no oscilatoria en los regímenes no regulares. El diseño en 1D se embarca en el problema de eludir el fenómeno de Gibbs observado en las proyecciones de tipo L^2. A continuación, estudiamos las condiciones de los parámetros de estabilización para evitar las oscilaciones globales debido al término convectivo. Combinamos los dos resultados (una conjetura) para tratar el problema CDR, cuya solución numérica sufre de oscilaciones numéricas del tipo global, Gibbs y dispersiva. También presentamos una extensión multidimensional del método HRPG utilizando los elementos finitos multi-lineales.

A continuación, proponemos un esquema compacto de orden superior (que incluye dos parámetros) en mallas estructuradas para la ecuación de Helmholtz. Haciendo igual ambos parámetros, se recupera la interpolación lineal del método de elementos finitos (FEM) de tipo Galerkin y el clásico método de diferencias finitas centradas. En 1D este esquema es idéntico al método AIM [140] y en 2D eligiendo el valor de 0.5 para ambos parámetros, se recupera el esquema compacto de cuarto orden de Padé generalizada en [81, 168] (con el parámetro $\gamma = 2$). Seguimos [10] para el análisis de este

esquema y comparamos su rendimiento en las mallas uniformes con el de *FEM cuasi-estabilizado* (QSFEM) [10]. Presentamos expresiones genéricas de los parámetros que garantiza una precisión dispersiva de sexto orden si ambos parámetros son distintos y de cuarto orden en caso de ser iguales. En este último caso, presentamos la expresión del parámetro que minimiza el error máximo de fase relativa en 2D. También proponemos una formulación de tipo Petrov–Galerkin que recupera los esquemas antes mencionados en mallas estructuradas. Presentamos estudios de convergencia del error en la norma de tipo L^2, la semi-norma de tipo H^1 y la norma Euclidiana tipo l^∞ y mostramos que la pérdida de estabilidad del operador de Helmholtz (*pollution effect*) es incluso pequeña para grandes números de onda.

Por último, presentamos una colección de métodos FE estabilizado para el problema de Stokes desarrollados a través del método FIC de primer orden y de segundo orden. Mostramos que varios métodos FE de estabilización existentes y conocidos como el método de penalización, el método de Galerkin de mínimos cuadrados (GLS) [93], el método PGP (estabilizado a través de la proyección del gradiente de presión) [35] y el método OSS (estabilizado a través de las sub-escalas ortogonales) [34] se recuperan del marco general de FIC. Desarrollamos una nueva familia de métodos FE, en adelante denominado como PLS (estabilizado a través del Laplaciano de presión) con las formas no lineales y consistentes de los parámetros de estabilización. Una característica distintiva de la familia de los métodos PLS es que son no lineales y basados en el residuo, es decir, los términos de estabilización dependerá de los residuos discretos del momento y/o las ecuaciones de incompresibilidad. Discutimos las ventajas y desventajas de estas técnicas de estabilización y presentamos varios ejemplos de aplicación.

Palabras clave : El problema de convección–difusión–reacción, La ecuación de Helmholtz, El problema de Stokes, Métodos de elementos finitos estabilizado, El método de Petrov–Galerkin con alta-resolución, Interpolación lineal de las esténciles del MEF y del MDF, Cálculo finito, Análisis de dispersión numérica.

Every morning in Africa, a gazelle wakes up.
It knows it must run faster than the fastest lion or it will be killed.
Every morning in Africa, a lion wakes up.
It knows it must outrun the slowest gazelle or it will starve to death.
It doesn't matter whether you are a lion or a gazelle;
when the sun comes up, you'd better be running.

— Herbert Eugene Caen.

PUBLICATIONS

This monograph contains some unpublished material, but is mainly based on the following publications wherein some ideas and figures have appeared previously.

I Prashanth Nadukandi, Eugenio Oñate, and Julio García. Analysis of a consistency recovery method for the 1D convection–diffusion equation using linear finite elements. *International Journal for Numerical Methods in Fluids*, 57(9):1291–1320, July 2008. ISSN 02712091. doi: 10.1002/fld.1863.

II Prashanth Nadukandi, Eugenio Oñate, and Julio García. A high-resolution Petrov–Galerkin method for the 1D convection–diffusion–reaction problem. *Computer Methods in Applied Mechanics and Engineering*, 199(9–12):525–546, January 2010. ISSN 00457825. doi: 10.1016/j.cma.2009.10.009.

III Prashanth Nadukandi, Eugenio Oñate, and Julio García. A fourth-order compact scheme for the Helmholtz equation: Alpha-interpolation of FEM and FDM stencils. *International Journal for Numerical Methods in Engineering*, 86(1):18–46, April 2011. ISSN 00295981. doi: 10.1002/nme.3043.

IV Eugenio Oñate, Prashanth Nadukandi, Sergio R. Idelsohn, Julio García and Carlos A. Felippa. A family of residual-based stabilized finite element methods for Stokes flows. *International Journal for Numerical Methods in Fluids*, 65(1–3):106–134, January 2011. ISSN 02712091. doi: 10.1002/fld.2468.

V Prashanth Nadukandi, Eugenio Oñate, and Julio García. A Petrov–Galerkin formulation for the alpha-interpolation of FEM and FDM stencils. Applications to the Helmholtz equation. *International Journal for Numerical Methods in Engineering*, 89(11):1367–1391, March 2012. ISSN 00295981. doi: 10.1002/nme.3291.

VI Prashanth Nadukandi, Eugenio Oñate, and Julio García. A high-resolution Petrov–Galerkin method for the convection–diffusion–reaction problem. Part II—A multidimensional extension. *Computer Methods in Applied Mechanics and Engineering*, 213–216:327–352, March 2012. ISSN 00457825. doi: 10.1016/j.cma.2011.10.003.

When eating bamboo sprouts, remember the man who planted them

— a Chinese Proverb

ACKNOWLEDGMENTS

The authors thanks Profs. Ramon Codina, Sergio Idelsohn, Carlos Felippa, Assad Oberai for many useful discussions.

The first author acknowledges the economic support received through the FI predoctoral grant (ref. 2006FI 00836) from the *Department of Universities, Research and Information Society* (Generalitat de Catalunya) and the *European Social Fund*.

This work was partially supported by the SAFECON project of the European Research Council (European Commission).

CONTENTS

LIST OF TABLES

Let auspicious thoughts come unto us from every direction.

— Rig Veda I:89.1.

INTRODUCTION

This monograph consists of three parts. The first part describes the work done on the convection–diffusion–reaction problem, the second one deals with the Helmholtz problem and the third one deals with the Stokes problem. In each part we propose a new stabilized finite element based Petrov–Galerkin method for the corresponding problem. The work presented in each part is independent from the rest of the parts and hence can be read arbitrarily.

The motivation for the current work is in the quest to develop new numerical methods capable of providing stable and accurate solutions to the problems of fluid mechanics and their interaction with structures. The work presented here is the first step towards this objective. This is done by the study of the convection–diffusion–reaction, the Helmholtz and the Stokes problems. These problems are of vital importance as they are the simplest models related to transport processes, wave propagation phenomenon and associated numerical difficulties that arise in fluid flow problems and fluid-structure interaction. The convection–diffusion–reaction equation is an ideal model problem to study the stabilization of singularly perturbed problems. Ten typical problems where the convection–diffusion phenomenon occurs is listed in [134]. For instance, a common source is the linearization of Navier–Stokes equations with large Reynolds number. The Helmholtz equation is an ideal model problem to study the so-called *pollution effect*. It is also the simplest model problem concerning wave propagation phenomenon, viz. acoustics, elastodynamics, fluid-structure interaction, electrodynamics etc. [106]. A simple one-dimensional fluid-structure interaction problem modeled using the Helmholtz equation was presented in [47, Section 5]. The Stokes problem is the simplest model describing incompressible flow of a viscous fluid and thus are ideal to study pressure stabilization, i. e.circumventing the *div-stability* condition (also known as the *Ladyzhenskaya-Babuska-Brezzi* condition) that the finite element spaces should otherwise satisfy. We refer to [75, Chapter 2] for an exposition of the same.

The point of departure was initially set to the analysis of an approach that has been gaining momentum in the literature: *consistency recovery procedures*[1] for lower-order stabilized finite element methods. For residual-based stabilization methods, the higher-order derivatives of the residual that appear in the stabilization terms vanish when lower-order finite elements are used. Consistency recovery methods have been advocated for these cases and have been shown to result in improved accuracy for some problems [111, 148]. To be precise, the improved accuracy was reported for other related unknowns of the problem (e. g.the pressure field) and not for the transported unknown (e. g.the velocity field). The first chapter of this monograph discusses the gain/loss by recovering the consistency of the discrete residual in the stabilization terms via the form that includes the convective projection variable (as in the OSS

1 also know as *residual correction methods*

method [34]) for the 1D convection-diffusion equation. The dispersion analysis of the semi-discrete and fully-discrete problems was done and no gain in the dispersion accuracy is found by including the convective projection variable. Further, the optimal expression of the stabilization parameter on uniform meshes for the steady-state case and including the convective projection variable revealed a strategical difficulty in the development of discontinuity capturing methods. Hence, we do not prefer this consistency recovery method in the stabilization of singularly perturbed problems.

In the second chapter we present the design of a high-resolution Petrov–Galerkin (HRPG) method for the 1D convection–diffusion–reaction problem. The problem is studied from a fresh point of view, including practical implications on the formulation of the maximum principle, M-Matrices theory, monotonicity and total variation diminishing (TVD) finite volume schemes. The prefix high-resolution is used here in the sense popularized by Harten [82] in the finite-difference and finite-volume community, i. e.second order accuracy for smooth/regular regimes and good shock-capturing in non-regular regimes. The HRPG method is designed using the divide and conquer strategy, i. e.the original problem is further divided into smaller model problems where the different types of numerical artifacts that plague the original problem are singled-out and the expressions for the stabilization parameters are derived/updated to treat them effectively. Several 1D examples are presented that support the design objective—stabilization with high-resolution.

In the third chapter we present a multi dimensional extension of the HRPG method using multi-linear block finite elements. In higher dimensions the solutions to the convection–diffusion–reaction problem might additionally develop characteristic layers. These layers are a unique feature of multi dimensions and hence have no instances in 1D. So we design a nondimensional element number that quantifies the characteristic layers. By quantification we mean that it should serve a similar purpose in the definition of the stabilization parameters as the element Peclet number does for the exponential layers. Although the structure of HRPG method in 1D is identical to the consistent approximate upwind Petrov–Galerkin method [68], in multi dimensions the former method has a unique structure. The distinction is that in general the upwinding is not streamline and the discontinuity-capturing is neither isotropic nor purely crosswind. In this line, we present anisotropic element length vectors and using them objective characteristic tensors associated with the HRPG method are defined. Except for the modification to include the new dimensionless number that quantifies the characteristic layers, the definition of the stabilization parameters calculated along the element length vectors are a direct extension of their counterparts in 1D. The strategy used to treat the artifacts about the characteristic layers is to treat them just like the artifacts found across the parabolic layers in the reaction-dominant case. Several 2D examples are presented that illustrate not only the advantages of the HRPG method but also its limitations. Of course, the advantages outscore the limitations.

The second part of this monograph deals with the work on the Helmholtz problem. Although this equation is related to some fluid-structure interaction problems, the events that lead to the work presented here had a modest kickoff. Chronologically, the work began just after the development of the HRPG method for the 1D convection–diffusion–reaction problem. As the properties of the Helmholtz equation is shared by the diffusion–production problem (obtained using a negative reaction co-efficient), we were curious to investigate if the HRPG form be efficient to solve this case. As

the HRPG form is nonlinear, one needs to choose a target solution to design the stabilization parameters so as to recover this target solution. In 1D the simplest target solution is the one that you get with the Galerkin FEM using a higher-order mass matrix [70, 71, 110]. This scheme has a dispersion accuracy of fourth-order in 1D but in higher dimensions it drops to second-order. Unfortunately all attempts to recover this target solution within the HRPG form were in vain. Although this exercise bore no fruit in favor of the HRPG method, we discovered alternate target solutions for the Helmholtz equation that had higher-order dispersion accuracy in multi dimensions.

In the fourth chapter we present some observations and related dispersion analysis of a simple domain-based higher-order compact numerical scheme (involving two parameters) for the Helmholtz equation. The stencil obtained by choosing the parameters as distinct was denoted therein as the 'nonstandard compact stencil'. Making the parameters equal, the nonstandard compact stencil simplifies to the alpha-interpolation of the equation stencils obtained by the Galerkin finite element method (FEM) and classical central finite difference method (FDM). For the Helmholtz equation, generic expressions for the parameters were given that guarantees a dispersion accuracy of sixth-order should the parameters be distinct and fourth-order should they be equal. Convergence studies of the relative error in the L^2 norm, the H^1 semi-norm and the l^∞ Euclidean norm are done and the pollution effect is found to be small.

In the fifth chapter we present a new Petrov–Galerkin method involving two parameters that yields on rectangular meshes the nonstandard compact stencil presented earlier in the fifth chapter. This Petrov–Galerkin method provides the counterparts of the later scheme on unstructured meshes and allows the treatment of natural boundary conditions (Neumann or Robin) and the source terms in a straight-forward manner. First, we present the test-functions that reproduce the mass matrix lumping technique within a Petrov–Galerkin setting. Then, we show that the use of these test functions in an appropriate variational setting reproduces the FDM stencil for the current problem on rectangular meshes. Next, we show that an appropriate combination of these test-functions with the standard FEM shape functions will yield the nonstandard compact stencil on rectangular meshes. Convergence studies of the relative error in the L^2 norm, the H^1 semi-norm and the l^∞ Euclidean norm show that the proposed Petrov–Galerkin method inherits the higher-order dispersion accuracy observed for the aforesaid numerical scheme.

The third and final part of this monograph deals with the work done on the Stokes problem. Here we present a collection of stabilized finite element (FE) methods derived via first-order and second-order finite calculus (FIC) procedures. It is shown that several well known existing stabilized FE methods such as the penalty technique, the Galerkin Least Square (GLS) method, the Pressure Gradient Projection (PGP) method and the orthogonal sub-scales (OSS) method are recovered from the general residual-based FIC stabilized form. A new family of Pressure Laplacian Stabilization (PLS) FE methods with consistent nonlinear forms of the stabilization parameters are derived. The distinct feature of the family of PLS methods is that they are nonlinear and residual-based, i. e.the stabilization terms depend on the discrete residuals of the momentum and/or the incompressibility equations. The advantages and disadvantages of these stabilization techniques are discussed and several examples of application are presented.

Part I

CONVECTION-DIFFUSION-REACTION PROBLEM

Experience is what you get when you don't get what you want.

— Dan Stanford.

1 ANALYSIS OF A CONSISTENCY RECOVERY METHOD

1.1 INTRODUCTION

In many transport processes arising in physical problems, convection essentially dominates diffusion. The design of numerical methods for such problems that reflect their almost hyperbolic nature and guarantee that the discrete solution satisfies the physical conditions is a subject that has been widely studied. In particular for the convection-diffusion problem the standard Galerkin finite element method leads to numerical instabilities for the convection dominated case. Several stabilization methods, for instance the Streamline-Upwind Petrov-Galerkin (SUPG), Galerkin Least Square (GLS), Sub-Grid Scale (SGS), SGS with Orthogonal Sub-scales (OSS) etc., have been designed to overcome this numerical instability. A thorough comparison of some of these methods from the point of view of their formulation and the motivations that lead to them can be found in [32]. Also stabilization procedures based on Finite Calculus (FIC) have been developed as a general purpose tool for improving the stability and accuracy of the convection-diffusion problem [141, 143, 145, 147]. A residual correction method based on FIC was presented in [148] and is shown to yield an equivalent formulation to an OSS form [34] with very little manipulation.

For the convection diffusion problem using the SUPG or FIC methods, the higher order term (here the diffusion term) that appears in the stabilization term vanish when simplicial elements are used. In [148] it is shown that for the elasticity problem, the form that includes the projected gradient of pressure into the stabilization terms (motivated by the OSS method) is essential to obtain accurate numerical results which converge in a more monotone manner and are less sensitive to the value of the stabilization parameter. In [111] a method was presented to globally reconstruct a continuous approximation to the diffusive flux for linear finite elements using a L^2 projection and shown, in some cases (when advection and diffusion are on a par), to greatly improve accuracy. It is important to note that again this improved accuracy is demonstrated for other related unknowns of the problem (like pressure) and not for the transported unknown (say velocity). Also when convection dominates diffusion there is little effect in the inclusion of the recovered diffusive flux. On the other hand, consistency recovery following the OSS philosophy is independent from the diffusive term. These observations are the motivation to investigate in detail the benefits of including similar projections for the convection diffusion problem following the OSS philosophy. In other words, we try to answer the question - what do we gain/loose by recovering the consistency of the discrete residual in the stabilization terms for the convection-dominated case via the introduction of OSS-type convective projection ?

The von Neumann analysis for the Galerkin and SUPG method is relatively well known [72]. Relevant literature on the type of analysis presented here may be found

in [26]. First, we present the von Neumann analysis for the 1D FIC method with recovered consistency (FIC_RC). This is achieved by including the convective projection into the stabilization term (motivated by the OSS method). It is then shown that in 1D the FIC and FIC_RC methods are equivalent to the SUPG and OSS methods respectively. Consequently the comparisons made between the former methods may be carried over to the later methods. The transient analysis is done by examining the discrete dispersion relation (DDR) of the stabilization methods. The explanation for the occurrence of wiggles/oscillations in the transient evolution of the numerical solution was explained by examining the dispersion relations of the continuous and discrete problems in [183]. It has been found that beyond a certain wavenumber ξ_d the continuous and the discrete dispersion relations diverge [44, 183]. This wavenumber (ξ_d) is referred to as the *phase departure wave number* in [44]. If the bandwidth of the amplitude spectra of any given initial function has wavenumbers greater than ξ_d, the initial function suffers a change of form (with wiggles/oscillations) in its transient evolution. Examining the respective DDRs, we seek to find if the stabilization methods provide any improvements in the solutions. A comparison of the DDR of the FIC_RC/OSS method is done with the DDRs of the Galerkin and FIC/SUPG methods for three standard time integration schemes. Also, the range of wavenumbers to which the DDR agrees with the continuous dispersion relation is shown to extend, should a consistent "effective" mass matrix be preferred to a lumped one. Next, it is shown that unlike the FIC/SUPG method, the FIC_RC/OSS method introduces a certain rearrangement in the equation stencils at nodes on and adjacent to the domain boundary. Thus using a uniform expression for the stabilization parameter (α) will lead to enhanced localized oscillations at the boundary. For the 1D steady-state problem , we present a new expression for α which is optimal for uniform grids and provides negligible damping when used in the transient mode. Unfortunately for non-uniform grids, it leads to weak node-to-node oscillations.

1.2 TRANSIENT CONVECTION DIFFUSION EQUATION

1.2.1 *Problem Statement*

The statement of the multi-dimensional problem is as follows:

$$\frac{\partial \phi}{\partial t} + \mathbf{u} \cdot \boldsymbol{\nabla} \phi - \boldsymbol{\nabla} \cdot (k \boldsymbol{\nabla} \phi) - f = 0 \qquad \text{in} \quad \Omega \tag{1.1a}$$

$$\phi(\mathbf{x}, t = 0) = \phi_o(\mathbf{x}) \qquad \text{in} \quad \Omega \tag{1.1b}$$

$$\phi = \phi^p \qquad \text{on} \quad \Gamma_D \tag{1.1c}$$

$$k \boldsymbol{\nabla} \phi \cdot \mathbf{n} = q^p \qquad \text{on} \quad \Gamma_N \tag{1.1d}$$

where, $\mathbf{u}$ is the convection velocity, k is the diffusion, f is the source, $\phi(\mathbf{x}, t)$ is the transported variable, $\phi_o(\mathbf{x})$ is the initial solution, ϕ^p and q^p are the prescribed values of ϕ and the diffusive flux at the Dirichlet and Neumann boundaries respectively. The

FIC formulation of this problem (neglecting the time stabilization terms [145]) is as follows:

$$r - \frac{1}{2}\mathbf{h}\cdot\boldsymbol{\nabla} r = 0 \qquad \text{in} \quad \Omega \tag{1.2a}$$

$$\phi(\mathbf{x}, t=0) = \phi_o(\mathbf{x}) \qquad \text{in} \quad \Omega \tag{1.2b}$$

$$\phi_h = \phi^p \qquad \text{on} \quad \Gamma_D \tag{1.2c}$$

$$k\boldsymbol{\nabla}\phi_h\cdot\mathbf{n} + \frac{1}{2}(\mathbf{h}\cdot\mathbf{n})r = q^p \qquad \text{on} \quad \Gamma_N \tag{1.2d}$$

$$r := \frac{\partial\phi}{\partial t} + \mathbf{u}\cdot\boldsymbol{\nabla}\phi - \boldsymbol{\nabla}\cdot(k\boldsymbol{\nabla}\phi) - f \tag{1.2e}$$

Let $(\cdot,\cdot)$ and $(\cdot,\cdot)_{\Gamma_N}$ denote the $L^2(\Omega)$ and $L^2(\Gamma_N)$ inner products respectively. The variational form of the problem (1.1) can be expressed as follows: Find $\phi : [0,T] \mapsto V$ such that $\forall w \in V_0$ we have,

$$a(w,\phi) = l(w) \tag{1.3a}$$

$$a(w,\phi) := \left(w, \frac{\partial\phi}{\partial t}\right) + (w, \mathbf{u}\cdot\boldsymbol{\nabla}\phi) + (\boldsymbol{\nabla} w, k\boldsymbol{\nabla}\phi) \tag{1.3b}$$

$$l(w) := (w,f) + (w,q^p)_{\Gamma_N} \tag{1.3c}$$

where $V := \{w : w \in H^1(\Omega) \text{ and } w = \phi^p \text{ on } \Gamma_D\}$, $V_0 := \{w : w \in H^1(\Omega) \text{ and } w = 0 \text{ on } \Gamma_D\}$. For the FIC equations the variational form can be expressed as follows: Find $\phi : [0,T] \mapsto V$ such that $\forall w \in V_0$ we have,

$$a(w,\phi) + \frac{1}{2}(\boldsymbol{\nabla}\cdot(\mathbf{h}w), r) = l(w) \tag{1.4}$$

The calculation of the residual contribution in the stabilization term can be simplified if we introduce the projection of the convective term π via an auxiliary equation defined as,

$$\pi = \mathbf{u}\cdot\boldsymbol{\nabla}\phi - r \tag{1.5}$$

We express the residual, r, that occurs in the stabilization term $(\boldsymbol{\nabla}\cdot(\mathbf{h}w), r)$ as a function of π. Thus π becomes an addendum to the set of unknowns to be found. The integral equation system is now augmented forcing that the residual r expressed in terms of π via Eq.(1.5) vanishes (in average) over the analysis domain. Problem (1.4) after the addendum of the convective projection π is expressed as follows: Find $\phi : [0,T] \mapsto V$ and $\pi \in H^1(\Omega)$ such that $\forall (w,z) \in (V_0(\Omega), H^1(\Omega))$,

$$a(w,\phi) + \frac{1}{2}(\boldsymbol{\nabla}\cdot(\mathbf{h}w), \mathbf{u}\cdot\boldsymbol{\nabla}\phi - \pi) = l(w) \tag{1.6a}$$

$$(z,\pi) = (z, \mathbf{u}\cdot\boldsymbol{\nabla}\phi) \tag{1.6b}$$

We remark that the projection of the convective term provides consistency to the formulation, i. e.the system of equations (1.6) have the residual form which vanishes for the exact solution. Henceforth we refer to the formulation given by the equation (1.6) as the FIC formulation with recovered consistency (FIC_RC). The convective projection is expected to capture the otherwise lost effect of higher order terms in the residual when simplicial elements are used. The introduction of the convective projection variable π was deduced from the OSS approach in [34].

1.2.2 *Dispersion Relation in 1D*

Any equation that admits plane wave solutions of the form $\exp[i(\omega t - \xi x)]$, but with the property that the speed of propagation of these waves is dependent on ξ, is generally referred to as a *dispersive equation*. Here ξ, ω are the angular wave-number and the frequency, respectively. The equation that expresses ω as a function of ξ is known as the *dispersion relation*. Generally the transient convection-diffusion equation is a dispersive equation. In the limit, when diffusion tends to zero and the equation morphs into a pure convection problem, it tend to become non-dispersive [183]. In 1D and for the sourceless case ($f = 0$), Eq.(1.1a) can be expressed as follows:

$$\frac{\partial \phi}{\partial t} + u\frac{\partial \phi}{\partial x} - k\frac{\partial^2 \phi}{\partial x^2} = 0 \tag{1.7}$$

For a given discretization of size ℓ in space and an increment θ in time, we express the *Courant* and *Peclet* numbers as $C = \frac{u\theta}{\ell}$ and $\gamma = \frac{u\ell}{2k}$. We write down the dispersion relation (Eq. 1.8) for the continuous problem (Eq. 1.7) by propagating the plane wave solution ϕ =$\exp[i(\omega t - \xi x)]$. From Eq.(1.8) we obtain, taking the limit $k \to 0$ or $\gamma \to \infty$, the dispersion relation for the pure convection problem.

$$\omega = u\xi + ik\xi^2 \tag{1.8a}$$

$$\omega\theta = C(\xi\ell) + i\frac{C}{2\gamma}(\xi\ell)^2 \tag{1.8b}$$

Now let us consider the amplification of the solution from time step n to $n+1$ and at some given spatial point. The *amplification parameter* β as defined in [44] is given by Eq.(1.9). It can be clearly seen that β is stationary in time and uniform in space. The amplification and phase shift per time step are given by the magnitude and argument of Eq.(1.9), respectively. Thus for the pure convection problem we can see that the amplification is unity, i. e.$|\beta| = 1$. The phase and group velocities are given by the Eqs. (1.10a) and (1.10b) respectively [73].

$$\beta = \frac{\phi_j^{n+1}}{\phi_j^n} = \exp[i\omega\theta] = \exp\left[-k\theta\xi^2\right]\exp[iu\theta\xi] = \exp\left[-\frac{C}{2\gamma}(\xi\ell)^2\right]\exp[iC(\xi\ell)] \tag{1.9}$$

$$V_p(\xi) = \frac{\omega(\xi)}{\xi} \tag{1.10a}$$

$$V_g(\xi) = \frac{\partial\omega(\xi)}{\partial\xi} \tag{1.10b}$$

1.3 FE DISCRETIZATION

1.3.1 *Semi-Discrete Form*

The semi-discrete (continuous in time, discrete in space) counterpart of the FIC method (1.4) can be written as follows: Find $\phi_h : [0, T] \mapsto V^h$ such that $\forall w_h \in V_0^h$ we have,

$$a(w_h, \phi_h) + \sum_e \frac{1}{2}\left(\boldsymbol{\nabla}\cdot(\mathbf{h}w_h), r_h\right)_{\Omega^e} = l(w_h) \tag{1.11}$$

where $V^h \subset V$ and $V_0^h \subset V_0$. The stabilization term in Eq.(1.11) has been expressed as a sum of the element contributions to allow for inter-element discontinuities in the term $\boldsymbol{\nabla} r_h$ of Eq.(1.2), where $r_h := r(\phi_h)$ is the residual of the FE approximation of the infinitesimal governing equation and $(\cdot,\cdot)_{\Omega^e}$ denote the $L^2(\Omega^e)$ inner product. Similarly the discrete counterpart of the FIC_RC method (1.6) can be written as: Find $\phi_h : [0,T] \mapsto V^h$ and $\pi_h \in H^1(\Omega)$ such that $\forall (w_h, z_h) \in (V_0^h(\Omega), H^1(\Omega))$,

$$a(w_h,\phi_h) + \sum_e \frac{1}{2}\left(\boldsymbol{\nabla}\cdot(\mathbf{h}w_h), \mathbf{u}\cdot\boldsymbol{\nabla}\phi_h - \pi_h\right)_{\Omega^e} = l(w_h) \tag{1.12a}$$

$$(z_h,\pi_h) = (z_h, \mathbf{u}\cdot\boldsymbol{\nabla}\phi_h) \tag{1.12b}$$

The variables in the Eqs.(1.11) and (1.12) interpolated by finite element shape functions N^a can be expressed as follows:

$$\phi_h = N^a\phi_a, \pi_h = N^a\pi_a, w_h = N^a w_a, z_h = N^a z_a \tag{1.13}$$

where a is the spatial node index. The discrete problems (1.11) and (1.12) can be written in matrix notation via Eqs.(1.14) and (1.15), respectively, as follows:

$$[\mathbf{M} + \mathbf{S2}]\dot{\Phi} + [\mathbf{C} + \mathbf{D} + \mathbf{S1} + \mathbf{S3}]\Phi = \mathbf{fg} + \mathbf{fs} \tag{1.14}$$

$$\mathbf{M}\dot{\Phi} + [\mathbf{C} + \mathbf{D} + \mathbf{S1}]\Phi - \mathbf{S2}\Pi = \mathbf{fg} \tag{1.15a}$$

$$\mathbf{M}\Pi - \mathbf{C}\Phi = \mathbf{o} \tag{1.15b}$$

where $\Phi := \{\phi_a\}$ and $\Pi := \{\pi_a\}$ represent the vector of nodal unknowns. The element contributions to the matrices and vectors in Eqs.(1.14) and (1.15) are given by

$$\mathbf{C}^e_{ab} = \left(N^a, \mathbf{u}\cdot\boldsymbol{\nabla}N^b\right)_{\Omega^e}, \qquad \mathbf{D}^e_{ab} = \left(\boldsymbol{\nabla}N^a, k\boldsymbol{\nabla}N^b\right)_{\Omega^e} \tag{1.16a}$$

$$\mathbf{M}^e_{ab} = \left(N^a, N^b\right)_{\Omega^e}, \qquad \mathbf{S1}^e_{ab} = \frac{1}{2}\left(\boldsymbol{\nabla}\cdot(\mathbf{h}N^a), \mathbf{u}\cdot\boldsymbol{\nabla}N^b\right)_{\Omega^e} \tag{1.16b}$$

$$\mathbf{S2}^e_{ab} = \frac{1}{2}\left(\boldsymbol{\nabla}\cdot(\mathbf{h}N^a), N^b\right)_{\Omega^e}, \qquad \mathbf{S3}^e_{ab} = -\frac{1}{2}\left(\boldsymbol{\nabla}\cdot(\mathbf{h}N^a), \boldsymbol{\nabla}\cdot(k\boldsymbol{\nabla}N^b)\right)_{\Omega^e} \tag{1.16c}$$

$$\mathbf{fg}^e_a = \left(N^a, f\right)_{\Omega^e} + \left(N^a, q^p\right)_{\Gamma_N}, \qquad \mathbf{fs}^e_a = \frac{1}{2}\left(\boldsymbol{\nabla}\cdot(\mathbf{h}N^a), f\right)_{\Omega^e} \tag{1.16d}$$

Note that Eq.(1.15b) correspond to the L^2-projection of the term $\mathbf{u}\cdot\boldsymbol{\nabla}\phi_h$ onto the space spanned by the shape functions. Whenever the later term admits discontinuities the projection is non-monotone [34]. A monotone projection of the convective term can be achieved if the mass matrix $\mathbf{M}$ that appears in Eq.(1.15b) is lumped. Also the expression for Π will have local support only when $\mathbf{M}$ is lumped. This feature allows us to study generic nodal equation stencils for the interior of the domain. Henceforth we always consider the Eq.(1.15b) with $\mathbf{M}$ lumped. Eqs.(1.14) and (1.15) may be expressed in a general form as shown in Eq.(1.17). Table 1 defines the corresponding matrices for the FIC and FIC_RC methods.

$$\mathbf{T}\dot{\Phi} + [\mathbf{C} + \mathbf{D} + \mathbf{S}]\Phi = \mathbf{f} \tag{1.17}$$

	FIC	FIC_RC
T-lumped	$\mathbf{M}_L + \mathbf{S}2$	$\mathbf{M}_L$
T-consistent	$\mathbf{M} + \mathbf{S}2$	$\mathbf{M}$
S	$\mathbf{S}1 + \mathbf{S}3$	$\mathbf{S}1 - \mathbf{S}2\mathbf{M}_L^{-1}\mathbf{C}$
f	**fg**+**fs**	**fg**

Table 1: Matrix definitions for FIC and FIC_RC methods

where matrix **T** is the 'effective' mass matrix.

Note that for the direction of **h** being the same as that of the velocity **u**, i. e.$\mathbf{h} = 2\tau\mathbf{u}$ and assuming τ constant within an element, the form of the stabilization term $\left(\boldsymbol{\nabla}\cdot(\mathbf{h}w_h), r_h\right)_{\Omega^e}$ in Eq.(1.11) is identical to that of the standard SUPG method. Thus with this choice of **h** the FIC and FIC_RC methods are identical to the SUPG and OSS (orthogonal sub-scales) methods respectively. The general direction of **h** introduces naturally stabilization along the streamlines and also along the directions of the gradient of the solution transverse to the velocity vector. The FIC formulation therefore incorporates the best features of the SUPG and the shock-capturing methods. Applications of the FIC-FEM formulation to a wide range of convection-diffusion problems with sharp gradients are presented in [154]. We remark that in 1D (assuming h constant within an element) the FIC and the FIC_RC methods are identical to the SUPG and OSS methods respectively. Thus the conclusions made between the FIC and FIC_RC methods may be carried over to those between SUPG and OSS methods.

1.3.2 *DDR in 1D*

The DDR of the semi-discrete problem (semi-DDR) and when the temporal terms are discretized using two class of time discretization schemes are investigated in this section. The time discretization schemes considered are the *trapezoidal scheme* and the *second-order backward differencing formula (BDF2)*. The effects on the numerical dispersion due to the choice of the form of the effective mass matrix **T** (lumped or consistent) in Eq.(1.17) are also studied. The flag *lumped* or *consistent* refers only to the matrix **T** as defined in Table 1 for the FIC and FIC_RC methods. The DDRs are written by inserting a plane wave solution of the form $\phi = \exp[i(\omega t - \xi x)]$ into the corresponding equation stencils. Taking the limit $\gamma \to \infty$ we recover the DDR for the pure convection problem.

1.3.2.1 *Semi-DDR*

We study the equation stencil for an interior node of the semi-discrete problem given by Eq.(1.17) with $f = 0$ in 1D. For a compact representation of the stencils we introduce the following definition,

$$(\star) := \left(\frac{u}{2}\right)(\phi_{j+1} - \phi_{j-1}) - \left(\frac{k}{\ell} + \frac{u\alpha}{2}\right)(\phi_{j+1} - 2\phi_j + \phi_{j-1}) \tag{1.18}$$

FIC/SUPG method:

$$(\frac{\ell}{6})\big[(\frac{-3\alpha}{2})\dot{\phi}_{j+1} + 6\dot{\phi}_j + (\frac{3\alpha}{2})\dot{\phi}_{j-1}\big] + (\star) = 0, \mathbf{T}\text{-lumped} \tag{1.19a}$$

$$(\frac{\ell}{6})\big[(1 - \frac{3\alpha}{2})\dot{\phi}_{j+1} + 4\dot{\phi}_j + (1 + \frac{3\alpha}{2})\dot{\phi}_{j-1}\big] + (\star) = 0, \mathbf{T}\text{-consistent} \tag{1.19b}$$

FIC_RC/OSS method:

$$\ell\ \dot{\phi}_j + (\star) + (\frac{u\alpha}{8})(\phi_{j+2} - 2\phi_j + \phi_{j-2}) = 0, \mathbf{T}\text{-lumped} \tag{1.20a}$$

$$(\frac{\ell}{6})(\dot{\phi}_{j+1} + 4\dot{\phi}_j + \dot{\phi}_{j-1}) + (\star) + (\frac{u\alpha}{8})(\phi_{j+2} - 2\phi_j + \phi_{j-2}) = 0, \mathbf{T}\text{-consistent} \tag{1.20b}$$

Making $\alpha = 0$ in Eqs. (1.19) and (1.20) we recover the standard Galerkin method. The Semi-DDRs for all the methods and for the **T**-lumped, **T**-consistent cases can be expressed in a generic and compact manner as follows:

$$\omega_h = -i\frac{B}{\theta A} \tag{1.21}$$

where,

$$A := \begin{cases} \dfrac{1}{C} & \text{Galerkin, FIC\_RC/OSS methods, } \mathbf{T}\text{-lumped} \\ \dfrac{2+\cos(\xi\ell)}{3C} & \text{Galerkin, FIC\_RC/OSS methods, } \mathbf{T}\text{-consistent} \\ \dfrac{1}{C} + i\dfrac{\alpha}{2C}\sin(\xi\ell) & \text{FIC/SUPG methods, } \mathbf{T}\text{-lumped} \\ \dfrac{2+\cos(\xi\ell)}{3C} + i\dfrac{\alpha}{2C}\sin(\xi\ell) & \text{FIC/SUPG methods, } \mathbf{T}\text{-consistent} \end{cases} \tag{1.22a}$$

$$B := \begin{cases} i\sin(\xi\ell) - 2\sin^2(\frac{\xi\ell}{2})\ \left(\dfrac{1}{\gamma}\right) & \text{Galerkin method} \\ i\sin(\xi\ell) - 2\sin^2(\frac{\xi\ell}{2})\ \left(\dfrac{1}{\gamma} + \alpha\right) & \text{FIC/SUPG methods} \\ i\sin(\xi\ell) - 2\sin^2(\frac{\xi\ell}{2})\ \left(\dfrac{1}{\gamma} + \alpha\sin^2(\frac{\xi\ell}{2})\right) & \text{FIC\_RC/OSS methods} \end{cases} \tag{1.22b}$$

1.3.2.2 *Trapezoidal Scheme*

The structure of the equation stencil for an interior node with respect to the spatial indices is the same as in the semi-discrete problem. Henceforth we express the fully discrete system in the matrix notation only.

$$\mathbf{T}\cdot\frac{\Phi^{n+1} - \Phi^n}{\theta} + [\mathbf{C} + \mathbf{D} + \mathbf{S}]\cdot\Phi^{n+\sigma} = \mathbf{0} \tag{1.23a}$$

$$\Phi^{n+\sigma} := \sigma\ \Phi^{n+1} + (1-\sigma)\ \Phi^n \tag{1.23b}$$

Making $\sigma = \{0, 0.5, 1\}$ we recover the *forward Euler, Crank-Nicholson* and *backward Euler* schemes respectively. The DDR for the trapezoidal scheme can be expressed in terms of A and B defined in Eq.(1.22) as follows,

$$\exp\left[i\omega_h\theta\right] = \frac{A + (1-\sigma)B}{A - \sigma B} \quad \text{or equivalently,} \tag{1.24a}$$

$$\tan\left(\frac{\omega_h\theta}{2}\right) = -i\frac{B}{2A + (1-2\sigma)B} \tag{1.24b}$$

1.3.2.3 *BDF2 Scheme*

The fully discrete system of equations after time discretization by the BDF2 scheme is given by,

$$\mathbf{T}\cdot\frac{3\Phi^{n+1} - 4\Phi^n + \Phi^{n-1}}{2\theta} + [\mathbf{C} + \mathbf{D} + \mathbf{S}]\cdot\Phi^{n+1} = \mathbf{O} \tag{1.25}$$

The DDR for the BDF2 scheme is a quadratic relation in $\exp[i\omega_h\theta]$. The solution to the quadratic equation gives two expressions for the DDR, which can be expressed as follows:

$$\exp\left[i\omega_h\theta\right] = \frac{2A + \sqrt{A^2 + 2AB}}{3A - 2B} \tag{1.26a}$$

$$\exp\left[i\omega_h\theta\right] = \frac{2A - \sqrt{A^2 + 2AB}}{3A - 2B} \tag{1.26b}$$

We remark that the solution given by Eq.(1.26b) predicts negative values of $\mathfrak{R}(\omega_h)$[1] for positive wave-numbers. Thus, we consider the solution given by Eq.(1.26a) as the only acceptable solution.

1.3.3 *DDR Plots*

The DDRs presented in the previous section represent the frequency as a function of six independent variables, i.e.$\omega_h := \omega_h(\xi, \ell, \theta, C, \gamma, \alpha)$. For a feasible graphical representation of the DDRs we freeze some of them and normalize the frequency and wavenumbers to retain maximum generality. In the DDR plots we consider only the pure convection problem ($k = 0$). The stabilization parameter $\alpha = 1.0$ is chosen. This corresponds to the optimal value for the SUPG/FIC method in 1D and for a uniform mesh. This choice is made for convenience and comparison of the effects of the stabilization term introduced by the considered methods. Note that the DDRs are periodic in ξ and the corresponding fundamental domain is $\xi \in [-\pi/\ell, \pi/\ell]$. The Nyquist frequency in space is $\xi_{nq} = \pi/\ell$ and in time is $\omega_{nq} = \pi/\theta$. Thus in the DDR plots we do not consider wavenumbers and frequencies beyond the Nyquist limits. It can be shown with respect to the exact dispersion relation (Eq. 1.8) that this condition corresponds to choosing $C \leqslant 1$. We normalize the wavenumber ξ by the Nyquist limit ξ_{nq}, i.e.$\xi^* = \xi/\xi_{nq}$. The frequency ω_h is normalized as $\omega_h^* = \omega_h/(u\xi_{nq}) = \omega_h\theta/(C\pi) = \omega_h\ell/(u\pi)$. The DDRs are now expressed with respect to the normalized wavenumber and frequency, i.e.$\omega_h^* := \omega_h^*(\xi^*, C)$. The plotting domain considered is $(\xi^*, C) = (0, 1) \times (0, 1)$.

1 The real and imaginary parts of ω_h are denoted as $\mathfrak{R}(\omega_h)$ and $\mathfrak{I}(\omega_h)$, respectively

1.3.3.1 *Semi-discrete case*

For the semi-discrete problem the DDR no longer depends on θ. The frequency ω_h^* is now only a function of ξ^*, i.e.$\omega_h^* := \omega_h^*(\xi^*)$. The amplification at time $t_n = n\theta$ is given by $\exp[-\mathfrak{I}(\omega_h)t_n] = \exp[-\mathfrak{I}(\omega_h^*)\pi(ut_n/\ell)] = \exp[-\mathfrak{I}(\omega_h^*)\pi Cn]$. This means that should $\mathfrak{I}(\omega_h^*) \neq 0$ the amplification at any given time is independent of the time step θ but dependent on the space discretization ℓ. Thus we present 1D plots for the following:

- Plot of $\mathfrak{R}(\omega_h^*)$ vs ξ^*. The departure wavenumber ξ_d^* is marked such that $\forall\ \xi^* \leqslant \xi_d^*$ we have $|\mathfrak{R}(\omega^* - \omega_h^*)| \leqslant 0.001$ (Figure 1).
- Amplification plots using $C = 0.1$ and at times θ, 2θ, 100θ, 200θ, and 300θ sec (Figure 2).

1.3.3.2 *Fully discrete case*

For the fully discrete case the time integration schemes considered are: the *backward Euler*, *Crank-Nicholson* and *BDF2*. The frequency is now a function of both ξ^* and C, i.e.$\omega_h^* := \omega_h^*(\xi^*, C)$. The amplification at time $t_n = n\theta$ is given by $\exp[-\mathfrak{I}(\omega_h^*)\pi(ut_n/\ell)]$; the same as for the semi-discrete case except for the fact that ω_h^* is now dependent on C also. Contour plots are presented for the following,

- $\log_{10}(|\mathfrak{R}(\omega^* - \omega_h^*)|)$ vs. (ξ^*, C). The contour of values $\{-5, -4, -3, -2, -1\}$ are shown (Figures 3, 5 and 7).
- $\log_{10}(\mathfrak{I}(\omega_h^*))$ vs. (ξ^*, C). The contour of values $\{-3, -2.5, -2, -1.5, -1, -0.5, 0\}$ are shown (Figures 4, 6 and 8).

Only the contour plots for the normalized group velocity, i.e.$\partial\omega_h^*/\partial\xi^*$ vs (ξ^*,C), for the Crank-Nicholson scheme have been presented (Figure 9). The original group velocity can be recovered as follows: $\partial\omega_h/\partial\xi = u\ \partial\omega_h^*/\partial\xi^*$.

1.3.4 *Discussion*

It can be seen from the DDR plots that every discrete model and also the semi-discrete model of the continuous problem diverges from the exact dispersion relation beyond a certain wave-number, here ξ_d^* (Figures 1, 3, 5 and 7). For the semi-discrete case, ξ_d^* is marked in the plots and for the fully discrete case ξ_d^* is a contour line given by the value -3. ξ_d^* is greater when we use a consistent mass matrix for the transient terms in the formulation. Thus, one should expect better phase fidelity over a wider range of wave numbers using a consistent mass matrix. The gain in the value of ξ_d^* from lumped $\mathbf{T}$ case to the consistent $\mathbf{T}$ case gradually decreases as the Courant number C increases (except for a certain range of the Courant number C for the FIC/SUPG method).

We now examine the differences in the DDR plots between the Galerkin method and the DDR plots of the FIC/SUPG and FIC_RC/OSS methods. It is interesting that the stabilization terms introduced by the FIC_RC/OSS method do not alter much the location of the phase departure wave-number ξ_d^*. On the other hand, the stabilization terms introduced by the FIC/SUPG method contribute to the $\mathbf{T}$ matrix (Table 1).

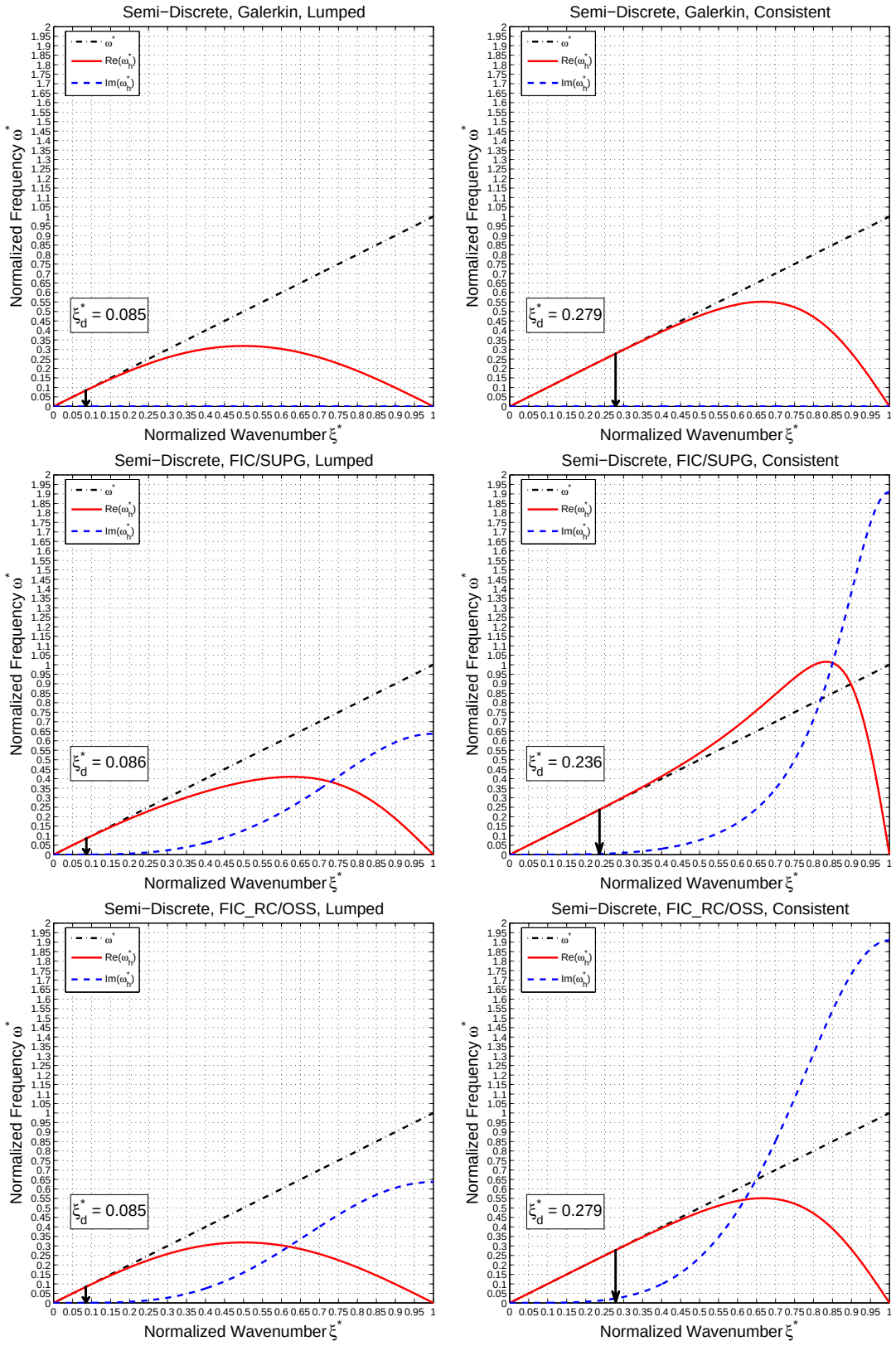


Figure 1: Plot of ω_h^* vs ξ^* for the semi-discrete problem. Frequencies ω^* and ω_h^* correspond to the continuous and discretized problems respectively. $\alpha = 1.0$ is used.

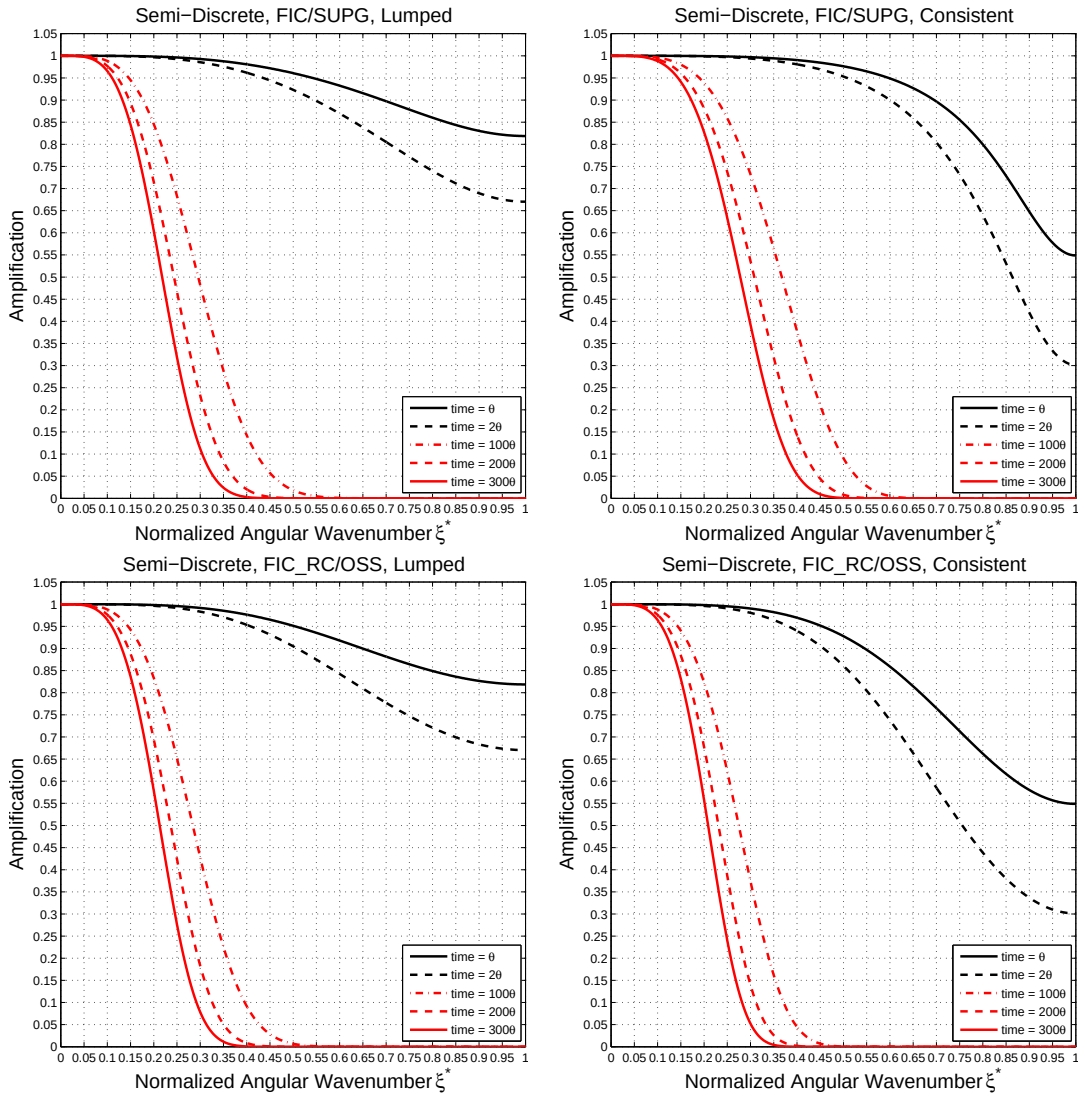


Figure 2: Amplification plots of the semi-discrete problem for the FIC/SUPG and FIC_RC/OSS methods. $\alpha = 1.0$ and $C = 0.1$ are used. The amplification for the Galerkin method is not shown here as it is equal to 1

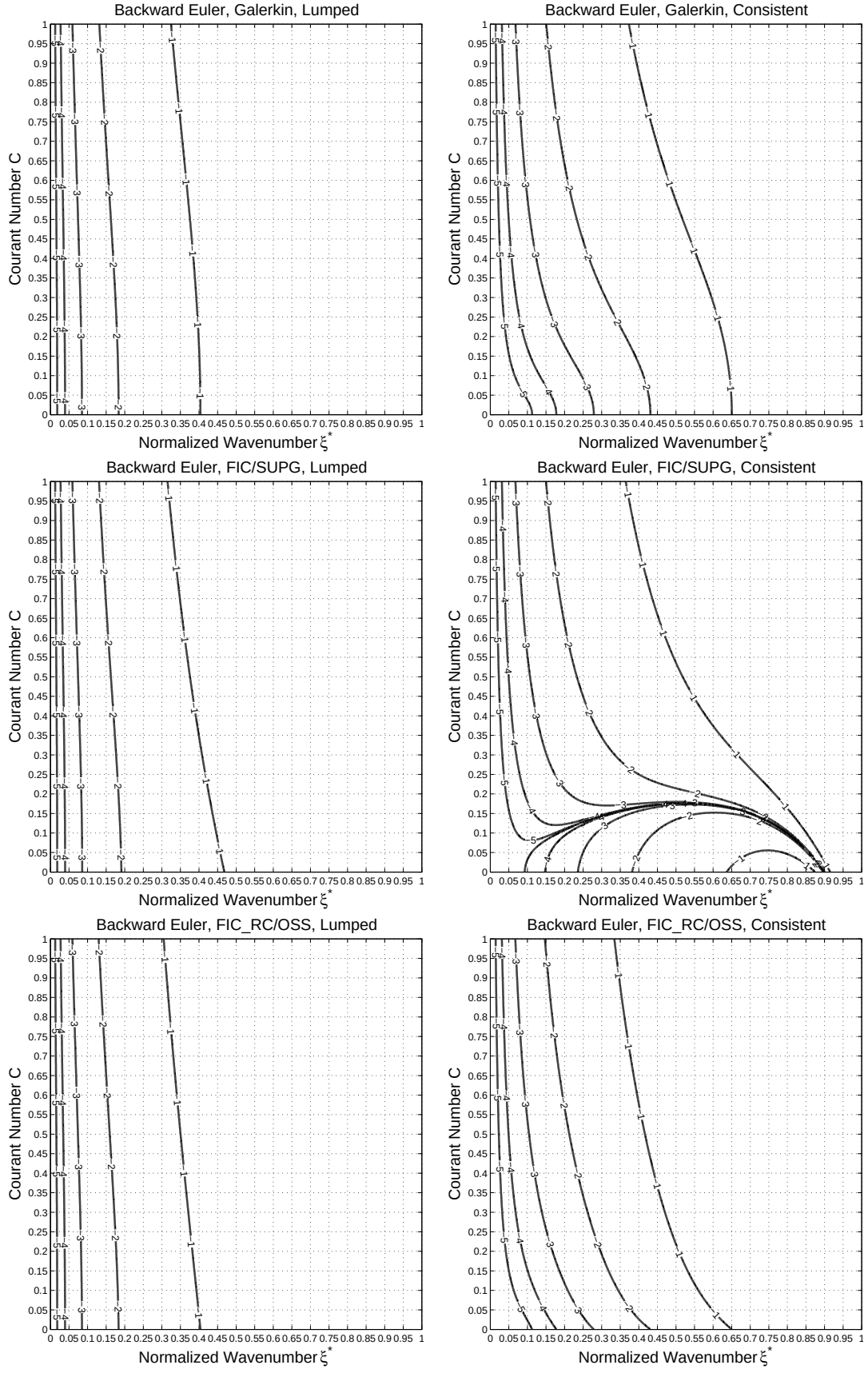


Figure 3: Contour plot of $\log_{10}[\Re(|\omega_h^* - \omega^*|)]$ for the backward Euler scheme. $\alpha = 1.0$ is used.

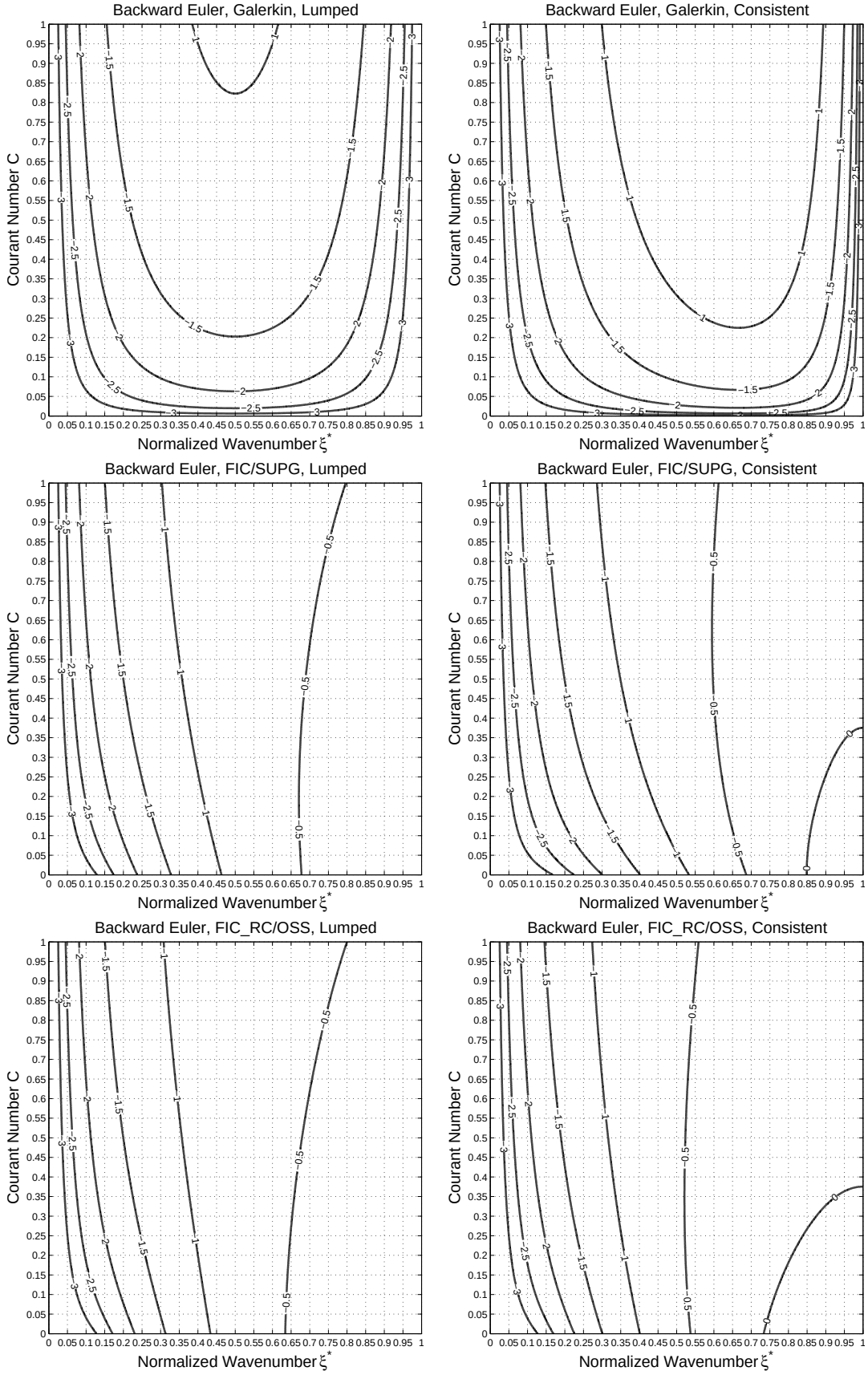


Figure 4: Contour plot of $\log_{10}[\mathfrak{I}(\omega_h^*)]$ for the backward Euler scheme. $\alpha = 1.0$ is used.

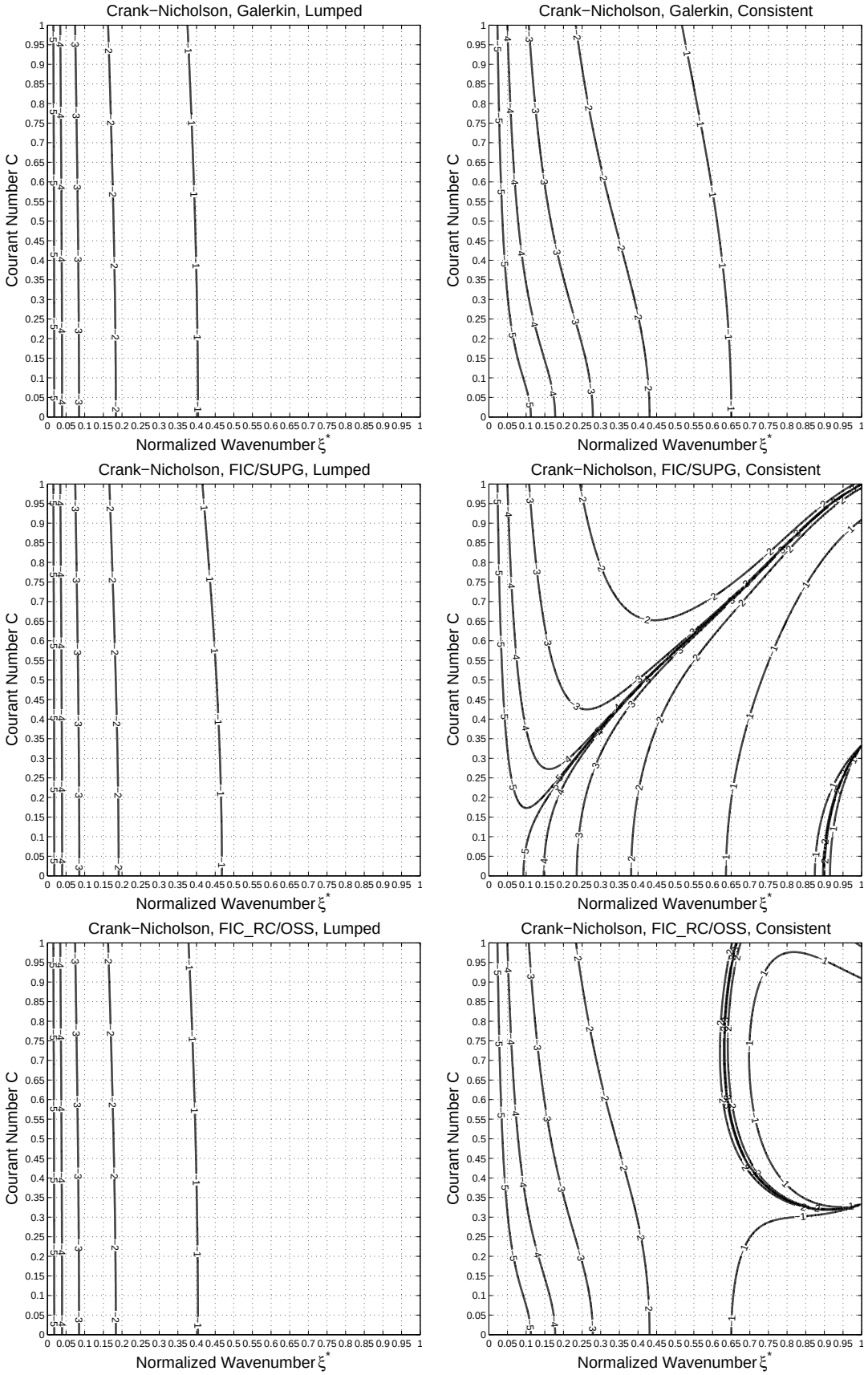


Figure 5: Contour plot of $\log_{10}[\Re(|\omega_h^* - \omega^*|)]$ for the Crank-Nicholson scheme. $\alpha = 1.0$ is used.

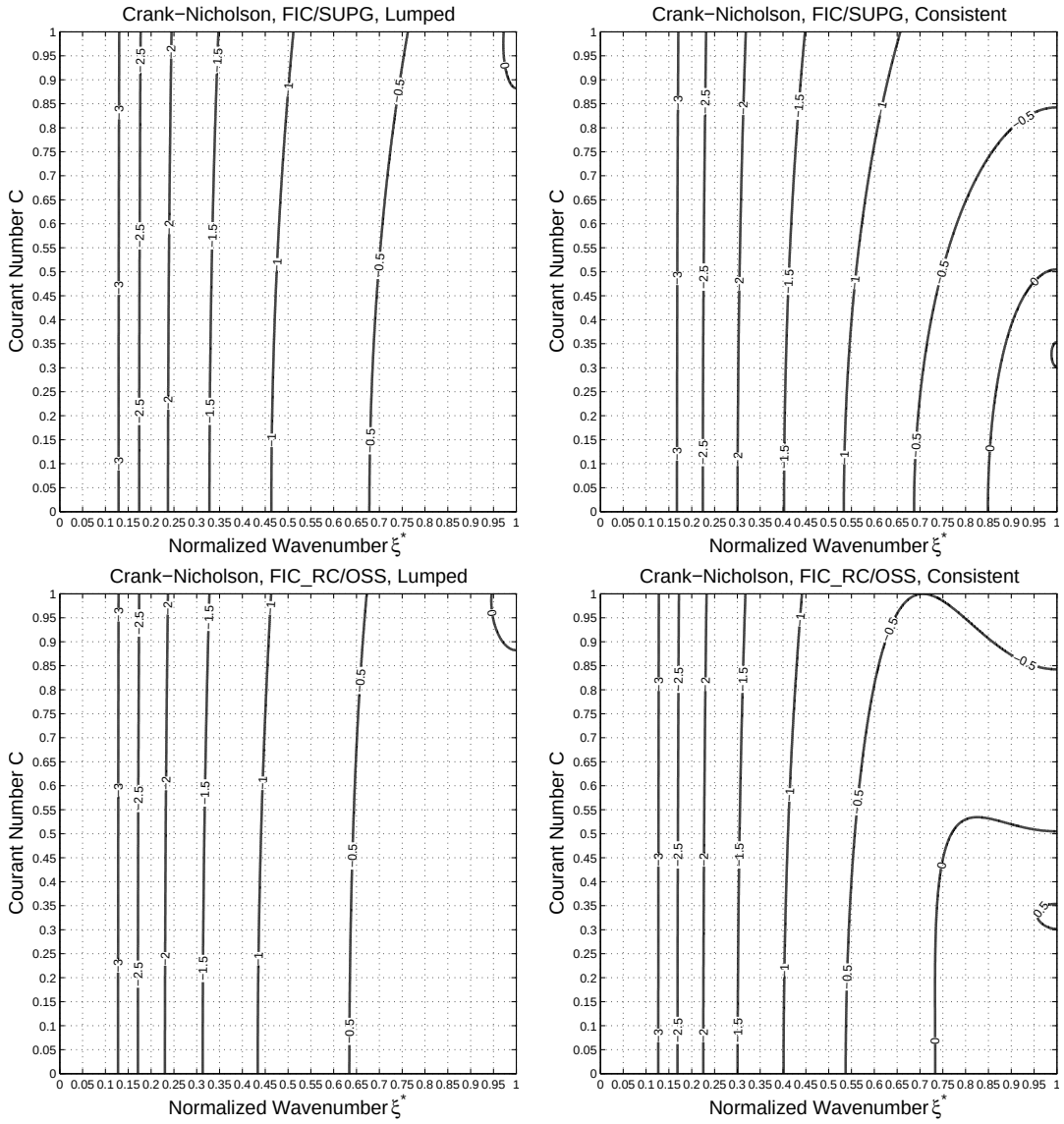


Figure 6: Contour plot of $\log_{10}[\mathfrak{I}(\omega_h^*)]$for the Crank-Nicholson scheme. $\alpha = 1.0$ is used. The plots for the Galerkin method is not shown here as $\mathfrak{I}(\omega_h^*) = 0$

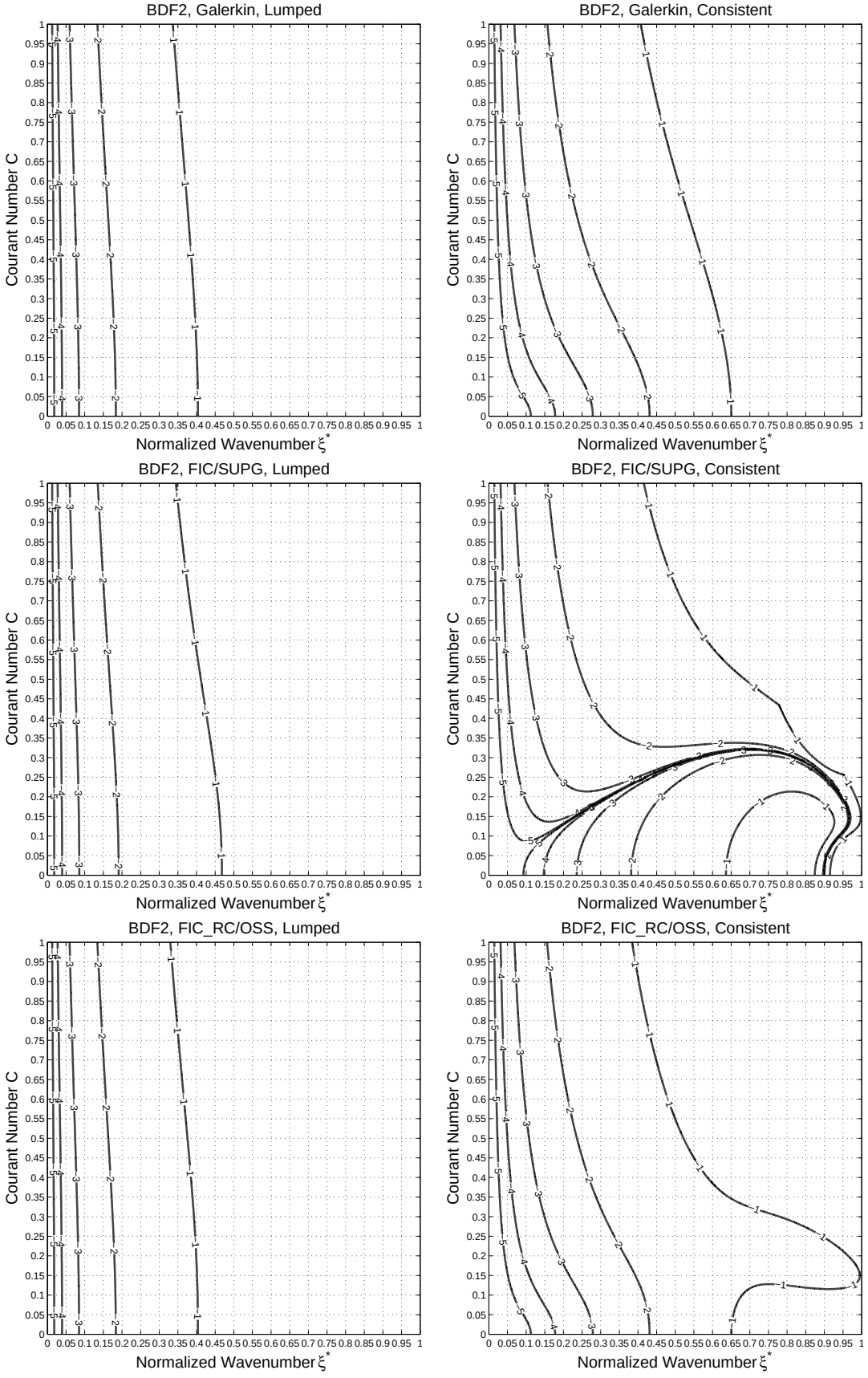


Figure 7: Contour plot of $\log_{10}[\Re(|\omega_h^* - \omega^*|)]$ for the BDF2 scheme. $\alpha = 1.0$ is used.

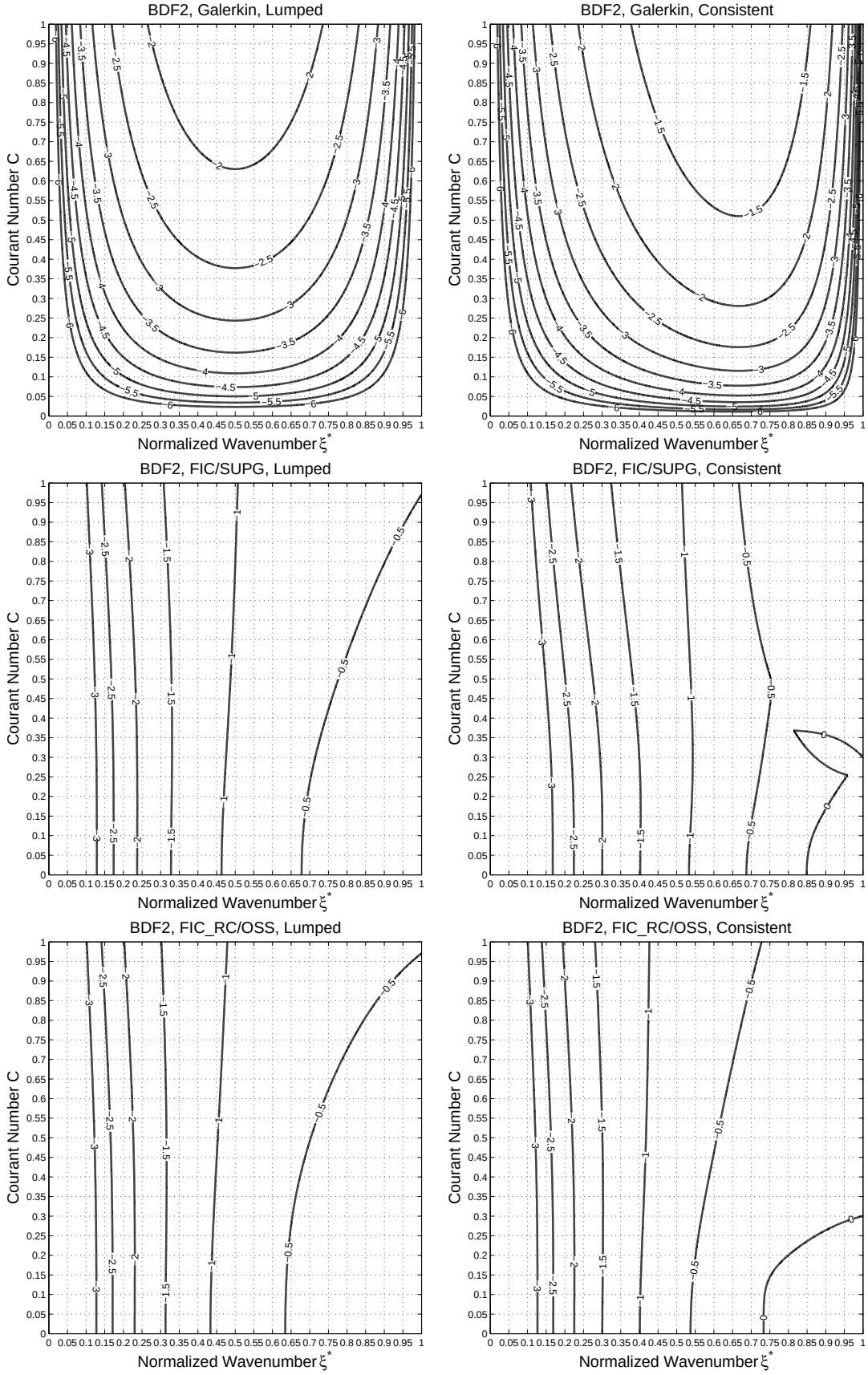


Figure 8: Contour plot of $\log_{10}[\mathfrak{I}(\omega_h^*)]$ for the BDF2 scheme. $\alpha = 1.0$ is used.

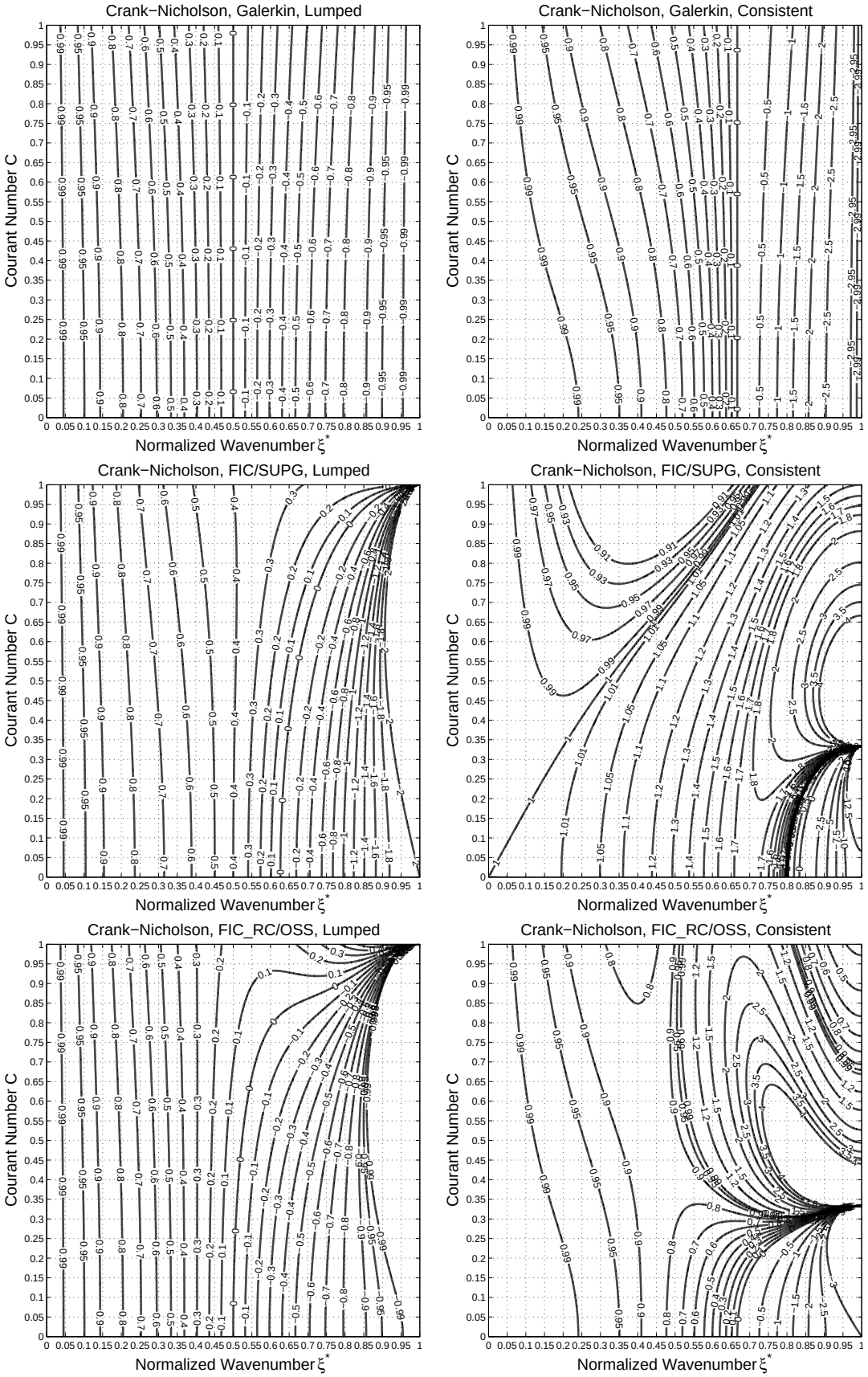


Figure 9: Real part of the normalized group velocity $\partial\omega_h^*/\partial\xi^*$ vs. ξ^* for the Crank Nicholson scheme; $\alpha = 1.0$ is used.

Although for higher Courant number C the effect on the location of ξ_d^* is negligible, significant alterations in ξ_d^* are found for a lower C and a consistent **T** matrix (Figures 3, 5 and 7). When **T** is lumped, all the methods give similar patterns for ξ_d^* indicating that stabilization terms have no role to play in the improvement of the DDRs.

Next, we examine the effect of the variation in C on the $\Re(\omega_h^*)$ vs ξ^* relation for the FIC and SUPG methods with consistent **T** matrix and the Crank Nicholson scheme. Again, we choose ξ_d^* using the criterion: $\forall\ \xi^* \leqslant \xi_d^*$ we have $|\Re(\omega^* - \omega_h^*)| \leqslant 0.001$. In Figure 5 this corresponds to the contour line for the value -3. It can been seen in the semi-discrete case that $\Re(\omega_h^*) > \Re(\omega^*)$ for lower wavenumbers (Figure 1). In addition, the error $\Re(\omega_h^* - \omega^*)$ first increases and later decreases. This behavior is exhibited in the fully discrete case too. Thus, there will be multiple contours of the same value in the $\log_{10}(|\omega_h^* - \omega^*|)$ vs (ξ^*, C) contour plots. In this case ξ_d^* is the smallest ξ^* on the contours. For instance, choosing $C = 0.4$ for the FIC and SUPG methods we find $\xi_d^* \approx 0.35$, where as for the Galerkin, FIC_RC and OSS methods we find $\xi_d^* < 0.2$ (Figure 5). Thus, there is a significant improvement in the DDR for the FIC and SUPG methods for this value of C. It is interesting to note that for $C \leqslant 0.1$ the variation of DDR with C is insignificant. Similar conclusion can be made for the backward Euler and BDF2 schemes from their respective DDR plots.

It can be seen from the amplification plots (Figures 2, 4, 6 and 8) that *for the same value of* α (here $\alpha = 1.0$) and using a consistent **T** matrix, the damping associated with the FIC/SUPG method is relatively less than that of the FIC_RC/OSS method, though the gain is not significant for low wavenumbers. On the other hand, using a lumped **T** matrix the difference in the amplification associated with FIC/SUPG and FIC_RC/OSS methods is insignificant. It can also be seen that unlike the notable differences in the location of ξ_d^*, the differences in the amplification due to lumping the **T** matrix is insignificant.

An important aspect is the effect of group velocity. This is more evident when the transported function is periodic and resembles a sinusoid wave train. For such functions, the Fourier transform is a narrow peak concentrated around the characteristic angular wave number of the function. For such problems the effect of the group velocity is more significant than that of phase velocity. If the DDR predicts a deviation in the group velocity then the wave train travels at that deviant velocity (Example 1.6.3).

1.4 Stabilization Parameters

In this section, the optimal expressions of the stabilization parameters for the FIC_RC method in 1D and on a uniform mesh are proposed. We consider the steady-state form of the discrete problem (1.17) for the sourceless case ($f = 0$). The equation stencil for an interior node j is as follows:

$$(\frac{u}{2})[\phi_{j+1} - \phi_{j-1}] - (k + \frac{uh}{2})[\frac{\phi_{j+1} - 2\phi_j + \phi_{j-1}}{\ell}] + (\frac{uh}{8})[\frac{\phi_{j+2} - 2\phi_j + \phi_{j-2}}{\ell}] = 0 \tag{1.27}$$

The analytical solution of the steady-state form of problem (1.17) in the 1D space with only Dirichlet boundary conditions and source $f = 0$ is,

$$\phi(x) = \phi_l^p + (\phi_r^p - \phi_l^p)\left[\frac{\exp[ux/k] - 1}{\exp[uL/k] - 1}\right] \tag{1.28}$$

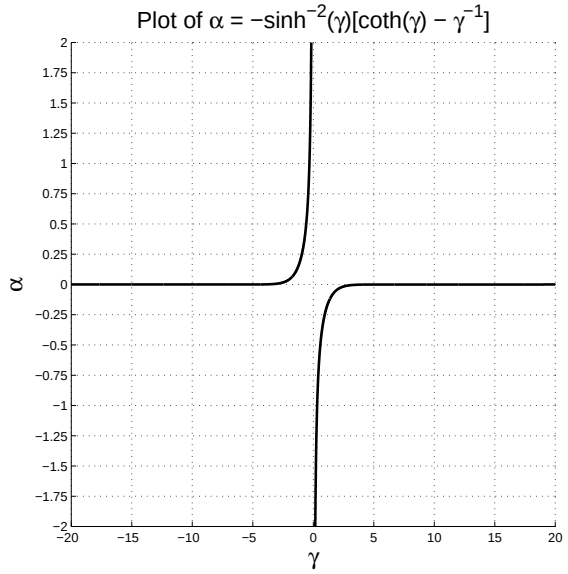


Figure 10: Optimal expression of α for the FIC_RC/OSS method. Plot of α vs. γ for the interior nodes.

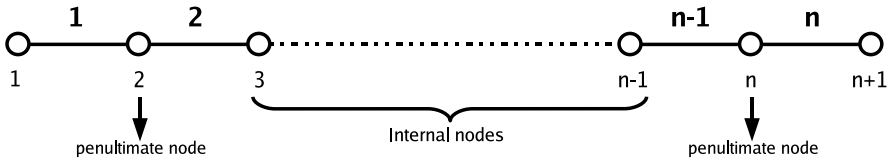


Figure 11: Node stencil for the 1D problem.

where L is the length of the 1D domain and ϕ_l^p,ϕ_r^p are the prescribed values of ϕ at the left and right ends of the domain respectively. We now express the characteristic length in terms of the element size as $h = \alpha\,\ell$. The optimal value of the stabilization parameter α is found by substituting the analytical solution into the stencil given in Eq.(1.27). This leads to the following expression of α for the interior nodes.

$$\alpha = \frac{-1}{\sinh^2(\gamma)}\left[\coth(\gamma) - \frac{1}{\gamma}\right] \tag{1.29}$$

where, γ is the element *Peclet number* given by $\gamma = \frac{u\ell}{2k}$. The plot of the parameter α vs. γ is shown in Figure 10.

It can be seen that $\alpha \approx 0$ for $\gamma \geqslant 3$. So for high Peclet numbers the formulation suggests to use very small values of α. That is for large values of γ the formulation breaks down into the standard Galerkin method without stabilization and one can expect spurious oscillations in the numerical solution. The clues to reason out this behavior can be found by examining the assembly of the linear system.

The 1D problem is discretized by linear elements and the final form of the finite element assembly is examined. For simplicity, the nodes are numbered from left to right as shown in Figure 11. For the interior nodes, the equation stencil is as given by Eq.(1.27). For the left penultimate boundary node, here node 2 as per the numbering scheme Figure 11, we find the following stencil:

$$(\frac{u}{2})[\phi_3 - \phi_1] - (k + \frac{uh}{2})[\frac{\phi_3 - 2\phi_2 + \phi_1}{\ell}] + (\frac{uh}{8})[\frac{\phi_4 - 3\phi_2 + 2\phi_1}{\ell}] = 0 \tag{1.30}$$

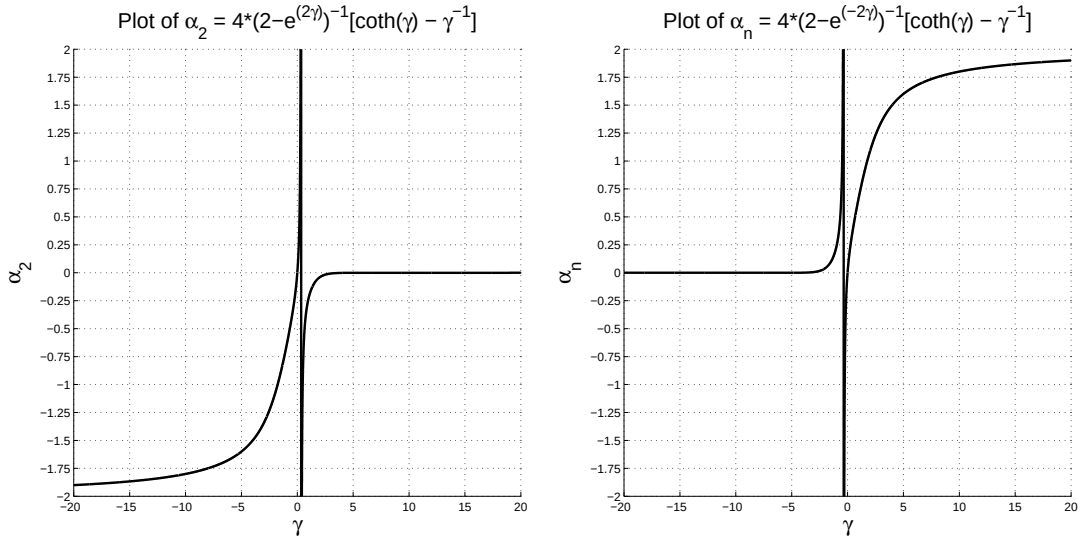


Figure 12: Plot of optimal α (for the penultimate boundary nodes) vs. γ for the FIC_RC/OSS method.

For the right penultimate boundary node, i.e the node n in Figure 11, we find the following stencil:

$$(\frac{u}{2})[\phi_{n+1} - \phi_{n-1}] - (k + \frac{uh}{2})[\frac{\phi_{n+1} - 2\phi_n + \phi_{n-1}}{\ell} + (\frac{uh}{8})[\frac{2\phi_{n+1} - 3\phi_n + \phi_{n-2}}{\ell}] = 0 \quad (1.31)$$

Thus, we see that there is a rearrangement of the equation stencils for the nodes that lie next to the boundary. The stencils for the boundary nodes also gets rearranged, but we do not consider them here as the focus here is to deal problems with Dirichlet boundary conditions. The deviation of the nodal equations for the penultimate nodes from the interior nodes is responsible for the spurious oscillations. It has been shown using a spectral analysis framework in [163, 164] that the asymmetry in the stencils brings about anti-diffusion even when they are used with central difference schemes and their effect is not localized, thus being responsible for spurious numerical oscillations. A simpler explanation would be that the rearrangement of the stencils near the boundary require different expressions for the stabilization parameter for those nodes.

The optimal values of the stabilization parameters for these penultimate nodes on a uniform mesh are as follows. Eqs.(1.32a) and (1.32b) correspond to the optimal values for the nodes 2 and n respectively. Figure 12 illustrates the variation of α with respect to γ.

$$\alpha_2 = \frac{4}{[2 - e^{2\gamma}]}\left[\coth(\gamma) - \frac{1}{\gamma}\right] \quad (1.32a)$$

$$\alpha_n = \frac{4}{[2 - e^{-2\gamma}]}\left[\coth(\gamma) - \frac{1}{\gamma}\right] \quad (1.32b)$$

An alternative to nodal stabilization parameters is to find optimal values of the stabilization parameters for the elements adjacent to the boundary. For elements laying

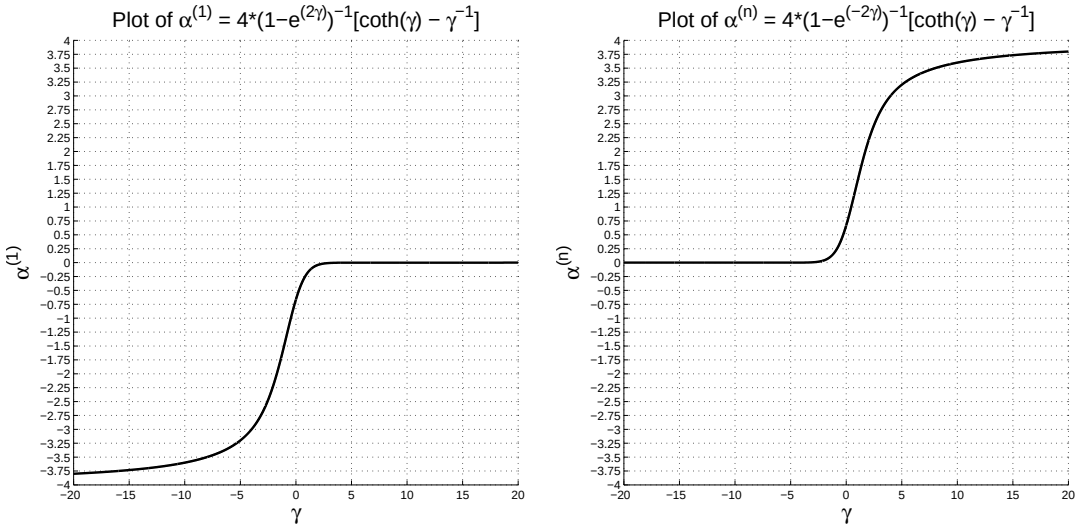


Figure 13: Plot of optimal α (for the elements adjacent to the boundary) vs. γ for the FIC_RC/OSS method.

in the domain interior we use the value given by Eq.(1.29). Denoting the stabilization parameters for elements 1, n and in the domain interior as $\alpha^{(1)}$, $\alpha^{(n)}$ and α respectively, we find the following equation stencils.

Node 2:

$$\begin{aligned}&(\frac{u}{2})[\phi_3 - \phi_1] - (\frac{k}{\ell})[\phi_3 - 2\phi_2 + \phi_1] \\ &- (\frac{u}{2})[\alpha\phi_3 - (\alpha^{(1)} + \alpha)\phi_2 + \alpha^{(1)}\phi_1] \\ &\qquad + (\frac{u}{8})[\alpha\phi_4 + (\alpha - \alpha^{(1)})\phi_3 - (\alpha + 2\alpha^{(1)})\phi_2 + (3\alpha^{(1)} - \alpha)\phi_1] = 0 \end{aligned} \tag{1.33}$$

Node n:

$$\begin{aligned}&(\frac{u}{2})[\phi_{n+1} - \phi_{n-1}] - (\frac{k}{\ell})[\phi_{n+1} - 2\phi_n + \phi_{n-1}] \\ &- (\frac{u}{2})[\alpha^{(n)}\phi_{n+1} - (\alpha^{(n)} + \alpha)\phi_n + \alpha\phi_{n-1}] \\ &\qquad + (\frac{u}{8})[(3\alpha^{(n)} - \alpha)\phi_{n+1} - (\alpha + 2\alpha^{(n)})\phi_n + (\alpha - \alpha^{(n)})\phi_{n-1} + \alpha\phi_{n-2}] = 0 \end{aligned} \tag{1.34}$$

The optimal values for $\alpha^{(1)}$ and $\alpha^{(n)}$ are given by the Eqs.(1.35a) and (1.35b) respectively. Figure 13 illustrates the variation of α with respect to γ.

$$\alpha^{(1)} = \frac{4}{[1 - e^{2\gamma}]}\left[\coth(\gamma) - \frac{1}{\gamma}\right] \tag{1.35a}$$

$$\alpha^{(n)} = \frac{4}{[1 - e^{-2\gamma}]}\left[\coth(\gamma) - \frac{1}{\gamma}\right] \tag{1.35b}$$

Remark: It is important to note that α is a signed parameter. For the domain interior, α is negative for positive values of γ! As the optimal α on a uniform grid for the FIC_RC/OSS method is close to zero everywhere within the domain interior (Figure

10) we do not introduce any artificial (unnatural) damping anywhere within the domain. The effect of boundary shock layers can also be captured efficiently by the new expressions of α at the boundary-adjacent elements. Thus, we see that in the transient mode the performance of the FIC_RC/OSS method using the proposed optimal α with boundary correction is similar to the standard Galerkin method. The former method also gives nodally exact solutions (on uniform grids) in the steady state mode unlike the spurious global oscillations produced by the later method. Of course, all these features are favorable addenda only if the bandwidth of the amplitude spectra of the transported function lies within the range of the phase departure wave number.

1.5 DISCRETE MAXIMUM PRINCIPLE

On further examination, the optimal expression (on uniform grids in 1D) of the stabilization parameter for the FIC_RC/OSS method illustrates certain subtle features related to the satisfaction of a discrete maximum principle (DMP). If the DMP holds for any numerical method then the maximum component in the modulus of the solution vector is bounded above by the maximum component in the modulus of the boundary data. It is well known that a sufficient condition for the satisfaction of the DMP is that the system matrix be an *M-matrix* or a matrix of '*positive type*'. We refer to [17, 40] for the definition of the latter matrices.

Note that for Peclet numbers greater than one, the system matrix (see Eq.(1.27)) associated with the FIC_RC/OSS method can never be cast into the form of an M-matrix or a matrix of 'positive type'. However, using the proposed optimal expression for α we get nodally exact solutions on a uniform mesh in 1D. It is remarkable that in this case the system matrix is not even a *monotone matrix* whose definition can be found in [17, 40]. Nevertheless, the FIC_RC/OSS system matrix obtained using the optimal α verifies the necessary and sufficient condition given in [186] for the DMP to hold. Unfortunately, this necessary and sufficient condition for a DMP to hold is difficult to identify a priori and thus poising a strategical difficulty in the design of discontinuity/shock capturing methods.

1.6 NUMERICAL EXAMPLES

1.6.1 *Example 1*

In this example, we study the effect of the stabilization introduced by FIC/SUPG and FIC_RC/OSS methods for transient problems. We study the pure convection problem primarily for two reasons. First, the problem becomes simple as the dispersion effect of natural diffusion is sidelined, thus allowing us to study the effects of the diffusion introduced by the stabilization methods. Next, the convection dominated problem is the primary concern of stabilization methods. The domain of interest is $x \in [0, 10]$. The problem data are $k = 1 \times 10^{-30}$, $u = 1.0$, time increment $\theta = 0.01$ and the 1D space is discretized by 100 linear elements with uniform mesh size. Thus, $\ell = 0.1$. The Peclet number for this problem is $\gamma = 5 \times 10^{28}$ ($\gamma \approx \infty$). The Courant number is $C = 0.1$. We have also chosen $\alpha = 1.0$, the optimal value for SUPG in this case, throughout the simulations. This choice is made just to study the effect of the artificial diffusion introduced by the FIC_RC/OSS method within the interior of the

domain using a non-optimal value for α. Recall that the optimal value of α for the FIC_RC/OSS method in this case is ≈ 0, thus behaving just like the standard Galerkin method in the domain interiors. The discretization in time is done by the following schemes: Crank-Nicholson, backward Euler and BDF2 schemes.

Figure 14 shows the results obtained when a narrow Gaussian pulse centered at $x = 3.0$ is taken as the initial solution. The equation for the initial solution is $\phi_o(x) = \exp[-8(x-3)^2]$. The numerical solution at time 3s is examined. The solutions obtained using a lumped and consistent **T** matrix in the formulation are also examined. The idea is to validate the conclusions that can be drawn from the DDR plots. The pulse width of the Gaussian function is chosen as to guarantee that the bandwidth of the amplitude spectra of this function is just less than $\xi_d^{consistent}$, the phase departure wave-number using a consistent **T** matrix. As $\xi_d^{lumped} \leqslant \xi_d^{consistent}$, one should expect incorrect superpositions of the wave trains due to their phase differences from the former. This leads to a train of wiggles as seen in the numerical solution when **T** is lumped (Figure 14).

It is interesting that the standard Galerkin method without any stabilization and by using a consistent **T** matrix yields very accurate results. The other methods using a consistent **T** create a slight bump at the foot of the Gaussian bell. This is because of the damping associated with those methods for the significant wavenumbers ($\xi^* \leqslant \xi_d^*$). We also notice that using a consistent **T**, the damping associated with the FIC/SUPG method is less than that for the FIC_RC/OSS method. The differences are insignificant for the **T**-lumped case.

1.6.2 *Example 2*

In this example, we illustrate the effect of the variation of the Courant number on the DDR. The problem data are the same as in Example 1.6.1. Time integration is performed by the Crank-Nicholson scheme. Two initial functions are considered : a narrow Gaussian pulse centered at $x = 3.0$ as defined in Example 1.6.1 and a square pulse function defined by $\phi_o(x) = 1.0$ if $x \in [2, 4]$ else $\phi_o(x) = 0.0$. The spectra of the square pulse is broad and the bandwidth extends beyond the ξ_d of all the methods considered here. In other words, in the absence of damping this function will exhibit numerical dispersion. In this example only the consistent **T** matrix is used.

First we note that for the higher Courant number (here $C = 1.0$) the Gaussian pulse exhibit numerical dispersion even when a consistent **T** matrix is used (Figure 15). As C is reduced to 0.4 and 0.1 we notice that the dispersion errors are minimized. As discussed earlier in §1.3.4 the FIC/SUPG method with $C = 0.4$ should exhibit a better performance over the FIC_RC/OSS method. Unfortunately the gain in the DDR for the higher wavenumbers does not materialize in the simulated results. The solutions for the FIC/SUPG and the FIC_RC/OSS methods are nearly the same for both the initial solutions (Figure 15). This is because all those wavenumbers suffer high damping. As $\mathfrak{I}(\omega_h^*)$ does not vary with C (Figure 6), Figure 2 may be referred for the amplification.

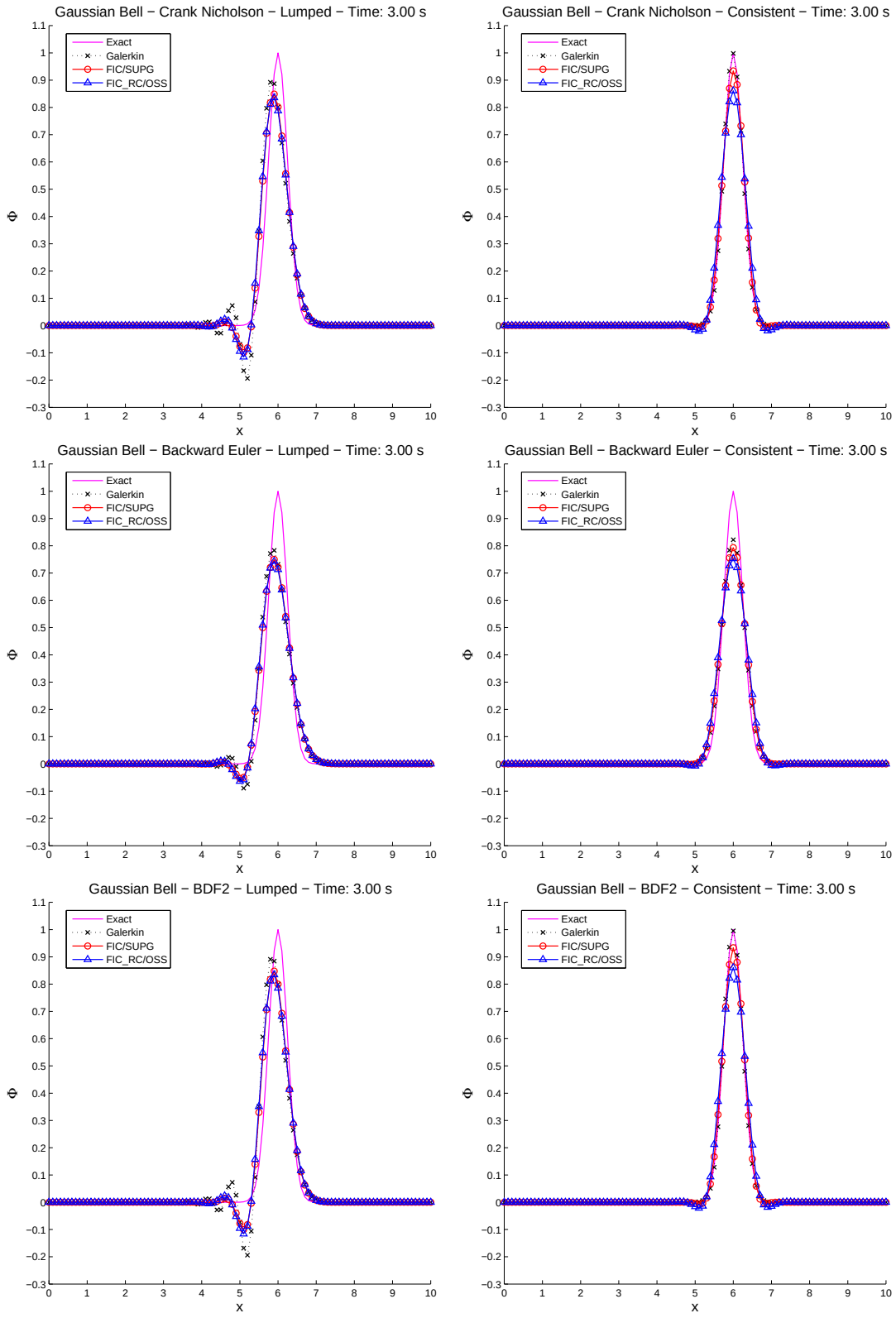


Figure 14: Example 1: transport of a Gaussian pulse. Solutions at 3s for the Galerkin, FIC/-SUPG and FIC_RC/OSS methods using lumped (on left) and consistent (on right) **T** matrix are shown; Time discretization is done by the Crank-Nicholson, backward Euler and BDF2 schemes. $\alpha = 1.0$ is used.

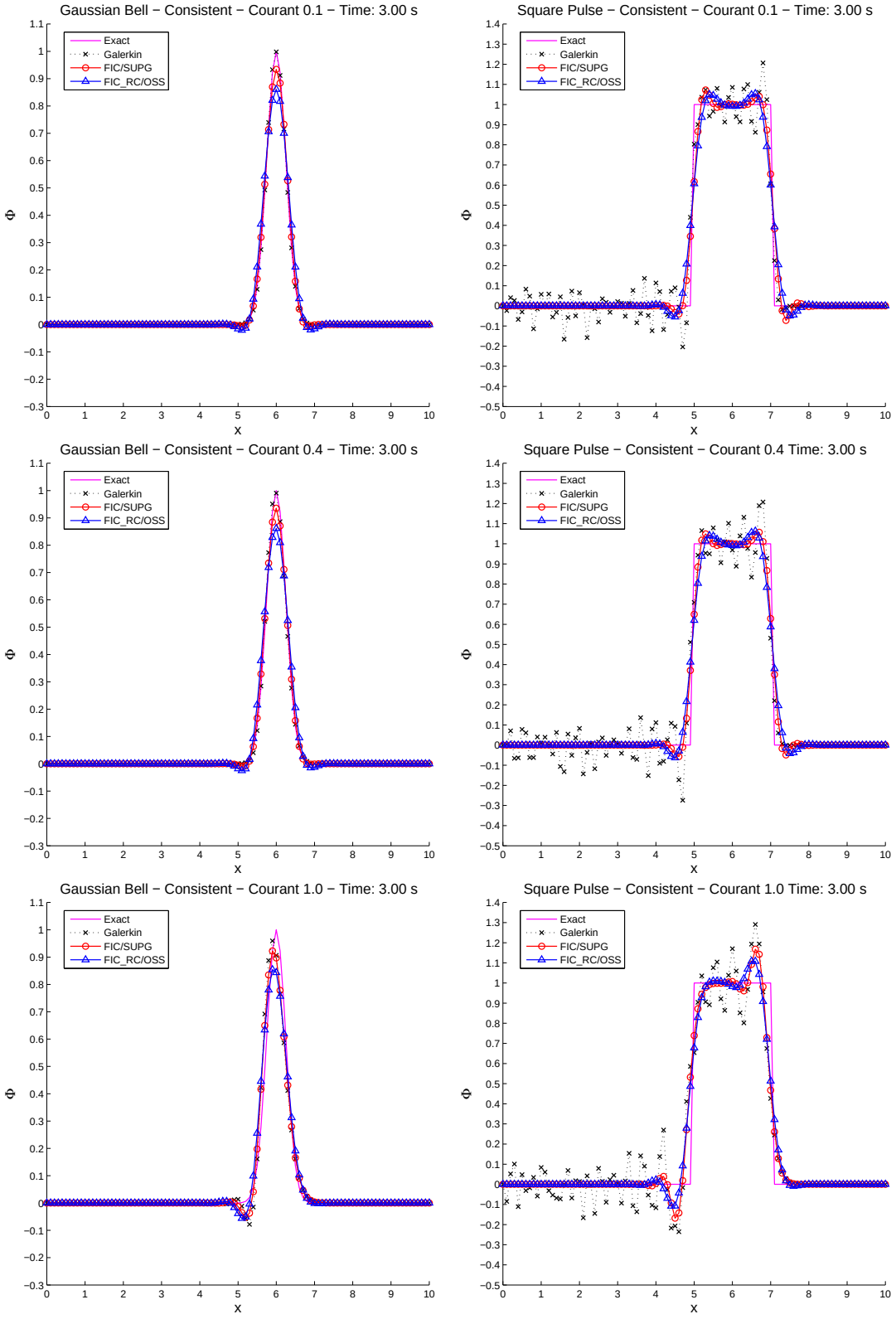


Figure 15: Example 2: transport of a Gaussian and square pulse. Solutions at 3s for the Galerkin, FIC/SUPG and FIC_RC/OSS methods using a consistent $\mathbf{T}$ matrix and for Courant number 0.1, 0.4 and 1.0; Time discretization performed by Crank-Nicholson scheme. $\alpha = 1.0$ is used.

1.6.3 *Example 3*

In this example, we study the transport of a periodic sine wave with angular wavenumber $\xi_o = 7.5$. The problem data is the same as in Example 1.6.1. The corresponding value of $\xi^* = 0.238$. Time integration is performed by the Crank-Nicholson scheme. The idea here is to study the effect of the group velocity in the numerical simulation. The initial function and the boundary condition prescribed at the left boundary are as follows:

$$f(x,0) = \sin(\xi_o x) \tag{1.36}$$

$$f(0,t) = \sin(u\xi_o t) \tag{1.37}$$

The DDR predicts a group velocity $V_g^{lumped} \approx 0.75$ and $V_g^{consistent} \approx 1.0$ for $\xi^* = 0.238$ using a lumped and consistent $\mathbf{T}$ matrices, respectively (Figure 9). This property is exhibited in the numerical solution where we find that the wave train moves at a different speed from the one assigned. The results are accurate using a consistent $\mathbf{T}$ matrix (Figure 16). The damping introduced by the stabilization methods are also in agreement with the one predicted by the DDR of the problem. Using a consistent $\mathbf{T}$ matrix the amplifications for the FIC/SUPG and the FIC_RC/OSS methods for $\alpha = 1.0$ and the wavenumber $\xi_o = 7.5$ are ≈ 0.7 and 0.3, respectively, after 3s (Figure 2). The corresponding values for the $\mathbf{T}$-lumped case are ≈ 0.4 and 0.32. The numerical results are in agreement with this prediction. We note that the numerical damping associated with the FIC/SUPG method is less than that of the FIC_RC/OSS method for the same value of α. An interesting result is that when the expressions (Eqs. 1.29,1.35) for α are used for the FIC_RC/OSS method, no damping takes place as it behaves similar to the Galerkin method in the interior domain. The numerical solution in this case coincides with that shown for the Galerkin method.

1.6.4 *Example 4*

In this example, we explore the performance of the stabilization parameter α given by Eqs. (1.29) and (1.35) for the steady state problem using the FIC_RC/OSS method. The domain of interest is $x \in [0.0, 1.0]$. The problem data are $k = 0.001$, $u = 1.0$ and $f = 1.0$. For the ease of notation and further reference we define the following:

$$\alpha_a = \left[\coth(\gamma) - \frac{1}{\gamma}\right]$$

$$\alpha_b = \begin{cases} \dfrac{4}{[1-e^{2\gamma}]}\left[\coth(\gamma) - \dfrac{1}{\gamma}\right], & \text{element 1} \\ \dfrac{4}{[1-e^{-2\gamma}]}\left[\coth(\gamma) - \dfrac{1}{\gamma}\right], & \text{element n} \\ \dfrac{-1}{\sinh^2(\gamma)}\left[\coth(\gamma) - \dfrac{1}{\gamma}\right], & \text{else} \end{cases}$$

First, we study the solution on a uniform mesh consisting of 20 linear elements ($\ell = 0.05$). We consider the cases when $f = 0$ and $f = 1.0$. The numerical solutions of the

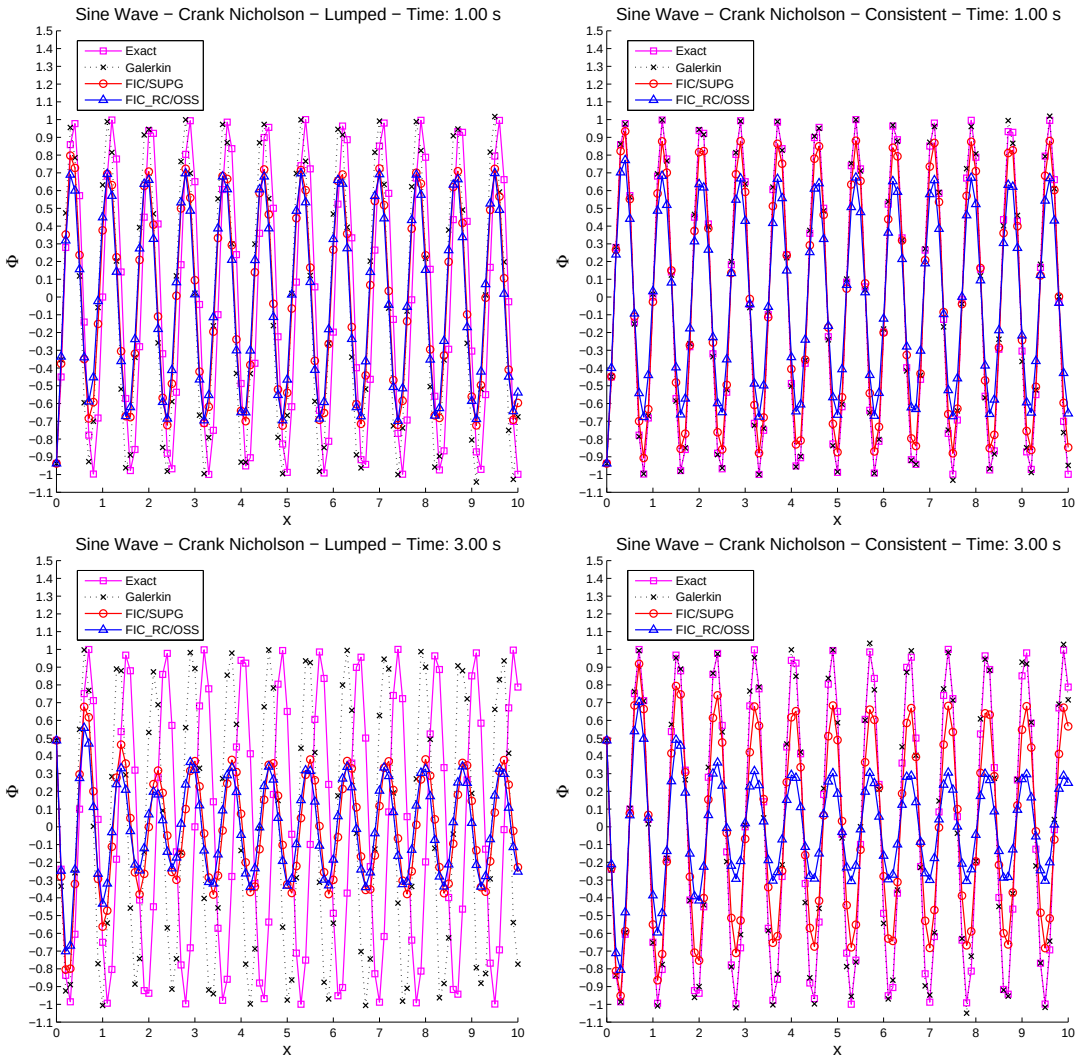


Figure 16: Example 3: transport of a sinusoidal wave. Solutions at 1s and 3s for the Galerkin, FIC/SUPG and FIC_RC/OSS methods using a lumped (on left) and consistent (on right) $\mathbf{T}$ matrix; Time discretization done by the Crank-Nicholson scheme. $\alpha = 1.0$ is used.

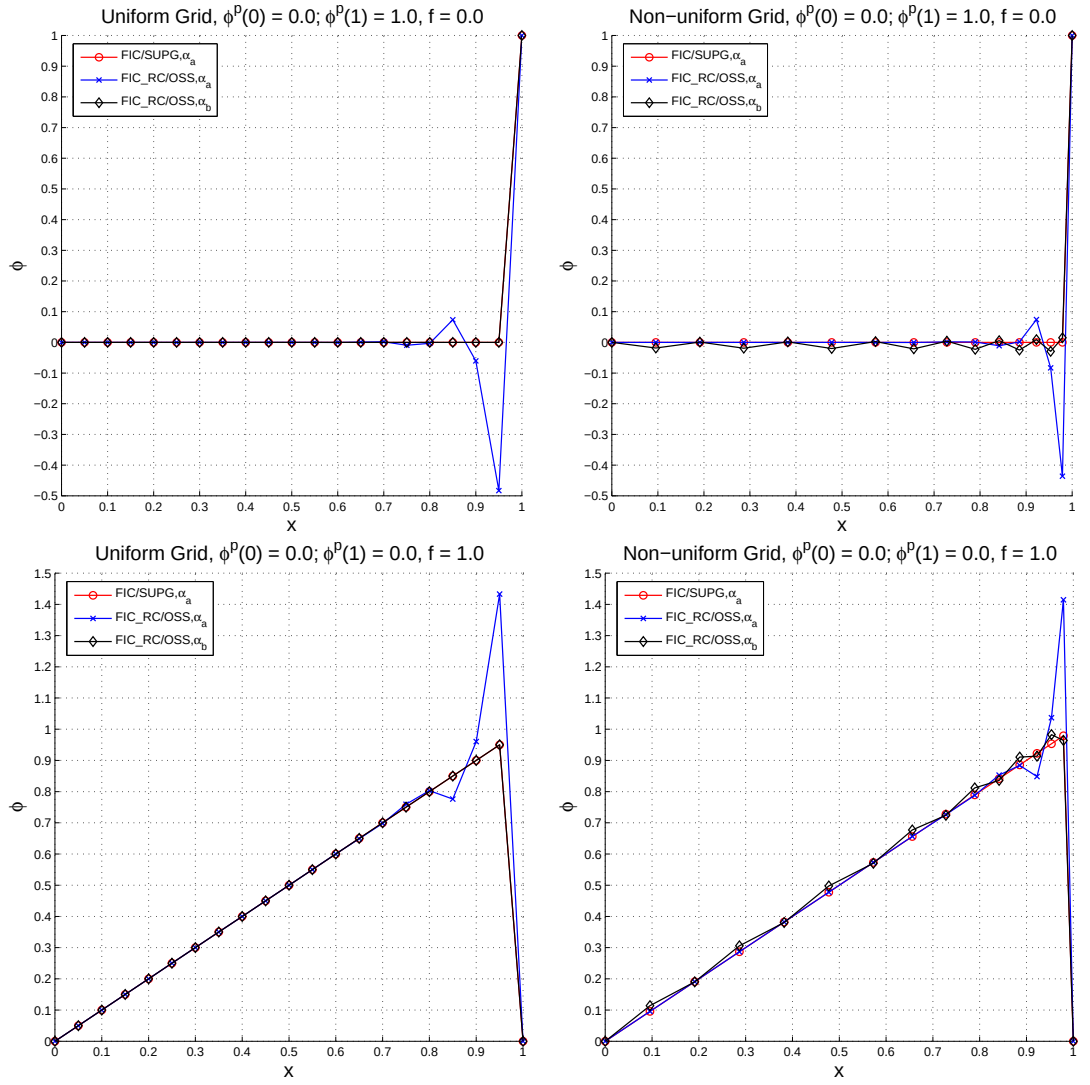


Figure 17: Example 4: steady-state solutions on uniform and nonuniform grids for the FIC/-SUPG and FIC_RC/OSS methods. $\alpha_a := [\coth(\gamma) - \gamma^{-1}]$. α_b is evaluated via the Eqs. (1.29) and (1.35)

FIC/SUPG method using the stabilization parameter as α_a and of the FIC_RC/OSS method using α_a and α_b are presented in Figure 17. We note that the new definition for the stabilization parameter α_b is optimal for the uniform mesh. The boundary correction introduced in Eq.(1.35) also takes effect.

Next, we study the solution on a non-uniform grid consisting of 15 elements. The node coordinates of the discrete 1D space are given by $x = \{0, 0.095, 0.191, 0.2866, 0.382, 0.477, 0.573, 0.656, 0.728, 0.789, 0.841, 0.885, 0.922, 0.953, 0.979, 1\}$. We note that the solution of the FIC/SUPG method is superior to that obtained by the FIC_RC/OSS method. The later method gives the sharp boundary oscillations using the parameter α_a (also appears on uniform grids). The solution of the FIC_RC/OSS method using α_b on the non-uniform grid is slightly corrupted with weak node-to-node spurious oscillations.

1.7 CONCLUSIONS

A detailed transient analysis of a consistency recovery method (FIC_RC/OSS) with respect to the Galerkin and FIC/SUPG methods has been done. The discrete dispersion relations for the above methods using the trapezoidal and BDF2 time integration schemes are presented. The phase departure wavenumber (ξ_d) is greater when a consistent matrix for the transient terms is used. ξ_d proportionally influences the range of wavenumbers that have a group velocity close to the convection velocity. The DDR plots predict that the gain in the value of ξ_d from the lumped $\mathbf{T}$ case to the consistent $\mathbf{T}$ case gradually decreases with the Courant number C. An exception to this is on a certain lower range of C for the FIC/SUPG method. Unfortunately, this gain is seldom realized as those higher wavenumbers are damped away. The contour plots of $\log[\Re(|\omega_h^* - \omega^*|)]$ are very similar for the Galerkin, FIC/SUPG and the FIC_RC/OSS methods except for the exception mentioned earlier. Neither is the change significant with the choice of the time integration schemes considered. This suggests that the role of the stabilization terms in the improvement of the DDR is insignificant and the enhancement can be achieved only by virtue of the resolution in space and time. It is shown that for the same value of the stabilization parameter α, the damping associated with the FIC_RC/OSS method is slightly greater than that of the FIC/SUPG method.

It is shown that, unlike the FIC/SUPG method, the FIC_RC/OSS method introduces a certain rearrangement in the equation stencils at nodes on and adjacent to the domain boundary. Thus using a uniform expression for α will lead to localized oscillations at the boundary. These oscillations are notable in the convection dominated case. When diffusion is at par with convection, the solution has a smooth profile and these local oscillations are insignificant even though they exist. A proposal for α that gives optimal results for the steady-state 1D convection diffusion problem on uniform grids is made for the FIC_RC/OSS method. An interesting result is that when the new expressions for α are used for the FIC_RC/OSS method, no damping takes place as $\alpha \approx 0$ in the domain interior. The numerical solution in the transient case coincides with that for the Galerkin method and in the steady state, unlike the Galerkin method, it is stable. Unfortunately, using this new expression for α, dispersion errors whenever present (for instance, the transport of a square pulse), cannot be controlled and also the steady state solution has weak node-to-node oscillations on a non-uniform grid. On the basis of these results, it appears that with respect to the stabilization of convection, the numerical performance of the FIC/SUPG method is better as one can control to some extent the dispersive errors and at the same time we can assure the stability of the steady-state solution.

Finally, it can be verified that the FIC_RC/OSS system matrix obtained using the optimal stabilization parameter (on uniform grids) is neither a monotone matrix nor a matrix of positive type. Although, it verifies the necessary and sufficient condition for a DMP to hold. Unfortunately, it is difficult to identify a priori the matrices that satisfy this necessary and sufficient condition unlike the matrices of 'positive type' that are easy to recognize. This poses a strategical difficulty in the design of discontinuity capturing methods. Due to these difficulties we currently do not prefer the consistency recovery method in the stabilization of convection.

What is karma? and what is not karma?
are questions that perplex the wisest of men.

— Bhagawad Gita IV:16.

2

A HIGH-RESOLUTION PETROV–GALERKIN METHOD IN 1D

2.1 INTRODUCTION

A singularly perturbed convection–diffusion–reaction problem is an initial-boundary value problem where the diffusion coefficient may take arbitrarily small values. The solution of this problem may exhibit transient and/or exponential boundary layers. In higher dimensions the solution may also exhibit characteristic boundary and/or interior layers. It is well known that the numerical solution of this problem by the Bubnov–Galerkin finite element method (FEM) is prone to exhibit *global*, *Gibbs* and *dispersive* oscillations. The solution of the stationary problem by the above method exhibits spurious global oscillations for the convection-dominated cases. The local Gibbs oscillations are exhibited along the characteristic layers for the convection-dominated cases. For the reaction-dominated cases Gibbs oscillations may be found near the Dirichlet boundaries and in the regions where the distributed source term is nonregular. The solution of the transient problem may exhibit dispersive oscillations should the initial solution and/or the distributed source term are nonregular.

In the context of variational formulations and weighted residual methods, control over the global instability has been achieved via the streamline-upwind Petrov–Galerkin (SUPG) [22, 92], Taylor–Galerkin [53], characteristic Galerkin [55, 127], Galerkin least squares (GLS) [95], bubble functions [12, 18, 19], variational multiscale (VMS) [91, 97], characteristic-based split (CBS) [193] and finite calculus (FIC) based methods [141]. A thorough comparison of some of these methods can be found in [32]. Close connections between the VMS method and stabilization via bubble functions was pointed out in [20]. It was shown that some of the above stabilized methods can be recovered using the FIC equations via an appropriate definition of the stabilization parameters [141, 146]. Nevertheless nonregular solutions continue to exhibit the Gibbs and dispersive oscillations.

Several shock-capturing nonlinear Petrov–Galerkin methods were proposed to control the Gibbs oscillations observed across characteristic internal/boundary layers for the convection-diffusion problem [23, 29, 46, 48–50, 68, 94, 116, 119, 132, 154]. A thorough review, comparison and state of the art of these and several other shock-capturing methods for the convection-diffusion equations, therein named as *spurious oscillations at layers diminishing* methods, was done in [114]. Reactive terms were not considered in the design of these methods and hence they fail to control the localized oscillations in the presence of these terms. Exceptions to this are the *consistent approximate upwind* (CAU) method [68], the methods presented in [23, 30] and those that take the CAU method as the starting point [48–50]. Nevertheless the expressions for the stabilization parameters therein were never optimized for reactive instability and often the solutions are over-diffusive in these cases.

In the quest to gain reactive stability several methods were built upon the existing frameworks of methods that control Global oscillations. Following the framework of the SUPG method linear Petrov–Galerkin methods were proposed for the convection–diffusion–reaction problem, viz. the DRD [180] and (SU+C)PG [100] methods. Based on the GLS method linear stabilized methods were proposed, viz. the GGLS method [61] for the diffusion–reaction problem and GLSGLS method [80] for the convection-diffusion-production problem. Within the framework of stabilization via bubbles the USFEM method [62] for the diffusion–reaction problem, the improved USFEM method [66] and the *link cutting bubbles* [21] for the convection–diffusion–reaction problem were proposed. Based on the VMS method linear stabilized methods were proposed for the convection–diffusion–reaction problem, viz. the ASGS method [33], the methods presented in [84, 85] and the SGS-GSGS method [86]. Using the FIC equations a nonlinear method based on a single stabilization parameter was proposed for the convection–diffusion–reaction problem [152, 155]. This nonlinear method, though initially formulated within the Petrov–Galerkin framework, subsequent modeling of a simplified form for the numerical nonlinear diffusion, deviated the method from being residual-based (consistency property is violated). Nodally exact Ritz discretizations of the 1D diffusion-absorption/production equations by variational FIC and modified equation methods using a single stabilization parameter were presented in [59]. Generally the homogeneous steady convection–diffusion–reaction problem in 1D has two fundamental solutions. Likewise, the characteristic equation associated with linear stabilized methods which result in compact stencils are quadratic and hence have two solutions. Thus in principle using two stabilization parameters (independent of the boundary conditions) linear stabilized methods which result in compact stencils can be designed to be nodally exact in 1D. Following this line several 'two-parameter methods' viz. (SU+C)PG, GLSGLS and SGS-GSGS methods were designed to be nodally exact for the stationary problem in 1D.

Control over the dispersive oscillations for the transient convection-diffusion problem via linear Petrov–Galerkin methods were discussed in [99] and using space-time finite elements in [189]. As for the linear methods, optimizing the expressions of the stabilization parameters to attain monotonicity will lead to solutions that are at most first-order accurate.

Out of the context of variational formulations and weighted residual methods, a vast literature exists on the design of high-resolution methods. These methods are characterized as *algebraic flux correction/limiting* methods and are usually developed with in the framework of finite-difference (FDM) or finite-volume (FVM) models. We refer to the books [89, 123, 124, 182] for a review of these methods and to the seminal papers in this field [16, 82, 83, 185, 191]. The book [120] describes the state of the art in the development of high-resolution schemes based on the Flux-Corrected Transport (FCT) paradigm for unstructured meshes and their generalization to the FEM. Nevertheless the use of these schemes were reported to be rather uncommon in spite of their enormous potential. We refer to the introduction in [161] that discusses the popularity of methods based on variational formulations and weighted residuals. Thus, as encouraged therein, the quest for the design of high-resolution methods based on variational/weighted-residual formulations is active to date.

In this chapter we present the design of a FIC-based nonlinear high-resolution Petrov–Galerkin (HRPG) method for the 1D convection–diffusion–reaction problem.

The prefix 'high-resolution' is used here in the sense popularized by Harten, i. e.second order accuracy for smooth/regular regimes and good shock-capturing in nonregular regimes. The goal is to design a numerical method within the context of the FIC variational formulation and weighted residuals which is capable of reproducing high-resolution numerical solutions for both the stationary (efficient control of global and Gibbs oscillations as seen in methods [21, 59, 80, 86, 100, 152, 155]) and transient regimes (efficient control of dispersive oscillations as seen in *algebraic flux correction/limiting* methods). In Section 2.2 we present the statement of the problem and the HRPG method in higher-dimensions. The statement in higher-dimensions is made only to distinguish the current method with the existing ones. The structure of the method in 1D is identical to the CAU method except for the definitions of the stabilization parameters. The method can be derived via the FIC approach [141] with an adequate (nonlinear) definition of the characteristic length. Thus the results presented here may be extended to these methods. In Section 2.4 we focus on the Gibbs phenomenon that is observed in L^2 projections. The design procedure embarks by defining a model L^2 projection problem and establishing the expression for the stabilization parameter to circumvent the Gibbs phenomenon. The target solution for the model problem is chosen to be the one obtained via the mass-lumping procedure. We remark that this solution is used to evaluate the stabilization terms introduced by the HRPG method and an expression for the stabilization parameter is defined that depends only on the problem data. In Section 2.5 we extend the methodology to the transient convection–diffusion–reaction problem. We split the design into four model problems and derive the stabilization parameters accordingly. Finally we arrive at an expression for the stabilization parameters depending only on the problem data and representing asymptotically the prior expressions derived for the model problems. We summarize the HRPG design in Section 2.5.6. In Section 2.5.7 several examples are presented that support the design objectives i. e.stabilization with high-resolution. Finally we arrive at some conclusions in Section 2.6.

2.2 HIGH-RESOLUTION PETROV–GALERKIN METHOD

The statement of the multidimensional convection–diffusion–reaction problem is as follows:

$$R(\phi) := \frac{\partial \phi}{\partial t} + \mathbf{u}\cdot\boldsymbol{\nabla}\phi - \boldsymbol{\nabla}\cdot(k\boldsymbol{\nabla}\phi) + s\phi - f(\mathbf{x}) = 0 \qquad \text{in} \quad \Omega \tag{2.1a}$$

$$\phi(\mathbf{x}, t=0) = \phi_0(\mathbf{x}) \qquad \text{in} \quad \Omega \tag{2.1b}$$

$$\phi = \phi^p \qquad \text{on} \quad \Gamma_D \tag{2.1c}$$

$$(k\boldsymbol{\nabla}\phi)\cdot\mathbf{n} + g^p = 0 \qquad \text{on} \quad \Gamma_N \tag{2.1d}$$

where $\mathbf{u}$ is the convection velocity, k,s are the diffusion and reaction coefficient respectively, $f(\mathbf{x})$ is the source, $\phi_0(\mathbf{x})$ is the initial solution, ϕ^p and g^p are the prescribed values of ϕ and the diffusive flux at the Dirichlet and Neumann boundaries respectively and $\mathbf{n}$ is the normal to the boundary.

The variational statement of the problem (2.1) can be expressed as follows: Find $\phi : [0, T] \mapsto V$ such that $\forall w \in V_0$ we have,

$$\Big(w, R(\phi)\Big)_\Omega + \Big(w, (k\boldsymbol{\nabla}\phi)\cdot\mathbf{n} + g^p\Big)_{\Gamma_N} = 0 \tag{2.2}$$

where, if H is the associated Hilbert space then $V := \{w : w \in H \text{ and } w = \phi^p \text{ on } \Gamma_D\}$, $V_0 := \{w : w \in H \text{ and } w = 0 \text{ on } \Gamma_D\}$, $(\cdot,\cdot)_\Omega$ and $(\cdot,\cdot)_{\Gamma_N}$ denote the $L^2(\Omega)$ and $L^2(\Gamma_N)$ inner products respectively. The problem (2.1) may also be expressed in the weak form as follows: Find $\phi : [0,T] \mapsto V$ such that $\forall w \in V_0$ we have,

$$a(w,\phi) = l(w) \tag{2.3a}$$

$$a(w,\phi) := \left(w, \frac{\partial\phi}{\partial t} + \mathbf{u}\cdot\boldsymbol{\nabla}\phi + s\phi\right)_\Omega + \left(\boldsymbol{\nabla}w, k\boldsymbol{\nabla}\phi\right)_\Omega \tag{2.3b}$$

$$l(w) := \left(w, f(x)\right)_\Omega - \left(w, g^p\right)_{\Gamma_N} \tag{2.3c}$$

The statement of the Galerkin method applied to the weak form of the problem (2.3) is: Find $\phi_h : [0,T] \mapsto V^h$ such that $\forall w_h \in V_0^h$ we have,

$$a(w_h,\phi_h) = l(w_h) \tag{2.4}$$

We follow [92] to describe a certain class of Petrov–Galerkin methods which account for weights that are discontinuous across element boundaries. The perturbed weighting function is written as $\tilde{w}_h = w_h + p_h$, where p_h is the perturbation that account for the discontinuities. The statement of these class of Petrov–Galerkin methods is as follows: Find $\phi_h : [0,T] \mapsto V^h$ such that $\forall w_h \in V_0^h$ we have,

$$a(w_h,\phi_h) + \sum_e \left(p_h, R(\phi_h)\right)_{\Omega_h^e} = l(w_h) \tag{2.5}$$

The HRPG method whose design in 1D is presented in the subsequent sections, may be defined as Eq.(2.5) along with the following definitions:

$$p_h := [\mathbf{h} + \mathbf{H}\cdot\hat{\mathbf{u}}^r]\cdot\boldsymbol{\nabla}w_h \tag{2.6a}$$

$$\mathbf{u}^r := \frac{R(\phi_h)}{|\boldsymbol{\nabla}\phi_h|^2}\boldsymbol{\nabla}\phi_h; \quad \Rightarrow \hat{\mathbf{u}}^r := \frac{\mathbf{u}^r}{|\mathbf{u}^r|} = \frac{\operatorname{sgn}[R(\phi_h)]}{|\boldsymbol{\nabla}\phi_h|}\boldsymbol{\nabla}\phi_h \tag{2.6b}$$

where, $\mathbf{h}$ and $\mathbf{H}$ are frame-independent linear *characteristic length* tensors that are defined based on the element geometry (see Section 3.3). We refer to the Table(2) for a comparison of the HRPG method with the SUPG, FIC and some of the existing shock-capturing methods. From Eqs.(2.6),(2.7) and Table(2) the HRPG method could be understood as the combination of upwinding plus a nonlinear discontinuity-capturing operator. The distinction is that in general the upwinding provided by $\mathbf{h}$ is not streamline and the discontinuity-capturing provided by $\mathbf{H}\cdot\hat{\mathbf{u}}^r$ is neither isotropic nor purely crosswind. Of course defining $\mathbf{h} := \tau\mathbf{u}$ and $\mathbf{H} := (\beta\ell)\mathbf{I}$ or $\mathbf{H} := (\beta\ell)[\mathbf{I} - \hat{\mathbf{u}}\otimes\hat{\mathbf{u}}]$ one would recover (except for the definitions of the stabilization parameters) the CAU and the CD methods respectively. We remark that one may arrive at the HRPG method via the finite-calculus (FIC) equations wherein the characteristic length is defined as $\mathbf{h}^{fic} := \mathbf{h} + \mathbf{H}\cdot\hat{\mathbf{u}}^r$. From this point of view the HRPG method can be presented as 'FIC-based'. More details are given in the next section.

Note that in 1D $\mathbf{u}^{\|} = \mathbf{u}$ and hence the performance of the DC method [94] is similar to that of the SUPG method. Also note that as the notion of crosswind directions does not exist in 1D, the CD method [29] is identical to the SUPG method. On the other hand the nonlinear shock-capturing terms introduced by the CAU method still exists in 1D and thus in principle are able to control the Gibbs and dispersive oscillations.

Method	Perturbation (p_h)	Remarks
SUPG[92]	$\tau \mathbf{u} \cdot \boldsymbol{\nabla} w_h$	
MH[132]	C_i^e	$C_i^e \in \{-\frac{1}{3}, \frac{2}{3}\}, i = 1,2,3$ $\sum C_i^e = 0$
DC[94]	$\tau_1 \mathbf{u} \cdot \boldsymbol{\nabla} w_h + \tau_2 \mathbf{u}^{\parallel} \cdot \boldsymbol{\nabla} w_h$	$\mathbf{u}^{\parallel} := \frac{\mathbf{u} \cdot \nabla \phi_h}{\lvert\boldsymbol{\nabla} \phi_h\rvert^2} \boldsymbol{\nabla} \phi_h$
CAU[68], CCAU[50]	$\tau_1 \mathbf{u} \cdot \boldsymbol{\nabla} w_h + \tau_2 \mathbf{u}^{r} \cdot \boldsymbol{\nabla} w_h$	$\mathbf{u}^{r} := \frac{R(\phi_h)}{\lvert\boldsymbol{\nabla} \phi_h\rvert^2} \boldsymbol{\nabla} \phi_h$
CD[29]	$\tau_1 \mathbf{u} \cdot \boldsymbol{\nabla} w_h + \alpha_2 \ell \boldsymbol{\nabla} w_h \cdot [\mathbf{I} - \hat{\mathbf{u}} \otimes \hat{\mathbf{u}}] \cdot \hat{\mathbf{u}}^{r}$	$\hat{\mathbf{u}} := \frac{\mathbf{u}}{\lvert\mathbf{u}\rvert}$ $\hat{\mathbf{u}}^{r} := \frac{\mathbf{u}^{r}}{\lvert\mathbf{u}^{r}\rvert} = \frac{\operatorname{sgn}[R(\phi_h)]}{\lvert\boldsymbol{\nabla} \phi_h\rvert} \boldsymbol{\nabla} \phi_h$
SAUPG[48], Mod.CAU[29]	$\tau[\lambda \mathbf{u} + (1-\lambda)\mathbf{u}^{r}] \cdot \boldsymbol{\nabla} w_h$	λ is a smoothness measure.
FIC[141]	$\mathbf{h}^{fic} \cdot \boldsymbol{\nabla} w_h$	here $\mathbf{h}^{fic}$ is a characteristic length vector which may be defined in a linear or nonlinear fashion.
HRPG	$[\mathbf{h} + \mathbf{H} \cdot \hat{\mathbf{u}}^{r}] \cdot \boldsymbol{\nabla} w_h$	$\mathbf{h}$,$\mathbf{H}$ are frame-independent linear characteristic length tensors based on the element geometry (see Section 3.3).

Table 2: Perturbations associated with Petrov–Galerkin methods

This feature does carry over to all the methods that have the shock-capturing term similar to that in the CAU method viz. the methods presented in [23, 48–50]. Unfortunately as pointed out in [114] and in Section 2.5.7.1 of this chapter, these methods are often over diffusive. The structure of the HRPG method in 1D is identical to the CAU method except for the definitions of the stabilization parameters. In the subsequent sections we design the stabilization parameters of the HRPG method to overcome the shortcomings of the earlier methods.

From Eqs.(2.5) and (2.6) the statement of the HRPG method in 1D can be expressed as follows: Find $\phi_h : [0,T] \mapsto V^h$ such that $\forall w_h \in V_0^h$ we have,

$$a(w_h,\phi_h) + \sum_e \left[\left(\frac{\alpha \ell}{2} \frac{dw_h}{dx}, R(\phi_h) \right)_{\Omega_h^e} + \left(\frac{\beta \ell}{2} \frac{|R(\phi_h)|}{|\nabla \phi_h|} \frac{dw_h}{dx}, \frac{d\phi_h}{dx} \right)_{\Omega_h^e} \right] = l(w_h) \quad (2.7)$$

where α , β are stabilization parameters to be defined later.

2.3 DERIVATION OF THE HRPG EXPRESSION VIA THE FIC PROCEDURE

The governing equations in the finite calculus (FIC) approach are derived by expressing the balance equations in a domain of finite size and retaining higher order terms. For the 1D convection–diffusion–reaction problem the FIC governing equations are written as [141]

$$R(\phi) - \frac{h}{2}\frac{\partial R(\phi)}{\partial x} = 0 \qquad \text{in} \quad \Omega \qquad (2.8a)$$

$$\phi(x, t=0) = \phi_0(x) \qquad \text{in} \quad \Omega \qquad (2.8b)$$

$$\phi - \phi^p = 0 \qquad \text{on} \quad \Gamma_D \qquad (2.8c)$$

$$k\frac{\partial \phi}{\partial x} + g^p + \frac{h}{2}R(\phi) = 0 \qquad \text{on} \quad \Gamma_N \qquad (2.8d)$$

where the characteristic length h is the dimension of the domain where balance of fluxes is enforced and $R(\phi)$ is defined in Eq.(2.1a).

The variational statement of Eqs.(2.8) can be written as: Find $\phi : [0,T] \mapsto V$ such that $\forall w \in V_0$ we have,

$$\left(w, R(\phi) - \frac{h}{2}\frac{\partial R(\phi)}{\partial x} \right)_\Omega + \left(w, k\frac{\partial \phi}{\partial x} + g^p + \frac{h}{2}R(\phi) \right)_{\Gamma_N} = 0 \qquad (2.9)$$

The corresponding weak form is: Find $\phi : [0,T] \mapsto V$ such that $\forall w \in V_0$ we have,

$$\left(w + \frac{h}{2}\frac{dw}{dx}, R(\phi) \right)_\Omega + \left(w, k\frac{\partial \phi}{\partial x} + g^p \right)_{\Gamma_N} = 0 \qquad (2.10)$$

In the derivation of Eq.(2.10) we have neglected the change of h within the elements. Clearly Eq.(2.10) can be seen as a Petrov–Galerkin form with the weighting function defined as $\tilde{w} := w + \frac{h}{2}\frac{dw}{dx}$. The term depending on h in Eq.(2.10) is usually computed in the element interiors only to avoid the discontinuities of the second derivatives terms in $R(\phi)$ along the element boundaries. The discretized form of Eq.(2.10) is therefore written as: Find $\phi_h : [0,T] \mapsto V^h$ such that $\forall w_h \in V_0^h$ we have,

$$a(w_h,\phi_h) + \sum_e \left(\frac{h}{2}\frac{dw_h}{dx}, R(\phi_h) \right)_{\Omega_h^e} = l(w_h) \qquad (2.11)$$

The characteristic length can be defined in a number of ways so as to provide an 'optimal' (stabilized) solution. In this work the following nonlinear expression is chosen for h:

$$h := \alpha\ell + \beta\ell\frac{\operatorname{sgn}[R(\phi_h)]}{|\boldsymbol{\nabla}\phi_h|}\boldsymbol{\nabla}\phi_h \tag{2.12}$$

where ℓ is the element length and α, β are stabilization parameters. A discussion of the alternatives for the definition of h in the FIC context can be found in [141, 152, 154, 155]. Substituting the expression of h into Eq.(2.11) gives the HRPG form of Eq.(2.7).

2.4 Gibbs phenomenon in L^2 projections

2.4.1 *Introduction*

Gibbs phenomenon is a spurious oscillation that occurs when using a truncated Fourier series or other eigen function series at a simple discontinuity. It is characterized by an initial overshoot and then a pattern of undershoot-overshoot oscillations that decrease in amplitude further from the discontinuity. In fact for any given function f and using the metric as the standard L^2 norm, the partial sum of order N of the Fourier series of f denoted as $S_N f$ is the *best approximation* of f in a subspace spanned by trigonometric polynomials of order N. Thus $S_N f$ is the L^2 projection of f in the considered subspace. This phenomenon is manifested due to the lack of completeness of the approximation space. Similar oscillations appear in the problem of finding the best approximation of a given discontinuous function in any subspace using the L^2 norm as the metric. On every discrete grid/mesh the maximum wavenumber that can be represented is limited by the Nyquist limit. The Nyquist frequency on a uniform grid with grid spacing ℓ is given by π/ℓ. Thus the span of the finite element basis functions associated with this mesh might be viewed as a truncated function series which might be expanded by refining the mesh. Hence the projection of a discontinuous function onto this finite element space exhibits the Gibbs phenomenon. As the amplitude spectrum of a discontinuous function decays only as fast as the *harmonic series*, which is not absolutely convergent, it is impossible to circumvent these oscillations by mere mesh refinement.

The variational statement of the L^2 projection problem is as follows: Find $\phi_h \in V^h$ such that $\forall w_h \in V_0^h$ we have,

$$\Big(w_h, \phi_h - f\Big)_{\Omega_h} = 0 \tag{2.13}$$

Where f is the given function which might admit discontinuities. If we denote the solution of Eq.(2.13) as $\phi_h = P_h^0 f$, we have,

$$\| P_h^0 f - f \|_{L^2(\Omega_h)} \leqslant \| w_h - f \|_{L^2(\Omega_h)} \tag{2.14}$$

In the following sections we consider the scaled L^2 projection problem defined by the residual $R(\phi) = s\phi - f$. In the context of the convection–diffusion–reaction problem, the problem data s physically represents the reaction coefficient. The variational statement of the scaled L^2 projection problem is as follows: Find $\phi_h \in V^h$ such that $\forall w_h \in V_0^h$ we have,

$$\Big(w_h, s\phi_h - f\Big)_{\Omega_h} = 0 \tag{2.15}$$

Taking $s = 1$ we recover the L^2 projection problem given by Eq.(2.13).

2.4.2 *Galerkin Method*

2.4.2.1 *FE discretization*

Discretization of the space by linear finite elements will lead to the approximation $\phi_h = N^a \Phi^a$ and the Eq.(2.15) reduces into the following system of equations.

$$s\mathbf{M} \cdot \mathbf{\Phi} = \mathbf{f} \tag{2.16}$$

$$\mathbf{M}^{ab} = \left(N^a, N^b\right)_{\Omega_h} \quad ; \quad \mathbf{f}^a = \left(N^a, f\right)_{\Omega_h} \tag{2.17}$$

It is well known that the Gibbs oscillations can be circumvented in the numerical solution if the standard row-lumping technique is performed on the mass matrix $\mathbf{M}$. Unfortunately, this operation though effective for this specific problem, it cannot be extended to other problems in general. The answer for not advocating this technique can be found in the Godunov's theorem: 'All linear monotone schemes are at most first order accurate'. The only way to circumvent this problem is to design a nonlinear method that would reproduce the same numerical solution as obtained by mass-lumping.

2.4.2.2 *Model Problem 1*

The 1D domain is chosen to be of unit length and discretized by $4N$ linear elements ($N > 5$). The function whose L^2 projection is sought is defined as follows:

$$f(x) = \begin{cases} 0 & \forall \quad x \in [0, 0.25 + \eta_1 \ell] \cup [0.75 - \eta_2 \ell, 1] \\ q & \text{else} \end{cases} \tag{2.18}$$

Where $\ell = 1/(4N)$ is the element length and $\eta_1, \eta_2 \in [0, 1]$ are parameters that determine the location of the simple discontinuity in the function f. The solution of Eq.(2.16) using a lumped mass matrix can be expressed as follows:

$$\mathbf{\Phi} = (\frac{q}{s})\{0, \cdots, 0, \frac{(1-\eta_1)^2}{2}, \frac{(2-\eta_1^2)}{2}, 1, \cdots, 1, \frac{(2-\eta_2^2)}{2}, \frac{(1-\eta_2)^2}{2}, 0, \cdots, 0\} \tag{2.19}$$

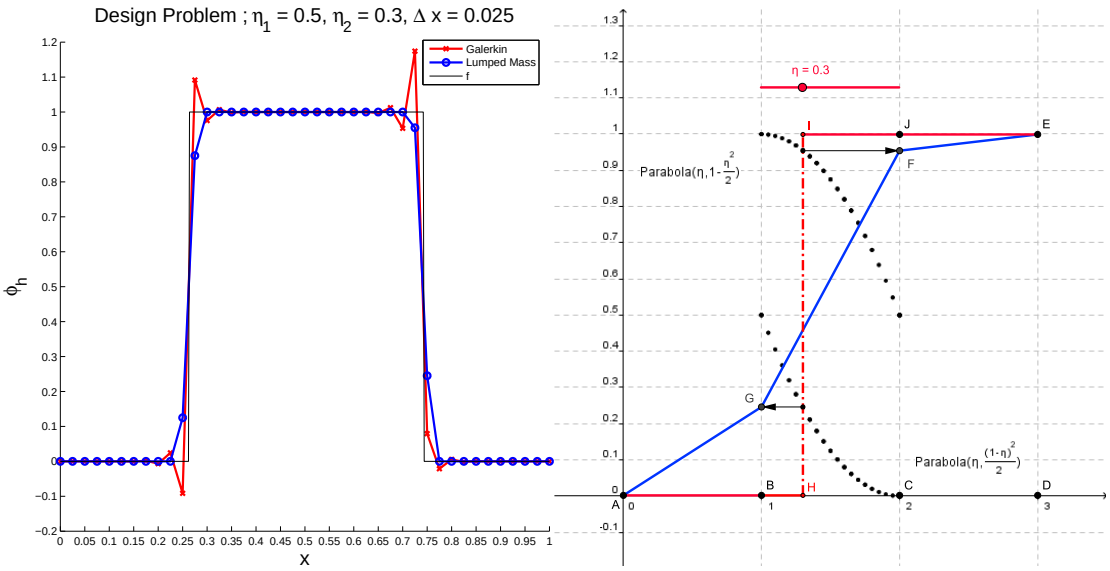
Figure (18a) illustrates the function $f(x)$ using $\eta_1 = 0.5$, $\eta_2 = 0.3$ and $q = s = 1$ alongside the numerical solution of the Eq.(2.16) using both the consistent and lumped mass matrix. Figure (18b) illustrates the profile characteristics of the monotone solution obtained via mass-lumping (Eq.2.19) with respect to the location of the discontinuity.

2.4.2.3 *Discrete Upwinding*

Let the discrete system of equations be represented in the matrix form as follows:

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b} \tag{2.20}$$

Discrete upwinding is an algebraic operation to convert the system matrix $\mathbf{A}$ into an M-matrix [120]. Discrete upwinding is the least diffusive linear operation to produce



(a) (b)

Figure 18: (a) The design problem ; (b) Characteristics of the monotone solution (AGFE). ABHIJE illustrates the discontinuous regime of $f(x)$

an M-matrix. We denote the discrete upwinding operation on any given matrix $\mathbf{A}$ by $\mathrm{DU}(\mathbf{A})$. The discrete upwinding operation is performed by adding to the matrix $\mathbf{A}$ a discrete diffusion matrix $\tilde{\mathbf{D}}$ as follows:

$$\tilde{d}_{ii} = -\sum_{j \neq i} \tilde{d}_{ij}; \qquad \tilde{d}_{ij} = \tilde{d}_{ji} = -\max\{0, a_{ij}, a_{ji}\} \tag{2.21}$$

$$\mathrm{DU}(\mathbf{A}) = \quad \tilde{\mathbf{A}} \quad = \mathbf{A} + \tilde{\mathbf{D}} \tag{2.22}$$

It is interesting to note that the discrete upwinding operation on the mass matrix $\mathbf{M}$ will result in the mass-lumping operation.

$$\mathbf{M}^e = \frac{\ell}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad ; \quad \tilde{\mathbf{D}}^e = \frac{\ell}{6}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{2.23}$$

$$\mathrm{DU}(\mathbf{M}^e) = \mathbf{M}^e + \tilde{\mathbf{D}}^e \quad = \quad \frac{\ell}{6}\begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} = \mathbf{M}_L^e \tag{2.24}$$

2.4.2.4 *Total variation*

The total variation of a function, say $\phi(x)$, in 1D is given by the following equation:

$$\mathrm{TV}(\phi) = \int_x |\nabla \phi| dx \tag{2.25}$$

Thus it can be seen that the total variation, as the name suggests, measures the total hike or drop in the function profile as we traverse the 1D domain. It can also be noticed that any spurious oscillation in the numerical approximation of ϕ would cause the total variation to increase. Harten proved that a monotone scheme is total variation non-increasing (TVD) and a TVD scheme is monotonicity preserving [82]. To date various high-resolution schemes have been designed based on the TVD concept

often using flux/slope limiters. If linear finite elements are used to approximate the numerical solution ϕ_h the total variation may be calculated as follows:

$$TV(\phi_h) = \sum_i |\Phi_{i+1} - \Phi_i| \tag{2.26}$$

For the problem under consideration, the sufficient conditions given by Harten in [82] for a numerical scheme to be TVD drops down to the condition for the system matrix to be an M-matrix. Thus in the design of the high-resolution Petrov–Galerkin method we use the total variation of the numerical solution as *a posteriori* verification condition. The total variation of the given discontinuous function $f(x)$ in the design problem is $TV(f) = 2$.

2.4.3 *HRPG design*

In this section we design the stabilization parameters of the HRPG method given by Eq.(2.7) and choosing $\alpha = 0$. For the model problem described in Section 2.4.2.2 the statement of the method is as follows: Find $\phi_h \in V^h$ such that $\forall w_h \in V_0^h$ we have,

$$\Big(w_h, R(\phi_h)\Big)_{\Omega_h} + \sum_e \Big(\frac{\beta\ell}{2}\frac{|R(\phi_h)|}{|\nabla\phi_h|}\frac{dw_h}{dx}, \frac{d\phi_h}{dx}\Big)_{\Omega_h^e} = 0 \tag{2.27}$$

Where, $R(\phi_h) := s\phi_h - f$ is the residual and β is a stabilization parameter to be defined later. If the domain is discretized by linear finite elements the Eq.(2.27) can be expressed in the matrix form for each element as follows:

$$\Big[s\mathbf{M}^e + \mathbf{S}^e\Big] \cdot \mathbf{\Phi}^e = \mathbf{f}_g^e \tag{2.28}$$

where the corresponding matrices are defined as,

$$\mathbf{M}^e = \Big(N^a, N^b\Big)_{\Omega_h^e} = \frac{s\ell}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{2.29}$$

$$\mathbf{S}^e = (\frac{\beta\ell}{2})\Big(\frac{|R(\phi_h)|}{|\nabla\phi_h|}\frac{dN^a}{dx}, \frac{dN^b}{dx}\Big)_{\Omega_h^e} = \frac{k^*(\phi_h)}{\ell}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{2.30}$$

$$\mathbf{f}_g^e = \Big(N^a, f\Big)_{\Omega_h^e} \tag{2.31}$$

$$k^*(\phi_h) = \frac{\beta}{2}\Big(\frac{|R(\phi_h)|}{|\nabla\phi_h|}, 1\Big)_{\Omega_h^e} \tag{2.32}$$

$$\mathbf{f}_g = q\ell\,\{0, \cdots, 0, \frac{(1-\eta_1)^2}{2}, \frac{(2-\eta_1^2)}{2}, 1, \cdots, 1, \frac{(2-\eta_2^2)}{2}, \frac{(1-\eta_2)^2}{2}, 0, \cdots, 0\} \tag{2.33}$$

In order to design the parameter β we assume that the method converges to the solution given by Eq.(2.19). This is a fair assumption as it can be seen in Eq.(2.27) that the nonlinear Petrov–Galerkin term is symmetric subjected to the linearization as shown in Eq.(2.30) and hence there exists a β such that the effect of this term is equivalent to the discrete diffusion introduced by the discrete upwinding operation.

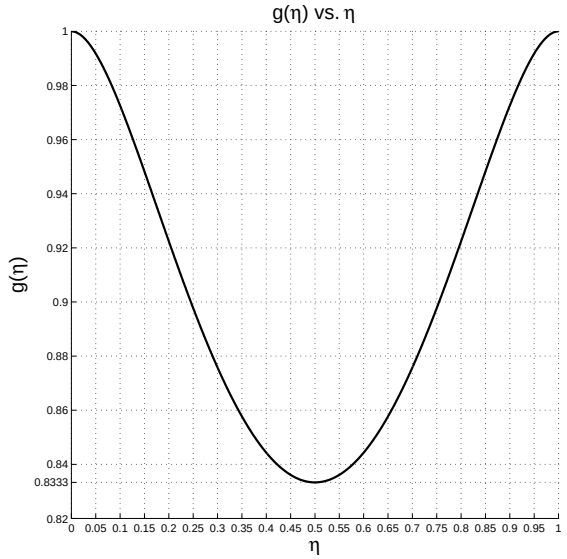


Figure 19: The plot of $g(\eta)$ for $\eta \in [0, 1]$.

If $\eta \in [0, 1]$ be a generic parameter to define the location of the simple discontinuity within the element, we have then for the element containing the discontinuity,

$$\begin{aligned}\left(\frac{|R(\phi_h)|}{|\nabla\phi_h|}, 1\right)_{\Omega_h^e} &= \left(\frac{s\ell^2}{2}\right)\left[\frac{1+2\eta-6\eta^2+8\eta^3-4\eta^4}{1+2\eta-2\eta^2}\right] \\ &= \left(\frac{s\ell^2}{2}\right)\left[\frac{1+2\eta(1-\eta)[1-2\eta(1-\eta)]}{[1+2\eta(1-\eta)]}\right] \end{aligned} \tag{2.34}$$

For the element adjacent to the element containing the discontinuity we have,

$$\left(\frac{|R(\phi_h)|}{|\nabla\phi_h|}, 1\right)_{\Omega_h^e} = \left(\frac{s\ell^2}{2}\right) \tag{2.35}$$

Thus, the nonlinear term in Eq.(2.27) and for the converged solution given by Eq.(2.19) can be expressed as follows:

$$\left(\frac{|R(\phi_h)|}{|\nabla\phi_h|}, 1\right)_{\Omega_h^e} = \begin{cases} (s\ell^2/2)\ g(\eta)\ , & \text{for elements with shock} \\ (s\ell^2/2)\ , & \text{for elements adjacent to the shock} \\ 0\ , & \text{else} \end{cases} \tag{2.36}$$

where, the function $g(\eta)$ is defined as,

$$g(\eta) := \left[\frac{1+2\eta(1-\eta)[1-2\eta(1-\eta)]}{[1+2\eta(1-\eta)]}\right] \tag{2.37}$$

$$\forall\ \eta \in [0, 1] \quad , \quad g(\eta) \in [(5/6), 1] \tag{2.38}$$

Figure (19) illustrates the plot of the function $g(\eta)$ vs η. To define the parameter β we require that for the elements in the vicinity of the discontinuity the nonlinear Petrov–Galerkin method reproduces the effect of discrete upwinding. The system matrix for the element containing the discontinuity is as follows,

$$\left[s\mathbf{M}^e + \mathbf{S}^e\right] = \left(\frac{s\ell}{6}\right)\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} + \left(\frac{\beta s\ell g(\eta)}{4}\right)\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{2.39}$$

To reproduce the effect of discrete upwinding the following relation should hold,

$$s\mathbf{M}^e + \mathbf{S}^e = \mathrm{DU}(s\mathbf{M}^e) = s\mathbf{M}^e + \tilde{\mathbf{D}}^e \tag{2.40}$$

$$\Rightarrow \left(\frac{\beta s \ell g(\eta)}{4}\right) \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \frac{s\ell}{6} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{2.41}$$

The expression for the parameter β may be expressed as follows:

$$\boxed{\beta g(\eta) \geqslant \frac{2}{3}} \tag{2.42}$$

Remarks:

- The definition of β satisfying the equation $\beta g(\eta) = \frac{2}{3}$ would exactly reproduce the solution of Eq.(2.16) using the lumped mass matrix.
- From the design point-of-view the definition of β involving the function $g(\eta)$ would imply *a priori* knowledge of the solution. Hence we define β using the extremum values of the function $g(\eta)$.

Thus from Eq.(2.38) and Eq.(2.42) we have,

$$\text{Type-I}: \quad \min\{g(\eta)\} = \frac{5}{6} \quad \Rightarrow \quad \boxed{\beta \geqslant \frac{4}{5}} \tag{2.43}$$

$$\text{Type-II}: \quad \max\{g(\eta)\} = 1 \quad \Rightarrow \quad \boxed{\beta \geqslant \frac{2}{3}} \tag{2.44}$$

With these definitions the design of the high-resolution Petrov–Galerkin (HRPG) method for the scaled L^2 projection problem is complete.

2.4.4 *Examples*

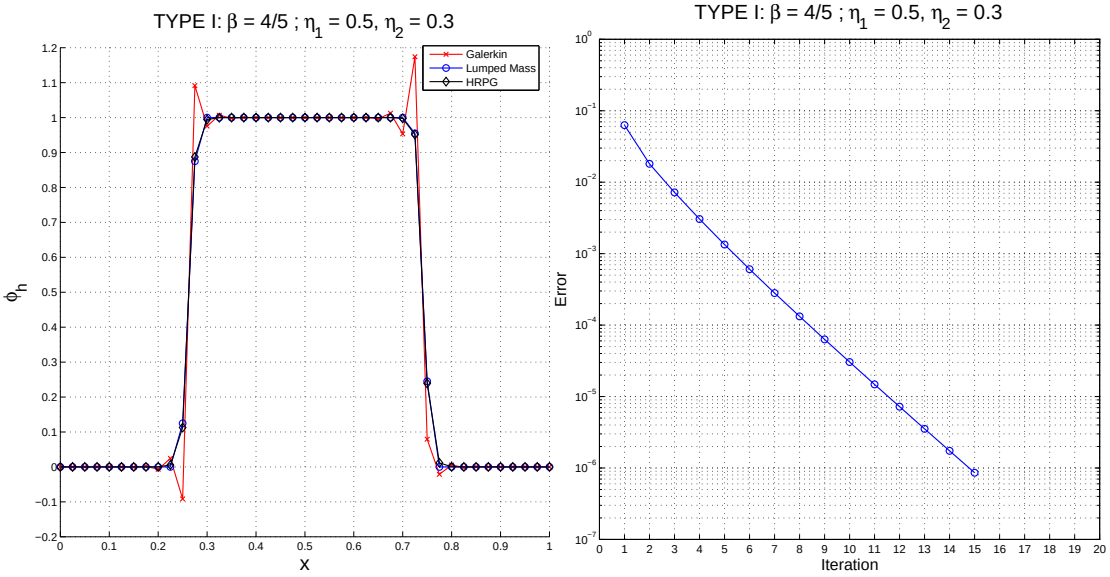
2.4.4.1 *Example 1*

We solve the problem described in Section 2.4.2.2 and using $q = s = 1$. The 1D domain is discretized into 40 linear elements. The numerical results of the HRPG method (Type I and II) are compared with the solutions obtained by the Galerkin method using both consistent and lumped mass matrix. Figs.(20,21) illustrate the results of HRPG Type I and Type II respectively for $\eta_1 = 0.5, \eta_2 = 0.3$. Figs.(22,23) illustrate the same for $\eta_1 = 1, \eta_2 = 0$. Both Type I and Type II effectively circumvent the Gibbs phenomenon. HRPG Type I method is clearly more diffusive, nevertheless monotonicity is guaranteed. HRPG Type II is monotone *to-the-eye*. A quantitative analysis based on the measured total variation is studied in Section 2.4.4.3.

2.4.4.2 *Example 2*

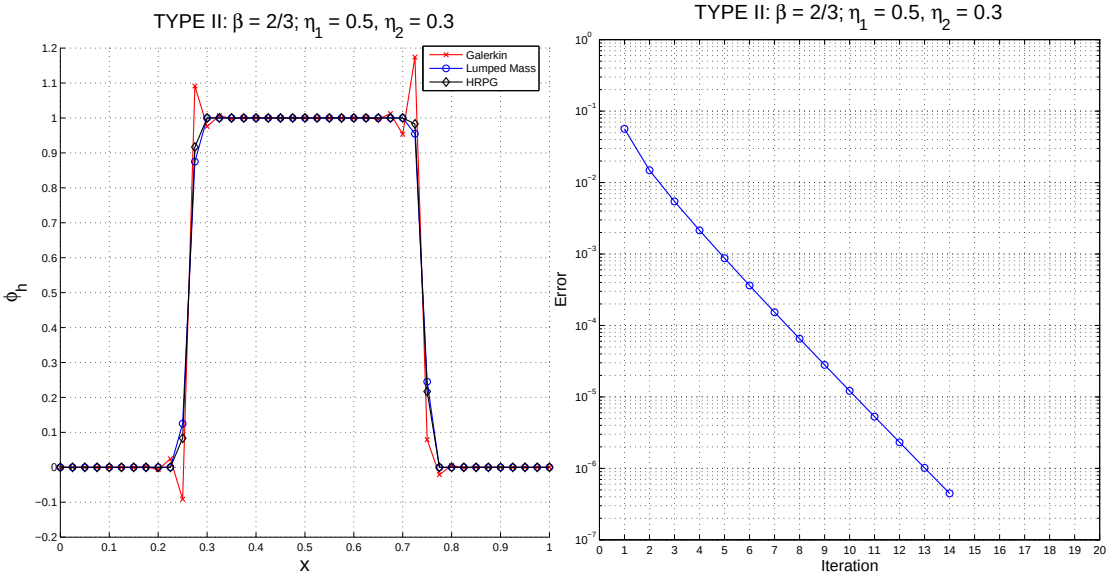
The analysis domain is the same as the problem described in Section 2.4.2.2 and using $s = 1$. The 1D domain is discretized into 100 linear elements. The function whose L^2 projection is sought is now defined as follows:

$$f(x) = \begin{cases} \cos(4\pi x - 2\pi) & \forall \quad x \in [0.25, 0.75] \\ 0 & \text{else} \end{cases} \tag{2.45}$$



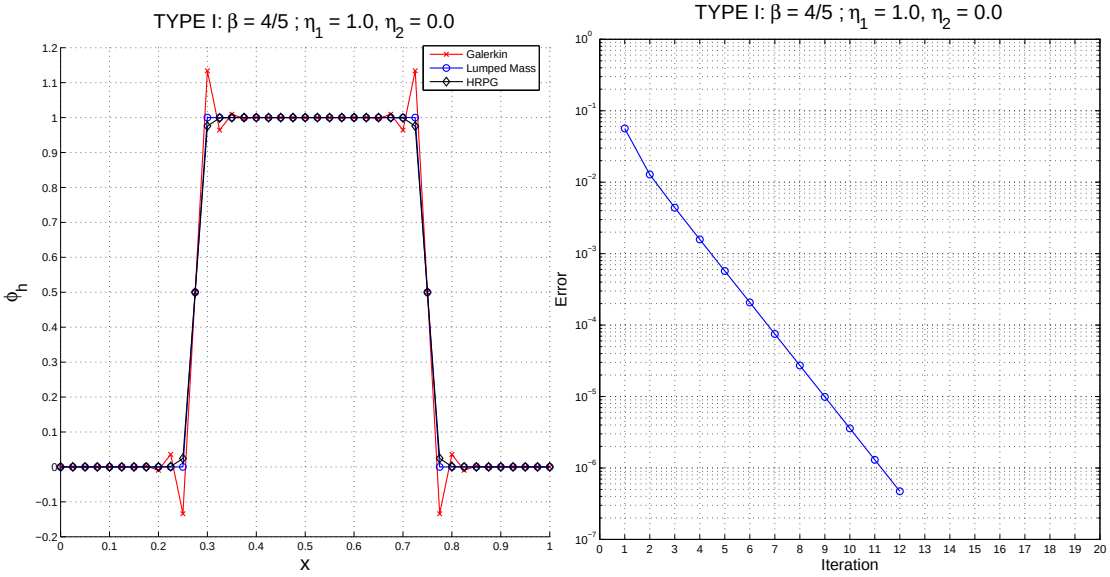
(a) (b)

Figure 20: Example 1: HRPG Type I, (a) numerical solution for $\eta_1 = 0.5, \eta_2 = 0.3$; (b) corresponding nonlinear convergence plot



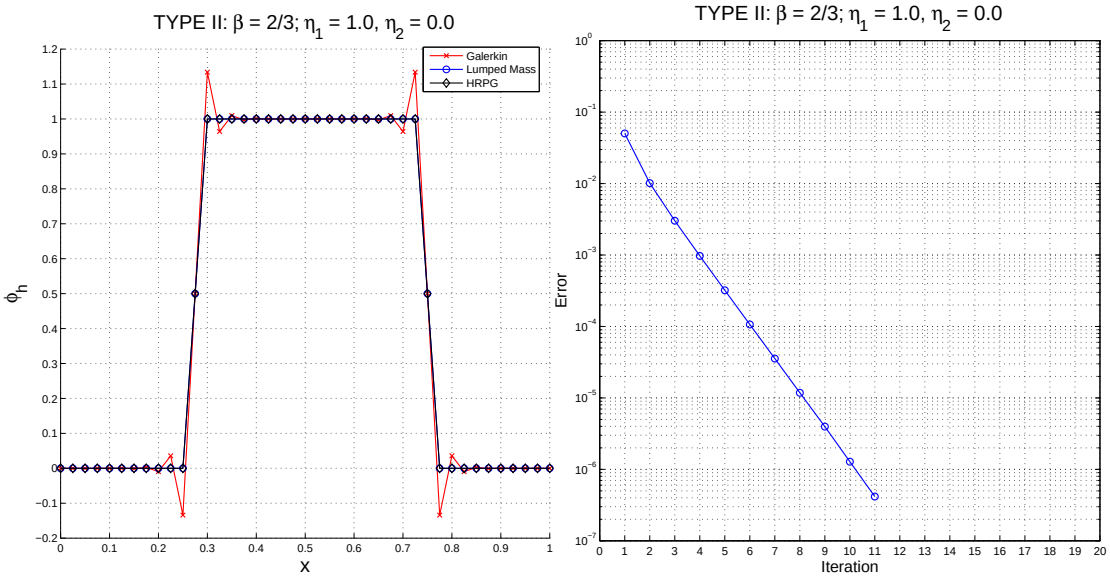
(a) (b)

Figure 21: Example 1: HRPG Type II, (a) numerical solution for $\eta_1 = 0.5, \eta_2 = 0.3$; (b) corresponding nonlinear convergence plot



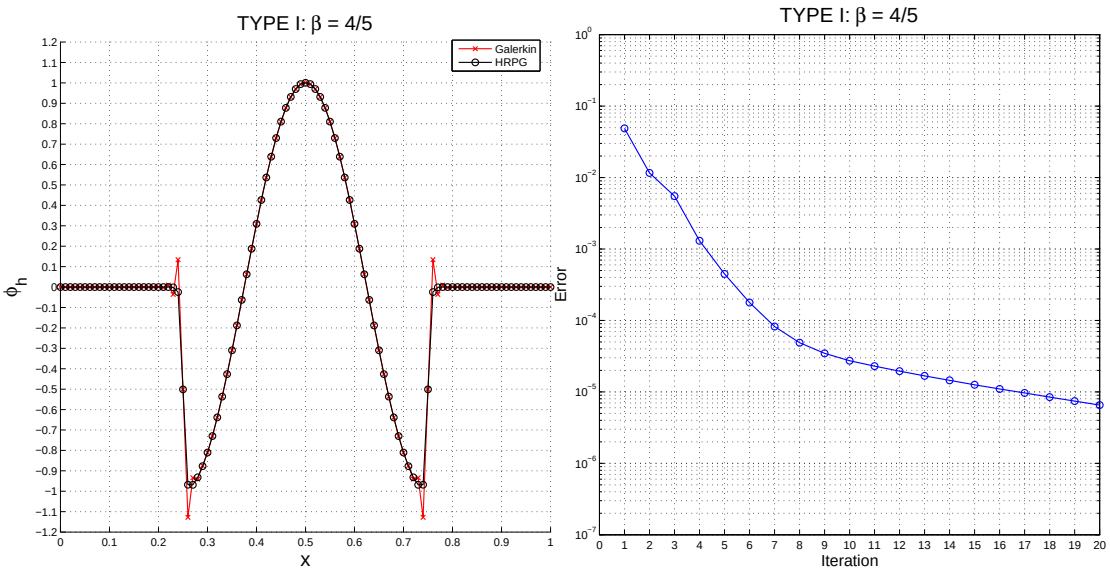
(a) (b)

Figure 22: Example 1: HRPG Type I, (a) numerical solution for $\eta_1 = 1.0, \eta_2 = 0.0$; (b) corresponding nonlinear convergence plot



(a) (b)

Figure 23: Example 1: HRPG Type II, (a) numerical solution for $\eta_1 = 1.0, \eta_2 = 0.0$; (b) corresponding nonlinear convergence plot



(a) (b)

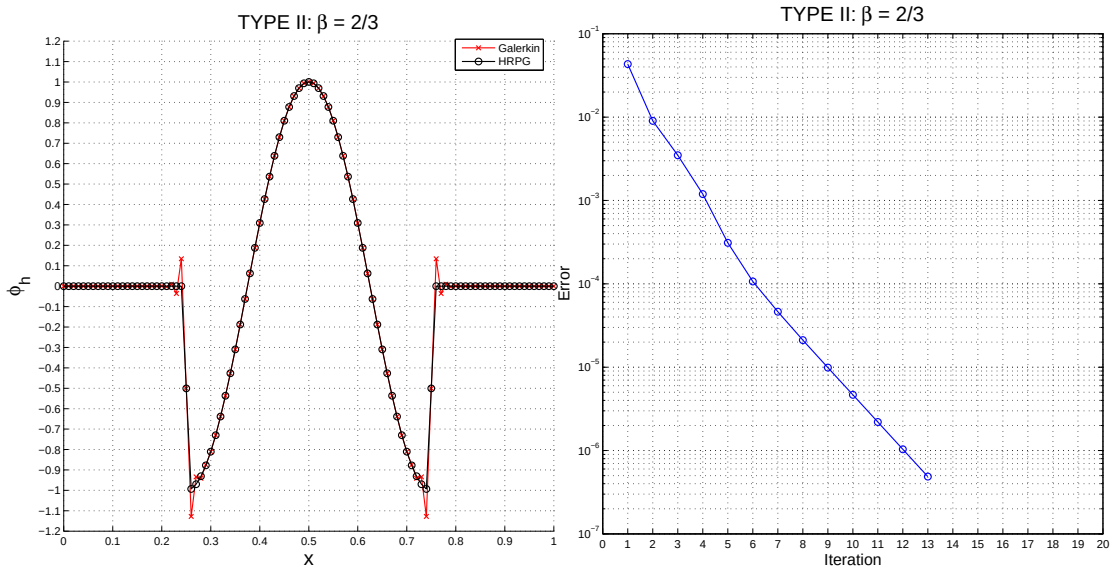
Figure 24: Example 2: HRPG Type I, (a) numerical solution with both smooth and shock regimes ; (b) corresponding nonlinear convergence plot

The above function has both smooth and shock regimes. Simple discontinuities are present at $x = 0.25$ and $x = 0.75$ and in the rest of the domain the function is smooth. This example studies the efficiency of the HRPG method for mixed regimes. Figures (24,25) illustrate that the accuracy of the solution in the smooth regime is not compromised while effectively circumventing the Gibbs phenomenon around the shocks.

2.4.4.3 *Example 3*

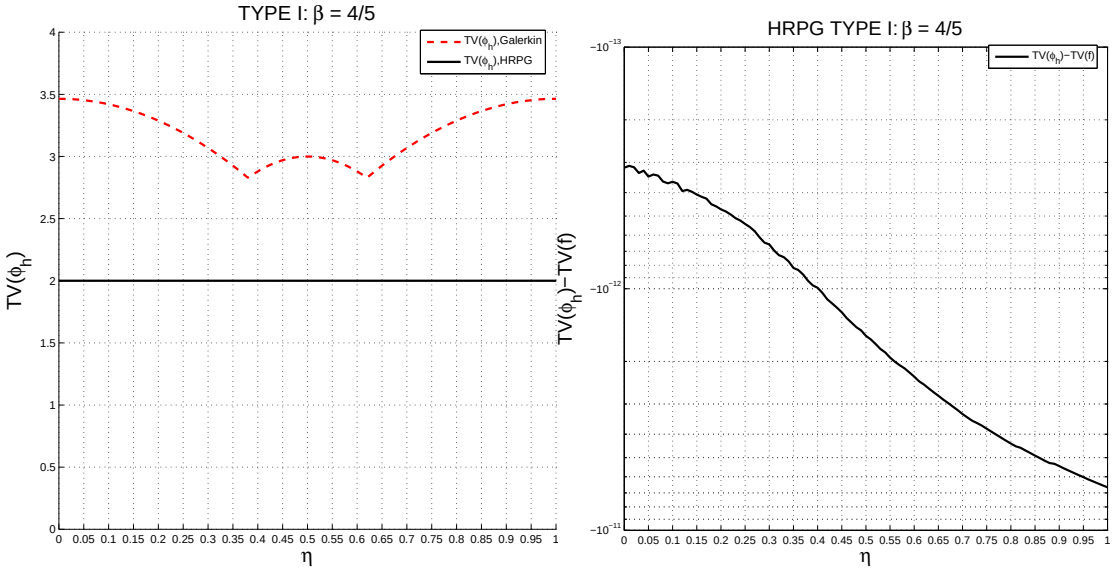
The problem considered in this example is the same as in Section 2.4.4.1. To reduce the variability of the problem data, we have chosen $\eta_1 = \eta_2$. As it is mentioned earlier in Section 2.4.2.4, the total variation (TV) of the numerical solution (ϕ_h) is a direct measure (though *a posteriori*) of the presence of spurious oscillations. HRPG Type I method guarantees that the system matrix for the current problem is an M-matrix for all values of η_1 and η_2. This is a sufficient condition to obtain a monotonicity-preserving solution. The system matrix using the HRPG Type II method is an M-matrix only when $\eta_1, \eta_2 = \{0, 1\}$. Thus we study $TV(\phi_h)$ to have a quantitative measure of performance for the Type I and Type II methods.

Figures (26,27) illustrate with respect to the Galerkin method the $TV(\phi_h)$ vs η plots for the HRPG Type I and Type II methods respectively. It is remarkable that both the methods measure $TV(\phi_h) = 2$ which is the same as $TV(f)$. A study of the error $TV(\phi_h) - TV(f)$ suggests (as expected) that for the Type I method $TV(\phi_h) < TV(f)$ and $TV(\phi_h) - TV(f) = O(1e\text{-}11)$. For the Type II method, $TV(\phi_h) > TV(f)$ and $TV(\phi_h) - TV(f) = O(1e\text{-}5)$ which is an acceptable tolerance. In the light of these results the method we currently prefer is HRPG Type II.



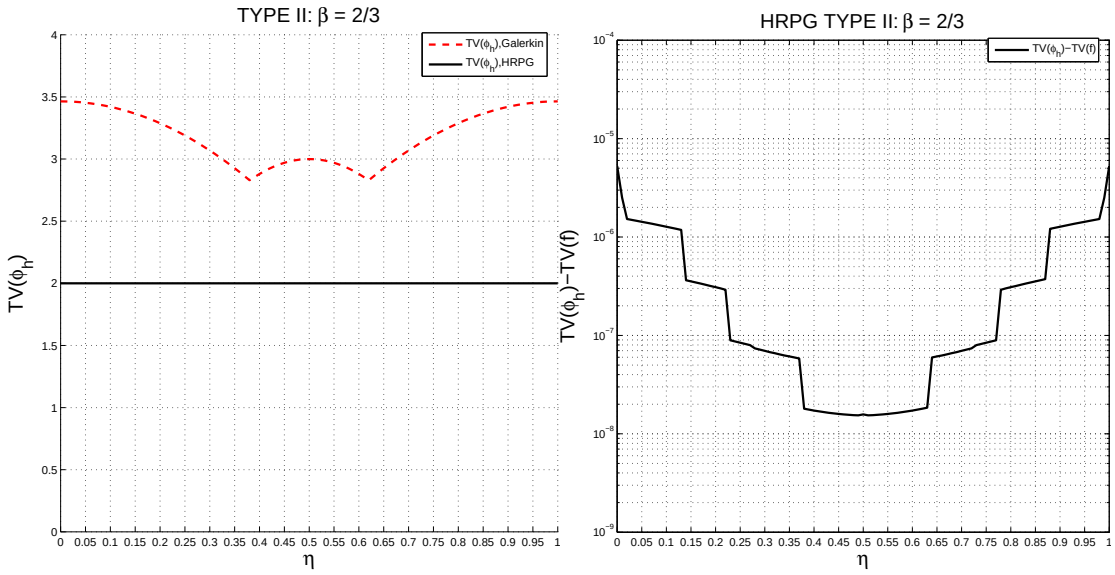
(a) (b)

Figure 25: Example 2: HRPG Type II, (a) numerical solution with both smooth and shock regimes ; (b) corresponding nonlinear convergence plot



(a) (b)

Figure 26: Example 3: HRPG Type I, (a) TV(ϕ_h) plot ; (b) TV(ϕ_h)-TV(f) plot



(a) (b)

Figure 27: Example 3: HRPG Type II, (a) $TV(\phi_h)$ plot ; (b) $TV(\phi_h)$-TV(f) plot

2.4.5 *Summary*

$$
\begin{aligned}
\text{Residual} \quad &: \quad R(\phi_h) := s\phi_h - f \\
\text{HRPG method} \quad &: \quad \Big(w_h, R(\phi_h)\Big)_{\Omega_h} + \sum_e \Big(\frac{\beta\ell}{2}\frac{|R(\phi_h)|}{|\nabla\phi_h|}\frac{dw_h}{dx}, \frac{d\phi_h}{dx}\Big)_{\Omega_h^e} = 0 \\
\text{Type I} \quad &: \quad \boxed{\beta \geqslant \frac{4}{5}} \rightarrow \text{M-matrix guaranteed} \\
\text{Type II} \quad &: \quad \boxed{\beta \geqslant \frac{2}{3}} \rightarrow \text{Total variation limit}
\end{aligned}
$$

The Gibbs phenomenon that arises in L^2 projections is studied for the Galerkin method in 1D using linear finite elements. A nonlinear Petrov–Galerkin method (HRPG) is formulated and the stabilization parameter is designed (Type I and Type II) so as to circumvent the Gibbs phenomenon and thus leading to a high-resolution method. The HRPG method is shown to perform well in the presence of both smooth and shock regimes in the solution. HRPG Type II method is shown to be in the total variation limit with acceptable tolerance of O(1e-5) and thus is essentially non-oscillatory (monotone *to-the-eye*). Hence it is currently the preferred choice for the extension of the HRPG design to the model problems in the subsequent sections.

2.5 CONVECTION–DIFFUSION–REACTION PROBLEM

2.5.1 *Galerkin method and discrete upwinding*

Consider the convection–diffusion–reaction problem given by Eq.(2.1) in 1D and subjected only to the Dirichlet boundary conditions. Discretization of the space by linear

finite elements will lead to the approximation $\phi_h = N^a \Phi^a$. For the Galerkin method we arrive at the following system of equations.

$$\mathbf{M}\dot{\Phi} + [u\mathbf{C} + k\mathbf{D} + s\mathbf{M}]\Phi = \mathbf{f} \tag{2.46}$$

where the element contributions to the above matrices and vector are given by,

$$\mathbf{M}^e_{ab} = \left(N^a, N^b\right)_{\Omega_h} = \frac{\ell}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \mathbf{C}^e_{ab} = \left(N^a, \nabla(N^b)\right)_{\Omega_h} = \frac{1}{2}\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} \tag{2.47}$$

$$\mathbf{f}^e_a = \left(N^a, f\right)_{\Omega_h} = \frac{\ell}{2}\begin{Bmatrix} 1 \\ 1 \end{Bmatrix}, \mathbf{D}^e_{ab} = \left(\nabla(N^a), \nabla(N^b)\right)_{\Omega_h} = \frac{1}{l}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{2.48}$$

The discrete upwinding process applied to the steady state of Eq.(2.46) will introduce a numerical diffusion k^{du} as follows:

$$\mathrm{DU}(u\mathbf{C} + k\mathbf{D} + s\mathbf{R}) = [u\mathbf{C} + k\mathbf{D} + s\mathbf{R}] + k^{du}\mathbf{D} \tag{2.49}$$

$$k^{du} = \max\left\{\left[\frac{|u|\ell}{2} + \frac{s\ell^2}{6} - k\right], 0\right\} = k\ \max\left\{\left[|\gamma| + \frac{\omega}{6} - 1\right], 0\right\} \tag{2.50}$$

The form of this numerical diffusion (Eq.(2.50)) is identical to that found in [152] using a FIC-based approach. The stabilization method presented in [152] introduces within each element an additional nonlinear diffusion as follows:

$$k^{fic} = k\ \max\left\{\left[\left(\frac{\mathrm{sgn}[\nabla\phi_h]}{\mathrm{sgn}[\Delta\phi_h]}\right)\gamma + \left(\frac{\mathrm{sgn}[\phi_h]}{\mathrm{sgn}[\Delta\phi_h]}\right)\frac{\omega}{6} - 1\right], 0\right\} \tag{2.51}$$

Clearly the form of Eq.(2.50) is an upper bound of the value of k^{fic} as defined in Eq.(2.51).

2.5.2 *Model problem 2*

Consider the steady diffusion–reaction problem with a distributed source term given by Eq.(2.18) and homogeneous Dirichlet boundary conditions:

$$R(\phi) := -k\Delta(\phi) + s\phi - f(x) \tag{2.52}$$

For the current problem we design the HRPG method with $\alpha = 0$. In the limit as $k \to 0$ the problem reduces to the scaled L^2 projection problem considered in Section 2.4. The discrete upwinding operation on the Galerkin method, will introduce an artificial diffusion equivalent to the following,

$$k^{du} = \max\left\{\left[\frac{s\ell^2}{6} - k\right], 0\right\} \tag{2.53}$$

Note that $\forall k \leqslant (s\ell^2/6)$ the critical non-oscillatory solution obtained via the discrete upwinding process is identical to the solution obtained with $k = 0$. Thus in order to design the parameter β we can use the solution given by Eq.(2.19) to estimate the

amount of nonlinear diffusion that would be introduced by the HRPG method. Thus,

$$k^*(\phi_h) := \frac{\beta}{2}\left(\frac{|R(\phi_h)|}{|\nabla\phi_h|}, 1\right)_{\Omega_h^e} = \beta\frac{s\ell^2}{4}g(\eta) \geqslant \max\left\{\left[\frac{s\ell^2}{6} - k\right], 0\right\} \tag{2.54}$$

We define a dimensionless element number $\omega := (s\ell^2/k)$ and consider $g(\eta) = 1$ (Type-II). Thus,

$$\beta \geqslant \max\left\{\frac{2}{3}\left[1 - \frac{6}{\omega}\right], 0\right\} \tag{2.55}$$

We remark that β depends only on the problem data and for the current model problem the nonlinear (residual-based) diffusion $k^*(\phi_h)$ implemented in the HRPG method is nonzero and equals k^{du} only for the elements in the vicinity of the discontinuity. In this way it differs from the form of Eq.(2.51) which for the current model problem introduces a nonlinear diffusion k^{fic} (Eq.(2.51) using $\gamma = 0$) for all elements.

2.5.3 *Model problem 3*

Consider the steady convection-diffusion problem,

$$R(\phi) := u\nabla(\phi) - k\Delta(\phi) \tag{2.56}$$

For the current problem we design the HRPG method with $\alpha = 0$. The discrete upwinding operation on the Galerkin method, will introduce an artificial diffusion equivalent to the following,

$$k^{du} = \max\left\{\left[\frac{|u|\ell}{2} - k\right], 0\right\} \tag{2.57}$$

The parameter β may be designed as follows:

$$\frac{|R(\phi_h)|}{|\nabla(\phi_h)|} = \frac{|u\nabla(\phi_h)|}{|\nabla(\phi_h)|} = |u| \tag{2.58}$$

$$\Rightarrow k^*(\phi_h) := \frac{\beta}{2}\left(\frac{|R(\phi_h)|}{|\nabla(\phi_h)|}, 1\right)_{\Omega_h^e} = \frac{\beta}{2}|u|\ell \geqslant \max\left\{\left[\frac{|u|\ell}{2} - k\right], 0\right\} \tag{2.59}$$

$$\Rightarrow \beta \geqslant \max\left\{\left[1 - \frac{1}{|\gamma|}\right], 0\right\} \tag{2.60}$$

Note that in contrast to the problem considered in Section 2.5.2 here we do not need the solution to estimate the nonlinear diffusion introduced by the HRPG method. The expression for β (Eq.2.60) is identical to the standard critical stabilization parameter obtained for upwind techniques (see [195]).

2.5.4 *Model problem 4*

Consider the steady convection–diffusion–reaction problem:

$$R(\phi) := u\nabla(\phi) - k\Delta(\phi) + s\phi - f(x) \tag{2.61}$$

For the current problem we design the HRPG method again with $\alpha = 0$. If linear finite elements were used, the residual obeys the following relation:

$$|R(\phi_h)| = |u\nabla(\phi_h) + s\phi_h - f| \leqslant |u\nabla(\phi_h)| + |s\phi_h - f| \tag{2.62}$$

$$\Rightarrow \frac{|R(\phi_h)|}{|\nabla(\phi_h)|} = \left|u + \frac{s\phi_h - f}{\nabla(\phi_h)}\right| \leqslant |u| + \frac{|s\phi_h - f|}{|\nabla(\phi_h)|} \tag{2.63}$$

As it can be seen in Eq.(2.63), to estimate the nonlinear diffusion introduced by the HRPG method we require the solution of the problem a priori. The simplest idea would be to use the nodally exact solution. Unlike the solution used in Section 2.5.2, the analytical solution of the current problem has a complex structure [152]. In order to retain simplicity in the design of β we make a conjuncture of the results obtained in Section 2.5.2 and Section 2.5.3. The conjuncture is made such that the designed expression for β would approach asymptotically the expressions obtained in Section 2.5.2 and Section 2.5.3 as $u \to 0$ and $s \to 0$ respectively.

Assume that $u \ll s$ and $f(x)$ be defined as in Eq.(2.18). Thus we may approximate the solution of the current problem to the one considered in Section 2.5.2. This assumption allows us to use the solution defined by Eq.(2.19) to approximately estimate the following expression:

$$\left(\frac{|s\phi_h - f|}{|\nabla(\phi_h)|}, 1\right)_{\Omega_h^e} \approx \frac{s\ell^2}{2} g(\eta) \tag{2.64}$$

As $u \ll s$ we make another approximation using Eq.(2.63) as follows:

$$\frac{|R(\phi_h)|}{|\nabla(\phi_h)|} \approx |u| + \frac{|s\phi_h - f|}{|\nabla(\phi_h)|} \tag{2.65}$$

Using the two approximations (Eq.2.64, Eq.2.65) the parameter β may be designed as follows:

$$k^*(\phi_h) := \frac{\beta}{2}\left(\frac{|R(\phi_h)|}{|\nabla(\phi_h)|}, 1\right)_{\Omega_h^e} \approx \frac{\beta}{2}\left[|u|\ell + \frac{s\ell^2}{2} g(\eta)\right] \geqslant \max\left\{\left[\frac{|u|\ell}{2} + \frac{s\ell^2}{6} - k\right], 0\right\} \tag{2.66}$$

$$\beta := \max\left\{\left[\frac{2}{3}\left(\frac{|\sigma|+3}{|\sigma|+2}\right) - \left(\frac{4}{\omega + 4|\gamma|}\right)\right], 0\right\} \Rightarrow \begin{cases} \lim\limits_{u\to 0} \beta = \max\left\{\frac{2}{3}\left[1 - \frac{6}{\omega}\right], 0\right\} \\ \lim\limits_{s\to 0} \beta = \max\left\{\left[1 - \frac{1}{|\gamma|}\right], 0\right\} \end{cases} \tag{2.67}$$

where, $\gamma := (u\ell/2k)$, $\omega := (s\ell^2/k)$ and $\sigma := (\omega/2\gamma) = (s\ell/u)$ are the element Peclect number, a velocity independent dimensionless number and the Damköler number respectively.

Remark : Eq.(2.66) does not mean that $k^*(\phi_h) = k^{du}$ for all elements and problem data. We remind that under the assumption $u \ll s$ and $f(x)$ defined as in Eq.(2.18), the solution to the current model problem is approximated as the one given by Eq.(2.19). This suggests that, similar to the model problems in Section 2.4.2.2 and Section 2.5.2, the nonlinear (residual-based) diffusion $k^*(\phi_h)$ equals k^{du} only for the elements in the vicinity of the layers. In general, as β is independent of ϕ_h, the only information known a priori is that $k^*(\phi_h)$ is proportional to the residual $R(\phi_h)$. The argument that this expression for β i. e.Eq.(2.67), would perform well $\forall u, s$ is a *conjecture* based on

the fact that we recover asymptotically the expressions for β i. e.Eq.(2.55) and Eq.(2.60) as $u \to 0$ and $s \to 0$ respectively. The efficiency of this expression for β is shown via various numerical examples (see Section 2.5.7).

2.5.5 *Model problem 5*

Consider the transient convection–diffusion–reaction problem:

$$R(\phi) := \dot{\phi} + u\nabla(\phi) - k\Delta(\phi) + s\phi - f(x) \tag{2.68}$$

We now design the parameter β associated with the nonlinear perturbation term when the linear perturbation terms exists (i. e.$\alpha \neq 0$). The HRPG method after discretization by linear finite elements will lead to the following system of equations.

$$\begin{aligned} &[\mathbf{M}]\dot{\Phi} + [u\mathbf{C} + k\mathbf{D} + s\mathbf{M}]\Phi - \mathbf{f_g} && \text{(Galerkin terms)} \\ &+\frac{\alpha\ell}{2}\left([\mathbf{C^t}]\dot{\Phi} + [u\mathbf{D} + s\mathbf{C^t}]\Phi - \mathbf{f_s}\right) && \text{(linear PG terms)} \\ &\quad +\frac{\beta}{2}\left(\frac{|R(\phi_h)|}{|\nabla(\phi)|}, 1\right)_{\Omega_h^e}[\mathbf{D}]\Phi = 0 && \text{(nonlinear PG terms)} \end{aligned} \tag{2.69}$$

The above equation may be rearranged as follows

$$\mathbf{M}\{\dot{\Phi} + s\Phi\} + \mathbf{C}\{u\Phi\} + \mathbf{C^t}\{\frac{\alpha\ell}{2}\dot{\Phi} + \frac{\alpha\ell s}{2}\Phi\} + (k + \frac{\alpha\ell u}{2} + k^*(\phi_h))\mathbf{D}\Phi = \mathbf{f_g} + \frac{\alpha\ell}{2}\mathbf{f_s} \tag{2.70}$$

where the expression for $k^*(\phi_h)$ is given by Eq.(2.32). We define for each element a measure δ with the dimensions of the reaction coefficient.

$$\delta := \dot{\Phi} \oslash \Phi \quad \Rightarrow \dot{\Phi} = \delta \odot \Phi \tag{2.71}$$

where the vector operators $\oslash$ and $\odot$ are understood to operate point-to-point division and multiplication respectively. Thus in the design of the parameter β, we may model δ as a non-linear reactive coefficient. For the fully discrete problem (after time discretization)δ may be approximated to an element-wise positive constant for simplification. This idea had been pointed out earlier in [99]. Also note that the convection matrix $\mathbf{C}$ is skew-symmetric. Hence the transposed matrix $\mathbf{C^t}$ introduces a negative convection effect [30]. Following this line, for each element we define the effective convection, diffusion and reaction coefficients as follows:

$$\tilde{u} := u - \frac{\alpha\ell s}{2} - \frac{\alpha\ell\delta}{2} \quad ; \quad \tilde{k} := k + \frac{\alpha\ell u}{2} \quad ; \quad \tilde{s} := s + \delta \tag{2.72}$$

The effective coefficients $\tilde{u}, \tilde{k}$ and $\tilde{s}$ will be used to design the parameter β. The following effective element dimensionless numbers may be defined:

$$\tilde{\gamma} := \frac{\tilde{u}\ell}{2\tilde{k}} \quad ; \quad \tilde{\omega} := \frac{\tilde{s}\ell^2}{\tilde{k}} \quad ; \quad \tilde{\sigma} := \frac{\tilde{s}\ell}{\tilde{u}} = \frac{\tilde{\omega}}{2\tilde{\gamma}} \tag{2.73}$$

The parameter β is now defined as in Eq.(2.67) using these effective element numbers as follows:

$$\beta := \max\left\{\left[\frac{2}{3}\left(\frac{|\tilde{\sigma}| + 3}{|\tilde{\sigma}| + 2}\right) - \left(\frac{4}{\tilde{\omega} + 4|\tilde{\gamma}|}\right)\right], 0\right\} \tag{2.74}$$

If the discretization in time is done using the implicit trapezoidal rule, we have

$$\dot{\Phi} = \frac{\tilde{\Phi} - \Phi^n}{\theta \Delta t} \quad ; \quad \Phi^{n+1} = \frac{1}{\theta}\tilde{\Phi} + \frac{\theta - 1}{\theta}\Phi^n \tag{2.75}$$

where, Δt is the time increment, $n, n+1$ denote the previous and current time steps and $\theta \in [0, 1]$ is a parameter that defines the scheme. $\theta = \{0, 0.5, 1\}$ define the forward Euler, implicit midpoint and the backward Euler methods. For the fully discrete system and within each element we evaluate the parameter δ and the residual as follows:

$$\delta \approx \frac{1}{\theta \Delta t} \frac{\| \tilde{\phi}_h - \phi_h^n \|_\infty^e}{\| \tilde{\phi}_h \|_\infty^e} \tag{2.76}$$

$$R(\tilde{\phi}_h) \approx \frac{\tilde{\phi}_h - \phi_h^n}{\theta \Delta t} + u \nabla(\tilde{\phi}_h) - k \Delta(\tilde{\phi}_h) + s\tilde{\phi}_h - f \tag{2.77}$$

where $\| \cdot \|_\infty^e$ is the L^∞norm. Note that as steady state is reached $\delta \to 0$. Thus for the steady state problem and using $\alpha = 0$ we recover the definition of the parameter β as given by Eq.(2.67).

It remains to define the parameter α that controls the fraction of linear perturbation term in the HRPG method. For the 1D convection–diffusion–reaction problem, the HRPG method using $\alpha = 0$ does solve a plethora of examples to give high-resolution stabilized results. Nevertheless for the transient problem the presence of the linear perturbation terms improves the convergence of the nonlinear iterations. Numerical experiments suggest $\alpha \in [0, 1/3]$ which means that the approximations/conjecture used in the design strategy does not hold for larger fractions of the linear perturbation term. The following expression for α was used in the examples to come.

$$\alpha := \lambda \ \mathrm{sgn}(u) \max\left\{ \left[1 - \frac{1}{|\gamma|} \right], 0 \right\} \quad ; \quad \lambda := \frac{1}{3(1 + \sqrt{|\sigma|})} \tag{2.78}$$

2.5.6 *Summary*

RESIDUAL

$$R(\phi_h) := \frac{\partial \phi_h}{\partial t} + u\nabla(\phi_h) - k\Delta(\phi_h) + s\phi_h - f(x)$$

THE HRPG METHOD

Find $\phi_h : [0, T] \mapsto V^h$ such that $\forall w_h \in V_0^h$ we have,

$$a(w_h, \phi_h) + \sum_e \left(\frac{\alpha\ell}{2}\frac{dw_h}{dx}, R(\phi_h)\right)_{\Omega_h^e} + \left(\frac{\beta\ell}{2}\frac{|R(\phi_h)|}{|\nabla\phi_h|}\frac{dw_h}{dx}, \frac{d\phi_h}{dx}\right)_{\Omega_h^e} = l(w_h)$$

$$\text{PG weight} \rightarrow w_h + \left[\frac{\alpha\ell}{2} + \frac{\beta\ell}{2}\,\text{sgn}[R(\phi_h)]\,\text{sgn}[\nabla(\phi_h)]\right]\frac{dw_h}{dx}$$

DEFINITIONS

$$\gamma := \frac{u\ell}{2k} \quad ; \quad \omega := \frac{s\ell^2}{k} \quad ; \quad \sigma := \frac{s\ell}{u}$$

$$R(\tilde{\phi}_h) \approx \frac{\tilde{\phi}_h - \phi_h^n}{\theta\Delta t} + u\nabla(\tilde{\phi}_h) - k\Delta(\tilde{\phi}_h) + s\tilde{\phi}_h - f$$

$$\phi_n^{n+1} = \left(\frac{1}{\theta}\right)\tilde{\phi}_h + \left(\frac{\theta - 1}{\theta}\right)\phi_h^n \quad ; \quad \Delta t = t^{n+1} - t^n \quad ; \quad \theta \in (0, 1)$$

$$\lambda := \frac{1}{3(1 + \sqrt{|\sigma|})} \quad ; \quad \delta \approx \frac{1}{\theta\Delta t}\frac{\| \tilde{\phi}_h - \phi_h^n \|_\infty}{\| \tilde{\phi}_h \|_\infty}$$

$$\boxed{\alpha := \lambda\ \text{sgn}(u)\max\left\{\left[1 - \frac{1}{|\gamma|}\right], 0\right\}}$$

$$\tilde{u} := u - \frac{\alpha\ell s}{2} - \frac{\alpha\ell\delta}{2} \quad ; \quad \tilde{k} := k + \frac{\alpha\ell u}{2} \quad ; \quad \tilde{s} := s + |\delta|$$

$$\tilde{\gamma} := \frac{\tilde{u}\ell}{2\tilde{k}} \quad ; \quad \tilde{\omega} := \frac{\tilde{s}\ell^2}{\tilde{k}} \quad ; \quad \tilde{\sigma} := \frac{\tilde{s}\ell}{\tilde{u}}$$

$$\boxed{\beta := \max\left\{\left[\frac{2}{3}\left(\frac{|\tilde{\sigma}| + 3}{|\tilde{\sigma}| + 2}\right) - \left(\frac{4}{\tilde{\omega} + 4|\tilde{\gamma}|}\right)\right], 0\right\}}$$

2.5.7 *Examples*

2.5.7.1 *Example 1*

We consider the convection–diffusion–reaction problem given by Eq.(2.1) in 1D. We study the steady-state case with the following data : $k = 1$ and $u, s \neq 0$. The 1D domain is taken as $x \in [0, 1]$ and it is discretized with eight two-node linear elements. The values of u and s are determined appropriately for different values of γ and ω. The results of the HRPG method (using both $\lambda = 0$ and $\lambda \neq 0$) are compared with that of the Galerkin, Galerkin with discrete upwinding (DU), SUPG, CAU, modified CAU and the FIC based stabilization method presented in [152]. The error in the nonlinear iterations was measured by the following norm:

$$\frac{\| \Phi^{i+1} - \Phi^i \|_e}{\| \Phi^{i+1} \|_e} \tag{2.79}$$

where, $\| \cdot \|_e$ is the standard Euclidean vector norm. A tolerance of `1e-5` was chosen as the termination criteria. A maximum of 30 iterations were allowed. Note that the number of iterations required by the nonlinear methods for convergence is displayed

next to the corresponding legends. The nonlinear iterations were initialized by the solution obtained by the DU method.

Figures 28 - 33 illustrate the solution obtained for the sourceless case ($f = 0$) and for $(\gamma, \omega) = \{(1,5), (1,20), (1,120), (2,2), (10,4), (10,20)\}$ respectively. The Dirichlet boundary conditions $\phi_L^p := \phi(x = 0) = 8$ and $\phi_R^p := \phi(x = 1) = 3$ were employed. The DU method is robust and provides stable solutions. Unfortunately the accuracy achieved is at most first-order and hence the solutions are generally over-diffusive. The FIC method presented in [152] provides more accurate solutions and remarkably the nonlinear iterations converge with just two iterations. Slight node-to-node oscillations around the exact solutions are observed for the case $\gamma = 10$, $\omega = 4$ viz. Figure 32a and is duly discussed in [152]. As expected the SUPG method provides good solutions to all except the reaction-dominated cases viz. Figures 29b,30b. The CAU and modified CAU methods succeed in circumventing the instabilities for the reaction-dominated cases but for these cases provide solutions that are more diffusive than that of the DU method. The HRPG method provides good solutions for all the cases considered. Note that for the reaction-dominated cases the solutions are less diffusive than the DU, CAU and modified CAU methods. Also note that the solutions obtained by taking $\lambda = 0$ is indistinguishable to that obtained by taking $\lambda \neq 0$. Nevertheless the nonlinear iterations converge faster for the latter.

Figures 34,35 illustrate the solution obtained for the sourceless case ($f = 0$) with $(\gamma, \omega) = (10, 200)$ and for Dirichlet boundary conditions $(\phi_L^p, \phi_R^p) = \{(0,1), (1,0)\}$ respectively. The FIC method of [152] provides nodally exact *to-the-eye* solutions and the nonlinear iterations converge within 2 iterations. The solutions obtained by the SUPG and CAU methods are indistinguishable and exhibit instabilities for the latter boundary conditions viz. Figure 35b. The modified CAU method circumvents these instabilities but instead provides solutions that are more diffusive than the DU method. The HRPG method (both $\lambda = 0$ and otherwise) succeed to provide stable solutions and are less diffusive than that obtained by the DU method. Figure 34 shows that the HRPG method with $\lambda = 0$ converges in just one iteration while using $\lambda \neq 0$ seven iterations were needed. This is a rare coincidence where the initial solution provided by the DU method and the solution of the HRPG method with $\lambda = 0$ are closer than the specified tolerance.

Figures 36,37 illustrate the solution obtained with $(\gamma, \omega) = (2, 0)$ and $(f, \phi_L^p, \phi_R^p) = \{(0,0,1), (u,0,0)\}$ respectively. As expected the SUPG method provides nodally exact solutions for these cases. The DU, CAU and HRPG methods provide stable solutions and are indistinguishable from each other. The modified CAU solution is very similar to the solutions of the former methods. For the sourceless case ($f = 0$) the solutions of the DU and FIC methods are indistinguishable. Unfortunately when $f \neq 0$ the nonlinear iterations associated with the latter fail to converge. We believe that this behavior is due to the increased nonlinearity associated with the definition of the stabilization parameters (See [152]).

2.5.7.2 *Example 2*

We consider again the convection–diffusion–reaction problem given by Eq.(2.1) in 1D. Now we study the transient pure-convection problem, i.e. $k, s, f = 0$. The Dirichlet

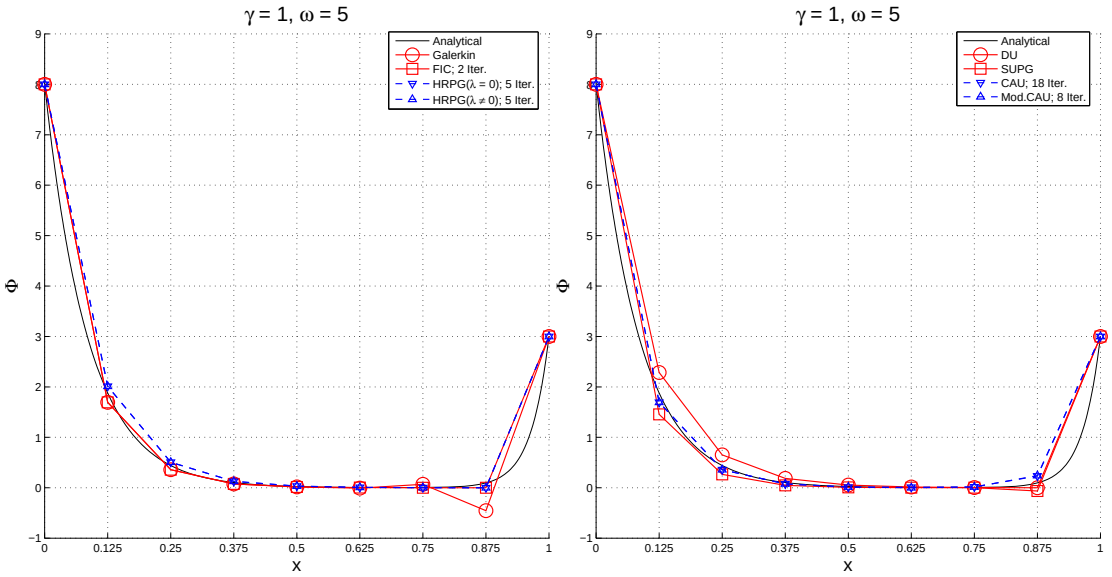


Figure 28: Steady state: $(\gamma, \omega, f, \phi_L^p, \phi_R^p) = (1, 5, 0, 8, 3)$. (a) Exact, Galerkin, FIC [152], HRPG ($\lambda = 0$) and HRPG ($\lambda \neq 0$) solutions; (b) Exact, DU, SUPG, CAU and Mod.CAU solutions

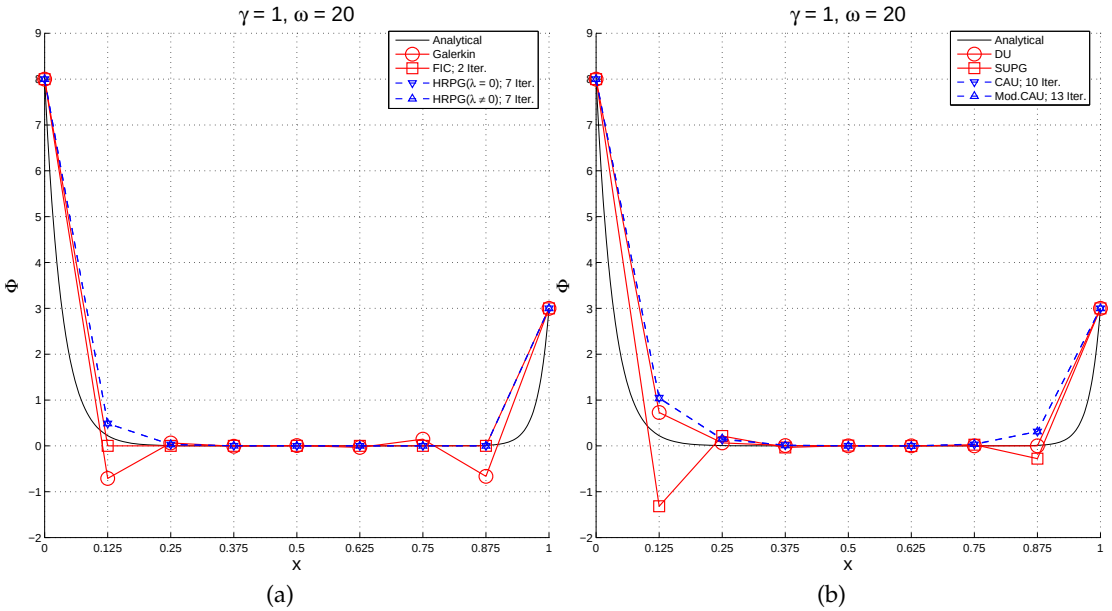


Figure 29: Steady state: $(\gamma, \omega, f, \phi_L^p, \phi_R^p) = (1, 20, 0, 8, 3)$. (a) Exact, Galerkin, FIC [152], HRPG ($\lambda = 0$) and HRPG ($\lambda \neq 0$) solutions; (b) Exact, DU, SUPG, CAU and Mod.CAU solutions

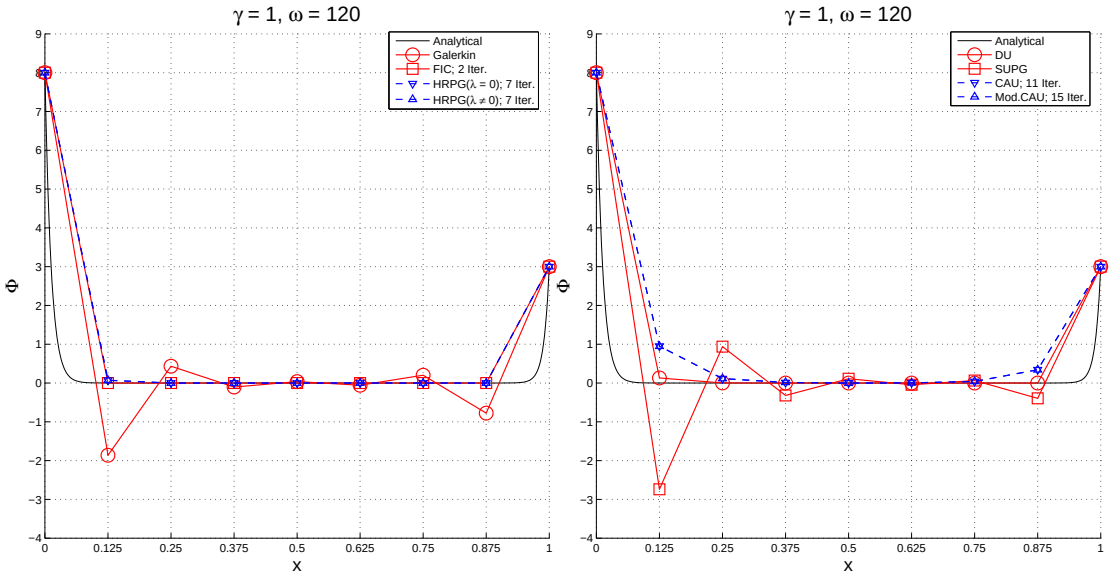


Figure 30: Steady state: $(\gamma, \omega, f, \phi_L^p, \phi_R^p) = (1, 120, 0, 8, 3)$. (a) Exact, Galerkin, FIC [152], HRPG ($\lambda = 0$) and HRPG ($\lambda \neq 0$) solutions; (b) Exact, DU, SUPG, CAU and Mod.CAU solutions

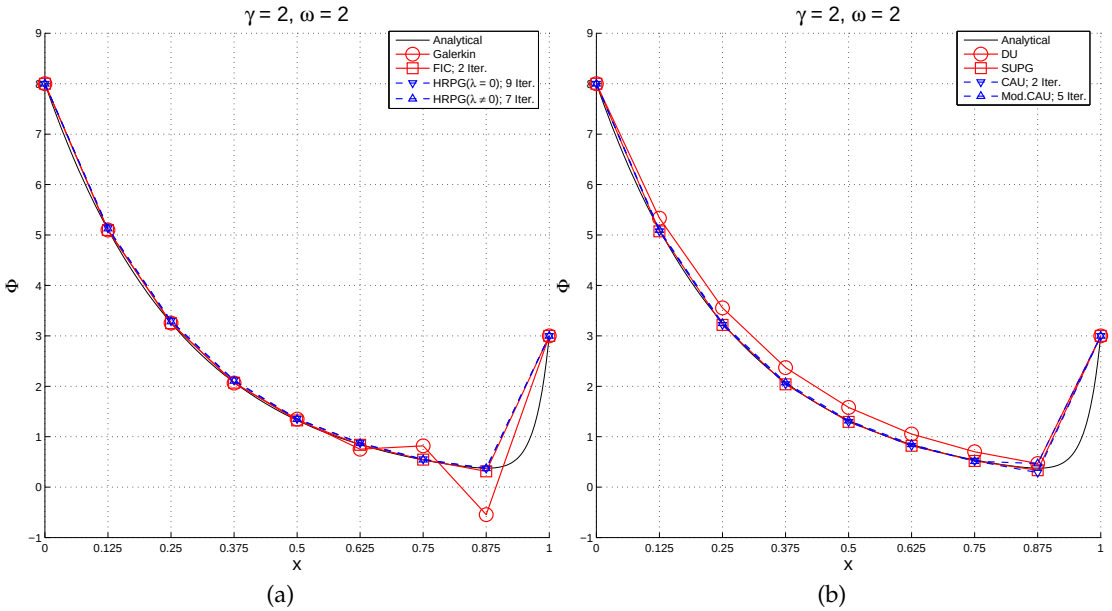
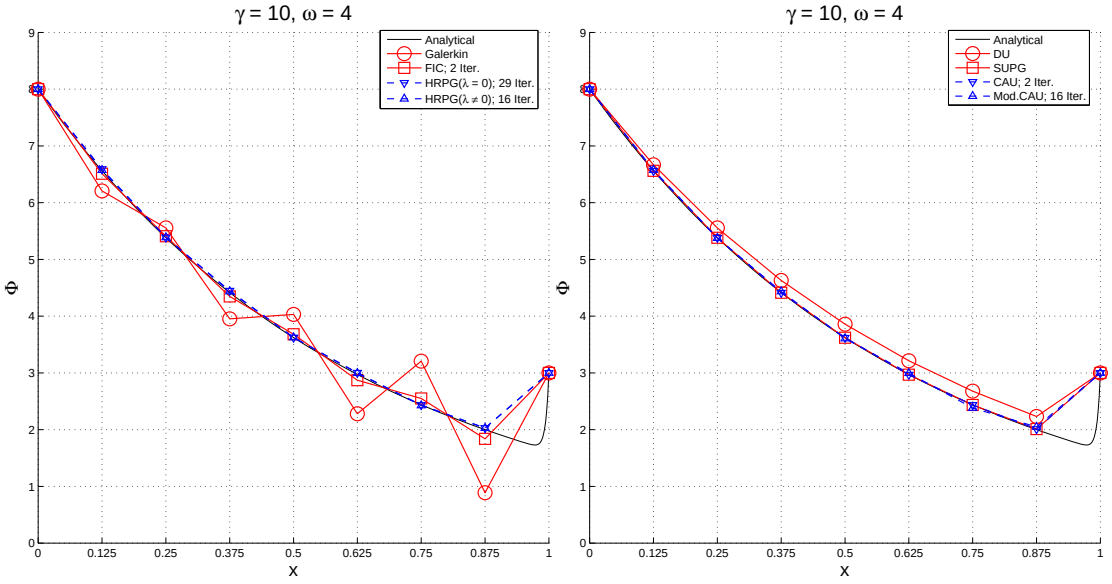
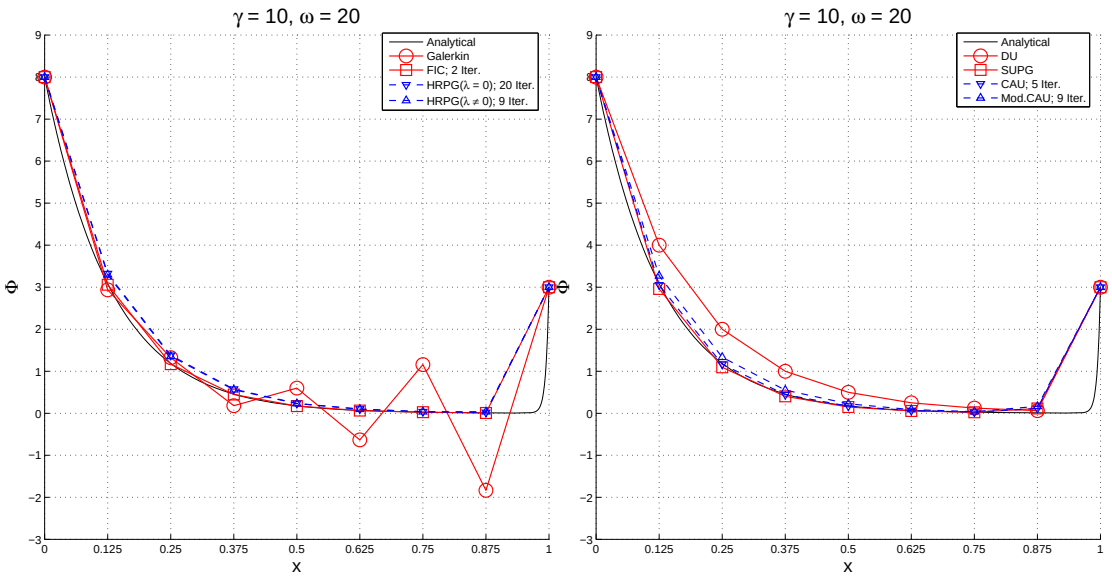


Figure 31: Steady state: $(\gamma, \omega, f, \phi_L^p, \phi_R^p) = (2, 2, 0, 8, 3)$. (a) Exact, Galerkin, FIC [152], HRPG ($\lambda = 0$) and HRPG ($\lambda \neq 0$) solutions; (b) Exact, DU, SUPG, CAU and Mod.CAU solutions



(a) (b)

Figure 32: Steady state: $(\gamma, \omega, f, \phi_L^p, \phi_R^p) = (10, 4, 0, 8, 3)$. (a) Exact, Galerkin, FIC [152], HRPG ($\lambda = 0$) and HRPG ($\lambda \neq 0$) solutions; (b) Exact, DU, SUPG, CAU and Mod.CAU solutions



(a) (b)

Figure 33: Steady state: $(\gamma, \omega, f, \phi_L^p, \phi_R^p) = (10, 20, 0, 8, 3)$. (a) Exact, Galerkin, FIC [152], HRPG ($\lambda = 0$) and HRPG ($\lambda \neq 0$) solutions; (b) Exact, DU, SUPG, CAU and Mod.CAU solutions

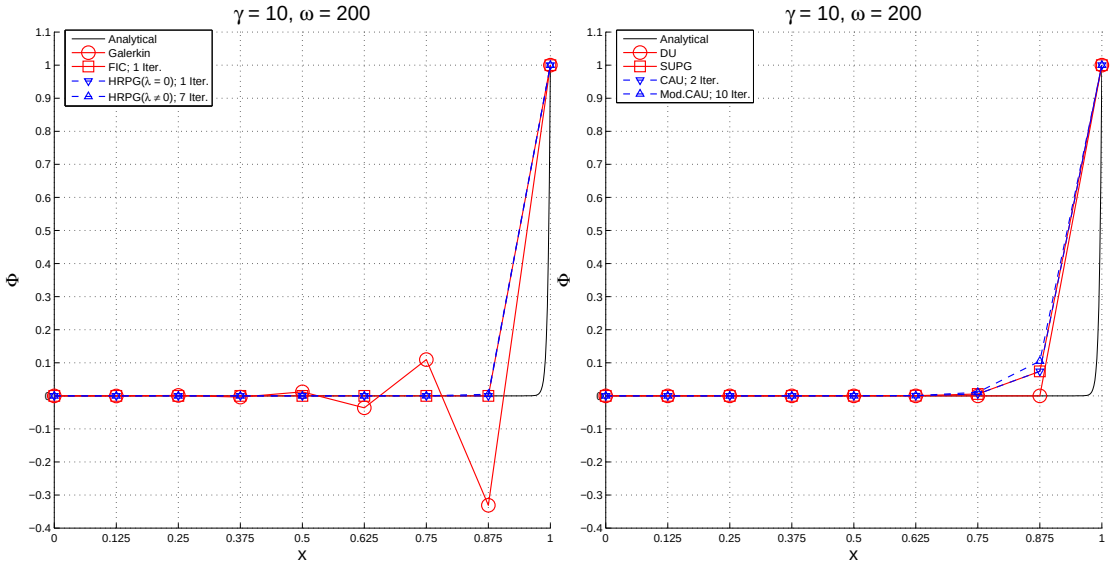


Figure 34: Steady state: $(\gamma, \omega, f, \phi_L^p, \phi_R^p) = (10, 200, 0, 0, 1)$. (a) Exact, Galerkin, FIC [152], HRPG ($\lambda = 0$) and HRPG ($\lambda \neq 0$) solutions; (b) Exact, DU, SUPG, CAU and Mod.CAU solutions

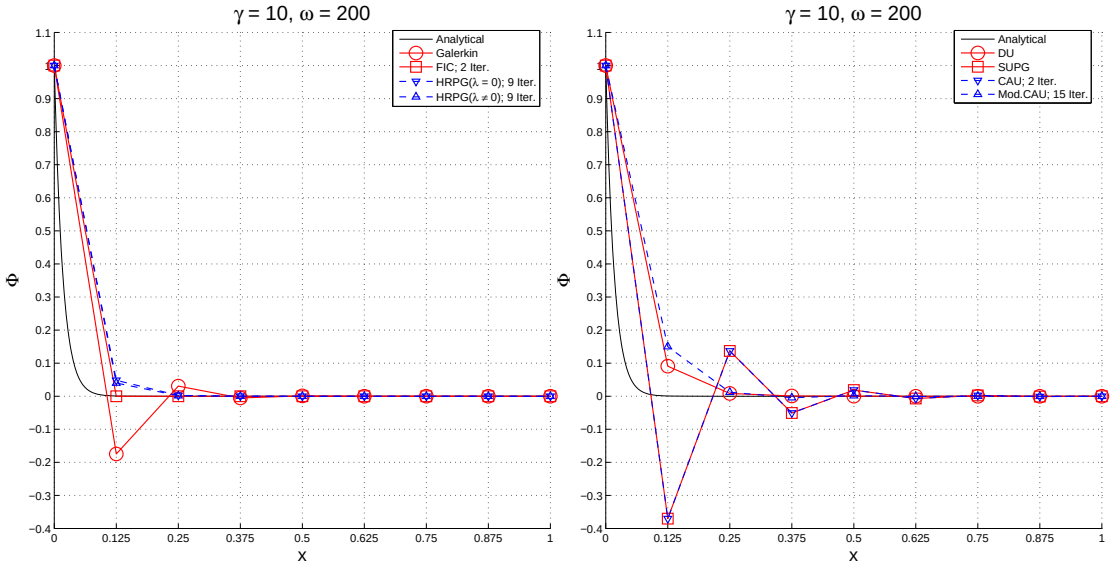
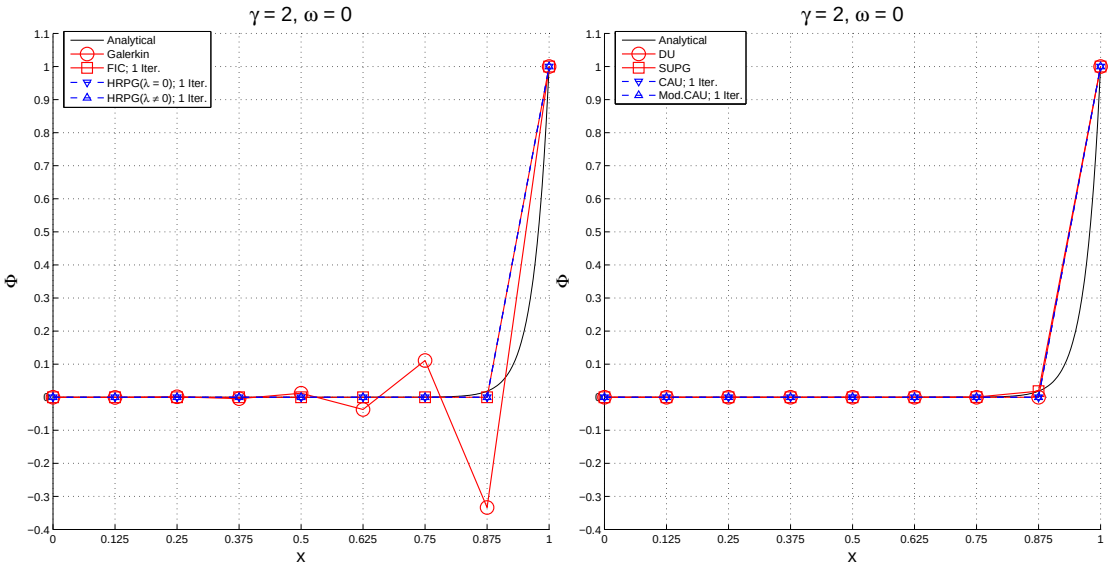
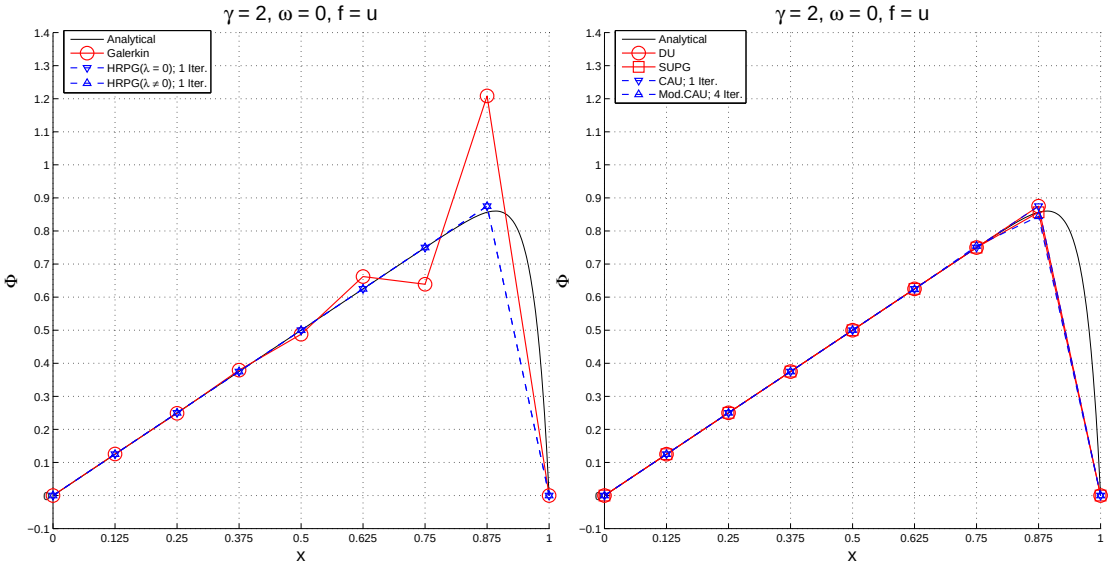


Figure 35: Steady state: $(\gamma, \omega, f, \phi_L^p, \phi_R^p) = (10, 200, 0, 1, 0)$. (a) Exact, Galerkin, FIC [152], HRPG ($\lambda = 0$) and HRPG ($\lambda \neq 0$) solutions; (b) Exact, DU, SUPG, CAU and Mod.CAU solutions



(a) (b)

Figure 36: Steady state: $(\gamma, \omega, f, \phi_L^p, \phi_R^p) = (2, 0, 0, 0, 1)$. (a) Exact, Galerkin, FIC [152], HRPG ($\lambda = 0$) and HRPG ($\lambda \neq 0$) solutions; (b) Exact, DU, SUPG, CAU and Mod.CAU solutions



(a) (b)

Figure 37: Steady state: $(\gamma, \omega, f, \phi_L^p, \phi_R^p) = (2, 0, u, 0, 0)$. (a) Exact, Galerkin, HRPG ($\lambda = 0$) and HRPG ($\lambda \neq 0$) solutions; (b) Exact, DU, SUPG, CAU and Mod.CAU solutions

boundary condition $\phi(x = 0) = 0$ is employed. The following initial condition was used:

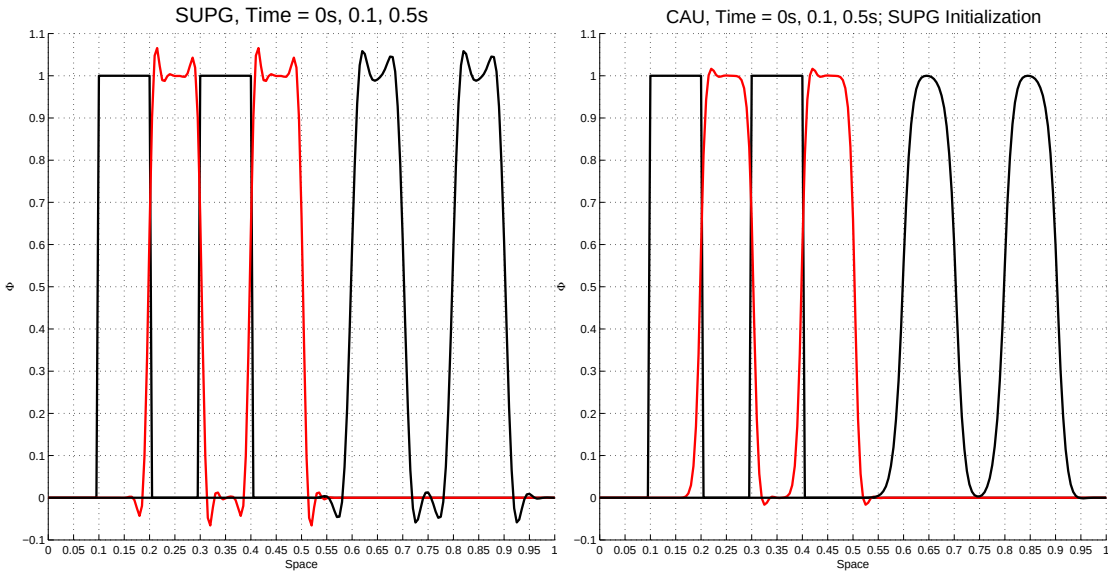
$$\phi(x, t = 0) = \begin{cases} 1 & \forall \quad x \in [0.1, 0.2] \cup [0.3, 0.4] \\ 0 & \text{else} \end{cases} \tag{2.80}$$

The above initial condition models a double rectangular pulse with simple discontinuities. The amplitude spectrum of this function decays only as fast as the *harmonic series* and hence is rich in high wave numbers. It is a challenging problem for the validation of any method for the control of dispersive oscillations and accuracy. The 1D domain is taken as $x \in [0, 1]$ and it is discretized with 200 two-node linear elements. The time step was chosen as $\Delta t = 0.001$s. This corresponds to a Courant number $C = 0.2$. The error was measured using Eq.(2.79) and a tolerance of `1e-4` was used. For the HRPG method with $\lambda = 0$ a tolerance of `1e-3` was used. The nonlinear iterations at every time step were initialized by the solution obtained by the SUPG method.

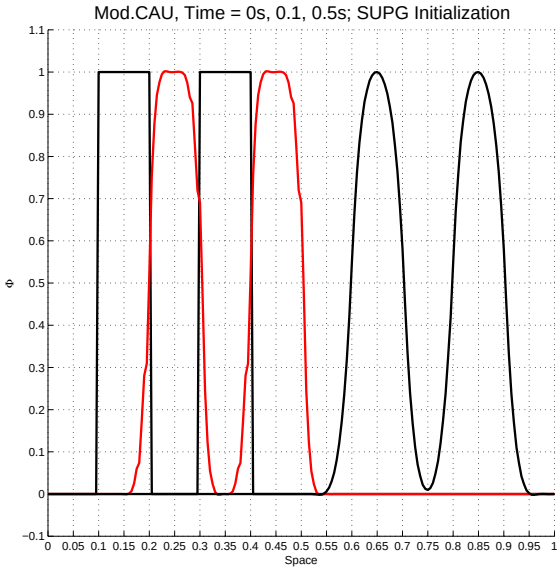
Figures 38,39 illustrate the solution obtained with the SUPG, CAU, modified CAU and HRPG methods. As expected the SUPG solution exhibits dispersive oscillations viz. Figure 38a. Appreciable control over the dispersive oscillations is obtained in the CAU method viz. Figure 38b. However slight crests and troughs do appear in the solution that gradually die out in time. These crests and troughs are reduced in the solution obtained by the modified CAU method at the cost of accuracy viz. Figure 38c. Best results were obtained with the HRPG method with $\lambda \neq 0$ which exhibits better control over the dispersive oscillations and maintains the symmetry of the initial solution viz. Figure 39b-c. On the other hand the HRPG method with $\lambda = 0$ is more diffusive than with $\lambda \neq 0$ and needs more iterations per time step for convergence viz. Figure 39a.

2.6 CONCLUSIONS

A high-resolution Petrov–Galerkin method is presented for the 1D convection–diffusion–reaction problem. The prefix 'high-resolution' is used here in the sense popularized by Harten, i. e.second order accuracy for smooth/regular regimes and good shock-capturing in nonregular regimes. The HRPG method could be understood as the combination of upwinding plus a nonlinear discontinuity-capturing operator. The distinction is that in general (multidimensions) the upwinding provided by $\mathbf{h}$ is not streamline and the discontinuity-capturing provided by $\mathbf{H} \cdot \hat{\mathbf{u}}^r$ is neither isotropic nor purely crosswind. The HRPG form can be considered as a particular class of the stabilized governing equations obtained via a finite calculus (FIC) procedure. For the 1D problem the HRPG method is similar to the CAU method with new definitions of the stabilization parameters. The 1D examples presented demonstrate that the method provides stabilized and essentially non-oscillatory i. e.monotone *to-the-eye* solutions for a wide range of the physical parameters and boundary conditions. It is interesting to note that the HRPG method without the linear upwinding term, i. e.using $\alpha = 0$ does solve all the steady-state examples to give high-resolution stabilized results. Nevertheless the presence of the linear perturbation terms improves the convergence of the nonlinear iterations especially for the transient problem.

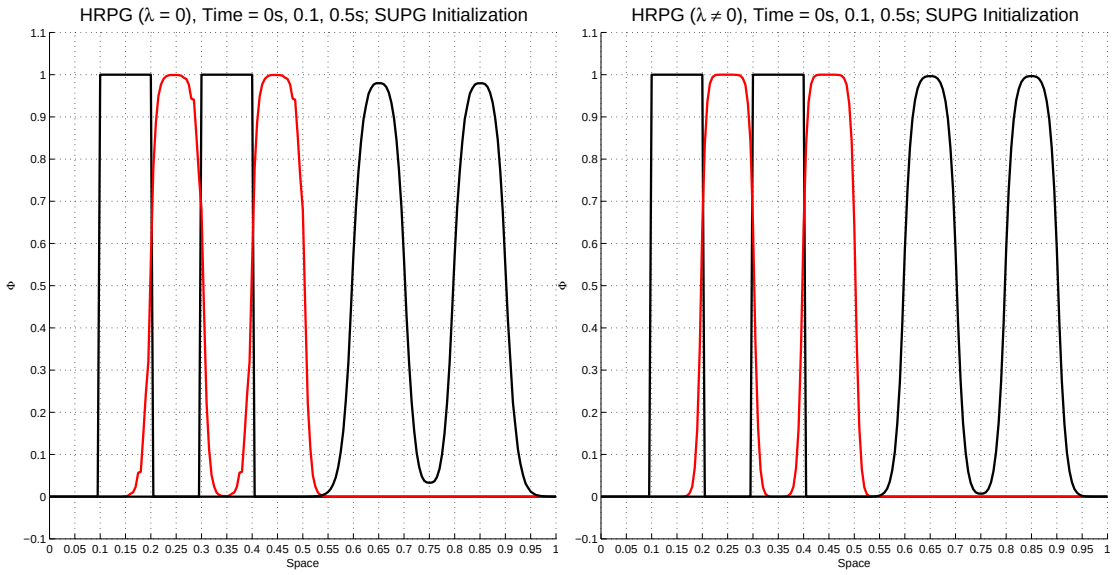


(a) (b)

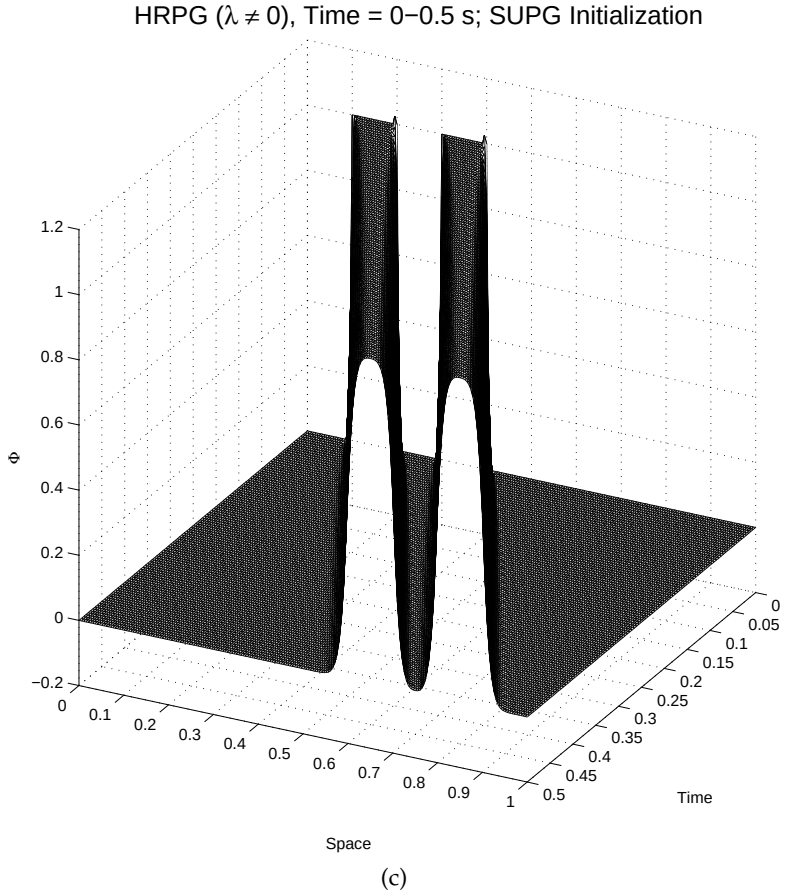


(c)

Figure 38: Transient pure convection; $u = 1m/s$, $\ell = 0.005m$, $\Delta t = 0.001s$. (a) SUPG solution; (b) CAU solution; (c) Mod.CAU solution



(a) (b)



(c)

Figure 39: Transient pure convection; $u = 1\text{m/s}$, $\ell = 0.005\text{m}$, $\Delta t = 0.001\text{s}$. (a) HRPG ($\lambda = 0$) solution; (b) HRPG ($\lambda \neq 0$) solution; (c) HRPG ($\lambda \neq 0$) solution (evolution plot)

yad bhāvam tad bhavati.
(As is the feeling, so is the result).

— a Sanskrit saying.

3

MULTIDIMENSIONAL EXTENSION OF THE HRPG METHOD

3.1 INTRODUCTION

This chapter is a continuation of chapter §2 wherein a nonlinear high-resolution Petrov–Galerkin (HRPG) method was presented for the convection–diffusion–reaction problem in 1D. In this chapter we develop the extension to multi dimensions of the HRPG method for the singularly perturbed convection–diffusion–reaction problem.

It is well known that the solution to the stationary convection–diffusion–reaction problem may develop two types of layers: *exponential* and *parabolic* layers. The exponential layers are usually found in the convection-dominant cases near the boundary or close to the regions where the source term is non regular. Parabolic layers, which are of larger width than exponential layers, are found in the reaction-dominant cases near the boundary or close to the regions where the source term is non regular and in the convection-dominated cases along the characteristics of the solution. The later characteristic internal/boundary layers are usually found only in higher dimensions and hence have no instances in 1D [175]. In other words we do not have a straightforward quantification of the characteristic layers in 1D. For this reason a direct extension of the definition of the stabilization parameters α, β derived for 1D will not be efficient to resolve these layers.

The numerical artifacts that are formed across the parabolic layers are usually manifested as the Gibbs phenomenon. Nevertheless there exists a subtle difference[1] between the numerical artifacts formed across the characteristic layers and those formed across the layers in the reaction-dominant cases. Consider a rectangular domain discretized by structured bilinear block finite elements. We use the following notation (introduced in [137]) to represent a generic compact stencil corresponding to any interior node (i,j) of the considered structured mesh.

$$\{\circ^{j+1}, \circ^{j}, \circ^{j-1}\}\mathbf{A}\{\circ^{i-1}, \circ^{i}, \circ^{i+1}\}^{t} = 0 \tag{3.1}$$

where $\mathbf{A}$ represents the matrix of the stencil coefficients. For instance, if the standard mass matrix obtained in the Galerkin FEM be assembled for a structured rectangular mesh then we may express the corresponding stencil as follows:

$$\mathbf{A}^{m} := \frac{\ell_2}{6}\{1,4,1\}^{t}\frac{\ell_1}{6}\{1,4,1\} = \frac{\ell_1\ell_2}{36}\begin{bmatrix} 1 & 4 & 1 \\ 4 & 16 & 4 \\ 1 & 4 & 1 \end{bmatrix} \tag{3.2}$$

1 related to the cause and size of these numerical artifacts

$$\{\circ^{j+1},\circ^{j},\circ^{j-1}\}\mathbf{A}^{m}\{\circ^{i-1},\circ^{i},\circ^{i+1}\}^{t} := \frac{\ell_1\ell_2}{36}\left\{\begin{array}{l}\Phi^{i-1,j+1}+4\Phi^{i,j+1}+\Phi^{i+1,j+1}+\\ 4\Phi^{i-1,j}+16\Phi^{i,j}+4\Phi^{i+1,j}+\\ \Phi^{i-1,j-1}+4\Phi^{i,j-1}+\Phi^{i+1,j-1}\end{array}\right\} \tag{3.3}$$

The stencil coefficient matrix associated with the convective term in the Galerkin FEM can be expressed as follows:

$$\mathbf{A}^{c} := \frac{\ell_2}{6}\{1,4,1\}^{t}\frac{u_1}{2}\{-1,0,1\} + \frac{u_2}{2}\{1,0,-1\}^{t}\frac{\ell_1}{6}\{1,4,1\} \tag{3.4}$$

Note that one may arrive at the terms in Eq.(3.2) and Eq.(3.4) via a 1D mass type averaging of their respective counterparts in 1D, i. e.replacing $(\ell_1/6)\{1,4,1\}$ with $(\ell_2/6)\{1,4,1\}^{t}$ $(\ell_1/6)\{1,4,1\}$ and $(u_1/2)\{-1,0,1\}$ with $(\ell_2/6)\{1,4,1\}^{t}(u_1/2)\{-1,0,1\}$ etc. Although this 1D mass type averaging leads to a higher-order approximation for smooth solution profiles, it unfortunately leads to the Gibbs phenomenon across layers. Unlike in the reaction-dominant case where it is the numerical solution that undergoes the 1D mass type averaging, in the convection-dominant case it is the derivatives of the numerical solution that undergoes the same. Thus, the Gibbs phenomenon across the characteristic layers in the later case is proportional to the variation in the derivatives of the solution across the characteristic layers. Despite this subtle difference in the Gibbs phenomenon associated with the characteristic layers in the convection-dominated case, we choose to treat them by the same strategy that we use to treat the numerical artifacts about the parabolic layers in the reaction-dominant case.

The outline of this chapter is as follows. In §3.2 we design a nondimensional element number that quantifies the characteristic internal/boundary layers. Anisotropic element length vectors $\mathbf{l}^{i}$ are introduced in §3.3 and using them objective characteristic tensors $\mathbf{h}$ and $\mathbf{H}$ associated with the HRPG method are defined. The stabilization parameters α^{i}, β^{i} used in the definition of $\mathbf{h}, \mathbf{H}$ are defined in §3.4 by a direct extension of their respective expressions in 1D. The definitions of β^{i} are updated to include the new dimensionless number introduced in §3.2. In Box 1 we summarize the HRPG method in multi dimensions. Several numerical examples are presented in §3.5 that throws light on the performance of the proposed method. Finally we arrive at some conclusions and outlook in §3.6

3.2 QUANTIFYING CHARACTERISTIC LAYERS

In this section we design a nondimensional element number that quantifies the characteristic internal/boundary layers. By quantification we mean that its should serve a similar purpose as the element Peclet number γ for the exponential layers in convection dominant cases and the dimensionless number $\omega := 2\gamma\sigma$ for the parabolic layers in the reaction dominant cases.

Consider the following singularly perturbed ($k \ll u$) convection–diffusion problem in 2D:

$$u\frac{\partial\phi}{\partial x} - k\left(\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}\right) = 0 \qquad \text{in } \Omega \tag{3.5a}$$

$$\phi = \phi^{p} \qquad \text{on } \Gamma \tag{3.5b}$$

where, Ω is a rectangular domain ABCD as shown in Figure 40a, Γ is the domain boundary and ϕ^p is the prescribed value of ϕ on Γ. The origin of the 2D axes is taken as the midpoint of AD. Consider $\phi^p = 0$ everywhere except on AD where it is defined as follows:

$$\phi^p(0, y) = f(y) := H(y + a) - H(y - a), a > 0 \tag{3.6a}$$

$$H(y) := \frac{1 + \text{sgn}(y)}{2} = \begin{cases} 0 & y < 0 \\ 0.5 & y = 0 \\ 1 & y > 0 \end{cases} \tag{3.6b}$$

The function $f(y)$ is discontinuous at $y = \pm a$ and its shape can be described as a rectangular pulse. A well known virtue of the solution $\phi(x, y)$ is that these discontinuities are immediately smoothed out in the interior of the domain, thus leading to parabolic layers along the characteristic lines of the problem [175]. In accordance with singular perturbation theory and by the method of matched asymptotic expansions [117], the leading term describing the characteristic layer is given by,

$$\phi(x, y) \approx \frac{1}{2}\left[\text{erf}\left(\sqrt{\frac{u}{4kx}}(y + a)\right) - \text{erf}\left(\sqrt{\frac{u}{4kx}}(y - a)\right)\right] \tag{3.7}$$

where erf represents the *error function* and is defined as follows:

$$\text{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2}\, dz \tag{3.8}$$

The approximation given in Eq.(3.7) is uniformly valid to $O(1)$ in a region away from the exponential layers formed near the boundary BC [117]. Figure 40b illustrates the solution given by Eq.(3.7) about a cross-section SS′ (cf. Figure 40a) located at a distance x from the boundary AD.

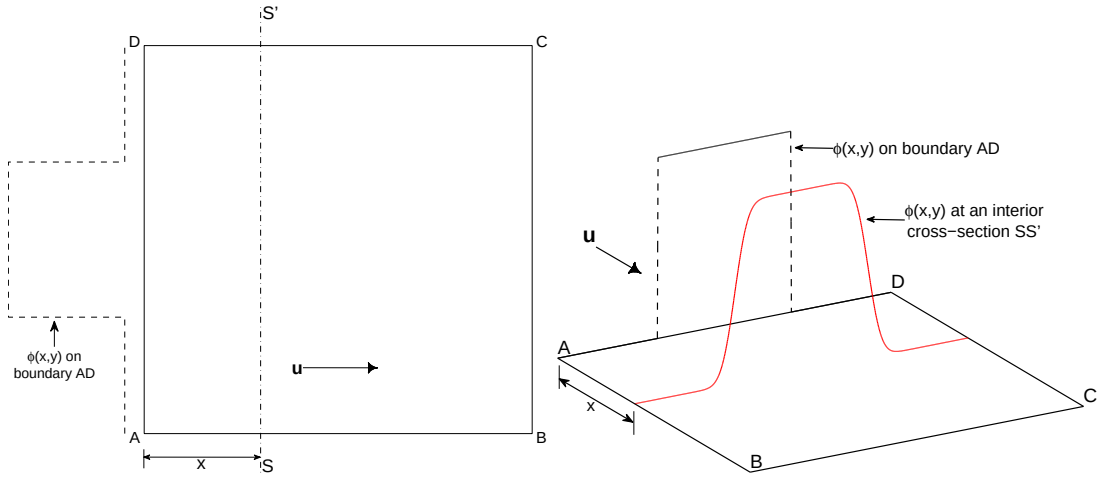


Figure 40: A singularly perturbed convection–diffusion problem. (a) The problem domain ABCD and boundary conditions; (b) The solution about a cross-section SS′ located at a distance x from the boundary AD

Consider now the heat equation posed on an infinite domain:

$$\frac{\partial \phi}{\partial t} - k\frac{\partial^2 \phi}{\partial y^2} = 0, \qquad \text{in } \Omega := \{(y,t) \mid y \in (-\infty,\infty), t \in [0,\infty)\} \tag{3.9a}$$

$$\phi(y, t=0) = f(y) \quad f(y) := [H(y+a) - H(y-a)]\,, a > 0 \tag{3.9b}$$

Note that we have initialized the solution with a function $f(y)$ that was used earlier in Eq.(3.6a) to prescribe the Dirichlet boundary condition. The exact solution for the problem (3.9) can be expressed as follows:

$$\phi(y,t) = \frac{1}{2}\left[\operatorname{erf}\left(\frac{y+a}{\sqrt{4kt}}\right) - \operatorname{erf}\left(\frac{y-a}{\sqrt{4kt}}\right)\right] \tag{3.10}$$

Clearly, replacing t with (x/u) in Eq.(3.10) we recover the leading term describing the characteristic layers given by Eq.(3.7). Note that (x/u) is the time required to travel a distance x along the characteristic lines. This resemblance is due to the fact that in regions far-away from the domain boundaries the convective and diffusive effects do not interact, i. e.convection just carries the diffusing solution along the characteristic lines [172].

Next, we try to relate the solution of the heat equation with the solution of the diffusion–reaction problem. The statement of the diffusion–reaction problem posed on an infinite domain is:

$$-k\frac{d^2\phi}{dy^2} + s\phi = sf(y) \quad \text{in } \Omega := \{y \mid y \in (-\infty,\infty)\} \tag{3.11a}$$

$$\phi(y) = 0 \qquad \text{at } y = \pm\infty \tag{3.11b}$$

The exact solution for the above problem can be expressed as follows:

$$\phi(y) = \frac{\operatorname{sgn}(y+a)}{2}\left[1 - e^{-\xi|y+a|}\right] - \frac{\operatorname{sgn}(y-a)}{2}\left[1 - e^{-\xi|y-a|}\right] \tag{3.12}$$

where, $\xi := \sqrt{s/k}$. Figures 41a and 41b illustrate the solution of the heat equation given by Eq.(3.10) and the solution of the diffusion–reaction problem given by Eq.(3.12) respectively. Clearly these two solutions have distinct profiles. Nevertheless, they share a common trait of possessing parabolic layers, i. e.the first-order derivatives in the direction perpendicular to the layers have magnitude $O(1/\sqrt{k})$. We refer to [175] for further details about parabolic and exponential layers.

Now we pose the following design problem: *Relate s and t such that the parabolic layers in the solution of the heat equation i. e.Eq.(3.10) and the solution of the diffusion–reaction problem i. e.Eq.(3.12) have the same width.*

In the following developments the width of the layer is taken as the distance within which the value of ϕ varies from 1% to 99% of $[\max(f(y)) - \min(f(y))]$. We choose $f(y) = H(y)$ to simplify the algebra. For this choice of $f(y)$ the solution of the heat equation and the diffusion–reaction problem can be expressed as in Eq.(3.13) and Eq.(3.14) respectively.

$$\phi(y,t) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{y}{\sqrt{4kt}}\right)\right] \tag{3.13}$$

$$\phi(y) = \frac{1}{2}\left[1 + \operatorname{sgn}(y)\left(1 + e^{\xi|y|}\right)\right] \tag{3.14}$$

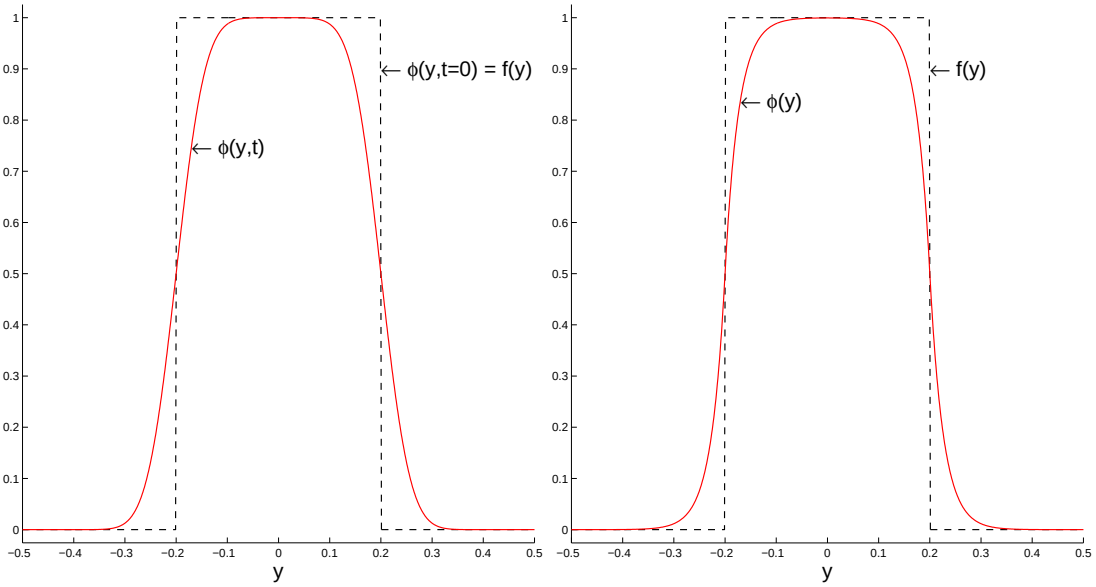


Figure 41: Parabolic layers in the solution of: (a) the heat equation given by Eq.(3.10) using $k = 0.01$ and $t = 0.1$; (b) the diffusion–reaction problem given by Eq.(3.12) using $k = 0.01$ and $s = 10\sqrt{2}$

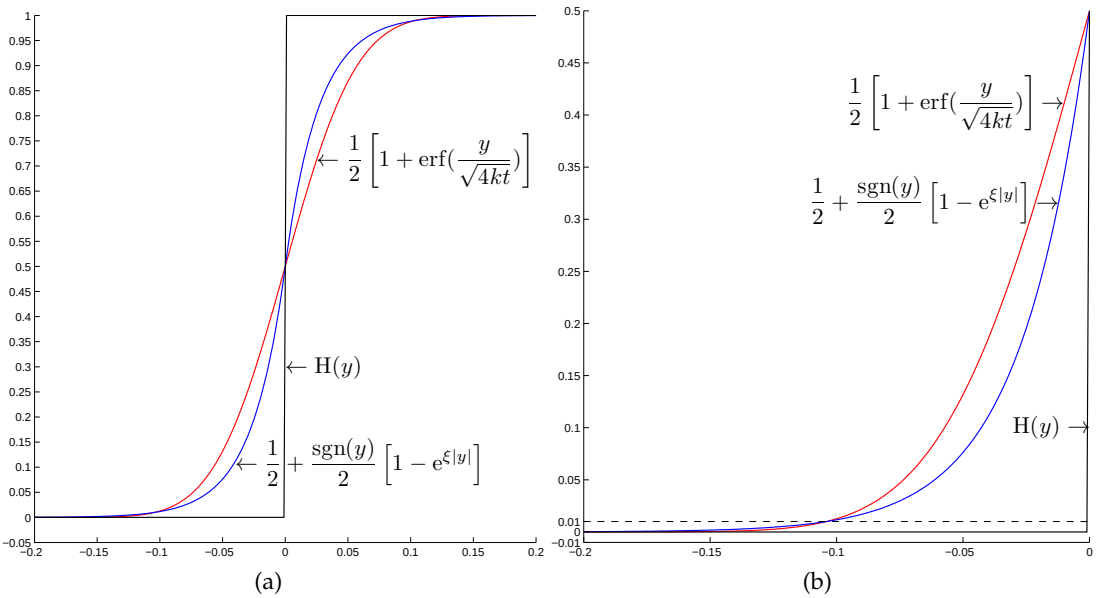


Figure 42: Matching the layers in the solution of the heat equation and the diffusion–reaction problem. (a) plot domain: $[-0.2, 0.2]$, $k = 0.01$, $t = 0.1$ and $s := (\sqrt{2}/t) = 10\sqrt{2}$; (b) plot domain: $[-0.2, 0]$, the two solutions always meet at a value equal to 0.01

Let $y = -y^*$ be the distance at which the solutions given by Eq.(3.13) and Eq.(3.14) have a value equal to 1% of $[\max(H(y)) - \min(H(y))]$, i. e. 0.01. Due to the inherent

symmetry of the problem, these solutions at $y = y^*$ will attain a value equal to 99% of $[max(H(y)) - min(H(y))]$, i. e.0.99. Thus we have,

$$\frac{1}{2}\left[1 + \text{erf}\left(\frac{-y^*}{\sqrt{4kt}}\right)\right] = \frac{1}{100} = \frac{e^{-\xi y^*}}{2} \tag{3.15}$$

Solving Eq.(3.15) we get the following equation relating s and t,

$$st = \frac{1}{4}\left[\frac{\ln(50)}{\text{erf}^{-1}(49/50)}\right]^2 \approx \sqrt{2} \quad \Rightarrow \boxed{s \approx \frac{\sqrt{2}}{t}} \tag{3.16}$$

The above relation between s and t guarantees that the parabolic layers that appear in the solutions of the heat equation and the diffusion–production problem will have the same width. In Figure 42 using Eq.(3.16) these solutions having the same layer width are compared.

We now address the initial objective of quantifying the characteristic layers found in the singularly perturbed convection–diffusion problem (3.5). Consider a fictitious reaction coefficient s_c and an associated dimensionless element number ω_c defined as below.

$$s_c := \frac{\sqrt{2}u}{x} \quad , \quad \omega_c = \frac{s_c \ell^2}{k} \tag{3.17}$$

where ℓ is an appropriate element length measure. We have used the substitution $t = (x/u)$ in Eq.(3.16) to arrive at the expression for s_c in Eq.(3.17). Recall that we have used earlier the same substitution in the solution of the heat equation to recover the leading term describing the characteristic layers in the solution of the convection–diffusion problem. We may use this fictitious reaction coefficient s_c to relate the characteristic layers of the convection–diffusion problem to similar[2] parabolic layers of the 1D diffusion–production problem. In this sense, the nondimensional element number ω_c quantifies the characteristic layers and could be used in the design of stabilization parameters to control the numerical artifacts about these layers.

Note that the value of s_c is a function of x, i. e.s_c is inversely proportional to the distance from the source of the discontinuity along the characteristic lines. In fact this is how the characteristic layers in the solution of the convection–diffusion problem behave, i. e.their width widens as we move away from the source of the discontinuity along the characteristic lines. However from the design point-of-view, a variable definition of s_c and hence of ω_c is inconvenient. This is due to the fact that the characteristic lines could be arbitrary curves governed by the velocity field and hence finding the distance x along these lines need not be straight-forward. Hence we redefine s_c and ω_c using an appropriate element characteristic length ℓ_c.

$$s_c := \frac{\sqrt{2}u}{\ell_c} \quad , \quad \omega_c = \frac{s_c \ell^2}{k} \tag{3.18}$$

3.3 OBJECTIVE CHARACTERISTIC TENSORS

In this section we present the objective characteristic tensors $\mathbf{h}$ and $\mathbf{H}$ used in the extension of the HRPG method to higher dimensions. In the developments to follow,

2 in the sense of matched layer widths

only the multi-linear block finite elements are considered. Here *objectivity* is to be understood in the sense popular in tensor calculus, i. e.independence of the method on the description of the reference frame and admissible node numbering permutations of the mesh.

Consider the following definition for the element length vectors $\mathbf{l}^i$:

$$\boxed{\mathbf{l}^i := \mathbf{J} \cdot \tilde{\mathbf{l}}^i} \quad ; \quad \mathbf{J}_{ij} := \frac{\partial x_i}{\partial \widetilde{x}_j} \quad ; \quad \tilde{\mathbf{l}}^1 := 2(1,0) \quad ; \quad \tilde{\mathbf{l}}^2 := 2(0,1) \tag{3.19}$$

where $\mathbf{J}$ represents the Jacobian matrix of bijective mappings from the local to global coordinate systems, x_i and $\widetilde{x}_i$ represent the global and local coordinates respectively and $\tilde{\mathbf{l}}^i$ are fixed vectors along the axes of the local frame. Figure 43 illustrates the element length vectors $\mathbf{l}^i$ obtained at any arbitrary point $P(\widetilde{x}_1, \widetilde{x}_2)$ within a 2D bilinear block finite element. The expression for the vectors $\mathbf{l}^i$ in 2D and at this point P can be simplified to the following:

$$\mathbf{l}^1 = \frac{1-\widetilde{x}_2}{2}\mathbf{E}^{12} + \frac{1+\widetilde{x}_2}{2}\mathbf{E}^{43} \quad ; \quad \mathbf{l}^2 = \frac{1-\widetilde{x}_1}{2}\mathbf{E}^{14} + \frac{1+\widetilde{x}_1}{2}\mathbf{E}^{23} \tag{3.20}$$

where $\mathbf{E}^{ab}$ is the edge vector pointing from node a to node b.

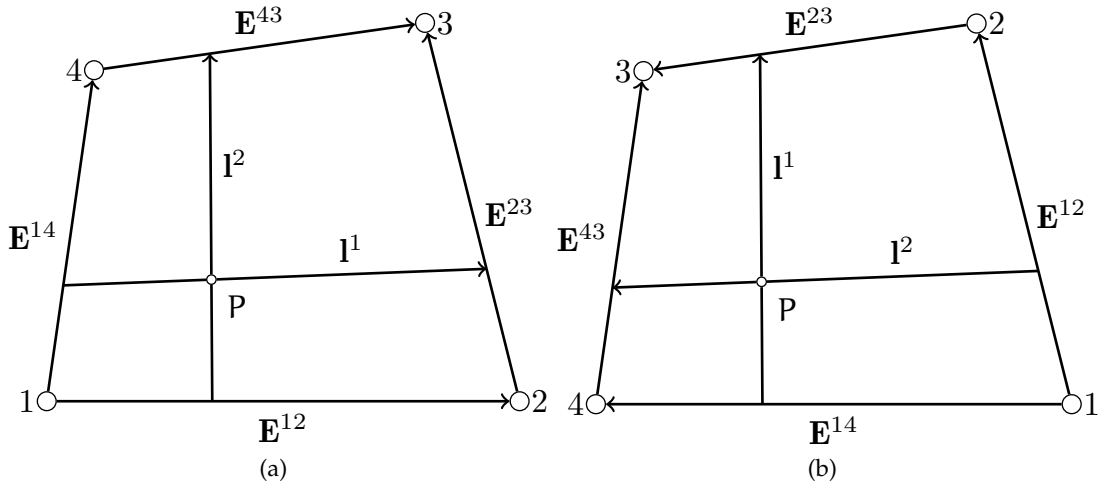


Figure 43: Anisotropic element length vectors $\mathbf{l}^i$ obtained at any arbitrary point $P(\widetilde{x}_1, \widetilde{x}_2)$ within a 2D bilinear block finite element. The sub-figures (a) and (b) illustrate $\mathbf{l}^i$ obtained for two admissible global node numbering permutations

Let α^i, β^i be stabilization parameters calculated along the element length vectors $\mathbf{l}^i$ and with the following properties: a) $(\mathbf{u} \cdot \mathbf{l}^i)\alpha^i \geqslant 0 \; \forall \; i$, b) $\beta^i \geqslant 0 \; \forall \; i$ and c) only scalars and free vectors[3] are used in their respective definitions. The definition of these parameters is delayed until §3.4. The characteristic tensors $\mathbf{h}$ and $\mathbf{H}$ are calculated as: $\mathbf{h} := 0.5\alpha^i\mathbf{l}^i$, $\mathbf{H} := 0.5(\beta^i/|\mathbf{l}^i|)[\mathbf{l}^i \otimes \mathbf{l}^i]$. Thus in 2D the characteristic tensors could be expressed as follows:

$$\mathbf{h} := \alpha^1\mathbf{l}^1 + \alpha^2\mathbf{l}^2 \quad ; \quad \mathbf{H} := \frac{\beta^1}{|\mathbf{l}^1|}[\mathbf{l}^1 \otimes \mathbf{l}^1] + \frac{\beta^2}{|\mathbf{l}^2|}[\mathbf{l}^2 \otimes \mathbf{l}^2] \tag{3.21}$$

3 If one is interested only in the magnitude and direction of the vector and does not think of it as situated at any particular location, then it is called a free vector

Using $\mathbf{h}, \mathbf{H}$ as defined above we calculate the perturbation p_h associated with the HRPG method as described earlier in Eq.(2.6), i. e.

$$p_h := [\mathbf{h} + \mathbf{H} \cdot \hat{\mathbf{u}}^r] \cdot \nabla w_h \tag{3.22}$$

$$\mathbf{u}^r := \frac{R(\phi_h)}{|\nabla \phi_h|^2} \nabla \phi_h; \quad \Rightarrow \hat{\mathbf{u}}^r := \frac{\mathbf{u}^r}{|\mathbf{u}^r|} = \frac{\operatorname{sgn}[R(\phi_h)]}{|\nabla \phi_h|} \nabla \phi_h \tag{3.23}$$

The definition of $\mathbf{h}$ and $\mathbf{H}$ given by Eq.(3.21) guarantees the objectivity of the HRPG method. Reference frame independence can be verified by the fact that the tensors $\mathbf{h}$ and $\mathbf{H}$ obey the same tensor transformation rules as any other free tensor associated with the problem, e. g.the velocity vector $\mathbf{u}$. Admissible node numbering permutations only swap one element length vector with the other (possibly with a change of sign) as shown in Figure 43b. Due to the properties of α^i, β^i and by their definition, the characteristic tensors $\mathbf{h}$ and $\mathbf{H}$ are invariant with respect to these swaps in $\mathbf{l}^i$.

Remark 1 : As noted in §2.2 the multidimensional HRPG method could be arrived at from the FIC equations with an appropriate definition of the characteristic length as $\mathbf{h}^{fic} := \mathbf{h} + \mathbf{H} \cdot \hat{\mathbf{u}}^r$.

Remark 2: A figure similar to Figure 43a were presented earlier in [92] (cf. Fig. 3.2, pp. 55), [22] (cf. Fig. 3.4, pp. 215) and [190] (cf. Figure 2, pp. 2205). Therein the element length vectors $\mathbf{l}^i$ evaluated at the centroid of the element were used to define a *scalar* element size measure. The distinction here is to use these $\mathbf{l}^i$ to arrive at *objective characteristic tensors* $\mathbf{h}$ and $\mathbf{H}$ that treat effectively the anisotropy of the finite element.

3.4 STABILIZATION PARAMETERS

Except for the modification to include the new dimensionless number introduced in §3.2 that quantifies the characteristic layers, the definition of the stabilization parameters α^i, β^i calculated along the element length vectors $\mathbf{l}^i$ are a direct extension of their counterparts in 1D summarized in §2.5.6. Following this line, in multi dimensions and along $\mathbf{l}^i$ we define the following nondimensional element numbers:

$$\gamma^i := \frac{\mathbf{u} \cdot \mathbf{l}^i}{2k} \quad ; \quad \omega^i := \frac{s|\mathbf{l}^i|^2}{k} \quad ; \quad \sigma^i := \frac{s|\mathbf{l}^i|^2}{\mathbf{u} \cdot \mathbf{l}^i} \tag{3.24}$$

Following Eq.(3.18), the fictitious reaction coefficient $\widehat{s}^i$ and the associated dimensionless number $\widehat{\omega}^i$ along $\mathbf{l}^i$ are calculated as follows.

$$\widehat{s}^i := \max_{j \neq i} \sqrt{2} \frac{|\mathbf{u} \cdot \mathbf{l}^j|}{|\mathbf{l}^j|^2} \quad ; \quad \widehat{\omega}^i := \frac{\widehat{s}^i |\mathbf{l}^i|^2}{k} \tag{3.25}$$

Following Eq.(2.78) the stabilization parameters α^i along $\mathbf{l}^i$ are calculated as follows.

$$\boxed{\alpha^i := \lambda^i \ \operatorname{sgn}(\mathbf{u} \cdot \mathbf{l}^i) \max\left\{ \left[1 - \frac{1}{|\gamma^i|} \right], 0 \right\}} \quad ; \quad \lambda^i := \frac{1}{3(1 + \sqrt{|\sigma^i|})} \tag{3.26}$$

Assuming that the discretization in time is done using the implicit trapezoidal rule and following Eq.(2.76) we calculate the nonlinear pseudo-reaction coefficient δ as follows.

$$\delta \approx \frac{1}{\theta \Delta t} \frac{\| \tilde{\phi}_h - \phi_h^n \|_\infty}{\| \tilde{\phi}_h \|_\infty} \tag{3.27}$$

Following Eq.(2.72) we define the effective convection, diffusion and reaction coefficients along $\mathbf{l}^i$ as follows.

$$\widetilde{u}^i := \frac{\mathbf{u}\cdot\mathbf{l}^i}{|\mathbf{l}^j|} - \frac{\alpha^i|\mathbf{l}^i|s}{2} - \frac{\alpha|\mathbf{l}^i|\delta}{2} \quad ; \quad \widetilde{k}^i := k + \frac{\alpha^i\mathbf{u}\cdot\mathbf{l}^i}{2} \quad ; \quad \widetilde{s} := s + \delta \tag{3.28}$$

Likewise following Eq.(2.73), the effective element dimensionless numbers along $\mathbf{l}^i$ can be calculated as,

$$\widetilde{\gamma}^i := \frac{|\widetilde{u}^i||\mathbf{l}^i|}{2\widetilde{k}^i} \quad ; \quad \widetilde{\sigma}^i := \frac{\widetilde{s}|\mathbf{l}^i|}{|\widetilde{u}^i|} \quad ; \quad \widetilde{\omega}^i := \frac{\widetilde{s}|\mathbf{l}^i|^2}{\widetilde{k}^i} \tag{3.29}$$

Finally, following Eq.(2.74) the stabilization parameters β^i along $\mathbf{l}^i$ are calculated using the dimensionless numbers $\widetilde{\gamma}^i, \widetilde{\sigma}^i, \widetilde{\omega}^i$ and $\widehat{\omega}^i$ as follows:

$$\boxed{\beta^i := \max\left\{\left[\frac{2}{3}\left(\frac{\widetilde{\sigma}^i+3}{\widetilde{\sigma}^i+2}\right) - \left(\frac{4}{\widetilde{\omega}^i+4\widetilde{\gamma}^i}\right)\right], \left[\frac{2}{3} - \frac{4}{\widehat{\omega}^i}\right], 0\right\}} \tag{3.30}$$

The inclusion of the term $(2/3) - (4/\widehat{\omega}^i)$ in the definition of β^i is the only modification from a straight-forward extension to multi dimensions of the definition of its counterpart in 1D. This expression follows from Eq.(2.55) and the justification is based on the strategy we employ to treat the numerical artifacts about the characteristic layers— to treat them just like the numerical artifacts about the parabolic layers in the reaction-dominant case.

3.5 EXAMPLES

In this section we present some examples in 2D for the convection–diffusion–reaction problem defined by Eq.(2.1). The domain Ω is discretized by considering both structured and unstructured meshes made up of just the bilinear block finite elements. The unstructured meshes are obtained by randomly perturbing the interior nodes of structured meshes with coordinates (x_i, y_i) as follows [60, 128]:

$$x_i' = x_i + \ell_1\delta\,\texttt{rand}() \quad ; \quad y_i' = y_i + \ell_2\delta\,\texttt{rand}() \tag{3.31}$$

where, (x_i', y_i') represent the corresponding coordinates of the unstructured mesh, ℓ_1, ℓ_2 represent the mesh sizes of the structured mesh, δ is a mesh distortion parameter and `rand()` is a function that returns random numbers uniformly distributed in the interval $[-1, 1]$. Figure 44 illustrates two types of unstructured meshes obtained by this procedure using a 20×20 square mesh and the parameter $\delta = 0.2$. In Figure 44a, $\delta = 0.2$ was chosen for all the internal nodes of the mesh. Whereas for the nodes adjacent to the boundary in the mesh shown in Figure 44b, the perturbation perpendicular to the boundary was set to zero. The unstructured meshes obtained using the former and later techniques are denoted as 'Type I' and 'Type II' respectively.

3.5.1 *Steady-state examples*

In this section we illustrate the performance of the HRPG method for the stationary convection–diffusion–reaction problem. Unless otherwise specified, in all the examples the following data is considered. The domain Ω is $[0, 1] \times [0, 1]$. Each example is

RESIDUAL

$$R(\phi_h) := \frac{\partial \phi_h}{\partial t} + \mathbf{u}\cdot\boldsymbol{\nabla}(\phi_h) - k\Delta(\phi_h) + s\phi_h - f(\mathbf{x})$$

HRPG METHOD

Find $\phi_h : [0,T] \mapsto V^h$ such that $\forall w_h \in V_0^h$ we have,

$$a(w_h,\phi_h) + \sum_e \Big(\mathbf{h}\cdot\boldsymbol{\nabla} w_h, R(\phi_h)\Big)_{\Omega_h^e} + \Big(\frac{|R(\phi_h)|}{|\boldsymbol{\nabla}\phi_h|}\mathbf{H}\cdot\boldsymbol{\nabla} w_h, \boldsymbol{\nabla}\phi_h\Big)_{\Omega_h^e} = l(w_h)$$

$$\text{HRPG weight} \rightarrow w_h + \left[\mathbf{h} + \frac{\operatorname{sgn}[R(\phi_h)]}{|\boldsymbol{\nabla}\phi_h|}\mathbf{H}\cdot\boldsymbol{\nabla}\phi_h\right]\cdot\boldsymbol{\nabla} w_h$$

DEFINITIONS

$$R(\widetilde{\phi}_h) \approx \frac{\widetilde{\phi}_h - \phi_h^n}{\theta\Delta t} + \mathbf{u}\cdot\boldsymbol{\nabla}(\widetilde{\phi}_h) - k\Delta(\widetilde{\phi}_h) + s\widetilde{\phi}_h - f$$

$$\phi_n^{n+1} = \Big(\frac{1}{\theta}\Big)\widetilde{\phi}_h + \Big(\frac{\theta-1}{\theta}\Big)\phi_h^n \quad ; \quad \Delta t = t^{n+1} - t^n \quad ; \quad \theta \in (0,1)$$

$$\boxed{\mathbf{l}^i := 2\mathbf{J}\cdot\widetilde{\mathbf{e}}^i} \quad ; \quad \mathbf{J}_{ij} := \frac{\partial x_i}{\partial \widetilde{x}_j} \quad ; \quad \widehat{s}^i := \max_{j\neq i}\sqrt{2}\frac{|\mathbf{u}\cdot\mathbf{l}^j|}{|\mathbf{l}^j|^2}$$

$$\gamma^i := \frac{\mathbf{u}\cdot\mathbf{l}^i}{2k} \quad ; \quad \sigma^i := \frac{s|\mathbf{l}^i|^2}{\mathbf{u}\cdot\mathbf{l}^i} \quad ; \quad \widehat{\omega}^i := \frac{\widehat{s}^i|\mathbf{l}^i|^2}{k}$$

$$\lambda^i := \frac{1}{3(1+\sqrt{|\sigma^i|})} \quad ; \quad \delta \approx \frac{1}{\theta\Delta t}\frac{\|\,\widetilde{\phi}_h - \phi_h^n\,\|_\infty}{\|\,\widetilde{\phi}_h\,\|_\infty}$$

$$\boxed{\alpha^i := \lambda^i\ \operatorname{sgn}(\mathbf{u}\cdot\mathbf{l}^i)\max\left\{\left[1 - \frac{1}{|\gamma^i|}\right], 0\right\}}$$

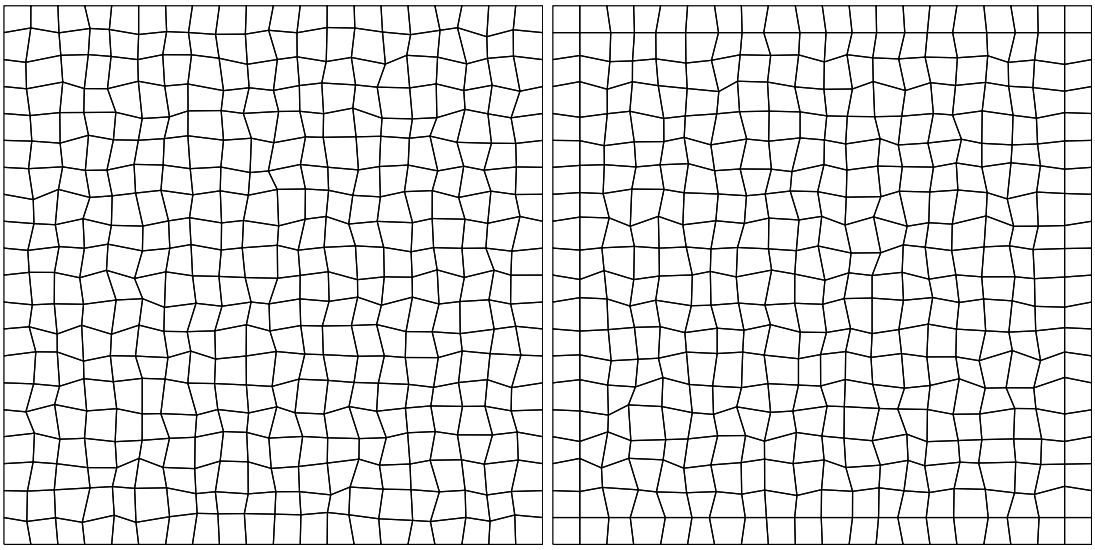
$$\widetilde{u}^i := \frac{\mathbf{u}\cdot\mathbf{l}^i}{|\mathbf{l}^i|} - \frac{\alpha^i|\mathbf{l}^i|s}{2} - \frac{\alpha|\mathbf{l}^i|\delta}{2} \quad ; \quad \widetilde{k}^i := k + \frac{\alpha^i\mathbf{u}\cdot\mathbf{l}^i}{2} \quad ; \quad \widetilde{s} := s + \delta$$

$$\widetilde{\gamma}^i := \frac{|\widetilde{u}^i||\mathbf{l}^i|}{2\widetilde{k}^i} \quad ; \quad \widetilde{\sigma}^i := \frac{\widetilde{s}|\mathbf{l}^i|}{|\widetilde{u}^i|} \quad ; \quad \widetilde{\omega}^i := \frac{\widetilde{s}|\mathbf{l}^i|^2}{\widetilde{k}^i}$$

$$\boxed{\beta^i := \max\left\{\left[\frac{2}{3}\left(\frac{\widetilde{\sigma}^i+3}{\widetilde{\sigma}^i+2}\right) - \left(\frac{4}{\widetilde{\omega}^i+4\widetilde{\gamma}^i}\right)\right], \left[\frac{2}{3} - \frac{4}{\widehat{\omega}^i}\right], 0\right\}}$$

$$\mathbf{h} := \frac{1}{2}\alpha^i\mathbf{l}^i \quad ; \quad \mathbf{H} := \frac{1}{2}\frac{\beta^i}{|\mathbf{l}^i|}[\mathbf{l}^i\otimes\mathbf{l}^i]$$

Box 1: Summary of the HRPG method in multi dimensions and considering the implicit trapezoidal rule for time integration



(a) (b)

Figure 44: Unstructured 20×20 meshes made of bilinear block finite elements. (a) Type I: all internal nodes of the mesh are perturbed using $\delta = 0.2$. (b) Type II: the perturbation perpendicular to the boundary was set to zero for the boundary-adjacent nodes of the mesh. For the rest of the cases $\delta = 0.2$ was chosen.

solved using four meshes, two of which are structured and the remaining two are unstructured. The structured meshes consists of 20×20 (uniform/square) and 40×20 (rectangular) bilinear elements respectively. The unstructured meshes are obtained from the considered uniform mesh via the two perturbation techniques described earlier and illustrated in Figure 44. The obtained solutions are illustrated as surface plots whose view is described as $(\theta^\circ, \psi^\circ)$, where θ° is the azimuthal angle with respect to the negative y-axis and ψ° is the elevation angle from the x–y plane.

Example 1: This is a classical steady-state problem introduced in [22] where the advection is skew to the mesh with downwind essential boundary conditions. The problem data is: $\mathbf{u} = (5, -9)$, $k = 10^{-8}$, $s = 0$ and $f = 0$. The boundary conditions are: $\phi = 1$ on $(x = 0, y > 0.7) \cup (x < 1, y = 1)$, $\phi = 0.5$ at $(x = 0, y = 0.7)$ and $\phi = 0$ on the rest of the boundary. This problem has exponential boundary layers at the outflow boundary and an internal characteristic layer. Figure 45 illustrates the solutions obtained by the HRPG method viewed at $(20^\circ, 20^\circ)$.

Example 2: This problem was studied in [152] wherein a nonuniform rotational velocity field is employed in a rectangular domain $\Omega := [-1, 1] \times [0, 1]$. Structured meshes of 40×20 (uniform/square) and 80×20 (rectangular) bilinear elements are used. The unstructured meshes are obtained from the uniform mesh via the two perturbation techniques described earlier. The problem data is: $\mathbf{u} = 10^4(y[1 - x^2], -x[1 - y^2])$, $k = 10^{-4}$, $f = 0$, $s = 0$. The boundary conditions are: $\phi = 1$ on $(x < -0.5, y = 0)$, $\phi = 0.5$ at $(x = -0.5, y = 0)$, $\phi = 0$ on $(-0.5 < x \leqslant 0, y = 0) \cup (x = 1, y)$ and on the rest of the boundary the Neumann condition $\mathbf{n} \cdot \nabla\phi = 0$ is imposed. The numerical solution of the HRPG method viewed at $(20^\circ, 20^\circ)$ is shown in Figure 46.

Example 3: This is a plain diffusion–reaction problem. The problem data is: $\mathbf{u} = (0, 0)$, $k = 10^{-8}$, $f = 1$, $s = 1$. The homogeneous boundary condition $\phi = 0$ is imposed

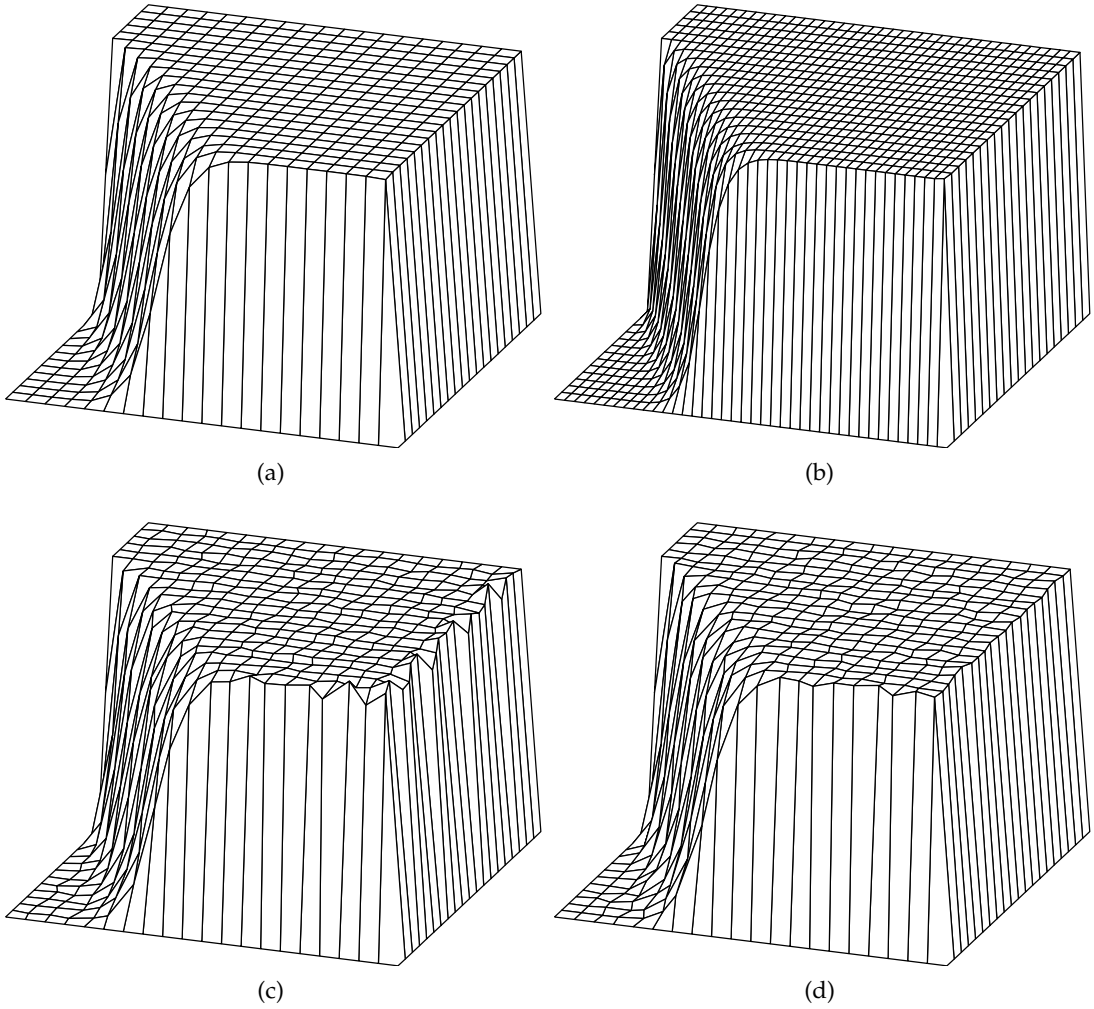


Figure 45: Example 1, advection skew to the mesh. The solution of the HRPG method viewed at $(20°, 20°)$ and using (a) a structured 20×20 mesh, (b) a structured 40×20 mesh, (c) an unstructured (Type I) 20×20 mesh, (d) an unstructured (Type II) 20×20 mesh.

everywhere. The numerical solution of the HRPG method viewed at $(-45°, 20°)$ is shown in Figure 47.

Example 4: This is a multidimensional modification of the convection–diffusion–reaction problem studied earlier in §2.5.7.1 and also in [136, 152]. The problem data is: $\mathbf{u} = (0.01, 0)$, $k = 10^{-4}$, $s = 4.8$ and $f = 0$. The boundary conditions are: $\phi = 1.0$ on $(x = 0, y) \cup (x, y = 0)$, $\phi = (3/8)$ on the rest of the boundary. The value of the element dimensionless numbers γ^1, ω^1 are 2.5 and 120 respectively. Recall that for similar problem data in 1D (cf. §2.5.7.1 and [136, 152]) the upwind numerical artifacts in the solution of Galerkin method were found to be enhanced in the solution of the SUPG method. The numerical solution of the HRPG method viewed at $(120°, 20°)$ is shown in Figure 48.

Example 5: This is a uniform advection problem with a constant source term introduced in [118]. The problem data is: $\mathbf{u} = (1, 0)$, $k = 10^{-8}$, $f = 1$, $s = 0$. The homogeneous boundary condition $\phi = 0$ is imposed everywhere. The exact solution develops

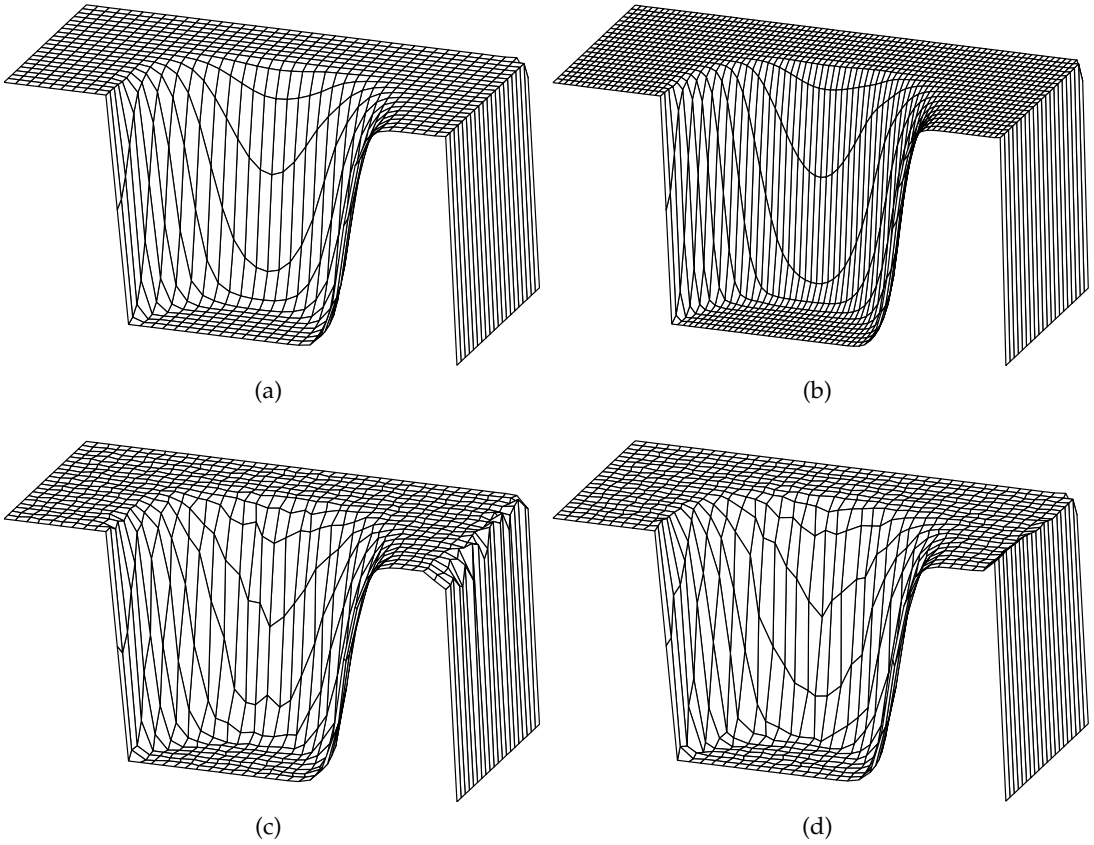


Figure 46: Example 2, nonuniform rotational advection. The solution of the HRPG method viewed at $(20°, 20°)$ and using (a) a structured 40×20 mesh, (b) a structured 80×20 mesh, (c) an unstructured (Type I) 40×20 mesh, (d) an unstructured (Type II) 40×20 mesh.

exponential layers at the outlet boundary $(x = 1, y)$ and characteristic boundary layers at $(x, y = 0)$ and $(x, y = 1)$. The numerical solution of the HRPG method viewed at $(-45°, 20°)$ is shown in Figure 49.

Example 6: This is a non-uniform advection problem with a constant source term introduced in [188]. The advection is caused by a unit angular velocity field. Structured meshes of 64×64 (uniform/square) and 128×64 (rectangular) bilinear elements are used. The unstructured meshes are obtained from the uniform mesh via the two perturbation techniques described earlier. The problem data is: $\mathbf{u} = (y, -x)$, $k = 10^{-6}$, $f = 1$, $s = 0$. The homogeneous boundary condition $\phi = 0$ is imposed everywhere. This problem has a complicated boundary layer. For instance close to the boundary $(x, y = 0)$ and with an increase in x, these layers gradually vary from being parabolic to exponential while maintaining a constant profile height $\phi(x, y = 0) \approx (\pi/2)$ (away from the corners). Close to the boundary $(x = 1, y)$ the solution profile is approximately $\phi(x = 1, y) \approx (\pi/2) - 2\tan^{-1}(y)$. The numerical solution of the HRPG method viewed at $(-200°, 20°)$ is shown in Figure 50.

Example 7: This is a uniform advection problem with a discontinuous source term introduced in [132]. The problem data is: $\mathbf{u} = (1, 0)$, $k = 10^{-8}$, $f(x \leqslant 0.5, y) = 1$, $f(x > 0.5, y) = -1$, $s = 0$. The homogeneous boundary condition $\phi = 0$ is imposed everywhere. Structured meshes of 30×30 (uniform/square) and 60×30 (rectangular)

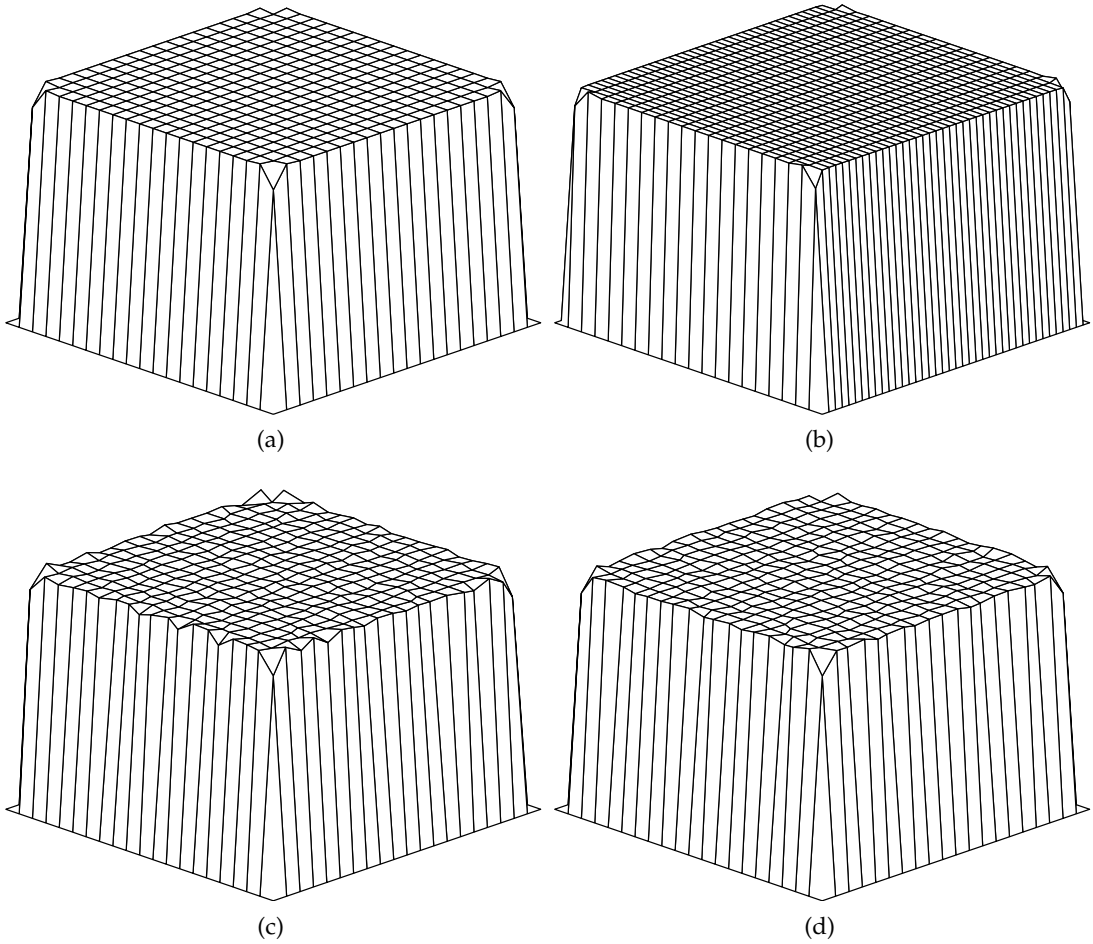


Figure 47: Example 3, a reaction–diffusion problem The solution of the HRPG method viewed at $(-45°, 20°)$ and using (a) a structured 20×20 mesh, (b) a structured 40×20 mesh, (c) an unstructured (Type I) 20×20 mesh, (d) an unstructured (Type II) 20×20 mesh.

bilinear elements are used. The unstructured meshes are obtained from the uniform mesh via the two perturbation techniques described earlier. The numerical solution of the HRPG method viewed at $(-10°, 20°)$ is shown in Figure 51.

3.5.2 *Transient examples*

Here we illustrate the performance of the HRPG method for the transient 2D pure convection problem. Only uniform bilinear finite elements are used here. Both of the examples presented here deal with the advection of solid bodies modeled with appropriate density functions. These problems are frequently used as test cases for advection algorithms demonstrating their treatment of dispersive oscillations and the overall solution accuracy.

Example 8: This is a test case introduced in the ERCOFTAC document [24]. A circular scalar bubble is initially positioned at the bottom of a square domain in a fixed constant velocity field directed at 45° toward the top right of the domain. The problem data is: $\mathbf{u} = (0.5, 0.5)$, $k = 10^{-30}$, $s = 0$ and $f = 0$. The domain $\Omega := [0, 3] \times [0, 3]$ is

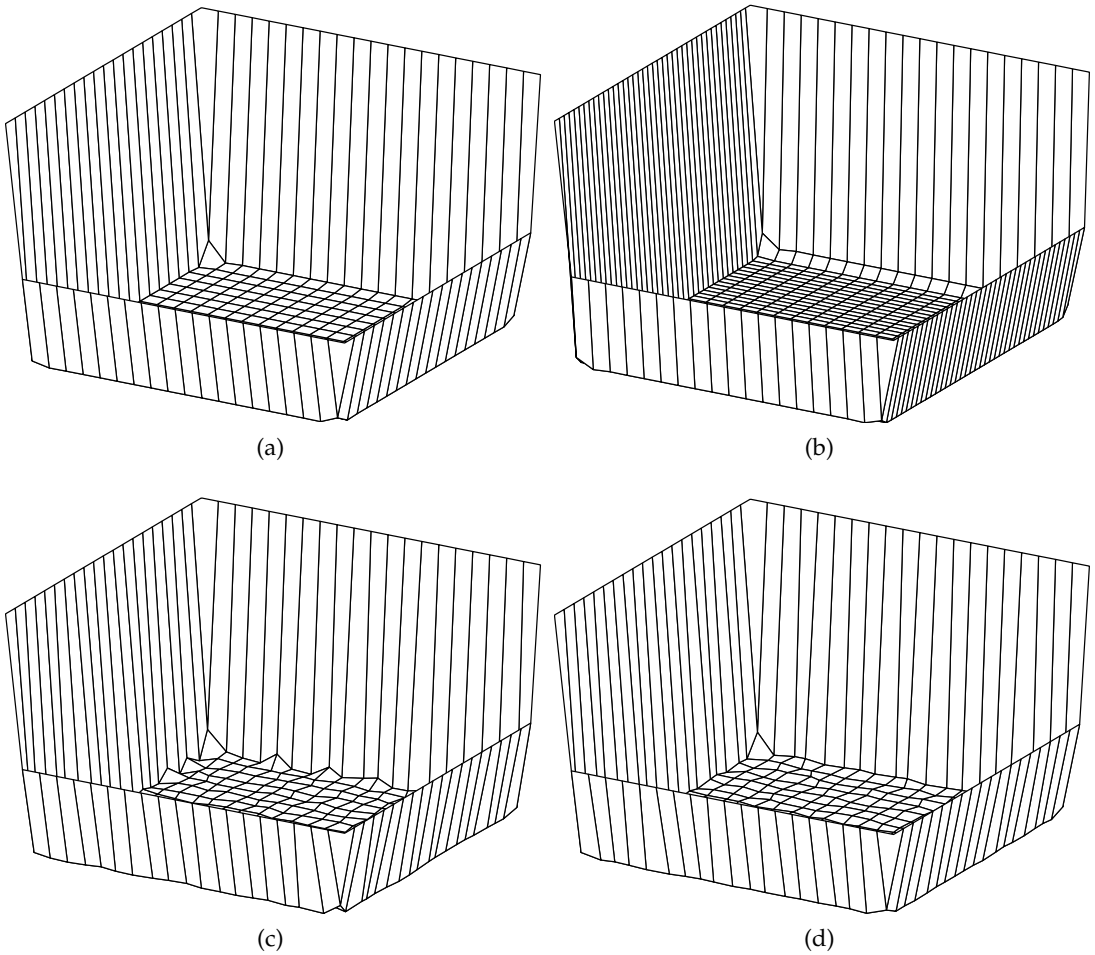


Figure 48: Example 4, a convection–diffusion–reaction problem with a dominant reaction term. The solution of the HRPG method viewed at $(120°, 20°)$ and using (a) a structured 20×20 mesh, (b) a structured 40×20 mesh, (c) an unstructured (Type I) 20×20 mesh, (d) an unstructured (Type II) 20×20 mesh.

discretized by a uniform mesh of 300×300 bilinear elements. The time integration is done using the implicit midpoint rule and is advanced at a time step of 0.005 seconds. This corresponds to an element CFL number of 0.25. Define a radius $R = 0.25$, an arbitrary position vector $\mathbf{r} := (x, y) \in \Omega$ and a specific position vector $\mathbf{r}^c := (0.5, 0.5) \in \Omega$. The initial solution can then be expressed as follows:

$$\phi(\mathbf{r}, t = 0) = H(R - |\mathbf{r} - \mathbf{r}^c|) \tag{3.32}$$

where $H()$ is the Heaviside function defined earlier in Eq.(3.6b) and $\mathbf{r}^c$ is the center of the circular scalar bubble. The initial solution viewed at $(40°, 20°)$ is shown in Figure 52a (elevation plot) and Figure 52c (contour plot). The Dirichlet boundary condition $\phi = 0$ is imposed at the inlet boundaries. The numerical solution of the HRPG method at time $t \in \{1, 2, 3, 4\}$ seconds and viewed at $(40°, 20°)$ is shown in Figure 53 (elevation plots) and Figure 54 (contour plots).

Example 9: This is a standard benchmark problem introduced in [125] that simulates the advection of a solid body subjected to a constant angular velocity field. The solid body is modeled with a scalar density function that has three shapes, viz. a slotted

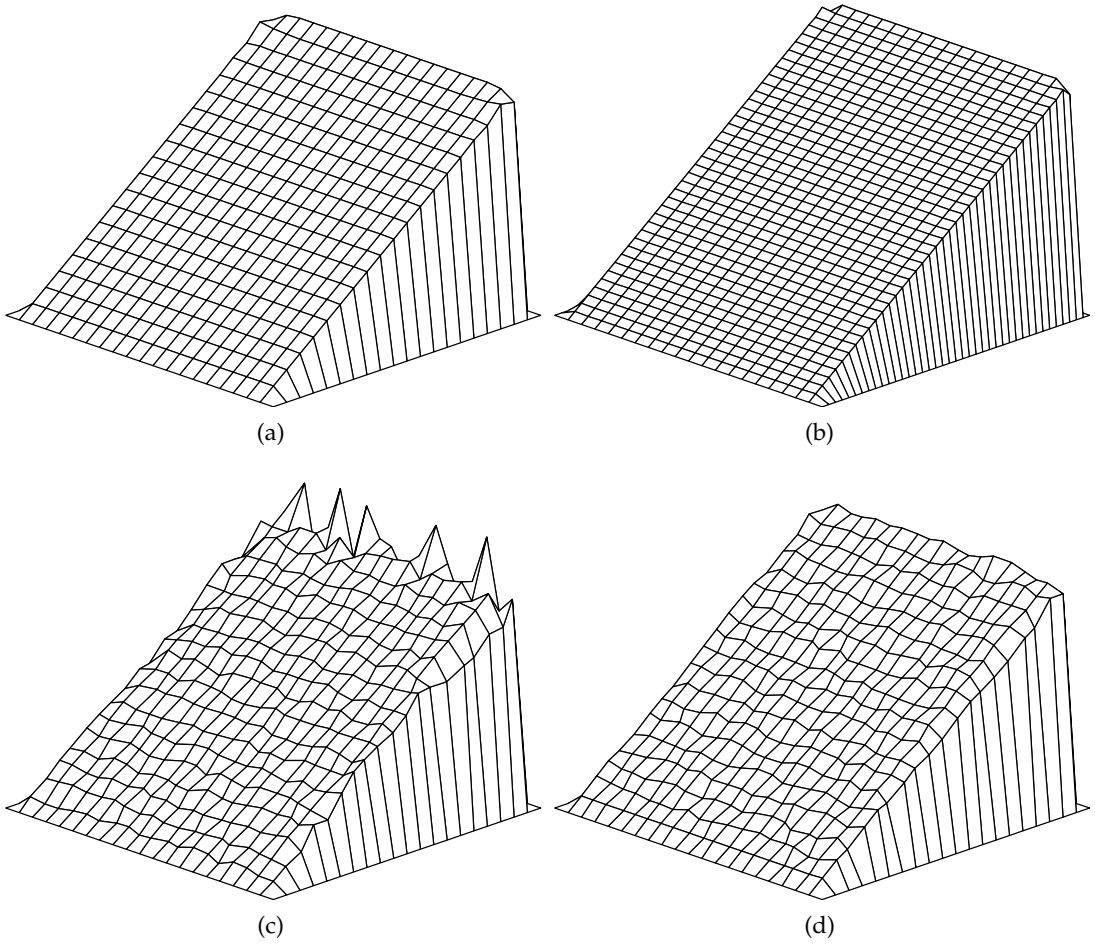


Figure 49: Example 5, uniform advection with a constant source term. The solution of the HRPG method viewed at $(-45°, 20°)$ and using (a) a structured 20×20 mesh, (b) a structured 40×20 mesh, (c) an unstructured (Type I) 20×20 mesh, (d) an unstructured (Type II) 20×20 mesh.

cylinder, a cone and a sinusoidal hump. The classical problem with just the slotted cylinder revolving about the center of the domain was proposed by Zalesak in the seminal paper [191] that extended the FCT method to multi dimensions. The problem data is: $\mathbf{u} = (0.5 - y, x - 0.5)$, $k = 10^{-30}$, $s = 0$ and $f = 0$. The domain $\Omega := [0,1] \times [0,1]$ is discretized using 200×200 uniform bilinear elements. The time integration is done using the implicit midpoint rule and is advanced at a time step of 0.001 seconds. This corresponds to a maximum element CFL number of 0.1. Define a radius $R = 0.15$, an arbitrary position vector $\mathbf{r} := (x, y) \in \Omega$ and a specific position vector $\mathbf{r}^a := (x^a, y^a) \in \Omega$ for some chosen point a. The initial solution can then be expressed as follows:

$$\begin{aligned}\phi(\mathbf{r}, t = 0) = H(R - |\mathbf{r} - \mathbf{r}^1|)\left[1 - H(0.025 - |x - x^1|)\, H(0.85 - y)\right] + \\ 1 - \min\left\{\frac{|\mathbf{r} - \mathbf{r}^2|}{R}, 1\right\} + \frac{1}{4}\left[1 + \cos\left(\pi \min\left\{\frac{|\mathbf{r} - \mathbf{r}^2|}{R}, 1\right\}\right)\right]\end{aligned} \tag{3.33}$$

where $H()$ is the Heaviside function defined earlier in Eq.(3.6b), $\mathbf{r}^1 = (0.5, 0.75)$, $\mathbf{r}^2 = (0.5, 0.25)$ and $\mathbf{r}^3 = (0.25, 0.5)$ are the position vectors corresponding to the center of the slotted cylinder, the cone and the sinusoidal hump respectively. The initial

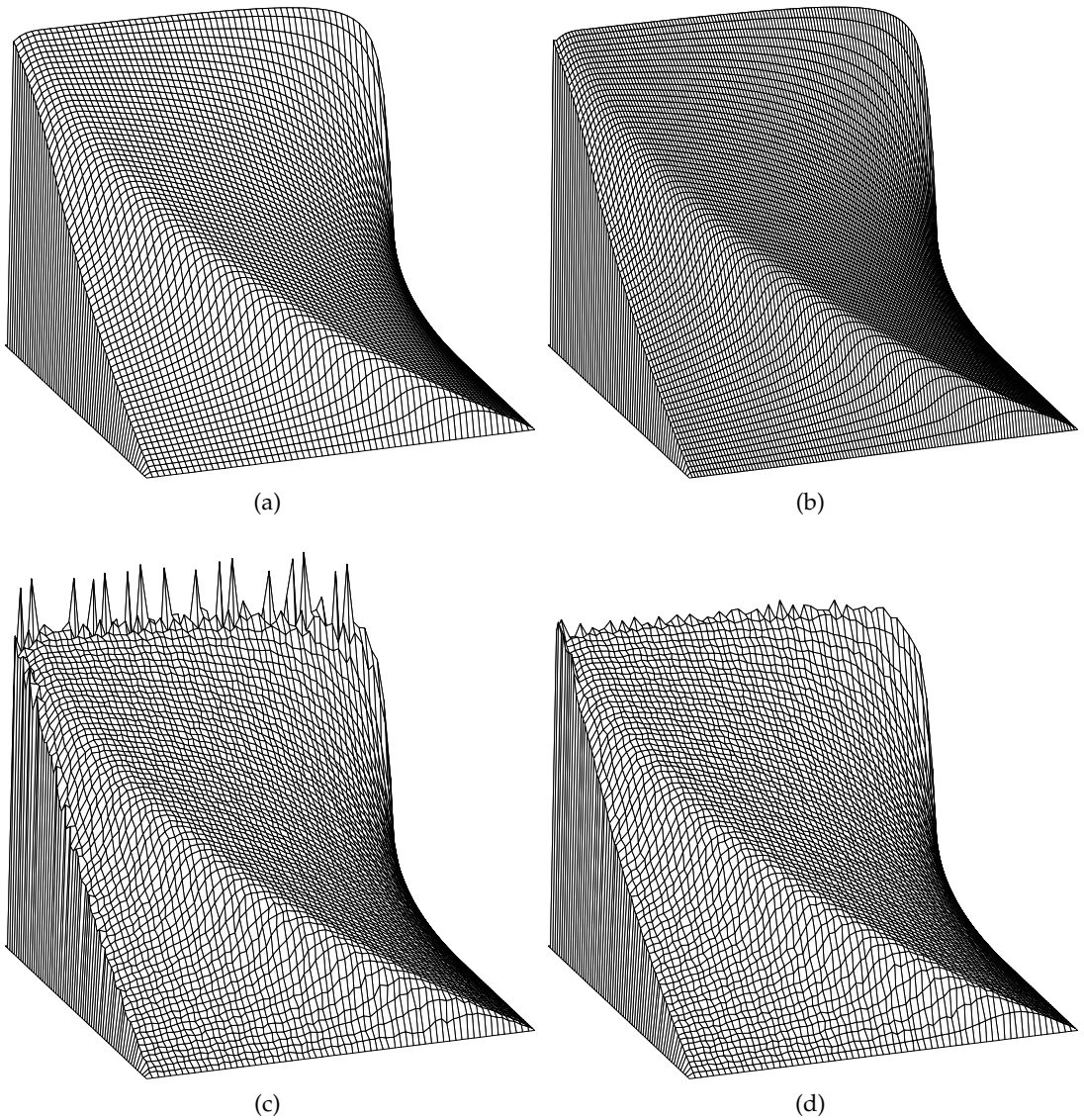


Figure 50: Example 6, non-uniform advection with a constant source term. The solution of the HRPG method viewed at $(-200°, 20°)$ and using (a) a structured 64×64 mesh, (b) a structured 128×64 mesh, (c) an unstructured (Type I) 64×64 mesh, (d) an unstructured (Type II) 64×64 mesh.

solution viewed at $(-20°, 20°)$ is shown in Figure 52b (elevation plot) and Figure 52d (contour plot). The Dirichlet boundary condition $\phi = 0$ is imposed at the inlet boundaries. Under the considered velocity field the initial solution completes a full revolution in 2π seconds. The numerical solution of the HRPG method at time $t = \{(\pi/2), \pi, (3\pi/2), 2\pi\}$ seconds and viewed at $(-20°, 20°)$ is shown in Figure 55 (elevation plots) and Figure 56 (contour plots).

3.5.3 *Discussion*

The HRPG method proposed here can be understood as the combination of upwinding plus a discontinuity-capturing operator. Also the discontinuity-capturing

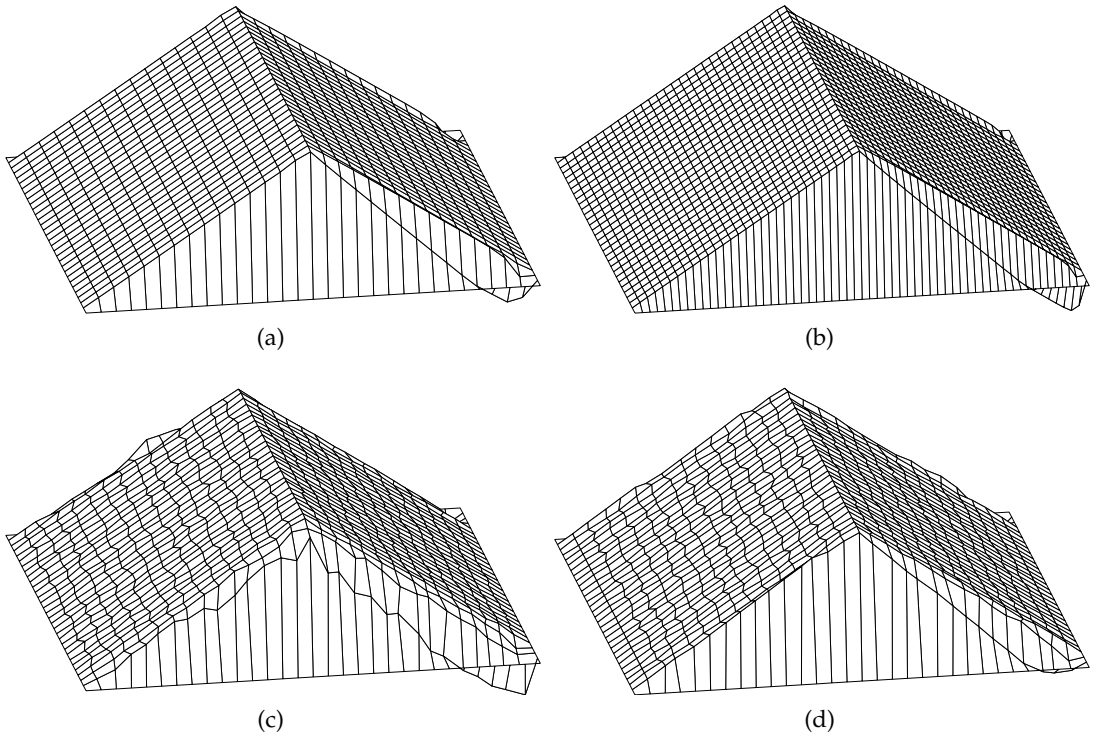


Figure 51: Example 7, uniform advection with a discontinuous source term. The solution of the HRPG method viewed at $(-10°, 20°)$ and using (a) a structured 30×30 mesh, (b) a structured 60×30 mesh, (c) an unstructured (Type I) 30×30 mesh, (d) an unstructured (Type II) 30×30 mesh.

term has the canonical form of the shock-capturing diffusion, i. e.it is proportional to $(|R(\phi_h)|/|\nabla\phi_h|)$. Nevertheless the finer structure of the HRPG method is distinct from the existing shock-capturing Petrov–Galerkin methods in the literature (cf. Table 2 in chapter 2). The distinction is that the upwinding provided by the characteristic tensor $\mathbf{h}$ is not streamline and the discontinuity capturing provided by the characteristic tensor $\mathbf{H}$ is neither isotropic nor purely crosswind.

It is clearly seen from the steady-state examples presented in the previous section that for structured meshes (both square and rectangular bilinear elements) the HRPG method reproduces a crisp resolution of the layers in the numerical solution. The good performance on rectangular elements (here considered with an aspect ratio of 2:1) is due to the anisotropic treatment of the stabilization terms involving the characteristic tensors $\mathbf{h}$ and $\mathbf{H}$. The solutions obtained by the HRPG method for the transient 2D advection examples advocate its good treatment of dispersive oscillations without compromising the solution-accuracy (cf. Figures 53 and 55). Also the symmetry of the initial data is well maintained (cf. Figures 54 and 56). Recall that the time integration was performed by the implicit midpoint rule which is a symplectic time integrator [76]. This choice was made to single-out the treatment of the geometrical symmetry in the initial data by the HRPG method.

Clearly on unstructured meshes we do not attain the same layer resolution quality as is obtained on the corresponding structured meshes. However the parabolic layers (characteristic and reactive layers) are captured satisfactorily. About the exponen-

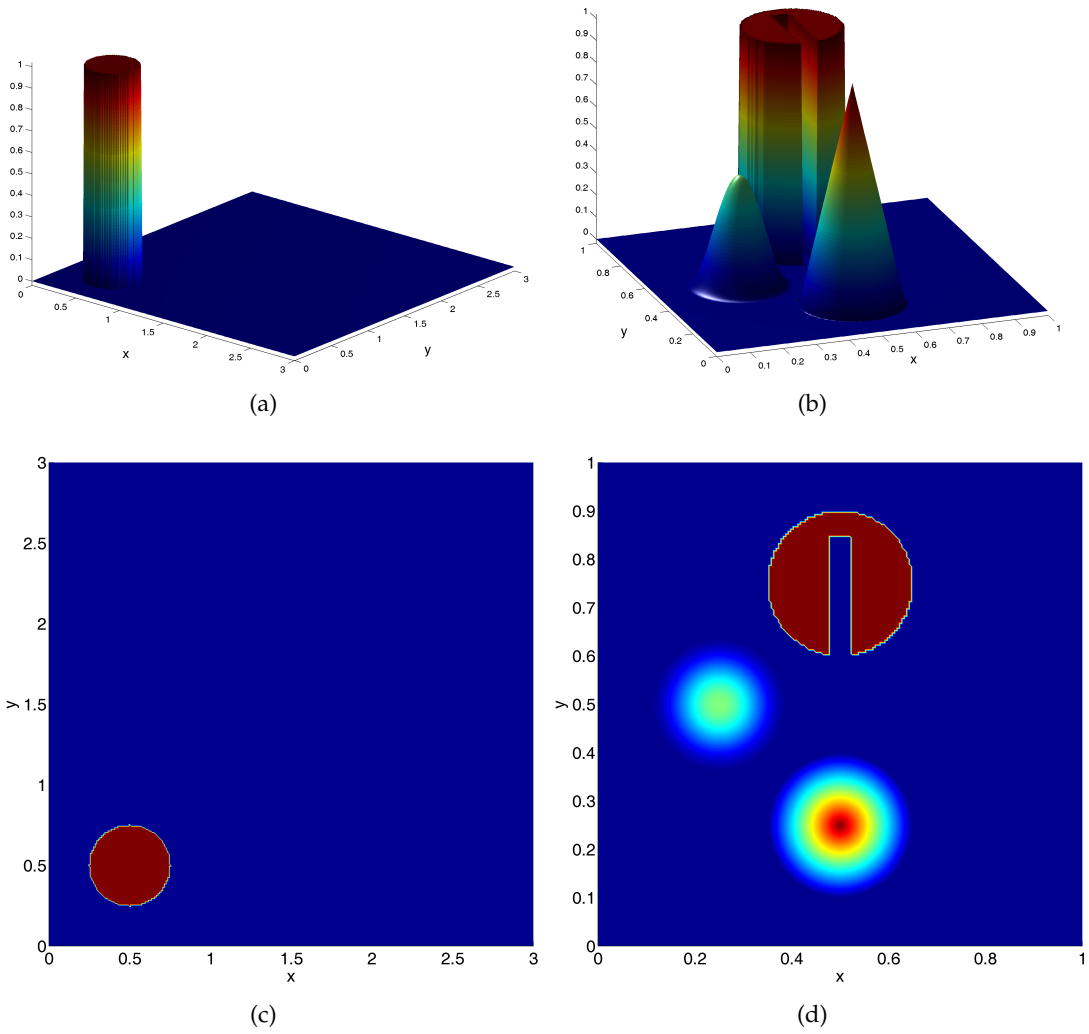


Figure 52: Initial data for the transient 2D advection examples. (a) Example 8, elevation plot viewed at $(40°, 20°)$, (b) Example 9, elevation plot viewed at $(-20°, 20°)$, (c) Example 8, contour plot, (d) Example 9, contour plot.

tial layers some overshoots and undershoots are observed using Type I unstructured meshes. These unwanted localized artifacts are conspicuous in the solutions of example 5 (Figure 49c) and example 6 (Figure 50c) suggesting that there is room for further improvement of the method. Nevertheless using Type II unstructured meshes where in the random perturbation of the mesh nodes perpendicular to the domain boundary is set to zero, these unwanted artifacts about the exponential layers are greatly reduced.

Figure 51 illustrates another shortcoming of the HRPG method that is conspicuous even when structured meshes are used. On one half of the domain (here the source term is positive) the obtained solutions have crisp layer resolutions, whereas in the remaining half (here the source term is negative) the numerical solution appears to be over-damped and even negative near the corners of the outlet boundary. This is a shortcoming suffered by all the shock-capturing techniques designed within the Petrov–Galekin framework (see Codina's monograph [31]) that rely on the canoni-

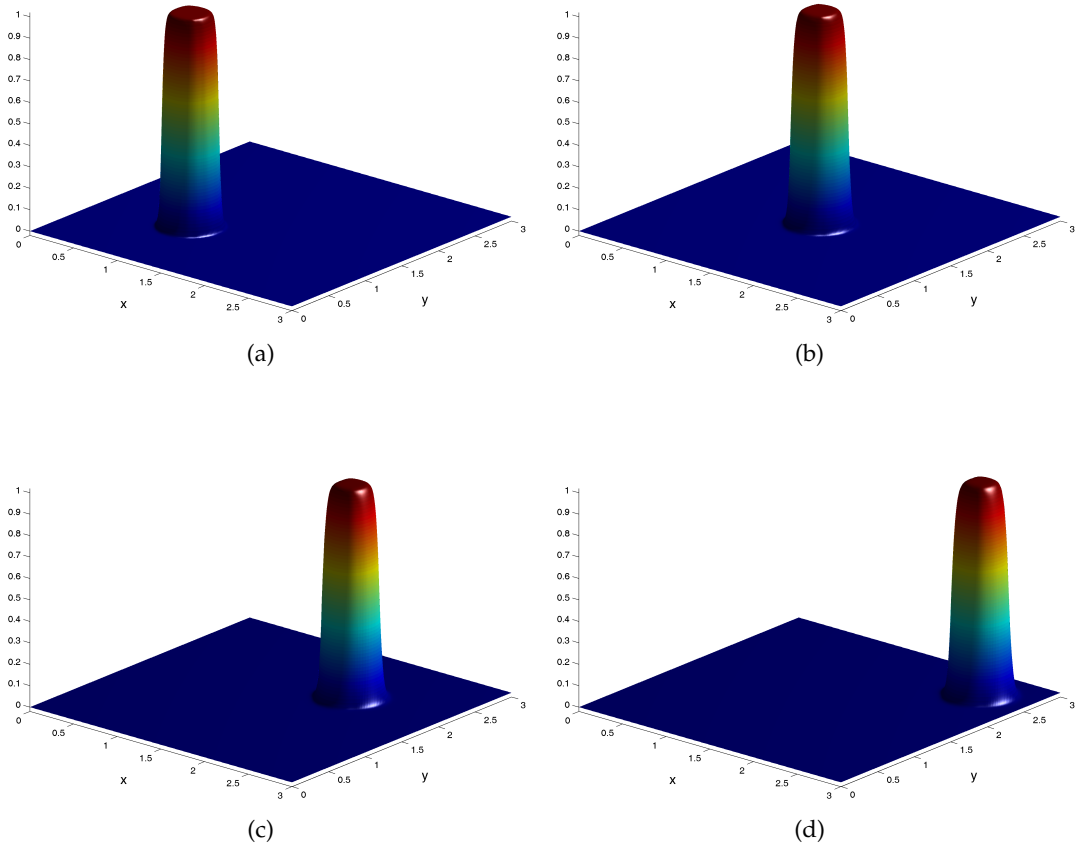


Figure 53: Example 8, transient pure convection skew to the mesh. The solution of the HRPG method viewed at $(40°, 20°)$ and at time (a) $t = 1$ s, (b) $t = 2$ s, (c) $t = 3$ s, (d) $t = 4$ s.

cal strategy of adding a positive shock-capturing diffusion. The following example illustrates why the aforesaid strategy fails to address this shortcoming.

Example 10: Consider a unit domain $\Omega := [0,1] \times [0,1]$ and the following problem data: $\mathbf{u} = (1,0)$, $k = 10^{-8}$, $s = 0$ and $f = -1$. The Dirichlet boundary conditions are: $\phi = 1$ on $(x = 0, y > 0) \cup (x, y = 1)$ and $\phi = 0$ on the rest of the boundary. The domain Ω is discretized using a structured mesh of 20×20 (uniform/square) bilinear elements. In the interior of the domain the exact solution has the profile of a flat surface with a slope of -1. Along the boundaries $(x, y = 0)$ and $(x, y = 1)$ the exact solution develops characteristic boundary layers and as a consequence within the width of these characteristic layers and near the corners of the outlet boundary $(x = 1, y)$, exponential layers are formed. Hence the solution of the plain Galekin FEM will be corrupted with global oscillations. The solutions obtained by the SUPG and the HRPG method are shown in Figure 57.

Note that the undershoots and overshoots in the solution of the SUPG method is identical across both characteristic layers (cf. Figures 57a and 57c). This is in agreement with the reasoning made in §3.1 related to the numerical artifacts across characteristic layers, i. e.unlike in the reaction-dominant case where it is the numerical solution

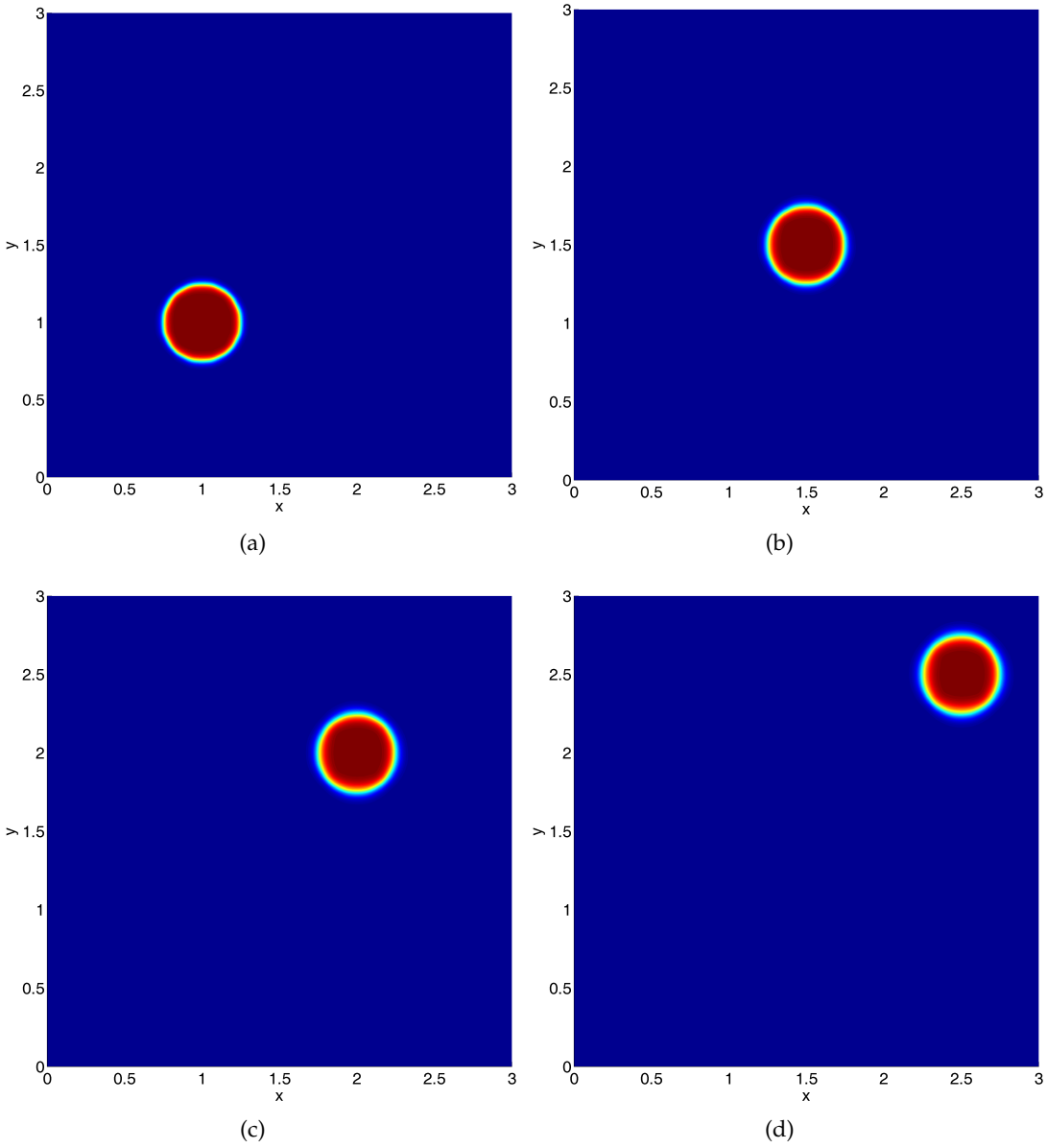


Figure 54: Example 8, transient pure convection skew to the mesh. The contour plots of the solution of the HRPG method at time (a) $t = 1$ s, (b) $t = 2$ s, (c) $t = 3$ s, (d) $t = 4$ s.

that undergoes the 1D mass type averaging, in the convection-dominant case it is the derivatives of the numerical solution that undergoes the same. Thus, the Gibbs phenomenon across the characteristic layers in the later case is proportional to the variation in the derivatives of the solution across these layers. In other words for the current problem it is the slope of ϕ_h and not the actual value of ϕ_h on the boundary that determines the observed artifacts. It can be clearly seen in Figure 57c that any method, that relies on the canonical strategy of adding a *positive* shock-capturing diffusion, will not be able to recover (near the boundary $(x, y = 0)$) the nodally exact interpolant from the initial SUPG solution. On the other hand, note that the artifacts near the boundary $(x, y = 1)$ has a profile similar to the one that would have been

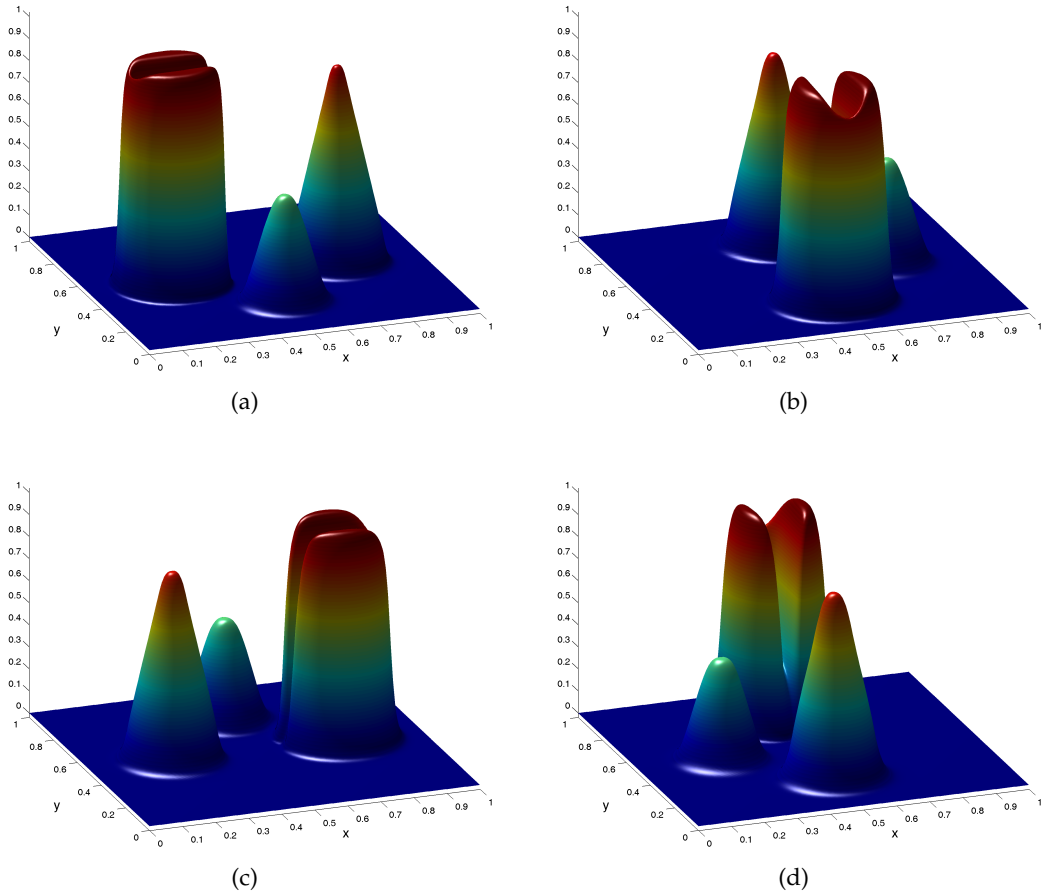


Figure 55: Example 9, rotation of solid bodies. The solution of the HRPG method viewed at $(-20°, 20°)$ and at time (a) $t = (\pi/2)$ s i. e.after a quarter-revolution, (b) $t = \pi$ s i. e.after a half-revolution, (c) $t = (3\pi/2)$ s, i. e.after three quarters of a revolution (d) $t = 2\pi$ s i. e.after a full-revolution.

observed for the L^2 projection of the exact solution onto the finite element space. It is for this reason that the aforesaid strategy succeeds in capturing these layers.

Obviously tailor-made solutions exist to treat this shortcoming. For instance, one such trick that recovers crisp resolution of these layers for the HRPG method and for the current problem (example 10) is to reverse the sign of the stabilization parameter β (along the y-axis) for all elements containing the boundary section $(x > 0.5, y = 0)$, thus enforcing a *negative* shock-capturing diffusion for these elements. Unfortunately it is difficult to generalize these tailor-made tricks to an arbitrary situation. An alternative would be to change the strategy to the one which directly treats the cause of the Gibbs phenomenon for both the reactive and characteristic layers[4]—Design the weights of a Petrov–Galerkin FEM such that the typical 1D mass type averaging in the Galerkin FEM (cf. Eq.(3.4)) be lumped in the regions across the layers. Research in this line is still under development and we delay its introduction to future works.

Remark: Fortunately, this idea which was born to treat this shortcoming in the convection–diffusion–reaction problem, has opened door to a class of higher-order

4 this idea is a fruit of the discussions with Prof. Ramon Codina

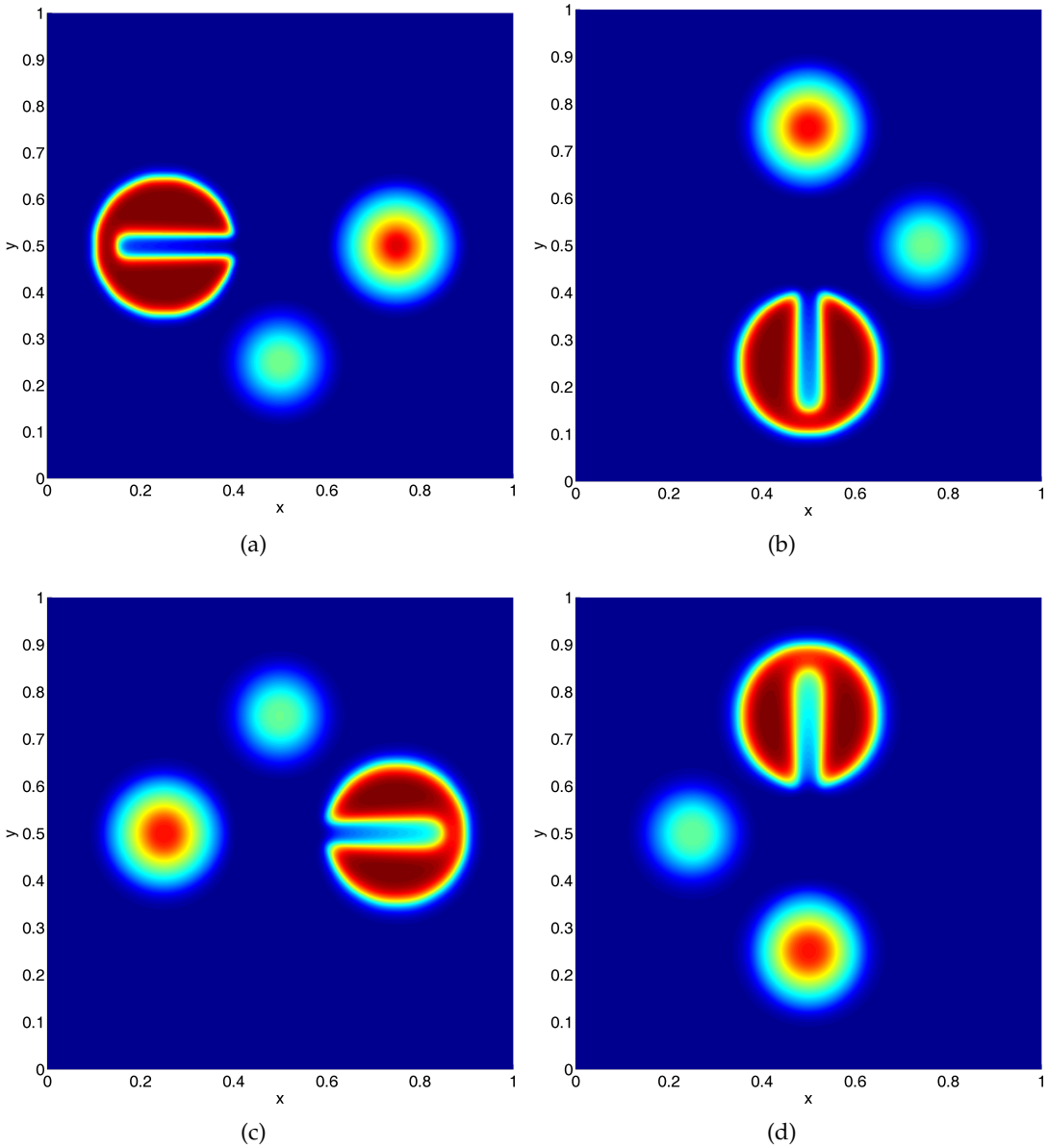


Figure 56: Example 9, rotation of solid bodies. The contour plots of the solution of the HRPG method at time (a) $t = (\pi/2)$ s i. e.after a quarter-revolution, (b) $t = \pi$ s i. e.after a half-revolution, (c) $t = (3\pi/2)$ s, i. e.after three quarters of a revolution, (d) $t = 2\pi$ s i. e.after a full-revolution.

compact Petrov–Galerkin FEM effective for the Helmholtz problem. The design of such a Petrov–Galerkin FEM and its applications to the Helmholtz equation is the subject matter of chapter 5.

3.6 CONCLUSIONS

We have developed a multi dimensional extension of the HRPG method presented earlier in chapter 2 for the 1D convection–diffusion–reaction problem. As the characteristic internal/boundary layers found in the convection-dominant case are a unique

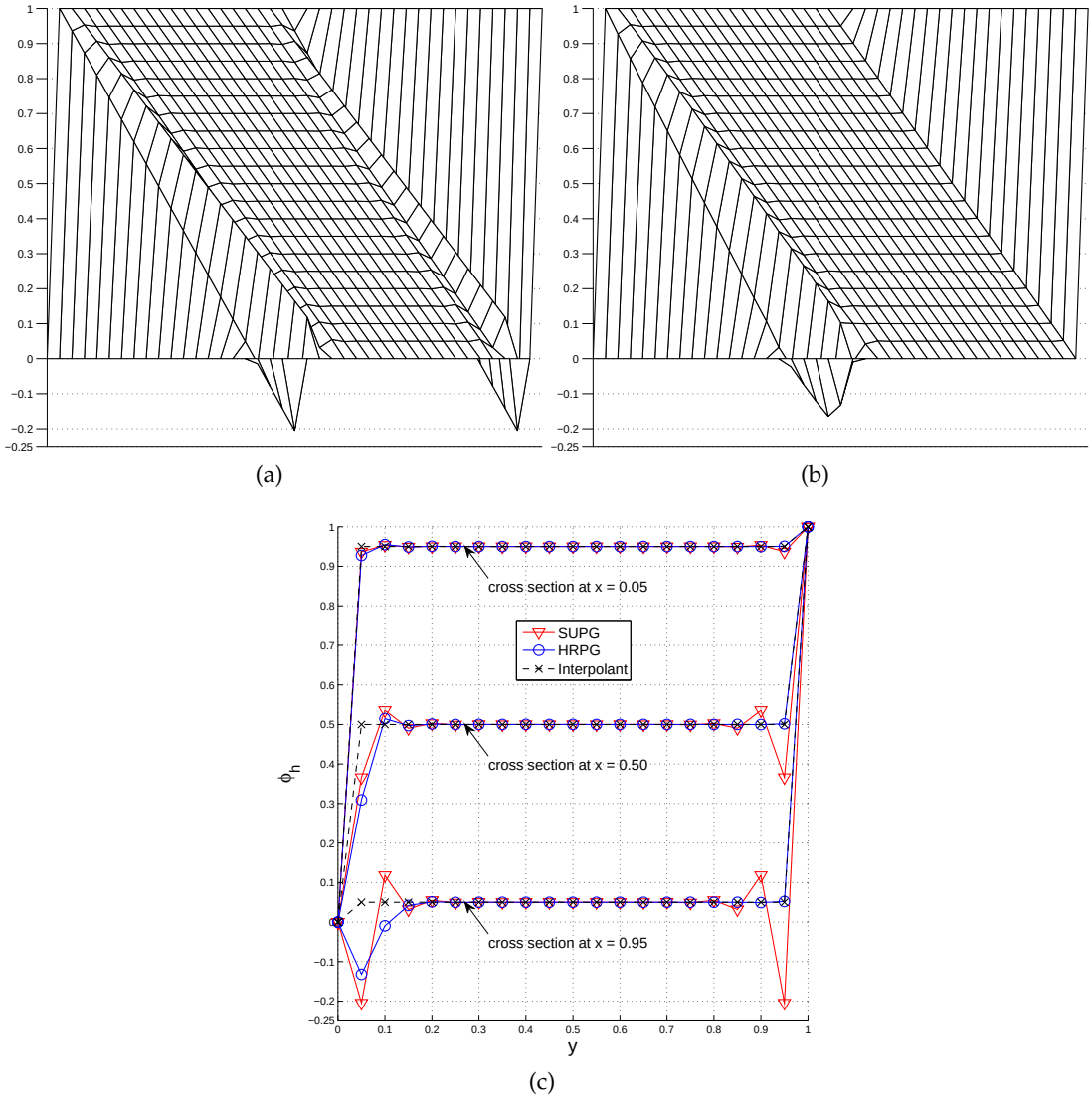


Figure 57: Example 10, uniform advection with a negative source term. The solution obtained on a uniform 20×20 mesh viewed at $(45°, 0°)$ and using (a) the SUPG method, (b) the HRPG method. (c) Comparison of the nodally exact interpolant at three different cross sections with the numerical solution obtained by the SUPG and the HRPG method.

feature of the solution in higher dimensions, they do not have any counterparts in 1D. Hence, a straight-forward extension of the stabilization parameters of the HRPG method derived for the 1D case will not be efficient to resolve these parabolic layers.

The numerical artifacts that are formed across the parabolic layers are usually manifested as the Gibbs phenomenon. The strategy we employ to treat the artifacts about the characteristic layers is to treat them just like the artifacts found across the parabolic layers in the reaction-dominant case. This is done by relating the characteristic layers in the convection–diffusion problem to the parabolic layers formed in a fictitious diffusion–reaction problem. The fictitious reaction coefficient in the later problem is designed such that the parabolic layers in both the problems have the same width. Using this fictitious reaction coefficient, we present a nondimensional element num-

ber that quantifies these characteristic layers. By quantification we mean that it should serve a similar purpose in the definition of the stabilization parameters as the element Peclet number does for the exponential layers.

Although the structure of HRPG method in 1D is identical to the consistent approximate upwind Petrov–Galerkin method [68], in multi dimensions the former method has a unique structure. The distinction is that in general the upwinding is not streamline and the discontinuity-capturing is neither isotropic nor purely crosswind. In this line, we present anisotropic element length vectors $\mathbf{l}^i$ and using them objective characteristic tensors associated with the HRPG method are defined. Only the multilinear block finite elements are considered in this study. Except for the modification to include the new dimensionless number that quantifies the characteristic layers, the definition of the stabilization parameters α^i, β^i calculated along the element length vectors $\mathbf{l}^i$ are a direct extension of their counterparts in 1D summarized earlier in §2.5.6.

Finally, several steady-state and transient examples are presented that throws light on the good performance of the proposed method.

Part II

HELMHOLTZ PROBLEM

Most of what we learn, we learn indirectly or by 'head-fake'.

— Randy Pausch.

4

ALPHA-INTERPOLATION OF FEM AND FDM

4.1 INTRODUCTION

In this chapter we study the Helmholtz equation given by $R(\phi) := f(\mathbf{x}) + \Delta\phi + \xi^2\phi = 0$ and subjected to Dirichlet boundary conditions. The solution ϕ to this equation is oscillatory and ξ is the wave number (spatial frequency) of ϕ. If λ_m is an eigenvalue of the operator $-\Delta$, then for $\xi \neq \sqrt{\lambda_m}$ the problem has a unique solution. On the contrary, i. e.for $\xi = \sqrt{\lambda_m}$ the problem is indefinite. In this case if the equation and the Dirichlet boundary conditions are homogeneous then we end up in a differential eigenvalue problem. It follows that the solution is not unique and can be represented as a scalar multiple of the eigenfunction corresponding to each eigenvalue. Let λ_m^h represent an eigenvalue of the problem after any appropriate discretization. Unlike the set of eigenvalues $\{\lambda_m\}$ which is infinite, the set $\{\lambda_m^h\}$ is finite and its dimension is equal to that of the discrete space. Thus when the wave number $\xi \to \sqrt{\lambda_m^h}$ the discrete problem tends to be indefinite. This case is usually referred to as the case of degeneracy and here the discrete problem is ill-conditioned.

As the current problem admits a variational principle, naturally, discretization methods based on variational formulations viz. the Galerkin and the Trefftz–Galerkin type methods have been preferred to other methods. The Galerkin type methods are domain-based wherein the integral statement involves only the weak form of the governing differential equation and the sub-space of test-functions are assumed to satisfy a priori the kinematic compatibility and essential boundary conditions. The Trefftz–Galerkin type methods are boundary-based and are formulated using the reciprocal principle wherein the integral statement involves only the kinematic compatibility and essential boundary conditions of the problem and the sub-space of test-functions are assumed to satisfy a priori the governing differential equation [27, 41].

In the context of the Galerkin type methods, the finite element method (FEM) is a powerful technique to systematically generate subspaces of test-functions (classically piecewise polynomial spaces). Some of the earlier works on the use of FEM for the numerical solution of the Helmholtz equation can be found in [5, 6, 9, 47, 107, 108, 130] and the references cited therein. In [5, 107] error estimates were given for the asymptotic ($\xi^2\ell$ assumed sufficiently small) and pre-asymptotic ($\xi\ell$ assumed sufficiently small) cases respectively. It was shown that for the discrete problem the LBB[1] constant can be expressed as $\gamma^h = \min\{|\lambda_m^h - \xi^2|/\lambda_m^h\}$ [47]. Thus for the continuous problem (visualized as $\ell \to 0$) the LBB constant can be expressed as $\gamma = \min\{|\lambda_m - \xi^2|/\lambda_m\}$ which in an average sense implies that γ is inversely proportional to the wavenumber ξ, i. e.$\gamma \propto \xi^{-1}$ [47, 107]. Thus for high wavenumbers and for the case of degeneracy ($\xi \to \sqrt{\lambda_m^h}$) the LBB constant for the discrete problem tends to be small which in

1 *Ladyzhenskaya-Babuska-Brezzi* constant

turn leads to a loss of stability. The loss of stability with respect to an increase in the wavenumber ξ is called the pollution effect which is impossible to avoid completely [9]. Nevertheless the pollution effect can be controlled unlike the loss of stability for the case of degeneracy where it is out of control.

Several stabilization methods were developed to control the pollution effect of the Galerkin FEM. The Galerkin least squares (GLS) method was extended to the Helmholtz equation in [79, 181]. In [79] the extension of the Galerkin gradient least squares (GGLS) method for the current problem was also studied. In order to retain stability for problems that involve the physics of both the convection–diffusion–reaction and Helmholtz equations the GLSGLS method was proposed [80]. Following the framework of the Generalized Finite Element Methods (GFEM) which were first introduced in [8] in a variational setting, the Quasi-Stabilized FEM (QSFEM) was proposed in [10]. To be precise, within an algebraic setting a 9-node interior stencil was designed such that the pollution effect is asymptotically minimal, thus leading to minimal phase error for arbitrary wave direction in 2D. The partition of unity method (PUM) was proposed in [7, 131] by which conforming subspaces of higher regularity can be generated out of a set of local approximation spaces. These local approximation spaces could be designed to include a priori knowledge about the local behavior of the solution. Recently, following the framework of PUM, a locally enriched FEM was proposed in [121] wherein it was shown that the Bessel functions of the first kind could be used to enrich the finite element space instead of the plane waves (as is done in PUM). Another stabilization approach consists of enriching the classical finite element spaces by bubble functions. Following this line the residual-free bubbles (RFB) method was extended to the Helmholtz equation in [67]. Another bubble-based method is the nearly optimal Petrov–Galerkin method (NOPG) presented in [13]. A comparison of the RFB and NOPG methods for the Helmholtz equation was done in [78]. Recently another GFEM was proposed in [173] in which the classical FEM is enriched by plane waves pasted into the finite element basis at each mesh vertex by the PUM. Also, this method allows the use of Cartesian meshes which may overlap the boundaries of the problem domain. This GFEM was further developed in [174] wherein the effects of using alternative handbook functions and mesh types is addressed. Based on the variational multiscale (VMS) method several stabilization methods were proposed, viz. the sub-grid FEM [159], the two sub-grid scale (SGS) models presented in [28], the residual-based FEM (RBFEM) [160] and more recently, the algebraic subgrid FEM (ASGS) [74] and the SGS-GSGS method [86]. Following the more general VMS method wherein the subscales are not modeled as bubbles, the RBFEM method also includes the residuals on the inter-element boundaries while retaining the sparsity of the Galerkin method. As in the GLSGLS method, the SGS-GSGS method attempts to stabilize the advection–diffusion–reaction/production problem and is designed to be nodally exact in 1D. Within the framework of the discontinuous Galerkin (DG) method, the discontinuous enrichment method (DEM) was proposed [57, 58] wherein the classical finite element spaces are enriched (as in bubble-based methods) via a set of local approximation spaces (as in the PUM-based methods). In the DEM, the continuity of the enrichment across element boundaries is enforced weakly by Lagrange multipliers (unlike the PUM-based methods) and it need not vanish at the element boundaries (unlike the bubble-based methods). Another DG method is presented in [3] wherein the continuity of the finite element spaces across

the element edges is relaxed and weakly enforced via two penalty parameters corresponding to possible jumps of the solution field and its gradient. These penalty parameters are designed to minimize the pollution error. Following the ideas of the former DG method [3], another discontinuous FEM was proposed in [129] (therein called as the DGB method). In the DGB method the classical finite element spaces are enriched via bubbles that are allowed to be discontinuous across subgrid patches. Following the DG method in [3] the continuity of the bubble spaces across interior patch boundaries is enforced weakly via two penalty parameters corresponding to possible jumps of the solution field and its gradient. Again, these penalty parameters are designed to minimize the pollution error. Nodally exact Ritz discretization of the 1D diffusion-absorption/production equations via variational finite calculus (FIC) and modified equation methods using a single stabilization parameter were presented in [59]. The Galerkin projected residual (GPR) method for the Helmholtz equation was presented in [51]. A survey of finite element methods for time-harmonic acoustics is done in [77].

Due to the abstractness in the definition of the QSFEM, it is often labeled as a finite difference method. Nevertheless, it provides solutions that are sixth-order accurate, i. e.O $\left((\xi\ell)^6\right)$ which is the best one can get on any compact stencil. Recently, a quasi-optimal Petrov–Galerkin (QOPG) method using bilinear finite elements was proposed in [128] that recovers the QSFEM stencil on square meshes. In the QOPG method the Galerkin FEM weights are perturbed by a quadratic bubble function defined over the macro-element. The parameters multiplying the bubble perturbations are found by solving local optimization problems involving a functional of the local truncation error. Later, following this line, a quasi-optimal finite difference method on generic unstructured meshes was proposed in [60].

Within the framework of the finite difference methods, several fourth-order compact schemes obtained through a generalization of the fourth-order Padé approximation were studied in [81, 168]. Following this line, two new FDMs were proposed in [165] that achieves sixth and eight-order accuracy respectively using a five-point (and hence non-compact) stencil in 1D. In [122] a new FDM with improved accuracy was proposed by modifying the central difference scheme (i. e.the classical FDM) by replacing the weight multiplying the central node with an optimal expression that used the Bessel's function of the first kind. The FLAME method was proposed in [184] that exploits the use of local approximating functions to define higher-order finite difference schemes on a chosen stencil. In particular on a compact stencil a sixth-order accurate scheme for the Helmholtz equation can be derived using the FLAME method. Sixth-order accurate FD schemes on a compact stencil for the Helmholtz equation were proposed in [135, 169, 176]. An alternate approach to derive FDMs is the global method of differential quadrature (DQ) [15]. Following this line, a polynomial-based DQ and a Fourier expansion-based DQ were derived for the Helmholtz equation in [166]. As higher-order polynomial or sinusoidal interpolation functions are employed, these methods reduce the restriction on the mesh resolution to the Nyquist limits, i. e.the rule of thumb for these methods is to provide at least two elements per wavelength.

Some of the earlier works in the context of the Trefftz–Galerkin type methods could be found in the seminal papers [87, 112, 162, 192]. A treatise on Trefftz type methods can be found in [88, 113]. Specifically, the Trefftz type methods were used for the Helmholtz equation in [25, 69, 126, 133, 170]. The case of degeneracy, i. e.$\xi \to \lambda_m^h$, is

considered for the first time in [126] and the error asymptote of the solution by the Trefftz method is given.

In this chapter we present some observations and related dispersion analysis of a domain-based fourth-order compact scheme for the Helmholtz equation. In other words, the phase error of the numerical solution and the local truncation error of this scheme for plane wave solutions diminish at the rate $O\left((\xi\ell)^4\right)$. The focus is on the approximation of the Helmholtz equation in the interior of the domain using compact stencils. The scheme consists in taking the alpha-interpolation of the Galerkin finite element method (FEM) and the classical finite difference method (FDM). This scheme has its origins in an old idea which marks the point of departure: to replace the consistent mass matrix $\mathbf{M}$ in the Galerkin FEM by a higher-order mass matrix $\mathbf{M}^{0.5} := (\mathbf{M} + \mathbf{M}_L)/2$, where $\mathbf{M}_L$ is the lumped mass matrix. This idea was proposed independently for eigenvalue problems by Goudreau [70, 71] and Ishihara [110]. In the later work the matrix $\mathbf{M}^{0.5}$ was denominated as the mixed-mass matrix and as a concluding remark the generalized mixed mass (GMM) scheme was proposed as an extension to the MM scheme where an α-interpolation of the mass matrices is done, i. e.$\mathbf{M}^{\alpha} := \alpha\mathbf{M} + (1-\alpha)\mathbf{M}_L$. This GMM scheme was later baptized as the alpha-interpolation method (AIM) [140] and was extended to the hollow waveguide analysis in [109] and the Schrodinger equation in [139]. For the simple 1D case our scheme mimics the AIM and in 2D making the choice $\alpha = 0.5$ we recover the generalized fourth-order compact Padé approximation [81, 168] (therein using the parameter $\gamma = 2$).

The chapter is organized as follows. In Section 4.2 we present the statement of the Helmholtz equation viewed as a diffusion–production problem. This is done only to facilitate future assimilation of ideas towards a generic method that would aim at stabilizing problems that involve the physics of both the convection–diffusion–reaction and Helmholtz equations. In Section 4.3 we present the analysis of the problem in 1D. The expressions for the numerical solution of our scheme and its relative phase error are given considering a generic definition of the parameter α given as a series expansion in terms of $(\xi\ell)$. A numerical example is given that illustrates not only the approximation properties of our scheme but also throws light on possible encounters with the zones of degeneracy. In Section 4.4 we present the 2D analysis of a nonstandard compact stencil which results from a two-parameter scheme wherein α-interpolations of the diffusion and production terms are done independently and it can model several methods (including QSFEM). This nonstandard compact stencil has an additional structure that reduces its abstractness and hence could be exploited for the extension of this stencil to unstructured meshes (cf. Section 4.4.5). We follow [10] for the analysis of this stencil and its performance on square meshes is compared with that of the quasi-stabilized FEM (QSFEM) [10]. Just like in 1D, we try to express the numerical solution of this stencil in 2D considering generic definitions of the parameters given as a series expansion in terms of $(\xi\ell)$. Using this expression for the numerical solution, the expressions for the relative phase and local truncation errors are given. In particular for our scheme, i. e.the α-interpolation of the FEM and FDM stencils an optimal expression for the parameter α is given. The dispersion plots in 2D and related discussion are done in Section 4.4.6. Some examples are presented in Section 4.4.7 which illustrate the pollution effect through convergence studies in the L^2 norm, H^1 semi-norm and the l^∞ Euclidean norms. Finally in Section 4.5 we

remark on the extension of our scheme to unstructured meshes and arrive at some conclusions.

4.2 PROBLEM STATEMENT

The statement of the multidimensional Helmholtz equation subjected to Dirichlet boundary conditions is as follows:

$$R(\phi) := k\Delta\phi + s\phi + f(\mathbf{x}) = 0 \qquad \text{in} \quad \Omega \tag{4.1a}$$

$$\phi = \phi^p \qquad \text{on} \quad \Gamma_D \tag{4.1b}$$

where $k > 0$, $s > 0$ are the diffusion and production coefficients respectively, $f(\mathbf{x})$ is the source and ϕ^p is the prescribed value of ϕ at the Dirichlet boundary. When $s < 0$ the Eq.(4.1) represents the diffusion-reaction problem that models the mass transfer processes with first-order chemical reactions and wherein s represents the reaction coefficient.

The variational statement of the problem (4.1) can be expressed as follows: Find $\phi \in V$ such that $\forall w \in V_0$ we have,

$$a(w, \phi) = l(w) \tag{4.2a}$$

$$a(w, \phi) := \int_\Omega (k\boldsymbol{\nabla} w \cdot \boldsymbol{\nabla}\phi - sw\phi)\, d\Omega \tag{4.2b}$$

$$l(w) := \int_\Omega w f(\mathbf{x})\, d\Omega \tag{4.2c}$$

where, $V := \{w : w \in H^1(\Omega) \text{ and } w = \phi^p \text{ on } \Gamma_D\}$ and $V_0 := \{w : w \in H^1(\Omega) \text{ and } w = 0 \text{ on } \Gamma_D\}$. The statement of the Galerkin method applied to the weak form (4.2) of the problem is: Find $\phi_h \in V^h$ such that $\forall w_h \in V_0^h$ we have,

$$a(w_h, \phi_h) = l(w_h) \tag{4.3}$$

where $V^h \subset V$ is a subspace obtained via any appropriate discretization. Discretization of the space by finite elements will lead to the approximation $\phi_h = N^a \Phi^a$ and Eq.(4.3) reduces into the following system of equations.

$$\Big[k\mathbf{D} - s\mathbf{M}\Big]\boldsymbol{\Phi} = \mathbf{f} \tag{4.4a}$$

$$D_{ab} = \int_\Omega \boldsymbol{\nabla} N^a \cdot \boldsymbol{\nabla} N^b\, d\Omega, \quad M_{ab} = \int_\Omega N^a N^b\, d\Omega, \quad f_a = \int_\Omega N^a f(\mathbf{x})\, d\Omega \tag{4.4b}$$

4.3 ANALYSIS IN 1D

4.3.1 *Introduction*

In this section we study the homogeneous Helmholtz equation in 1D subjected to Dirichlet boundary conditions. The problem (4.1) in 1D can be written as:

$$k\frac{d^2\phi}{dx^2} + s\phi = 0 \quad \text{in} \quad \Omega \tag{4.5a}$$

$$\phi(x = 0) = \Phi^l \quad ; \quad \phi(x = L) = \Phi^r \quad \text{on} \quad \Gamma_D \tag{4.5b}$$

where L is the length of the 1D domain and Φ^l, Φ^r are the Dirichlet boundary data at the left and right domain boundaries respectively. The solution to Eq.(4.5) when $s > 0$ is harmonic and is expressed as:

$$\phi(x) = \frac{\Phi^l \sin(\xi_o L - \xi_o x) + \Phi^r \sin(\xi_o x)}{\sin(\xi_o L)} \tag{4.6}$$

where $\xi_o := \sqrt{s/k}$ is the angular wave number. We also list the eigenvalues of this problem which can be expressed as $\lambda_m := (m\pi/L)^2 \ \forall \ m \in \{1, 2, 3, \ldots\}$. The element contributions to the matrices given in Eq.(4.4) using 2-node linear finite elements are,

$$\mathbf{D}^e = \frac{1}{\ell}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \quad ; \quad \mathbf{M}^e = \frac{\ell}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{4.7}$$

where ℓ is the corresponding element length. If the discretization is uniform the equation stencil for the problem (4.5a) corresponding to each interior node can be expressed as follows,

$$\left(\frac{k}{\ell}\right)(-\Phi^{i-1} + 2\Phi^i - \Phi^{i+1}) - \left(\frac{s\ell}{6}\right)(\Phi^{i-1} + 4\Phi^i + \Phi^{i+1}) = 0 \tag{4.8}$$

If the mass matrix $\mathbf{M}$ is lumped then the equation stencil corresponding to any interior node can be written as follows.

$$\left(\frac{k}{\ell}\right)(-\Phi^{i-1} + 2\Phi^i - \Phi^{i+1}) - s\ell\Phi^i = 0 \tag{4.9}$$

This is also the stencil we get using the classical finite difference method[2] (FDM).

4.3.2 *α-Interpolation of the Galerkin-FEM and the classical FDM*

Define a free parameter α and consider the α-interpolation of the stencils obtained by the Galerkin FEM and the classical FDM methods for the problem (4.5):

$$\begin{aligned}(1-\alpha)\left[\left(\frac{k}{\ell}\right)(-\Phi^{i-1} + 2\Phi^i - \Phi^{i+1}) - \left(\frac{s\ell}{6}\right)(\Phi^{i-1} + 4\Phi^i + \Phi^{i+1})\right] \\ +\alpha\left[\left(\frac{k}{\ell}\right)(-\Phi^{i-1} + 2\Phi^i - \Phi^{i+1}) - s\ell\Phi^i\right] = 0\end{aligned} \tag{4.10a}$$

$$\begin{aligned}\Rightarrow \left(\frac{k}{\ell}\right)(-\Phi^{i-1} + 2\Phi^i - \Phi^{i+1}) - (1-\alpha)\left(\frac{s\ell}{6}\right)(\Phi^{i-1} + 4\Phi^i + \Phi^{i+1}) \\ - \alpha s\ell\Phi^i = 0\end{aligned} \tag{4.10b}$$

$$\Rightarrow \left(\frac{k}{\ell} - \alpha\frac{s\ell}{6}\right)(-\Phi^{i-1} + 2\Phi^i - \Phi^{i+1}) - \left(\frac{s\ell}{6}\right)(\Phi^{i-1} + 4\Phi^i + \Phi^{i+1}) = 0 \tag{4.10c}$$

Remark: In 1D we can arrive at the above equations through an alternative argument: Consider the Galerkin FEM method using the α-interpolated mass matrix $\mathbf{M}^\alpha$. The later argument leads to the AIM. A particular case (taking $\alpha = 0.5$) is the mixed-mass (MM) scheme proposed by Ishihara applied to the Helmholtz equation [110]. The

2 by classical FDM we refer to the central difference scheme

mixed-mass matrix ($\mathbf{M}^{0.5}$) was earlier referred to as the higher-order-mass matrix by Goudreau [70]. In 1D a stencil equivalent to the MM scheme will be obtained using the compact fourth-order Padé approximation to problem (4.5) [81, 168].

We can guess that a solution to Eq.(4.10) takes the form $\Phi^i := \phi(x_i) = \exp(i\xi^h x_i)$. Substituting this solution into Eq.(4.10) and defining $\lambda := \exp(i\xi^h \ell)$ we get the characteristic equation of the stencil:

$$\lambda^2 - 2\left(\frac{6-(2+\alpha)\omega}{6+(1-\alpha)\omega}\right)\lambda + 1 = 0 \tag{4.11}$$

where $\omega := (s\ell^2/k) = (\xi_o \ell)^2$ is a dimensionless element number. The solution to Eq.(4.11) can be expressed as follows.

$$\lambda := e^{i\xi^h \ell} = f_\alpha \pm \sqrt{(f_\alpha)^2 - 1} = f_\alpha \pm i\sqrt{1-(f_\alpha)^2} \quad ; \quad f_\alpha := \left(\frac{6-(2+\alpha)\omega}{6+(1-\alpha)\omega}\right) \tag{4.12}$$

Note that if $|f_\alpha| \leqslant 1$ then the solution given by Eq.(4.12) is real (i.e. $\xi^h \in \mathbb{R}$). This solution can be expressed as a series expansion in terms of ω as follows:

$$\begin{aligned}\xi^h \ell = \cos^{-1}(f_\alpha) = \cos^{-1}\left(\frac{6-(2+\alpha)\omega}{6+(1-\alpha)\omega}\right) = \sqrt{\omega}\left[1 - \left(\frac{2\alpha-1}{24}\right)\omega\right. \\ \left. + \left(\frac{20\alpha^2-20\alpha+9}{1920}\right)\omega^2 + \left(\frac{280\alpha^3-420\alpha^2+378\alpha-103}{193536}\right)\omega^3 + O\left(\omega^4\right)\right]\end{aligned} \tag{4.13}$$

Should the expression for α be written as a generic series expansion in terms of ω given by $\alpha = \sum_{m=0}^{\infty} a_m \omega^m$, then the solution ξ^h can be written as shown below.

$$\begin{aligned}\xi^h \ell = \sqrt{\omega}\left[1 + \left(\frac{1-2a_0}{24}\right)\omega + \left(\frac{20a_0^2-20a_0+9}{1920} + \frac{a_1}{12}\right)\omega^2\right. \\ \left. + \left(\frac{280a_0^3-420a_0^2+378a_0-103}{193536} + \frac{(2a_0-1)a_1}{48} + \frac{a_1}{12}\right)\omega^3 + O\left(\omega^4\right)\right]\end{aligned} \tag{4.14}$$

where a_m are coefficients independent of ω. The relative phase error of the above solution can be expressed as shown below.

$$\frac{\xi^h - \xi_o}{\xi_o} = \frac{\xi^h \ell - \sqrt{\omega}}{\sqrt{\omega}} = \left[\left(\frac{1-2a_0}{24}\right)\omega + \left(\frac{20a_0^2-20a_0+9}{1920} + \frac{a_1}{12}\right)\omega^2 + O\left(\omega^3\right)\right] \tag{4.15}$$

Note that for the choice $a_0 = 1/2$, the relative phase error diminishes at the rate of $O\left(\omega^2\right)$ or equivalently $O\left((\xi_o \ell)^4\right)$. Further, making the choice $a_1 = -1/40$, the relative phase error now diminishes at the rate of $O\left(\omega^3\right)$ or equivalently $O\left((\xi_o \ell)^6\right)$. Fortunately in 1D it is possible to choose α such that the solution given by Eq.(4.12) be nodally exact (i.e. $\xi^h \ell = \xi_o \ell = \sqrt{\omega}$). The expression for α that reproduces this effect, say α_e, can be written as follows:

$$f_{\alpha_e} = \cos(\xi_o \ell) = \cos(\sqrt{\omega}) \quad \Rightarrow \quad \alpha_e = \frac{6}{\omega} - \left(\frac{2+\cos(\sqrt{\omega})}{1-\cos(\sqrt{\omega})}\right) \tag{4.16}$$

The optimal parameter α_e can be expressed as a series expansion in terms of ω as shown in Eq.(4.17). Truncating the series up to the first n terms would yield a scheme whose relative phase error diminishes at the rate of $O\left(\omega^{n+1}\right)$ or equivalently $O\left((\xi_o \ell)^{2n+2}\right)$.

$$\alpha_e \approx \frac{1}{2} - \frac{\omega}{40} - \frac{\omega^2}{1008} - \frac{\omega^3}{28800} - \frac{\omega^4}{887040} - \frac{691\omega^5}{19813248000} + O\left(\omega^6\right) \tag{4.17}$$

4.3.3 *Dispersion plots in 1D*

In this section we consider $\alpha \in \{0, 1, 0.5, \alpha_e\}$ and study their dispersion plots. The subscripts c, l, m are flags used for the expressions obtained using $\alpha = \{0, 1, 0.5\}$ respectively. These cases correspond for the stencils that arise using the consistent, lumped and mixed (higher-order) mass matrices respectively. The subscript e is used to flag the choice $\alpha = \alpha_e$, the optimal expression for α, in order to attain nodally exact numerical solutions in 1D. For the graphical representation of $f(\omega)$ and $\xi(\omega)$ we normalize some of these fields as follows:

$$\omega^* := \frac{\omega}{\pi^2} \quad ; \quad \xi^* := \frac{\xi}{\xi_{nq}} = \frac{\xi \ell}{\pi} \tag{4.18}$$

Restricting the domain to $\omega^* \in [0, 1]$ guarantees that the Nyquist frequency[3] of the discretization ((ξ_{nq})) is always greater than the frequency of the exact solution (ξ_o). Thus for every wave length of the harmonic solution we ensure the presence of at least two elements. The Nyquist-Shannon sampling theorem states that this minimum resolution of the mesh is essential to allow a perfect reconstruction of the solution using sinusoidal interpolation. However, using linear interpolation at least 4 elements per wavelength ($\xi_o \ell \leqslant (\pi/2)$ or $\omega^* \leqslant (1/4)$) are needed to capture the sinusoidal profile. As a rule of thumb at least 8 to 10 elements per wavelength are recommended for a decent representation of the solution using linear interpolation [79, 183]. The latter resolution of the mesh is guaranteed by restricting the domain to $\omega^* \in [0, 1/16]$.

Figures 58a and 58b illustrate the plot of $f(\omega^*)$ for $\omega^* \in [0, 1]$ and $\omega^* \in [0, 1/4]$ respectively. As expected a higher-order convergence of $f_m \to f_e$ is observed as $\omega^* \to 0$. Also for both the domains $f(\omega^*) \leqslant 0$ and in particular for the latter domain i. e.$\omega^* \in [0, 1/4]$, we see that $|f(\omega^*)| < 1$. Figures 58c and 58d illustrate the plot of $\xi^*(\omega^*)$ for $\omega^* \in [0, 1]$ and $\omega^* \in [0, 1/4]$ respectively. Whenever $|f(\omega^*)| > 1$, Eq.(4.12) suggests that $\lambda := \exp(i\xi^h \ell) \in \mathbb{R}$. This implies that ξ^h is a complex number ($\xi^h \in \mathbb{C}$) with the real part $\Re(\xi^h) = (n\pi/\ell), n \in \{0, 1, 2, \ldots\}$ and the imaginary part $\Im(\xi) \neq 0$. As the Nyquist frequency in space is $\xi_{nq} = \pi/\ell$, the real part is either $\Re(\xi^h) = 0$ for $f(\omega^*) \geqslant 0$ or $\Re(\xi^h) = (\pi/\ell)$ for $f(\omega^*) < 0$. Thus whenever $|f(\omega^*)| > 1$ we find $\Re(\xi^h) = (\pi/\ell)$, i. e.$\Re(\xi^{h*}) = 1$ (see Figure 58c). Also as $\Im(\xi) \neq 0$ the numerical solutions will be subjected to amplification intrinsic to the discretization (the one studied in the von Neumann analysis). Finally, whenever $|f(\omega^*)| \leqslant 1$, the solution ξ^h is real ($\xi^h \in \mathbb{R}$) and all the considered schemes are devoid of any amplification intrinsic to the discretization. In Figure 58d we observe that for $\omega^* \in [0, 1/16]$, the graphs of ξ^*_e and ξ^*_m are indistinguishable.

4.3.4 *Examples*

We consider the problem defined in Eq.(4.5) with the following problem data: $k =$ `1e-3`, $s = 1$, $L = 1$, $\Phi^l = 3$, $\Phi^r = 1$. Thus the exact solution of the problem given by Eq.(4.6) has an angular wave number $\xi_o = 10\sqrt{10}$. The discretization of the space is done by linear finite elements and is uniform. We solve the problem using $\alpha \in \{0, 1, 0.5\}$ and the subscripts c, l, m are used to flag them respectively. Four meshes of

3 `http://en.wikipedia.org/wiki/Nyquist_frequency`. Here frequency is to be understood in the spatial context, i. e.the wavenumber

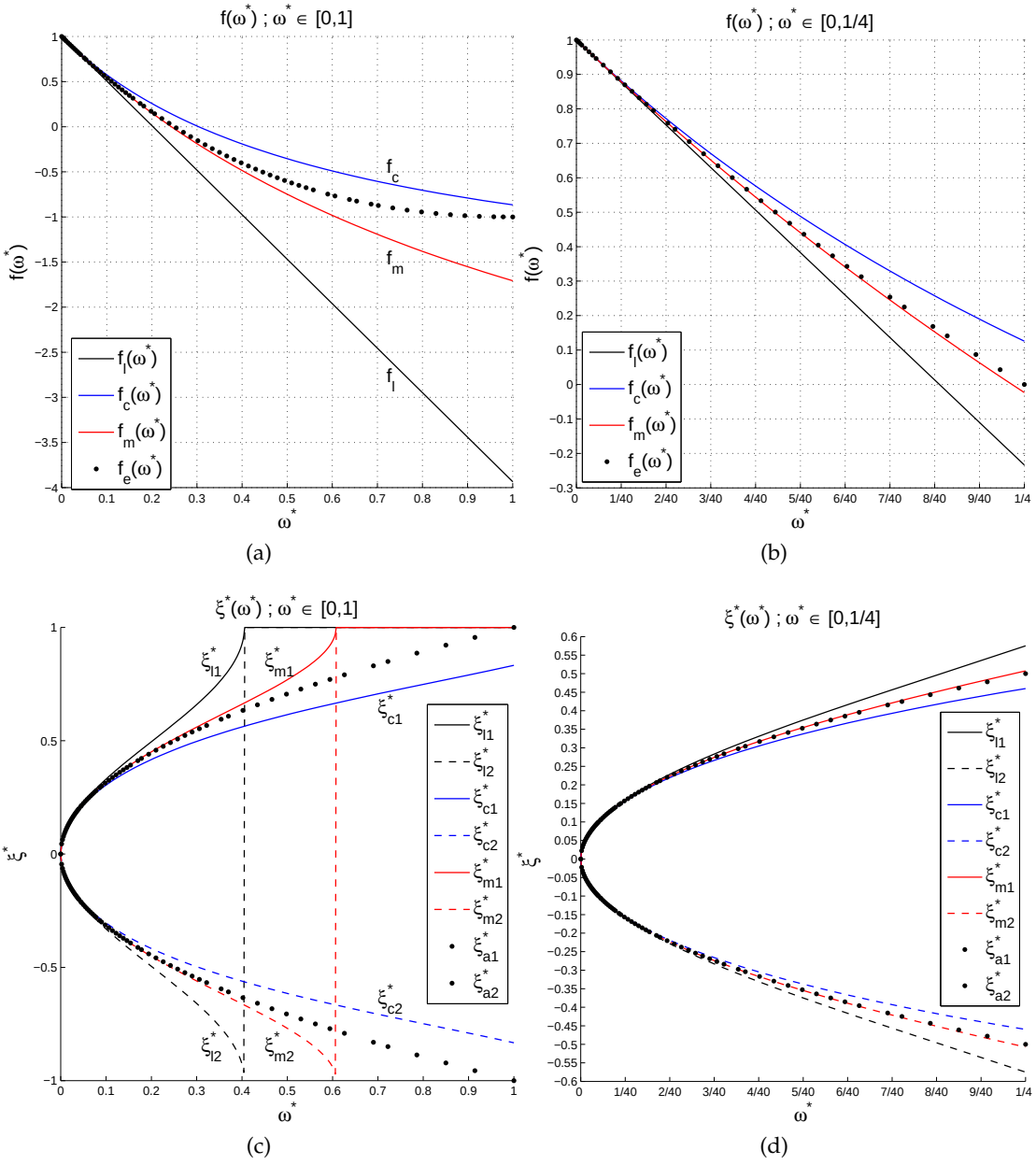


Figure 58: Plots of $f(\omega^*)$ and $\xi^*(\omega^*)$. (a) Domain: $\omega^* \in [0,1]$; (b) Domain: $\omega^* \in [0,1/4]$; (c) Domain: $\omega^* \in [0,1]$; (d) Domain: $\omega^* \in [0,1/4]$

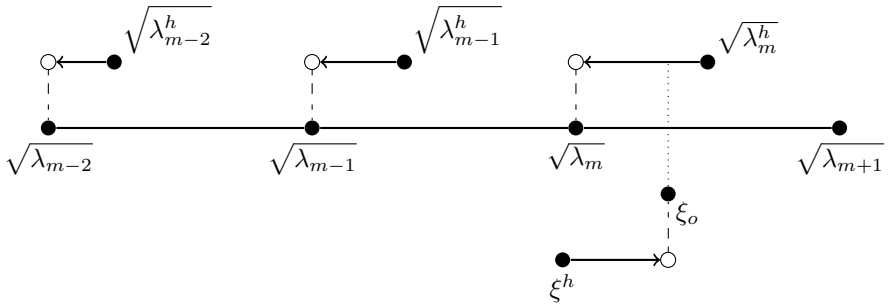


Figure 59: A schematic diagram that illustrates the encounter of a zone of degeneracy on mesh refinement ($\ell \to 0$). As the value of $\sqrt{\lambda^h_m}$ crosses ξ_o on its path towards $\sqrt{\lambda_m}$, the discrete LBB constant takes values arbitrarily close to zero.

different resolution viz. 41, 81, 162 and 323 elements are considered. These meshes guarantee the presence of at least 8, 16, 32 and 64 elements per wavelength of the harmonic solution respectively. All the meshes restrict the domain of ω^* to [0,1/16].

Figure 60 illustrates the plots of the numerical solutions obtained using a consistent, lumped and semi-lumped mass matrices denoted by Φ^c_h, Φ^l_h and Φ^m_h respectively, against the exact solution of the problem denoted by Φ^a. In Figure 60a the solutions Φ^c_h and Φ^l_h are out-of-phase and as expected the phase accuracy improves on mesh refinement (Figures 60b-d). We observe a remarkable error in the amplitude of these solutions. Note that there is no intrinsic amplification for all the schemes and the errors in the angular wave numbers ξ^*_c, ξ^*_l are small (Figure 58d). The amplitude of the solution depends not only on the intrinsic amplification of the scheme but also on the wave number ξ and on the applied Dirichlet boundary conditions. Thus we may conclude that small errors in the wave number of the computed solution may result in huge errors in their amplitude. An alternative explanation to this behavior can be given via the following argument. First note that $(\sqrt{\lambda_{10}} = 31.4159) < (\xi_o = 10\sqrt{10} = 31.6227) < (\sqrt{\lambda_{11}} = 34.5575)$. It is possible that the discrete eigenvalue $\sqrt{\lambda^h_{10}}$ for the initial course mesh/grid is greater than ξ_o and on further mesh refinement it approaches $\sqrt{\lambda_{10}}$ by crossing ξ_o. This explains the observation that the numerical solution Φ^c_h on mesh refinement first explodes (as it enters the zone of degeneracy) and then gradually converges to the exact solution. Figure 59 illustrates schematically[4] the encounter of a zone of degeneracy on mesh refinement ($\ell \to 0$) while as the value of $\sqrt{\lambda^h_m}$ crosses ξ_o on its path towards $\sqrt{\lambda_m}$. Nevertheless, the convergence of the discrete wavenumber ($\xi^h \to \xi_o$) need not be affected in this process. This argument also suggest that this phenomenon could have been equally observed for the solutions Φ^l_h and Φ^m_h should their corresponding discrete eigenvalues cross ξ_o.

On the other hand, the solution Φ^m_h could represent approximately the profile of the exact solution even on the coarsest mesh (see Figure 60a). Figures 60b-d show that on further mesh refinements Φ^m_h is indistinguishable from the analytical solution.

4 A similar figure was presented earlier in [47] (c.f. Figure 1, pp. 74)

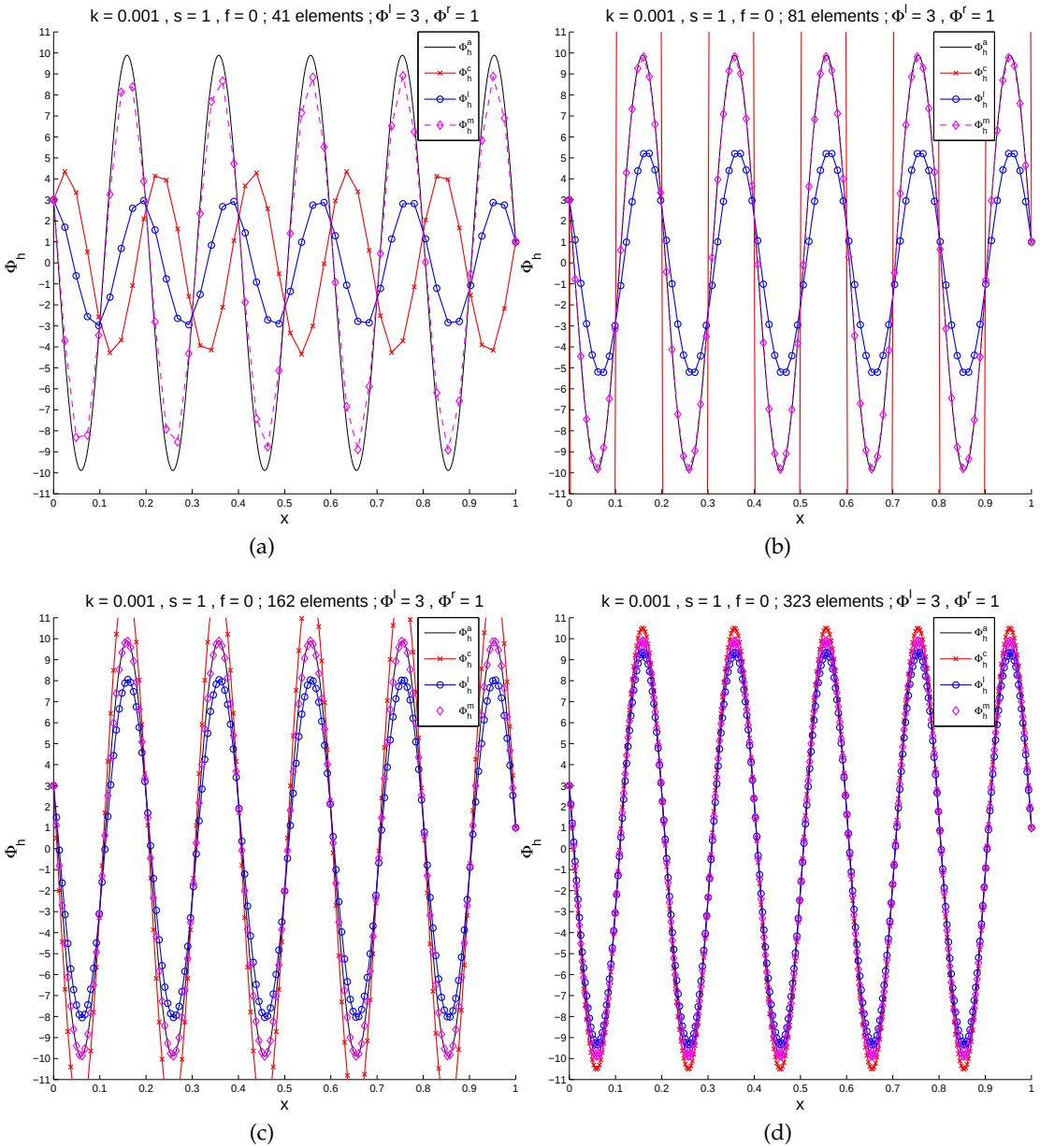


Figure 60: Numerical solution Φ_h using a mesh with at least: (a) 8 elements per wave length ; (b) 16 elements per wave length ; (c) 32 elements per wave length ; (d) 64 elements per wave length. In figures (c) and (d) the solution Φ_h^m effectively coincides with the exact solution and the solutions Φ_m^c and Φ_m^l bound the exact solution from above and below respectively.

4.4 ANALYSIS IN 2D

4.4.1 *Introduction*

In multidimensions the general solution to the problem (4.1) considering a linear source $f(\mathbf{x})$ may be expressed as follows:

$$\phi(\mathbf{x}) = \frac{f}{s} + \sum_{\theta} C_\theta \exp(i\boldsymbol{\xi}^\theta \cdot \mathbf{x}) \tag{4.19a}$$

$$|\boldsymbol{\xi}^\theta| = \xi_o \quad \Rightarrow \quad \boldsymbol{\xi}^\theta := (\xi_1^\theta, \xi_2^\theta) = (\xi_o \cos(\theta), \xi_o \sin(\theta)) \tag{4.19b}$$

where, C_θ represents a generic constant independent of the spatial coordinates. Generally it is not possible to arrive at an expression for C_θ in the closed form. Nevertheless this detail is not needed in the Fourier analysis of these problems. Eliminating θ from Eq.(4.19b) we arrive at the characteristic equation of the continuous problem (4.1):

$$(\xi_1^\theta)^2 + (\xi_2^\theta)^2 = \xi_o^2 \tag{4.20}$$

4.4.2 *Galerkin FEM using rectangular bilinear finite elements*

The element contributions to the matrices given in Eq.(4.4) using 4-node rectangular bilinear finite elements are,

$$\mathbf{D}^e = \frac{\ell_2}{6\ell_1}\begin{bmatrix} 2 & -2 & -1 & 1 \\ -2 & 2 & 1 & -1 \\ -1 & 1 & 2 & -2 \\ 1 & -1 & -2 & 2 \end{bmatrix} + \frac{\ell_1}{6\ell_2}\begin{bmatrix} 2 & 1 & -1 & -2 \\ 1 & 2 & -2 & -1 \\ -1 & -2 & 2 & 1 \\ -2 & -1 & 1 & 2 \end{bmatrix} \tag{4.21a}$$

$$\mathbf{M}^e = \frac{\ell_1 \ell_2}{36}\begin{bmatrix} 4 & 2 & 1 & 2 \\ 2 & 4 & 2 & 1 \\ 1 & 2 & 4 & 2 \\ 2 & 1 & 2 & 4 \end{bmatrix} \tag{4.21b}$$

where ℓ_1, ℓ_2 are the corresponding element lengths along the 2D axes. Restraining the discretization to be uniform, we can arrive at an equation stencil for every interior node of the mesh. We use the following notation (described earlier in §3.1) to represent a generic compact stencil obtained for the (i, j) node on a rectangular grid.

$$\{\circ^{j+1}, \circ^{j}, \circ^{j-1}\}\mathbf{A}\{\circ^{i-1}, \circ^{i}, \circ^{i+1}\}^t = 0 \tag{4.22}$$

where $\mathbf{A}$ represents the matrix of the stencil coefficients. We can guess that a solution to Eq.(4.22) takes the form $\Phi^{i,j} := \phi(x_1^i, x_2^j) = \exp[i(\xi_1^h x_1^i + \xi_2^h x_2^j)]$. Substituting this solution into Eq.(4.22) and defining $\lambda_1 := \exp(i\xi_1^h \ell_1)$ and $\lambda_2 := \exp(i\xi_2^h \ell_2)$ we get the characteristic equation of the generic stencil(4.22):

$$\left\{\lambda_2, 1, \lambda_2^{-1}\right\} \mathbf{A} \left\{\lambda_1^{-1}, 1, \lambda_1\right\}^t = 0 \tag{4.23}$$

The stencil for the Galerkin FEM method corresponding to any interior node (i,j) can be written as Eq.(4.22) with the following definition of the stencil coefficient matrix ($\mathbf{A}$):

$$\begin{aligned}\mathbf{A}^{fem} := &\frac{k\ell_2}{6\ell_1}\{1,4,1\}^t\{-1,2,-1\} + \frac{k\ell_1}{6\ell_2}\{-1,2,-1\}^t\{1,4,1\} \\ &- \frac{s\ell_1\ell_2}{36}\{1,4,1\}^t\{1,4,1\}\end{aligned} \tag{4.24}$$

The stencil for the classical FDM method corresponding to any interior node (i,j) can be written as Eq.(4.22) with the following definition of $\mathbf{A}$:

$$\begin{aligned}\mathbf{A}^{fdm} := &\frac{k\ell_2}{6\ell_1}\{0,6,0\}^t\{-1,2,-1\} + \frac{k\ell_1}{6\ell_2}\{-1,2,-1\}^t\{0,6,0\} \\ &- \frac{s\ell_1\ell_2}{36}\{0,6,0\}^t\{0,6,0\}\end{aligned} \tag{4.25}$$

The characteristic equation associated with the stencil for the Galerkin FEM can be written as Eq.(4.23) using the definition of $\mathbf{A}$ given by Eq.(4.24). Likewise the characteristic equation associated with the stencil for the classical FDM can be written as Eq.(4.23) using the definition of $\mathbf{A}$ given by Eq.(4.25).

4.4.3 *A nonstandard compact stencil in 2D*

Define two free parameters α_1, α_2 and consider the following definition of $\mathbf{A}$:

$$\begin{aligned}\mathbf{A}^{\alpha_1,\alpha_2} := &(1-\alpha_1)\frac{k\ell_2}{6\ell_1}\{1,4,1\}^t\{-1,2,-1\} + \alpha_1\frac{k\ell_2}{6\ell_1}\{0,6,0\}^t\{-1,2,-1\} \\ &+ (1-\alpha_1)\frac{k\ell_1}{6\ell_2}\{-1,2,-1\}^t\{1,4,1\} + \alpha_1\frac{k\ell_1}{6\ell_2}\{-1,2,-1\}^t\{0,6,0\} \\ &- (1-\alpha_2)\frac{s\ell_1\ell_2}{36}\{1,4,1\}^t\{1,4,1\} - \alpha_2\frac{s\ell_1\ell_2}{36}\{0,6,0\}^t\{0,6,0\}\end{aligned} \tag{4.26}$$

Note that taking $\alpha_1 = \alpha_2 = \alpha$ we arrive at a stencil that is the α-interpolation of the FEM and FDM stencils, i. e. $\mathbf{A}^{\alpha,\alpha} = (1-\alpha)\mathbf{A}^{fem} + \alpha\mathbf{A}^{fdm}$. Likewise taking $\alpha_1 = 0$ and $\alpha_2 = \alpha$ we arrive at a stencil that results from the Galerkin FEM method using an α-interpolated mass matrix $\mathbf{M}^\alpha := (1-\alpha)\mathbf{M} + \alpha\mathbf{M}_L$. We remark that unlike in 1D where both choices resulted in the same stencil, in 2D the obtained stencils are different.

Next we relate this nonstandard stencil with the compact fourth-order Padé approximation in 2D. A generalized version of the same was studied in [81, 168] and the associated stencil coefficient matrix of the scheme $\mathbf{A}^\gamma$ can be expressed as follows:

$$\begin{aligned}\mathbf{A}^\gamma := &-k\left[\{0,1,0\} + \frac{\{1,-2,1\}}{12}\right]^t\frac{\{1,-2,1\}}{\ell_1^2} - k\frac{\{1,-2,1\}^t}{\ell_2^2}\left[\{0,1,0\} + \frac{\{1,-2,1\}}{12}\right] \\ &- s\left[\{0,1,0\} + \frac{\{1,-2,1\}}{12}\right]^t\left[\{0,1,0\} + \frac{\{1,-2,1\}}{12}\right] - s(\gamma-1)\frac{\{1,-2,1\}^t}{12}\frac{\{1,-2,1\}}{12}\end{aligned} \tag{4.27}$$

where γ is a free parameter. The standard compact fourth-order Padé scheme in 2D is obtained by selecting $\gamma = 1$. Other alternatives viz. $\gamma = 0$ and $\gamma = 2$ were presented

in [40](cf. Appendix, Table VI, p.542). After some algebraic rearrangement, matrix $\mathbf{A}^{\gamma}$ given in Eq.(4.27) can be re-written equivalently as follows:

$$\mathbf{A}^{\gamma} := \frac{k}{2}\left[\frac{\{1,4,1\}}{6}+\frac{\{0,6,0\}}{6}\right]^{t}\frac{\{-1,2,-1\}}{\ell_1^2}+\frac{k}{2}\frac{\{-1,2,-1\}^{t}}{\ell_2^2}\left[\frac{\{1,4,1\}}{6}+\frac{\{0,6,0\}}{6}\right] - \frac{s}{2}\left[\frac{\{1,4,1\}^{t}}{6}\frac{\{1,4,1\}}{6}+\frac{\{0,6,0\}^{t}}{6}\frac{\{0,6,0\}}{6}\right]-s(\gamma-2)\frac{\{1,-2,1\}^{t}}{12}\frac{\{1,-2,1\}}{12} \tag{4.28}$$

Note that by selecting $\gamma = 2$ we obtain a stencil that is equivalent to the one obtained by taking the average of the FEM and the FDM stencils. Thus,

$$\mathbf{A}^{2} = \frac{1}{\ell_1\ell_2}\mathbf{A}^{0.5,0.5} \tag{4.29}$$

We now relate this nonstandard stencil for square meshes with the compact scheme proposed by Vichnevetsky and Bowles [187] in order to reduce the anisotropy related to the numerical dispersion. This scheme was studied in [183] and the conditions for appropriate numerical isotropy were determined therein. Also, this scheme was used to synthesize an equivalent transmission-line matrix (TLM) [115] model for the Maxwell's equations in [167]. The associated stencil coefficient matrix $\mathbf{A}^{vb}$ can be written as follows.

$$\mathbf{A}^{vb} := \frac{\gamma k}{\ell^2}\begin{bmatrix} 0 & -1 & 0\\ -1 & 4 & -1\\ 0 & -1 & 0\end{bmatrix}+\frac{(1-\gamma)k}{2\ell^2}\begin{bmatrix} -1 & 0 & -1\\ 0 & 4 & 0\\ -1 & 0 & -1\end{bmatrix}-s\begin{bmatrix} 0 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0\end{bmatrix} \tag{4.30}$$

where γ is the associated interpolation parameter. Note that for $\gamma = 1$ we recover the classical FDM (i. e.the second-order central difference scheme) and for $\gamma = 0$ we get a similar scheme but with the stencil inclined at 45° and hence with the mesh size $\sqrt{2}\ell$. Note that we recover the Galerkin FEM contribution of the term $-k\Delta\phi$ by choosing $\gamma = (1/3)$. Making the substitution $\gamma = (1+2\alpha)/3$ in Eq.(4.30) and after some algebraic rearrangement, matrix $\mathbf{A}^{vb}$ can be re-written equivalently as follows:

$$\mathbf{A}^{vb} := k\left[\frac{(1-\alpha)}{6}\{1,4,1\}+\frac{\alpha}{6}\{0,6,0\}\right]^{t}\frac{\{-1,2,-1\}}{\ell^2} + k\frac{\{-1,2,-1\}^{t}}{\ell^2}\left[\frac{(1-\alpha)}{6}\{1,4,1\}+\frac{\alpha}{6}\{0,6,0\}\right]-s\frac{\{0,6,0\}^{t}}{6}\frac{\{0,6,0\}}{6} \tag{4.31}$$

This is precisely what we get using an α-interpolated (Galerkin FEM and classical FDM) diffusion matrix in the classical FDM stencil. Thus, on square meshes we can relate $\mathbf{A}^{vb}$ with the nonstandard stencil as shown below.

$$\mathbf{A}^{vb} = \frac{1}{\ell^2}\mathbf{A}^{\alpha,1} \tag{4.32}$$

Using the definition of $\mathbf{A}$ given by Eq.(4.26), the characteristic equation associated to the resulting stencil is given by Eq.(4.33) and on simplification we arrive at Eq.(4.34).

$$\left\{\lambda_2, 1, \lambda_2^{-1}\right\} \mathbf{A}^{\alpha_1,\alpha_2} \left\{\lambda_1^{-1}, 1, \lambda_1\right\}^t = 0 \tag{4.33}$$

$$\begin{aligned} \Rightarrow & \left(\frac{[(1-\alpha_1)(\lambda_2^2+4\lambda_2+1)+6\alpha_1\lambda_2](-1+2\lambda_1-\lambda_1^2)}{6\omega_1}\right) \\ & + \left(\frac{(-\lambda_2^2+2\lambda_2-1)[(1-\alpha_1)(1+4\lambda_1+\lambda_1^2)+6\alpha_1\lambda_1]}{6\omega_2}\right) \\ & - \left(\frac{[(1-\alpha_2)(\lambda_2^2+4\lambda_2+1)(1+4\lambda_1+\lambda_1^2)+36\alpha_2\lambda_2\lambda_1]}{36}\right) = 0 \end{aligned} \tag{4.34}$$

where ω_1, ω_2 are two dimensionless element numbers defined as follows:

$$\omega_1 := \frac{s\ell_1^2}{k} = (\xi_o \ell_1)^2 \quad ; \quad \omega_2 := \frac{s\ell_2^2}{k} = (\xi_o \ell_2)^2 \tag{4.35}$$

Unlike in 1D, the characteristic equations of the stencils in 2D have infinite solutions (fundamental frequencies (ξ_1^h, ξ_2^h)) for every (ω_1, ω_2) pair. For every choice of the pair (ω_1, ω_2), these solutions will trace well-defined contours in the $\xi_1^h - \xi_2^h$ plane. The solutions to Eq.(4.34) are symmetric about the origin and the axes. This statement can be easily verified due to the fact that by replacing the pair (λ_1, λ_2) with $(\lambda_1^{\pm 1}, \lambda_2^{\pm 1})$ in Eq.(4.34) we end up in the same equation. Thus we may conclude that if (ξ_1^h, ξ_2^h) is a solution to Eq.(4.34) then $(\pm\xi_1^h, \pm\xi_2^h)$ are also solutions to the same. Obviously this statement also extends to the characteristic equation of the continuous problem (4.20) which additionally has a rotational symmetry (i.e.if $(\xi_1^\theta, \xi_2^\theta)$ is a solution then $(\xi_2^\theta, \xi_1^\theta)$ is also a solution). These contour lines are circular for the continuous problem and their radius equals to the chosen ξ_o value. Rotational symmetry for the solution (ξ_1^h, ξ_2^h) is attained should the element lengths be the same, i.e.$\ell_1 = \ell_2 = \ell$. In this case the stencil coefficient matrix $\mathbf{A}^{\alpha_1,\alpha_2}$ is symmetric and after scaling down by k it can be expressed as follows:

$$\frac{\mathbf{A}^{\alpha_1,\alpha_2}}{k} = \begin{bmatrix} A_2 & A_1 & A_2 \\ A_1 & A_0 & A_1 \\ A_2 & A_1 & A_2 \end{bmatrix}^{\alpha_1,\alpha_2} \quad ; \quad \begin{aligned} A_0^{\alpha_1,\alpha_2} &:= \frac{8}{3} - \frac{4\omega}{9} + \frac{4\alpha_1}{3} - \frac{5\omega\alpha_2}{9} \\ A_1^{\alpha_1,\alpha_2} &:= -\frac{1}{3} - \frac{\omega}{9} - \frac{2\alpha_1}{3} + \frac{\omega\alpha_2}{9} \\ A_2^{\alpha_1,\alpha_2} &:= -\frac{1}{3} - \frac{\omega}{36} + \frac{\alpha_1}{3} + \frac{\omega\alpha_2}{36} \end{aligned} \tag{4.36}$$

where, $\omega := (s\ell^2/k) = (\xi_o \ell)^2$.

Finally we relate this nonstandard stencil with methods that have a symmetric stencil coefficient matrix $\mathbf{A}^{sym}$ defined as:

$$\mathbf{A}^{sym} := \begin{bmatrix} A_2 & A_1 & A_2 \\ A_1 & A_0 & A_1 \\ A_2 & A_1 & A_2 \end{bmatrix} \tag{4.37}$$

Let $g_1 = (4A_1/A_0)$ and $g_2 = (4A_2/A_0)$ and if $(g_1 + g_2 + 1) \neq 0$ then we can obtain $\mathbf{A}^{sym}$ (possibly scaled by a factor) from $\mathbf{A}^{\alpha_1,\alpha_2}$ by selecting α_1 and α_2 as follows:

$$\alpha_1 := \frac{4(g_1+g_2+1)+\omega(g_1-4g_2)}{8(g_1+g_2+1)} \quad ; \quad \alpha_2 := \frac{12(g_1+g_2+1)+\omega(2-g_1-4g_2)}{2\omega(g_1+g_2+1)}$$

(4.38)

For instance consider the QSFEM method [10] for which the expressions for g_1 and g_2 can be written as shown below.

$$g_1 := \frac{2(c_1 s_1 - c_2 s_2)}{c_2 s_2 (c_1 + s_1) - c_1 s_1 (c_2 + s_2)} \quad ; \quad g_2 := \frac{(c_2 + s_2 - c_1 - s_1)}{c_2 s_2 (c_1 + s_1) - c_1 s_1 (c_2 + s_2)} \tag{4.39a}$$

$$\begin{aligned} c_1 &:= \cos\left[\sqrt{\omega}\cos\left(\frac{\pi}{16}\right)\right] \quad c_2 := \cos\left[\sqrt{\omega}\cos\left(\frac{3\pi}{16}\right)\right] \\ s_1 &:= \cos\left[\sqrt{\omega}\sin\left(\frac{\pi}{16}\right)\right] \quad s_2 := \cos\left[\sqrt{\omega}\sin\left(\frac{3\pi}{16}\right)\right] \end{aligned} \tag{4.39b}$$

4.4.4 *Numerical solution, phase error and local truncation error*

In this section we will deal only with the case when $\ell_1 = \ell_2 = \ell$. Here we present the solution to Eq.(4.34) for a given α_1, α_2 expressed as a generic series expansion in terms of ω as follows:

$$\alpha_1 := \sum_{m=0}^{\infty} a_m \omega^m \approx a_0 + a_1\omega + a_2\omega^2 + a_3\omega^3 + O\left(\omega^4\right) \tag{4.40a}$$

$$\alpha_2 := \sum_{m=0}^{\infty} b_m \omega^m \approx b_0 + b_1\omega + b_2\omega^2 + b_3\omega^3 + O\left(\omega^4\right) \tag{4.40b}$$

where a_m, b_m are coefficients independent of ω. Following [10] the solution $\xi^h := (\xi_1^h, \xi_2^h)$ can also be expressed as a series expansion in terms of ω:

$$\begin{Bmatrix} \xi_1^h \ell \\ \xi_2^h \ell \end{Bmatrix} = R(a_m, b_m, \beta, \omega) \begin{Bmatrix} \cos(\beta) \\ \sin(\beta) \end{Bmatrix} \tag{4.41a}$$

$$R := \sqrt{\omega}\left[1 + \sum_{m=1}^{\infty} r_m(a_i, b_i, \beta)\omega^m\right] \approx \sqrt{\omega}\left[1 + r_1\omega + r_2\omega^2 + r_3\omega^3 + O\left(\omega^4\right)\right] \tag{4.41b}$$

where, r_m are coefficients independent of ω and will be determined later in this section. Recall that the numerical solution in 1D given by Eq.(4.13) or Eq.(4.14) obeys the above series expansion in terms of ω. Figure 61 illustrates schematically the contour traced by the numerical solution $P^h(\xi_1^h \ell, \xi_2^h \ell)$ and compares it with the contour of the exact solution $P(\xi_1^\beta \ell, \xi_2^\beta \ell)$. In [10] the denomination 'dist(β)' was used for the distance between P^h and P, i. e.dist$(\beta) := R - \sqrt{\omega}$. Therein 'dist($\beta$)' was used as a measure of the approximation quality of the solution and from it error estimates were derived that bound the solution from below. The relative phase error of the solution along any direction β is given by,

$$\frac{\|\xi^h\| - \|\xi^\beta\|}{\|\xi^\beta\|} = \frac{R - \sqrt{\omega}}{\sqrt{\omega}} = \frac{\text{dist}(\beta)}{\sqrt{\omega}} = \sum_{m=1}^{\infty} r_m\, \omega^m \approx \left[r_1\omega + r_2\omega^2 + r_3\omega^3 + O\left(\omega^4\right)\right] \tag{4.42}$$

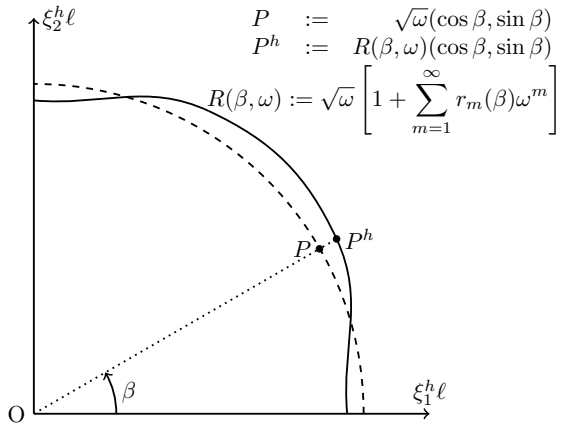


Figure 61: A schematic diagram of the contours traced by the numerical solution $P^h(\xi_1^h \ell, \xi_2^h \ell)$ and the exact solution $P(\xi_1^\beta \ell, \xi_2^\beta \ell)$.

Substituting $P^h(\xi_1^h \ell, \xi_2^h \ell)$ into the stencil corresponding to $\mathbf{A}^{\alpha_1,\alpha_2}$ given in Eq.(4.36) we get:

$$A_0^{\alpha_1,\alpha_2} + 2A_1^{\alpha_1,\alpha_2}[\cos(R\cos\beta) + \cos(R\sin\beta)] + 4A_2^{\alpha_1,\alpha_2}\cos(R\cos\beta)\cos(R\sin\beta) = 0 \tag{4.43}$$

Using the definitions of α_1, α_2 and R given in Eq.(4.40) and Eq.(4.41b) respectively, the left hand side (LHS) of Eq.(4.43) can be expanded as a series in terms of ω as shown in Eq.(4.44a). The first four coefficients of this series can be expressed as shown in Eq.(4.44b) and Eq.(4.44c) respectively.

$$\text{LHS} = \sum_{m=0}^{\infty} S_m(a_i, b_j, r_k, \beta)\omega^m \tag{4.44a}$$

$$S_0 = S_1 = 0 \quad ; \quad S_2 = 2r_1 + \left(\frac{3 + 2a_0 - 8b_0}{48}\right) + \left(\frac{1 - 2a_0}{48}\right)\cos(4\beta) \tag{4.44b}$$

$$\begin{aligned} S_3 = 2r_2 + r_1^2 + \left(\frac{2a_0 - 4b_0 - 1}{12}\right) r_1 + \left(\frac{24a_1 - 96b_1 - 2a_0 + 8b_0 - 5}{576}\right) \\ + \left[\left(\frac{10a_0 - 7 - 120a_1}{2880}\right) + \left(\frac{1 - 2a_0}{12}\right) r_1\right]\cos(4\beta) \end{aligned} \tag{4.44c}$$

The local truncation error of the solution along any direction β is found by substituting the exact solution $P(\xi_1^\beta \ell, \xi_2^\beta \ell)$ into the stencil corresponding to $\mathbf{A}^{\alpha_1,\alpha_2}$ given in Eq.(4.36). This is equivalent to substituting $r_k = 0 \; \forall \; k$ in the expression for LHS given in Eq.(4.44a). Thus using the result $S_0 = S_1 = 0$, the relative truncation error $\mathbb{T}$ along any direction β is given by:

$$\mathbb{T} := \frac{\text{LHS}|_P}{\omega} = \sum_{2}^{\infty} S_m(a_i, b_j, r_k = 0, \beta)\omega^{m-1} \approx \left[S_2\omega + S_3\omega^2 + S_4\omega^3 + O\left(\omega^4\right)\right] \tag{4.45}$$

We now present the expressions for the unknowns r_k. Clearly all the coefficients S_m should be zero for Eq.(4.43) to hold. We can solve for the unknowns r_k by imposing

the conditions $S_m = 0 \; \forall \; m$. Thus the first two unknowns in Eq.(4.41b) viz. r_1 and r_2 can be expressed as follows:

$$r_1 = -\left(\frac{3+2a_0-8b_0}{96}\right) - \left(\frac{1-2a_0}{96}\right)\cos(4\beta) \tag{4.46a}$$

$$\begin{aligned} r_2 = -\frac{r_1^2}{2} - \left(\frac{2a_0-4b_0-1}{24}\right) r_1 - \left(\frac{24a_1-96b_1-2a_0+8b_0-5}{1152}\right) \\ - \left[\left(\frac{10a_0-7-120a_1}{5760}\right) + \left(\frac{1-2a_0}{24}\right) r_1\right] \cos(4\beta) \end{aligned} \tag{4.46b}$$

Note that we obtain the condition $r_1 = 0$ if and only if a_0 and b_0 satisfy the condition $a_0 = b_0 = (1/2)$. Further we obtain the condition $r_2 = 0$ if and only if a_1 and b_1 satisfy the condition $a_1 = (-1/60)$ and $b_1 = (-1/40)$. For these choices of a_0, a_1, b_0 and b_1 the first five coefficients in $\{S_m\}$ can be simplified as follows:

$$S_0 = S_1 = 0 \quad ; \quad S_2 = 2r_1 \quad ; \quad S_3 = 2r_2 + r_1^2 - \frac{r_1}{6} \tag{4.47a}$$

$$\begin{aligned} S_4 = 2r_3 + 2r_1 r_2 - \frac{r_2}{6} - \frac{r_1}{720} - \frac{5r_1^2}{12} \\ - \left[\frac{5}{55296} - \left(\frac{a_2-4b_2}{24}\right) + \left(\frac{1+576a_2}{13824}\right)\cos(4\beta) + \frac{\cos(8\beta)}{387072}\right] \end{aligned} \tag{4.47b}$$

Likewise, by imposing the condition $S_4 = 0$ the unknown r_3 in Eq.(4.41b) can be simplified to the following:

$$r_3 = \left[\frac{5}{110592} - \left(\frac{a_2-4b_2}{48}\right) + \left(\frac{1+576a_2}{27648}\right)\cos(4\beta) + \frac{\cos(8\beta)}{774144}\right] \tag{4.48}$$

Clearly it is impossible to obtain the condition $r_3 = 0$ and this fact was pointed out earlier in [10]. To conclude this section we summarize the salient results. The parameters α_1 and α_2 that appear in $\mathbf{A}^{\alpha_1,\alpha_2}$ can be chosen such that the numerical solution be sixth-order accurate, i. e.$O\left((\xi_o \ell)^6\right)$ or equivalently $O\left(\omega^3\right)$. Recall that this is the maximum order of accuracy that can be attained on any compact stencil [10]. All such α_1 and α_2 should obey the following series expansion in terms of ω.

$$\alpha_1 = \frac{1}{2} - \frac{\omega}{60} + \sum_{m=2}^{\infty} a_m \omega^m \quad ; \quad \alpha_2 = \frac{1}{2} - \frac{\omega}{40} + \sum_{m=2}^{\infty} b_m \omega^m \tag{4.49}$$

The relative phase and local truncation errors of these schemes can be expressed as follows:

$$\frac{\|\xi^h\| - \|\xi^\beta\|}{\|\xi^\beta\|} = r_3 \omega^3 + O\left(\omega^4\right) \quad ; \quad \mathbb{T} = -2r_3\omega^3 + O\left(\omega^4\right) \tag{4.50}$$

where r_3 is given in Eq.(4.48). As $a_m, b_m \; (m \geqslant 2)$ can be chosen arbitrarily, infinitely many sixth-order schemes can be designed through $\mathbf{A}^{\alpha_1,\alpha_2}$. Of course some particular choice of a_m, b_m may yield a scheme with better features. For instance, a_m, b_m may be chosen such that the local truncation error $\mathbb{T}$ be zero along some chosen directions.

4.4.5 α-Interpolation of the FEM and the FDM in 2D

In this section we consider the case $\alpha_1 = \alpha_2 = \alpha$, that results in a scheme which is the α-interpolation of the FEM and the FDM stencils. Here the coefficients $a_m = b_m \; \forall \; m$. Recall that a necessary condition to obtain a sixth-order scheme is $a_1 = (-1/60)$ and $b_1 = (-1/40)$. Thus an immediate consequence is that this α-interpolation scheme can be at the best fourth-order accurate. Nevertheless, a compromise to the loss in accuracy is that the condition $\alpha_1 = \alpha_2 = \alpha$ imposes an additional structure to the scheme that may be exploited. For instance, this additional structure might throw light on the extension of this scheme to unstructured meshes. Precisely, should it be possible to design a Petrov–Galerkin method that would yield the FDM stencil on a structured mesh, then this scheme can be extended to unstructured meshes in a straight-forward manner. We show that indeed it is possible to design such a Petrov–Galerkin method using just the lowest-order block finite elements in chapter 5 and [138].

We now discuss the salient features of this scheme, i. e.the case $\alpha_1 = \alpha_2 = \alpha$. It is possible to choose α such that the local truncation error along any direction θ be zero. Let this choice be denominated as α_θ and it can be expressed as follows:

$$\alpha_\theta := \frac{6(c_\theta + s_\theta + 2c_\theta s_\theta - 4) + \omega(2c_\theta + 2s_\theta + c_\theta s_\theta + 4)}{12(1 - c_\theta - s_\theta + c_\theta s_\theta) + \omega(2c_\theta + 2s_\theta + c_\theta s_\theta - 5)} \quad ; \quad \begin{aligned} c_\theta &:= \cos(\sqrt{\omega}\cos(\theta)) \\ s_\theta &:= \cos(\sqrt{\omega}\sin(\theta)) \end{aligned} \tag{4.51}$$

Note that choosing $\theta = 0$ we would recover the expression for α given in Eq.(4.16) which results in solutions that are nodally exact in 1D. The expression for α_θ can be written as a series expansion in terms of ω as shown below:

$$\alpha_\theta = \sum_{m=0}^{\infty} a_m \omega^m \approx \frac{1}{2} - \left[\frac{5 + \cos(4\theta)}{3 + \cos(4\theta)}\right]\frac{\omega}{60} - \left[\frac{35 + 28\cos(4\theta) + \cos(8\theta)}{3 + \cos(4\theta)}\right]\frac{\omega^2}{16128} + O\left(\omega^3\right) \tag{4.52}$$

Recall that the choice $a_0 = (1/2)$ will make the coefficient $r_1 = 0$ and hence using the expression for α_θ we will always obtain fourth-order accurate solutions on uniform meshes. The expression for the coefficient r_2 given in Eq.(4.46b) can now be simplified to the following.

$$r_2 = \frac{1}{1440}\left[\frac{\cos(4\theta) - \cos(4\beta)}{3 + \cos(4\theta)}\right] \tag{4.53}$$

The relative phase and local truncation errors of this scheme can be expressed as follows:

$$\frac{\|\xi^h\| - \|\xi^\beta\|}{\|\xi^\beta\|} = r_2\omega^2 + O\left(\omega^3\right) \quad ; \quad \mathbb{T} = -2r_2\omega^2 + O\left(\omega^3\right) \tag{4.54}$$

where r_2 is given in Eq.(4.53). So far the direction θ, along which the local truncation error is made zero, is arbitrary. We now try to optimize the solution error with respect to θ. Ideally the function to optimize could be either the relative phase or local truncation errors and the optimization problem can be posed as follows:

$$\min_\theta \max_\beta |\mathbb{T}| \quad \text{(or)} \quad \min_\theta \max_\beta \left|\frac{\|\xi^h\| - \|\xi^\beta\|}{\|\xi^\beta\|}\right| \tag{4.55}$$

Unfortunately, this is a difficult problem to solve in the closed form as it is a nonlinear function of ω and the location of the minimum might vary with ω. We conjuncture that in the pre-asymptotic range (i. e.$\xi_o \ell \ll 1$ or equivalently $\omega \ll 1$) the location of the minimum in the $\theta - \beta$ space is independent of ω. Thus, under this assumption the minimization of the relative phase or local truncation errors is essentially equivalent to the minimization of the coefficient of the lowest order term, i. e.here r_2. Hence we choose to optimize the coefficient r_2 instead. The redefined problem and its solution is given below.

$$\min_{\theta} \max_{\beta} |r_2| = \min_{\theta} \max_{\beta} \frac{|\cos(4\theta) - \cos(4\beta)|}{1440(3 + \cos(4\theta))} = \min_{\theta} \frac{1 + |\cos(4\theta)|}{1440(3 + \cos(4\theta))} = \frac{1}{4320} \tag{4.56a}$$

$$\begin{aligned} &\max_{\beta} \quad \text{occurs at} \quad |\cos(4\beta)| = 1 \quad \Rightarrow \beta = \frac{m\pi}{4} \qquad ; \quad m = \{0, 1, 2, \ldots\} \\ &\min_{\theta} \quad \text{occurs at} \quad |\cos(4\theta)| = 0 \quad \Rightarrow \theta = \frac{(2n+1)\pi}{8} \qquad ; \quad n = \{0, 1, 2, \ldots\} \end{aligned} \tag{4.56b}$$

Thus, for a given θ the maximum error in the stencil will be found for some $\beta \in \{0, (\pi/4), (\pi/2)\}$. That maximum error along the direction β takes a minimum value should the chosen direction (where the truncation error is made zero) be some $\theta \in \{(\pi/8, 3\pi/8)\}$. Note that due to the inherent symmetries in the stencil the expressions for $\alpha_{(\pi/8)}$ and $\alpha_{(3\pi/8)}$ are equivalent.

4.4.6 *Dispersion plots in* 2D

For a feasible graphical representation and comparison of the solutions to the characteristic equations we plot the $\xi_1 - \xi_2$ contours for some values of (ω_1, ω_2) only. Here and henceforth the superscripts $\{\theta, h\}$ are dropped in order to refer to the contour plots of both the continuous and discrete problems simultaneously. In order to retain generality to the plots the quantities $\omega_1, \omega_2, \xi_1^\theta, \xi_2^\theta, \xi_1^h$ and ξ_2^h are normalized as follows:

$$\omega_1^* := \frac{\omega_1}{\pi^2} \quad ; \quad \xi_1^{\theta *} := \frac{\xi_1^\theta}{\xi_1^{nq}} = \frac{\xi_1^\theta \ell_1}{\pi} \quad ; \quad \xi_1^{h*} := \frac{\xi_1^h}{\xi_1^{nq}} = \frac{\xi_1^h \ell_1}{\pi} \tag{4.57a}$$

$$\omega_2^* := \frac{\omega_2}{\pi^2} \quad ; \quad \xi_2^{\theta *} := \frac{\xi_2^\theta}{\xi_2^{nq}} = \frac{\xi_2^\theta \ell_2}{\pi} \quad ; \quad \xi_2^{h*} := \frac{\xi_2^h}{\xi_2^{nq}} = \frac{\xi_2^h \ell_2}{\pi} \tag{4.57b}$$

$$\Rightarrow \lambda_1 := e^{i\xi_1^h \ell_1} = e^{i\pi \xi_1^{h*}} \quad ; \quad \lambda_2 := e^{i\xi_2^h \ell_2} = e^{i\pi \xi_2^{h*}} \tag{4.57c}$$

where ξ_1^{nq}, ξ_2^{nq} are the Nyquist frequencies of the discretization along the 2D axes. Using these normalized quantities the characteristic equations of the continuous and

discrete problems given by Eq.(4.20) and Eq.(4.34) can be expressed as Eq.(4.58) and Eq.(4.59) respectively.

$$\frac{(\xi_1^{\theta*})^2}{\omega_1^*} + \frac{(\xi_2^{\theta*})^2}{\omega_2^*} = 1 \tag{4.58}$$

$$\left(\frac{[(1-\alpha_1)(\lambda_2^2+4\lambda_2+1)+6\alpha_1\lambda_2](-1+2\lambda_1-\lambda_1^2)}{6\pi^2\omega_1^*}\right) + \left(\frac{(-\lambda_2^2+2\lambda_2-1)[(1-\alpha_1)(1+4\lambda_1+\lambda_1^2)+6\alpha_1\lambda_1]}{6\pi^2\omega_2^*}\right) - \left(\frac{[(1-\alpha_2)(\lambda_2^2+4\lambda_2+1)(1+4\lambda_1+\lambda_1^2)+36\alpha_2\lambda_2\lambda_1]}{36}\right) = 0 \tag{4.59}$$

For every choice of the pair (ω_1^*, ω_2^*) the solution to Eq.(4.58) will trace elliptic contours with the center at the origin in the $\xi_1^* - \xi_2^*$ plane. Due to the inherent symmetry of the solutions the dispersion plots are presented just in the first quadrant. Similar to 1D, we require that the Nyquist frequencies of the discretization in 2D are always greater than the frequencies of the exact solution, i. e.$\min\{\xi_1^{nq}, \xi_2^{nq}\} \geqslant \xi_o$. Note that the following expressions are equivalent ($\equiv$): $\min\{\xi_1^{nq}, \xi_2^{nq}\} \equiv \max\{\ell_1, \ell_2\} \equiv \max\{\omega_1^*, \omega_2^*\}$. Thus restricting the domain to $\max\{\omega_1^*, \omega_2^*\} \in [0, 1]$ guarantees this requirement :

$$\omega^* := \max\{\omega_1^*, \omega_2^*\} \in [0, 1] \Leftrightarrow \min\{\xi_1^{nq}, \xi_2^{nq}\} \geqslant \xi_o \tag{4.60}$$

Likewise, a mesh resolution of at least 8 elements per wavelength is guaranteed by restricting the domain to $\omega^* \in [0, 1/16]$. We study the following four cases concerned with the choice of the (α_1, α_2) pair:

I: $\alpha_1 = \alpha_2 = (1/2)$. This case corresponds to the equation stencil associated with $\mathbf{A}^{0.5,0.5} := (\mathbf{A}^{fem} + \mathbf{A}^{fdm})/2$. Thus the discrete system obtained here is the average of the systems obtained from the Galerkin FEM and the classical FDM. Recall that we can also obtain this stencil using the generalized Padé approximation in 2D and choosing the parameter $\gamma = 2$.

II: $\alpha_1 = \alpha_2 = \alpha_\theta$ and $\theta = 0$. This case corresponds to the α-interpolation of the Galerkin FEM and the classical FDM. The local truncation error is zero along the direction $\theta = 0$ whenever $\ell_1 = \ell_2$.

III: $\alpha_1 = \alpha_2 = \alpha_\theta$ and $\theta = (\pi/8)$. This case also corresponds to the α-interpolation of the Galerkin FEM and the classical FDM. The local truncation error is zero along the direction $\theta = (\pi/8)$ whenever $\ell_1 = \ell_2$. Recall that choosing $\theta = (\pi/8)$ leads to an optimized expression for the coefficient r_2.

IV: QSFEM, $\alpha_1 \neq \alpha_2 \neq 0$ and given by Eq.(4.38) and Eq.(4.39). This case corresponds to the quasi-stabilized FEM presented in [10]. The local truncation error is zero along the directions $\theta = (\pi/16)$ and $\theta = (3\pi/16)$ whenever $\ell_1 = \ell_2$.

Note that for cases I,II and III the relative phase and local truncation errors of the numerical solution diminish at a fourth-order rate i.e $O\left((\xi_o\ell)^4\right)$ or equivalently $O\left(\omega^2\right)$. For the case IV, i.e the QSFEM, these errors diminish at a sixth-order rate i.e $O\left((\xi_o\ell)^6\right)$ or equivalently $O\left(\omega^3\right)$.

In Figures 62 and 63 we plot the solutions to the characteristic equations of the continuous and discrete problems given by Eq.(4.58) and Eq.(4.59) respectively. The contours of the continuous problem are drawn using the *dashed* line-style and the corresponding contour value displayed in a single text-box. Labeled *solid* line-style is used to display the contours of the discrete problem. Each figure is further divided into four sub-figures viz. (a)-(d) which correspond to the considered four cases I–IV. Within each sub-figure the contours plots of the continuous and discrete problems are plotted and compared. In Figure 62 we plot the $\xi_1^* - \xi_2^*$ contours keeping $\omega_1^* = \omega_2^*$ i. e.$\ell_1 = \ell_2$. The plotting domain considered here is $(\xi_1^*, \xi_2^*) = [0, 0.55] \times [0, 0.55]$. In Figure 63 we plot the $\xi_1^* - \xi_2^*$ contours keeping $\omega_2^* = 0.49\omega_1^*$ i. e.$\ell_2 = 0.7\ell_2$. The plotting domain considered here is again $(\xi_1^*, \xi_2^*) = [0, 0.55] \times [0, 0.55]$. In both the figures, contours are drawn for the values of $\omega^* \in \{(1/4), (1/9), (1/16), (1/25)\}$. These values of ω^* guarantee the presence of at least four, six, eight and ten elements per wavelength respectively. Note that except for the contour value $\omega^* = (1/4)$ in case I, the rest of the contours of the numerical solution are indistinguishable from their continuous counterparts. This is due to the fact that the relative local truncation error is of the order of `1e-3` which is small with respect to the scale of the plotting domain.

In order to quantify better the relative local truncation errors of the solutions, we compare them in Figure 64. This figure is further divided into four sub-figures viz. (a)-(d) which correspond to the considered four values of ω^* respectively, i. e.$\omega^* \in \{(1/4), (1/9), (1/16), (1/25)\}$. Within each sub-figure the relative local truncation errors of the considered four cases viz. I–IV are plotted vs. the direction β. Now we can clearly distinguish the errors related to the four cases. However in these figures the error associated with the case IV, i. e.the QSFEM is indistinguishable from zero at this scale. In Figure 65 the relative local truncation errors of the solutions are plotted in the log-scale. The sub-figures are organized just like in Figure 64. Note that in figures 64 and 65 the relative local truncation errors converge monotonically with respect to ω^*, i. e.the plots of the errors with respect to the direction β maintain their shape. This supports the conjuncture made in Section 4.4.5 that in the pre-asymptotic range the location of the mini-max error is independent of ω^*. Also we note that choosing $\theta = (\pi/8)$ in the expression for α_θ, the maximum error is less than the one choosing $\theta = 0$.

4.4.7 *Examples*

We consider the problem defined by Eq.(4.1) with the following problem data: $k \in \{\texttt{1e-3}, \texttt{1e-4}\}$, $s = 1$, $f = 0$ and the domain $\Omega = [0, 1] \times [0, 1]$. The Dirichlet boundary conditions are assigned such that the exact solution of Eq.(4.1) is $\phi(\mathbf{x}) = \sin(\boldsymbol{\xi}^\beta \cdot \mathbf{x})$, where β is the chosen direction of wave propagation, $\boldsymbol{\xi}^\beta := \xi_o(\cos(\beta), \sin(\beta))$ and $\xi_o := \sqrt{s/k}$. Thus for the chosen values of k, the wavenumber ξ_o takes the values in $\{10\sqrt{10}, 100\}$. The following wave directions are considered: $\beta \in \{(\pi/9), (\pi/4)\}$. Seven uniform meshes ($\ell_1 = \ell_2$) of different resolution are considered such that there are at least four, six, eight, ten, twelve, fourteen and sixteen elements per wavelength respectively. If the element length is chosen such that there are exactly n elements per wavelength, then the value of $\xi^* = (2/n)$ and $\omega^* = (2/n)^2$. As it can be seen all these

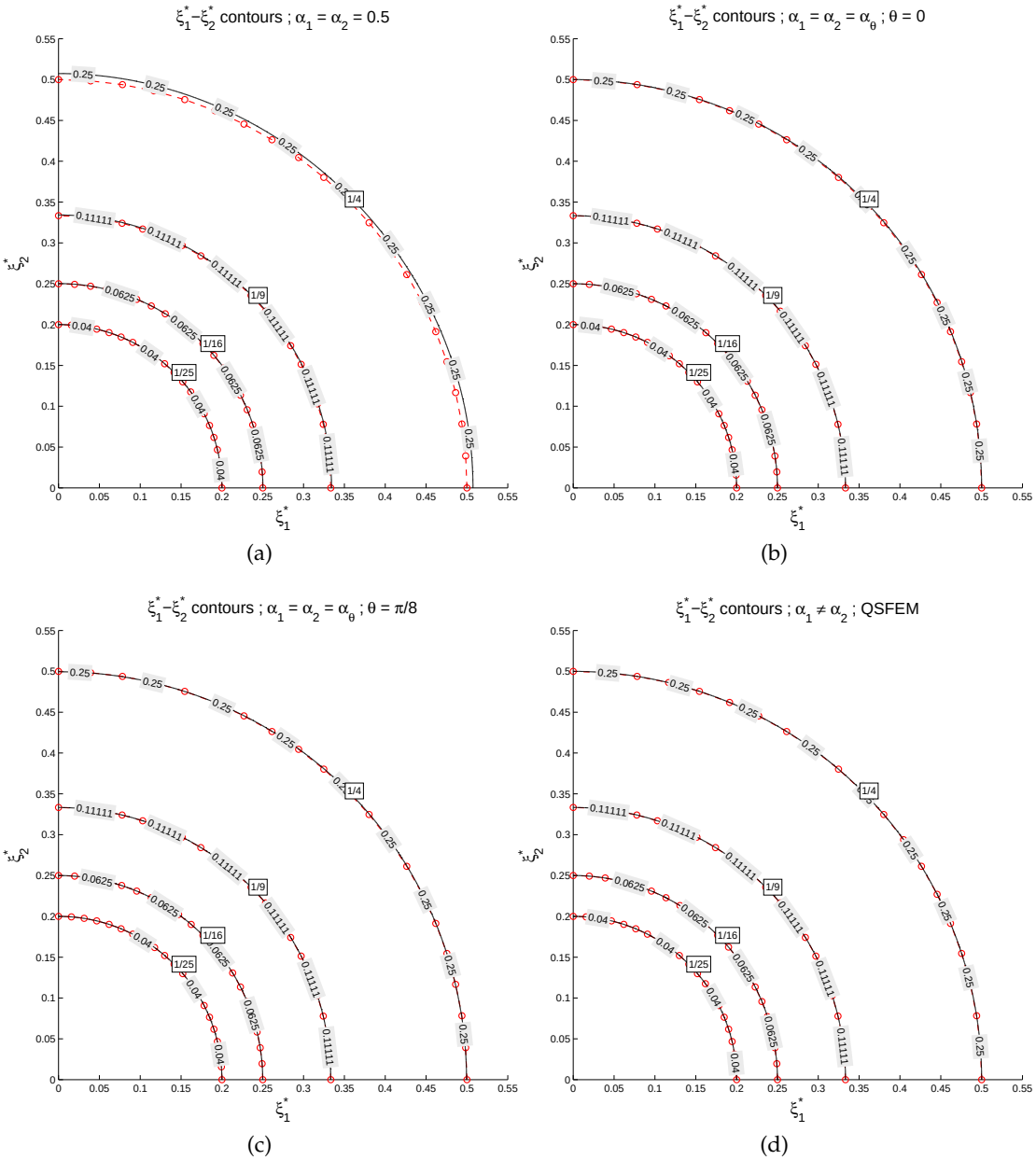


Figure 62: $\xi_1^* - \xi_2^*$ contours for $\omega^* \in \{(1/4), (1/9), (1/16), (1/25)\}$ and $\omega_1^* = \omega_2^*$. The *dashed* and *solid* line-styles correspond to the solutions of the continuous and discrete problems respectively. (a) Case I: $\alpha_1 = \alpha_2 = 0.5$; (b) Case II: $\alpha_1 = \alpha_2 = \alpha_\theta$ and $\theta = 0$; (c) Case III: $\alpha_1 = \alpha_2 = \alpha_\theta$ and $\theta = (\pi/8)$; (d) Case IV: QSFEM, $\alpha_1 \neq \alpha_2 \neq 0$ and given by Eq.(4.38) and Eq.(4.39)

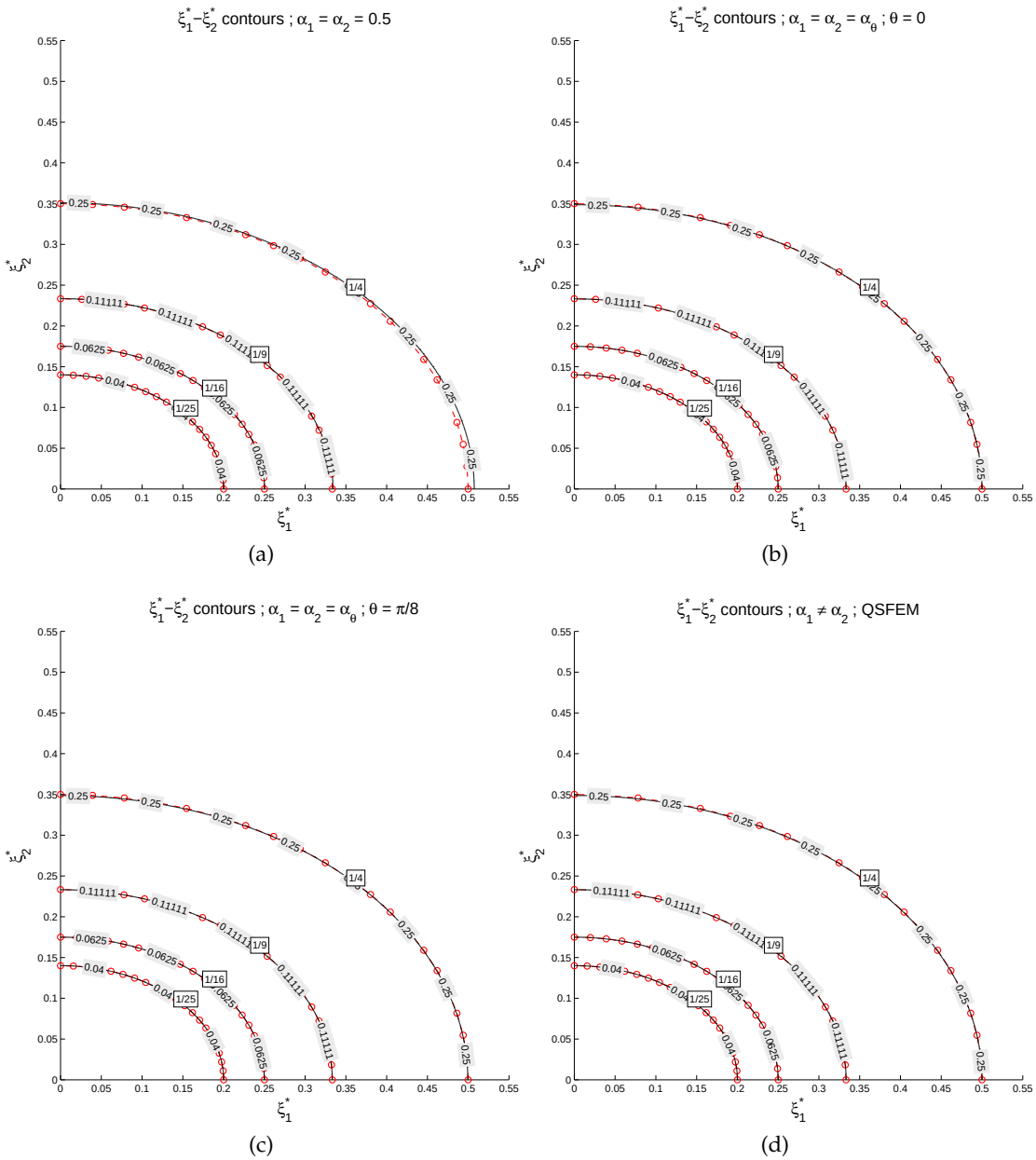


Figure 63: $\xi_1^* - \xi_2^*$ contours for $\omega^* \in \{(1/4), (1/9), (1/16), (1/25)\}$ and $\omega_2^* = 0.49\omega_1^*$. The *dashed* and *solid* line-styles correspond to the solutions of the continuous and discrete problems respectively. (a) Case I: $\alpha_1 = \alpha_2 = 0.5$; (b) Case II: $\alpha_1 = \alpha_2 = \alpha_\theta$ and $\theta = 0$; (c) Case III: $\alpha_1 = \alpha_2 = \alpha_\theta$ and $\theta = (\pi/8)$; (d) Case IV: QSFEM, $\alpha_1 \neq \alpha_2 \neq 0$ and given by Eq.(4.38) and Eq.(4.39)

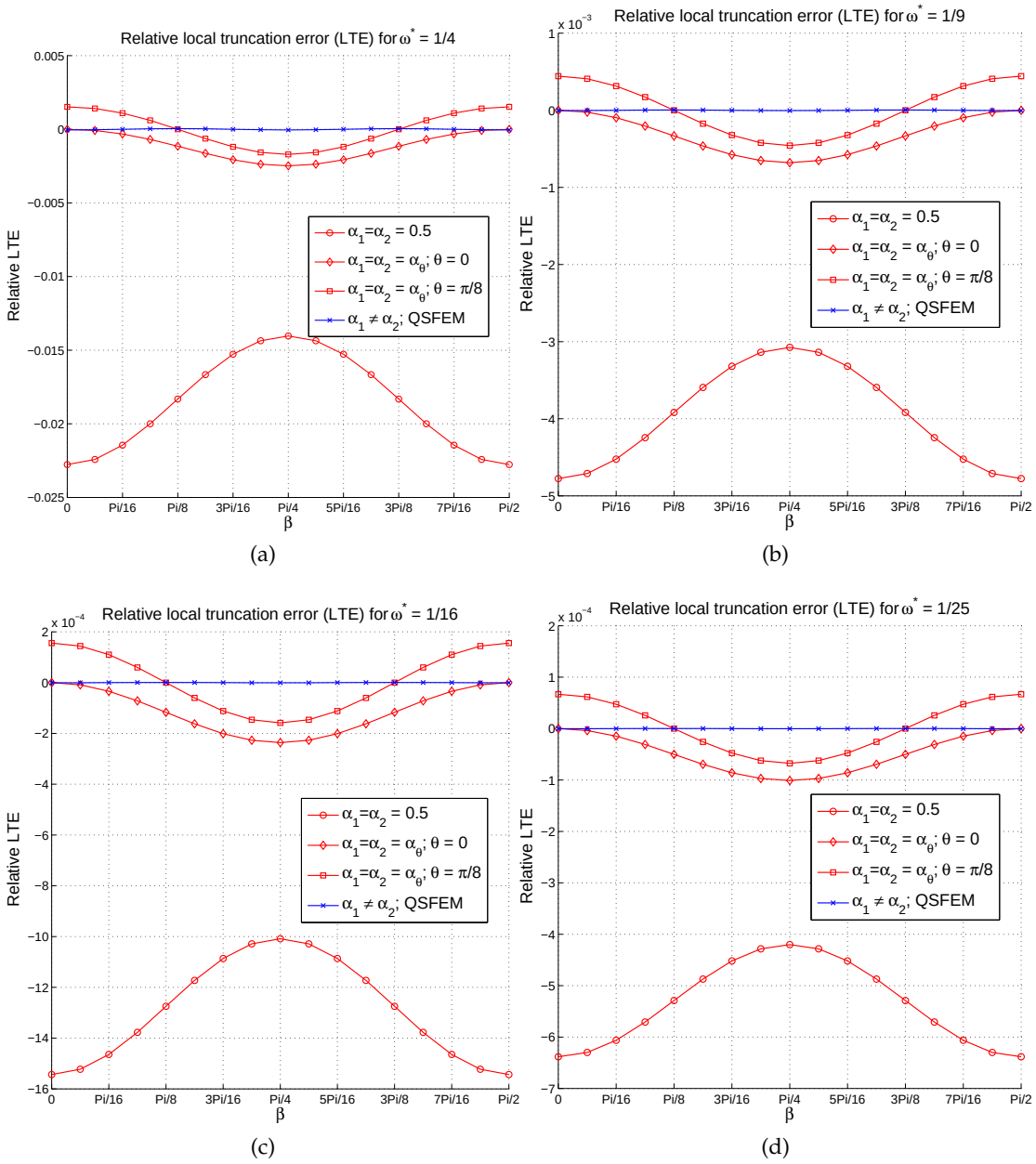


Figure 64: Relative local truncation error plots using $\ell_1 = \ell_2$. Comparisons are made for the considered four cases viz. I–IV and for (a) $\omega^* = (1/4)$; (b) $\omega^* = (1/9)$; (c) $\omega^* = (1/16)$; (d) $\omega^* = (1/25)$

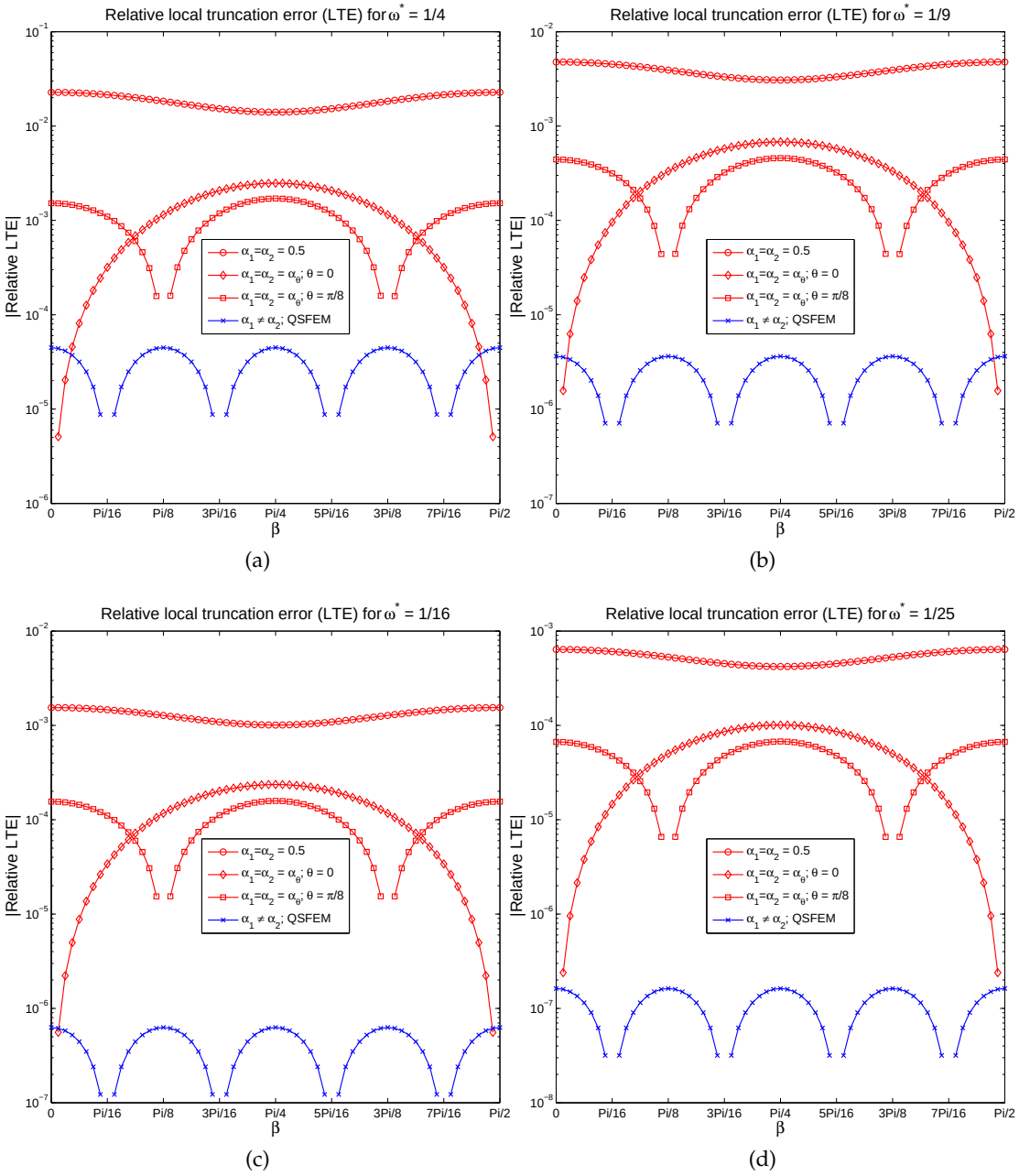


Figure 65: Log-scaled relative local truncation error plots using $\ell_1 = \ell_2$. Comparisons are made for the considered four cases viz. I–IV and for (a) $\omega^* = (1/4)$; (b) $\omega^* = (1/9)$; (c) $\omega^* = (1/16)$; (d) $\omega^* = (1/25)$

meshes restrict the domain of ω^* to $[0, 1/4]$. For these considerations we study the convergence of the relative error in the following error norms:

$$\text{L}^2 \text{ norm} \qquad \frac{\|\phi - \phi_h\|_0}{\|\phi\|_0} := \frac{[\int_\Omega (\phi - \phi_h)^2 \, d\Omega]^{1/2}}{[\int_\Omega \phi^2 \, d\Omega]^{1/2}} \tag{4.61a}$$

$$\text{H}^1 \text{ semi-norm} \qquad \frac{\|\phi - \phi_h\|_1}{\|\phi\|_1} := \frac{[\int_\Omega |\boldsymbol{\nabla}(\phi - \phi_h)|^2 \, d\Omega]^{1/2}}{[\int_\Omega |\boldsymbol{\nabla}\phi|^2 \, d\Omega]^{1/2}} \tag{4.61b}$$

$$\text{l}^\infty \text{ Euclidean norm} \qquad \frac{|\Phi_e - \Phi_h|_\infty}{|\Phi_e|_\infty} := \frac{\max_i |\Phi_e^i - \Phi_h^i|}{\max_i |\Phi_e^i|} \tag{4.61c}$$

In the convergence studies done here, the numerical solutions corresponding to the four cases viz. I–IV, are compared with the following solutions: the nodally exact interpolant denoted by $I_h\phi$ and the best approximations with respect to the L^2 norm and the H^1 semi-norm denoted by $P_h^0\phi$ and $P_h^1\phi$ respectively. The solutions $I_h\phi$, $P_h^0\phi$ and $P_h^1\phi$ can be found as shown in Eq.(4.62).

$$I_h\phi := N^a \Phi_e^a \tag{4.62a}$$

$$\int_\Omega w_h(\phi - P_h^0\phi) \, d\Omega = 0 \; \forall \; w_h \in V_0^h \Rightarrow \; \| \phi - P_h^0\phi \|_0 \leqslant \| \phi - \phi_h \|_0 \; \forall \; \phi_h \in V^h \tag{4.62b}$$

$$\int_\Omega \boldsymbol{\nabla} w_h \cdot \boldsymbol{\nabla}(\phi - P_h^1\phi) \, d\Omega = 0 \; \forall \; w_h \in V_0^h \Rightarrow \; \| \phi - P_h^1\phi \|_1 \leqslant \| \phi - \phi_h \|_1 \; \forall \; \phi_h \in V^h \tag{4.62c}$$

As here the exact solution is sinusoidal, we have used a third-order Gauss quadrature rule to evaluate the expressions in Eq.(4.61) and Eq.(4.62). Figures 66a and 66b illustrate the convergence of the relative error considering the wavenumber $\xi_o = 10\sqrt{10} \approx 31.62$, the wave direction $\beta = (\pi/9)$ and using the L^2 norm and H^1 semi-norm respectively. Figures 66c and 66d illustrate the same but now considering the wave direction $\beta = (\pi/4)$. Clearly the errors in the L^2 norm and H^1 semi-norm corresponding to all the cases are greater than that of the respective best approximations. The error lines corresponding to the cases II–IV show a convergence trend indistinguishable from the error line of $I_h\phi$. On coarse meshes the error line corresponding to case I deviates significantly from the error line of $I_h\phi$. Nevertheless, it quickly recovers the convergence trend of the later on finer meshes.

Figures 67a and 67b illustrate the convergence of the relative error considering the wavenumber $\xi_o = 100$, the wave direction $\beta = (\pi/9)$ and using the L^2-norm and H^1-seminorm respectively. Figures 67c and 67d illustrate the same but now considering the wave direction $\beta = (\pi/4)$. Note that a higher value of ξ_o does introduce the 'pollution-effect' in the error lines as they deviate more from the error line of $I_h\phi$. However the pollution effect is very small for cases II and III and is practically nil for case IV (sixth-order dispersion accuracy).

Figure 68 illustrates the convergence of the relative error in the l^∞ Euclidean norm. As a nodally exact solution requires that the dispersion error be zero, one may expect that the order of convergence in the l^∞ Euclidean norm be the same as that of the corresponding dispersion error. In fact the same is observed for the solutions corresponding to all the cases. The error lines of cases I–III converge at a fourth-order rate and that of case IV converges at a sixth-order rate. The error lines of the best

approximations in the L^2 norm ($P_h^0\phi$) and the H^1 semi-norm ($P_h^1\phi$) converge at a second-order rate. The relative error of $P_h^0\phi$ is always greater that that of $P_h^1\phi$. The pollution effect is now clearly visible for all the cases. Irrespective of the wave direction β, the error lines of all the cases shift higher with an increase in the wavenumber ξ_o. Meanwhile, the location of the error lines of $P_h^0\phi$ and $P_h^1\phi$ are practically unaffected by an increase in ξ_o (no pollution). As the magnitudes of the relative error in the l^∞ Euclidean norm for cases II–IV is small (with respect to relative error of $I_h\phi$ in the the L^2 norm and the H^1 semi-norm) for both the values of ξ_o, the pollution effect is hardly visible for these cases considering the relative error in the L^2 norm and the H^1 semi-norm.

Remark: As discussed in Section 4.3.4 and pointed out earlier in [47], though the discrete LBB constant in an average sense is inversely proportional to ξ_o, it has a more complicated behavior that tends its value to zero should ξ_o approach the zones of degeneracy (see Figure 59). Thus pollution effects may be found not only for higher wavenumbers but also in those situations where the wavenumber ξ_o approaches the zones of degeneracy. Of course, the higher the dispersion accuracy the closer will be the discrete eigenvalues to their continuous counterparts and narrower will be the zones of degeneracy. Also, if only the Dirichlet boundary conditions are prescribed (as is the case here), spurious amplitude and/or phase modulations might occur to satisfy them in spite of small dispersion errors [79]. For the presented scheme, we have found vestiges of this behavior along the wave direction $\beta = 0$.

4.5 CONCLUSIONS AND OUTLOOK

A fourth-order compact scheme on structured meshes is presented for the Helmholtz equation. The scheme consists in taking the α-interpolation of the Galerkin FEM and the classical FDM. For the 2D analysis of this scheme a generic nonstandard compact stencil involving two parameters α_1, α_2 is considered. In particular this nonstandard compact stencil can model the aforementioned scheme (choosing $\alpha_1 = \alpha_2 = \alpha$) and also the QSFEM which has a dispersion accuracy of sixth-order. The expression for the numerical solution of this nonstandard stencil is given considering generic expressions for α_1, α_2 written as a series expansion in terms of $\omega := (\xi_o \ell)^2$. Using this result, we provide the expressions for the phase and local truncation errors of this nonstandard compact stencil. In particular for our scheme it is shown that these errors diminish at the rate $O\left((\xi_o \ell)^4\right)$ or equivalently $O\left(\omega^2\right)$. An expression for the parameter α is given that minimizes the relative phase error in the pre-asymptotic range ($\xi_o \ell$ small). Also, by this choice the local truncation error of the scheme along the direction $\beta = (\pi/8)$ is made zero. Convergence studies of the relative error in the L^2 norm, the H^1 semi-norm and the l^∞ Euclidean norm are done and the pollution effect is found to be small. In particular, using the optimal expression for α the relative error of our scheme in the l^∞ Euclidean norm (for the considered examples and using at least ten elements per wavelength) is found to be around or less than one percent.

The abstractness in the definition of the QSFEM hinders its extension to unstructured meshes. This is a common problem faced by all the sixth-order methods proposed within the framework of the FDM. The recently proposed QOPG method addresses this issue and is able to attain a dispersion accuracy of the same order as

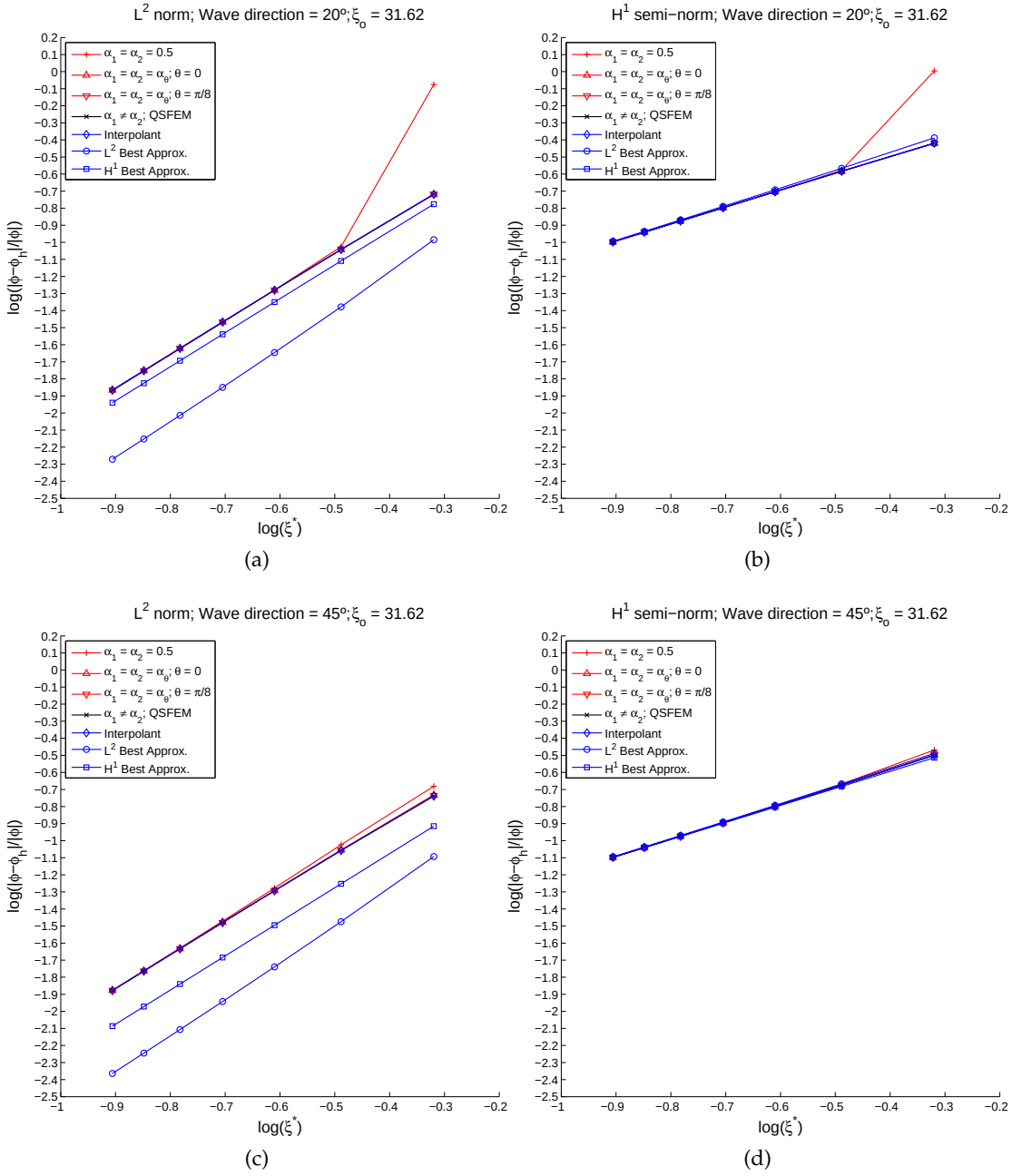


Figure 66: Convergence of the relative error considering $\xi_o = 10\sqrt{10}$ and for mesh resolutions that guarantee at least n elements per wavelength, where $n \in \{4, 6, 8, 10, 12, 14, 16\}$. The considered norms and the wave directions are: (a) L^2 norm, $\beta = (\pi/9)$; (b) H^1 semi-norm, $\beta = (\pi/9)$; (c) L^2 norm, $\beta = (\pi/4)$ and (d) H^1 semi-norm, $\beta = (\pi/4)$

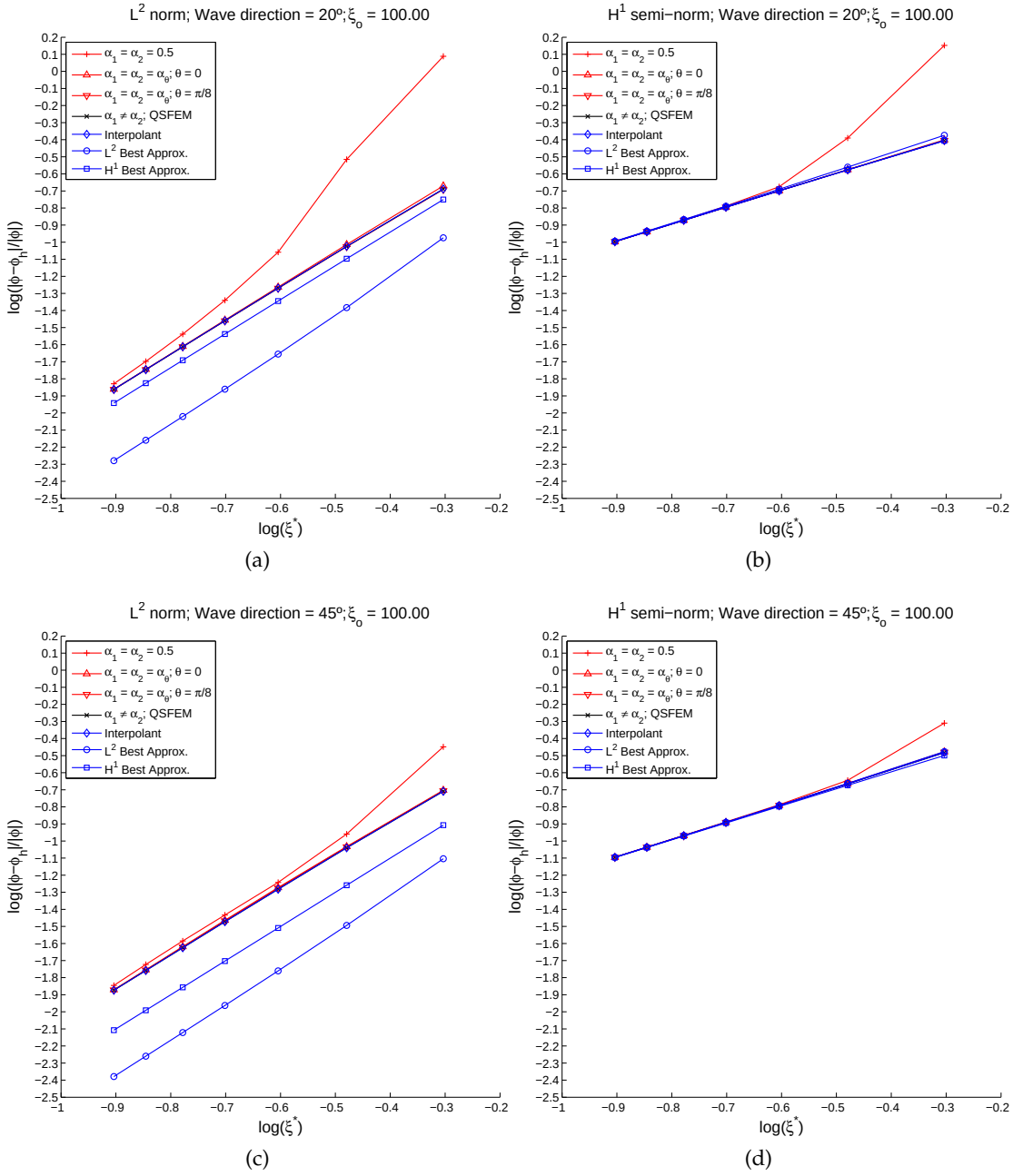


Figure 67: Convergence of the relative error considering $\xi_o = 100$ and for mesh resolutions that guarantee at least n elements per wavelength, where $n \in \{4, 6, 8, 10, 12, 14, 16\}$. The considered norms and the wave directions are: (a) L^2 norm, $\beta = (\pi/9)$; (b) H^1 semi-norm, $\beta = (\pi/9)$; (c) L^2 norm, $\beta = (\pi/4)$ and (d) H^1 semi-norm, $\beta = (\pi/4)$

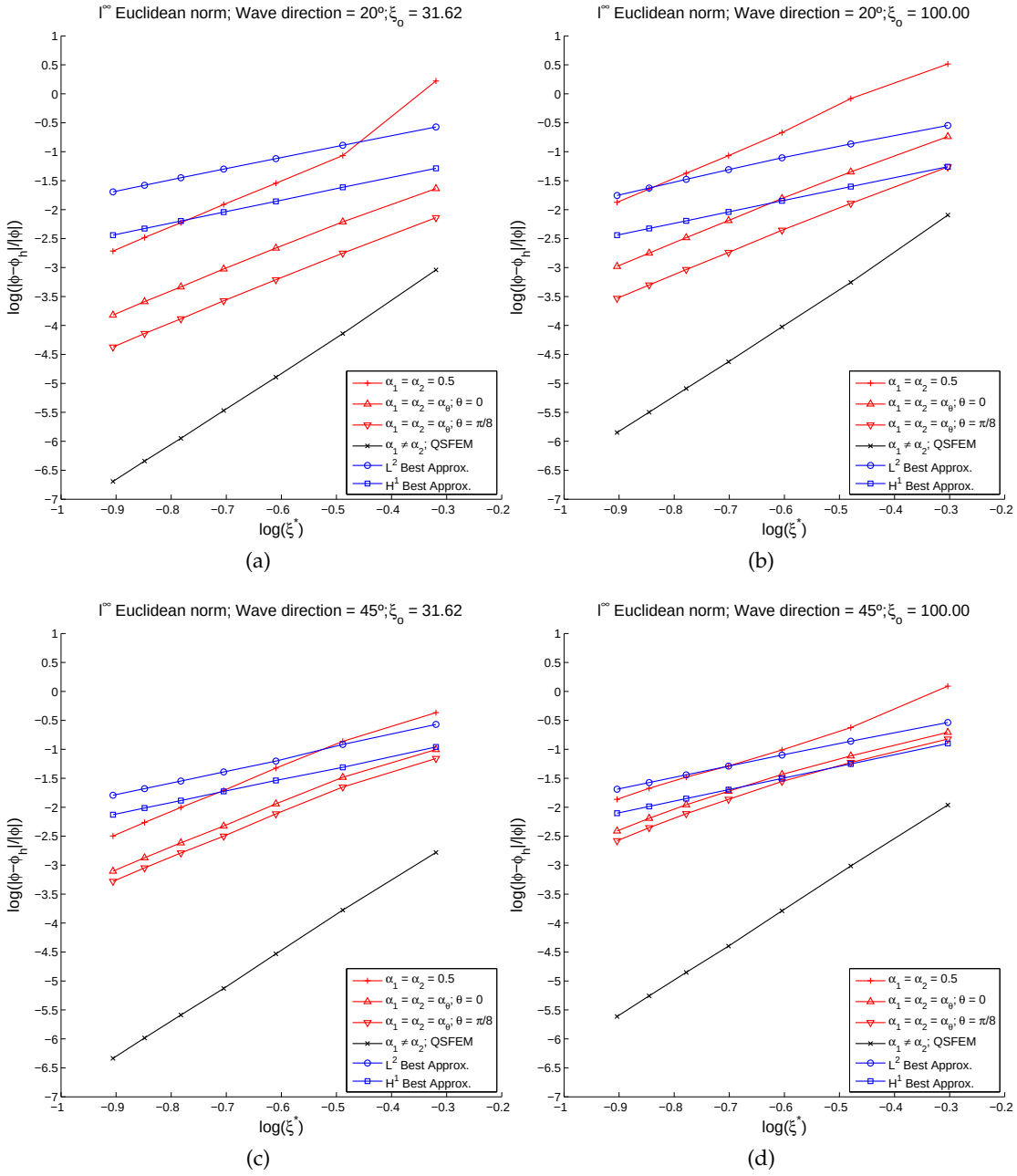


Figure 68: Convergence of the relative error in the l^∞ Euclidean norm using: (a) $\xi_o = 10\sqrt{10}$, $\beta = (\pi/9)$; (b) $\xi_o = 100$, $\beta = (\pi/9)$; (c) $\xi_o = 10\sqrt{10}$, $\beta = (\pi/4)$; (d) $\xi_o = 100$, $\beta = (\pi/4)$. The considered mesh resolutions guarantee at least n elements per wavelength, where $n \in \{4, 6, 8, 10, 12, 14, 16\}$.

the QSFEM on square meshes. Nevertheless it uses a quadratic bubble perturbation function defined over a macro-element and the parameters multiplying these bubbles are found by solving local optimization problems involving a functional of the local truncation error. Alternate methods that achieve this objective were proposed earlier within a variational setting and with similar implementation/computational cost, viz. the RBFEM [160], the DGB method [129], the GPR method [51] etc. Can this path to obtain the QSFEM be simplified? That is the outlook of this chapter.

Recall that the nonstandard compact stencil studied here has an additional structure that reduces its abstractness. This additional structure throws light on the extension of this stencil to unstructured meshes. In chapter 5 and [138] a new Petrov–Galerkin method involving two parameters viz. α_1, α_2 is presented which yields this nonstandard compact stencil on rectangular meshes. Making the two parameters equal, i. e. $\alpha_1 = \alpha_2 = \alpha$, we recover the compact stencil obtained by the α-interpolation of the Galerkin FEM and the classical central FDM. This Petrov–Galerkin method provides the counterparts of these two schemes on unstructured meshes and allows the treatment of natural boundary conditions (Neumann or Robin) and the source terms in a straight-forward manner. This we believe would open door to design higher-order Petrov–Galerkin methods which can be an alternative to the existing higher-order methods for the Helmholtz equation.

In general, ideas and experience shall interact permanently.

— Lothar Collatz.

5 PETROV–GALERKIN FORMULATION

5.1 INTRODUCTION

This chapter is a continuation of chapter 4[1] wherein a simple domain-based higher-order compact numerical scheme involving two parameters viz. α_1, α_2 was presented for the Helmholtz equation. The stencil obtained by choosing the parameters as distinct, i. e.$\alpha_1 \neq \alpha_2$ was denoted therein as the 'nonstandard compact stencil'. Taking $\alpha_1 = \alpha_2 = \alpha$, the nonstandard compact stencil simplifies to the α-interpolation of the equation stencils obtained by the Galerkin finite element method (FEM) and classical central finite difference method (FDM). For the Helmholtz equation, generic expressions for the parameters were given that guarantees a dispersion accuracy of sixth-order should $\alpha_1 \neq \alpha_2$ and fourth-order should $\alpha_1 = \alpha_2$. As the findings reported therein and the corresponding analysis was done for compact stencils, the contribution of the Galerkin FEM to the equation stencil corresponds to the choice of the lowest order rectangular block finite elements. By blocks we mean Cartesian product of intervals and by lowest order we refer to multilinear finite-element (FE) interpolation on these blocks. In this chapter we extend this scheme to unstructured meshes. We shall only consider the lowest order finite elements, i. e., linear FE interpolation on simplices and multilinear FE interpolation on blocks. The focus of this chapter is twofold: a) to design a Petrov–Galerkin (PG) method that reproduces on structured meshes the aforesaid numerical scheme and b) to study if this PG method using lowest order FE on unstructured meshes inherits the higher-order dispersion accuracy observed for the same on structured meshes.

Consider the multidimensional Helmholtz equation subjected to the following boundary conditions:

$$\mathcal{L}\phi := -\Delta\phi - \xi_o^2\phi = f(\mathbf{x}) \quad \text{in} \quad \Omega \tag{5.1a}$$

$$\phi = \phi^p \quad \text{on} \quad \Gamma_D \tag{5.1b}$$

$$\mathbf{n}\cdot\boldsymbol{\nabla}\phi - \mathcal{M}\phi = q \quad \text{on} \quad \Gamma_R \tag{5.1c}$$

where ξ_o is the wavenumber, $f(\mathbf{x})$ is the source term, ϕ^p is the prescribed value of ϕ on the Dirichlet boundary Γ_D. The operator $\mathcal{M}$ models either the Dirichlet-to-Neumann (DtN) map should the boundary-value problem (BVP) be posed on a domain with an exterior DtN boundary or the Neumann/Robin boundary conditions should the BVP be posed on an interior domain. Consider the following case: an interior BVP with only Dirichlet boundary conditions and $f(\mathbf{x}) = 0$; discretization of Eq.(5.1) is done by using either the Galerkin FEM or the FDM; a structured rectangular mesh/grid in 2D and bilinear FE interpolations on them are used. For the considered case we use the

1 Henceforth all references to chapter 4 will be cited as [137] which is the published version of the same.

following notation (described earlier in §3.1) to represent a generic compact stencil corresponding to any interior node (i,j) of a structured mesh/grid.

$$\{\circ^{j+1},\circ^{j},\circ^{j-1}\}\mathbf{A}\{\circ^{i-1},\circ^{i},\circ^{i+1}\}^{t}=0 \tag{5.2}$$

where $\mathbf{A}$ represents the matrix of the stencil coefficients. The equation stencil for the Galerkin FEM method corresponding to any interior node (i,j) can be written as Eq.(5.2) with the following definition of the stencil coefficient matrix ($\mathbf{A}$):

$$\begin{aligned}\mathbf{A}^{fem} := &\frac{\ell_2}{6\ell_1}\{1,4,1\}^{t}\{-1,2,-1\}+\frac{\ell_1}{6\ell_2}\{-1,2,-1\}^{t}\{1,4,1\}\\ &-\frac{\xi_o^2\ell_1\ell_2}{36}\{1,4,1\}^{t}\{1,4,1\}\end{aligned} \tag{5.3}$$

The stencil for the classical FDM method corresponding to any interior node (i,j) can be written as Eq.(5.2) with the following definition of $\mathbf{A}$:

$$\begin{aligned}\mathbf{A}^{fdm} := &\frac{\ell_2}{6\ell_1}\{0,6,0\}^{t}\{-1,2,-1\}+\frac{\ell_1}{6\ell_2}\{-1,2,-1\}^{t}\{0,6,0\}\\ &-\frac{\xi_o^2\ell_1\ell_2}{36}\{0,6,0\}^{t}\{0,6,0\}\end{aligned} \tag{5.4}$$

Naturally, the first step is to ascertain the feasibility of a PG method that reproduces on structured meshes the equation stencil corresponding to the classical FDM. Note that one may arrive at Eq.(5.4) from Eq.(5.3) by lumping the 1D mass type distributions within the equation stencil wherever they occur, i. e.replacing $\{1,4,1\}^{t}\{-1,2,-1\}$ by $\{0,6,0\}^{t}\{-1,2,-1\}$. Following this observation we may pose some questions which will mark the turning points of the current exposition: a) Is it possible to reproduce the mass matrix lumping technique within a PG setting ? b) Will the use of such test functions/weights for the current problem reproduce the FDM stencil and further the α-interpolation of the FEM and FDM stencils ? c) What is the possibility to recover the nonstandard compact stencil? and d) Will a direct extension of these weights to unstructured FEs result in a PG method that inherits the higher-order dispersion accuracy observed for the aforesaid numerical scheme ? These questions will be addressed in the subsequent sections.

This chapter is organized as follows. In Section 5.2 we recover the effect of mass lumping within a PG setting. The weights that incorporate this effect are also shown to lump the 1D mass type distributions within the equation stencil as discussed above. These weights are found to be discontinuous at the inter-element boundary and integration-by-parts needs to be done for an integral form of Eq.(5.1a) involving these weights. In Section 5.3 we present the variational setting needed to provide a well defined weak form for the problem. In Section 5.4.2 considering lowest order block FEs, we provide models for the weights on the inter-element boundaries that would recover the nonstandard compact stencil on structured meshes. Some examples are presented in Section 5.6 which illustrate the pollution effect through convergence studies in the L^2 norm, the H^1 semi-norm and the l^∞ Euclidean norms. Finally in Section 5.7 we arrive at some conclusions.

5.2 MASS LUMPING

Attempts to arrive at mass lumping procedures within a variational setting is not new. 'Legal' mass lumping emanating from residual-free bubbles had been presented in

[64, 65]. In this section we try to recover the mass matrix lumping process via a PG approach. In other words, the objective is to design the test functions W^a such that the following equation holds:

$$\int_\Omega W^a N^b \, d\Omega = \mathbf{M}_L^{ab} \tag{5.5}$$

where, $\mathbf{M}_L$ is the lumped mass matrix. For the sake of simplicity, the weights are designed to be piecewise polynomials of the same degree as that of the FE shape functions. This choice allows us to express the test functions in terms of the FE shape functions as:

$$W^a := \mathbb{W}^{ab} N^b \tag{5.6}$$

where, $\mathbb{W}$ is a matrix of constant coefficients. Using Eq.(5.6), the solution to the problem given by Eq.(5.5) can be expressed as follows.

$$\int_\Omega W^a N^b \, d\Omega = \mathbb{W}^{ac} \int_\Omega N^c N^b \, d\Omega = \mathbb{W}\mathbf{M} \quad ; \quad \mathbb{W}\mathbf{M} = \mathbf{M}_L \Rightarrow \boxed{\mathbb{W} = \mathbf{M}_L \mathbf{M}^{-1}} \tag{5.7}$$

The standard row lumping technique used to obtain $\mathbf{M}_L$ from the consistent mass matrix $\mathbf{M}$ can be expressed as follows.

$$\mathbf{M}_L^{ab} := \delta^{ab} \sum_c \mathbf{M}^{ac} = \delta^{ab} \sum_c \int_\Omega N^a N^c \, d\Omega = \delta^{ab} \int_\Omega N^a \, d\Omega \tag{5.8}$$

The fact that the finite element (FE) shape functions $\{N^a\}$ being a partition of unity is used to arrive at the last part of Eq.(5.8). If $\mathbf{M}_L$ be obtained via the row lumping technique, then the test functions defined by Eq.(5.6) also form a partition of unity. This statement can be verified as follows:

$$\begin{aligned} \sum_a W^a = \sum_a \mathbb{W}^{ab} N^b = \sum_a \mathbf{M}_L^{ac} \mathbf{M}^{-cb} N^b = \sum_a \mathbf{M}^{ac} \mathbf{M}^{-cb} N^b \\ = \sum_a \delta^{ab} N^b = \sum_a N^a = 1 \end{aligned} \tag{5.9a}$$

$$\boxed{\sum_a W^a = \sum_a N^a = 1} \tag{5.9b}$$

The process of defining test functions $\{W^a\}$ via Eqs.(5.6) and (5.7) is generic up to to the feasibility of a suitable lumping technique. Unfortunately, techniques other than the row lumping procedure generally will not render these test functions as a partition of unity. Thus, further developments in this line are restricted to those finite elements where the row lumping procedure makes sense, i. e.linear FE interpolation on simplices and multilinear FE interpolation on blocks. Thus for these FE discretization of the solution, PG test functions can be designed that would result in the mass lumping operation. Considering the fact that the test functions can model constants (as they are a partition of unity) and the exposition is done within the framework of PG methods, consistency and conservation properties are guaranteed.

The design problem for the test functions given by Eq.(5.5) may be posed either at the global or local level. Posing the problem at the global level results in weights that

Shape functions N^a	Test functions $\widetilde{W}^a$	Remarks
$\frac{1+\bar{\xi}^a\xi}{2}$	$\frac{1+3\bar{\xi}^a\xi}{2}$	1D linear FE. $\{\bar{\xi}^a\}=\{-1,1\}$
$\left(\frac{1+\bar{\xi}^a\xi}{2}\right)\left(\frac{1+\bar{\eta}^a\eta}{2}\right)$	$\left(\frac{1+3\bar{\xi}^a\xi}{2}\right)\left(\frac{1+3\bar{\eta}^a\eta}{2}\right)$	2D rectangular bilinear FE. $\{\bar{\xi}^a\}=\{-1,1,1,-1)\}$, $\{\bar{\eta}^a\}=\{-1,-1,1,1)\}$
$\{(1-\xi-\eta),\xi,\eta\}$	$\{(3-4\xi-4\eta),(4\xi-1),(4\eta-1)\}$	2D linear triangle FE.

Table 3: Test functions corresponding to some finite elements

are piecewise continuous (C^0 functions) but with a non-local support. This observation is straightforward given the following facts: the matrix $\mathbb{W}$ is full and the shape functions $\{N^a\}$ being piecewise continuous. Thus the sparse structure of the resulting algebraic system is lost. Also the space spanned by these test functions will always be non-zero on the domain boundary. This poses difficulty in providing a simple and well-defined weak formulation for the problem subjected to Dirichlet boundary conditions.

On the other hand, posing the problem at the local (element) level will result in weights with a local support but with a loss in C^0 continuity at the element edges. Note that in this approach the global weights (assembled in a piecewise manner) are no longer a linear combination of the global FE shape functions. Henceforth the restriction of the test functions W^a to the element interiors and edges will be denoted by $\widetilde{W}^a$ and $\widehat{W}^a$ respectively. We remark that in this approach the solution to Eq.(5.5) posed at the element level will be used only to define $\widetilde{W}^a$. This is done for two reasons, viz. a) to ensure that the test functions be a partition of unity on the element edges and b) to be able to model $\widehat{W}^a$ such that we recover the sparsity pattern of the Galerkin FEM. The later condition also allows us to construct test function spaces that vanish at the Dirichlet boundary.

In this work we have opted for the later approach, i. e.the design problem for the weights is posed at the element level. The weights $\widetilde{W}^a$ corresponding to three different element types is listed in Table 3. Figure 69 illustrates the construction of the weights corresponding to the 1D linear FE shape functions. Note the loss of C^0 continuity at the element edges in Fig.69b.

Consider the diffusion term $\int_K \boldsymbol{\nabla}\widetilde{W}^a \cdot \boldsymbol{\nabla} N^b \, d\Omega$ for an element K of a rectangular FE mesh in 2D. Using the test functions $\widetilde{W}^a$ defined in Table 3 we get,

$$\int_K \boldsymbol{\nabla}\widetilde{W}^a \cdot \boldsymbol{\nabla} N^b \, d\Omega = \int_K \frac{\partial \widetilde{W}^a}{\partial x}\frac{\partial N^b}{\partial x} \, d\Omega + \int_K \frac{\partial \widetilde{W}^a}{\partial y}\frac{\partial N^b}{\partial y} \, d\Omega \tag{5.10a}$$

$$= \frac{3\ell_2}{2\ell_1}\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix} + \frac{3\ell_1}{2\ell_2}\begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \tag{5.10b}$$

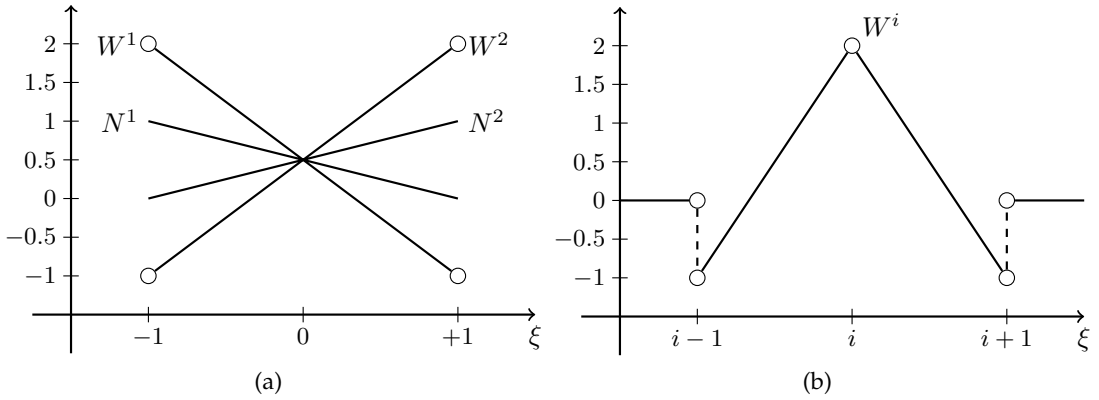


Figure 69: Test functions corresponding to the 1D linear FE that results in mass lumping. (a) Comparison of the weights $\{W^a\}$ within an element and with respect to the 1D FE shape functions $\{N^a\}$. (b) Illustration of the weight corresponding to an arbitrary node i assembled element-wise. The open circles in these illustrations signify that the weights have not yet been defined at these points.

The stencil coefficient matrix $\mathbf{A}^d$ corresponding to the assembly of the above element matrices can be expressed as follows:

$$\mathbf{A}^d = \frac{3\ell_2}{\ell_1}\{0,1,0\}^t\{-1,2,-1\} + \frac{3\ell_1}{\ell_2}\{-1,2,-1\}^t\{0,1,0\} \tag{5.11}$$

Note that except for a scalar multiple (here '3') the stencil coefficient matrix $\mathbf{A}^d$ is the same as that corresponding to the diffusive term in $\mathbf{A}^{fdm}$. $\mathbf{A}^d$ differs here by a scalar multiple because we have not yet considered the jumps in the values of the test functions at the element edges. Nevertheless this example sheds light on the first two questions raised in Section 5.1, i. e.using these weights the 1D mass type distributions found in the equation stencils are lumped, thus opening way to design FEM based PG methods that would yield the FDM stencil. Naturally, the next step is to ascertain the feasibility of a model for $\widehat{W}^a$ which when used together with $\widetilde{W}^a$ within a well-defined variational setting would end up in the classical FDM.

Recall that to arrive at the form $\int_K \boldsymbol{\nabla}\widetilde{W}^a \cdot \boldsymbol{\nabla}N^b \, d\Omega$ integration by parts needs to be done for an integral form of Eq.(5.1a) containing discontinuous test functions. This is the distinction of the current work from existing stabilized FEM based PG methods that follows the theoretical framework originally proposed for the Streamline–Upwind/Petrov–Galerkin (SUPG) method [92]. Thus the framework of Discontinuous–Galerkin (DG) methods is most appropriate to provide a variational setting for the current work. The distinction (apart from the trivial one[2]) of the current work with the DG methods is illustrated via a schematic representation of the same in Fig. 70. Fig. 70a illustrates a generic DG method. Recall that the weights on either side of an element edge in a DG method are not only discontinuous but also independent. The same applies to the trial solutions (ϕ) and in addition to this, models $\widehat{\phi}$ for ϕ are specified on the element edges. For conservative DG methods, $\widehat{\phi}$ which is sometimes named as the *scalar numerical flux*, is single valued on the element edges [4]. On the other hand, Fig. 70b illustrates the current PG method. Note that the test functions

2 DG belongs to the class of Galerkin methods where the test functions and trial solutions belong to the same function space

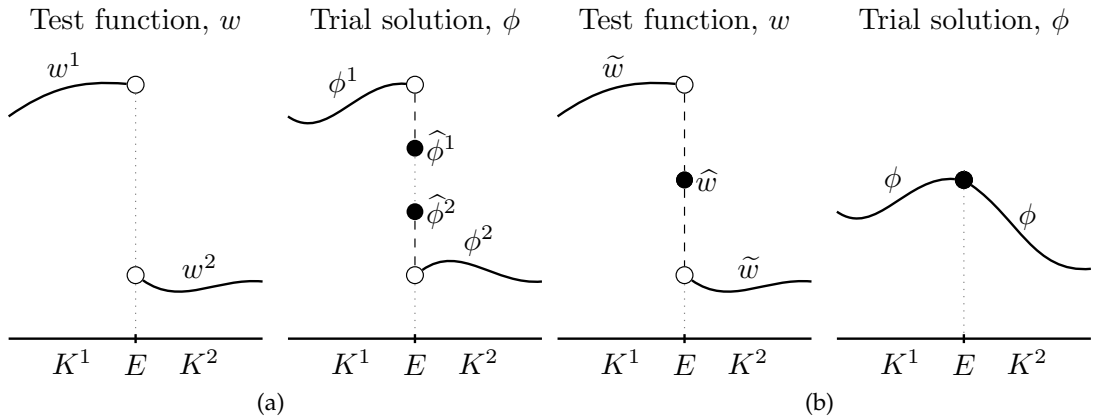


Figure 70: Comparison of the test function w and trial solution ϕ of a generic Discontinuous–Galerkin (DG) method with those of the current Petrov–Galerkin (PG) method. Schematic representations of w and ϕ for (a) a DG method and (b) the current PG method. Note that unlike for the DG method, the w and ϕ for the current PG method are not independent on either sides of the edge E.

(w) remain discontinuous but they are no longer independent. In addition to this, we also specify single-valued models $\widehat{w}$ for w on the element edges. The trial solutions for the current PG method are the standard FE solutions which are C^0-continuous and are not independent on either sides of the element edge. In the next section, following the framework of DG methods and the notations used therein [4], we present the variational setting for the current PG method.

5.3 VARIATIONAL SETTING

In this section we present the variational setting for PG methods with discontinuous test functions (C^{-1} functions) and standard finite-element trial solutions (C^0 functions) which can be schematically represented as in Fig. 70b. First, we recall some standard notations which are used in the developments that follows. Let $\mathcal{T}_h = \{K\}$ be a regular family of elements K generating a partition of Ω and the summation over all elements will be indicated by $\sum_K$. The piecewise integral $\sum_K \int_K$ will be denoted by $\int_{\mathcal{T}_h}$. The collection of all element edges (including edges on the boundary) will be written as $\mathcal{E}_h = \{E\}$. The set of internal and boundary edges will be denoted by $\mathcal{E}_h^o = \{E_o\}$ and $\mathcal{E}_h^\partial = \{E_\partial\}$ respectively. The piecewise integral $\sum_E \int_E$ will be written as $\int_{\mathcal{E}_h}$, using $\int_{\mathcal{E}_h^o}$ and $\int_{\mathcal{E}_h^\partial}$ when the edges are interior or on the boundary respectively. The restriction of a function φ to K is denoted by $\varphi|_K$.

Suppose now that the elements K^1 and K^2 share an element edge E, and let $\mathbf{n}^1$ and $\mathbf{n}^2$ be the normals to E exterior to K^1 and K^2 respectively. For an arbitrary scalar function φ, possibly discontinuous across E, the jump and the average operators are defined as follows.

$$[\![\varphi]\!] := \mathbf{n}^1 \varphi|_{\partial K^1 \cap E} + \mathbf{n}^2 \varphi|_{\partial K^2 \cap E} \tag{5.12a}$$

$$\{\varphi\} := \frac{1}{2}\left(\varphi|_{\partial K^1 \cap E} + \varphi|_{\partial K^2 \cap E}\right) \tag{5.12b}$$

whereas for an arbitrary vector function $\boldsymbol{\sigma}$ these operators are defined as follows.

$$[\![\boldsymbol{\sigma}]\!] := \mathbf{n}^1 \cdot \boldsymbol{\sigma}|_{\partial K^1 \cap E} + \mathbf{n}^2 \cdot \boldsymbol{\sigma}|_{\partial K^2 \cap E} \tag{5.13a}$$

$$\{\boldsymbol{\sigma}\} := \frac{1}{2}\left(\boldsymbol{\sigma}|_{\partial K^1 \cap E} + \boldsymbol{\sigma}|_{\partial K^2 \cap E}\right) \tag{5.13b}$$

Note that $\{\varphi\}$ and $[\![\boldsymbol{\sigma}]\!]$ are scalar quantities while $[\![\varphi]\!]$ and $\{\boldsymbol{\sigma}\}$ are vectors perpendicular to the edge E. To deal with the sums of the form $\sum_K \int_{\partial K} \varphi(\mathbf{n} \cdot \boldsymbol{\sigma})$, we use the average and jump operators defined in Eq.(5.12) and Eq.(5.13). A straightforward computation leads to the following formula:

$$\sum_K \int_{\partial K} \varphi(\mathbf{n} \cdot \boldsymbol{\sigma}) \, d\Gamma = \int_{\mathcal{E}_h^o} \left(\{\varphi\}[\![\boldsymbol{\sigma}]\!] + [\![\varphi]\!] \cdot \{\boldsymbol{\sigma}\}\right) d\Gamma + \int_{\mathcal{E}_h^\partial} \varphi(\mathbf{n} \cdot \boldsymbol{\sigma}) \, d\Gamma \tag{5.14}$$

Consider an arbitrary element K with boundary ∂K and define two sub-domains within it viz. K_o and K_ε as shown in Fig 71a. The boundary that K_ε shares with K_o is denoted by ∂K_o. The external normals to ∂K and ∂K_o are denoted by $\mathbf{n}$ and $\mathbf{n}^{o+}$ respectively. The normal $\mathbf{n}^{o-} := -\mathbf{n}^{o+}$. The piecewise integral $\int_{K_o} + \int_{K_\varepsilon}$ will be denoted by $\int_{\mathcal{K}_h}$. The free parameter ε characterizes the width of the K_ε sub-domain. As shown in Fig. 71b we split the definition of the test function w over the sub-domains as follows:

$$w|_{K_o} := \widetilde{w} \quad ; \quad w|_{K_\varepsilon} := \tau \tag{5.15}$$

Likewise, we split the definition of w on the respective boundaries as follows:

$$w|_{\partial K_o} := \widetilde{w}|_{\partial K_o} = \tau|_{\partial K_o} \quad ; \quad w|_{\partial K} := \widehat{w} \tag{5.16}$$

Thus, as shown in Fig. 71b, letting $\varepsilon \to 0$ we arrive at a class of test functions which were represented schematically in Fig. 70b. Following this line, the strategy to present the weak form of the problem would be to first present it assuming ε to be finite and then taking the limit $\varepsilon \to 0$.

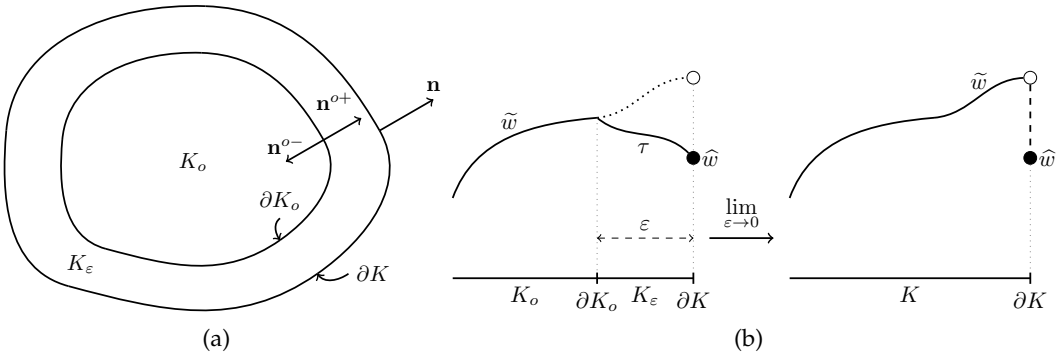


Figure 71: Schematic diagrams of an arbitrary element $K \in \mathcal{T}_h$ and a test function defined over it. (a) The element K is further divided into two sub-domains K_o and K_ε. (b) The test function w is defined piecewise as follows: $\widetilde{w}$ over K_o, τ over K_ε, $\widetilde{w}|_{\partial K_o} = \tau|_{\partial K_o}$ and $\widehat{w}$ on ∂K. The free parameter ε characterizes the width of the K_ε domain.

We now proceed to present a well-defined weak formulation of the continuous problem 5.1 wherein the test functions could possibly be discontinuous across $\mathcal{E}_h$. Multiplying Eq.(5.1a) by test functions[3] w and integrating formally on K, we get

$$\int_K w(\mathcal{L}\phi - f)\,d\Omega = -\int_K w\Delta\phi\,d\Omega - \int_K w(\xi_o^2\phi + f)\,d\Omega \tag{5.17a}$$

$$= -\int_{K_o} \widetilde{w}\Delta\phi\,d\Omega - \int_{K_\varepsilon} \tau\Delta\phi\,d\Omega - \int_K w(\xi_o^2\phi + f)\,d\Omega \tag{5.17b}$$

Integrating by parts the terms $\widetilde{w}\Delta\phi$ and $\tau\Delta\phi$ we get,

$$\int_K w(\mathcal{L}\phi - f)\,d\Omega = \int_{K_o} (\boldsymbol{\nabla}\widetilde{w}\cdot\boldsymbol{\nabla}\phi)\,d\Omega + \int_{K_\varepsilon} (\boldsymbol{\nabla}\tau\cdot\boldsymbol{\nabla}\phi)\,d\Omega - \int_K w(\xi_o^2\phi + f)\,d\Omega$$
$$- \int_{\partial K_o} \widetilde{w}(\mathbf{n}^{o+}\cdot\boldsymbol{\nabla}\phi)\,d\Gamma - \int_{\partial K_o} \tau(\mathbf{n}^{o-}\cdot\boldsymbol{\nabla}\phi)\,d\Gamma - \int_{\partial K} \widehat{w}(\mathbf{n}\cdot\boldsymbol{\nabla}\phi)\,d\Gamma \tag{5.18}$$

Using the formula given by Eq.(5.14) on the boundary ∂K_o we get,

$$\int_K w(\mathcal{L}\phi - f)\,d\Omega = \int_{\mathcal{K}_h} (\boldsymbol{\nabla}w\cdot\boldsymbol{\nabla}\phi)\,d\Omega - \int_K w(\xi_o^2\phi + f)\phi\,d\Omega$$
$$- \int_{\partial K_o} \Big(\{w\}[\![\boldsymbol{\nabla}\phi]\!] + [\![w]\!]\cdot\{\boldsymbol{\nabla}\phi\}\Big)\,d\Gamma - \int_{\partial K} \widehat{w}(\mathbf{n}\cdot\boldsymbol{\nabla}\phi)\,d\Gamma \tag{5.19}$$

The variational statement of the problem obtained by using the method of weighted residuals is: Find $\phi \in U$, such that $\forall w \in V$ we have

$$\int_\Omega w(\mathcal{L}\phi - f)\,d\Omega + \int_{\Gamma_R} \widehat{w}(\mathbf{n}\cdot\boldsymbol{\nabla}\phi - \mathcal{M}\phi - q)\,d\Gamma = 0 \tag{5.20}$$

Using Eq.(5.19) and the following equivalent ($\equiv$) expressions: $\int_\Omega \equiv \int_{\mathcal{T}_h} := \sum_K \int_K$ and $\int_{\mathcal{E}_h^\partial} \widehat{w}\,(\star)\,d\Gamma \equiv \int_{\Gamma_R} \widehat{w}\,(\star)\,d\Gamma$, the variational statement can be rewritten as follows:

$$\int_{\mathcal{T}_h} w(\xi_o^2\phi + f)\,d\Omega + \int_{\Gamma_R} \widehat{w}(\mathcal{M}\phi + q)\,d\Gamma = \sum_K \int_{\mathcal{K}_h} (\boldsymbol{\nabla}w\cdot\boldsymbol{\nabla}\phi)\,d\Omega$$
$$- \sum_K \int_{\partial K_o} \Big(\{w\}[\![\boldsymbol{\nabla}\phi]\!] + [\![w]\!]\cdot\{\boldsymbol{\nabla}\phi\}\Big)\,d\Gamma - \int_{\mathcal{E}_h^o} \Big(\{\widehat{w}\}[\![\boldsymbol{\nabla}\phi]\!] + [\![\widehat{w}]\!]\cdot\{\boldsymbol{\nabla}\phi\}\Big)\,d\Gamma \tag{5.21}$$

We obtain the weak form of the continuous problem 5.1 by invoking the continuity requirement $[\![\boldsymbol{\nabla}\phi|_{\mathcal{E}_h}]\!] = 0$ for the solution ϕ in Eq.(5.21). At this point we also invoke the continuity requirements for the considered test functions w in Eq.(5.21), i. e. $[\![w|_{\partial K_o}]\!] = 0$ and $[\![\widehat{w}]\!] = 0$. The weak form thus obtained is specific to this particular choice of test functions.

$$\int_{\mathcal{T}_h} w(\xi_o^2\phi + f)\,d\Omega + \int_{\Gamma_R} \widehat{w}(\mathcal{M}\phi + q)\,d\Gamma = \sum_K \int_{\mathcal{K}_h} (\boldsymbol{\nabla}w\cdot\boldsymbol{\nabla}\phi)\,d\Omega \tag{5.22}$$

After any appropriate discretization, the weak form of the discretized problem is obtained by replacing the continuous unknowns by their discrete counterparts,

$$\int_{\mathcal{T}_h} w_h(\xi_o^2\phi_h + f)\,d\Omega + \int_{\Gamma_R} \widehat{w}_h(\mathcal{M}\phi_h + q)\,d\Gamma = \sum_K \int_{\mathcal{K}_h} (\boldsymbol{\nabla}w_h\cdot\boldsymbol{\nabla}\phi_h)\,d\Omega \tag{5.23}$$

3 Whenever the solution ϕ is complex-valued, the complex-conjugate of w is used instead.

In the above discrete weak form the continuity constraint $[\![\nabla\phi_h|_{\mathcal{E}_h^o}]\!] = 0$ is enforced in a weak way. Consider the term $\nabla w_h \cdot \nabla\phi_h$ in Eq.(5.23) and re-integrating by parts both the terms in the piecewise integral $\int_{\mathcal{K}_h}$ we get,

$$\int_{\mathcal{K}_h} (\nabla w_h \cdot \nabla\phi_h)\, d\Omega = \int_{K_o} (\nabla\widetilde{w}_h \cdot \nabla\phi_h)\, d\Omega + \int_{K_\varepsilon} (\nabla\tau_h \cdot \nabla\phi_h)\, d\Omega \tag{5.24a}$$

$$= -\int_{K_o} \widetilde{w}_h \Delta\phi_h\, d\Omega - \int_{K_\varepsilon} \tau_h \Delta\phi_h\, d\Omega + \int_{\partial K} \widehat{w}_h (\mathbf{n} \cdot \nabla\phi_h)\, d\Gamma \tag{5.24b}$$

$$= -\int_{\mathcal{K}_h} w_h \Delta\phi_h\, d\Omega + \int_{\partial K} \widehat{w}_h (\mathbf{n} \cdot \nabla\phi_h)\, d\Gamma \tag{5.24c}$$

To arrive at Eq.(5.24b) we have used the following results for the boundary ∂K_o: $[\![w_h|_{\partial K_o}]\!] = 0$ and $[\![\nabla\phi_h|_{\partial K_o}]\!] = 0$. Using the result in Eq.(5.16) and re-integrating by parts only the $\nabla\tau_h \cdot \nabla\phi_h$ term in the piecewise integral $\int_{\mathcal{K}_h}$ of Eq.(5.23) we get,

$$\begin{aligned}\int_{\mathcal{K}_h} (\nabla w_h \cdot \nabla\phi_h)\, d\Omega = \int_{K_o} (\nabla\widetilde{w}_h \cdot \nabla\phi_h)\, d\Omega - \int_{K_\varepsilon} \tau_h \Delta\phi_h\, d\Omega \\ - \int_{\partial K_o} \widetilde{w}_h (\mathbf{n}^{o+} \cdot \nabla\phi_h)\, d\Gamma + \int_{\partial K} \widehat{w}_h (\mathbf{n} \cdot \nabla\phi_h)\, d\Gamma\end{aligned} \tag{5.25}$$

As the parameter $\varepsilon \to 0$ we see that $\mathbf{n}^{o+} \to \mathbf{n}$, $\mathbf{n}^{o-} \to -\mathbf{n}$ and $\partial K_o \to \partial K$. Also note that $\forall\varphi \in L^2(K)$, $\int_{K_o} \varphi\, d\Omega \to \int_K \varphi\, d\Omega$ and $\int_{K_\varepsilon} \varphi\, d\Omega \to 0$. Thus, taking the limit $\varepsilon \to 0$, we have:

$$\lim_{\varepsilon\to 0} \int_{\mathcal{K}_h} (\nabla w_h \cdot \nabla\phi_h)\, d\Omega = \int_K (\nabla\widetilde{w}_h \cdot \nabla\phi_h)\, d\Omega + \lim_{\varepsilon\to 0} \int_{K_\varepsilon} (\nabla\tau_h \cdot \nabla\phi_h)\, d\Omega \tag{5.26a}$$

$$= -\int_K \widetilde{w}_h \Delta\phi_h\, d\Omega + \int_{\partial K} \widehat{w}_h (\mathbf{n} \cdot \nabla\phi_h)\, d\Gamma \tag{5.26b}$$

$$= \int_K (\nabla\widetilde{w}_h \cdot \nabla\phi_h)\, d\Omega + \int_{\partial K} (\widehat{w}_h - \widetilde{w}_h)(\mathbf{n} \cdot \nabla\phi_h)\, d\Gamma \tag{5.26c}$$

Thus in the limit $\varepsilon \to 0$, the discrete weak form given by Eq.(5.23) can be expressed equivalently by the following forms:

$$\int_{\mathcal{T}_h} w_h(\xi_o^2 \phi_h + f)\, d\Omega + \int_{\Gamma_R} \widehat{w}_h(\mathcal{M}\phi_h + q)\, d\Gamma = \sum_K \lim_{\varepsilon\to 0} \int_{\mathcal{K}_h} (\boldsymbol{\nabla} w_h \cdot \boldsymbol{\nabla}\phi_h)\, d\Omega \tag{5.27a}$$

$$= \int_{\mathcal{T}_h} (\boldsymbol{\nabla}\widetilde{w}_h \cdot \boldsymbol{\nabla}\phi_h)\, d\Omega + \sum_K \lim_{\varepsilon\to 0} \int_{K_\varepsilon} (\boldsymbol{\nabla}\tau_h \cdot \boldsymbol{\nabla}\phi_h)\, d\Omega \tag{5.27b}$$

$$= \boxed{-\int_{\mathcal{T}_h} \widetilde{w}_h \Delta\phi_h\, d\Omega + \sum_K \int_{\partial K} \widehat{w}_h (\mathbf{n}\cdot\boldsymbol{\nabla}\phi_h)\, d\Gamma} \tag{5.27c}$$

$$= -\int_{\mathcal{T}_h} \widetilde{w}_h \Delta\phi_h\, d\Omega + \int_{\mathcal{E}_h^o} \widehat{w}_h [\![\boldsymbol{\nabla}\phi_h]\!]\, d\Gamma + \int_{\Gamma_R} \widehat{w}_h (\mathbf{n}\cdot\boldsymbol{\nabla}\phi_h)\, d\Gamma \tag{5.27d}$$

$$= \boxed{\int_{\mathcal{T}_h} (\boldsymbol{\nabla}\widetilde{w}_h \cdot \boldsymbol{\nabla}\phi_h)\, d\Omega + \sum_K \int_{\partial K} (\widehat{w}_h - \widetilde{w}_h)(\mathbf{n}\cdot\boldsymbol{\nabla}\phi_h)\, d\Gamma} \tag{5.27e}$$

$$\begin{aligned} = \int_{\mathcal{T}_h} (\boldsymbol{\nabla}\widetilde{w}_h \cdot \boldsymbol{\nabla}\phi_h)\, d\Omega - \int_{\mathcal{E}_h^o} \left(\{\widetilde{w}_h - \widehat{w}_h\}[\![\boldsymbol{\nabla}\phi_h]\!] + [\![\widetilde{w}_h]\!]\cdot\{\boldsymbol{\nabla}\phi_h\}\right) d\Gamma \\ - \int_{\mathcal{E}_h^\partial} \widetilde{w}_h (\mathbf{n}\cdot\boldsymbol{\nabla}\phi_h)\, d\Gamma + \int_{\Gamma_R} \widehat{w}_h (\mathbf{n}\cdot\boldsymbol{\nabla}\phi_h)\, d\Gamma \end{aligned} \tag{5.27f}$$

Note that in the limit $\varepsilon \to 0$ the test functions w_h develop sharp Dirac-type layers at the element boundaries $\mathcal{E}_h$. Hence the integral $\int_{K_\varepsilon}$ that appears in Eq.(5.27a) and Eq.(5.27b) does not vanish as $\varepsilon \to 0$. On the other hand, from Eq.(5.27c) we note that the sparsity of the resulting discrete system depends on the employed model for $\widehat{w}_h$. Clearly, if the value of $\widehat{w}_h$ be designed to be zero on the boundary of a patch of elements associated to any given node, then from Eq.(5.27c) we see that the resulting discrete system will have a sparsity pattern equivalent to that of the Galerkin FEM. Also note that on a generic block finite element, the $\widetilde{w}_h \Delta\phi_h$ term will not vanish and needs to be evaluated. Nevertheless, using the weak form expressed as in Eq.(5.27e) the extra labor just involves the evaluation of the element boundary integrals. This can be easily incorporated within an 'assemble-by-elements' data structure.

5.4 BLOCK FINITE ELEMENTS

5.4.1 *1D linear FE*

In this section we will provide models for the PG weights on the elements edges. Consider the 1D linear FE and corresponding PG weights specification illustrated in Fig. 69. In Fig. 69b, let $\widehat{W^i}|_{i-1}, \widehat{W^i}|_i$ and $\widehat{W^i}|_{i+1}$ be the models for the weight W^i on the edges $i-1, i$ and $i+1$ respectively. For these weights W^i to be a partition of unity also on the element edges the following relation should hold:

$$\widehat{W^i}|_{i-1} + \widehat{W^i}|_i + \widehat{W^i}|_{i+1} = 1 \tag{5.28}$$

There exists an infinity of solutions for Eq.(5.28) but only the choice $\{\widehat{W^i}|_{i-1}, \widehat{W^i}|_i, \widehat{W^i}|_{i+1}\} = \{0, 1, 0\}$ will result in a discrete system that has the same sparsity structure as that of the Galerkin FEM or the classical FDM. Also, the space spanned by these

weights can be restricted to zero on the Dirichlet boundary without being trivially zero inside the domain and thus, justifying their admittance in weak formulations. Using the later definition the PG weights corresponding to the 1D linear FEs can be represented as shown in Fig. 72

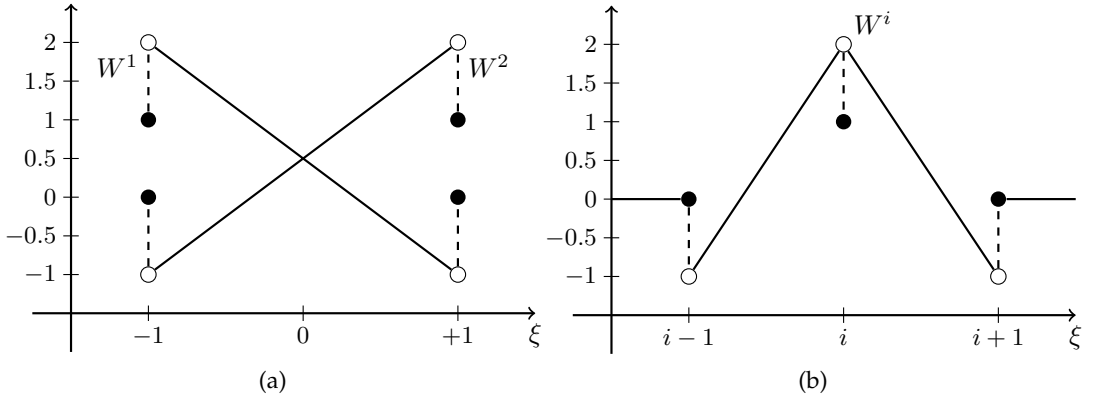


Figure 72: A model for the PG weights on the element edges corresponding to the 1D linear FE. (a) Illustration of the weights $\{W^a\}$ within an element. (b) The weight corresponding to an arbitrary node i assembled element-wise. The filled circles in these illustrations represent the chosen model for the weights on the element edges.

Discretization of the space by finite elements will lead to the approximation $\phi_h = N^a \Phi^a$ and using the PG weights (shown in Fig. 72) in any of the weak forms presented in Eq.(5.27), the following equation stencil is obtained:

$$\left(\frac{1}{\ell}\right)(-\Phi^{i-1} + 2\Phi^i - \Phi^{i+1}) - \xi_o^2 \ell \Phi^i = 0 \tag{5.29}$$

This is precisely the equation stencil obtained by using either the classical FDM or the Galerkin FEM wherein the mass matrix is lumped. Define a free parameter α and consider the following definition of the PG weights within an element:

$$W_\alpha^a = \begin{cases} \alpha\left(\frac{1+3\xi\bar{\xi}^a}{2}\right) + (1-\alpha)\left(\frac{1+\xi\bar{\xi}^a}{2}\right) = \left(\frac{1+(1+2\alpha)\xi\bar{\xi}^a}{2}\right) & -1 < \xi < 1 \\ \left(\frac{1+\xi\bar{\xi}^a}{2}\right) & \xi = \pm 1 \end{cases} \tag{5.30}$$

In the above equation choosing $\alpha = 0$ we will recover the weights W_0^a as the standard 1D FE shape functions N^a. Likewise, choosing $\alpha = 1$, the obtained weights W_1^a is the same as shown in Fig. 72. Using the PG weights defined by Eq.(5.30) in any of the weak forms presented in Eq.(5.27), the following equation stencil is obtained:

$$\begin{aligned}(1-\alpha)\left[\left(\frac{1}{\ell}\right)(-\Phi^{i-1} + 2\Phi^i - \Phi^{i+1}) - \left(\frac{\xi_o^2\ell}{6}\right)(\Phi^{i-1} + 4\Phi^i + \Phi^{i+1})\right] \\ + \alpha\left[\left(\frac{1}{\ell}\right)(-\Phi^{i-1} + 2\Phi^i - \Phi^{i+1}) - \xi_o^2\ell\Phi^i\right] = 0\end{aligned} \tag{5.31}$$

$$\begin{aligned}\Rightarrow \left(\frac{1}{\ell}\right)(-\Phi^{i-1} + 2\Phi^i - \Phi^{i+1}) - (1-\alpha)\left(\frac{\xi_o^2\ell}{6}\right)(\Phi^{i-1} + 4\Phi^i + \Phi^{i+1}) \\ - \alpha\xi_o^2\ell\Phi^i = 0\end{aligned} \tag{5.32}$$

Equation (5.31) is precisely the α-interpolation of the stencils obtained by the Galerkin FEM and the classical FDM methods in 1D. For the 1D case using linear FE and as shown in Eq.(5.32), it is equivalent to the Galerkin FEM wherein an alpha-interpolated mass matrix is used.

5.4.2 2D *bilinear FE*

Consider now the 2D bilinear rectangular FE and define the PG weights over these blocks as the Cartesian product of its 1D counterparts defined earlier in Eq.(5.30). Thus we get,

$$W_\alpha^a = \begin{cases} \left(\frac{1+(1+2\alpha)\xi\bar{\xi}^a}{2}\right)\left(\frac{1+(1+2\alpha)\eta\bar{\eta}^a}{2}\right) & (\xi,\eta) \in (-1,1)\times(-1,1) \\ \left(\frac{1+(1+2\alpha)\xi\bar{\xi}^a}{2}\right)\left(\frac{1+\eta\bar{\eta}^a}{2}\right) & (\xi,\eta) \in (-1,1)\times\{\pm 1\} \\ \left(\frac{1+\xi\bar{\xi}^a}{2}\right)\left(\frac{1+(1+2\alpha)\eta\bar{\eta}^a}{2}\right) & (\xi,\eta) \in \{\pm 1\}\times(-1,1) \\ \left(\frac{1+\xi\bar{\xi}^a}{2}\right)\left(\frac{1+\eta\bar{\eta}^a}{2}\right) & (\xi,\eta) \in \{\pm 1\}\times\{\pm 1\} \end{cases} \tag{5.33}$$

Note that on the element edges whenever $N^a = 0$, we have simultaneously $\widehat{W}^a = 0$. Likewise, on the edges whenever the expression for α is single-valued, we have simultaneously a single-valued model for $\widehat{W}^a$. In this way it is possible to retain the sparsity pattern of the Galerkin FEM. On a structured mesh in 2D and using the PG weights defined by Eq.(5.33) in any of the weak forms presented in Eq.(5.27), the stencil corresponding to any interior node (i,j) can be written as Eq.(5.2) with the following definition of the stencil coefficient matrix ($\mathbf{A}$):

$$\begin{aligned} \mathbf{A}^\alpha := {} & (1-\alpha)\frac{\ell_2}{6\ell_1}\{1,4,1\}^t\{-1,2,-1\} + \alpha\frac{\ell_2}{6\ell_1}\{0,6,0\}^t\{-1,2,-1\} \\ & + (1-\alpha)\frac{\ell_1}{6\ell_2}\{-1,2,-1\}^t\{1,4,1\} + \alpha\frac{\ell_1}{6\ell_2}\{-1,2,-1\}^t\{0,6,0\} \\ & - (1-\alpha)\frac{\xi_o^2\ell_1\ell_2}{36}\{1,4,1\}^t\{1,4,1\} - \alpha\frac{\xi_o^2\ell_1\ell_2}{36}\{0,6,0\}^t\{0,6,0\} \end{aligned} \tag{5.34}$$

This is precisely the α-interpolation of the FEM and FDM stencils in 2D, i.e.$\mathbf{A}^\alpha = (1-\alpha)\mathbf{A}^{fem} + \alpha\mathbf{A}^{fdm}$. Unlike in 1D where we had a unique way to model $\widehat{W}^a$ so as to retain the sparsity pattern of the Galerkin-FEM, in 2D many alternatives models exits. However, all acceptable models for $\widehat{W}^a$ have to be a partition of unity for every element and be single-valued on the element edges. In the following we show one of the many possible models for $\widehat{W}^a$ and for an unstructured 2D bilinear block FE. Let α_1, α_2 be two free parameters and consider the following definition for the PG weights:

$$W_{\alpha_1,\alpha_2}^a = \begin{cases} \widetilde{W}_{\alpha_2}^a := \mathbb{W}_{\alpha_2}^{ab} N^b & (\xi,\eta) \in (-1,1)\times(-1,1) \\ \left(\frac{1+(1+2\alpha_1)\xi\bar{\xi}^a}{2}\right)\left(\frac{1+\eta\bar{\eta}^a}{2}\right) & (\xi,\eta) \in (-1,1)\times\{\pm 1\} \\ \left(\frac{1+\xi\bar{\xi}^a}{2}\right)\left(\frac{1+(1+2\alpha_1)\eta\bar{\eta}^a}{2}\right) & (\xi,\eta) \in \{\pm 1\}\times(-1,1) \\ \left(\frac{1+\xi\bar{\xi}^a}{2}\right)\left(\frac{1+\eta\bar{\eta}^a}{2}\right) & (\xi,\eta) \in \{\pm 1\}\times\{\pm 1\} \end{cases} \tag{5.35}$$

where $\mathbb{W}^{ab}_{\alpha_2} := (1-\alpha_2)\delta^{ab} + \alpha_2 \mathbb{W}^{ab}$, δ^{ab} is the Kronecker delta and $\mathbb{W}^{ab}$ is the matrix of constant coefficients given by Eq.(5.7). Note that $\widehat{W}^a$ is designed to be a partition of unity. Choosing a single valued expression for α_1 on each element edge implies a single-valued model for $\widehat{W}^a$ on the same. Thus, should any length scale appear within the expression for α_1, then it should be proportional to the corresponding edge length. On a structured mesh in 2D and using the PG weights defined by Eq.(5.35) in any of the weak forms presented in Eq.(5.27), the stencil corresponding to any interior node (i,j) can be written as Eq.(5.2) with the following definition of the stencil coefficient matrix ($\mathbf{A}$):

$$\begin{aligned}\mathbf{A}^{\alpha_1,\alpha_2} := (1-\alpha_1)\frac{\ell_2}{6\ell_1}\{1,4,1\}^t\{-1,2,-1\} + \alpha_1\frac{\ell_2}{6\ell_1}\{0,6,0\}^t\{-1,2,-1\} \\ + (1-\alpha_1)\frac{\ell_1}{6\ell_2}\{-1,2,-1\}^t\{1,4,1\} + \alpha_1\frac{\ell_1}{6\ell_2}\{-1,2,-1\}^t\{0,6,0\} \\ - (1-\alpha_2)\frac{\xi_o^2\ell_1\ell_2}{36}\{1,4,1\}^t\{1,4,1\} - \alpha_2\frac{\xi_o^2\ell_1\ell_2}{36}\{0,6,0\}^t\{0,6,0\}\end{aligned} \tag{5.36}$$

This is precisely the nonstandard compact stencil presented in [137]. Note that taking $\alpha_1 = \alpha_2 = \alpha$ we recover the stencil given by Eq.(5.34), i. e.the α-interpolation of the FEM and FDM stencils in 2D: $\mathbf{A}^\alpha = (1-\alpha)\mathbf{A}^{fem} + \alpha\mathbf{A}^{fdm}$. Choosing $\alpha_1 = \alpha_2 = 0.5$ we get a stencil that is the average of the FEM and FDM stencils in 2D and can be shown to be equal [137] to the stencil obtained by the generalized fourth-order compact Padé approximation [81, 168] (therein using the parameter $\gamma = 2$). Likewise taking $\alpha_1 = 0$ and $\alpha_2 = \alpha$ we arrive at a stencil that results from the Galerkin FEM method using an α-interpolated mass matrix $\mathbf{M}^\alpha := (1-\alpha)\mathbf{M} + \alpha\mathbf{M}_L$.

5.4.3 *Stabilization Parameters*

Considering square meshes/grids (i. e.$\ell_1 = \ell_2$) the parameters α_1 and α_2 that appear in $\mathbf{A}^{\alpha_1,\alpha_2}$ can be chosen such that the numerical solution be sixth-order accurate, i. e.$O\left((\xi_o\ell)^6\right)$ or equivalently $O\left(\omega^3\right)$. Recall that this is the maximum order of accuracy that can be attained on any compact stencil [10]. All such α_1 and α_2 should obey the following series expansion in terms of ω [137].

$$\alpha_1 = \frac{1}{2} - \frac{\omega}{60} + \sum_{m=2}^{\infty} a_m\omega^m \quad ; \quad \alpha_2 = \frac{1}{2} - \frac{\omega}{40} + \sum_{m=2}^{\infty} b_m\omega^m \tag{5.37}$$

where a_m, b_m are coefficients independent of ω. The relative phase $\mathbb{P}$ and local truncation $\mathbb{T}$ errors of these schemes can be expressed as follows:

$$\mathbb{P} = r_3\omega^3 + O\left(\omega^4\right) \quad ; \quad \mathbb{T} = -2r_3\omega^3 + O\left(\omega^4\right) \tag{5.38}$$

$$r_3 = \left[\frac{5}{110592} - \left(\frac{a_2 - 4b_2}{48}\right) + \left(\frac{1+576a_2}{27648}\right)\cos(4\beta) + \frac{\cos(8\beta)}{774144}\right] \tag{5.39}$$

As a_m, b_m $(m \geqslant 2)$ can be chosen arbitrarily, infinitely many sixth-order schemes can be designed through $\mathbf{A}^{\alpha_1,\alpha_2}$. Of course some particular choice of a_m, b_m may yield a scheme with better features. For instance, a_m, b_m may be chosen such that the local truncation error $\mathbb{T}$ be zero along some chosen directions. Choosing $\alpha_1 = \alpha_2 = \alpha$, i. e.$a_m = b_m$ $\forall m$ we recover the α-interpolation of the Galerkin FEM and

FDM. Further details on the choice of the parameters to recover various stencils can be found in [137].

As most of the expressions for α_1, α_2 optimized for square meshes need not be optimal for unstructured meshes, in the current work we consider only the simplest expressions that would guarantee fourth-order (Eq.(5.40a)) and sixth-order (Eq.(5.40b)) dispersion accuracy on square meshes. On unstructured meshes the expressions for α_1, α_2 corresponding to these two choices can be written as follows:

$$\alpha_1 = \alpha_2 = \frac{1}{2} \tag{5.40a}$$

$$\alpha_1 = \frac{1}{2} - \frac{\widehat{\omega}}{60} \quad ; \quad \alpha_2 = \frac{1}{2} - \frac{\widetilde{\omega}}{40} \tag{5.40b}$$

where $\widehat{\omega} := (\xi_o \widehat{\ell})^2$ and $\widetilde{\omega} := (\xi_o \widetilde{\ell})^2$. $\widehat{\ell}$ and $\widetilde{\ell}$ represent the models used for the length measures corresponding to the element edges and the interior. In the current study for each element we have chosen $\widehat{\ell}$ equal to the edge length (will vary from edge to edge) and $\widetilde{\ell}$ equal to the maximum edge length. Note that using this model, α_1 is always single-valued on the edges. On square meshes using Eq.(5.40b) we recover α_1, α_2 as given in Eq.(5.37) up to the first two terms which is sufficient to attain sixth-order dispersion accuracy.

5.5 SIMPLICIAL FINITE ELEMENTS

Consider a rectangular domain discretized by structured simplicial FEs. Such discretization would typically yield stencils as shown in figure 73. The stencils with the hypotenuse oriented along left and right are labeled using the markers $o = l$ and $o = r$ respectively.

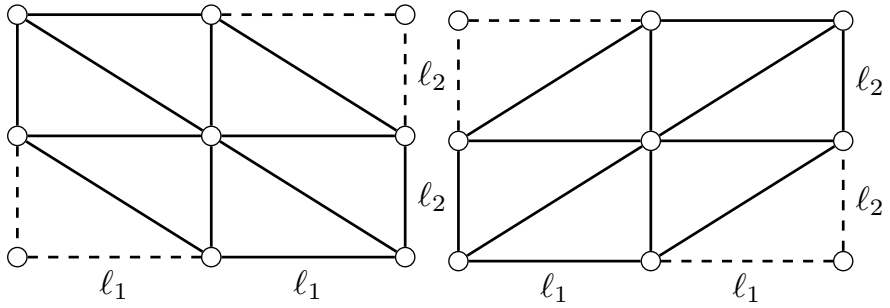


Figure 73: Stencils obtained by using a structured simplicial finite element mesh with the hypotenuse oriented/tilted along (a) left, i.e. $o = l$ and (b) right , i.e. $o = r$. 'o' is a flag that indicates the stencil tilt.

The equation stencil for the Galerkin FEM corresponding to any interior node (i, j) can be written as Eq.(5.2) with the following definition of stencil coefficient matrix:

$$\mathbf{A}^{fem} = \frac{\ell_2}{\ell_1}\begin{bmatrix} 0 & 0 & 0 \\ -1 & 2 & -1 \\ 0 & 0 & 0 \end{bmatrix} + \frac{\ell_1}{\ell_2}\begin{bmatrix} 0 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 0 \end{bmatrix} - \frac{\xi_o^2 \ell_1 \ell_2}{12}\begin{bmatrix} \delta_{ol} & 1 & \delta_{or} \\ 1 & 6 & 1 \\ \delta_{or} & 1 & \delta_{ol} \end{bmatrix} \tag{5.41}$$

where δ_{ol} and δ_{or} are Kronecker deltas and 'o' is a flag that indicates the stencil tilt. Note that using simplicial FEs the contribution of the diffusion term in Eq.(5.41) is identical to that obtained in the FDM stencil given by Eq.(5.4). Thus, the stencil obtained via an α-interpolation of the Galerkin FEM and the FDM stencils will lead to the following stencil coefficient matrix:

$$\mathbf{A}^{\alpha} = \frac{\ell_2}{\ell_1}\begin{bmatrix} 0 & 0 & 0 \\ -1 & 2 & -1 \\ 0 & 0 & 0 \end{bmatrix} + \frac{\ell_1}{\ell_2}\begin{bmatrix} 0 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 0 \end{bmatrix} - \frac{\xi_o^2\ell_1\ell_2}{12}\begin{bmatrix} (1-\alpha)\delta_{ol} & (1-\alpha) & (1-\alpha)\delta_{or} \\ (1-\alpha) & 6(1+\alpha) & (1-\alpha) \\ (1-\alpha)\delta_{or} & (1-\alpha) & (1-\alpha)\delta_{ol} \end{bmatrix} \tag{5.42}$$

We see that using simplicial FEs in 2D, the α-interpolation of Galerkin FEM and FDM is equivalent to the alpha-interpolation method (AIM) [110, 140]. In the AIM, the consistent mass matrix $\mathbf{M}$ that appears in the Galerkin FEM is replaced by the α-interpolated mass matrix $\mathbf{M}^{\alpha} := (1-\alpha)\mathbf{M} + \alpha\mathbf{M}_L$. Consider the following definition of the PG weights for simplicial FEs,

$$W_{\alpha}^{a} = \begin{cases} \widetilde{W}_{\alpha}^{a} := \mathbb{W}_{\alpha}^{ab}\mathrm{N}^{b} & \text{in the element interior} \\ \widehat{W}^{a} := \mathrm{N}^{a} & \text{on the element edges} \end{cases} \tag{5.43}$$

Using the above weights for simplicial FEs in any of the weak forms presented in Eq.(5.27), we recover the AIM within a PG setting. In particular for structured simplicial FE meshes we recover the stencil given in Eq.(5.42). We can guess that a solution to any generic stencil takes the form $\Phi^{i,j} := \phi(x_1^i, x_2^j) = \exp[\mathrm{i}(\xi_1^h x_1^i + \xi_2^h x_2^j)]$. Substituting this solution into the stencil formed by $\mathbf{A}^{\alpha}$ given in Eq.(5.42) and defining $\lambda_1 := \exp(\mathrm{i}\xi_1^h\ell_1)$ and $\lambda_2 := \exp(\mathrm{i}\xi_2^h\ell_2)$ we get the characteristic equation as follows:

$$\begin{aligned} \frac{[2-\lambda_1-\lambda_1^{-1}]}{\omega_1} + \frac{[2-\lambda_2-\lambda_2^{-1}]}{\omega_2} &= \frac{(1+\alpha)}{2} + \frac{(1-\alpha)}{12}\left(\lambda_1+\lambda_1^{-1}+\lambda_2+\lambda_2^{-1}\right) \\ &+ \frac{(1-\alpha)}{12}\left(\delta_{ol}[\lambda_1\lambda_2^{-1}+\lambda_1^{-1}\lambda_2] + \delta_{or}[\lambda_1\lambda_2+\lambda_1^{-1}\lambda_2^{-1}]\right) \end{aligned} \tag{5.44}$$

where, $\omega_1 := (\xi_o\ell_1)^2$ and $\omega_2 := (\xi_o\ell_2)^2$. For the dispersion analysis of $\mathbf{A}^{\alpha}$ given in Eq.(5.42) we restrict to the case $\ell_1 = \ell_2 = \ell$. In this case, the stencil coefficient matrix $\mathbf{A}^{\alpha}$ simplifies to,

$$\mathbf{A}^{\alpha} = \begin{bmatrix} \delta_{ol}A_2 & A_1 & \delta_{or}A_2 \\ A_1 & A_0 & A_1 \\ \delta_{or}A_2 & A_1 & \delta_{ol}A_2 \end{bmatrix} \quad ; \quad \begin{aligned} A_0 &:= 4-(1+\alpha)(\omega/2) \\ A_1 &:= -1-(1-\alpha)(\omega/12) \\ A_2 &:= -(1-\alpha)(\omega/12) \end{aligned} \tag{5.45}$$

where, $\omega := (\xi_o\ell)^2$. The characteristic equation given in Eq.(5.44) now gets simplified to the following:

$$A_0 + 2A_1[\cos(\xi_1^h\ell) + \cos(\xi_2^h\ell)] + 2A_2\cos(\xi_1^h\ell \pm \xi_2^h\ell) = 0 \tag{5.46}$$

The '$\pm$' that appears in the above equation corresponds to the cases $o = r$ and $o = l$ respectively (see figure 73). The parameter α may be expressed as a generic series expansion in terms of ω as follows:

$$\alpha := \sum_{m=0}^{\infty} a_m\omega^m \approx a_0 + a_1\omega + a_2\omega^2 + a_3\omega^3 + O\left(\omega^4\right) \tag{5.47}$$

where a_m, b_m are coefficients independent of ω. Following the approach used in [137] which was originally presented in [10], the relative phase ($\mathbb{P}$) and local truncation ($\mathbb{T}$) errors along any direction β can be written as,

$$\mathbb{P} = r_1\omega + O\left(\omega^2\right) \quad ; \quad \mathbb{T} = -2r_1\omega + O\left(\omega^2\right) \tag{5.48}$$

$$r_1 := \frac{(a_0 - 1)}{24}[2 \pm \sin(2\beta)] + \left[\frac{3 + \cos(4\beta)}{96}\right] \tag{5.49}$$

Clearly it is impossible to obtain the condition $r_1 = 0$ by a choice of the coefficient a_0 that is independent of the angle β. Thus, unlike for structured bilinear block FEs, unfortunately for structured simplicial FEs the pollution is essentially of the same order as for those of the Galerkin FEM, the FDM and the GLS-FEM [79, 181]. Nevertheless, just like for the GLS-FEM, the coefficient a_0 can be chosen so as to arrive at a higher-order modification of the interior stencil of the Galerkin FEM. Similar studies for eigenvalue problems using the AIM with simplicial FEs was done in [109, 110, 139, 140].

Remark: Following the approach taken for bilinear block FEs, it is possible to provide different models for the PG weights on the elements edges. This idea will be explored in future works.

5.6 EXAMPLES

In this section we present some examples in 2D for the problem defined by Eq.(5.1) and considering the following problem data: the wavenumber $\xi_o \in \{50, 100\}$, the source $f = 0$, the direction of wave propagation $\beta = (\pi/9)$ and the domain $\Omega = [0, 1] \times [0, 1]$. The domain Ω is discretized by considering both structured and unstructured meshes made up of just the bilinear block-FEs. The unstructured meshes are obtained by randomly perturbing the interior nodes of structured meshes with coordinates (x_i, y_i) as follows [60, 128]:

$$x_i^{'} = x_i + \ell_1\delta\,\texttt{rand}() \quad ; \quad y_i^{'} = y_i + \ell_2\delta\,\texttt{rand}() \tag{5.50}$$

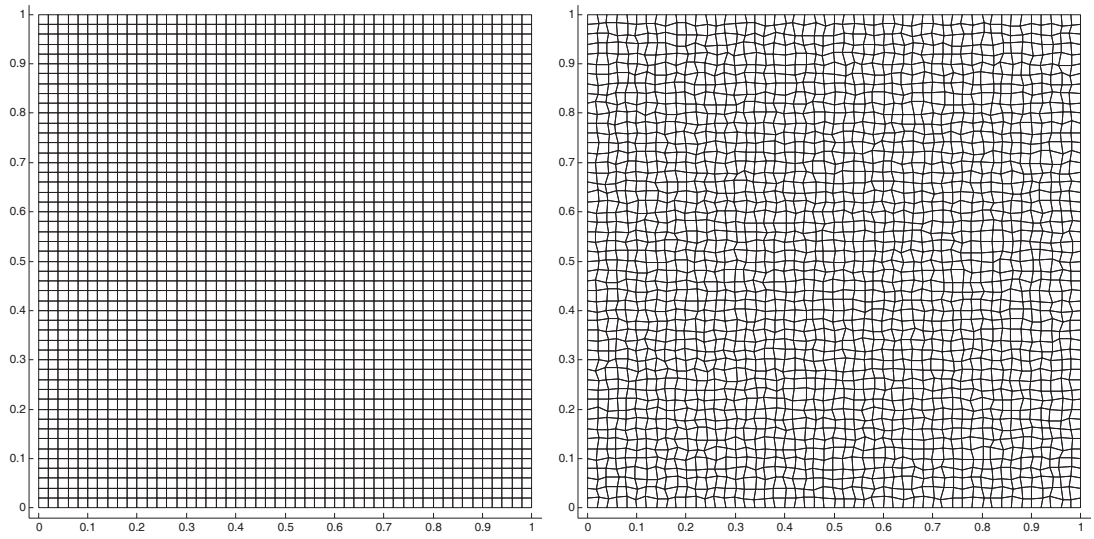
where, $(x_i^{'}, y_i^{'})$ represent the corresponding coordinates of the unstructured mesh, δ is a mesh distortion parameter and rand() is a function that returns random numbers uniformly distributed in the interval $[-1, 1]$. Figure 74 illustrates an instance of an unstructured mesh obtained by this procedure using a 50×50 square mesh and the parameter $\delta = 0.2$.

We consider the following four cases concerned with the choice of the stabilization parameters α_1, α_2:

I: $\alpha_1 = \alpha_2 = 0$. This case corresponds to the Galerkin FEM.

II: $\alpha_1 = \alpha_2 = 1$. This case on rectangular meshes corresponds to the classical FDM. We denote this case as FDM/PG as it is obtained within a Petrov–Galerkin framework. FDM/PG is a straight-forward extension of the FDM to unstructured meshes.

III: $\alpha_1 = \alpha_2 = (1/2)$. This case corresponds to a discrete system that is equivalent to the average of the Galerkin FEM and the FDM/PG. On rectangular meshes



(a) (b)

Figure 74: Meshes made of bilinear block-FEs. (a) Structured mesh, $\delta = 0$. (b) Unstructured mesh, $\delta = 0.2$.

we obtain the stencil associated with $(\mathbf{A}^{fem} + \mathbf{A}^{fdm})/2$, which is equivalent to the one obtained using the generalized Padé approximation in 2D [81, 168]. The dispersion accuracy on square meshes is of fourth-order.

IV: $\alpha_1 \neq \alpha_2 \neq 0$ and given by Eq.(5.40b). On rectangular meshes this case yields the nonstandard compact stencil presented in [137]. Recall that on square meshes these expressions for the parameters α_1, α_2, guarantee sixth-order dispersion accuracy.

For these considerations we study the convergence of the relative error in the following norms:

$$\text{L}^2 \text{ norm} \qquad \frac{\|\phi - \phi_h\|_0}{\|\phi\|_0} := \frac{[\int_\Omega |\phi - \phi_h|^2 \, d\Omega]^{1/2}}{[\int_\Omega |\phi|^2 \, d\Omega]^{1/2}} \tag{5.51a}$$

$$\text{H}^1 \text{ semi-norm} \qquad \frac{\|\phi - \phi_h\|_1}{\|\phi\|_1} := \frac{[\int_\Omega |\boldsymbol{\nabla}(\phi - \phi_h)|^2 \, d\Omega]^{1/2}}{[\int_\Omega |\boldsymbol{\nabla}\phi|^2 \, d\Omega]^{1/2}} \tag{5.51b}$$

$$l^\infty \text{ Euclidean norm} \qquad \frac{|\Phi_e - \Phi_h|_\infty}{|\Phi_e|_\infty} := \frac{\max_i |\Phi_e^i - \Phi_h^i|}{\max_i |\Phi_e^i|} \tag{5.51c}$$

where Φ_e is the exact solution sampled at the nodes of the mesh. In the convergence studies done here, the numerical solutions corresponding to the four cases viz. I-IV, are compared with the following solutions: the nodally exact interpolant denoted by $I_h\phi$ and the best approximations with respect to the L^2 norm and the H^1 semi-norm

denoted by $P_h^0\phi$ and $P_h^1\phi$ respectively. The solutions $I_h\phi$, $P_h^0\phi$ and $P_h^1\phi$ can be found as shown in Eq.(5.52).

$$I_h\phi := N^a \Phi_e^a \tag{5.52a}$$

$$\int_\Omega w_h(\phi - P_h^0\phi)\, d\Omega = 0 \;\forall\; w_h \in V_0^h \quad \Rightarrow \;\| \phi - P_h^0\phi \|_0 \leqslant \| \phi - \phi_h \|_0 \;\forall\; \phi_h \in V^h \tag{5.52b}$$

$$\int_\Omega \boldsymbol{\nabla} w_h \cdot \boldsymbol{\nabla}(\phi - P_h^1\phi)\, d\Omega = 0 \;\forall\; w_h \in V_0^h \quad \Rightarrow \;\| \phi - P_h^1\phi \|_1 \leqslant \| \phi - \phi_h \|_1 \;\forall\; \phi_h \in V^h \tag{5.52c}$$

As the exact solution ϕ is sinusoidal, we have used a third-order Gauss quadrature rule to evaluate the expressions involving ϕ in Eq.(5.51) and Eq.(5.52).

5.6.1 *Example 1: Dirichlet boundary conditions*

In this example, only the Dirichlet boundary conditions are prescribed such that the exact solution is $\phi(\mathbf{x}) = \sin(\xi^\beta \cdot \mathbf{x})$, where $\xi^\beta := \xi_o(\cos(\beta), \sin(\beta))$. Structured meshes with $n \times n$ square elements are considered with n given by the following expression.

$$n = \texttt{ceil}(50 \times 2^{m/8}) \quad ; \quad m \in \{0, 1, 2, \ldots 28\} \tag{5.53}$$

where $\texttt{ceil}(m)$ is a function that returns the nearest integer greater than or equal to m. Unstructured meshes are obtained corresponding to each structured mesh using the procedure described earlier. For these considerations we present the plots of the relative error vs. the mesh-size.

Figure 75 illustrates the convergence of the relative error in the L^2 norm. Clearly the error lines of the considered solutions are bounded from below by the error line of $P_h^0\phi$ (L^2-BA) and show a tendency to become parallel to the error line of $P_h^0\phi$ as $\ell \to 0$. Figures 75a and 75b show the L^2 error considering $\xi_o = 50$ and for structured ($\delta = 0$) and unstructured ($\delta = 0.2$) meshes respectively. As expected the error lines corresponding to cases I and II differs substantially from those of $I_h\phi$, $P_h^0\phi$ and $P_h^1\phi$. The error lines corresponding to cases III and IV are very close to that of $I_h\phi$. As the solution in case IV has sixth-order dispersion accuracy on square meshes it is almost the same as $I_h\phi$. On unstructured meshes the quality of the solution in case IV deteriorates and is similar to that of case III. Figures 75c and 75d show the error lines considering $\xi_o = 100$ and for the choices $\delta = 0$ and $\delta = 0.2$ respectively. As expected all the error lines corresponding to cases I-IV deviate further from the error lines of $I_h\phi$, $P_h^0\phi$ and $P_h^1\phi$ (the pollution effect). On square meshes the solution of case IV shows the least deviation and is practically identical to $I_h\phi$ (Figure 75c). The pollution associated with the solution of case III is similar to that of cases I and II on coarse meshes but it diminishes rapidly on further mesh refinement. Again, on unstructured meshes the quality of the solution in case IV deteriorates showing an appreciable deviation from the error lines of $I_h\phi$, $P_h^0\phi$ and $P_h^1\phi$ and is similar to that of case III (Figure 75d). A distinctive feature in these plots is the formation of spikes in the error lines. Their presence is more evident for higher wavenumbers and on unstructured meshes where the dispersion errors are relatively higher. As here we have prescribed only the Dirichlet boundary conditions the numerical solutions might suffer spurious

amplitude and/or phase modulations to satisfy them [79]. Encounters with zones of degeneracy (wherein the solution might blow up) also contributes to huge errors in the amplitude [47, 79, 137]. Fortunately, these spurious modulations reduce should other choices for the boundary conditions be employed viz. an exterior problem with DtN boundary conditions [79], an interior problem with Robin boundary conditions [10].

Figure 76 illustrates the convergence of the relative error in the H^1 semi-norm. Clearly the error lines of the considered solutions are bounded from below by the error line of $P_h^1\phi$ (H^1-BA). Unlike the errors measured in the L^2 norm, the errors measured in the H^1 semi-norm show a tendency to merge with the error line of $P_h^1\phi$. Figures 76a and 76b ($\xi_o = 50$) show that the error lines of case III and IV are practically the same as of $I_h\phi$, $P_h^0\phi$ and $P_h^1\phi$. Figures 76c and 76d ($\xi_o = 100$) show that the deviations of the error lines of cases III and IV from the error line of $P_h^1\phi$ even though they exist, it is smaller than that observed using the L^2 norm.

Figure 77 illustrates the convergence of the relative error in the l^∞ Euclidean norm which is a measure of nodal exactness. Figures 77a and 77c show that on structured meshes ($\delta = 0$) the error lines of case III and IV converge at a rate of fourth and sixth order respectively. Figures 77b and 77d show that on unstructured meshes ($\delta = 0.2$) the higher order accuracy of case IV deteriorates and has a trend similar to that of case III. Also, in an average sense both the cases III and IV have second-order convergence rate similar to $P_h^0\phi$ and $P_h^1\phi$. For the wavenumber $\xi_o = 50$ the errors found for the cases III and IV are similar to that of $P_h^0\phi$ (Figure 77b).

5.6.2 *Example 2: Robin boundary conditions*

In this example, only the Robin boundary conditions are prescribed such that the exact solution is $\phi(\mathbf{x}) = \exp(i\boldsymbol{\xi}^\beta \cdot \mathbf{x})$, where $\boldsymbol{\xi}^\beta := \xi_o(\cos(\beta), \sin(\beta))$. The operator $\mathcal{M}$ that appears in Eq.(5.1c) is chosen as $\mathcal{M} := i\xi_o$. Thus, $q(\mathbf{x}) := i(\mathbf{n}\cdot\boldsymbol{\xi}^\beta - \xi_o)\exp(i\boldsymbol{\xi}^\beta \cdot \mathbf{x})$. Structured meshes with $n \times n$ square elements are considered with n given by the following expression.

$$n = \texttt{ceil}\left(\frac{m\xi_o}{2\pi}\right) \quad ; \quad m \in \{10, 10.5, 11, 11.5, \ldots 25\} \tag{5.54}$$

Choosing n by the above expression guarantees the presence of at least m elements per wavelength. Unstructured meshes are obtained corresponding to each structured mesh using the procedure described earlier. For these considerations, we present the plots of the relative error vs. ξ^*, where $\xi^* := (\xi_o \ell/\pi)$. The choice of ξ^* as the abscissa in the plots allows us to single out the pollution effect.

Figures 78, 79 and 80 illustrate the convergence of the relative error in the L^2 norm, the H^1 semi-norm and the l^∞ Euclidean norm respectively. Clearly, all the spurious modulations that appeared in the error lines considering only the Dirichlet boundary conditions (Figures 75, 76 and 77) diminish when the Robin boundary conditions are prescribed. Also, in all the figures (78, 79 and 80) by freezing the value of δ and increasing the value of ξ_o we observe the following trait. The location of the error lines of $I_h\phi$, $P_h^0\phi$ and $P_h^1\phi$ is practically unaffected by an increase in ξ_o (no pollution). As expected the error lines of cases I and II not only are located high above the error lines of $I_h\phi$, $P_h^0\phi$ and $P_h^1\phi$ but also shift higher with an increase in ξ_o (pollution effect).

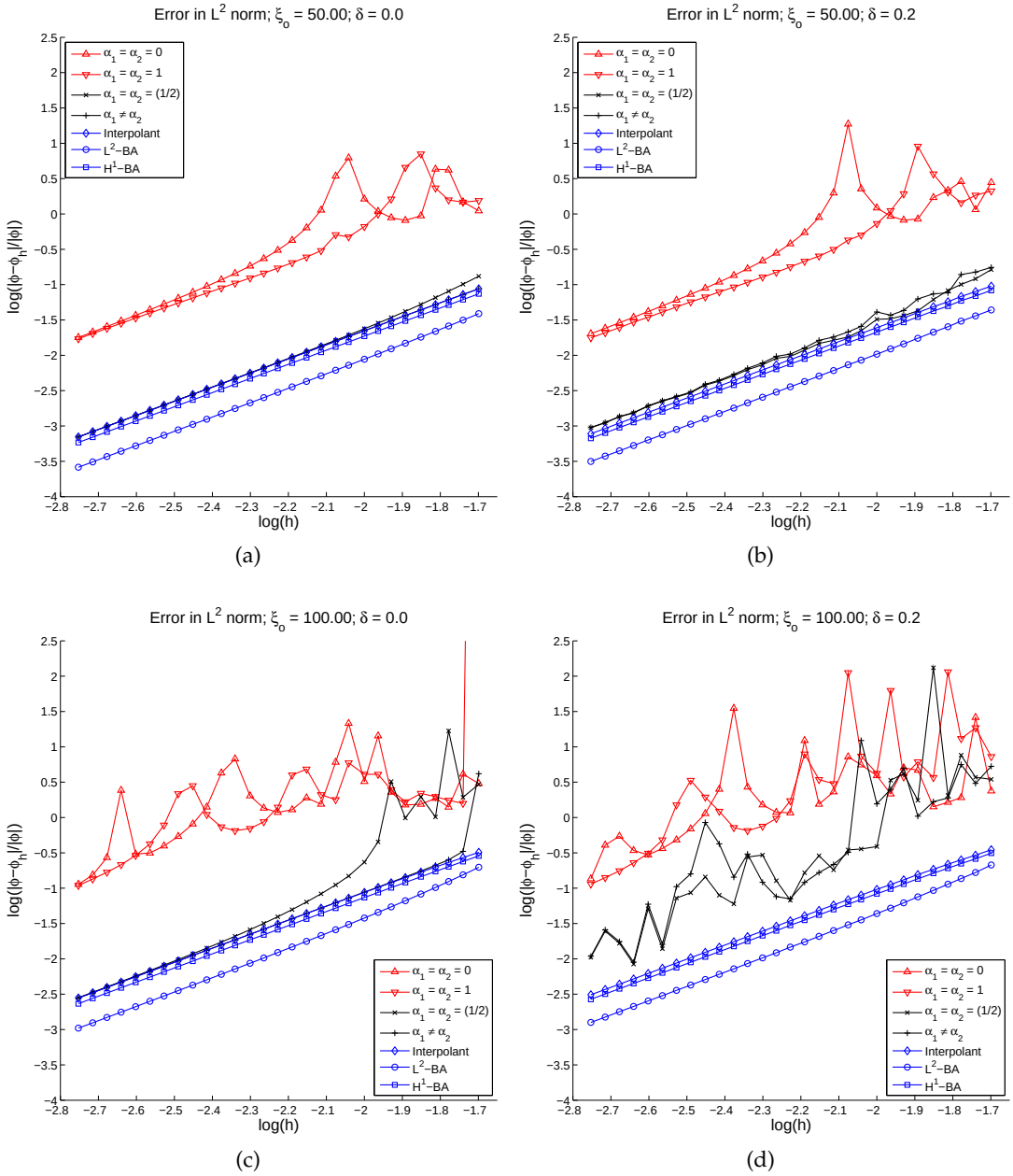


Figure 75: Convergence of the relative error in the L^2 norm using $\beta = (\pi/9)$ and Dirichlet boundary conditions. The wavenumber ξ_o and the mesh distortion parameter used are: (a) $\xi_o = 50$, $\delta = 0$; (b) $\xi_o = 50$, $\delta = 0.2$; (c) $\xi_o = 100$, $\delta = 0$ and (d) $\xi_o = 100$, $\delta = 0.2$.

On uniform meshes ($\delta = 0$) the error lines of cases III and IV not only are located close to the respective best approximations but also show negligible upward shift with an increase in ξ_o (small pollution). Clearly, on uniform meshes the performance of case IV is relatively better than that of case III (although the difference is small). The pollution effect is more visible for these cases on unstructured meshes ($\delta = 0.2$). In the L^2 norm the error lines of cases III and IV show an accuracy at par with $I_h\phi$ and $P_h^1\phi$ (figures 78c and 78d). In the H^1 semi-norm the error lines of cases III and IV are practically the same as those corresponding to $I_h\phi$, $P_h^0\phi$ and $P_h^1\phi$ (figures 79c and

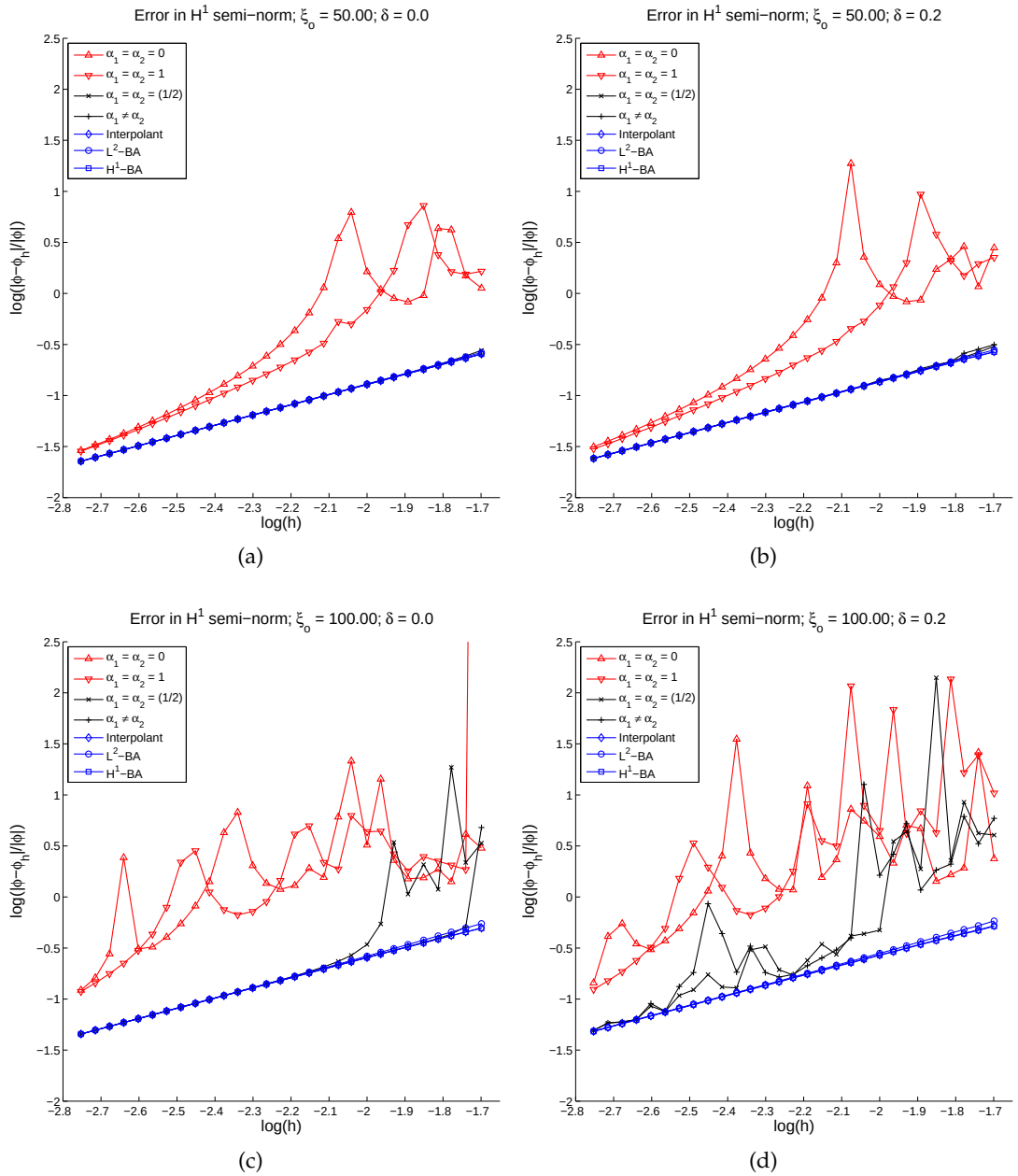


Figure 76: Convergence of the relative error in the H^1 semi-norm using $\beta = (\pi/9)$ and Dirichlet boundary conditions. The wavenumber ξ_o and the mesh distortion parameter used are: (a) $\xi_o = 50$, $\delta = 0$; (b) $\xi_o = 50$, $\delta = 0.2$; (c) $\xi_o = 100$, $\delta = 0$ and (d) $\xi_o = 100$, $\delta = 0.2$.

79d). In the l^∞ Euclidean norm the error lines of cases III and IV are close to the error line of $P_h^0\phi$ (figures 80c and 80d). Further, in Figure 80 note that in an average sense all the error lines have second-order convergence rate in the l^∞ Euclidean norm. This result is due to the error in the approximation of the Robin boundary condition. Thus, unlike in Figure 77 wherein the error lines of cases III and IV showed fourth-order and sixth-order convergence rates respectively, here it drops to second-order.

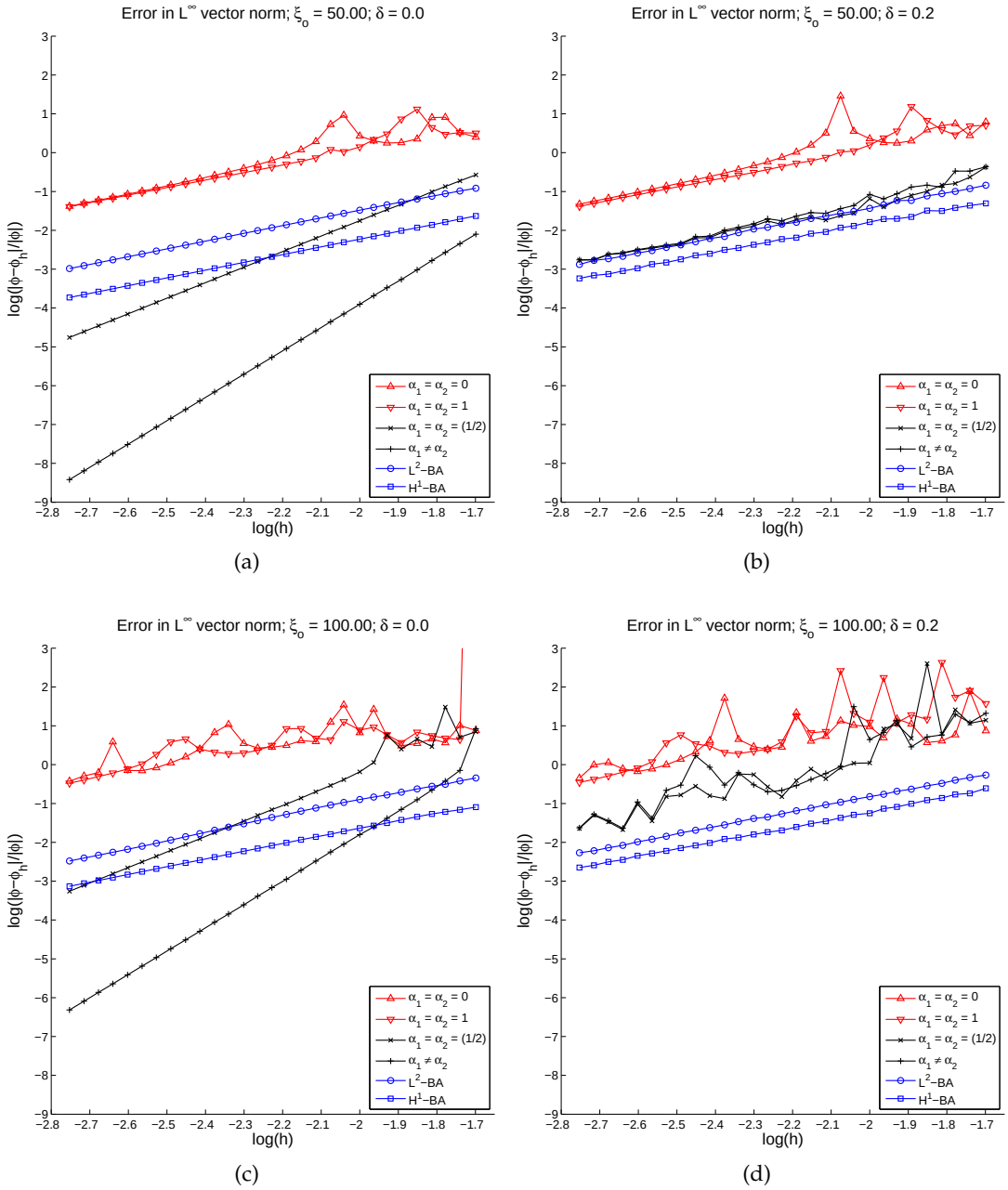


Figure 77: Convergence of the relative error in the l^∞ Euclidean norm using $\beta = (\pi/9)$ and Dirichlet boundary conditions. The wavenumber ξ_o and the mesh distortion parameter used are: (a) $\xi_o = 50$, $\delta = 0$; (b) $\xi_o = 50$, $\delta = 0.2$; (c) $\xi_o = 100$, $\delta = 0$ and (d) $\xi_o = 100$, $\delta = 0.2$.

5.7 CONCLUSIONS

A new Petrov–Galerkin (PG) method involving two parameters viz. α_1, α_2 is presented which yields the following schemes on rectangular meshes: a) a compact stencil obtained by the α-interpolation of the Galerkin FEM and the classical central FDM, should the two parameters be made equal, i. e.$\alpha_1 = \alpha_2 = \alpha$ and b) the nonstandard compact stencil presented in [137] for the Helmholtz equation if the parameters are

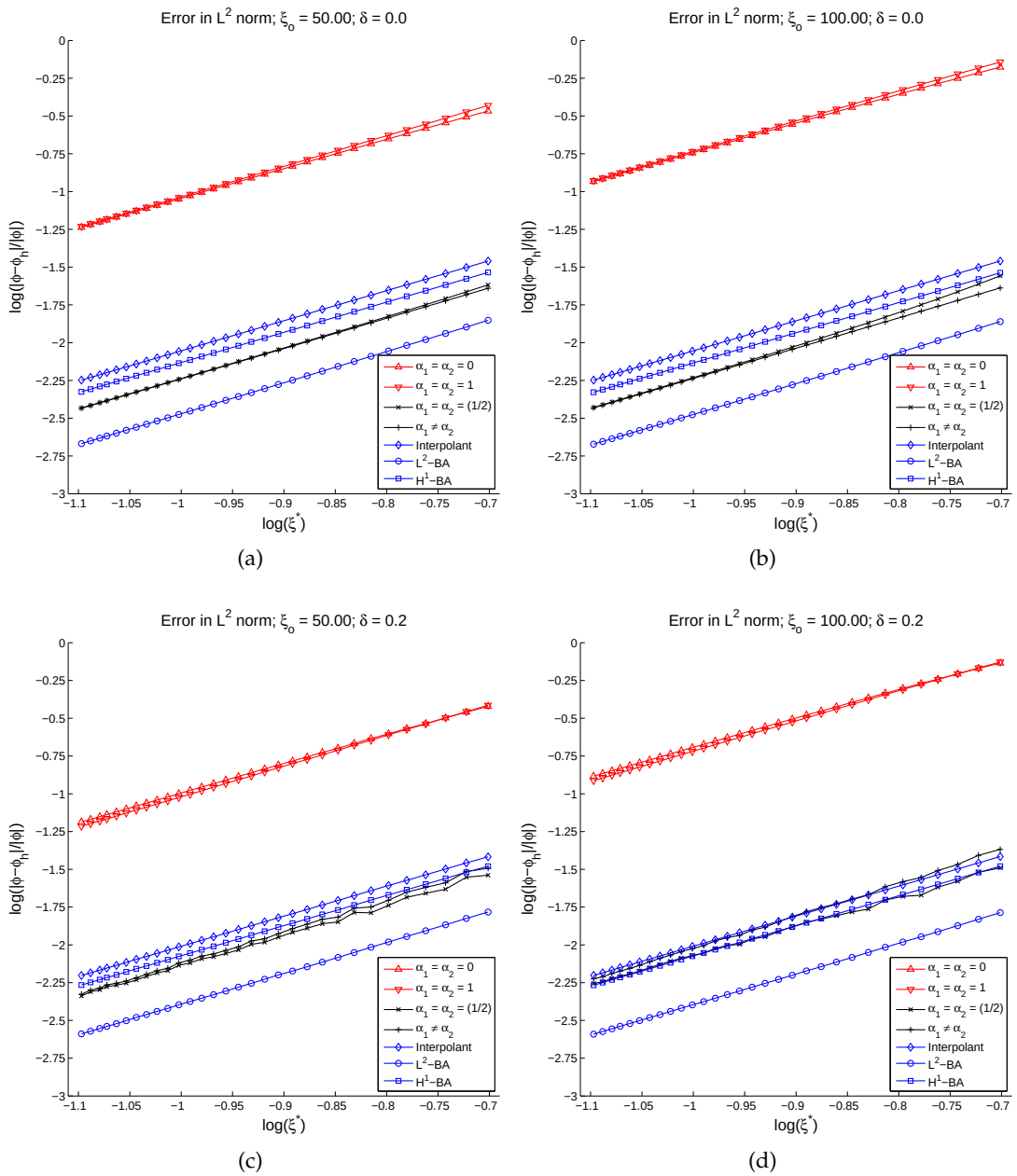


Figure 78: Convergence of the relative error in the L^2 norm using $\beta = (\pi/9)$ and Robin boundary conditions. The wavenumber ξ_o and the mesh distortion parameter used are: (a) $\xi_o = 50$, $\delta = 0$; (b) $\xi_o = 100$, $\delta = 0$; (c) $\xi_o = 50$, $\delta = 0.2$ and (d) $\xi_o = 100$, $\delta = 0.2$.

distinct, i. e.$\alpha_1 \neq \alpha_2$. On square meshes, these two schemes were shown to provide solutions to the Helmholtz equation that have a dispersion accuracy of fourth and sixth order respectively [137]. Thus, this Petrov–Galerkin method yields in a straightforward manner the counterparts of these two schemes on unstructured meshes.

The salient features of this new PG method include the following. The solution space is discretized by standard C^0-continuous finite elements. The test functions/weights are piecewise polynomials of the same degree as the FE shape functions and are generally discontinuous at the inter-element boundaries. Models for the weights

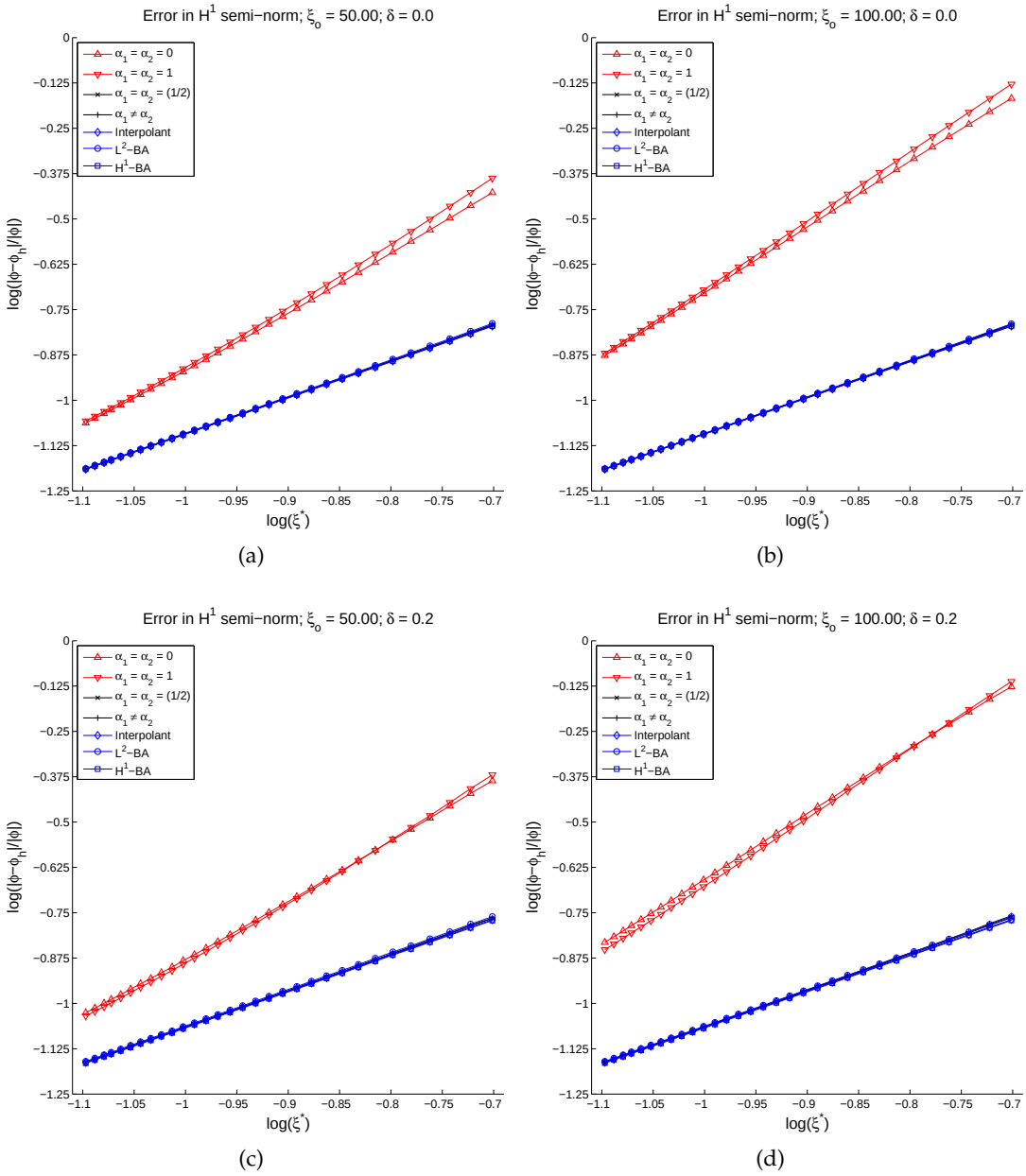


Figure 79: Convergence of the relative error in the H^1 semi-norm using $\beta = (\pi/9)$ and Robin boundary conditions. The wavenumber ξ_o and the mesh distortion parameter used are: (a) $\xi_o = 50$, $\delta = 0$; (b) $\xi_o = 100$, $\delta = 0$; (c) $\xi_o = 50$, $\delta = 0.2$ and (d) $\xi_o = 100$, $\delta = 0.2$.

on the inter-element boundaries are provided such that the sparsity pattern is the same as that for the Galerkin FEM. The parameters α_1, α_2 determine the shape of the weights on the inter-element boundaries and the element interior respectively. The choice $\alpha_1 = \alpha_2 = 0$ yield weights that are identical to the FE shape functions and hence we recover the Galekin FEM. The weights are a partition of unity only in the sense that they add up to unity. As the row lumping technique for the FEM mass matrices is a critical step in the design of these weights (to fulfill the partition of unity constraint), the current PG method is restricted only to those FEs where this tech-

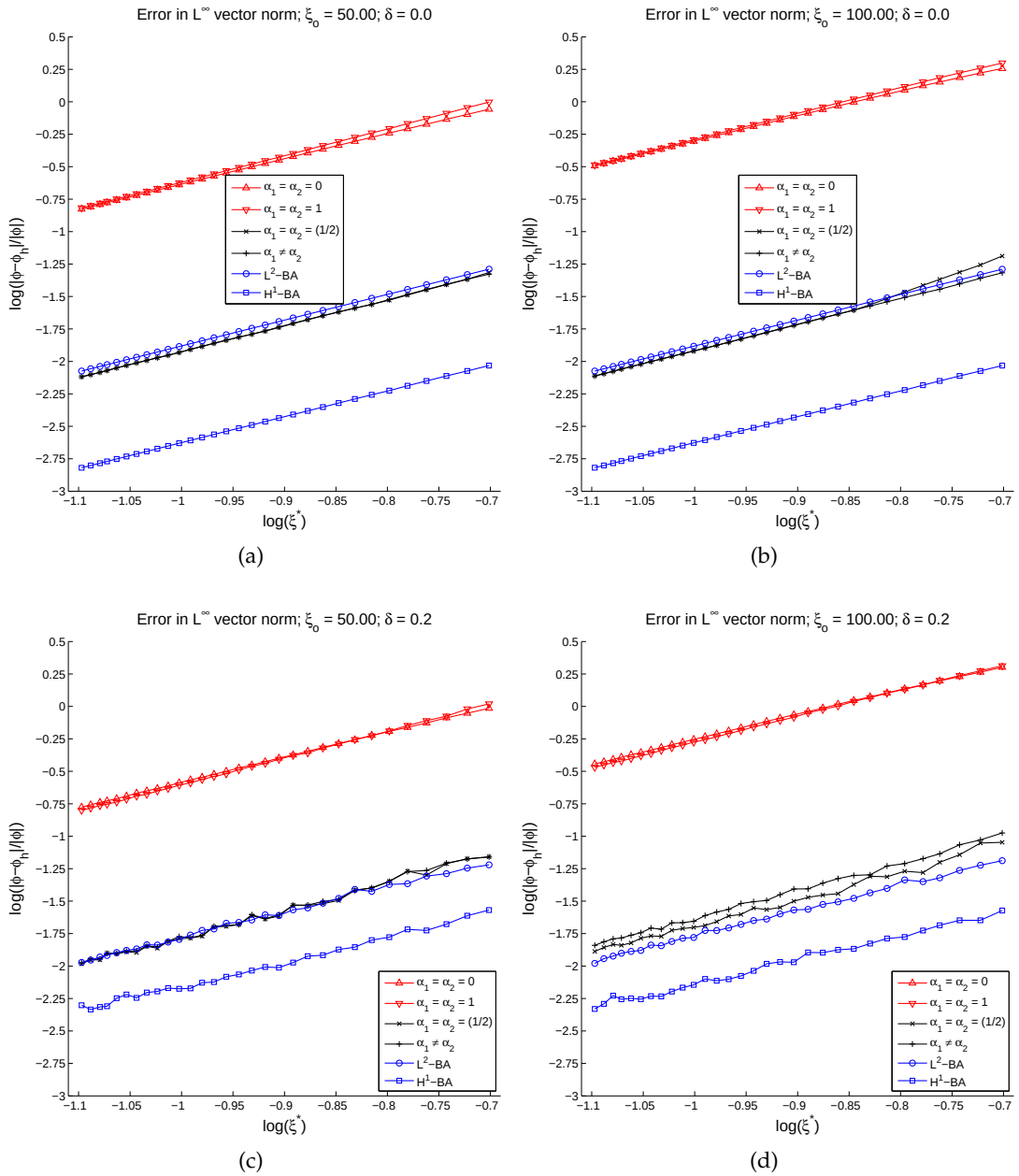


Figure 80: Convergence of the relative error in the l^∞ Euclidean norm using $\beta = (\pi/9)$ and Robin boundary conditions. The wavenumber ξ_o and the mesh distortion parameter used are: (a) $\xi_o = 50$, $\delta = 0$; (b) $\xi_o = 100$, $\delta = 0$; (c) $\xi_o = 50$, $\delta = 0.2$ and (d) $\xi_o = 100$, $\delta = 0.2$.

nique makes sense, i. e.linear interpolation on simplices and multilinear interpolation on blocks.

The α-interpolation of FEM and FDM on a rectangular domain discretized by structured simplicial FE mesh would yield a scheme identical to the alpha-interpolation method (AIM) [110, 140] wherein the mass matrix that appears in the Galerkin FEM is replaced by an α-interpolated mass matrix. In the current PG method we recover

the AIM (even on unstructured simplicial meshes) by making the choice $\alpha_1 = 0$. Unfortunately in this case the dispersion accuracy drops to second-order.

Recall that on square meshes many existing higher-order compact schemes (including the QSFEM [10]) can be recovered by an appropriate choice of the parameters α_1, α_2 [137]. As most of the expressions for α_1, α_2 optimized for square meshes need not be optimal for unstructured meshes in the presented examples we have considered only the simplest expressions that would guarantee fourth-order (choosing $\alpha_1 = \alpha_2 = (1/2)$) and sixth-order ($\alpha_1, \alpha_2$ given by Eq.(5.40b)) dispersion accuracy on square meshes. Convergence studies of the solution error corresponding to these two choices are done to quantify the pollution effect and comparisons are made with respect to the errors of the Galekin FEM, the nodally exact interpolant and the best approximations in the L^2 norm and the H^1 semi-norm respectively. Both the Dirichlet and Robin boundary conditions were considered in the examples. The wavenumbers $\xi_o = 50$ and $\xi_o = 100$ were chosen to represent values in the mid-frequency and high-frequency range respectively.

For the Dirichlet problem, the results on square meshes verify the higher-order dispersion accuracy and the low pollution effect. However on unstructured meshes the dispersion accuracy of the current PG method drops down to second-order (verified by the errors in the l^∞ Euclidean norm). Also, the performance of both the choices for the parameters α_1, α_2 is similar on unstructured grids. For the mid-frequency range, i. e.$\xi_o = 50$ the errors in the l^∞ Euclidean norm for both the parameter choices is close to the error of the best approximation in the L^2 norm. In the high-frequency range, i. e.$\xi_o = 100$, the improvement with respect to the Galerkin FEM is significant. However, the solutions exhibit spurious modulations indicating that there is still room for improvement.

For the Robin problem, these spurious modulations in the solutions cease to exist. The pollution effect on square meshes is greatly reduced and on unstructured meshes it is small. Also, the location of the error lines of the current PG method is in between the error lines of pollution-free solutions, viz. the nodally-exact FE interpolant and the best approximations in the L^2 norm and the H^1 semi-norm, thus indicating high accuracy.

The additional cost of implementation of the current PG method is just the evaluation of the element boundary integrals. All the algebraic evaluations are done at the element level unlike the QOPG method [128] where it is done at the patch level. This feature allows the current PG method to be easily incorporated within an 'assemble-by-elements' data structure. The choice of the parameters $\alpha_1 = \alpha_2 = (1/2)$ render the current PG method independent of the problem and mesh data. In this sense and for this choice, the current PG method could be labeled 'parameter-free'.

Part III

STOKES PROBLEM

O Lord, just as different rivers with distinct sources join the sea, the different paths which people take through different tendencies, various though they appear, crooked or straight, all lead to Thee!

— Hymn 7, Shiva Mahimnha Stotram.

6 PRESSURE LAPLACIAN STABILIZATION

6.1 INTRODUCTION

Many stabilization procedures for solving incompressible problems in fluid mechanics using the finite element method (FEM) have been proposed [1, 2, 11, 14, 22, 34–39, 42, 43, 52, 54, 56, 63, 90, 93, 96, 98, 141, 142, 171, 177–179, 195]. Earlier procedures were based on the so-called penalty approach. This method assumes a pseudo-compressible behavior for the flow with a relationship between the volumetric strain rate ε_v and the pressure p expressed as [54, 195]

$$\varepsilon_v = \frac{1}{\alpha} p \tag{6.1}$$

where α is a large number playing the role of an "artificial" bulk parameter for the fluid. Clearly for $\alpha \to \infty$ the full incompressibility condition $\varepsilon_v \to 0$ is recovered. Another family of stabilized methods added to the standard incompressibility equation (either in the strong form or in the variational expression) a Laplacian of pressure term scaled by a stabilization coefficient depending on physical parameters and the time step increment. Some of these stabilization methods are described in [54, 195]. A similar stabilization procedure adds to the variational equation a local L^2 polynomial pressure projection multiplied by the inverse of the kinematic viscosity [52]. These approaches are *inconsistent* since the stabilization term does not vanish for the exact solution, which can lead to errors in the pressure distribution and in the conservation of the total volume for some problems. An improved stabilization technique adds to the incompressibility condition a term that is a function of the discretized momentum equations, thus ensuring consistency. A popular procedure of this kind is the Galerkin Least Square (GLS) method [56, 93]. In the GLS procedure the stabilized variational expression for the incompressibility equation has the following form

$$\int_\Omega q \varepsilon_v d\Omega - \int_\Omega \tau(\boldsymbol{\nabla}^T q)\bar{\mathbf{r}}_m d\Omega + \text{boundary terms} = 0 \tag{6.2}$$

where Ω is the analysis domain, q are test functions, $\boldsymbol{\nabla}$ is the gradient operator, τ is a stabilization parameter and $\bar{\mathbf{r}}_m$ is a vector containing the discrete residuals of the momentum equations written as $\mathbf{r}_m = \mathbf{0}$. The boundary terms in Eq.(6.2) are added to ensure the consistency of the method [56]. The GLS method is an efficient and accurate stabilization procedure for incompressible flows provided the boundary terms are properly accounted for in Eq.(6.2).

Another residual-based stabilization technique is the so-called *pressure-gradient projection* (PGP) stabilization [35, 36]). In the PGP method, pressure gradients are projected onto a continuous field and the difference between the actual gradients and

their own projections generates the stabilization terms. This is equivalent to replacing the variational expression for the incompressibility equation by the following equation

$$\int_\Omega q\varepsilon_v d\Omega - \int_\Omega \tau(\boldsymbol{\nabla}^T q)(\boldsymbol{\nabla} p + \boldsymbol{\pi}) = 0 \tag{6.3}$$

where $\boldsymbol{\pi}$ is a continuous function (termed the pressure-gradient projection vector) obtained by projecting the pressure gradient $\boldsymbol{\nabla} p$ on the velocity field, and τ is a stabilization parameter.

The term $(\boldsymbol{\nabla} p + \boldsymbol{\pi})$ in Eq.(6.3) can be interpreted as a residual term if we write the momentum equations as $\mathbf{r}_m = \boldsymbol{\nabla} p + \boldsymbol{\pi}$. The total number of discrete unknowns is increased by the $\boldsymbol{\pi}$ field, which is discretized via pressure shape functions. For completeness, the set of governing equations is extended with additional equations requiring the vanishing of the sum $(\boldsymbol{\nabla} p + \boldsymbol{\pi})$ in a weighted residual sense. This provides the equations for computing the $\boldsymbol{\pi}$ variables.

A variant of the PGP technique is the *orthogonal sub-scales* (OSS) method [11, 34, 39]. The variational expression for the incompressibility constraint is written in the OSS method as

$$\int_\Omega q\varepsilon_v d\Omega - \int_\Omega \tau(\boldsymbol{\nabla}^T q)(\bar{\mathbf{r}}_m + \boldsymbol{\pi}) d\Omega = 0 \tag{6.4}$$

where $\bar{\mathbf{r}}_m$ is the discrete residual of the momentum equations and $\boldsymbol{\pi}$ are additional variables that are now interpreted as the projection of the momentum residuals into the velocity space (without boundary conditions). The term $\bar{\mathbf{r}}_m + \boldsymbol{\pi}$ represents an enhanced approximation to the exact momentum residuals $\mathbf{r}_m$. Consistency is preserved by enforcing that the sum $\bar{\mathbf{r}}_m + \boldsymbol{\pi}$ vanishes in a weighted residual sense. This also provides the closure equations for computing the $\boldsymbol{\pi}$ variables.

PGP and OSS stabilization methods are useful for homogeneous flows lacking free-surfaces but encounter difficulties to satisfy incompressibility for fluids with heterogeneous (and discontinuous) physical properties [45, 103, 104] and, in some cases, for free-surface flows when pressure segregation techniques are used for solving the Navier-Stokes equations. Furthermore, PGP and OSS methods increase the number of problem variables ($\mathbf{u}$, p and $\boldsymbol{\pi}$) as well as the connectivity (bandwidth) of the stabilization matrices to be solved.

The stabilization parameter τ in the GLS, PGP and OSS methods is typically chosen as a function of the viscosity and the mesh size [1, 2, 11, 14, 22, 34–39, 42, 43, 52, 54, 56, 63, 90, 93, 96, 98, 141, 142, 171, 177–179, 195]. However, the optimal definition of the stabilization parameter is still a challenge in these methods.

A *pressure Laplacian stabilization* (PLS) method was introduced in [158], that adds two stabilization terms to the variational form of the incompressibility equation: (1) a pressure Laplacian, and (2) a boundary integral. Both terms are multiplied by residual dependent stabilization parameters which emerge naturally from the formulation. Consistency is preserved since the stabilization parameters vanish for the exact solution. The Laplace matrix and the boundary matrix are computed at element level. Because pressure gradient continuity is not enforced, as it happens, for instance, in standard PGP methods, the treatment of heterogeneous multi-fluid problems, such as mixing, is facilitated.

The aim of this chapter is to show that many of the stabilized methods described in the previous lines, and some new ones, can be derived starting from the modified mass balance equation obtained via first and second order finite calculus (FIC) procedures. The FIC technique is based on writing the balance equations in mechanics in a domain of *finite size* and retaining higher order terms in the Taylor series expansions used for expressing the derivative field in the vicinity of a fixed point in the domain. The resulting modified balance equation contains the traditional terms of infinitesimal theory plus additional terms that depend on the dimensions of the balance domain and the derivatives of the infinitesimal balance equations [141]. Clearly, as the dimensions of the balance domain tend to zero the classical balance laws of mechanics are recovered. The interest of the additional terms in the FIC expressions is that they naturally lead to the stabilized numerical schemes (such as stabilized FEM) in fluid and solid mechanics without the need of introducing additional assumptions [101, 102, 105, 141–144, 148–151, 153, 156, 157]. The FIC approach therefore is presented here as *a parent procedure* for deriving a family of old and new residual-based stabilized methods for the analysis of Stokes flows.

An apparent drawback of some of the residual-based stabilized methods presented in this chapter is that the resulting stabilized equation is nonlinear (due to the residual dependence of the stabilization parameters) and this requires using an iterative solution scheme. Preliminary results obtained for simple Stokes flow problems solved with the PLS and PGP methods show that the convergence of the PLS solution is typically found in 2-3 iterations [158]. Also, the nonlinearity can be easily handled within a time integration scheme in transient problems, or in practical problems where other non-linearities might appear due to the presence of convective terms in the momentum equations or non-linear material behavior. In the last part of this chapter, the performance of the PLS method is compared with that of the GLS, PLS and OSS techniques for some relatively simple but illustrative examples of application to Stokes flow problems.

6.2 GOVERNING EQUATIONS

The equations for an incompressible Stokes flow are expressed in the usual manner as:

Momentum

$$\rho \frac{Dv_i}{Dt} - \frac{\partial \sigma_{ij}}{\partial x_j} - b_i = 0 \quad \text{on } \Omega \tag{6.5}$$

Mass balance (incompressibility)

$$\varepsilon_v := \frac{\partial v_i}{\partial x_i} = 0 \quad \text{on } \Omega \quad , \quad i = 1,2,3 \tag{6.6}$$

In Eqs.(6.5) and (6.6), Ω is the analysis domain with a boundary Γ, v_i is the velocity along the i^{th} coordinate direction, ρ is the density, σ_{ij} are the Cauchy stresses, b_i are the body forces (typically $b_i = \rho g_i$ where g_i is the component of the gravity along the i^{th} direction). In our work we assume that $\frac{Dv_i}{Dt} = \frac{\partial v_i}{\partial t}$, i. e.convective derivative terms are neglected, as it is usual in Stokes flows and Lagrangian descriptions of incompressible continua [101, 102, 105, 149, 150, 157, 194].

The problem is completed with the *boundary conditions* for velocities and tractions, i. e.

$$v_i - v_i^p = 0 \quad \text{on } \Gamma_u \tag{6.7a}$$
$$\sigma_{ij} n_j - t_i^p = 0 \quad \text{on } \Gamma_t \tag{6.7b}$$

where v_i^p denote the prescribed velocities on the Dirichlet boundary Γ_u and t_i^p are the traction forces acting on the Neumann boundary Γ_t, with the normal vector $\mathbf{n} = [n_1, n_2, n_3]^T$ (for 3D problems). The total boundary is $\Gamma := \Gamma_u \cup \Gamma_t$.

In Eqs.(6.5)–(6.7) and in the following, summation convention for repeated indices in products and derivatives is used unless otherwise specified.

Following standard practice, Cauchy stresses are split into deviatoric and pressure components as

$$\sigma_{ij} = s_{ij} + p\delta_{ij} \tag{6.8}$$

where s_{ij} are deviatoric stresses, $p = \sigma_{ii}/3$ is the pressure (assumed here to be positive if the mean normal stress is tensile) and δ_{ij} is the Kronecker delta.

We will also assume the constitutive equations of an isotropic Newtonian viscous fluid for which deviatoric stresses are related to deformation rates ε_{ij} by

$$s_{ij} = 2\mu \left(\varepsilon_{ij} - \frac{1}{3}\varepsilon_v \delta_{ij} \right) \tag{6.9a}$$

where μ is the fluid viscosity and

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) \quad , \quad \varepsilon_v := \varepsilon_{ii} \tag{6.9b}$$

6.3 INTEGRAL FORM OF THE MOMENTUM EQUATIONS

The weighted residual form of Eqs.(6.5) and (6.7) is

$$\int_\Omega w_i \left[\rho \frac{\partial v_i}{\partial t} - \frac{\partial \sigma_{ij}}{\partial x_j} - b_i \right] d\Omega + \int_{\Gamma_t} w_i (\sigma_{ij} n_j - t_i^p) d\Gamma = 0 \tag{6.10}$$

where w_i are components of an appropriate test function.

Integrating by parts the term involving σ_{ij} in Eq.(6.10) and substituting Eq.(6.8) into the expression for σ_{ij} gives an integral expression of the momentum equations as

$$\int_\Omega \left[w_i \rho \frac{\partial v_i}{\partial t} + \frac{\partial w_i}{\partial x_j} s_{ij} + \frac{\partial w_i}{\partial x_i} p \right] d\Omega - \int_\Omega w_i b_i \, d\Omega - \int_{\Gamma_t} w_i t_i^p \, d\Gamma = 0 \tag{6.11}$$

Eq.(6.11) is the starting point for the finite element discretization of the momentum equations.

6.4 STABILIZED FORM OF THE INCOMPRESSIBILITY EQUATION USING FINITE CALCULUS

Here we present the finite calculus (FIC) technique and two stabilized forms for the mass balance equation using first and second order FIC techniques.

6.4.1 *Finite calculus*

The finite calculus technique is based on writing the balance equations of mechanics in a domain of *finite size* and retaining higher order terms in the Taylor series expansion used for expressing the integrand (i. e.the classical balance equation in infinitesimal theory) in the vicinity of a fixed point in the domain. The resulting modified balance equation contains the traditional terms of infinitesimal theory plus additional terms that depend on the dimensions of the balance domain and the derivatives of the infinitesimal balance equations [141–143].

First we recall some standard identities/results which will be used to arrive at the FIC modified balance equation. Consider an arbitrary finite-size balance domain Ω_b used to express the balance of fluxes/mass about an arbitrary point P (see Figure 81a). The centroid G of the domain Ω_b is found as,

$$\mathbf{x}^G := \frac{\int_{\Omega_b} \mathbf{x}\, d\Omega}{\int_{\Omega_b} d\Omega} \Rightarrow \int_{\Omega_b} (\mathbf{x} - \mathbf{x}^G)\, d\Omega = 0 \tag{6.12}$$

The moment of inertia of the domain Ω_b is expressed as,

$$\mathbf{I} := \int_{\Omega_b} [(\mathbf{x}\cdot\mathbf{x})\boldsymbol{\delta} - (\mathbf{x}\otimes\mathbf{x})]\; d\Omega = \int_{\Omega_b} \left(|\mathbf{x}|^2\delta_{ij} - x_i x_j\right)\; d\Omega \tag{6.13}$$

Clearly, $\mathbf{I}$ is real, symmetric and has a full rank. The spectral theorem for tensors guarantee the existence of principle axes for $\mathbf{I}$. By orienting the reference frame along the principle axes, the products of inertia become zero, i. e.$I_{ij} = 0\ \forall i \neq j$. In other words, the tensor $\mathbf{I}$ is diagonalized in this configuration.

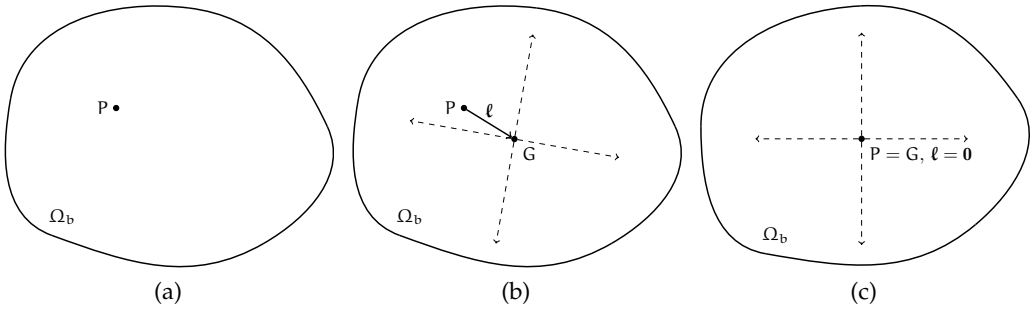


Figure 81: (a) An arbitrary finite-size balance domain Ω_b used in FIC to express the balance of fluxes/mass about an arbitrary point P. (b) The centroid G and the principle axes of Ω_b are now shown. The distance vector from P to G is represented by $\boldsymbol{\ell}$. (c) The balance domain Ω_b is translated such that P = G and is rotated about P such that its principle axes are aligned with the reference axes. In this configuration the products of inertia of Ω_b are zero.

The balance laws of mechanics in the domain Ω_b can be expressed as follows,

$$\int_{\Omega_b} R(\mathbf{x})\, d\Omega = 0 \tag{6.14}$$

where $R(\mathbf{x}) = 0$ is the classical balance equation in infinitesimal theory. Using the multidimensional Taylor series expansion, the residual $R(\mathbf{x})$ can be expressed with respect to the point P as follows,

$$R(\mathbf{x}) = R(\mathbf{x}^P) + \widetilde{\mathbf{x}} \cdot \boldsymbol{\nabla} R(\mathbf{x}^P) + \frac{1}{2}\widetilde{\mathbf{x}} \cdot \mathbf{H}R(\mathbf{x}^P) \cdot \widetilde{\mathbf{x}} + \text{h.o.t} \tag{6.15a}$$

$$\widetilde{\mathbf{x}} := \mathbf{x} - \mathbf{x}^P \tag{6.15b}$$

where $\mathbf{H}(\cdot)$ and h.o.t represent the Hessian operator and the higher-order terms respectively. Substituting Eq.(6.15a) in Eq.(6.14) we have,

$$\int_{\Omega_b} \left[R(\mathbf{x}^P) + \widetilde{\mathbf{x}} \cdot \boldsymbol{\nabla} R(\mathbf{x}^P) + \frac{1}{2}\widetilde{\mathbf{x}} \cdot \mathbf{H}R(\mathbf{x}^P) \cdot \widetilde{\mathbf{x}} + \text{h.o.t} \right] \, d\Omega = 0 \tag{6.16a}$$

$$\Rightarrow R(\mathbf{x}^P) \int_{\Omega_b} d\Omega + \boldsymbol{\nabla} R(\mathbf{x}^P) \cdot \int_{\Omega_b} \widetilde{\mathbf{x}} \, d\Omega + \mathbf{H}R(\mathbf{x}^P) : \int_{\Omega_b} \frac{1}{2}(\widetilde{\mathbf{x}} \otimes \widetilde{\mathbf{x}}) \, d\Omega + \text{h.o.t} = 0 \tag{6.16b}$$

Using Eq.(6.12) in Eq.(6.16b) we get,

$$R(\mathbf{x}^P) + \boldsymbol{\ell} \cdot \boldsymbol{\nabla} R(\mathbf{x}^P) + \mathbf{L} : \mathbf{H}R(\mathbf{x}^P) + \text{h.o.t} = 0 \tag{6.17a}$$

$$\boldsymbol{\ell} := (\mathbf{x}^G - \mathbf{x}^P) \quad ; \quad \mathbf{L} := \frac{\int_{\Omega_b} (\widetilde{\mathbf{x}} \otimes \widetilde{\mathbf{x}}) \, d\Omega}{2 \int_{\Omega_b} d\Omega} \tag{6.17b}$$

Now as this point P can be arbitrarily chosen, the FIC-modified balance equation can be written after dropping out the higher-order terms (h.o.t) as follows:

$$\boxed{R(\mathbf{x}) + \boldsymbol{\ell} \cdot \boldsymbol{\nabla} R(\mathbf{x}) + \mathbf{L} : \mathbf{H}R(\mathbf{x}) = 0} \tag{6.18}$$

As the finite balance-domain Ω_b may be arbitrarily chosen for a given point, generally $\boldsymbol{\ell}$ and $\mathbf{L}$ may vary arbitrarily from one point to the other. However we may choose to fix the shape of Ω_b for all the points that belong to a particular sub-domain of the problem. For instance consider a partition of the problem domain into finite elements. Then for each element K we may fix some shape for Ω_b and hence obtain a constant value for $\boldsymbol{\ell}$ and $\mathbf{L}$ within K. We may also choose a fixed shape of Ω_b for the entire problem domain and thus obtaining a constant $\boldsymbol{\ell}$ and $\mathbf{L}$ everywhere. Henceforth we will only consider the choice of Ω_b to be fixed for any given element K but may vary from one element to the other.

Remark: Choosing the shape for Ω_b and the location of the sampling point P within Ω_b defines the FIC model and the associated FIC-modified balance equation.

Choosing Ω_b such that its centroid G coincides with P, we obtain $\boldsymbol{\ell} := \mathbf{x}^G - \mathbf{x}^P = 0$ (see Figure 81b). In this FIC model, the first order terms in the FIC-modified residual vanishes. So we have,

$$\boxed{R(\mathbf{x}) + \mathbf{L} : \mathbf{H}R(\mathbf{x}) = 0} \tag{6.19}$$

Note that the matrix $\mathbf{L}$ defined earlier in Eq.(6.17b) has full rank and it can be diagonalized by orienting Ω_b such that its principle axes are aligned with the coordinate axes (see Figure 81c). This is directly related to the fact that the off-diagonal elements of $\mathbf{L}$ and moment of inertia $\mathbf{I}$ of Ω_b have the same expressions (except for a change of sign).

6.4.2 *First order FIC form of the incompressibility equation*

The first order FIC form for the incompressibility equation ($R(\mathbf{x}) := \varepsilon_v$) is found by retaining only the first order derivatives in the FIC-modified balance equation given by Eq.(6.18). To fix ideas we choose the balance domain Ω_b to be rectangular and the sampling point P situated at either the top-right or bottom-left corner. The resulting expression is

$$\varepsilon_v \pm \frac{1}{2}\mathbf{h}^T \boldsymbol{\nabla} \varepsilon_v = 0 \tag{6.20}$$

For 2D problems $\mathbf{h} = [h_1, h_2]^T$ where h_1, h_2 are the dimension of the rectangular domain Ω_b. The sign in Eq.(6.20) is positive or negative depending on whether the sampling point P is the bottom-left or the top-right corner node of Ω_b respectively. The sign in Eq.(6.20) is irrelevant in practice. The original derivation of Eq.(6.20) from a different point-of-view can be found in [141].

6.4.3 *Higher order FIC form of the incompressibility equation*

The higher order FIC incompressibility equation is found by using the FIC model that gives the FIC-modified balance equation given by Eq.(6.19). As discussed earlier one can arrive at this form by choosing the balance domain Ω_b such that its centroid G coincides with the sampling point P (see Figure 81). Again, to fix ideas we choose a rectangular balance domain Ω_b and the sampling point P situated at its center. For this choice we have,

$$\boldsymbol{\ell} = \mathbf{o} \quad , \quad \mathbf{L} = \frac{1}{24}\begin{bmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{bmatrix} \tag{6.21}$$

Thus for 2D problems the higher order FIC incompressibility equation is written as,

$$\varepsilon_v + \frac{h_1^2}{24}\frac{\partial^2 \varepsilon_v}{\partial x_1^2} + \frac{h_2^2}{24}\frac{\partial^2 \varepsilon_v}{\partial x_2^2} = 0 \tag{6.22}$$

The original derivation of Eq.(6.22) from a different point-of-view is shown in [158]. Clearly for the infinitesimal case $h_1 = h_2 = 0$, the standard incompressibility equation ($\varepsilon_v = 0$) is recovered for both Eqs.(6.20) and (6.22).

Eqs.(6.20) and (6.22) can be interpreted as *non-local* mass balance equations incorporating the size of the domain used to enforce the mass balance condition and space derivatives of the volumetric strain rate. The FIC mass balance equations can be extended to account for temporal stabilization terms. These terms, however, are disregarded here as they have not been found to be relevant for the problems investigated so far.

6.5 ON THE PROPORTIONALITY BETWEEN THE PRESSURE AND THE VOLUMETRIC STRAIN RATE

Let us assume a relationship between the pressure and the volumetric strain rate typical for "compressible" and "quasi-incompressible" fluids, as

$$\frac{1}{K}p = \varepsilon_v \tag{6.23}$$

where K is the bulk modulus. Clearly for a fully incompressible fluid $K = \infty$ and $\varepsilon_v = 0$. For finite, although very large, values of K the following expression is readily deduced from Eq.(6.23)

$$\frac{1}{K}\boldsymbol{\nabla} p = \boldsymbol{\nabla}\varepsilon_v \tag{6.24}$$

where $\boldsymbol{\nabla}$ is the gradient operator. For 2D problems, $\boldsymbol{\nabla} = \left[\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}\right]^T$.

Eq.(6.24) shows that pressure and volumetric strain rate gradients are co-directional for any $K \neq 0$. We will assume that this property also holds for the full incompressible case (at least for values of K comfortably representable on the computer without overflow). From Eqs.(6.23) and (6.24) we deduce

$$\frac{\boldsymbol{\nabla}\varepsilon_v}{|\boldsymbol{\nabla}\varepsilon_v|} = \frac{\boldsymbol{\nabla} p}{|\boldsymbol{\nabla} p|} \quad , \quad \frac{1}{\varepsilon_v}\frac{\partial \varepsilon_v}{\partial x_i} = \frac{1}{p}\frac{\partial p}{\partial x_i} \tag{6.25a}$$

and hence

$$\frac{\partial \varepsilon_v}{\partial x_i} = \frac{\varepsilon_v}{p}\frac{\partial p}{\partial x_i} = \frac{|\boldsymbol{\nabla}\varepsilon_v|}{|\boldsymbol{\nabla} p|}\frac{\partial p}{\partial x_i} \tag{6.25b}$$

6.6 A PENALTY-TYPE STABILIZED FORMULATION

The first order FIC mass balance equation (6.20) is written as (taking the negative sign)

$$\varepsilon_v = \frac{h_j}{2}\frac{\partial \varepsilon_v}{\partial x_j} \tag{6.26}$$

Using Eq.(6.25b), the FIC term in the r.h.s. of Eq.(6.26) can be expressed as follows

$$\varepsilon_v = \left(\frac{h_j}{2}\frac{\varepsilon_v}{p^2}\frac{\partial p}{\partial x_j}\right) p = \frac{1}{\alpha} p \tag{6.27}$$

where

$$\alpha = \frac{2p^2}{h_j \varepsilon_v}\left(\frac{\partial p}{\partial x_j}\right)^{-1} \tag{6.28}$$

is a pressure stabilization parameter [54, 194, 195].

Eq.(6.27) is equivalent to the so-called penalty formulation (see Eq.(6.1)) for which the pressure-volumetric strain rate relationship is expressed as $p = \alpha\varepsilon_v$ where α a penalty parameter that plays the role of a large "artificial" bulk modulus.

Clearly, $\alpha \to \infty$ for values of $\varepsilon_v \to 0$. However, the full incompressibility condition ($\varepsilon_v = 0$) at element level is obtained on rare occasions only. Hence, the form for α of Eq.(6.28) provides a consistent (residual-based) definition for the penalty stabilization parameter. Nevertheless, upper and lower cut-off values for the values for α should be imposed to prevent volumetric locking when ε_v or $\frac{\partial p}{\partial x_i}$ are equal to zero or the vanishing of α in zones where p is zero [54, 194, 195].

The weighed residual form of Eq.(6.27) is

$$\int_\Omega q(\varepsilon_v - \frac{1}{\alpha}p)d\Omega = 0 \tag{6.29}$$

where q are adequate test functions.

6.7 Galerkin-least squares (GLS) formulation

The starting point is now the higher order FIC incompressibility equation (6.22). The weighted residual form of this equation is

$$\int_\Omega q\left(\varepsilon_v + \frac{h_i^2}{24}\frac{\partial^2 \varepsilon_v}{\partial x_i^2}\right) d\Omega = 0 \tag{6.30}$$

Integration by parts of the second term in Eq.(6.30) gives (for 2D problems)

$$\int_\Omega q\varepsilon_v d\Omega - \int_\Omega \left(\sum_{i=1}^{2}\frac{h_i^2}{24}\frac{\partial q}{\partial x_i}\frac{\partial \varepsilon_v}{\partial x_i}\right) d\Omega + \int_\Gamma \frac{q}{24}\left(\sum_{i=1}^{2} n_i h_i^2 \frac{\partial \varepsilon_v}{\partial x_i}\right) d\Gamma = 0 \tag{6.31}$$

where n_i are the components of the unit normal vector to the boundary Γ.

In the derivation of Eq.(6.31), space derivatives of the characteristic lengths h_1 and h_2 have been neglected. This is correct if we assume that the value of the characteristic lengths is fixed at each point in space. In any case, this assumption does not invalidate the derivation, as long as the discretized formulation converges to correct velocity and pressure fields satisfying the momentum and incompressibility equations in a weighted residual sense and up to the desired order of accuracy.

The term $\frac{\partial \varepsilon_v}{\partial x_i}$ in Eq.(6.31) is expressed as follows. The momentum equations (6.5) can be written using Eqs.(6.8) and (6.9a)

$$\rho\frac{\partial v_i}{\partial t} - \frac{\partial}{\partial x_j}\left(2\mu\varepsilon_{ij}\right) + \frac{2}{3}\mu\frac{\partial \varepsilon_v}{\partial x_i} - \frac{\partial p}{\partial x_i} - b_i = 0 \tag{6.32}$$

From Eq.(6.32) we deduce (neglecting space variations of the viscosity)

$$\frac{\partial \varepsilon_v}{\partial x_i} = \frac{3}{2\mu}r_{m_i} \quad \text{and} \quad \nabla\varepsilon_v = \frac{3}{2\mu}\mathbf{r}_m \tag{6.33}$$

where

$$r_{m_i} := -\rho\frac{\partial v_i}{\partial t} + \frac{\partial}{\partial x_j}\left(2\mu\varepsilon_{ij}\right) + \frac{\partial p}{\partial x_i} + b_i \tag{6.34a}$$

is the form of the momentum residuals used in the subsequent derivations. Note that $r_{m_i} = 0$ for the exact incompressible solution.

For 2D problems

$$\mathbf{r}_m = [r_{m_1}, r_{m_2}]^T \tag{6.34b}$$

Substituting $\partial\varepsilon_v/\partial x_i$ from Eq.(6.33) into (6.31) gives

$$\int_\Omega q\varepsilon_v d\Omega - \int_\Omega \left(\sum_{i=1}^{2} \tau_i \frac{\partial q}{\partial x_i} r_{m_i}\right) d\Omega + \int_\Gamma q \left(\sum_{i=1}^{2} \tau_i n_i r_{m_i}\right) d\Gamma = 0 \tag{6.35}$$

with

$$\tau_i = \frac{h_i^2}{16\mu} \tag{6.36}$$

The form of Eq.(6.35) is equivalent to that obtained in the Galerkin Least Square (GLS) formulation [93] with the boundary integral modification presented in [56]. The expression for the stabilization τ_i of Eq.(6.36) is similar to that typically found in the stabilization literature for Stokes flows [54, 56, 93, 195].

Expansion of the residual term within the second integral yields the standard Laplacian of pressure plus additional terms, i. e.

$$\begin{aligned}
&\int_\Omega q\varepsilon_v d\Omega - \int_\Omega \left(\sum_{i=1}^{2} \tau_i \frac{\partial q}{\partial x_i}\frac{\partial p}{\partial x_i}\right) d\Omega \\
&- \int_\Omega \sum_{i=1}^{2} \left(\tau_i \frac{\partial q}{\partial x_i}\left[-\rho\frac{\partial v_i}{\partial t} + \frac{\partial}{\partial x_j}(2\mu\varepsilon_{ij}) + b_i\right]\right) d\Omega \\
&\qquad + \int_\Gamma q \left(\sum_{i=1}^{2} \tau_i n_i \left[-\rho\frac{\partial v_i}{\partial t} + \frac{\partial p}{\partial x_j} + \frac{\partial}{\partial x_i}(2\mu\varepsilon_{ij}) + b_i\right]\right) d\Gamma = 0
\end{aligned} \tag{6.37}$$

Clearly for linear FE approximations the viscous terms vanish in Eq.(6.37).

We note that the FIC approach presented here introduces the GLS-type stabilization terms just in the incompressibility equation. This is a difference with the standard GLS method which also introduces stabilization terms in the momentum equations [93]. These terms provide symmetry of the global system of equations and are useful for analysis of Navier-Stokes flows. However, they are typically unnecessary for the analysis of Stokes flows. An exception is some transient problems when small time steps are used [90].

6.8 PRESSURE LAPLACIAN STABILIZATION (PLS) METHOD

6.8.1 *Variational form of the mass balance equation in the PLS method*

Using the relationships in Eq.(6.25b) we can write the second integral in Eq.(6.31) as

$$\begin{aligned}
\int_\Omega \left(\sum_{i=1}^{2} \frac{h_i^2}{24}\frac{\partial q}{\partial x_i}\frac{\partial \varepsilon_v}{\partial x_i}\right) d\Omega &= \int_\Omega \left(\sum_{i=1}^{2} h_i^2 \frac{\partial q}{\partial x_i}\frac{\partial p}{\partial x_i}\right) \frac{|\boldsymbol{\nabla}\varepsilon_v|}{24|\boldsymbol{\nabla} p|} d\Omega \\
&= \int_\Omega \left(\sum_{i=1}^{2} \tau_i \frac{\partial q}{\partial x_i}\frac{\partial p}{\partial x_i}\right) d\Omega
\end{aligned} \tag{6.38}$$

with the stabilization parameters given by

$$\tau_i = \frac{h_i^2 |\boldsymbol{\nabla}\varepsilon_v|}{24|\boldsymbol{\nabla}p|} \quad , \quad i = 1,2 \text{ (for 2D problems)} \tag{6.39}$$

Substituting Eqs.(6.38) into (6.31) we write the stabilized mass balance equation as

$$\int_\Omega q\varepsilon_v d\Omega - \int_\Omega (\boldsymbol{\nabla}^T q)\mathbf{D}_v \boldsymbol{\nabla} p d\Omega + \int_\Gamma q g d\Gamma = 0 \tag{6.40}$$

For 2D problems

$$\mathbf{D}_v = \begin{bmatrix} \tau_1 & 0 \\ 0 & \tau_2 \end{bmatrix} \quad \text{and} \quad g = \sum_{i=1}^{2} \frac{h_i^2}{24} n_i \frac{\partial \varepsilon_v}{\partial x_i} \tag{6.41}$$

where $\mathbf{D}_v$ is a matrix of stabilization parameters.

6.8.2 *Computation of the stabilization parameters in the PLS method*

From Eq.(6.33) we deduce (neglecting space variations of the viscosity)

$$\frac{2}{3}\mu|\boldsymbol{\nabla}\varepsilon_v| = |\mathbf{r}_m| \tag{6.42}$$

From the first order FIC mass balance equation (6.20) (with the negative sign) we deduce

$$\frac{1}{2}h_\xi|\boldsymbol{\nabla}\varepsilon_v| = \varepsilon_v \tag{6.43}$$

where h_ξ is the projection of $\mathbf{h}$ along the gradient of ε_v, i. e.

$$h_\xi = \frac{h_i}{|\boldsymbol{\nabla}\varepsilon_v|}\frac{\partial \varepsilon_v}{\partial x_i} = \frac{\mathbf{h}^T \boldsymbol{\nabla}\varepsilon_v}{|\boldsymbol{\nabla}\varepsilon_v|} \tag{6.44}$$

Eqs.(6.42) and (6.43) are consistently modified as follows

$$\frac{2}{3}\mu|\mathbf{v}||\boldsymbol{\nabla}\varepsilon_v| = |\mathbf{v}||\mathbf{r}_m| \tag{6.45}$$

$$\frac{1}{2}ph_\xi|\boldsymbol{\nabla}\varepsilon_v| = p\varepsilon_v \tag{6.46}$$

In Eq.(6.45) $\mathbf{v}$ is the velocity vector.

The r.h.s. of Eqs.(6.45) and (6.46) represents the power of the residual forces in the momentum equations and of the volumetric strain rate, respectively. Note that the product ph_ξ in Eq.(6.46) is always positive, as $p\varepsilon_v \geqslant 0$ (see Eq.(6.23)). Positiveness of $p\varepsilon_v$ pointwise is however ensured in the computation by taking the modulus of this product in the subsequent expressions.

From Eqs.(6.45) and (6.46) we deduce

$$|\boldsymbol{\nabla}\varepsilon_v| = \frac{|p\varepsilon_v| + |\mathbf{v}||\mathbf{r}_m|}{\frac{1}{2}|ph_\xi| + \frac{2}{3}\mu|\mathbf{v}|} \tag{6.47}$$

Substituting Eq.(6.47) into (6.39) gives the expression for the stabilization parameters as

$$\boxed{\tau_i = \frac{h_i^2(|p\varepsilon_v| + |\mathbf{v}||\mathbf{r}_m|)}{(12|ph_\xi| + 16\mu|\mathbf{v}|)|\nabla p|}} \tag{6.48}$$

The expression for τ_i in Eq.(6.48) will vanish for values of v_i and p satisfying *exactly* the incompressibility equation ($\varepsilon_v = 0$) and the momentum equations ($\mathbf{r}_m = \mathbf{0}$). Clearly for the discrete problem, the stabilization parameters depend on the numerical errors in the approximation for ε_v and $\mathbf{r}_m$. In practice, it is advisable to choose a cut-off value for the lower and upper bounds for τ_i avoiding very small or too large values of the stabilization parameter (for instance in zones where ∇p is small). In the examples shown in this chapter we have chosen the following limiting band: $10^{-8} \leqslant \tau_i \leqslant 10^5$.

REMARK 1. Using just Eq.(6.42) for defining $|\nabla \varepsilon_v|$ and substituting this into Eq.(6.39) gives

$$\tau_i = \frac{h_i^2 |\mathbf{r}_m|}{16\mu |\nabla p|} \tag{6.49}$$

In absence of body forces and assuming steady state conditions and a linear FE approximation, then $|\mathbf{r}_m| = |\nabla p|$ and

$$\tau_i = \frac{h_i^2}{16\mu} \tag{6.50}$$

which coincides with Eq.(6.36) deduced for the GLS method. The dimension of τ_i is $\frac{s \times m^3}{kg}$. The expression for τ_i in Eq.(6.50) is typically found in the stabilized FEM literature for Stokes flow [1, 2, 14, 36, 54, 90, 141, 142, 177, 195].

REMARK 2. Other residual-based expressions for the stabilization parameters τ_i in the PLS method can be found. For instance, two alternative expressions for τ_i are

$$\tau_i = \frac{h_i^2}{|\nabla p|} \frac{\left(\rho|\mathbf{v}| + \frac{\rho|h_\xi|}{2\Delta t}\right)|\varepsilon_v| + |\mathbf{r}_m|}{\left(12\rho|h_\xi \mathbf{v}| + 6\rho\frac{h_\xi^2}{\Delta t} + 16\mu\right)} \tag{6.51a}$$

and

$$\tau_i = h_i^2 \left[\frac{\rho|h_\xi \mathbf{v}| + \mu}{24\rho|h_\xi \mathbf{v}||\frac{p}{\varepsilon_v}| + 16\mu^2 \frac{|\nabla p|}{|\mathbf{r}_m|}}\right] \tag{6.51b}$$

The derivation of above expressions can be found in [158].

6.8.3 *PLS boundary stabilization term*

From the relationships in Eq.(6.25b) we express the boundary term g of Eq.(6.41) as

$$g = \sum_{i=1}^{2} \frac{h_i^2}{24} n_i \frac{\partial \varepsilon_v}{\partial x_i} = \sum_{i=1}^{2} \frac{h_i^2}{24} n_i \frac{|\nabla \varepsilon_v|}{|\nabla p|} \frac{\partial p}{\partial x_i} = \sum_{i=1}^{2} \tau_i n_i \frac{\partial p}{\partial x_i} \tag{6.52}$$

The boundary integral in Eq.(6.40) can therefore be expressed in terms of the pressure gradient components using Eq.(6.52) as

$$\int_\Gamma q \left(\sum_{i=1}^{2} \tau_i n_i \frac{\partial p}{\partial x_i} \right) d\Gamma \tag{6.53}$$

where all the terms within the integral are computed at the boundary Γ.

REMARK 3. For $h_i = h_j = h$ then $\tau_i = \tau$. In this case, the boundary integral (6.40) can be expressed as

$$\int_\Gamma q\tau \frac{\partial p}{\partial n} d\Gamma \tag{6.54}$$

where $\frac{\partial p}{\partial n} = n_i \frac{\partial p}{\partial x_i}$ is the gradient of the pressure along the direction normal to the boundary.

We note that all the expressions for τ_i given in the previous equations are "solution dependent". The nonlinear definition of the stabilization parameters can be useful for overcoming the limitations of the standard definitions of τ_i. A recent evidence of the usefulness of solution-dependent stabilization parameters can be found in [90].

6.9 PRESSURE-GRADIENT PROJECTION (PGP) FORMULATION

An alternative stabilized formulation can be derived from the higher order FIC equations by introducing the so-called *pressure-gradient projection* variables. The resulting stabilized mass balance equations can be derived is a number of ways [35, 36, 54, 142, 195]. Here we show how a PGP (for pressure-gradient projection) method can be readily obtained following the higher order FIC approach.

From the momentum equations it can be found

$$\frac{h_i^2}{24} \frac{\partial \varepsilon_v}{\partial x_i} = \tau_i r_{m_i} \tag{6.55}$$

where r_{m_i} is defined in Eq.(6.34a). An expression for the stabilization parameter τ_i is [35, 36, 54, 142, 195]

$$\tau_i = \frac{h_i^2}{24} \left[\frac{\rho l^2}{4\Delta t} + \frac{2\mu}{3} \right]^{-1} \tag{6.56}$$

and l is a typical grid distance. The expression for τ_i of Eq.(6.56) can be obtained as a particular case of Eq.(6.51a).

For relatively fine grids the numerical solution is insensitive to the values of $h_i := \alpha_i l$ [148]. In [158] we found that good results are obtained for the range of values of h_i such that $\sqrt{2} \leqslant \alpha_i \leqslant \sqrt{6}$, where l is a typical grid distance.

For the steady state problems solved in this work the form of τ_i of Eq.(6.36) is used with $\alpha_i = \alpha = \sqrt{6}$.

Note that the stabilization parameters in the PGP method are constant for each element. This is an important difference versus the PLS and penalty formulations

of previous sections, where a nonlinear (and consistent) form for the stabilization parameters is used.

In the standard PGP method the momentum residuals r_{m_i} are split as $r_{m_i} := \frac{\partial p}{\partial x_i} + \pi_i$ where π_i are the so-called pressure-gradient projection variables [35, 36]. In our work we use a slight different approach and split the momentum equations as

$$r_{m_i} := \frac{\partial p}{\partial x_i} + \frac{1}{\tau_i}\pi_i \tag{6.57}$$

where

$$\pi_i = \tau_i \left(-\rho \frac{\partial v_i}{\partial t} + \frac{\partial s_{ij}}{\partial x_j} + b_i \right) \tag{6.58}$$

is the i^{th} pressure-gradient projection weighted by the i^{th} stabilization parameter. The π_i's are now taken as additional variables which are discretized with the standard FEM in the same manner as for the pressure.

The split of Eq.(6.57) ensures that the term $\frac{1}{\tau_i}\pi_i$ is discontinuous between adjacent elements after discretization. This is essential for accurately capturing high discontinuous pressure gradient jumps typical of fluids with heterogeneous physical properties (either the viscosity or the pressure) [45, 104, 105]. In this manner the term $\frac{1}{\tau_i}\pi_i$ can match the discrete pressure gradient term $\frac{\partial p}{\partial x_i}$ which is naturally discontinuous between elements for a linear approximation of the pressure.

Substituting Eq.(6.55) into the second and third integral of Eq.(6.31) and using (6.57) gives (for 2D problems)

$$\int_\Omega q \varepsilon_v d\Omega - \int_\Omega \sum_{i=1}^{2} \frac{\partial q}{\partial x_i} \left(\tau_i \frac{\partial p}{\partial x_i} + \pi_i \right) d\Omega + \int_\Gamma q \sum_{i=1}^{2} n_i \left(\tau_i \frac{\partial p}{\partial x_i} + \pi_i \right) d\Gamma = 0 \tag{6.59}$$

The boundary integral in Eq.(6.59) is typically neglected in PGP formulations and will be disregarded from here onward.

The following additional equations are introduced for computing the pressure gradient projection variables π_i

$$\int_\Omega \sum_{i=1}^{2} \bar{w}_i \left(\frac{\partial p}{\partial x_i} + \frac{1}{\tau_i}\pi_i \right) d\Omega = 0 \tag{6.60}$$

where $\bar{w}_i = q$ is usually taken.

Recall that the term $\frac{\partial p}{\partial x_i} + \frac{1}{\tau_i}\pi_i$ (no sum in i) is an alternative expression for the momentum residual equation (see Eq.(6.57)). This term is enforced to vanish in an average sense via Eq.(6.60).

6.10 ORTHOGONAL SUB-SCALES (OSS) FORMULATION

The OSS method can be readily derived by writing the momentum residuals as

$$r_{m_i} = (\bar{r}_{m_i} + \pi_i) \tag{6.61}$$

where $\bar{r}_{m_i}$ are the *discrete residuals* of the momentum equations and π_i are additional variables that are now interpreted as the projection of the discrete momentum residuals into the velocity space (without boundary conditions) [11, 34, 39].

The variational form for the incompressibility equations is obtained by substituting Eq.(6.61) into (6.35). This gives (neglecting the boundary terms)

$$\int_\Omega q\varepsilon_\nu d\Omega - \int_\Omega \sum_{i=1}^{2} \frac{\partial q}{\partial x_i}\tau_i(\bar{r}_{m_i} + \pi_i)d\Omega = 0 \tag{6.62}$$

The π_i variables are computed by the following additional equations

$$\int_\Omega \sum_{i=1}^{2} \bar{w}_i\tau_i(\bar{r}_{m_i} + \pi_i)d\Omega = 0 \tag{6.63}$$

with $\bar{w}_i = q$. Eq.(6.63) enforces the consistency of the method in an average sense.

The stabilization parameters τ_i are computed as in the PGP method of previous section.

REMARK 4. . For linear FE interpolations the term $\bar{r}_{m_i}$ is simply

$$\bar{r}_{m_i} = -\rho\frac{\partial \bar{v}_i}{\partial t} + \frac{\partial \bar{p}}{\partial x_i} + b_i \tag{6.64}$$

where $\bar{v}_i$ and $\bar{p}$ are the approximate FE values for the velocities and the pressure.

6.11 PLS+π METHOD

The PLS method presented in §6.8 can be substantially enhanced if the momentum residuals appearing in the expression for the stabilization parameters τ_i are computed using Eq.(6.61). The resulting expression for τ_i is

$$\tau_i = \frac{h_i^2\left(|p\varepsilon_\nu| + |\mathbf{v}||\bar{\mathbf{r}}_m + \boldsymbol{\pi}|\right)}{\left(12|ph_\xi| + 16\mu|\mathbf{v}|\right)|\boldsymbol{\nabla} p|} \tag{6.65}$$

This $\boldsymbol{\pi}$ variables are computed via Eq.(6.63) following the discretization procedure described in the next section.

The expression for τ_i of Eq.(6.65) increases the efficiency and accuracy of the PLS method (see Example 6.15.5).

6.12 FINITE ELEMENT DISCRETIZATION

6.12.1 *Discretized equations*

The domain Ω is discretized with a mesh of triangles or quadrilaterals (for 2D) and tetrahedra or hexahedra (for 3D).

For the penalty, GLS and PLS formulations, the velocities and the pressure are interpolated over each element using the same approximation as (for 3D problems)

$$\mathbf{v} = \left\{ \begin{array}{c} v_1 \\ v_2 \\ v_3 \end{array} \right\} = \sum_{j=1}^{n} \mathbf{N}_j \bar{\mathbf{v}}_j \quad , \quad p = \sum_{j=1}^{n} N_j \bar{p}_j \tag{6.66}$$

For the PGP and OSS formulations the $\boldsymbol{\pi}_i$ variables are also interpolated using the shape functions N_i as

$$\boldsymbol{\pi} = \left\{ \begin{array}{c} \pi_1 \\ \pi_2 \\ \pi_3 \end{array} \right\} = \sum_{j=1}^{n} \mathbf{N}_j \bar{\boldsymbol{\pi}}^j \tag{6.67}$$

In the above equations $\mathbf{N}_j = N_j \mathbf{I}_3$, N_j is the standard shape function for node j, $\mathbf{I}_3$ is the 3×3 unit matrix, n is the number of nodes in the element (i.e. $n = 3/4$ for linear triangles/tetrahedra) and $\bar{\mathbf{v}}_j$, $\bar{p}_j$ and $\bar{\boldsymbol{\pi}}^j$ are the values of the velocity vector, the pressure and the $\boldsymbol{\pi}$ variables vector at node j, respectively. Indeed, any other FE approximation for $\mathbf{v}$, p and $\boldsymbol{\pi}$ can be used.

Substituting the approximations (6.66) and (6.67) into the weak form of the momentum equations (Eq.(6.10)) and in the adequate variational expression for the incompressibility equation, gives the following global system of equations (for all the methods considered in this chapter)

$$\bar{\mathbf{M}}\dot{\bar{\mathbf{a}}} + \mathbf{H}\bar{\mathbf{a}} = \mathbf{f} \tag{6.68}$$

where $\dot{\bar{\mathbf{a}}} = \frac{d}{dt}\bar{\mathbf{a}}$ and the different matrices and vectors for the different stabilized FE methods are

Penalty, GLS, PLS methods

$$\bar{\mathbf{a}} = \left\{ \begin{array}{c} \bar{\mathbf{v}} \\ \bar{\mathbf{p}} \end{array} \right\}, \quad \bar{\mathbf{M}} = \left[\begin{array}{cc} \mathbf{M} & \mathbf{o} \\ \hat{\mathbf{M}} & \mathbf{o} \end{array} \right], \quad \mathbf{H} = \left[\begin{array}{cc} \mathbf{K} & \mathbf{Q} \\ (\mathbf{Q}^T + \mathbf{R}) & \mathbf{S} \end{array} \right], \quad \mathbf{f} = \left\{ \begin{array}{c} \mathbf{f}_v \\ \mathbf{f}_p \end{array} \right\} \tag{6.69a}$$

where

$$\begin{array}{ll} \textit{Penalty method}: & \mathbf{S} = \mathbf{P}, \quad \mathbf{R} = \hat{\mathbf{M}} = \mathbf{o}, \quad \mathbf{f}_p = \mathbf{o} \\ \textit{GLS method}: & \mathbf{S} = -\mathbf{L} + \mathbf{B} \\ \textit{PLS method}: & \mathbf{S} = -\mathbf{L} + \mathbf{B}, \quad \mathbf{R} = \hat{\mathbf{M}} = \mathbf{o}, \quad \mathbf{f}_p = \mathbf{o} \end{array} \tag{6.69b}$$

PGP method

$$\bar{\mathbf{a}} = \left\{ \begin{array}{c} \bar{\mathbf{v}} \\ \bar{\mathbf{p}} \\ \bar{\boldsymbol{\pi}} \end{array} \right\}, \quad \bar{\mathbf{M}} = \left[\begin{array}{ccc} \mathbf{M} & \mathbf{o} & \mathbf{o} \\ \mathbf{o} & \mathbf{o} & \mathbf{o} \\ \mathbf{o} & \mathbf{o} & \mathbf{o} \end{array} \right], \quad \mathbf{H} = \left[\begin{array}{ccc} \mathbf{K} & \mathbf{Q} & \mathbf{o} \\ \mathbf{Q}^T & -\mathbf{L} & -\mathbf{C} \\ \mathbf{o} & -\mathbf{C}^T & -\mathbf{T} \end{array} \right], \quad \mathbf{f} = \left\{ \begin{array}{c} \mathbf{f}_v \\ \mathbf{o} \\ \mathbf{o} \end{array} \right\} \tag{6.70}$$

OSS method

$\bar{\mathbf{a}}$ and $\mathbf{H}$ as in Eq.(6.70)

$$\bar{\mathbf{M}} = \left[\begin{array}{ccc} \mathbf{M} & \mathbf{o} & \mathbf{o} \\ \mathbf{M}_p & \mathbf{o} & \mathbf{o} \\ \mathbf{M}_\pi & \mathbf{o} & \mathbf{o} \end{array} \right], \quad \mathbf{f} = \left\{ \begin{array}{c} \mathbf{f}_v \\ \mathbf{f}_p \\ \mathbf{f}_\pi \end{array} \right\} \quad \text{with} \quad \mathbf{f}_\pi = \left\{ \begin{array}{c} \mathbf{f}_{\pi_1} \\ \mathbf{f}_{\pi_2} \\ \mathbf{f}_{\pi_3} \end{array} \right\} \tag{6.71}$$

The matrices and vectors in Eqs.(6.69)–(6.71) are formed by assembling the element contributions given in Box 1 for 3D problems.

$$\mathbf{M}^e_{ij} = \int_{\Omega^e} \rho \mathbf{N}_i \mathbf{N}_j \, d\Omega, \quad \mathbf{K}^e_{ij} = \int_{\Omega^e} \mathbf{G}_i^T \mathbf{D} \mathbf{G}_j \, d\Omega, \quad \mathbf{Q}^e_{ij} = \int_{\Omega^e} \mathbf{G}_i^T \mathbf{m} N_j \, d\Omega$$

with

$$\mathbf{G}_i = \hat{\mathbf{G}} \mathbf{N}_i \quad \text{with} \quad \hat{\mathbf{G}} = \begin{bmatrix} \frac{\partial}{\partial x_1} & 0 & 0 & \frac{\partial}{\partial x_2} & \frac{\partial}{\partial x_3} & 0 \\ 0 & \frac{\partial}{\partial x_2} & 0 & \frac{\partial}{\partial x_1} & 0 & \frac{\partial}{\partial x_3} \\ 0 & 0 & \frac{\partial}{\partial x_3} & 0 & \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} \end{bmatrix}^T$$

$$\mathbf{m} = \begin{Bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \end{Bmatrix}, \quad \mathbf{D} = \mu \begin{bmatrix} 2\mathbf{I}_3 & 0 \\ 0 & \mathbf{I}_3 \end{bmatrix}$$

$$P^e_{ij} = \int_{\Omega} \frac{1}{\alpha} N_i N_j \, d\Omega, \quad L^e_{ij} = \int_{\Omega^e} \tau_k \frac{\partial N_i}{\partial x_k} \frac{\partial N_j}{\partial x_k} d\Omega, \quad B^e_{ij} = \int_{\Gamma} \sum_{k=1}^{3} \left(\tau_k n_k N_i \frac{\partial N_j}{\partial x_k} \right) d\Gamma$$

$$\hat{\mathbf{M}}^e_{ij} = \int_{\Omega^e} \mathbf{m}^T \mathbf{G}_i \rho [\tau] \mathbf{N}_j \, d\Omega - \int_{\Gamma^e} N_i \rho \boldsymbol{\tau}_n^T \mathbf{N}_j \, d\Gamma$$

$$(\mathbf{M}^e_\pi)_{ij} = -\int_{\Omega^e} \rho \mathbf{N}_i^T [\tau] \mathbf{N}_j \, d\Omega$$

$\mathbf{M}^e_p$ as $\hat{\mathbf{M}}^e$ neglecting the boundary term.

$$\mathbf{R}^e_{ij} = -\int_{\Omega^e} (\boldsymbol{\nabla}^T N_i)[\tau] \hat{\mathbf{G}}^T (\mathbf{D}\mathbf{G}_j) d\Omega + \int_{\Gamma^e} N_i \boldsymbol{\tau}_n^T \hat{\mathbf{G}}^T (\mathbf{D}\mathbf{G}_j)) d\Gamma$$

$$\boldsymbol{\tau} = [\tau_1, \tau_2, \tau_3]^T \quad , \quad [\tau] = \begin{bmatrix} \tau_1 & 0 & 0 \\ 0 & \tau_2 & 0 \\ 0 & 0 & \tau_3 \end{bmatrix}, \quad \boldsymbol{\tau}_n = [\tau]\mathbf{n}$$

$$\textbf{PGP}: \qquad \mathbf{C}^e_{ij} = \int_{\Omega^e} \mathbf{m}^T \mathbf{G}_i \mathbf{N}_j \, d\Omega, \quad \mathbf{T}^e_{ij} = \int_{\Omega^e} \mathbf{N}_i [\tau]^{-1} \mathbf{N}_j \, d\Omega$$

$$\textbf{OSS}: \qquad \mathbf{C}^e_{ij} = \int_{\Omega^e} \mathbf{m}^T \mathbf{G}_i [\tau] \mathbf{N}_j \, d\Omega, \quad \mathbf{T}^e_{ij} = \int_{\Omega^e} \mathbf{N}_i [\tau] \mathbf{N}_j \, d\Omega$$

$$\mathbf{f}^e_{v_i} = \int_{\Omega^e} N_i \mathbf{b} \, d\Omega + \int_{\Gamma^e_t} N_i \mathbf{t}^p \, d\Gamma, \quad i, j = 1, 2, 3$$

$$f^e_{p_i} = \int_{\Omega^e} \left(\sum_{j=1}^{3} \tau_j \frac{\partial N_i}{\partial x_j} b_j \right) d\Omega - \int_{\Gamma^e} N_i \left(\sum_{j=1}^{3} \tau_j n_j b_j \right) d\Gamma, \quad (\mathbf{f}_{\pi_j})_i = -\int_{\Omega} N_i \tau_j \mathbf{b} \, d\Omega$$

$\bar{f}_{p_i}$ as f_{p_i} neglecting the boundary term.
$\mathbf{I}_3 : 3 \times 3$ unit matrix , $\mathbf{b} = [b_1, b_2, b_3]^T$, $\mathbf{t}^p = [t^p_1, t^p_2, t^p_3]^T$
Γ^e_t: boundary of element e coincident with the external Neumann boundary

Box 2: Element expressions for the matrices and vectors in Eqs.(6.69)–(6.71) for 3D problems. The expression for τ_i changes for the different stabilized methods

REMARK 5. For linear triangles, matrices $\mathbf{M}$, $\mathbf{P}$ and $\mathbf{T}$ are computed with a three point Gauss quadrature. The rest of the matrices and vectors in Box 1 are computed with just a one-point quadrature. A higher order quadrature might be required in some cases for integrating more accurately the nonlinear terms involving the stabilization parameters τ_i.

REMARK 6. The contribution of the stabilization terms to the stiffness matrix $\mathbf{K}$ which are typical in the standard GLS formulation are not taken into account in this work. These terms are irrelevant for the analysis of the steady-state Stokes problems presented in this chapter.

6.12.2 *Solution schemes for penalty, GLS, PLS methods*

For penalty, GLS, PLS and combined methods, a monolithic transient solution of Eq.(6.68) can be found using the following iterative scheme

$$^{j+1}\bar{\mathbf{a}}^{n+1} = \left[{}^{j}\bar{\mathbf{H}}^{n+1}\right]^{-1}\left(\mathbf{f} + \frac{1}{\Delta t}\bar{\mathbf{M}}\mathbf{a}^{n}\right) \tag{6.72a}$$

with

$$\bar{\mathbf{H}} = \mathbf{H} + \frac{1}{\Delta t}\bar{\mathbf{M}} \tag{6.72b}$$

In Eq.(6.72a) $(\cdot)^n$ and $(\cdot)^{n+1}$ denote values at times t and $t+\Delta t$, respectively while the upper left index j denotes the iteration number; i. e. ${}^{j}(\cdot)^{n+1}$ denotes values at time $t+\Delta t$ and the jth iteration.

For the steady state problems solved in this work we have found the velocity and pressure variables simultaneously by inverting the system

$$\mathbf{H}\bar{\mathbf{a}} = \mathbf{f} \tag{6.73}$$

Clearly, for the PLS method the solution of Eq.(6.73) must be found iteratively, as the stabilization parameters are a function of the velocity and the pressure. A simple direct iteration scheme gives

$$^{j+1}\bar{\mathbf{a}} = [{}^{j}\mathbf{H}]^{-1}\mathbf{f} \tag{6.74}$$

6.12.3 *Solution scheme for PGP and OSS methods*

The solution of Eqs.(6.68) for the PGP and OSS methods is typically performed via an iterative staggered scheme.

The $\bar{\boldsymbol{\pi}}$ nodal variables can be eliminated from Eqs.(6.68) as follows

$$\textit{PGP method}: \quad \bar{\boldsymbol{\pi}} = -\mathbf{T}^{-1}\mathbf{C}^{T}\bar{\mathbf{p}} \tag{6.75a}$$

$$\textit{OSS method}: \quad \bar{\boldsymbol{\pi}} = -\mathbf{T}^{-1}\left(\mathbf{M}_{\pi}\dot{\bar{\mathbf{v}}} + \mathbf{C}^{T}\bar{\mathbf{p}} - \mathbf{f}_{\pi}\right) \tag{6.75b}$$

Substituting $\bar{\boldsymbol{\pi}}$ from Eqs.(6.75) into the second row of Eqs.(6.68) yields the following system of two equations for $\bar{\mathbf{v}}$ and $\bar{\mathbf{p}}$

$$\textit{PGP method}: \qquad \begin{aligned} &\mathbf{M}\dot{\bar{\mathbf{v}}} + \mathbf{K}\bar{\mathbf{v}} + \mathbf{Q}\bar{\mathbf{p}} = \mathbf{f}_v \\ &\mathbf{Q}^T\bar{\mathbf{v}} - (\mathbf{L} - \hat{\mathbf{L}})\bar{\mathbf{p}} = \mathbf{o} \end{aligned} \tag{6.76a}$$

$$\textit{OSS method}: \qquad \begin{aligned} &\mathbf{M}\dot{\bar{\mathbf{v}}} + \mathbf{K}\bar{\mathbf{v}} + \mathbf{Q}\bar{\mathbf{p}} = \mathbf{f}_v \\ &(\bar{\mathbf{M}}_p - \mathbf{C}\mathbf{T}^{-1}\mathbf{M}_\pi)\dot{\bar{\mathbf{v}}} + \mathbf{Q}^T\bar{\mathbf{v}} - (\mathbf{L} - \hat{\mathbf{L}})\bar{\mathbf{p}} = \bar{\mathbf{f}}_p - \mathbf{T}^{-1}\mathbf{f}_\pi \end{aligned} \tag{6.76b}$$

In Eqs.(6.76) $\hat{\mathbf{L}} = \mathbf{C}\mathbf{T}^{-1}\mathbf{C}^T$ is the *discrete pressure Laplace matrix*. This matrix has a wider bandwidth than the Laplace pressure matrix $\mathbf{L}$ in Eqs.(6.69b). The difference between $\mathbf{L}$ and $\hat{\mathbf{L}}$ provides the necessary stabilization for the accurate solution of Eqs.(6.76).

For the steady state case the solution for the velocity and pressure is found simultaneously by Eq.(6.73) with

$$\mathbf{H} = \begin{bmatrix} \mathbf{K} & \mathbf{Q} \\ \mathbf{Q}^T & (\hat{\mathbf{L}} - \mathbf{L}) \end{bmatrix} \quad , \quad \mathbf{a} = \begin{Bmatrix} \bar{\mathbf{v}} \\ \bar{\mathbf{p}} \end{Bmatrix} \quad , \quad \mathbf{f} = \begin{Bmatrix} \mathbf{f}_v \\ \hat{\mathbf{f}}_p \end{Bmatrix} \tag{6.77}$$

where $\hat{\mathbf{f}}_p = \mathbf{o}$ for the PGP method and $\hat{\mathbf{f}}_p = \bar{\mathbf{f}}_p - \mathbf{T}^{-1}\mathbf{f}_\pi$ for the OSS method.

REMARK 7. The elimination of the $\boldsymbol{\pi}$ variables via Eq.(6.75) can be simplified by using a diagonal form of matrix $\mathbf{T}$ obtained as $\mathbf{T}_d = \text{diag}(\mathbf{T})$. This, however, does not affect the bandwidth of matrix $\hat{\mathbf{L}}$.

REMARK 8. The matrix multiplying $\dot{\bar{\mathbf{v}}}$ in the second part of Eq.(6.76b) vanishes for the case of constant density.

6.13 DEFINITION OF THE CHARACTERISTIC LENGTHS

In §6.4.3 we had chosen a rectangular balance domain with the sampling point located at its centroid to arrive at the higher order FIC-modified incompressibility equation. Clearly, the expression for the tensor $\mathbf{L}$ obtained with this choice is anisotropic. In this case, we have to associate the definition of the balance domain to a procedure that guarantees the objectivity of the resulting FIC-based method.

Unlike for the convection–diffusion–reaction problem where an anisotropic definition of the characteristic length tensors was shown to provide better solutions (see Chapter 3), preliminary numerical experiments with the Stokes problem suggest otherwise. That is, no gain is obtained with respect to the quality of the solution (both $\mathbf{v}$ and p) by defining $\mathbf{L}$ in an anisotropic manner. It is for this reason that we have chosen a constant value for h_i defined (for 2D problems) as

$$h_i = h^e \quad \text{with } h^e = [\eta A^e]^{1/2} \tag{6.78}$$

where A^e is the element area, $\eta = 1$ for multilinear elements and $\eta = 2$ for simplicial elements. This corresonds to the choice of a circular FIC balance domian with the sampling point located at its centroid and with radius $h/\sqrt{6}$. So we have,

$$\boldsymbol{\ell} = \mathbf{o} \quad , \quad \mathbf{L} = \frac{1}{24}\begin{bmatrix} h^2 & 0 \\ 0 & h^2 \end{bmatrix} \tag{6.79}$$

6.14 SOME COMMENTS ON THE DIFFERENT STABILIZATION METHODS

6.14.1 *Penalty method*

For the penalty method the system matrix $\mathbf{H}$ is symmetrical. Matrix $\mathbf{M}$ becomes ill-conditioned for very large values of the penalty parameter α. A cut-off for α when the solution approaches the incompressibility limit is therefore mandatory to avoid the volumetric locking problem [54, 195].

6.14.2 *GLS method*

For the GLS method presented here matrices $\hat{\mathbf{M}}$, $\mathbf{R}$ and $\mathbf{B}$ are non symmetrical. Symmetry of matrices $\bar{\mathbf{M}}$ and $\mathbf{H}$ can be simply found by shifting the non symmetrical terms to the r.h.s. of Eqs.(6.68). This introduces a nonlinearity in the solution scheme.

6.14.3 *PLS method*

For the PLS method the boundary stabilization matrix $\mathbf{B}$ is non-symmetrical. Symmetry of the system matrix $\mathbf{H}$ can be recovered by shifting the boundary terms to the r.h.s. of Eq.(6.68). This gives

$$\mathbf{H} = \begin{bmatrix} \mathbf{K} & \mathbf{Q} \\ \mathbf{Q}^T & -\mathbf{L} \end{bmatrix} \quad , \quad \mathbf{f} = \begin{Bmatrix} \mathbf{f}_v \\ \mathbf{q}_p \end{Bmatrix} \tag{6.80a}$$

For 3D problems

$$q^e_{p_i} = \int_{\Gamma^e} N_i \left(\sum_{j=1}^{3} \tau_j n_j \frac{\partial p}{\partial x_j} \right) d\Gamma \tag{6.80b}$$

The boundary force vector $\mathbf{q}_p$ is computed at each iteration as part of the iterative solution process. Preliminary experiences in applying this method show that this does not increase the total number of iterations.

6.14.4 *PGP and OSS method*

All matrices in the PGP method are symmetrical. For the OSS method matrix $\bar{\mathbf{M}}$ is not symmetrical. Symmetry of matrix $\bar{\mathbf{M}}$ can be found by shifting the non symmetrical terms (involving time derivatives of the velocities) to the r.h.s. of Eq.(6.68).

6.14.5 *Comparison between the different stabilization methods*

1. In the PGP and OSS methods the stabilization parameter is typically taken as constant (at least for homogeneous meshes and constant viscous fluids). In the PLS method, however, the stabilization parameter τ_i varies as a function of the volumetric strain rate and the residual of the momentum equations.

2. In the PGP and OSS methods the amount of stabilization is variable in space. This variation is introduced by the difference between the Laplace pressure matrix $\mathbf{L}$ and the discrete pressure Laplace matrix $\hat{\mathbf{L}}$. In the PLS method (and also in the penalty method presented here) the amount of stabilization is also variable in space, but the variation is introduced by the consistent stabilization parameters.

3. The consistency in the GLS methods is guaranteed by introducing the discrete residual of the momentum equations in the stabilized mass balance equation. Consistency in the PGP and OSS methods is enforced by introducing additional equations representing the vanishing of the momentum residuals in an average sense. In the PLS method the consistency is guaranteed by the expression of the stabilization parameters which also vanish for the exact solution (i. e.for $\varepsilon_v = 0$ and $r_{m_i} = 0$).

4. The PLS method is nonlinear due to the dependence of the stabilization parameters with the volumetric strain rate, the pressure, the pressure gradient and the residual of the momentum equations. The PGP and OSS methods are non linear due to the definition of the π_i variables which are a function of the pressure field. The GLS method can be considered as a linear method.

5. All methods, except the penalty method, introduce a boundary stabilization term. Accounting for this term is relevant in the GLS and PLS methods when lower order finite elements are used. On the contrary, consistency recovery techniques as proposed in [111] are required. Unlike for the GLS method, it is observed that this dependence on such consistency recovery techniques is weak for the PLS method. In other words, the adverse effect of excluding the boundary stabilization term is found to be small for the PLS method (see Examples 6.15.3 and 6.15.4).

6. In PGP and OSS methods the boundary stabilization term is usually neglected. This simplification is acceptable on external boundaries, but cannot be neglected at internal interfaces with a jump in the physical properties.

7. PGP and OSS methods typically yield identical results for problems when the force term b_i belongs to the space of finite element functions and linear (or bilinear) elements are used [35].

6.15 EXAMPLES OF APPLICATION

We present a number of examples of simple steady-state Stokes flow problems. The aim is to validate and compare the accuracy and efficiency of some of the methods presented in this chapter. The methods compared are:

- PLS method of §6.8 using the expression for τ_i of Eq.(6.48). The effect of including or not the boundary integral (BI) terms of Eq.(6.40) has been studied.
- PLS+π method of §6.11. This method was used just in the manufactured flow problem (Example 6.15.5).

- GLS method of §6.7, including and excluding the boundary integral (BI) terms of Eq.(6.37).
- OSS method of §6.10 with consistent and diagonal forms of matrix $\mathbf{T}$.

The problems solved are the following:

i) Hydrostatic flow problem for a single fluid in a square domain.

ii) Two-fluid hydrostatic problem in a square domain.

iii) Poiseuille flow in a trapezoidal domain.

iv) Lid driven cavity flow problem.

v) Manufactured flow problem in a trapezoidal domain.

For all problems the nodal velocities and pressures have been found simultaneously *under steady-state conditions* by solving Eq.(6.73). For the GLS and the OSS methods the solution is found in a *single step*. For the PLS method the direct iteration scheme of Eq.(6.74) is used. The first PLS (and PLS+π) solution is found in all cases using a constant value of $\tau_i = \tau = 10^{-5}$. This roughly corresponds to the value of τ_i for the GLS and OSS methods given by Eq.(6.36).

The convergence of the nonlinear iterations for the PLS method is measured in the Euclidean vector norm for velocities and pressure measured as

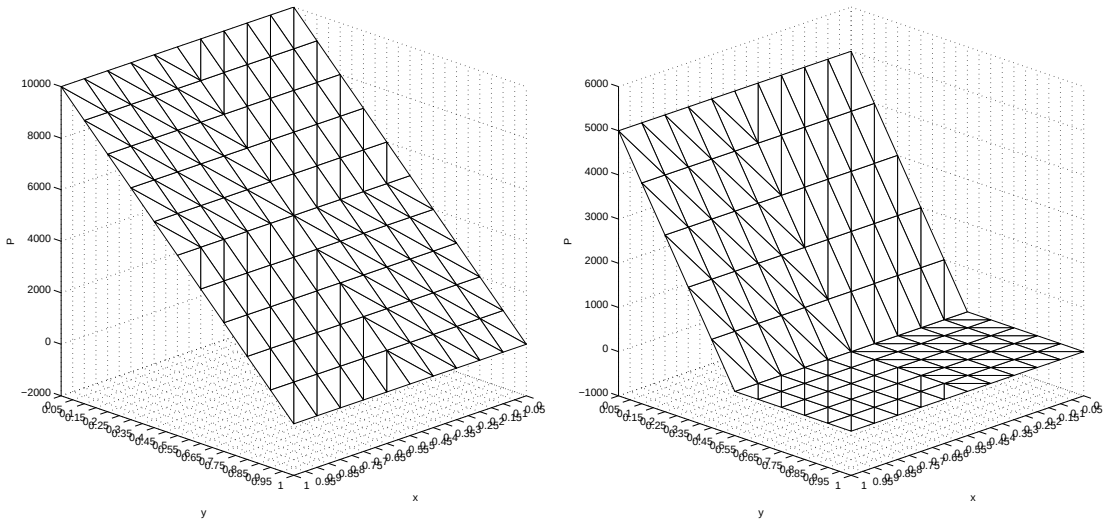
$$\frac{\|^j\bar{\boldsymbol{\Phi}} - {}^{j-1}\bar{\boldsymbol{\Phi}}\|_{L^2}}{\|^j\bar{\boldsymbol{\Phi}}\|_{L^2}} \leqslant \epsilon \quad \text{with} \quad \boldsymbol{\Phi}_k = \begin{Bmatrix} \bar{v}_{1_k} \\ \bar{v}_{2_k} \\ \bar{p}_k \end{Bmatrix} \quad \text{and} \quad \|\bar{\boldsymbol{\Phi}}\|_{L^2} = \left[\sum_a (\bar{\boldsymbol{\Phi}}^a)^2\right]^{1/2}$$

In our work we have chosen $\epsilon = 10^{-4}$ for examples i)–iv) and $\epsilon = 10^{-3}$ for example v). A comparison of the PLS and PGP methods for problems i, ii and iv is reported in [158].

6.15.1 *Hydrostatic flow problem for a single fluid in a square domain*

We solve for the pressure distribution in a square container filled with water. The body forces are $b_1 = b_2 = \rho g$ with values of the density and gravity constant equal to $\rho = 1000$ Kg/m^3 and $g = -10$ m/s^2, respectively. The viscosity is $\mu = 10^{-3}$ Ns/m^2. The normal velocity has been prescribed to zero at the bottom line and the two vertical walls. The nodes on the top surface are allowed to move freely. The solution for this simple problem is $\mathbf{v} = \mathbf{o}$ an hydrostatic distribution of the pressure which is independent of the fluid viscosity. The problem is solved with a $2 \times 10 \times 10$ mesh of 3-node triangles.

Figure 82a shows the pressure distribution obtained by all methods. A converged solution which approximates practically the exact hydrostatic distribution is found with the PLS method in *just two iterations*.



(a) (b)

Figure 82: (a) Hydrostatic flow problem for a single fluid in square domain. (b) Two-fluid hydrostatic problem in square domain. Pressure distribution for both problems obtained with PLS (with and without BI), GLS (with and without BI) and OSS (using $\mathbf{T}$ and $\mathbf{T}_d$). Results for all methods coincide.

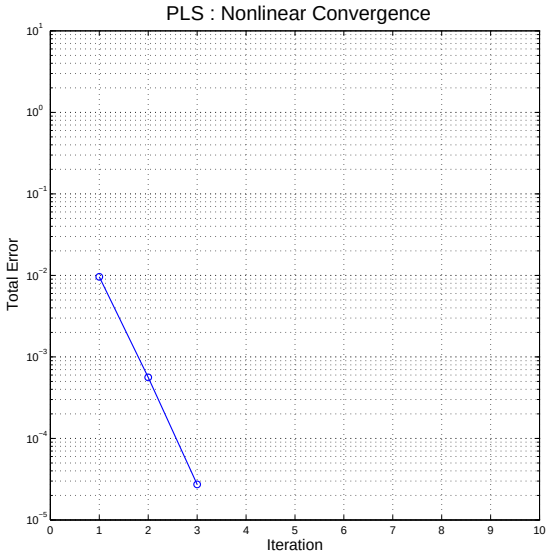


Figure 83: Two-fluid hydrostatic problem. Convergence of the PLS method (with and without BI).

6.15.2 *Two-fluid hydrostatic problem in a square domain*

The same square container of the previous example is considered assuming that the upper half is filled with a liquid of density $\rho = 10^{-3}$ Kg/m^3. The viscosity is the same for both fluids with $\mu = 10^{-3}$ Ns/m^{-2}. The body forces, the boundary conditions and the mesh are the same as for the previous example. The exact analytical solution is $\mathbf{v} = 0$ in the whole container and a linear distribution of the pressure ranging from

$p = 0$ at the top ($x_2 = 1.0$ m) to $p = 10^{-2}$ Pa at $x_2 = 0.5$ m; and again a linear distribution of the pressure from $p = 10^{-2}$ Pa at $x_2 = 0.5$ m to $p = 5000$ Pa at $x_2 = 0$.

Results for the pressure distribution are shown in Figure 82b. Numerical results for all methods studied coincide.

The converged solution for the PLS method is obtained in *three iterations*. The convergence history is shown in Figure 83.

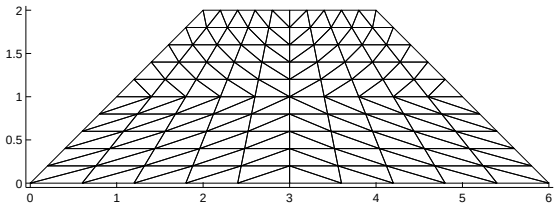


Figure 84: Trapezoidal discretized by a symmetrical mesh of $2 \times 10 \times 10$ 3-node triangles.

6.15.3 *Poiseuille flow in a trapezoidal domain*

A trapezoidal domain Ω is considered with corner nodes given by: (0,0),(6,0),(4,2) and (2,2). The domain is discretized with a mesh of $2 \times 10 \times 10$ 3-node triangles (Figure 84). A parabolic profile for the horizontal velocity is prescribed at both the inlet and outlet lines.

Figure 85a shows the pressure distribution obtained with PLS method (with and without BI), the GLS method (with BI) and the OSS method (using $\mathbf{T}$ and $\mathbf{T}_d$). Results for all these methods coincide. Figure 85b shows the GLS results *not including the BI term*. Note the inaccuracies in the pressure distribution near the edge.

The same trend is observed for the results of the distribution of the v_1 velocity shown in Figure 86. Note the slight increase in accuracy obtained in the PLS method by including the BI term.

The convergence of the PLS solution with and without taking into account the BI terms is shown in Figure 87. The convergence improves when the BI terms are included (3 iterations versus 5 iterations).

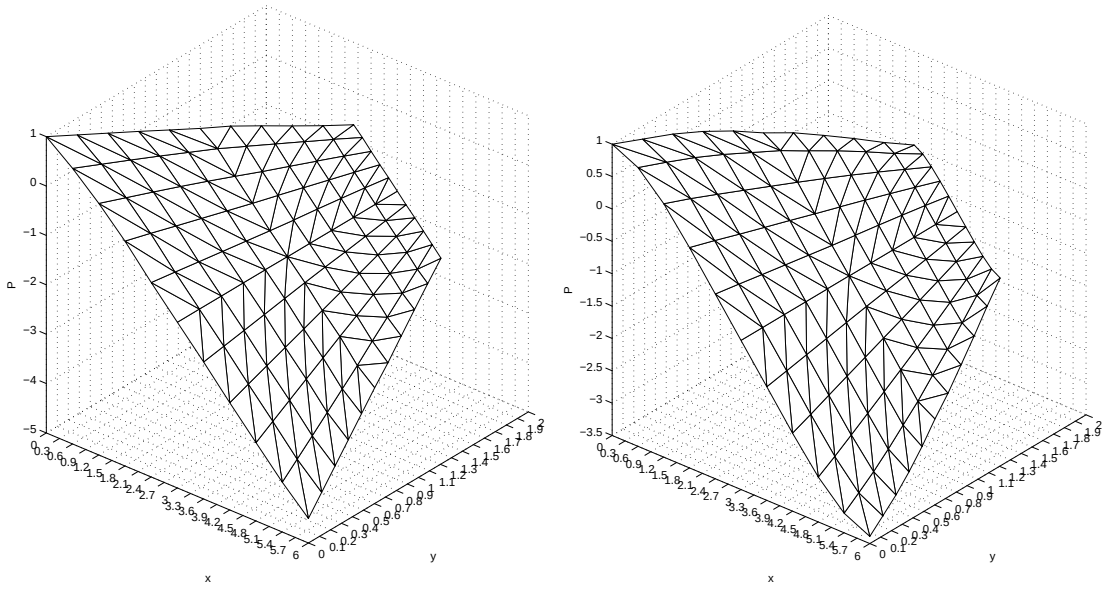
6.15.4 *Lid driven cavity problem*

The flow in a driven square cavity of 1×1 m^2 is studied.

The horizontal velocity on the top surface nodes has been prescribed to $v_1^p(x_1, 1) = 1$ m/s. The vertical velocity has also been prescribed to zero at all nodes on the top surface with the exception of the central node with coordinates (0.5,1) which is left free to move in the vertical direction. The normal velocity at the bottom line and the two vertical walls has been prescribed to zero. The physical properties are $\rho = 10^{-10}$ Kg/m^3, $g = 0$ N/m^2, $\mu = 1$ Ns/m^2.

It can be easily verified that, for the material properties chosen, the value of the stabilization parameter τ_i for the PLS method is approximately constant over the whole analysis domain and equal to

$$\tau_i = \tau \simeq \frac{h_i^2}{16\mu} = \frac{10^{-2}}{16} = 6.25 \times 10^{-3} \frac{m^3 s}{Kg} \tag{6.81}$$



(a) (b)

Figure 85: Poiseuille flow in a trapezoidal domain. (a) Pressure distribution obtained with PLS (with and without BI), GLS (with BI) and OSS (using $\mathbf{T}$ and $\mathbf{T}_d$). (b) Pressure distribution obtained with GLS without BI

Figure 88 shows the pressure distributions in the cavity for all the methods studied using a $2 \times 20 \times 20$ mesh of 3-node triangles. An analysis of the results of Figure 88 shows that (a) the PLS method captures better the singularity of the pressure values at the top corner nodes, (b) the effect of the BI terms is irrelevant for the PLS method but has a positive influence in the GLS method in terms of a better capture of the pressure singularity, (c) the diagonal form of matrix $\mathbf{T}$ introduce a slight diffusion in the OSS results. The pressure contour lines for the considered methods is shown in Figure 89.

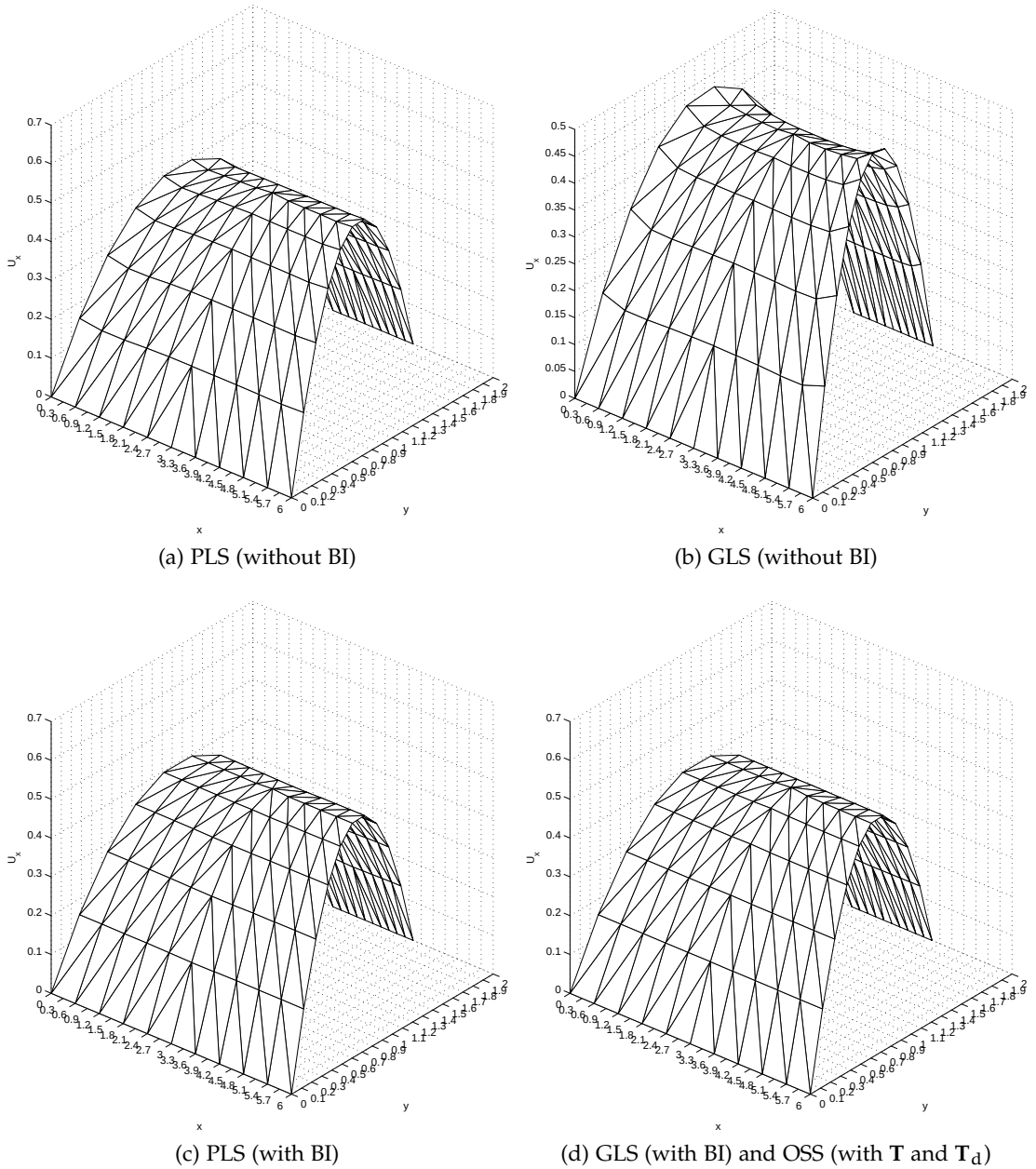
Figure 90 shows the convergence of the PLS results. The convergence curve is practically the same accounting or not for the BI terms. Convergence is slower in this case due to the pressure singularity at the top corners.

6.15.5 *Manufactured flow problem in a trapezoidal domain*

The trapezoidal domain of Figure 84 is discretized by a series of symmetrical mesh consisting of $2 \times n \times n$ 3-node triangular elements and using $n \in \{10, 12, 14, 16, 18, 20, 24, 32, 36, 40\}$. The following relative error norms are used to study the convergence rates of the considered methods

$$e^h_{vH^1} = \frac{\|\mathbf{v} - \bar{\mathbf{v}}\|_1}{\|\mathbf{v}\|_1} = \frac{\sqrt{\int_\Omega \boldsymbol{\nabla}(\mathbf{v} - \bar{\mathbf{v}}) : \boldsymbol{\nabla}(\mathbf{v} - \bar{\mathbf{v}}) \, d\Omega}}{\sqrt{\int_\Omega \boldsymbol{\nabla}\mathbf{v} : \boldsymbol{\nabla}\mathbf{v} \, d\Omega}} \tag{6.82a}$$

$$e^h_{pL^2} = \frac{\|p - \bar{p}\|_0}{\|p\|_0} = \frac{\sqrt{\int_\Omega (p - \bar{p})^2 \, d\Omega}}{\sqrt{\int_\Omega p^2 \, d\Omega}} \tag{6.82b}$$



(a) PLS (without BI)

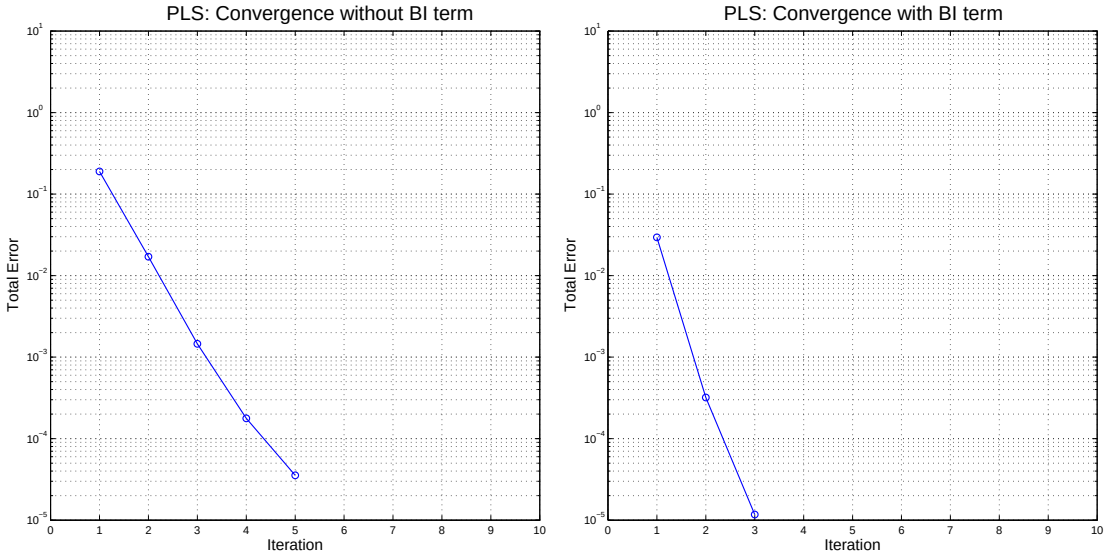
(b) GLS (without BI)

(c) PLS (with BI)

(d) GLS (with BI) and OSS (with $\mathbf{T}$ and $\mathbf{T}_d$)

Figure 86: Poiseuille flow in a trapezoidal domain. Results for the velocity v_1 (denoted u_x in the figure obtained with (a) PLS (without BI), (b) GLS (without BI), (c) PLS (with BI), (d) GLS (with BI) and OSS (with $\mathbf{T}$ and $\mathbf{T}_d$)

where, $(\bar{\mathbf{v}}, \bar{p})$ is the finite element approximation of the exact solution $(\mathbf{v}, p)$. The numerical solutions corresponding to the GLS, PLS, PLS+π and OSS methods, are compared with the following solutions: the nodally exact interpolant denoted by $(I_h\mathbf{v}, I_h p)$ and the best approximation (BA) with respect to the L^2 norm (for pressure) and the H^1 semi-norm (for velocity) denoted by $P_h^0 p$ and $P_h^1\mathbf{v}$ respectively. The solutions $I_h\mathbf{v}$, $I_h p$, $P_h^0 p$ and $P_h^1\mathbf{v}$ can be found as shown in Eq.(6.83).



(a) (b)

Figure 87: Poiseuille flow in a trapezoidal domain. Convergence of the PLS solution. (a) Without BI term. (b) With BI terms

$$I_h \mathbf{v} := N^a \mathbf{v}^a; \quad I_h p := N^a p^a \tag{6.83a}$$

$$\int_\Omega q^h (p - P_h^0 p)\, d\Omega = 0 \;\forall\; q^h \in Q^h \Rightarrow \parallel p - P_h^0 p \parallel_0 \leqslant \parallel p - \bar{p} \parallel_0 \forall\; \bar{p} \in Q^h \tag{6.83b}$$

$$\int_\Omega \boldsymbol{\nabla} \mathbf{v}^h : \boldsymbol{\nabla}(\mathbf{v} - P_h^1 \mathbf{v})\, d\Omega = 0 \;\forall\; \mathbf{v}^h \in V^h \Rightarrow \parallel \mathbf{v} - P_h^1 \mathbf{v} \parallel_1 \leqslant \parallel \mathbf{v} - \bar{\mathbf{v}} \parallel_1 \forall\; \bar{\mathbf{v}} \in V^h \tag{6.83c}$$

where, $(\mathbf{v}^a, p^a)$ represent the nodal values of the exact solution $(\mathbf{v}, p)$, $Q^h \subset L^2(\Omega)$ and $V^h \subset H_0^1(\Omega)$. In the current example Q^h and V^h are the solution spaces spanned by the 3-node triangle shape functions.

Consider a manufactured flow problem in which choosing the force term $\mathbf{f} = (f_1, f_2)$, with $f_1 = \mu(6x - 17)$, $f_2 = 0$, we have the exact solution to the Stokes problem as $\mathbf{v} = (v_1, v_2)$, with $v_1 = y(2-y)/2$, $v_2 = 0$ and $p = \mu(3x^2 - 18x + 1)$. Note that changing the magnitude of μ, any inf-sup stable Galerkin-FEM will give numerical solutions that would scale proportional to the exact solution. Thus, for this manufactured problem, the relative errors $e^h_{vH^1}$ and $e^h_{pL^2}$ will be independent of μ.

The $e^h_{vH^1}$ and $e^h_{pL^2}$ convergence rates for the GLS, PLS, PLS+π and OSS methods using $\mu = 1$ are shown in Figure 91.

Figure 91a illustrates the $e^h_{vH^1}$ error lines for the considered methods. The error line of the GLS method is slightly shifted above the error lines of $I_h \mathbf{v}$ (label: 'Interpolant') and $P_h^1 \mathbf{v}$ (label: 'H^1-BA'). The error line of the OSS method shows a slight deviation from those of $I_h \mathbf{v}$ and $P_h^1 \mathbf{v}$, but it quickly merges with the later error lines on mesh refinement. The error line of the PLS method shows an improvement over the GLS method on coarse meshes but this advantage is lost on finer meshes wherein the two

error lines merge. The error line of the PLS+π method practically coincides with the error lines of $I_h\mathbf{v}$ and $P_h^1\mathbf{v}$.

Figure 91b illustrates the $e^h_{pL^2}$ error lines for the considered methods. The error line of the GLS method not only shows the greatest deviation from the error lines of I_hp (label: 'Interpolant') and P_h^0p (label: 'L^2-BA') but also a sub-optimal convergence rate. The OSS method on the other hand show a convergence rate similar to that of I_hp and P_h^0p and the improvement over the GLS method is clear. The PLS and PLS+π methods show convergence rates similar to that of the GLS and the OSS methods respectively. Nevertheless, the error lines of the former two methods are located closer to the error lines of I_hp and P_h^0p showing improved accuracy in the considered norms. The increase in accuracy is particularly relevant for the PLS+π method as clearly seen in Figure 91b.

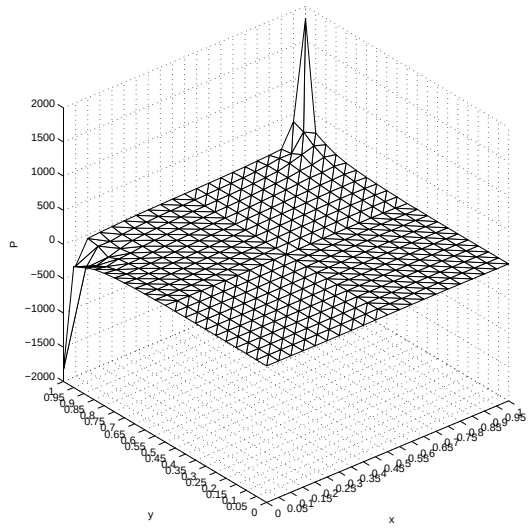
The PLS results for each of the meshes studied were obtained in some 10 iterations whereas typically 7 iterations were needed to obtain each of the PLS+π solutions.

6.16 CONCLUSIONS

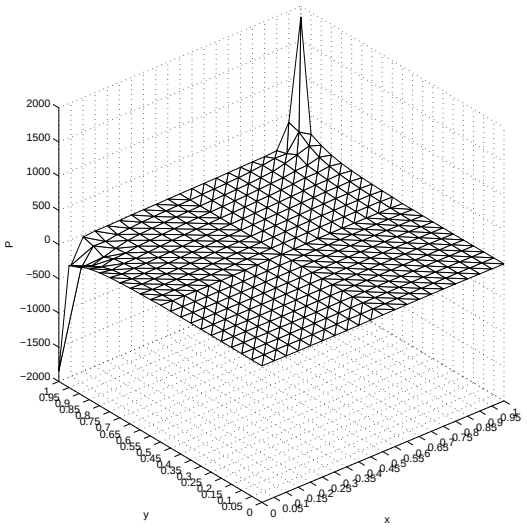
We have presented a family of stabilized finite element (FE) methods derived via first and second order the finite calculus (FIC) procedures. We have shown that several well known existing stabilized FE methods such as the penalty technique, the Galerkin Least Square (GLS) method, the Pressure Gradient Projection (PGP) method and the Orthogonal Sub-Scales (OSS) method are recovered from the general residual-based FIC stabilized form. New stabilized FE methods such as the Pressure Laplacian Stabilization (PLS) and the PLS+π method with consistent nonlinear forms of the stabilization parameters have been derived. The distinct feature of the PLS and PLS+π methods is that the stabilization terms depend on the discrete residuals of the momentum and the incompressibility equations.

The numerical results obtained for the Stokes problems solved in this work show that the PLS method and, in particular, the PLS+π method provide accurate solutions that improve in several cases the results of the traditional GLS and OSS methods. Results presented in [158] indicate that the PLS methods has also a superior performance than the PGP method for some problems.

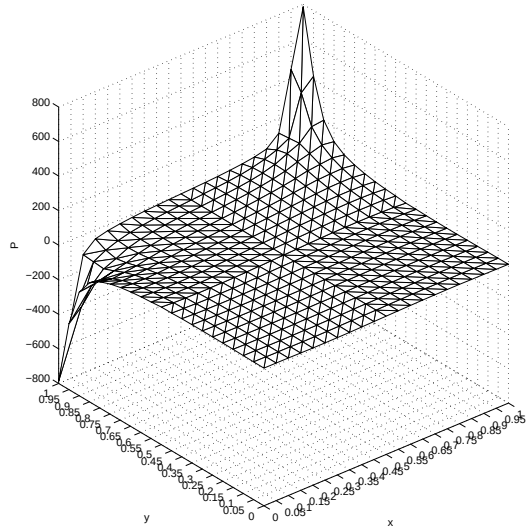
The prize to be paid for this increase in accuracy is the higher computational cost associated to the nonlinear solution which is intrinsic to the PLS method. The potential and advantages of the PLS method will be clearer for transient problems solved via staggered schemes or for nonlinear flow problems which invariably require an iterative scheme.



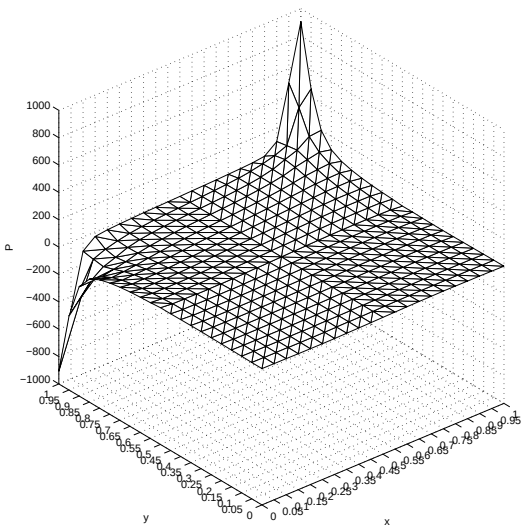
(a) PLS (without BI)



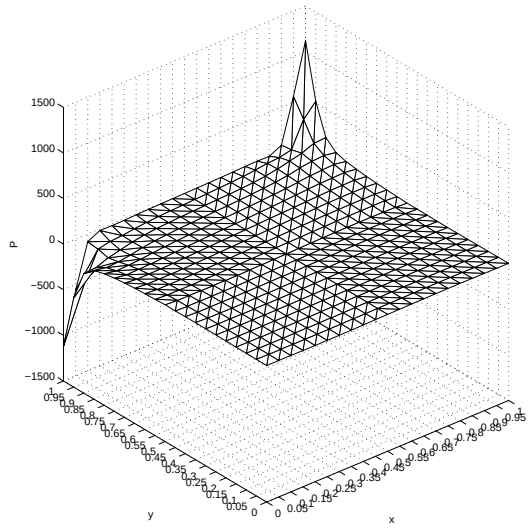
(b) PLS (with BI)



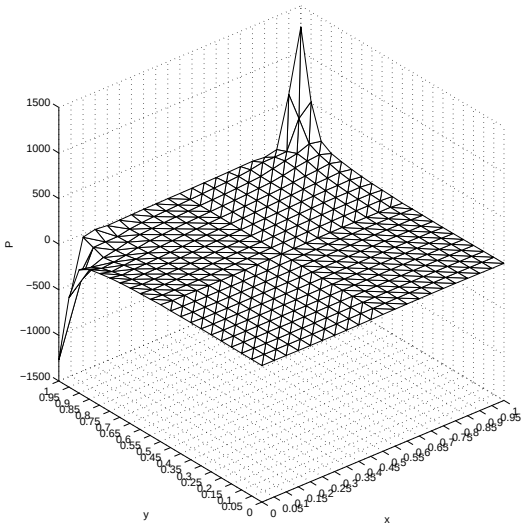
(c) GLS (without BI)



(d) GLS (with BI)

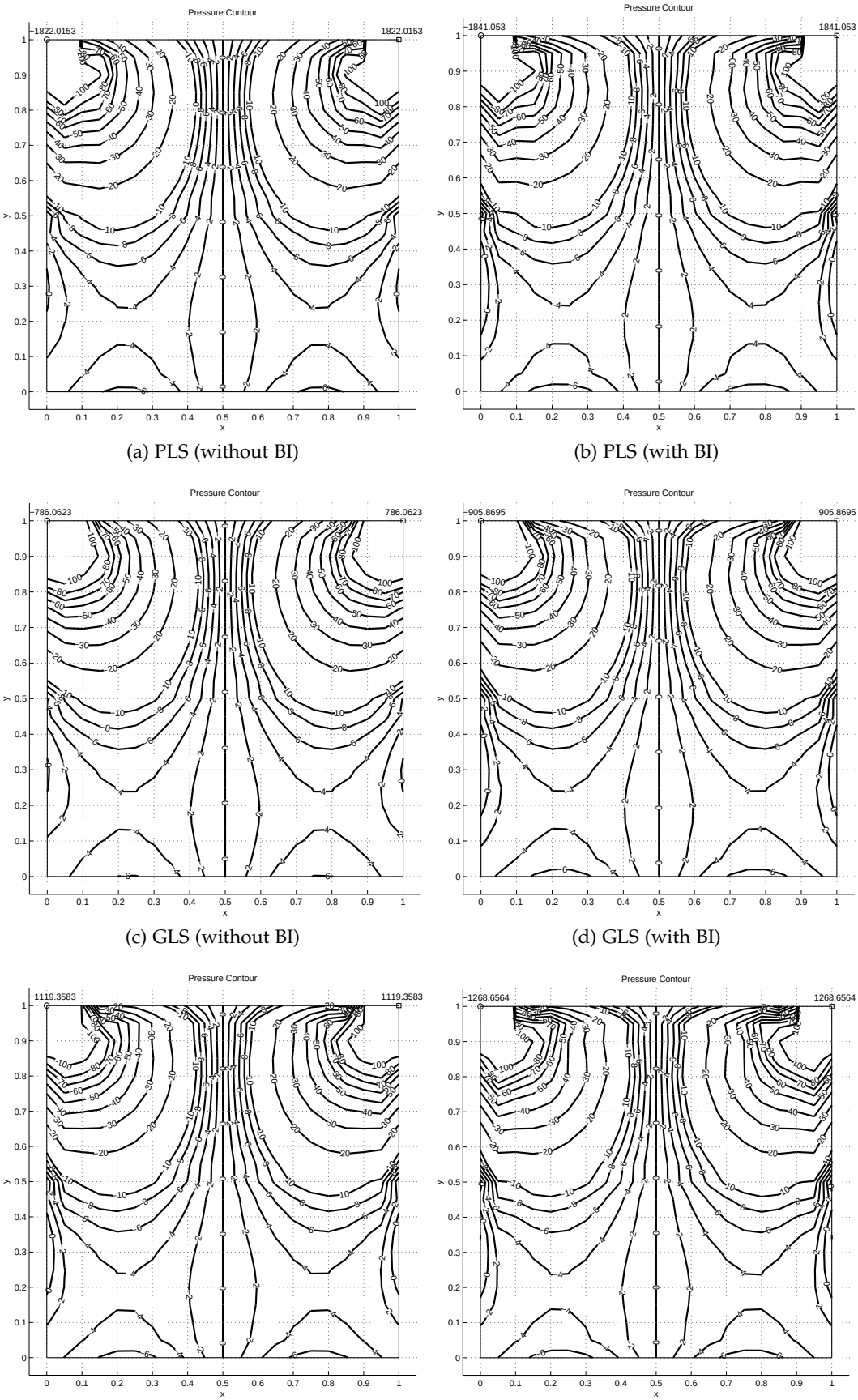


(e) OSS (with $\mathbf{T}_{\mathrm{d}}$)



(f) OSS (with $\mathbf{T}$)

Figure 88: Pressure elevation plots for the lid driven cavity problem.



(a) PLS (without BI)

(b) PLS (with BI)

(c) GLS (without BI)

(d) GLS (with BI)

(e) OSS (with $\mathbf{T}_d$)

(f) OSS (with $\mathbf{T}$)

Figure 89: Pressure contour plots for the lid driven cavity problem.

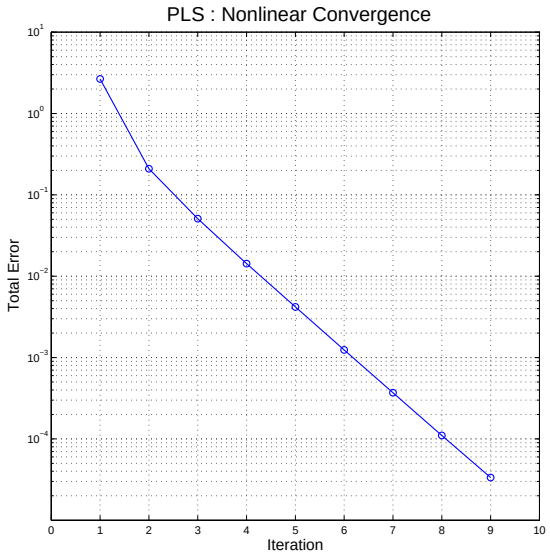
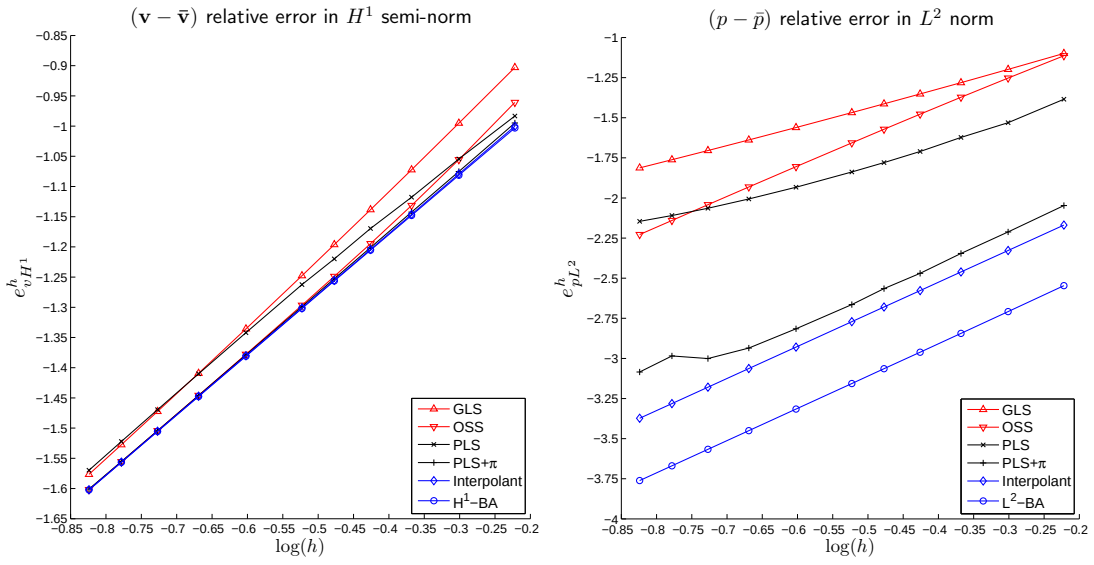


Figure 90: Lid driven cavity problem. Convergence of PLS results. Convergence curve with or without BI terms are similar



(a) (b)

Figure 91: Convergence rates for the GLS, OSS, PLS and PLS+π methods: (a) the relative errors in the velocity ($e^h_{vH^1}$) and (b) the relative errors in the pressure ($e^h_{pL^2}$)

I have fought the good fight, I have finished the race, I have kept the faith.

— 2 Timothy 4:7.

CONCLUSIONS

In this monograph, we have proposed the following new stabilized finite element (FE) based Petrov–Galerkin (PG) methods:

1. A high-resolution Petrov–Galerkin (HRPG) method for the singularly perturbed convection–diffusion–reaction (CDR) problem. The prefix high-resolution is used here in the sense popularized by Harten [82]—second-order accuracy for smooth regimes and good shock-capturing in non-regular regimes.

2. A PG method that reproduces on structured meshes the alpha-interpolation of the Galerkin finite element method (FEM) and the classical central finite difference method (FDM). Particularly for the Helmholtz problem, this method is capable of providing numerical solutions with a higher-order dispersion accuracy.

3. A FIC-based pressure Laplacian stabilization (PLS) method for the Stokes problem.

In the following, we describe the salient features of the work presented in the respective chapters. We refer to the individual chapter conclusions, viz. §1.7, §2.6, §3.6, §4.5, §5.7 and §6.16 for a more detailed summary of the same.

Chapter 1: To fix ideas prior to the development of the HRPG method, we have done a detailed analysis of a consistency recovery method for the 1D convection–diffusion equation stabilized using the SUPG [22, 92] or the FIC [141] method. The residual correction is done by including a convective projection variable into the stabilization term (motivated by the OSS [34] method). No gain in the dispersion accuracy is found by this procedure. For the steady-state case, the optimal expression of the stabilization parameter on uniform meshes is derived. An ad hoc extension of this stabilization parameter to non-uniform meshes failed to preclude the occurrence of weak node-to-node oscillations. Further, the discrete system obtained on uniform meshes using this optimal parameter is neither a matrix of positive-type nor a monotone matrix. Nevertheless, it verifies the necessary and sufficient condition given in [186] for a discrete maximum principle to hold. Unfortunately, all the conditions on the discrete system, except for it being a matrix of positive-type, are difficult to identify a priori. This poses a strategical difficulty in the design of shock-capturing methods and hence, this consistency recovery procedure is not preferred in the development of the HRPG method.

Chapter 2: The weak form associated with the HRPG method consists of the standard Galerkin terms, a linear upwinding term and a nonlinear shock-capturing term. The structure of the HRPG method in 1D (except for the definition of the stabilization parameters) is identical to that of the consistent approximate upwind (CAU) Petrov–Galerkin method [68]. The distinction is that in multi-dimensions the upwinding is

not streamline and the shock-capturing term is neither isotropic nor purely crosswind. Dropping the linear upwinding term, the expression in 1D of the stabilization parameter β multiplying the shock-capturing term is found by relating it with the diffusion introduced by the discrete-upwinding operation [120] on the Galerkin terms. It was pointed out earlier by Idelsohn et al. [99] that the transient term can be modeled as an instantaneous reaction term. Further, it was pointed out earlier by Codina [30] that the linear upwinding term can be interpreted as to contribute additional convection (negative upwind direction) and diffusion (rank one tensor) effects. Using these ideas the effective convection, diffusion and reaction coefficients (for the transient problem and using the linear upwinding term) are calculated. Thus, for the transient case and/or including the linear upwinding term, it is these effective coefficients that are used in the expression for β derived earlier. For the steady-state case, β depends only on the problem data, whereas for the transient case a nonlinear dependence exists which guarantees the independence of the steady-state solution on the time step used. Several 1D examples are presented that illustrate the good treatment of the global, Gibbs and dispersive oscillations that otherwise plague the numerical solution of the singularly perturbed CDR problem.

Chapter 3: A multi-dimensional extension of the HRPG method using multi-linear block FEs is presented. First, we design a nondimensional element number that quantifies the characteristic layers which are found only in higher dimensions. This is done by matching the width of the characteristic layers to the width of the parabolic layers found for a fictitious 1D reaction–diffusion problem. The nondimensional element number is then defined using this fictitious reaction coefficient, the diffusion coefficient and an appropriate element size. Next, we introduce anisotropic element length vectors $\mathbf{l}^i$ and the stabilization parameters α^i, β^i calculated along these $\mathbf{l}^i$. Except for the modification to include the new dimensionless number that quantifies the characteristic layers, the definition of α^i, β^i are a direct extension of their counterparts in 1D. Using α^i, β^i and $\mathbf{l}^i$, objective characteristic tensors associated with the HRPG method are defined. The numerical artifacts across the characteristic layers are manifested as the Gibbs phenomenon. Hence, we treat them just like the artifacts formed across the parabolic layers in the reaction-dominant case. Several 2D examples are presented that support the design objective—stabilization with high-resolution.

Chapter 4: We present a simple domain-based higher-order compact scheme involving two parameters α_1, α_2 on structured meshes for the Helmholtz equation. Making the parameters equal, we obtain the linear interpolation of the stencils obtained using the Galerkin FEM and the classical FDM. The stencil obtained by taking the parameters as distinct is denoted as the 'nonstandard compact stencil'. In this case, the diffusion and production terms obtained from the Galerkin FEM and the classical FDM stencils are interpolated distinctly. Generic expressions for the parameters are given that guarantee a dispersion accuracy of sixth-order should the parameters be distinct and fourth-order should they be equal. In the later case, an expression for the parameter is given that minimizes the maximum relative phase error of the scheme. Many existing higher-order compact schemes proposed within an algebraic setting can be recovered from the current scheme, viz. the alpha-interpolation method [110, 140] ($\alpha_1 = 0$, $\alpha_2 = \alpha$), the Vichnevetsky and Bowles scheme [187] ($\alpha_1 = \alpha$, $\alpha_2 = 1$), the fourth-order generalized Padé approximation studied in [81, 168] ($\alpha_1 = \alpha_2 = 0.5$) and the sixth-order quasi-stabilized FEM [10]. Recall that it is impossible to circumvent the

pollution effect associated with the Helmholtz equation (hence one can only propose quasi-stabilized methods) and that the maximum dispersion accuracy attainable on compact meshes is of sixth-order [10]. We remark that the proposed nonstandard compact stencil has an additional structure that reduces its abstractness. This additional structure (alternate point-of-view) is the key to develop PG methods which naturally extend these schemes to unstructured meshes.

Chapter 5: The basic idea is to design test functions whose inner product with the standard FE shape functions result in the lumped mass matrix. These test functions are designed to have the following features: a) to be piecewise polynomials of the same degree as the FE shape functions, b) to be a partition of unity (only in the sense that they add up to unity) and c) to have a compact support. The last condition allows us to construct test function spaces that vanish at the Dirichlet boundaries and thus, advocating its admittance into weak formulations. However, this condition makes these test functions discontinuous at the element boundaries. As the row lumping technique is a critical step in the design of these test functions (to fulfill the partition of unity constraint), the current work is restricted only to those FEs where this technique makes sense—simplicial FEs and multi-linear block FEs. We show that using these test functions with an appropriate single-valued model on the element boundaries, it is possible to recover the classical FDM stencil of the Helmholtz equation on structured meshes. The linear interpolation on the element boundaries (specified by α_1) and the element interiors (specified by α_2) of these test functions with the standard FE shape functions, will result in a new class of test functions. These new test functions define the proposed PG method involving two parameters α_1, α_2 that yields the nonstandard compact stencil of the Helmholtz equation on structured meshes. Making $\alpha_1 = \alpha_2$ we recover the linear interpolation of the Galerkin FEM and the classical FDM stencils on structured meshes. The proposed PG method provides the counterparts of these two schemes on unstructured meshes and allows the treatment of natural boundary conditions (Neumann or Robin) and the source terms in a straight-forward manner. The additional cost of implementation of the proposed PG method is just the evaluation of the element boundary integrals. All the algebraic evaluations are done at the element level unlike the QOPG method [128] where it is done at the patch level. This feature allows the proposed PG method to be easily incorporated within an 'assemble-by-elements' data structure. The choice of the parameters $\alpha_1 = \alpha_2 = (1/2)$ render the proposed PG method independent of the problem and mesh data. In this sense and for this choice, the proposed PG method could be labeled 'parameter-free'.

Chapter 6: New stabilized FE methods such as the pressure Laplacian stabilization (PLS) method and the PLS+π method with consistent nonlinear forms of the stabilization parameters are derived for the Stokes problem using the finite calculus (FIC) procedures. The distinct feature of these methods is that the stabilization terms depend on the discrete residuals of the momentum and the incompressibility equations. The PLS and PLS+π methods also introduce a boundary stabilization term. A similar boundary integral modification for the GLS method was proposed in [56]. Therein, it was shown that accounting for this term is relevant when lower-order FEs are used. On the contrary, residual correction techniques as proposed in [111] are required. Unlike for the GLS method, the adverse effect of excluding the boundary stabilization term is found to be small for the PLS method. We believe this is due to the nonlinear fine-tuning of the stabilization terms in the PLS method. The performance of the PLS

method is enhanced by adding to the momentum residual that appears in the stabilization terms its fine-scale projected counterpart (motivated by the OSS method) and thus, resulting in the PLS+π method. The numerical results obtained for the Stokes problems solved in this work show that the PLS method and, in particular, the PLS+π method provide accurate solutions that improve in several cases the results of the traditional GLS and OSS methods. The prize to be paid for this increase in accuracy is the higher computational cost associated to the nonlinear solution which is intrinsic to the PLS method. The potential and advantages of the PLS method will be clearer for transient problems solved via staggered schemes or for nonlinear flow problems which invariably require an iterative scheme.

OUTLOOK

In the following we list some aspects that are not addressed in this monograph which naturally constitute the future lines of work.

1. Simplicial FEs are not considered in the multi-dimensional extension of the HRPG method for the CDR problem. The primary obstacle in this line is to identify anisotropic element length vectors and the subsequent definition of the characteristic tensors that not only guarantee the objectivity of the method but also yield numerical solutions with a crisp layer resolution. The current proposal when applied to simplicial FEs violates the objectivity with respect to the admissible node numbering permutations.

2. For the new PG method proposed for the Helmholtz problem, it will be interesting to study the dispersion accuracy of stencils made up of a symmetric patch of simplicial elements. For instance, patches that have a honeycomb structure might yield stencils with higher-order dispersion accuracy. Following the approach taken for bilinear block FEs, it is possible to provide different models for the PG weights on the element boundaries. The benefits if any of this idea will be explored in future works. Finally, the application of the new PG method to engineering problems involving time-harmonic wave propagation phenomenon will also be explored.

3. For the convection–diffusion problem, the semi-discrete dispersion plots corresponding to the FIC/SUPG method (see Figure 1) reveals a possible gain in the dispersion accuracy by using an alpha-interpolated mass matrix in the transient terms. This is unlike what is observed for the Galerkin and the FIC_RC/OSS methods, where the best dispersion accuracy was obtained using the consistent mass matrix. Such an alpha-interpolated mass matrix can be attained in a variationally consistent manner using the PG formulation presented in chapter 5. Further studies will be done for the CDR problem exploring the benefits of using the test functions presented in chapter 5 to treat the unsolved issues related to the HRPG method—Gibbs phenomenon of the solution derivatives.

4. Study the Navier-Stokes equations wherein the momentum equation and the incompressibility constraint are stabilized using the HRPG and the PLS methods, respectively. Clearly, these two methods fall under the umbrella of FIC-based methods, of course with more elaborate stabilization terms.

BIBLIOGRAPHY

[1] J. E. Akin and Tayfun E. Tezduyar. Calculation of the advective limit of the SUPG stabilization parameter for linear and higher-order elements. *Computer Methods in Applied Mechanics and Engineering*, 193(21-22):1909–1922, May 2004. ISSN 00457825. doi: 10.1016/j.cma.2003.12.050. URL http://linkinghub.elsevier.com/retrieve/pii/S004578250400060X. (Cited on pages 157, 158, and 168.)

[2] J. E. Akin, Tayfun E. Tezduyar, M. Ungor, and Sanjay Mittal. Stabilization Parameters and Smagorinsky Turbulence Model. *Journal of Applied Mechanics*, 70 (1):2, 2003. ISSN 00218936. doi: 10.1115/1.1526569. URL http://link.aip.org/link/JAMCAV/v70/i1/p2/s1&Agg=doi. (Cited on pages 157, 158, and 168.)

[3] Gustavo Benitez Alvarez, Abimael F. D. Loula, Eduardo Gomes Dutra do Carmo, and Fernando Alves Rochinha. A discontinuous finite element formulation for Helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 195(33-36):4018–4035, July 2006. ISSN 00457825. doi: 10.1016/j.cma.2005.07.013. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782505003439. (Cited on pages 98 and 99.)

[4] Douglas N. Arnold, Franco Brezzi, Bernardo Cockburn, and L. Donatella Marini. Unified Analysis of Discontinuous Galerkin Methods for Elliptic Problems. *SIAM Journal on Numerical Analysis*, 39(5):1749–1779, 2002. ISSN 00361429. doi: 10.1137/S0036142901384162. URL http://link.aip.org/link/SJNAAM/v39/i5/p1749/s1&Agg=doi. (Cited on pages 133 and 134.)

[5] Abdul Kadir Aziz and A. Werschulz. On the Numerical Solutions of Helmholtz's Equation by the Finite Element Method. *SIAM Journal on Numerical Analysis*, 17(5):681–686, 1980. ISSN 00361429. doi: 10.1137/0717058. URL http://link.aip.org/link/SJNAAM/v17/i5/p681/s1&Agg=doi. (Cited on page 97.)

[6] Abdul Kadir Aziz, R. B. Kellogg, and A. B. Stephens. A two point boundary value problem with a rapidly oscillating solution. *Numerische Mathematik*, 53(1-2):107–121, January 1988. ISSN 0029-599X. doi: 10.1007/BF01395880. URL http://www.springerlink.com/index/10.1007/BF01395880. (Cited on page 97.)

[7] Ivo M. Babuska and J. M. Melenk. The partition of unity method. *International Journal for Numerical Methods in Engineering*, 40(4):727–758, February 1997. ISSN 0029-5981. doi: 10.1002/(SICI)1097-0207(19970228)40:4<727::AID-NME86>3.0.CO;2-N. URL http://doi.wiley.com/10.1002/(SICI)1097-0207(19970228)40:4<727::AID-NME86>3.0.CO;2-N. (Cited on page 98.)

[8] Ivo M. Babuska and J. E. Osborn. Generalized Finite Element Methods: Their Performance and Their Relation to Mixed Methods. *SIAM Journal on Numerical Analysis*, 20(3):510–536, 1983. ISSN 00361429. doi: 10.1137/0720034. URL http://link.aip.org/link/SJNAAM/v20/i3/p510/s1&Agg=doi. (Cited on page 98.)

[9] Ivo M. Babuska and Stefan A. Sauter. Is the Pollution Effect of the FEM Avoidable for the Helmholtz Equation Considering High Wave Numbers? *SIAM Journal on Numerical Analysis*, 34(6):2392–2423, 1997. ISSN 00361429. doi: 10.1137/S0036142994269186. URL http://link.aip.org/link/SJNAAM/v34/i6/p2392/s1&Agg=doi. (Cited on pages 97 and 98.)

[10] Ivo M. Babuska, Frank Ihlenburg, E.T. Paik, and Stefan A. Sauter. A Generalized Finite Element Method for solving the Helmholtz equation in two dimensions with minimal pollution. *Computer Methods in Applied Mechanics and Engineering*, 128(3-4):325–359, December 1995. ISSN 00457825. doi: 10.1016/0045-7825(95)00890-X. URL http://linkinghub.elsevier.com/retrieve/pii/004578259500890X. (Cited on pages iii, iv, v, 98, 100, 112, 114, 117, 141, 144, 147, 154, 190, and 191.)

[11] Santiago Badia and Ramon Codina. On a multiscale approach to the transient Stokes problem: Dynamic subscales and anisotropic space–time discretization. *Applied Mathematics and Computation*, 207(2):415–433, January 2009. ISSN 00963003. doi: 10.1016/j.amc.2008.10.059. URL http://linkinghub.elsevier.com/retrieve/pii/S0096300308008291. (Cited on pages 157, 158, and 170.)

[12] Claudio Baiocchi, Franco Brezzi, and Leopoldo Penna Franca. Virtual bubbles and Galerkin-least-squares type methods (Ga.L.S.). *Computer Methods in Applied Mechanics and Engineering*, 105(1):125–141, May 1993. ISSN 00457825. doi: 10.1016/0045-7825(93)90119-I. URL http://linkinghub.elsevier.com/retrieve/pii/004578259390119I. (Cited on page 37.)

[13] P Barbone and Isaac Harari. Nearly H^1-optimal finite element methods. *Computer Methods in Applied Mechanics and Engineering*, 190(43-44):5679–5690, August 2001. ISSN 00457825. doi: 10.1016/S0045-7825(01)00191-8. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782501001918. (Cited on page 98.)

[14] Yuri Bazilevs, Victor M. Calo, Tayfun E. Tezduyar, and Thomas J. R. Hughes. YZβ discontinuity capturing for advection-dominated processes with application to arterial drug delivery. *International Journal for Numerical Methods in Fluids*, 54(6-8):593–608, June 2007. ISSN 02712091. doi: 10.1002/fld.1484. URL http://doi.wiley.com/10.1002/fld.1484. (Cited on pages 157, 158, and 168.)

[15] Richard Bellman, B. G. Kashef, and J Casti. Differential quadrature: A technique for the rapid solution of nonlinear partial differential equations. *Journal of Computational Physics*, 10(1):40–52, August 1972. ISSN 00219991. doi: 10.1016/0021-9991(72)90089-7. URL http://linkinghub.elsevier.com/retrieve/pii/0021999172900897. (Cited on page 99.)

[16] Jay P. Boris and David L. Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works. *Journal of Computational Physics*, 11(1):38–69, January 1973. ISSN 00219991. doi: 10.1016/0021-9991(73)90147-2. URL http://linkinghub.elsevier.com/retrieve/pii/0021999173901472. (Cited on page 38.)

[17] James H. Bramble and Bert E. Hubbard. New monotone type approximations for elliptic problems. *Mathematics of Computation*, 18(87):349–349, September 1964. ISSN 0025-5718. doi: 10.1090/S0025-5718-1964-0165702-X. URL http://www.ams.org/jourcgi/jour-getitem?pii=S0025-5718-1964-0165702-X. (Cited on page 29.)

[18] Franco Brezzi and Alessandro Russo. Choosing bubbles for advection–diffusion problems. *Mathematical Models and Methods in Applied Sciences*, 4(4):571–587, 1994. (Cited on page 37.)

[19] Franco Brezzi, Marie-Odile Bristeau, Leopoldo Penna Franca, Michel Mallet, and Gilbert Rogé. A relationship between stabilized finite element methods and the Galerkin method with bubble functions. *Computer Methods in Applied Mechanics and Engineering*, 96(1):117–129, April 1992. ISSN 00457825. doi: 10.1016/0045-7825(92)90102-P. URL http://linkinghub.elsevier.com/retrieve/pii/004578259290102P. (Cited on page 37.)

[20] Franco Brezzi, Leopoldo Penna Franca, Thomas J. R. Hughes, and Alessandro Russo. $b = \int g$. *Computer Methods in Applied Mechanics and Engineering*, 145(3-4):329–339, June 1997. ISSN 00457825. doi: 10.1016/S0045-7825(96)01221-2. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782596012212. (Cited on page 37.)

[21] Franco Brezzi, Guillermo Hauke, L. D. Marini, and Giancarlo Sangalli. Link-cutting bubbles for the stabilization of convection–diffusion–reaction problems. *Mathematical Models and Methods in Applied Sciences*, 13(3):445–461, 2003. (Cited on pages 38 and 39.)

[22] Alexander N Brooks and Thomas J. R. Hughes. Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 32(1-3):199–259, September 1982. ISSN 00457825. doi: 10.1016/0045-7825(82)90071-8. URL http://linkinghub.elsevier.com/retrieve/pii/0045782582900718. (Cited on pages 37, 76, 79, 157, 158, and 189.)

[23] Erik Burman and Alexandre Ern. Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion–reaction equation. *Computer Methods in Applied Mechanics and Engineering*, 191(35):3833–3855, July 2002. ISSN 00457825. doi: 10.1016/S0045-7825(02)00318-3. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782502003183. (Cited on pages 37 and 42.)

[24] Michael Casey and Torsten Wintergerste, editors. *ERCOFTAC best practice guidelines for Industrial Computational Fluid Dynamics*. Ercoftac, Brussels, 2000. (Cited on page 82.)

[25] Olivier Cessenat and Bruno Despres. Application of an Ultra Weak Variational Formulation of Elliptic PDEs to the Two-Dimensional Helmholtz Problem. *SIAM Journal on Numerical Analysis*, 35(1):255, 1998. ISSN 00361429. doi: 10.1137/S0036142995285873. URL http://link.aip.org/link/SJNAAM/v35/i1/p255/s1&Agg=doi. (Cited on page 99.)

[26] Mark A. Christon, Mario J. Martinez, and Thomas E. Voth. Generalized Fourier analyses of the advection–diffusion equation—Part I: one-dimensional domains. *International Journal for Numerical Methods in Fluids*, 45(8):839–887, July 2004. ISSN 0271-2091. doi: 10.1002/fld.719. URL http://doi.wiley.com/10.1002/fld.719. (Cited on page 8.)

[27] Shung Chu-hua. The general derivation of Ritz method and Trefftz method in elastomechanics. *Applied Mathematics and Mechanics*, 3(5):739–748, 1982. ISSN 0253-4827. doi: 10.1007/BF01875738. URL http://www.springerlink.com/index/10.1007/BF01875738. (Cited on page 97.)

[28] J Cipolla. Subgrid modeling in a Galerkin method for the Helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 177(1-2):35–49, July 1999. ISSN 00457825. doi: 10.1016/S0045-7825(98)00276-X. URL http://linkinghub.elsevier.com/retrieve/pii/S004578259800276X. (Cited on page 98.)

[29] Ramon Codina. A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection–diffusion equation. *Computer Methods in Applied Mechanics and Engineering*, 110(3-4):325–342, December 1993. ISSN 00457825. doi: 10.1016/0045-7825(93)90213-H. URL http://linkinghub.elsevier.com/retrieve/pii/004578259390213H. (Cited on pages 37, 40, and 41.)

[30] Ramon Codina. A shock-capturing anisotropic diffusion for the finite element solution of the diffusion–convection–reaction equation. In Kenneth Morgan, Eugenio Oñate, Jacques Périaux, Jaume Peraire, and Olgierd C Zienkiewicz, editors, *Finite Elements in Fluids: new trends and applications*, volume 1, pages 67–75. CIMNE, Barcelona, 1993. ISBN 84-87867-29-4. (Cited on pages 37, 57, and 190.)

[31] Ramon Codina. *A Finite Element Formulation for the Numerical Solution of the Convection–Diffusion Equation*. CIMNE, Barcelona, first edition, 1993. ISBN 84-87867-17-0. URL http://www.cimne.com/tiendaCIMNE/ProductosCon.asp?id_prod=56. (Cited on page 87.)

[32] Ramon Codina. Comparison of some finite element methods for solving the diffusion–convection–reaction equation. *Computer Methods in Applied Mechanics and Engineering*, 156(1-4):185–210, April 1998. ISSN 00457825. doi: 10.1016/S0045-7825(97)00206-5. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782597002065. (Cited on pages 7 and 37.)

[33] Ramon Codina. On stabilized finite element methods for linear systems of convection–diffusion–reaction equations. *Computer Methods in Applied Mechanics and Engineering*, 188(1-3):61–82, July 2000. ISSN 00457825. doi: 10.1016/S0045-7825(00)00177-8. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782500001778. (Cited on page 38.)

[34] Ramon Codina. Stabilization of incompressibility and convection through orthogonal sub-scales in finite element methods. *Computer Methods in Applied Mechanics and Engineering*, 190(13-14):1579–1599, December 2000. ISSN 00457825.

doi: 10.1016/S0045-7825(00)00254-1. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782500002541. (Cited on pages iii, v, 2, 7, 9, 11, 157, 158, 170, and 189.)

[35] Ramon Codina and Jordi Blasco. A finite element formulation for the Stokes problem allowing equal velocity-pressure interpolation. *Computer Methods in Applied Mechanics and Engineering*, 143(3-4):373–391, April 1997. ISSN 00457825. doi: 10.1016/S0045-7825(96)01154-1. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782596011541. (Cited on pages iii, v, 157, 169, 170, and 177.)

[36] Ramon Codina and Jordi Blasco. Stabilized finite element method for the transient Navier–Stokes equations based on a pressure gradient projection. *Computer Methods in Applied Mechanics and Engineering*, 182(3-4):277–300, February 2000. ISSN 00457825. doi: 10.1016/S0045-7825(99)00194-2. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782599001942. (Cited on pages 157, 168, 169, and 170.)

[37] Ramon Codina and Olgierd C. Zienkiewicz. CBS versus GLS stabilization of the incompressible Navier–Stokes equations and the role of the time step as stabilization parameter. *Communications in Numerical Methods in Engineering*, 18 (2):99–112, December 2001. ISSN 10698299. doi: 10.1002/cnm.470. URL http://doi.wiley.com/10.1002/cnm.470.

[38] Ramon Codina, Mariano Vázquez, and Olgierd C. Zienkiewicz. A general algorithm for compressible and incompressible flows. Part III: The semi-implicit form. *International Journal for Numerical Methods in Fluids*, 27(1-4):13–32, January 1998. ISSN 0271-2091. doi: 10.1002/(SICI)1097-0363(199801)27:1/4<13::AID-FLD647>3.0.CO;2-8. URL http://doi.wiley.com/10.1002/%28SICI%291097-0363%28199801%2927%3A1/4%3C13%3A%3AAID-FLD647%3E3.0.CO%3B2-8.

[39] Ramon Codina, Javier Principe, Oriol Guasch, and Santiago Badia. Time dependent subscales in the stabilized finite element approximation of incompressible flow problems. *Computer Methods in Applied Mechanics and Engineering*, 196(21-24):2413–2430, April 2007. ISSN 00457825. doi: 10.1016/j.cma.2007.01.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782507000035. (Cited on pages 157, 158, and 170.)

[40] Lothar Collatz. *The numerical treatment of differential equations*. Springer, Berlin, 3rd edition, 1960. (Cited on pages 29 and 110.)

[41] Richard Courant and David Hilbert. *Methods of mathematical physics*, volume 1. Wiley-Interscience, New York, 1989. (Cited on page 97.)

[42] Marcela A. Cruchaga and Eugenio Oñate. A finite element formulation for incompressible flow problems using a generalized streamline operator. *Computer Methods in Applied Mechanics and Engineering*, 143(1-2):49–67, April 1997. ISSN 00457825. doi: 10.1016/S0045-7825(97)84579-3. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782597845793. (Cited on pages 157 and 158.)

[43] Marcela A. Cruchaga and Eugenio Oñate. A generalized streamline finite element approach for the analysis of incompressible flow problems including moving surfaces. *Computer Methods in Applied Mechanics and Engineering*, 173(1-2): 241–255, April 1999. ISSN 00457825. doi: 10.1016/S0045-7825(98)00272-2. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782598002722. (Cited on pages 157 and 158.)

[44] Sanford Davis. A Space-Time Discretization Procedure for Wave Propagation Problems. In *Nasa Technical Memorandum NASA-TM-102215*. 1989. (Cited on pages 8 and 10.)

[45] Monica de Mier-Torrecilla. *Numerical Simulation of Multi-Fluid Flows with the Particle Finite Element Method*. Phd thesis, Technical University of Catalonia (UPC), 2010. (Cited on pages 158 and 170.)

[46] P. A. B. de Sampaio and A. L. G. A. Coutinho. A natural derivation of discontinuity capturing operator for convection–diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 190(46-47):6291–6308, September 2001. ISSN 00457825. doi: 10.1016/S0045-7825(01)00229-8. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782501002298. (Cited on page 37.)

[47] Leszek Demkowicz. Asymptotic convergence in finite and boundary element methods: part 1: theoretical results. *Computers & Mathematics with Applications*, 27(12):69–84, June 1994. ISSN 08981221. doi: 10.1016/0898-1221(94)90087-6. URL http://linkinghub.elsevier.com/retrieve/pii/0898122194900876. (Cited on pages 1, 97, 106, 124, and 147.)

[48] Eduardo Gomes Dutra do Carmo and Gustavo Benitez Alvarez. A new stabilized finite element formulation for scalar convection–diffusion problems: the streamline and approximate upwind/Petrov–Galerkin method. *Computer Methods in Applied Mechanics and Engineering*, 192(31-32):3379–3396, August 2003. ISSN 00457825. doi: 10.1016/S0045-7825(03)00292-5. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782503002925. (Cited on pages 37, 41, and 42.)

[49] Eduardo Gomes Dutra do Carmo and Gustavo Benitez Alvarez. A new upwind function in stabilized finite element formulations, using linear and quadratic elements for scalar convection–diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 193(23-26):2383–2402, June 2004. ISSN 00457825. doi: 10.1016/j.cma.2004.01.015. URL http://linkinghub.elsevier.com/retrieve/pii/S004578250400088X.

[50] Eduardo Gomes Dutra do Carmo and Augusto Cesar Galeão. Feedback Petrov–Galerkin methods for convection-dominated problems. *Computer Methods in Applied Mechanics and Engineering*, 88(1):1–16, June 1991. ISSN 00457825. doi: 10.1016/0045-7825(91)90231-T. URL http://linkinghub.elsevier.com/retrieve/pii/004578259190231T. (Cited on pages 37, 41, and 42.)

[51] Eduardo Gomes Dutra do Carmo, Gustavo Benitez Alvarez, Abimael F. D. Loula, and Fernando Alves Rochinha. A nearly optimal Galerkin projected

residual finite element method for Helmholtz problem. *Computer Methods in Applied Mechanics and Engineering*, 197(13-16):1362–1375, February 2008. ISSN 00457825. doi: 10.1016/j.cma.2007.11.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782507004562. (Cited on pages 99 and 128.)

[52] Clark R. Dohrmann and Pavel B. Bochev. A stabilized finite element method for the Stokes problem based on polynomial pressure projections. *International Journal for Numerical Methods in Fluids*, 46(2):183–201, September 2004. ISSN 0271-2091. doi: 10.1002/fld.752. URL http://doi.wiley.com/10.1002/fld.752. (Cited on pages 157 and 158.)

[53] Jean Donea. A Taylor–Galerkin method for convective transport problems. *International Journal for Numerical Methods in Engineering*, 20(1):101–119, January 1984. ISSN 0029-5981. doi: 10.1002/nme.1620200108. URL http://doi.wiley.com/10.1002/nme.1620200108. (Cited on page 37.)

[54] Jean Donea and Antonio Huerta. *Finite Element Methods for Flow Problems*. Wiley, Chichester, 2003. ISBN 978-0-471-49666-3. (Cited on pages 157, 158, 164, 165, 166, 168, 169, and 176.)

[55] Jim Douglas, Jr. and Thomas F. Russell. Numerical Methods for Convection-Dominated Diffusion Problems Based on Combining the Method of Characteristics with Finite Element or Finite Difference Procedures. *SIAM Journal on Numerical Analysis*, 19(5):871, 1982. ISSN 00361429. doi: 10.1137/0719063. URL http://link.aip.org/link/SJNAAM/v19/i5/p871/s1&Agg=doi. (Cited on page 37.)

[56] Jean-Jacques Droux and Thomas J. R. Hughes. A boundary integral modification of the Galerkin least squares formulation for the Stokes problem. *Computer Methods in Applied Mechanics and Engineering*, 113(1-2):173–182, March 1994. ISSN 00457825. doi: 10.1016/0045-7825(94)90217-8. URL http://linkinghub.elsevier.com/retrieve/pii/0045782594902178. (Cited on pages 157, 158, 166, and 191.)

[57] Charbel Farhat, Isaac Harari, and Leopoldo Penna Franca. The discontinuous enrichment method. *Computer Methods in Applied Mechanics and Engineering*, 190(48):6455–6479, September 2001. ISSN 00457825. doi: 10.1016/S0045-7825(01)00232-8. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782501002328. (Cited on page 98.)

[58] Charbel Farhat, Isaac Harari, and U Hetmaniuk. A discontinuous Galerkin method with Lagrange multipliers for the solution of Helmholtz problems in the mid-frequency regime. *Computer Methods in Applied Mechanics and Engineering*, 192(11-12):1389–1419, March 2003. ISSN 00457825. doi: 10.1016/S0045-7825(02)00646-1. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782502006461. (Cited on page 98.)

[59] Carlos A. Felippa and Eugenio Oñate. Nodally exact Ritz discretizations of 1D diffusion–absorption and Helmholtz equations by variational FIC and modified equation methods. *Computational Mechanics*, 39(2):91–111, January 2007. ISSN

0178-7675. doi: 10.1007/s00466-005-0011-z. URL http://www.springerlink.com/index/10.1007/s00466-005-0011-z. (Cited on pages 38, 39, and 99.)

[60] Daniel T. Fernandes and Abimael F. D. Loula. Quasi optimal finite difference method for Helmholtz problem on unstructured grids. *International Journal for Numerical Methods in Engineering*, 82:1244–1281, 2009. ISSN 00295981. doi: 10.1002/nme.2795. URL http://doi.wiley.com/10.1002/nme.2795. (Cited on pages 77, 99, and 144.)

[61] Leopoldo Penna Franca and Eduardo Gomes Dutra do Carmo. The Galerkin gradient least-squares method. *Computer Methods in Applied Mechanics and Engineering*, 74(1):41–54, September 1989. ISSN 00457825. doi: 10.1016/0045-7825(89)90085-6. URL http://linkinghub.elsevier.com/retrieve/pii/0045782589900856. (Cited on page 38.)

[62] Leopoldo Penna Franca and Charbel Farhat. Bubble functions prompt unusual stabilized finite element methods. *Computer Methods in Applied Mechanics and Engineering*, 123(1-4):299–308, June 1995. ISSN 00457825. doi: 10.1016/0045-7825(94)00721-X. URL http://linkinghub.elsevier.com/retrieve/pii/004578259400721X. (Cited on page 38.)

[63] Leopoldo Penna Franca and Sergio L. Frey. Stabilized finite element methods: II. The incompressible Navier–Stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 99(2-3):209–233, September 1992. ISSN 00457825. doi: 10.1016/0045-7825(92)90041-H. URL http://linkinghub.elsevier.com/retrieve/pii/004578259290041H. (Cited on pages 157 and 158.)

[64] Leopoldo Penna Franca and Alessandro Russo. Deriving upwinding, mass lumping and selective reduced integration by residual-free bubbles. *Applied Mathematics Letters*, 9(5):83–88, September 1996. ISSN 08939659. doi: 10.1016/0893-9659(96)00078-X. URL http://linkinghub.elsevier.com/retrieve/pii/089396599600078X. (Cited on page 131.)

[65] Leopoldo Penna Franca and Alessandro Russo. Mass lumping emanating from residual-free bubbles. *Computer Methods in Applied Mechanics and Engineering*, 142(3-4):353–360, March 1997. ISSN 00457825. doi: 10.1016/S0045-7825(96)01137-1. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782596011371. (Cited on page 131.)

[66] Leopoldo Penna Franca and F. Valentin. On an improved unusual stabilized finite element method for the advective–reactive–diffusive equation. *Computer Methods in Applied Mechanics and Engineering*, 190(13-14):1785–1800, December 2000. ISSN 00457825. doi: 10.1016/S0045-7825(00)00190-0. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782500001900. (Cited on page 38.)

[67] Leopoldo Penna Franca, Charbel Farhat, Antonini P. Macedo, and Michel Lesoinne. Residual-free bubbles for the Helmholtz equation. *International Journal for Numerical Methods in Engineering*, 40(21):4003–4009, November 1997. ISSN 0029-5981. doi: 10.1002/(SICI)1097-0207(19971115)40:21<4003::AID-NME199>3.

o.CO;2-Z. URL http://doi.wiley.com/10.1002/(SICI)1097-0207(19971115)40:21<4003::AID-NME199>3.0.CO;2-Z. (Cited on page 98.)

[68] Augusto Cesar Galeão and Eduardo Gomes Dutra do Carmo. A consistent approximate upwind Petrov–Galerkin method for convection-dominated problems. *Computer Methods in Applied Mechanics and Engineering*, 68(1):83–95, May 1988. ISSN 00457825. doi: 10.1016/0045-7825(88)90108-9. URL http://linkinghub.elsevier.com/retrieve/pii/0045782588901089. (Cited on pages iii, iv, 2, 37, 41, 93, and 189.)

[69] Charles I. Goldstein. The weak element method applied to Helmholtz type equations. *Applied Numerical Mathematics*, 2(3-5):409–426, October 1986. ISSN 01689274. doi: 10.1016/0168-9274(86)90043-7. URL http://linkinghub.elsevier.com/retrieve/pii/0168927486900437. (Cited on page 99.)

[70] Gerald Lee Goudreau. *Evaluation of discrete methods for the linear dynamic response of elastic and viscoelastic solids*. PhD thesis, University of California, Berkeley, CA, 1970. (Cited on pages 3, 100, and 103.)

[71] Gerald Lee Goudreau and Robert L. Taylor. Evaluation of numerical integration methods in elastodynamics. *Computer Methods in Applied Mechanics and Engineering*, 2(1):69–97, February 1973. ISSN 00457825. doi: 10.1016/0045-7825(73)90023-6. URL http://linkinghub.elsevier.com/retrieve/pii/0045782573900236. (Cited on pages 3 and 100.)

[72] P M Gresho and R L Sani. The advection–diffusion equation. In *Incompressible flow and finite element method: Advection–Diffusion and Isothermal Laminar Flow*, volume 1. John Wiley & Sons, Chichester, 2000. ISBN 978-0-471-49249-8. (Cited on page 7.)

[73] Roger Grimshaw. Group velocity. In A C Scott, editor, *Encyclopedia of Nonlinear Science*, pages 385–388. Taylor and Francis, New York, 2004. ISBN 9780203647417. (Cited on page 10.)

[74] Oriol Guasch and Ramon Codina. An algebraic subgrid scale finite element method for the convected Helmholtz equation in two dimensions with applications in aeroacoustics. *Computer Methods in Applied Mechanics and Engineering*, 196(45-48):4672–4689, September 2007. ISSN 00457825. doi: 10.1016/j.cma.2007.06.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782507002381. (Cited on page 98.)

[75] Max D. Gunzburger. *Finite Element Methods for Viscous Incompressible Flows: A Guide to Theory, Practice, and Algorithms*. Computer Science and Scientific Computing. Academic Press, London, 1989. ISBN 0123073502. (Cited on page 1.)

[76] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin/Heidelberg, 2006. ISBN 3-540-30663-3. doi: 10.1007/3-540-30666-8. URL http://www.springerlink.com/index/10.1007/3-540-30666-8. (Cited on page 86.)

[77] Isaac Harari. A survey of finite element methods for time-harmonic acoustics. *Computer Methods in Applied Mechanics and Engineering*, 195(13-16):1594–1607, February 2006. ISSN 00457825. doi: 10.1016/j.cma.2005.05.030. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782505002975. (Cited on page 99.)

[78] Isaac Harari and Kirill Gosteev. Bubble-based stabilization for the Helmholtz equation. *International Journal for Numerical Methods in Engineering*, 70(10):1241–1260, June 2007. ISSN 00295981. doi: 10.1002/nme.1930. URL http://doi.wiley.com/10.1002/nme.1930. (Cited on page 98.)

[79] Isaac Harari and Thomas J. R. Hughes. Finite element methods for the helmholtz equation in an exterior domain: Model problems. *Computer Methods in Applied Mechanics and Engineering*, 87(1):59–96, May 1991. ISSN 00457825. doi: 10.1016/0045-7825(91)90146-W. URL http://linkinghub.elsevier.com/retrieve/pii/004578259190146W. (Cited on pages 98, 104, 124, 144, and 147.)

[80] Isaac Harari and Thomas J. R. Hughes. Stabilized finite element methods for steady advection–diffusion with production. *Computer Methods in Applied Mechanics and Engineering*, 115(1-2):165–191, 1994. ISSN 00457825. doi: 10.1016/0045-7825(94)90193-7. URL http://linkinghub.elsevier.com/retrieve/pii/0045782594901937. (Cited on pages 38, 39, and 98.)

[81] Isaac Harari and Eli Turkel. Accurate Finite Difference Methods for Time-Harmonic Wave Propagation. *Journal of Computational Physics*, 119(2):252–270, July 1995. ISSN 00219991. doi: 10.1006/jcph.1995.1134. URL http://linkinghub.elsevier.com/retrieve/doi/10.1006/jcph.1995.1134. (Cited on pages iii, iv, 99, 100, 103, 109, 141, 145, and 190.)

[82] Amiram Harten. High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics*, 49(3):357–393, March 1983. ISSN 00219991. doi: 10.1016/0021-9991(83)90136-5. URL http://linkinghub.elsevier.com/retrieve/pii/0021999183901365. (Cited on pages 2, 38, 45, 46, and 189.)

[83] Amiram Harten, B Engquist, S Osher, and S Chakravarthy. Uniformly high order accurate essentially non-oscillatory schemes, III. *Journal of Computational Physics*, 71(2):231–303, August 1987. ISSN 00219991. doi: 10.1016/0021-9991(87)90031-3. URL http://linkinghub.elsevier.com/retrieve/pii/0021999187900313. (Cited on page 38.)

[84] Guillermo Hauke. A simple subgrid scale stabilized method for the advection–diffusion–reaction equation. *Computer Methods in Applied Mechanics and Engineering*, 191(27-28):2925–2947, April 2002. ISSN 00457825. doi: 10.1016/S0045-7825(02)00217-7. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782502002177. (Cited on page 38.)

[85] Guillermo Hauke and Antonio García-Olivares. Variational subgrid scale formulations for the advection–diffusion–reaction equation. *Computer Methods in Applied Mechanics and Engineering*, 190(51-52):6847–6865, October 2001. ISSN 00457825. doi: 10.1016/S0045-7825(01)00262-6. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782501002626. (Cited on page 38.)

[86] Guillermo Hauke, Giancarlo Sangalli, and Mohamed H Doweidar. Combining adjoint stabilized methods for the advection–diffusion–reaction problem. *Mathematical Models and Methods in Applied Sciences*, 17(2):305–326, 2007. (Cited on pages 38, 39, and 98.)

[87] Ismael Herrera. Trefftz method. In Carlos A Brebbia, editor, *Topics in boundary element research: Basic principles and applications*, volume 1, pages 230–258. Springer–Verlag, New York, 1984. (Cited on page 99.)

[88] Ismael Herrera. Trefftz method: A general theory. *Numerical Methods for Partial Differential Equations*, 16(6):561–580, November 2000. ISSN 0749-159X. doi: 10.1002/1098-2426(200011)16:6<561::AID-NUM4>3.0.CO;2-V. URL http://doi.wiley.com/10.1002/1098-2426(200011)16:6<561::AID-NUM4>3.0.CO;2-V. (Cited on page 99.)

[89] Charles Hirsch. *Numerical computation of internal and external flows*, volume 1 and 2. Wiley, 1990. (Cited on page 38.)

[90] M.-C. Hsu, Yuri Bazilevs, Victor M. Calo, Tayfun E. Tezduyar, and Thomas J. R. Hughes. Improving stability of stabilized and multiscale formulations in flow simulations at small time steps. *Computer Methods in Applied Mechanics and Engineering*, 199(13-16):828–840, February 2010. ISSN 00457825. doi: 10.1016/j.cma.2009.06.019. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782509002254. (Cited on pages 157, 158, 166, 168, and 169.)

[91] Thomas J. R. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Computer Methods in Applied Mechanics and Engineering*, 127(1-4):387–401, November 1995. ISSN 00457825. doi: 10.1016/0045-7825(95)00844-9. URL http://linkinghub.elsevier.com/retrieve/pii/0045782595008449. (Cited on page 37.)

[92] Thomas J. R. Hughes and Alexander N Brooks. A theoretical framework for Petrov–Galerkin methods with discontinuous weighting functions: application to the streamline upwind procedure. In R.H. Gallagher, D.M. Norrie, J.T. Oden, and Olgierd C. Zienkiewicz, editors, *Finite Elements in Fluids*, volume IV, pages 47–65. John Wiley and Sons Ltd, Chichester, 1982. (Cited on pages 37, 40, 41, 76, 133, and 189.)

[93] Thomas J. R. Hughes, Leopoldo Penna Franca, and Marc Balestra. A new finite element formulation for computational fluid dynamics: V. Circumventing the Babuska-Brezzi condition: a stable Petrov–Galerkin formulation of the stokes problem accommodating equal-order interpolations. *Computer Methods in Applied Mechanics and Engineering*, 59(1):85–99, November 1986. ISSN 00457825. doi: 10.1016/0045-7825(86)90025-3. URL http://linkinghub.elsevier.com/retrieve/pii/0045782586900253. (Cited on pages iii, v, 157, 158, and 166.)

[94] Thomas J. R. Hughes, Michel Mallet, and Akira Mizukami. A new finite element formulation for computational fluid dynamics: II. Beyond SUPG. *Computer Methods in Applied Mechanics and Engineering*, 54(3):341–355, March 1986.

ISSN 00457825. doi: 10.1016/0045-7825(86)90110-6. URL http://linkinghub.elsevier.com/retrieve/pii/0045782586901106. (Cited on pages 37, 40, and 41.)

[95] Thomas J. R. Hughes, Leopoldo Penna Franca, and Gregory M Hulbert. A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/least-squares method for advective–diffusive equations. *Computer Methods in Applied Mechanics and Engineering*, 73:173–189, 1989. doi: 10.1016/0045-7825(89)90111-4. (Cited on page 37.)

[96] Thomas J. R. Hughes, Guillermo Hauke, and Kenneth E. Jansen. *Stabilized finite element methods in fluids: Inspirations, origins, status and recent developments.*, page 317. CIMNE, Barcelona, 1994. ISBN 84-87867-45-6. (Cited on pages 157 and 158.)

[97] Thomas J. R. Hughes, Gonzalo R. Feijóo, Luca Mazzei, and Jean-Baptiste Quincy. The variational multiscale method—a paradigm for computational mechanics. *Computer Methods in Applied Mechanics and Engineering*, 166(1-2):3–24, November 1998. ISSN 00457825. doi: 10.1016/S0045-7825(98)00079-6. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782598000796. (Cited on page 37.)

[98] Thomas J. R. Hughes, Luca Mazzei, and Kenneth E. Jansen. Large Eddy Simulation and the variational multiscale method. *Computing and Visualization in Science*, 3(1-2):47–59, May 2000. ISSN 1432-9360. doi: 10.1007/s007910050051. URL http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s007910050051. (Cited on pages 157 and 158.)

[99] Sergio R. Idelsohn, Juan Carlos Heinrich, and Eugenio Oñate. Petrov–Galerkin methods for the transient advective–diffusive equation with sharp gradients. *International Journal for Numerical Methods in Engineering*, 39(9):1455–1473, May 1996. ISSN 0029-5981. doi: 10.1002/(SICI)1097-0207(19960515)39:9<1455::AID-NME912>3.0.CO;2-0. URL http://doi.wiley.com/10.1002/%28SICI%291097-0207%2819960515%2939%3A9%3C1455%3A%3AAID-NME912%3E3.0.CO%3B2-0. (Cited on pages 38, 57, and 190.)

[100] Sergio R. Idelsohn, N. Nigro, M. Storti, and G. Buscaglia. A Petrov–Galerkin formulation for advection–reaction–diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 136(1-2):27–46, September 1996. ISSN 00457825. doi: 10.1016/0045-7825(96)01008-0. URL http://linkinghub.elsevier.com/retrieve/pii/0045782596010080. (Cited on pages 38 and 39.)

[101] Sergio R. Idelsohn, Eugenio Oñate, and Facundo Del Pin. The particle finite element method: a powerful tool to solve incompressible flows with free-surfaces and breaking waves. *International Journal for Numerical Methods in Engineering*, 61(7):964–989, October 2004. ISSN 0029-5981. doi: 10.1002/nme.1096. URL http://doi.wiley.com/10.1002/nme.1096. (Cited on page 159.)

[102] Sergio R. Idelsohn, Julio Marti, A. Limache, and Eugenio Oñate. Unified Lagrangian formulation for elastic solids and incompressible fluids: Application to fluid–structure interaction problems via the PFEM. *Computer Methods*

in Applied Mechanics and Engineering, 197(19-20):1762–1776, March 2008. ISSN 00457825. doi: 10.1016/j.cma.2007.06.004. URL http://linkinghub.elsevier.com/retrieve/pii/S004578250700237X. (Cited on page 159.)

[103] Sergio R. Idelsohn, Monica de Mier-Torrecilla, Norberto Nigro, and Eugenio Oñate. On the analysis of heterogeneous fluids with jumps in the viscosity using a discontinuous pressure field. *Computational Mechanics*, 46(1):115–124, December 2009. ISSN 0178-7675. doi: 10.1007/s00466-009-0448-6. URL http://www.springerlink.com/index/10.1007/s00466-009-0448-6. (Cited on page 158.)

[104] Sergio R. Idelsohn, Monica de Mier-Torrecilla, and Eugenio Oñate. Multi-fluid flows with the Particle Finite Element Method. *Computer Methods in Applied Mechanics and Engineering*, 198(33-36):2750–2767, July 2009. ISSN 00457825. doi: 10.1016/j.cma.2009.04.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782509001534. (Cited on pages 158 and 170.)

[105] Sergio R. Idelsohn, Facundo Del Pin, Riccardo Rossi, and Eugenio Oñate. Fluid–structure interaction problems with strong added-mass effect. *International Journal for Numerical Methods in Engineering*, 80(10):1261–1294, December 2009. ISSN 00295981. doi: 10.1002/nme.2659. URL http://doi.wiley.com/10.1002/nme.2659. (Cited on pages 159 and 170.)

[106] Frank Ihlenburg. *Finite Element Analysis of Acoustic Scattering*, volume 132 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1998. ISBN 0-387-98319-8. doi: 10.1007/b98828. URL http://www.springerlink.com/index/10.1007/b98828. (Cited on page 1.)

[107] Frank Ihlenburg and Ivo M. Babuska. Finite element solution of the Helmholtz equation with high wave number Part I: The h-version of the FEM. *Computers & Mathematics with Applications*, 30(9):9–37, November 1995. ISSN 08981221. doi: 10.1016/0898-1221(95)00144-N. URL http://linkinghub.elsevier.com/retrieve/pii/089812219500144N. (Cited on page 97.)

[108] Frank Ihlenburg and Ivo M. Babuska. Finite Element Solution of the Helmholtz Equation with High Wave Number Part II: The h-p Version of the FEM. *SIAM Journal on Numerical Analysis*, 34(1):315–358, 1997. ISSN 00361429. doi: 10.1137/S0036142994272337. URL http://link.aip.org/link/SJNAAM/v34/i1/p315/s1&Agg=doi. (Cited on page 97.)

[109] Masatoshi Ikeuchi, Kazuo Inoue, Hideo Sawami, and Hiroshi Niki. Arbitrarily Shaped Hollow Waveguide Analysis by the α-Interpolation Method. *SIAM Journal on Applied Mathematics*, 40(1):90–98, 1981. ISSN 00361399. doi: 10.1137/0140007. URL http://link.aip.org/link/SMJMAP/v40/i1/p90/s1&Agg=doi. (Cited on pages 100 and 144.)

[110] Kazuo Ishihara. Convergence of the Finite Element Method Applied to the Eigenvalue Problem $\Delta u + \lambda u = 0$. *Publications of the Research Institute for Mathematical Sciences*, 13(1):47–60, 1977. ISSN 00345318. URL http://ci.nii.ac.jp/naid/110004714382/en/. (Cited on pages 3, 100, 102, 143, 144, 153, and 190.)

[111] Kenneth E. Jansen, S. Collis, C. Whiting, and Farzin Shakib. A better consistency for low-order stabilized finite element methods. *Computer Methods in Applied Mechanics and Engineering*, 174(1-2):154–170, May 1999. ISSN 00457825. doi: 10.1016/S0045-7825(98)00284-9. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782598002849. (Cited on pages 1, 7, 177, and 191.)

[112] J Jirousek. Basis for development of large finite elements locally satisfying all field equations. *Computer Methods in Applied Mechanics and Engineering*, 14(1): 65–92, April 1978. ISSN 00457825. doi: 10.1016/0045-7825(78)90013-0. URL http://linkinghub.elsevier.com/retrieve/pii/0045782578900130. (Cited on page 99.)

[113] J Jirousek and A. P. Zielinski. Survey of trefftz-type element formulations. *Computers & Structures*, 63(2):225–242, April 1997. ISSN 00457949. doi: 10.1016/S0045-7949(96)00366-5. URL http://linkinghub.elsevier.com/retrieve/pii/S0045794996003665. (Cited on page 99.)

[114] Volker John and Petr Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review. *Computer Methods in Applied Mechanics and Engineering*, 196(17-20):2197–2215, March 2007. ISSN 00457825. doi: 10.1016/j.cma.2006.11.013. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782506003926. (Cited on pages 37 and 42.)

[115] P.B. Johns and R.L. Beurle. Numerical solution of 2-dimensional scattering problems using a transmission-line matrix. *Proceedings of the Institution of Electrical Engineers*, 118(9):1203–1208, 1971. ISSN 00203270. doi: 10.1049/piee.1971.0217. URL http://link.aip.org/link/PIEEAH/v118/i9/p1203/s1&Agg=doi. (Cited on page 110.)

[116] Claes Johnson, Anders Szepessy, and Peter Hansbo. On the convergence of shock-capturing streamline diffusion finite element methods for hyperbolic conservation laws. *Mathematics of Computation*, 54(189):107–129, 1990. URL http://www.jstor.org/stable/2008684. (Cited on page 37.)

[117] Jirair Kevorkian. Matched Asymptotic Expansions. In *Partial Differential Equations: Analytical Solution Techniques*, chapter 8.2. Springer–Verlag, second edition, 1999. (Cited on page 71.)

[118] F. Kikuchi and T. Ushijima. Theoretical analysis of some finite element schemes for convective diffusion equations. In R.H. Gallagher, D.M. Norrie, J.T. Oden, and Olgierd C. Zienkiewicz, editors, *Finite Elements in Fluids*, volume IV. John Wiley and Sons Ltd, Chichester, 1982. (Cited on page 80.)

[119] Petr Knobloch. Improvements of the Mizukami–Hughes method for convection–diffusion equations. *Computer Methods in Applied Mechanics and Engineering*, 196(1-3):579–594, December 2006. ISSN 00457825. doi: 10.1016/j.cma.2006.06.004. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782506001964. (Cited on page 37.)

[120] Dmitri Kuzmin, Rainald Löhner, and Stefan Turek. *Flux-corrected transport: principles, algorithms and applications*. Springer, 2005. (Cited on pages 38, 44, and 190.)

[121] O Laghrouche and M Mohamed. Locally enriched finite elements for the Helmholtz equation in two dimensions. *Computers & Structures*, May 2008. ISSN 00457949. doi: 10.1016/j.compstruc.2008.04.006. URL http://linkinghub.elsevier.com/retrieve/pii/S0045794908000989. (Cited on page 98.)

[122] Larry A. Lambe, Richard Luczak, and John W. Nehrbass. A new finite difference method for the Helmholtz equation using symbolic computation. *International Journal of Computational Engineering Science*, 4(1):121–144, 2003. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.7268. (Cited on page 99.)

[123] C B Laney. *Computational gas dynamics*. Cambridge University Press, 1998. (Cited on page 38.)

[124] Randall J LeVeque. *Numerical methods for conservation laws*. Birkhauser-Verlag, Basel, 1990. (Cited on page 38.)

[125] Randall J. LeVeque. High-Resolution Conservative Algorithms for Advection in Incompressible Flow. *SIAM Journal on Numerical Analysis*, 33(2):627–665, 1996. ISSN 00361429. doi: 10.1137/0733033. URL http://link.aip.org/link/SJNAAM/v33/i2/p627/s1&Agg=doi. (Cited on page 83.)

[126] Z.C. LI. The Trefftz method for the Helmholtz equation with degeneracy. *Applied Numerical Mathematics*, 58(2):131–159, February 2008. ISSN 01689274. doi: 10.1016/j.apnum.2006.11.004. URL http://linkinghub.elsevier.com/retrieve/pii/S0168927406002108. (Cited on pages 99 and 100.)

[127] Rainald Löhner, K. Morgan, and Olgierd C. Zienkiewicz. The solution of non-linear hyperbolic equation systems by the finite element method. *International Journal for Numerical Methods in Fluids*, 4(11):1043–1063, November 1984. ISSN 0271-2091. doi: 10.1002/fld.1650041105. URL http://doi.wiley.com/10.1002/fld.1650041105. (Cited on page 37.)

[128] Abimael F. D. Loula and Daniel T. Fernandes. A quasi optimal Petrov–Galerkin method for Helmholtz problem. *International Journal for Numerical Methods in Engineering*, 80(12):1595–1622, December 2009. ISSN 00295981. doi: 10.1002/nme.2677. URL http://doi.wiley.com/10.1002/nme.2677. (Cited on pages 77, 99, 144, 154, and 191.)

[129] Abimael F. D. Loula, Gustavo Benitez Alvarez, Eduardo Gomes Dutra do Carmo, and Fernando Alves Rochinha. A discontinuous finite element method at element level for Helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 196(4-6):867–878, January 2007. ISSN 00457825. doi: 10.1016/j.cma.2006.07.008. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782506002192. (Cited on pages 99 and 128.)

[130] S. P. Marin. *A Finite Element Method for Problems Involving the Helmholtz Equation in Two Dimensional Exterior Regions*. PhD thesis, Carnegie-Mellon University, Pittsburgh, PA, 1978. (Cited on page 97.)

[131] J. M. Melenk and Ivo M. Babuska. The partition of unity finite element method: Basic theory and applications. *Computer Methods in Applied Mechanics and Engineering*, 139(1-4):289–314, December 1996. ISSN 00457825. doi: 10.1016/S0045-7825(96)01087-0. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782596010870. (Cited on page 98.)

[132] Akira Mizukami and Thomas J. R. Hughes. A Petrov–Galerkin finite element method for convection-dominated flows: An accurate upwinding technique for satisfying the maximum principle. *Computer Methods in Applied Mechanics and Engineering*, 50(2):181–193, August 1985. ISSN 00457825. doi: 10.1016/0045-7825(85)90089-1. URL http://linkinghub.elsevier.com/retrieve/pii/0045782585900891. (Cited on pages 37, 41, and 81.)

[133] P. Monk and D. Wang. A least-squares method for the Helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 175(1-2):121–136, June 1999. ISSN 00457825. doi: 10.1016/S0045-7825(98)00326-0. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782598003260. (Cited on page 99.)

[134] Keith William Morton. *Numerical solution of convection–diffusion problems*, volume 12 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall, London, 1996. ISBN 0412564408. (Cited on page 1.)

[135] Majid Nabavi, M.H. Kamran Siddiqui, and Javad Dargahi. A new 9-point sixth-order accurate compact finite-difference method for the Helmholtz equation. *Journal of Sound and Vibration*, 307(3-5):972–982, November 2007. ISSN 0022460X. doi: 10.1016/j.jsv.2007.06.070. URL http://linkinghub.elsevier.com/retrieve/pii/S0022460X07004877. (Cited on page 99.)

[136] Prashanth Nadukandi, Eugenio Oñate, and Julio García. A high-resolution Petrov–Galerkin method for the 1D convection–diffusion–reaction problem. *Computer Methods in Applied Mechanics and Engineering*, 199(9-12):525–546, January 2010. ISSN 00457825. doi: 10.1016/j.cma.2009.10.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782509003545. (Cited on page 80.)

[137] Prashanth Nadukandi, Eugenio Oñate, and Julio Garcia. A fourth-order compact scheme for the Helmholtz equation: Alpha-interpolation of FEM and FDM stencils. *International Journal for Numerical Methods in Engineering*, 86(1):18–46, April 2011. ISSN 00295981. doi: 10.1002/nme.3043. URL http://doi.wiley.com/10.1002/nme.3043. (Cited on pages 69, 129, 141, 142, 144, 145, 147, 150, 151, and 154.)

[138] Prashanth Nadukandi, Eugenio Oñate, and Julio García. A Petrov-Galerkin formulation for the alpha interpolation of FEM and FDM stencils: Applications to the Helmholtz equation. *International Journal for Numerical Methods in Engineering*, 89(11):1367–1391, November 2012. ISSN 00295981. doi: 10.1002/nme.3291. URL http://doi.wiley.com/10.1002/nme.3291. (Cited on pages 115 and 128.)

[139] Mitsunobu Nakamura and Masahide Hirasawa. Eigenvalues of the Schrödinger Equation by the α-Interpolation Method. *SIAM Journal on Applied Mathematics*, 43(6):1286–1293, 1983. ISSN 00361399. doi: 10.1137/0143086. URL http://link.aip.org/link/SMJMAP/v43/i6/p1286/s1&Agg=doi. (Cited on pages 100 and 144.)

[140] Hiroshi Niki, Hideo Sawami, Masatoshi Ikeuchi, and Naotaka Okamoto. The alpha interpolation method for the solution of an eigenvalue problem. *Journal of Computational and Applied Mathematics*, 8(1):15–19, March 1982. ISSN 03770427. doi: 10.1016/0771-050X(82)90002-X. URL http://linkinghub.elsevier.com/retrieve/pii/0771050X8290002X. (Cited on pages iii, iv, 100, 143, 144, 153, and 190.)

[141] Eugenio Oñate. Derivation of stabilized equations for numerical solution of advective–diffusive transport and fluid flow problems. *Computer Methods in Applied Mechanics and Engineering*, 151(1-2):233–265, January 1998. ISSN 00457825. doi: 10.1016/S0045-7825(97)00119-9. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782597001199. (Cited on pages iii, iv, 7, 37, 39, 41, 42, 43, 157, 158, 159, 161, 163, 168, and 189.)

[142] Eugenio Oñate. A stabilized finite element method for incompressible viscous flows using a finite increment calculus formulation. *Computer Methods in Applied Mechanics and Engineering*, 182(3-4):355–370, February 2000. ISSN 00457825. doi: 10.1016/S0045-7825(99)00198-X. URL http://linkinghub.elsevier.com/retrieve/pii/S004578259900198X. (Cited on pages 157, 158, 168, and 169.)

[143] Eugenio Oñate. Possibilities of finite calculus in computational mechanics. *International Journal for Numerical Methods in Engineering*, 60(1):255–281, May 2004. ISSN 0029-5981. doi: 10.1002/nme.961. URL http://doi.wiley.com/10.1002/nme.961. (Cited on pages 7 and 161.)

[144] Eugenio Oñate and Julio García. A finite element method for fluid–structure interaction with surface waves using a finite calculus formulation. *Computer Methods in Applied Mechanics and Engineering*, 191(6-7):635–660, December 2001. ISSN 00457825. doi: 10.1016/S0045-7825(01)00306-1. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782501003061. (Cited on page 159.)

[145] Eugenio Oñate and M. Manzán. A general procedure for deriving stabilized space-time finite element methods for advective–diffusive problems. *International Journal for Numerical Methods in Fluids*, 31(1):203–221, September 1999. ISSN 0271-2091. doi: 10.1002/(SICI)1097-0363(19990915)31:1<203::AID-FLD964>3.0.CO;2-Z. URL http://doi.wiley.com/10.1002/%28SICI%291097-0363%2819990915%2931%3A1%3C203%3A%3AAID-FLD964%3E3.0.CO%3B2-Z. (Cited on pages 7 and 9.)

[146] Eugenio Oñate and M. Manzán. Stabilization techniques for finite element analysis of convection–diffusion problems. In B. Sundén and G. Comini, editors, *Computational Analysis of Convection Heat Transfer*, pages 71–117. WIT Press, Southampton, UK, 2000. ISBN 1853127345. (Cited on page 37.)

[147] Eugenio Oñate, Julio García, and Sergio R. Idelsohn. Computation of the stabilization parameter for the finite element solution of advective–diffusive problems. *International Journal for Numerical Methods in Fluids*, 25(12):1385–1407, December 1997. ISSN 0271-2091. doi: 10.1002/(SICI)1097-0363(19971230)25:12<1385::AID-FLD678>3.0.CO;2-7. URL http://doi.wiley.com/10.1002/%28SICI%291097-0363%2819971230%2925%3A12%3C1385%3A%3AAID-FLD678%3E3.0.CO%3B2-7. (Cited on page 7.)

[148] Eugenio Oñate, Robert L. Taylor, Olgierd C. Zienkiewicz, and J. Rojek. A residual correction method based on finite calculus. *Engineering Computations*, 20(5/6):629–658, 2003. ISSN 0264-4401. doi: 10.1108/02644400310488790. URL http://www.emeraldinsight.com/10.1108/02644400310488790. (Cited on pages 1, 7, 159, and 169.)

[149] Eugenio Oñate, Sergio R. Idelsohn, Facundo Del Pin, and Romain Aubry. The particle finite element method. An overview. *International Journal of Computational Methods*, 1(2):267–307, 2004. doi: 10.1142/S0219876204000204. URL http://www.worldscinet.com/ijcm/01/0102/S0219876204000204.html. (Cited on page 159.)

[150] Eugenio Oñate, Jerzy Rojek, Robert L. Taylor, and Olgierd C. Zienkiewicz. Finite calculus formulation for incompressible solids using linear triangles and tetrahedra. *International Journal for Numerical Methods in Engineering*, 59 (11):1473–1500, March 2004. ISSN 0029-5981. doi: 10.1002/nme.922. URL http://doi.wiley.com/10.1002/nme.922. (Cited on page 159.)

[151] Eugenio Oñate, Julio García, Sergio R. Idelsohn, and Facundo Del Pin. Finite calculus formulations for finite element analysis of incompressible flows. Eulerian, ALE and Lagrangian approaches. *Computer Methods in Applied Mechanics and Engineering*, 195(23-24):3001–3037, April 2006. ISSN 00457825. doi: 10.1016/j.cma.2004.10.016. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782505002379. (Cited on page 159.)

[152] Eugenio Oñate, Juan Miquel, and Guillermo Hauke. Stabilized formulation for the advection–diffusion–absorption equation using finite calculus and linear finite elements. *Computer Methods in Applied Mechanics and Engineering*, 195(33-36):3926–3946, July 2006. ISSN 00457825. doi: 10.1016/j.cma.2005.07.020. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782505003877. (Cited on pages 38, 39, 43, 54, 56, 59, 60, 61, 62, 63, 64, 65, 79, and 80.)

[153] Eugenio Oñate, Aleix Valls, and Julio García. FIC/FEM Formulation with Matrix Stabilizing Terms for Incompressible Flows at Low and High Reynolds Numbers. *Computational Mechanics*, 38(4-5):440–455, April 2006. ISSN 0178-7675. doi: 10.1007/s00466-006-0060-y. URL http://www.springerlink.com/index/10.1007/s00466-006-0060-y. (Cited on page 159.)

[154] Eugenio Oñate, Fransisco Zarate, and Sergio R. Idelsohn. Finite element formulation for convective–diffusive problems with sharp gradients using finite calculus. *Computer Methods in Applied Mechanics and Engineering*, 195(13-16): 1793–1825, February 2006. ISSN 00457825. doi: 10.1016/j.cma.2005.05.036. URL

http://linkinghub.elsevier.com/retrieve/pii/S0045782505003075. (Cited on pages 12, 37, and 43.)

[155] Eugenio Oñate, Juan Miquel, and Fransisco Zarate. Stabilized solution of the multidimensional advection–diffusion–absorption equation using linear finite elements. *Computers & Fluids*, 36(1):92–112, January 2007. ISSN 00457930. doi: 10.1016/j.compfluid.2005.07.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0045793005001313. (Cited on pages 38, 39, and 43.)

[156] Eugenio Oñate, Aleix Valls, and Julio García. Modeling incompressible flows at low and high Reynolds numbers via a finite calculus–finite element approach. *Journal of Computational Physics*, 224(1):332–351, May 2007. ISSN 00219991. doi: 10.1016/j.jcp.2007.02.026. URL http://linkinghub.elsevier.com/retrieve/pii/S0021999107000794. (Cited on page 159.)

[157] Eugenio Oñate, Sergio R. Idelsohn, Miguel A. Celigueta, and Riccardo Rossi. Advances in the particle finite element method for the analysis of fluid–multibody interaction and bed erosion in free surface flows. *Computer Methods in Applied Mechanics and Engineering*, 197(19-20):1777–1800, March 2008. ISSN 00457825. doi: 10.1016/j.cma.2007.06.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782507002368. (Cited on page 159.)

[158] Eugenio Oñate, Sergio R. Idelsohn, and Carlos A. Felippa. Consistent pressure Laplacian stabilization for incompressible continua via higher-order finite calculus. *International Journal for Numerical Methods in Engineering*, pages n/a–n/a, September 2010. ISSN 00295981. doi: 10.1002/nme.3021. URL http://doi.wiley.com/10.1002/nme.3021. (Cited on pages 158, 159, 163, 168, 169, 178, and 184.)

[159] Assad A. Oberai and Peter M. Pinsky. A multiscale finite element method for the Helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 154(3-4):281–297, March 1998. ISSN 00457825. doi: 10.1016/S0045-7825(97)00130-8. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782597001308. (Cited on page 98.)

[160] Assad A. Oberai and Peter M. Pinsky. A residual-based finite element method for the Helmholtz equation. *International Journal for Numerical Methods in Engineering*, 49(3):399–419, September 2000. ISSN 0029-5981. doi: 10.1002/1097-0207(20000930)49:3<399::AID-NME844>3.0.CO;2-5. URL http://doi.wiley.com/10.1002/1097-0207(20000930)49:3<399::AID-NME844>3.0.CO;2-5. (Cited on pages 98 and 128.)

[161] J. T. Oden. Historical comments on finite elements. In *Proceedings of the ACM conference on history of scientific and numeric computation*, pages 125–130, Princeton, New Jersey USA, 1987. (Cited on page 38.)

[162] Milton E. Rose. Weak-element approximations to elliptic differential equations. *Numerische Mathematik*, 24(3):185–204, June 1975. ISSN 0029-599X. doi: 10.1007/BF01436591. URL http://www.springerlink.com/index/10.1007/BF01436591. (Cited on page 99.)

[163] Tapan K Sengupta, Gaurav Ganeriwal, and S. De. Analysis of central and upwind compact schemes. *Journal of Computational Physics*, 192(2):677–694, December 2003. ISSN 00219991. doi: 10.1016/j.jcp.2003.07.015. URL http://linkinghub.elsevier.com/retrieve/pii/S0021999103004133. (Cited on page 27.)

[164] Tapan K Sengupta, Gaurav Ganeriwal, and Anurag Dipankar. High Accuracy Compact Schemes and Gibbs' Phenomenon. *Journal of Scientific Computing*, 21(3):253–268, December 2004. ISSN 0885-7474. doi: 10.1007/s10915-004-1317-2. URL http://www.springerlink.com/openurl.asp?id=doi:10.1007/s10915-004-1317-2. (Cited on page 27.)

[165] Tony W. H. Sheu, C. F. Chen, and L. W. Hsieh. A highly accurate Helmholtz scheme for modeling scattering wave. *Computer Methods in Applied Mechanics and Engineering*, 193(52):5573–5583, December 2004. ISSN 00457825. doi: 10.1016/j.cma.2003.08.013. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782504002907. (Cited on page 99.)

[166] C. Shu and H. Xue. Solution of Helmholtz equation by differential quadrature method. *Computer Methods in Applied Mechanics and Engineering*, 175(1-2):203–212, June 1999. ISSN 00457825. doi: 10.1016/S0045-7825(98)00370-3. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782598003703. (Cited on page 99.)

[167] N. R. S. Simons and A. A. Sebak. Spatially weighted numerical models for the two-dimensional wave equation: FD algorithm and synthesis of the equivalent TLM model. *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, 6(1):47–65, February 1993. ISSN 0894-3370. doi: 10.1002/jnm.1660060106. URL http://doi.wiley.com/10.1002/jnm.1660060106. (Cited on page 110.)

[168] I Singer and Eli Turkel. High-order finite difference methods for the Helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 163(1-4):343–358, September 1998. ISSN 00457825. doi: 10.1016/S0045-7825(98)00023-1. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782598000231. (Cited on pages iii, iv, 99, 100, 103, 109, 141, 145, and 190.)

[169] I Singer and Eli Turkel. Sixth order accurate finite difference schemes for the Helmholtz equation. *Journal of Computational Acoustics*, 14(3):339–351, 2006. (Cited on page 99.)

[170] Malgorzata Stojek. *Finite T-elements for the Poisson and Helmholtz equations*. PhD thesis, École Polytechnique Fédérale de Lausanne, Lausanne, 1996. (Cited on page 99.)

[171] Mario Storti, Norberto Nigro, and Sergio R. Idelsohn. Steady state incompressible flows using explicit schemes with an optimal local preconditioning. *Computer Methods in Applied Mechanics and Engineering*, 124(3):231–252, July 1995. ISSN 00457825. doi: 10.1016/0045-7825(95)00787-2. URL http://linkinghub.elsevier.com/retrieve/pii/0045782595007872. (Cited on pages 157 and 158.)

[172] Gilbert Strang. Diffusion, Convection, and Finance. In *Computational Science and Engineering*, chapter 6.5. Wellesley-Cambridge Press, first edition, 2007. URL http://math.mit.edu/cse/. (Cited on page 72.)

[173] T Strouboulis, Ivo M. Babuska, and R Hidajat. The generalized finite element method for Helmholtz equation: Theory, computation, and open problems. *Computer Methods in Applied Mechanics and Engineering*, 195(37-40):4711–4731, July 2006. ISSN 00457825. doi: 10.1016/j.cma.2005.09.019. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782505005037. (Cited on page 98.)

[174] T Strouboulis, R Hidajat, and Ivo M. Babuska. The generalized finite element method for Helmholtz equation. Part II: Effect of choice of handbook functions, error due to absorbing boundary conditions and its assessment. *Computer Methods in Applied Mechanics and Engineering*, 197(5):364–380, January 2008. ISSN 00457825. doi: 10.1016/j.cma.2007.05.019. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782507002289. (Cited on page 98.)

[175] Martin Stynes. Steady-state convection-diffusion problems. *Acta Numerica*, 14: 445–508, May 2005. ISSN 0962-4929. doi: 10.1017/S0962492904000261. URL http://www.journals.cambridge.org/abstract_S0962492904000261. (Cited on pages 69, 71, and 72.)

[176] Godehard Sutmann. Compact finite difference schemes of sixth order for the Helmholtz equation. *Journal of Computational and Applied Mathematics*, 203 (1):15–31, June 2007. ISSN 03770427. doi: 10.1016/j.cam.2006.03.008. URL http://linkinghub.elsevier.com/retrieve/pii/S0377042706001622. (Cited on page 99.)

[177] Tayfun E. Tezduyar. Computation of moving boundaries and interfaces and stabilization parameters. *International Journal for Numerical Methods in Fluids*, 43 (5):555–575, October 2003. ISSN 0271-2091. doi: 10.1002/fld.505. URL http://doi.wiley.com/10.1002/fld.505. (Cited on pages 157, 158, and 168.)

[178] Tayfun E. Tezduyar and Yasuo Osawa. Finite element stabilization parameters computed from element matrices and vectors. *Computer Methods in Applied Mechanics and Engineering*, 190(31):411–430, April 2001. ISSN 00457825. doi: 10.1016/S0045-7825(00)00211-5. URL http://linkinghub.elsevier.com/retrieve/pii/S0045782500002115.

[179] Tayfun E. Tezduyar, Sanjay Mittal, S. E. Ray, and Randy H Shih. Incompressible flow computations with stabilized bilinear and linear equal-order-interpolation velocity-pressure elements. *Computer Methods in Applied Mechanics and Engineering*, 95(2):221–242, March 1992. ISSN 00457825. doi: 10.1016/0045-7825(92)90141-6. URL http://linkinghub.elsevier.com/retrieve/pii/0045782592901416. (Cited on pages 157 and 158.)

[180] T.E. Tezduyar and Y.J. Park. Discontinuity-capturing finite element formulations for nonlinear convection–diffusion–reaction equations. *Computer Methods in Applied Mechanics and Engineering*, 59(3):307–325, December 1986. ISSN 00457825. doi: 10.1016/0045-7825(86)90003-4. URL http://linkinghub.elsevier.com/retrieve/pii/0045782586900034. (Cited on page 38.)

[181] Lonny L. Thompson and Peter M. Pinsky. A Galerkin least-squares finite element method for the two-dimensional Helmholtz equation. *International Journal for Numerical Methods in Engineering*, 38(3):371–397, February 1995. ISSN 0029-5981. doi: 10.1002/nme.1620380303. URL http://doi.wiley.com/10.1002/nme.1620380303. (Cited on pages 98 and 144.)

[182] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics*. Springer-Verlag, 1999. (Cited on page 38.)

[183] Lloyd N. Trefethen. Group Velocity in Finite Difference Schemes. *SIAM Review*, 24(2):113, 1982. ISSN 00361445. doi: 10.1137/1024038. URL http://link.aip.org/link/SIREAD/v24/i2/p113/s1&Agg=doi. (Cited on pages 8, 10, 104, and 110.)

[184] Igor Tsukerman. A class of difference schemes with flexible local approximation. *Journal of Computational Physics*, 211(2):659–699, January 2006. ISSN 00219991. doi: 10.1016/j.jcp.2005.06.011. URL http://linkinghub.elsevier.com/retrieve/pii/S0021999105002925. (Cited on page 99.)

[185] Bram van Leer. Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method. *Journal of Computational Physics*, 32 (1):101–136, 1979. ISSN 00219991. doi: 10.1016/0021-9991(79)90145-1. URL http://linkinghub.elsevier.com/retrieve/pii/0021999179901451. (Cited on page 38.)

[186] Richard S. Varga. On a Discrete Maximum Principle. *SIAM Journal on Numerical Analysis*, 3(2):355, 1966. ISSN 00361429. doi: 10.1137/0703029. URL http://link.aip.org/link/SJNAAM/v3/i2/p355/s1&Agg=doi. (Cited on pages 29 and 189.)

[187] Robert Vichnevetsky and John B Bowles. *Fourier Analysis of Numerical Approximations of Hyperbolic Equations*, volume 5. SIAM, Philadelphia, PA, studies in edition, 1982. (Cited on pages 110 and 190.)

[188] Jinchao Xu and Ludmil Zikatanov. A monotone finite element scheme for convection-diffusion equations. *Mathematics of Computation*, 68(228):1429–1447, May 1999. ISSN 0025-5718. doi: 10.1090/S0025-5718-99-01148-5. URL http://www.ams.org/journal-getitem?pii=S0025-5718-99-01148-5. (Cited on page 81.)

[189] C.-C. Yu and Juan Carlos Heinrich. Petrov–Galerkin methods for the time-dependent convective transport equation. *International Journal for Numerical Methods in Engineering*, 23(5):883–901, May 1986. ISSN 0029-5981. doi: 10.1002/nme.1620230510. URL http://doi.wiley.com/10.1002/nme.1620230510. (Cited on page 38.)

[190] C.-C. Yu and Juan Carlos Heinrich. Petrov–Galerkin method for multidimensional, time-dependent, convective-diffusion equations. *International Journal for Numerical Methods in Engineering*, 24(11):2201–2215, November 1987. ISSN 0029-5981. doi: 10.1002/nme.1620241112. URL http://doi.wiley.com/10.1002/nme.1620241112. (Cited on page 76.)

[191] Steven T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of Computational Physics*, 31(3):335–362, June 1979. ISSN 00219991. doi: 10.1016/0021-9991(79)90051-2. URL http://linkinghub.elsevier.com/retrieve/pii/0021999179900512. (Cited on pages 38 and 84.)

[192] A. P. Zielinski and Olgierd C. Zienkiewicz. Generalized finite element analysis with T-complete boundary solution functions. *International Journal for Numerical Methods in Engineering*, 21(3):509–528, March 1985. ISSN 0029-5981. doi: 10.1002/nme.1620210310. URL http://doi.wiley.com/10.1002/nme.1620210310. (Cited on page 99.)

[193] Olgierd C. Zienkiewicz and Ramon Codina. A general algorithm for compressible and incompressible flow—Part I. the split, characteristic-based scheme. *International Journal for Numerical Methods in Fluids*, 20(8-9):869–885, April 1995. ISSN 0271-2091. doi: 10.1002/fld.1650200812. URL http://doi.wiley.com/10.1002/fld.1650200812. (Cited on page 37.)

[194] Olgierd C. Zienkiewicz and Robert L. Taylor. *The Finite Element Method for Solid and Structural Mechanics*, volume 2. Butterworth-Heinemann, Oxford, 6 edition, 2005. ISBN 0-7506-6321-9. (Cited on pages 159, 164, and 165.)

[195] Olgierd C. Zienkiewicz, Robert L. Taylor, and Perumal Nithiarasu. *The finite element method for fluid dynamics*. Elsevier Butterworth-Heinemann, 2005. (Cited on pages 55, 157, 158, 164, 165, 166, 168, 169, and 176.)

COLOPHON

This monograph was typeset with LaTeX 2ε using André Miede's style package available via CTAN as "classicthesis". The package uses Hermann Zapf's *Palatino* and *Euler* type faces (Type 1 PostScript fonts *URW Palladio L* and *FPL* were used). The typographic style follows Bringhurst's suggestions as presented in his book—*The Elements of Typographic Style*.